# Off-Policy Evaluation for Ranking Policies under Deterministic Logging Policies

**Koichi Tanaka**[1], **Kazuki Kawamura**[2], **Takanori Muroi**[2], **Yusuke Narita**[3], **Yuki Sasamoto**[2], **Kei Tateno**[2], **Takuma Udagawa**[2], **Wei-Wei Du**[2], **Yuta Saito**[4]

Keio University[1], Sony Group Corporation[2], Yale University[3], Hanuku-kaso, Co., Ltd.[4]

## Off-Policy Evaluation for Ranking Policies

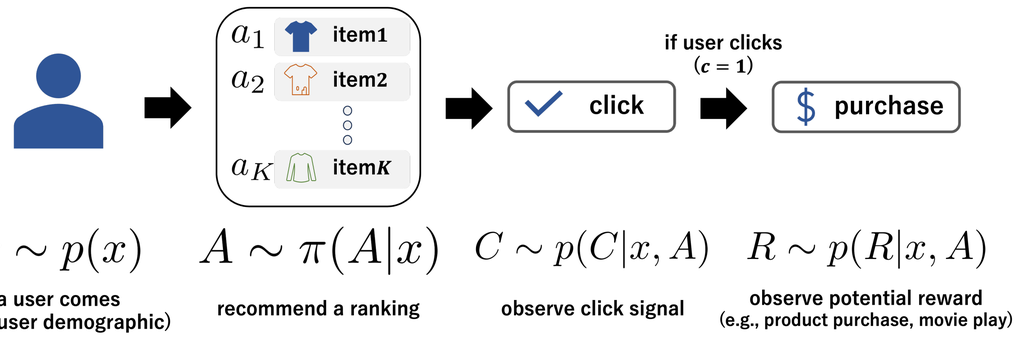


**Figure 1: The setting of off-policy evaluation for ranking policies.**

We formulate OPE for ranking policies in the *contextual bandit* process, where a ranking policy repeatedly observes a context $x$, produces a ranking $A$, and observes click signals $C$ and potential reward $R$. As ranking policies interact with the environment, they collect logged data $\mathcal{D} := \{(x_i, A_i, C_i, C_i R_i)\}_{i=1}^n$ that is valuable for *Off-Policy Evaluation* (OPE). OPE allows researchers to estimate the performance of new policies using only logged data.

The effectiveness of a policy $\pi$ is measured through its *value*, which is defined as follows.

$$V(\pi) = \mathbb{E}_{p(x)\pi(A|x)p(C,R|x,A)}\left[\sum_{k=1}^{K} C(k)R(k)\right] = \mathbb{E}_{p(x)\pi(A|x)}\left[\sum_{k=1}^{K} q_k(x, A)\right], \quad (1)$$

where $q_k(x, A) = \mathbb{E}[C(k)R(k) \mid x, A]$ is the *position-wise* expected reward function regarding position $k$.

In OPE, we aim to construct an estimator $\hat{V}(\pi; \mathcal{D})$ that can accurately estimate the policy value using only the logged data $\mathcal{D}$. The accuracy of an estimator is typically measured by the mean squared error (MSE):

$$\mathrm{MSE}(\hat{V}) := \mathbb{E}_{p(\mathcal{D})}\left[\left(\hat{V}(\pi; \mathcal{D}) - V(\pi)\right)^2\right] = \mathrm{Bias}\left[\hat{V}(\pi; \mathcal{D})\right]^2 + \mathrm{Var}\left[\hat{V}(\pi; \mathcal{D})\right],$$

## Limitations of Existing Methods

First, the ranking-wise IPS estimator is considered the most naive baseline. IPS reweights observed rewards by the ratio of ranking-wise probabilities under two policies as

$$\hat{V}_{\mathrm{IPS}}(\pi; \mathcal{D}) := \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(A_i|x_i)}{\pi_0(A_i|x_i)} \sum_{k=1}^{K} C_i(k)R_i(k) \quad (2)$$

where $w(x, A) = \pi(A|x)/\pi_0(A|x)$ is called the *ranking-wise importance weight*. This estimator is unbiased under the *ranking-wise common support* condition in the following.

**Condition 1** (Ranking-wise Common Support)**.** *The logging policy $\pi_0$ has common support if $\pi(A|x) > 0 \Rightarrow \pi_0(A|x) > 0$ for all $A \in \Delta(\prod_K(\mathcal{A}))$ and $x \in \mathcal{X}$.*

Although ranking-wise IPS only requires the common support for an unbiased evaluation, this condition is never satisfied under a deterministic logging policy, as shown in Table 1. **As results, IPS suffers from sever bias under a deterministic logging policy**.

Under the *independence condition*, which posits that users interact with items without being influenced by other items in a ranking, the independent IPS (IIPS) estimator can be defined as follows.

$$\hat{V}_{\mathrm{IIPS}}(\pi; \mathcal{D}) := \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \frac{\pi(A_i(k)|x_i)}{\pi_0(A_i(k)|x_i)} C_i(k)R_i(k) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} w(x_i, A_i(k))C_i(k)R_i(k), \quad (3)$$

where $\pi(A(k)|x) = \sum_{A'} \pi(A'|x)\mathbb{I}\{A'(k) = A(k)\}$. Although IIPS imposes weaker stochasticity requirements than Condition 1 (needed for unbiasedness of ranking-wise IPS), **it is still severely violated under a fully deterministic logging policy**.

**Table 1:** A toy example of importance weight under completely deterministic logging policy.

**(a)** Logging and new policy

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $A_6$ |
|---|---|---|---|---|---|---|
| $k=1$ | $a_1$ | $a_1$ | $a_2$ | $a_2$ | $a_3$ | $a_3$ |
| $k=2$ | $a_2$ | $a_3$ | $a_1$ | $a_3$ | $a_1$ | $a_2$ |
| $k=3$ | $a_3$ | $a_2$ | $a_3$ | $a_1$ | $a_2$ | $a_1$ |
| $\pi(A\|x_1)$ | 0.1 | 0.3 | 0.3 | 0.1 | 0.0 | 0.2 |
| $\pi_0(A\|x_1)$ | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

**(b)** Importance weights of IPS

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $A_6$ |
|---|---|---|---|---|---|---|
| $w(x_1, A)$ | 0.1 | NA | NA | NA | NA | NA |

**(c)** Importance weights of IIPS

| $w(x_1, A(k))$ | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| $k=1$ | 0.4 | NA | NA |
| $k=2$ | NA | 0.3 | NA |
| $k=3$ | NA | NA | 0.4 |

**(d)** click probability

| $p_c(x_1, a, A)$ | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $A_6$ |
|---|---|---|---|---|---|---|
| $k=1$ | 0.8 | 0.5 | 0.7 | 0.2 | 0.4 | 0.4 |
| $k=2$ | 0.5 | 0.6 | 0.6 | 0.5 | 0.3 | 0.4 |
| $k=3$ | 0.2 | 0.1 | 0.5 | 0.4 | 0.2 | 0.1 |

**(e)** importance weights of CIPS

| | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| $p_c(x_1, a, \pi)$ | 0.55 | 0.39 | 0.48 |
| $p_c(x_1, a, \pi_0)$ | 0.8 | 0.5 | 0.2 |
| $w(x_1, a, \pi, \pi_0)$ | 0.6875 | 0.78 | 2.4 |

## Click-based Inverse Propensity Score (CIPS)

Our proposed estimator exploits the intrinsic stochasticity of user click behavior rather than relying on the stochasticity of the logging policy. The key idea behind our method is that, even if the logging policy is fully deterministic, the user may still view each action with a positive probability.

We then propose our new estimator, **Click-based Inverse Propensity Score (CIPS)**, which leverages the ratio of click probabilities as a new form of importance weighting.

$$\hat{V}_{\mathrm{CIPS}}(\pi; \mathcal{D}) := \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}} \frac{p_c(x_i, a, \pi)}{p_c(x_i, a, \pi_0)} \cdot C_i(a)R_i(a) \quad (4)$$

where $p_c(x, a, \pi)/p_c(x, a, \pi_0)$ is the *click importance weight*, and $p_c(x, a, \pi) = \mathbb{E}_{\pi(A|x)}[p_c(x, a, A)] = \mathbb{E}_{\pi(A|x)}[\mathbb{E}[C(a)|x, A]]$ is the marginalized probability that a user with context $x$ clicks action $a$ under a ranking policy $\pi$. CIPS needs the following support condition for its unbiased estimation.

**Condition 2** (Click-wise Common Support)**.** *The logging policy $\pi_0$ has click-wise common support if $p_c(x, a, \pi) > 0 \Rightarrow p_c(x, a, \pi_0) > 0$ for all $a \in \mathcal{A}$ and $x \in \mathcal{X}$.*

This condition is more likely to hold because click probabilities are inherently stochastic. As discussed earlier, existing estimators fail to meet their respective support conditions in this setting by definition. In contrast, **Condition 2 is satisfied even when the logging policy is deterministic, intuitively showing that CIPS can mitigate bias by leveraging the stochasticity of click behavior**.

## Empirical Evaluation

We evaluate CIPS under deterministic logging policies using synthetic data, aiming to evaluate and compare the proposed methods, such as IPS, IIPS, and RIPS.

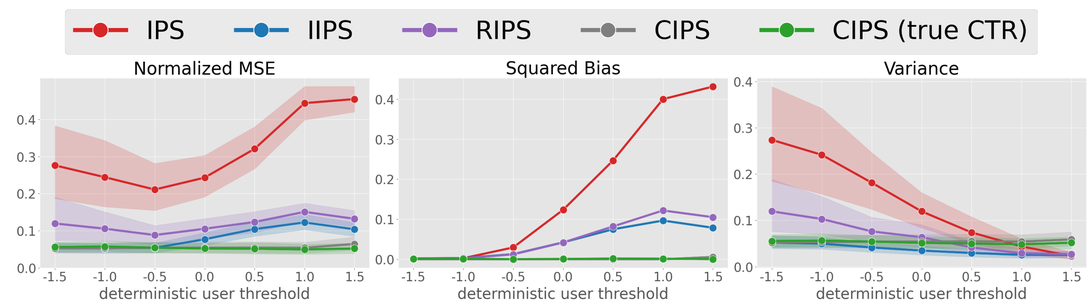


**Figure 2 : MSE, Squared Bias, and Variance with varying deterministic user thresholds.**

Figure 2 reports the performance of the estimators for varying the stochasticity of the logging policy. Varying user thresholds over $[-1.5, \ldots, 1.0, 1.5]$ changes the proportion of users with deterministic logging policies over $[0.07, \ldots, 0.84, 0.93]$. We find that CIPS maintains low bias even at large user thresholds, in contrast to the baselines, which exhibit substantial bias under near deterministic logging.