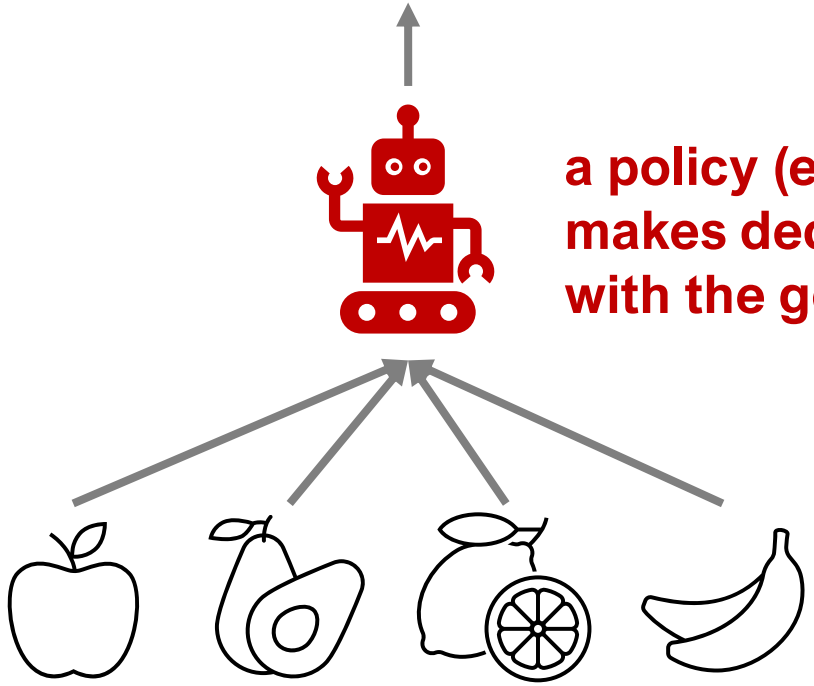
Evaluating the Robustness of Off-Policy Evaluation

Yuta Saito, Takuma Udagawa, **Haruka Kiyohara**,
Kazuki Mogi, Yusuke Narita, Kei Tateno

Haruka Kiyohara, Tokyo Institute of Technology
https://sites.google.com/view/harukakiyohara

Machine decision making in recommenders



a policy (e.g., contextual bandit)
makes decisions to recommend items,
with the goal of maximizing the (expected) reward

The system also produces logged data

$$\mathcal{D} := \{(x_i, a_i, r_i)\}_{i=1}^{n}$$

logged bandit feedback collected by a *behavior* policy π_b

$$x_i \sim p(x) \quad a_i \sim \pi_b(a \mid x_i)$$

$$r_i \sim p(r \mid x_i, a_i)$$

Motivation:
We want to evaluate the future policies using the logged data.

Off-Policy Evaluation (OPE)

In OPE, we aim to evaluate the performance of a new *evaluation* policy π_e using logged bandit feedback collected by the *behavior* policy π_b. **distribution shift**

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

hyperparameters of the OPE estimator $\hat{V}$

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

expected reward obtained by running on π_e the real system

Off-Policy Evaluation (OPE)

In OPE, we aim to evaluate the performance of a new *evaluation* policy π_e using logged bandit feedback collected by the *behavior* policy π_b. **distribution shift**

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

hyperparameters of the OPE estimator $\hat{V}$

An accurate OPE is beneficial, because it..

- avoids deploying poor policies without A/B tests
- identifies promising new policies among many candidates

Growing interest in OPE!

Inverse Provability Weighting (IPW) [Strehl+, 2010]

IPW mitigates the distribution shift between π_b and π_e using importance sampling.

$$\hat{V}_{\mathrm{IPW}}(\pi_e; \mathcal{D}) := \mathbb{E}_n[\rho(x_i, a_i) r_i]$$

where $\rho(x, a) := \pi_e(a \mid x) / \pi_b(a \mid x)$

$\mathbb{E}_n[\cdot]$: empirical average

Unbiased*, but large variance. *when π_b is known or accurately estimated

Hyperparameter: $\hat{\pi}_b$ (when π_b is unknown)

Doubly Robust (DR) [Dudík+, 2014]

DR tackles the variance of IPW by leveraging baseline estimation $\hat{q}$ and performing importance weighting only on its residual.

$$\hat{V}_{\mathrm{DR}}(\pi_e; \mathcal{D}, \hat{q}) := \mathbb{E}_n[\mathbb{E}_{a\sim\pi_e(a|x)}[\hat{q}(x_i, a)] + \rho(x_i, a_i)(r_i - \hat{q}(x_i, a_i))]$$

baseline **importance weighting on the residual**

where $\mathbb{E}[r \mid x, a] \approx \hat{q}(x, a)$

Unbiased* and lower variance than IPW. *when π_b is known or accurately estimated

Hyperparameter: $\hat{\pi}_b$ (when π_b is unknown) + $\hat{q}$

Pessimistic Shrinkage (IPWps, DRps) [Su+, 2020]

IPWps and DRps further reduce the variance by clipping large importance weights.

$$\hat{V}_{\mathrm{IPWps}}(\pi_e; \mathcal{D}, \lambda) := \mathbb{E}_n[\min\{\rho(x_i, a_i), \lambda\} r_i]$$

clipped importance weight

$$\hat{V}_{\mathrm{DRps}}(\pi_e; \mathcal{D}, \hat{q}, \lambda) := \mathbb{E}_n[\mathbb{E}_{a\sim\pi_e(a|x)}[\hat{q}(x_i, a)] + \min\{\rho(x_i, a_i), \lambda\}(r_i - \hat{q}(x_i, a_i))]$$

Lower variance than IPW / DR.

Hyperparameter: $\hat{\pi}_b$ (when π_b is unknown) (, $\hat{q}$) + λ

Many estimators with different hyperparameters

OPE Estimators	Hyperparameters
Direct Method (DM)	$\hat{q}, K$
Inverse Probability Weighting with Pessimistic Shrinkage (IPWps) [Strehl+, 2010] [Su+, 2020]	$\lambda, (\hat{\pi}_b)$
Self-Normalized Inverse Probability Weighting (SNIPW) [Swaminathan & Joachims, 2015]	$(\hat{\pi}_b)$
Doubly Robust with Pessimistic Shrinkage (DRps) [Dudík+, 2014] [Su+, 2020]	$\hat{q}, K, \lambda, (\hat{\pi}_b)$
Self-Normalized Doubly Robust (SNDR)	$\hat{q}, K, (\hat{\pi}_b)$
Switch Doubly Robust (Switch-DR) [Wang+, 2017]	$\hat{q}, K, \tau, (\hat{\pi}_b)$
Doubly Robust with Optimistic Shrinkage (DRos) [Su+, 2020]	$\hat{q}, K, \lambda, (\hat{\pi}_b)$

Note: $\hat{q}$ is an estimator for the mean reward function constructed by an arbitrary machine learning method. K is the number of folds in the cross-fitting procedure [Narita+, 2021]. $\hat{\pi}_b$ is an estimated behavior policy. This is unnecessary when we know the true behavior policy, and thus it is in parentheses. τ and λ are non-negative hyperparameters for defining the corresponding estimators.

Estimator Selection:
Which OPE estimator (and hyperparameters) should be used in practice?

What properties are desirable in practice?

- An estimator that works without significant hyperparameter tuning.
  .. because hyperparameters may depend on the logged data and evaluation policy, which might also entail risks for overfitting.

- An estimator that is stably accurate across various evaluation policies.
  .. because we need to evaluate various candidate policies to choose from.

- An estimator that shows acceptable errors in the worst case.
  .. because uncertainty of estimation is of great interest.

We want to evaluate the estimators' robustness to the possible changes in configurations such as hyperparameters and evaluation policies!

Is conventional evaluation sufficient?

Conventional OPE experiments compare *mean-squared-error* to evaluate the performance (estimation accuracy) of OPE estimators.

$$\mathrm{MSE}(\hat{V}; \pi_e, \theta) := \mathbb{E}_{\mathcal{D}} \left[\left(V(\pi_e) - \hat{V}(\pi_e; \mathcal{D}, \theta) \right)^2 \right]$$

evaluate only on a single set of configurations

Pitfall: fails to evaluate the estimators' robustness for configuration changes..
(such as hyperparameters θ and evaluation policy π_e)

Towards more informative evaluation for practice

To tackle the issues in conventional experimental procedure, we propose *Interpretable evaluation for offline evaluation (IEOE)*, which can..

- ✓ evaluate the estimators' robustness to the possible configuration changes
- ✓ provide a visual interpretation of the distribution of estimation errors
- ✓ be easily implemented using our open-source Python software, *pyIEOE*

Interpretable evaluation for offline evaluation (IEOE)

① set configurations spaces (hyperparameters θ and evaluation policies π_e)

② for each random seed s, sample configurations

③ calculate the estimators' squared error on the sampled configurations

④ obtain an error distribution

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 $\mathcal{S}$ ①

Output: empirical CDF, $\hat{F}_Z$, of the squared error (SE)

$\mathcal{Z} \leftarrow \emptyset$
for $s \in \mathcal{S}$ **do**
  $\theta \leftarrow \phi(\Theta; s)$
  $\pi_e \leftarrow \mathrm{Unif}(\Pi_e; s)$
  $\mathcal{D}^* \leftarrow \mathrm{Bootstrap}(\mathcal{D}; s)$
  ② (the three lines above)
  $z' \leftarrow \mathrm{SE}(\hat{V}; \mathcal{D}^*, \pi_e, \theta)$ — ③
  $\mathcal{Z} \leftarrow \mathcal{Z} \cup \{z'\}$
end for
Estimate F_Z using $\mathcal{Z}$ (by Eq. 1) — ④

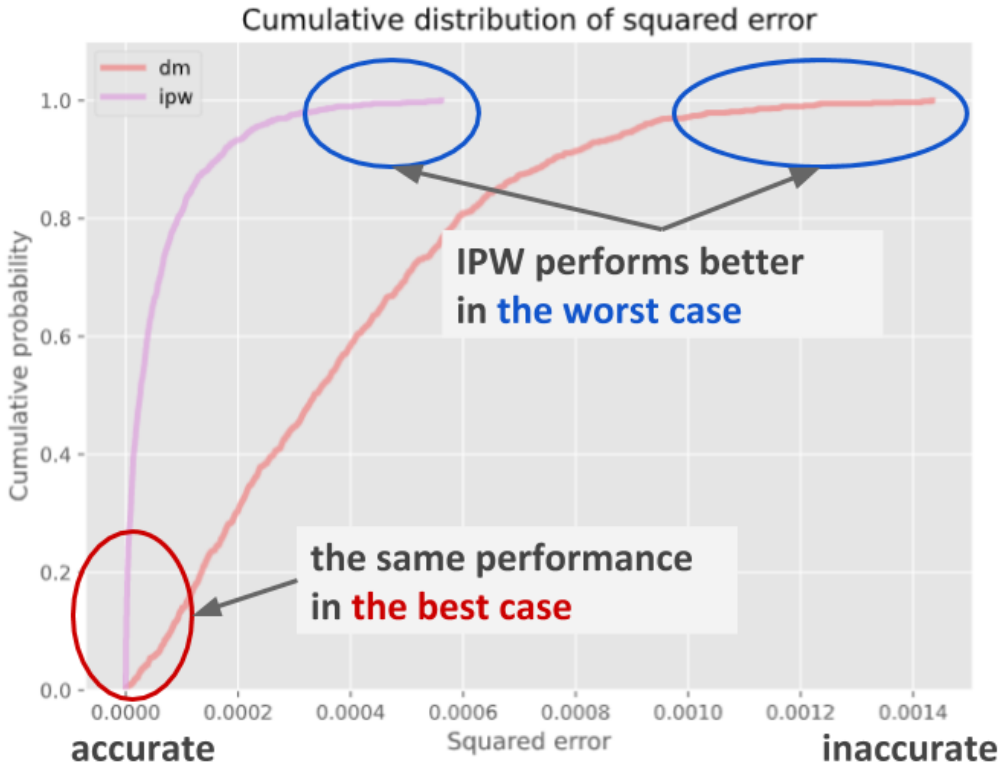
Visual comparison of OPE estimators

After gaining squared errors, we approximate *cumulative distribution function (CDF)*.

$$\hat{F}_Z(z) := \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\{z_i \leq z\} \ (\approx \mathbb{P}(Z \leq z))$$

$$\mathcal{Z} := \{z_1, \ldots, z_m\}$$

We can interpret how the estimators are robust across the given configurations.

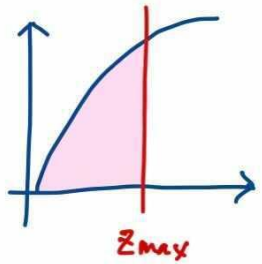


Quantitative performance measure

Based on the CDF, we can define some summary scores, which are useful for quantitative performance comparisons.

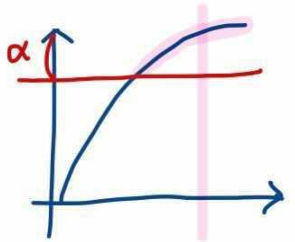
- *Area under the curve (AU-CDF)* compares the estimators' squared errors below the threshold.

$$\text{AU-CDF}(z_{\max}) := \int_0^{z_{\max}} F_Z(z)dz$$



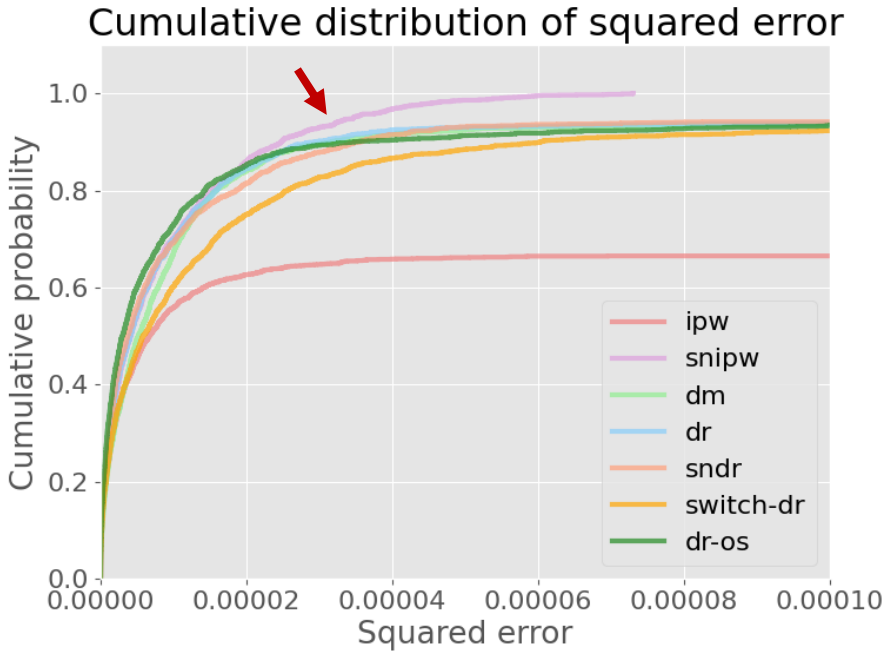
- *Conditional value-at-risk (CVaR)* compares the expected values of the estimators' squared error in the worst α x 100 % trials.

$$\text{CVaR}_\alpha(Z) := \mathbb{E}[Z \mid Z \geq F_Z^{-1}(\alpha)]$$



Experiments in a real-world application

- We applied IEOE to estimator selection in real e-commerce platform.
- The result demonstrates that SNIPW is stable across various configurations.
  (the conclusion may change when we consider different applications)



OPE Estimators	Mean (typical metric)	AU-CDF	$CVaR_{0.7}$	Std
DM	8.70	0.946	10.92	$35.94^{\dagger}$
IPWps	$29.45^{\dagger}$	$0.648^{\dagger}$	$31.96^{\dagger}$	$29.84^{\diamond}$
SNIPW	1.00^{*}	1.000^{*}	1.00^{*}	1.00^{*}
DRps	8.16	$0.953^{\diamond}$	10.27	34.54
SNDR	$7.45^{\diamond}$	0.942	$9.35^{\diamond}$	32.19
Switch-DR	8.16	$0.953^{\diamond}$	10.27	34.54
DRos	8.16	$0.953^{\diamond}$	10.27	34.54

*The values are normalized by that of SNIPW.

The platform now uses SNIPW based on our analysis!

Thank you for listening!

Find out more (e.g., synthetic and public data experiments) in the full paper!

contact: kiyohara.h.aa@m.titech.ac.jp

Direct Method (DM)

DM estimates mean reward function.

$$\hat{V}_{\mathrm{DM}}(\pi_e; \mathcal{D}, \hat{q}) := \mathbb{E}_n[\mathbb{E}_{a\sim\pi_e(a|x)}[\hat{q}(x_i, a)]]$$

where $\mathbb{E}[r \mid x, a] \approx \hat{q}(x, a)$

$\mathbb{E}_n[\cdot]$: empirical average

Large bias*, small variance. *due to inaccuracy of $\hat{q}$

Hyperparameter: $\hat{q}$

Self-Normalization (SNIPW, SNDR) [Swaminathan & Joachims, 2015]

SNIPW and SNDR address the variance issue of IPW and DR by using self-normalized value for importance weights.

$$\hat{V}_{\mathrm{SNIPW}}(\pi_e; \mathcal{D}) := \frac{\mathbb{E}_n[\rho(x_i, a_i) r_i]}{\mathbb{E}_n[\rho(x_i, a_i)]}$$

self-normalization

$$\hat{V}_{\mathrm{SNDR}}(\pi_e; \mathcal{D}, \hat{q}) := \mathbb{E}_n \left[\mathbb{E}_{a \sim \pi_e(a|x)}[\hat{q}(x_i, a)] + \frac{\rho(x_i, a_i)}{\mathbb{E}_n[\rho(x_i, a_i)]} (r_i - \hat{q}(x_i, a_i)) \right]$$

Consistent* and lower variance than IPW / DR. *when π_b is known or accurately estimated

Hyperparameter: $\hat{\pi}_b$ (when π_b is unknown) (, $\hat{q}$)

Switch-DR [Wang+, 2017]

Switch-DR interpolates between DM and DR ($\tau \to 0$ to DM, $\tau \to \infty$ to DR).

$$\hat{V}_{\mathrm{SwitchDR}}(\pi_e; \mathcal{D}, \hat{q}, \tau) := \mathbb{E}_n[\mathbb{E}_{a\sim\pi_e(a|x)}[\hat{q}(x_i, a)] + \rho(x_i, a_i)\mathbb{I}\{\rho(x_i, a_i) \leq \tau\}(r_i - \hat{q}(x_i, a_i))]$$

use importance weighting only when the weight is small

Lower variance than DR.

Hyperparameter: $\hat{\pi}_b$ (when π_b is unknown), $\hat{q}$ + τ

DR with Optimistic Shrinkage (DRos) [Su+, 2020]

DRos use new weight function to bridge DM and DR ($\lambda \to 0$ to DM, $\lambda \to \infty$ to DR).

$$\hat{V}_{\mathrm{DRos}}(\pi_e; \mathcal{D}, \hat{q}, \lambda) := \mathbb{E}_n[\mathbb{E}_{a\sim\pi_e(a|x)}[\hat{q}(x_i, a)] + \hat{\rho}(x_i, a_i; \lambda)(r_i - \hat{q}(x_i, a_i))]$$

where $$\hat{\rho}(x, a; \lambda) := \frac{\lambda}{\rho^2(x, a) + \lambda}\rho(x, a)$$

weight function to minimize error bounds

Minimize sharp bounds of mean-squared-error.

Hyperparameter: $\hat{\pi}_b$ (when π_b is unknown), $\hat{q}$ + λ

Conclusion

- We studied evaluation of off-policy evaluation (OPE).
- When applying OPE to a real-world problem, we need to identify a robust estimator that works without significant hyperparameter tuning.
- We develop *Interpretable evaluation for offline evaluation (IEOE)* to provide fruitful insights on the estimators' robustness.

We believe that IEOE will help practitioners to select a reliable OPE estimator!

References

[Strehl+, 2010] Alex Strehl, John Langford, Sham Kakade, and Lihong Li. Learning from Logged Implicit Exploration Data. NeurIPS, 2010. https://arxiv.org/abs/1003.0120

[Dudík+, 2014] Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly Robust Policy Evaluation and Optimization. Statistical Science, 2014. https://arxiv.org/abs/1503.02834

[Swaminathan & Joachims, 2015] Adith Swaminathan and Thorsten Joachims. The Self-Normalized Estimator for Counterfactual Learning. NeurIPS, 2015. https://dl.acm.org/doi/10.5555/2969442.2969600

[Wang+, 2017] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and Adaptive Off-policy Evaluation in Contextual Bandits. ICML, 2017. https://arxiv.org/abs/1612.01205

[Su+, 2020] Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly Robust Off-policy Evaluation with Shrinkage. ICML, 2020. https://arxiv.org/abs/1907.09623

[Narita+, 2021] Yusuke Narita, Shota Yasui, Kohei Yata. Debiased Off-Policy Evaluation for Recommendation Systems. RecSys, 2021. https://arxiv.org/abs/2002.08536