

A RELATED WORK

In this section, we distinguish our work from related studies and clarify the contributions of our methods and analysis.

First, we review OPE for ranking policies [Li et al. \(2018\)](#); [Kiyohara et al. \(2022\)](#); [McInerney et al. \(2020\)](#); [Kiyohara et al. \(2023\)](#) under the standard setting with sufficiently stochastic logging policies. Prior work mainly addresses the severe variance caused by large ranking spaces. To reduce the high variance of the “naïve”, ranking-wise IPS estimator, recent methods introduce user behavior assumptions such as independence [Li et al. \(2018\)](#) or cascade models [McInerney et al. \(2020\)](#); [Kiyohara et al. \(2022\)](#) to improve the bias–variance trade-off. In contrast, we target the severe bias introduced by deterministic logging policies rather than the variance problem. In this setting, all existing estimators fail drastically because they rely on the logging policy’s stochasticity and suffer from substantial bias due to the lack of exploration in the logged data.

Second, we compare our contributions with studies addressing the deficient support problem in the typical (non-ranking) OPE formulation [Sachdeva et al. \(2020\)](#); [Felicioni et al. \(2022\)](#). Deficient support refers to a milder situation where the logging policy assigns zero probability to some actions that the new policy may select. This situation is problematic for OPE because importance weighting cannot evaluate the reward of actions that have no probability of being selected under the logging policy. Compared to the deficient support problem, our work addresses an even more challenging setting, a fully deterministic logging policy. Deterministic logging is a strict special case of deficient support, eliminating all stochasticity, which makes existing methods severely ineffective and motivates our development of a new approach tailored to this scenario.

Third, we distinguish our contributions from studies on OPE for large action spaces [Saito et al. \(2023\)](#); [Saito & Joachims \(2022\)](#); [Taufiq et al. \(2023\)](#); [Guzman-Olivares et al. \(2025\)](#); [Kiyohara et al. \(2024\)](#). These studies aim to deal with the severe variance problem caused by large action spaces, similar to the typical OPE studies for ranking [Li et al. \(2018\)](#); [McInerney et al. \(2020\)](#); [Kiyohara et al. \(2022\)](#); [2023](#). Their key technique is to marginalize importance weights leveraging observed action embeddings to prevent the weights from taking excessively large values. Our proposed estimators also marginalize importance weights, but we do so using click probabilities rather than embeddings. While this shares the general idea, our goal is to address the lack of stochasticity in deterministic logging policies rather than variance in large action spaces. Furthermore, methods for large action spaces require additional information, such as pre-observed action embeddings, which our method does not require.

Finally, we mention a recent work addressing OPE in multiple domains [Natsubori et al. \(2025\)](#). That study uses logged data collected from multiple domains (e.g., multiple countries or hospitals) to compensate for the lack of stochasticity in the logging policy of the target domain. Although it tackles a similar problem to ours, it depends on multi-domain logged data, which is rarely available. Our method addresses deterministic logging without requiring data from multiple domains by exploiting the intrinsic stochasticity of user click behavior.

B EXTENSION TO CDR

We now extend CIPS to a more sophisticated estimator, **Click-based Doubly Robust (CDR)**, by incorporating a regression model for the expected potential reward.

$$\begin{aligned} & \hat{V}_{\text{CDR}}(\pi; \mathcal{D}) \\ &:= \frac{1}{n} \sum_{i=1}^n \sum_{a \in \mathcal{A}} \frac{p_c(x_i, a, \pi)}{p_c(x_i, a, \pi_0)} \cdot (C_i(a)R_i(a) - p_c(x_i, a, A)\hat{q}_r(x_i, a)) + \mathbb{E}_{\pi(A|x_i)} [p_c(x_i, a, A)\hat{q}_r(x_i, a)], \end{aligned} \quad (14)$$

where $p_c(x, a, A) = \mathbb{E}[C(a) \mid x, A]$.

We first analyze the bias of CDR under Conditions [3.1](#) and [3.2](#) both with and without access to the true click probability.

Theorem B.1. *Under Condition [3.1](#) and [3.2](#) CDR is unbiased, i.e., $\mathbb{E}_{p(\mathcal{D})}[\hat{V}_{\text{CDR}}(\pi; \mathcal{D})] = V(\pi)$. See Appendix for the proof.*

Theorem B.2. Under Condition 3.1 and 3.2 CDR has the following bias for a given estimated click probability.

$$\text{Bias}(\hat{V}_{\text{CDR}}) = \text{Bias}(\hat{V}_{\text{CIPS}}) = \mathbb{E}_{p(x)} \left[\sum_a p_c(x, a, \pi_0) \left(\frac{\hat{p}_c(x, a, \pi)}{\hat{p}_c(x, a, \pi_0)} - \frac{p_c(x, a, \pi)}{p_c(x, a, \pi_0)} \right) q_r(x, a) \right]$$

Theorems B.1 and B.2 show that CDR produces exactly the same bias as CIPS, regardless of the accuracy of the regression model.

Next, we analyze the variance of CDR and show that it is often smaller than that of CIPS.

Theorem B.3 (Variance of CDR). Under Conditions 3.1 and 3.2 CDR has the following variance.

$$\begin{aligned} n\mathbb{V}_{\mathcal{D}}[\hat{V}_{\text{CDR}}] &= \sum_{a \in \mathcal{A}} \left\{ \mathbb{E}_{p(x)\pi_0(A|x)} \left[w^2(x, a, \pi, \pi_0) \sigma^2(x, a, A) \right] \right. \\ &\quad \left. + \mathbb{E}_{p(x)} \left[w^2(x, a, \pi, \pi_0) \cdot \Delta_r^2(x, a) \mathbb{V}_{\pi_0(A|x)}[p_c(x, a, A)] \right] + \mathbb{V}_{p(x)} \left[p_c(x, a, \pi) q_r(x, a) \right] \right\}, \end{aligned}$$

where $\Delta_r(x, a) = q_r(x, a) - \hat{q}_r(x, a)$.

The key difference between the variance of CIPS (Theorem 3.3) and CDR lies in the term $\Delta_r(x, a)$, which captures the error of the potential reward model. When $\hat{q}_r(x, a)$ is more accurate than simple zero-filling, CDR is expected to reduce variance relative to CIPS.

Remark on Deterministic Logging Policies. When the logging policy is deterministic, CDR reduces to CIPS, implying that no variance reduction can be achieved. This is consistent with the variance analysis in Theorem B.3; in this case, the variance term related to the logging policy becomes zero, canceling the contribution of Δ . A similar phenomenon holds for the standard DR estimator, where under deterministic logging policies the Δ -dependent term vanishes and DR coincides with IPS in terms of variance. Consequently, the benefit of CDR arises not under fully deterministic logging, but rather in settings where the logging policy is partly deterministic or stochastic with limited support. In such cases, while IPS and its variants may suffer from severe bias, CIPS remains motivated, and CDR can further reduce variance relative to CIPS.

C EXAMPLES OF IMPORTANCE WEIGHTS UNDER COMPLETELY DETERMINISTIC LOGGING POLICY

Here, we provide useful intuition for the common support violation. Table 4 presents a toy example with $\mathcal{X} = \{x_1\}$, $\mathcal{A} = \{a_1, a_2, a_3\}$, and $K = 3$. This example is the same as Table 2 in the main text. In Table 4, we also consider importance weights of RIPS as well as those of ranking-wise IPS and position-wise IPS. As we discussed in the main text, the ranking-wise common support fails in 5 of 6 cases, whereas the position-wise counterpart fails in 6 of 9, indicating that IIPS satisfies its support condition slightly more often than IPS. Table 4d shows importance weights of RIPS. We can see that the common support of RIPS fails in 3 of 15 cases, suggesting that RIPS has severe bias under deterministic logging policies. This also indicates that RIPS satisfies its support condition slightly more often than IPS, but less often than IIPS. We can consider RIPS as an estimator that lies between IPS and IIPS. Thus, all baseline estimators in ranking OPE suffer from violation of support conditions, resulting in introducing substantial bias. However, CIPS achieves low-bias OPE under deterministic logging policies by leveraging click importance weights.

D OMITTED PROOFS

Here, we provide the derivations and proofs that are omitted in the main text.

D.1 PROOF OF THEOREM 3.1

We show that CIPS is unbiased under Conditions 3.1 and 3.2

$$\mathbb{E}_{p(x)\pi_0(A|x)p(C,R|x,A)} \left[\frac{1}{n} \sum_{i=1}^n \sum_{a \in \mathcal{A}} \frac{p_c(x_i, a_i, \pi)}{p_c(x_i, a_i, \pi_0)} C_i(a) R_i(a) \right]$$

Table 4: A toy example of importance weight under completely deterministic logging policy.

(a) Logging and new policy							(b) Importance weights of IPS						
	A_1	A_2	A_3	A_4	A_5	A_6		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	$w(x_1, A)$	0.1	NA	NA	NA	NA	NA
$k = 2$	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							
$\pi(A x_1)$							(c) Importance weights of IIPS						
	0.1	0.3	0.3	0.1	0.0	0.2	$w(x_1, A(k))$	a_1	a_2	a_3			
$\pi_0(A x_1)$	1.0	0.0	0.0	0.0	0.0	0.0							
							$k = 1$	0.4	NA	NA			
							$k = 2$	NA	0.3	NA			
							$k = 3$	NA	NA	0.4			
(d) Importance weights of RIPS													
	$A(1 : 1)$	a_1	a_2	a_3									
	$w(x_1, A(1 : 1))$	0.4	NA	NA									
	$A(1 : 2)$	(a_1, a_2)	(a_1, a_3)	(a_2, a_1)	(a_2, a_3)	(a_3, a_1)	(a_3, a_2)						
	$w(x_1, A(1 : 2))$	0.1	NA	NA	NA	NA	NA						
	$A(1 : 3)$	(a_1, a_2, a_3)	(a_1, a_3, a_2)	(a_2, a_1, a_3)	(a_2, a_3, a_1)	(a_3, a_1, a_2)	(a_3, a_2, a_1)						
	$w(x_1, A(1 : 3))$	0.1	NA	NA	NA	NA	NA						

$$\begin{aligned}
&= \mathbb{E}_{p(x)\pi_0(A|x)} \left[\sum_{a \in \mathcal{A}} \frac{p_c(x, a, \pi)}{p_c(x, a, \pi_0)} \mathbb{E}[C(a)R(a)|x, A] \right] \\
&= \mathbb{E}_{p(x)\pi_0(A|x)} \left[\sum_{a \in \mathcal{A}} \frac{p_c(x, a, \pi)}{p_c(x, a, \pi_0)} \mathbb{E}[C(a)|x, A] \mathbb{E}[R(a)|x] \right] \\
&= \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} \frac{p_c(x, a, \pi)}{p_c(x, a, \pi_0)} \mathbb{E}[R(a)|x] \sum_A \pi_0(A|x) \mathbb{E}[C(a)|x, A] \right] \\
&= \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} \frac{p_c(x, a, \pi)}{p_c(x, a, \pi_0)} \mathbb{E}[R(a)|x] p_c(x, a, \pi_0) \right] \\
&= \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} p_c(x, a, \pi) \mathbb{E}[R(a)|x] \right] \\
&= \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} \sum_A \pi(A|x) \mathbb{E}[C(a)|x, A] \mathbb{E}[R(a)|x] \right] \\
&= \mathbb{E}_{p(x)\pi(A|x)p(C, R|x, A)} \left[\sum_{a \in \mathcal{A}} C(a)R(a) \right] \\
&= V(\pi)
\end{aligned}$$

D.2 PROOF OF THEOREM 3.2

$$\text{Bias}(\hat{V}_{\text{CIPS}}) = \mathbb{E}_{p(x)\pi_0(A|x)p(C, R|x, A)} \left[\frac{1}{n} \sum_{i=1}^n \sum_{a \in \mathcal{A}} \frac{\hat{p}_c(x_i, a_i, \pi)}{\hat{p}_c(x_i, a_i, \pi_0)} C_i(a) R_i(a) \right] - V(\pi)$$

$$\begin{aligned}
&= \mathbb{E}_{p(x)\pi_0(A|x)p(C,R|x,A)} \left[\sum_{a \in \mathcal{A}} \frac{\hat{p}_c(x, a, \pi)}{\hat{p}_c(x, a, \pi_0)} C(a)R(a) \right] - \mathbb{E}_{p(x)\pi(A|x)p(C,R|x,A)} \left[\sum_{a \in \mathcal{A}} C(a)R(a) \right] \\
&= \mathbb{E}_{p(x)\pi_0(A|x)} \left[\sum_{a \in \mathcal{A}} \frac{\hat{p}_c(x, a, \pi)}{\hat{p}_c(x, a, \pi_0)} \mathbb{E}[C(a)|x, A] \mathbb{E}[R(a)|x, A] \right] \\
&\quad - \mathbb{E}_{p(x)\pi(A|x)} \left[\sum_{a \in \mathcal{A}} \mathbb{E}[C(a)|x, A] \mathbb{E}[R(a)|x, A] \right] \\
&= \mathbb{E}_{p(x)\pi_0(A|x)} \left[\sum_{a \in \mathcal{A}} \frac{\hat{p}_c(x, a, \pi)}{\hat{p}_c(x, a, \pi_0)} \mathbb{E}[C(a)|x, A] \mathbb{E}[R(a)|x] \right] \\
&\quad - \mathbb{E}_{p(x)\pi(A|x)} \left[\sum_{a \in \mathcal{A}} \mathbb{E}[C(a)|x, A] \mathbb{E}[R(a)|x] \right] \tag{15} \\
&= \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} \frac{\hat{p}_c(x, a, \pi)}{\hat{p}_c(x, a, \pi_0)} q_r(x, a) \sum_A \pi_0(A|x) \mathbb{E}[C(a)|x, A] \right] \\
&\quad - \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} q_r(x, a) \sum_A \pi(A|x) \mathbb{E}[C(a)|x, A] \right] \\
&= \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} \frac{\hat{p}_c(x, a, \pi)}{\hat{p}_c(x, a, \pi_0)} q_r(x, a) p_c(x, a, \pi_0) \right] - \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} q_r(x, a) p_c(x, a, \pi) \right] \\
&= \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} p_c(x, a, \pi_0) \left(\frac{\hat{p}_c(x, a, \pi)}{\hat{p}_c(x, a, \pi_0)} - \frac{p_c(x, a, \pi)}{p_c(x, a, \pi_0)} \right) q_r(x, a) \right],
\end{aligned}$$

where we use Condition 3.2 in Eq. 15.

D.3 PROOF OF THEOREM 3.3

We can derive the variance of CIPS by setting $\hat{q}_r(x, a) = 0$ in the variance of CDR.

$$\begin{aligned}
n\mathbb{V}_{\mathcal{D}} [\hat{V}_{\text{CIPS}}] &= \sum_{a \in \mathcal{A}} \left\{ \mathbb{E}_{p(x)\pi_0(A|x)} [w^2(x, a, \pi, \pi_0) \sigma^2(x, a, A)] \right. \\
&\quad \left. + \mathbb{E}_{p(x)} [w^2(x, a, \pi, \pi_0) q_r(x, a) \mathbb{V}_{\pi_0(A|x)} [q_c(x, a, A)]] + \mathbb{V}_{p(x)} [p_c(x, a, \pi) q_r(x, a)] \right\},
\end{aligned}$$

where $\sigma^2(x, a, A) = \mathbb{V}[C(a)R(a) | x, A]$.

D.4 PROOF OF THEOREM B.1

We show that CDR is unbiased under Conditions 3.1 and 3.2

$$\begin{aligned}
&\mathbb{E}_{p(x)\pi_0(A|x)p(C,R|x,A)} \left[\frac{1}{n} \sum_{i=1}^n \sum_{a \in \mathcal{A}} \frac{p_c(x_i, a, \pi)}{p_c(x_i, a, \pi_0)} \cdot (C_i(a)R_i(a) - p_c(x_i, a, A)\hat{q}_r(x_i, a)) + \mathbb{E}_{\pi(A|x_i)} [p_c(x_i, a, A)\hat{q}_r(x_i, a)] \right] \\
&= \mathbb{E}_{p(x)\pi_0(A|x)p(C,R|x,A)} \left[\sum_{a \in \mathcal{A}} \frac{p_c(x, a, \pi)}{p_c(x, a, \pi_0)} \cdot (C(a)R(a) - p_c(x, a, A)\hat{q}_r(x, a)) + \mathbb{E}_{\pi(A|x)} [p_c(x, a, A)\hat{q}_r(x, a)] \right] \\
&= \mathbb{E}_{p(x)\pi_0(A|x)} \left[\sum_{a \in \mathcal{A}} \frac{p_c(x, a, \pi)}{p_c(x, a, \pi_0)} \cdot (\mathbb{E}[C(a)R(a)|x, A] - p_c(x, a, A)\hat{q}_r(x, a)) + \mathbb{E}_{\pi(A|x)} [p_c(x, a, A)\hat{q}_r(x, a)] \right] \\
&= \mathbb{E}_{p(x)\pi_0(A|x)} \left[\sum_{a \in \mathcal{A}} \frac{p_c(x, a, \pi)}{p_c(x, a, \pi_0)} \cdot (p_c(x, a, A)q_r(x, a) - p_c(x, a, A)\hat{q}_r(x, a)) + \mathbb{E}_{\pi(A|x)} [p_c(x, a, A)\hat{q}_r(x, a)] \right]
\end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} \frac{p_c(x, a, \pi)}{p_c(x, a, \pi_0)} \cdot (q_r(x, a) - \hat{q}_r(x, a)) \sum_A \pi_0(A|x) p_c(x, a, A) + \mathbb{E}_{\pi(A|x)} [p_c(x, a, A) \hat{q}_r(x, a)] \right] \\
&= \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} \frac{p_c(x, a, \pi)}{p_c(x, a, \pi_0)} \cdot (q_r(x, a) - \hat{q}_r(x, a)) p_c(x, a, \pi_0) + p_c(x, a, \pi) \hat{q}_r(x, a) \right] \\
&= \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} p_c(x, a, \pi) \cdot (q_r(x, a) - \hat{q}_r(x, a)) + p_c(x, a, \pi) \hat{q}_r(x, a) \right] \\
&= \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} p_c(x, a, \pi) q_r(x, a) \right] \\
&= \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} \sum_A \pi(A|x) p_c(x, a, A) q_r(x, a) \right] \\
&= \mathbb{E}_{p(x) \pi(A|x)} \left[\sum_{a \in \mathcal{A}} p_c(x, a, A) q_r(x, a) \right] \\
&= \mathbb{E}_{p(x) \pi(A|x) p(C, R|x, A)} \left[\sum_{a \in \mathcal{A}} C(a) R(a) \right] \\
&= V(\pi)
\end{aligned}$$

D.5 PROOF OF THEOREM B.2

$$\begin{aligned}
&\text{Bias}(\hat{V}_{\text{CDR}}) \\
&= \mathbb{E}_{p(\mathcal{D})} \left[\frac{1}{n} \sum_{i=1}^n \sum_{a \in \mathcal{A}} \frac{\hat{p}_c(x_i, a, \pi)}{\hat{p}_c(x_i, a, \pi_0)} \cdot (C_i(a) R_i(a) - \hat{p}_c(x_i, a, A) \hat{q}_r(x_i, a)) + \mathbb{E}_{\pi(A|x_i)} [\hat{p}_c(x_i, a, A) \hat{q}_r(x_i, a)] \right] - V(\pi) \\
&= \mathbb{E}_{p(\mathcal{D})} \left[\sum_{a \in \mathcal{A}} \frac{\hat{p}_c(x, a, \pi)}{\hat{p}_c(x, a, \pi_0)} \cdot (C(a) R(a) - \hat{p}_c(x, a, A) \hat{q}_r(x, a)) + \mathbb{E}_{\pi(A|x)} [\hat{p}_c(x, a, A) \hat{q}_r(x, a)] \right] - V(\pi) \\
&= \mathbb{E}_{p(x) \pi_0(A|x)} \left[\sum_{a \in \mathcal{A}} \frac{\hat{p}_c(x, a, \pi)}{\hat{p}_c(x, a, \pi_0)} \cdot (p_c(x, a, A) q_r(x, A) - \hat{p}_c(x, a, A) \hat{q}_r(x, a)) + \hat{p}_c(x, a, \pi) \hat{q}_r(x, a) \right] - V(\pi) \\
&= \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} \frac{\hat{p}_c(x, a, \pi)}{\hat{p}_c(x, a, \pi_0)} \cdot (p_c(x, a, \pi_0) q_r(x, A) - \hat{p}_c(x, a, \pi_0) \hat{q}_r(x, a)) + \hat{p}_c(x, a, \pi) \hat{q}_r(x, a) \right] - V(\pi) \\
&= \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} \frac{\hat{p}_c(x, a, \pi)}{\hat{p}_c(x, a, \pi_0)} p_c(x, a, \pi_0) q_r(x, A) \right] - \mathbb{E}_{p(x)} [p_c(x, a, \pi) q_r(x, a)] \\
&= \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} p_c(x, a, \pi_0) \left(\frac{\hat{p}_c(x, a, \pi)}{\hat{p}_c(x, a, \pi_0)} - \frac{p_c(x, a, \pi)}{p_c(x, a, \pi_0)} \right) q_r(x, a) \right] \\
&= \text{Bias}(\hat{V}_{\text{CIPS}}) \tag{16}
\end{aligned}$$

D.6 PROOF OF THEOREM B.3

$$\begin{aligned}
&n \mathbb{V}_{\mathcal{D}} [\hat{V}_{\text{CDR}}] \\
&= n \mathbb{V}_{\mathcal{D}} \left[\frac{1}{n} \sum_{i=1}^n \sum_{a \in \mathcal{A}} \frac{p_c(x_i, a, \pi)}{p_c(x_i, a, \pi_0)} \cdot (C_i(a) R_i(a) - p_c(x_i, a, A) \hat{q}_r(x_i, a)) + \mathbb{E}_{\pi(A|x_i)} [p_c(x_i, a, A) \hat{q}_r(x_i, a)] \right]
\end{aligned}$$

$$\begin{aligned}
&= \mathbb{V}_{\mathcal{D}} \left[\sum_{a \in \mathcal{A}} \frac{p_c(x, a, \pi)}{p_c(x, a, \pi_0)} \cdot (C(a)R(a) - p_c(x, a, A)\hat{q}_r(x, a)) + \mathbb{E}_{\pi(A|x)} [p_c(x, a, A)\hat{q}_r(x, a)] \right] \\
&= \mathbb{E}_{p(x)\pi_0(A|x)} \left[\mathbb{V}_{p(C, R|x, A)} \left[\sum_{a \in \mathcal{A}} w(x, a, \pi, \pi_0) \cdot (C(a)R(a) - p_c(x, a, A)\hat{q}_r(x, a)) + \mathbb{E}_{\pi(A|x)} [p_c(x, a, A)\hat{q}_r(x, a)] \right] \right] \\
&\quad + \mathbb{V}_{p(x)\pi_0(A|x)} \left[\mathbb{E}_{p(C, R|x, A)} \left[\sum_{a \in \mathcal{A}} w(x, a, \pi, \pi_0) \cdot (C(a)R(a) - p_c(x, a, A)\hat{q}_r(x, a)) + \mathbb{E}_{\pi(A|x)} [p_c(x, a, A)\hat{q}_r(x, a)] \right] \right] \\
&= \mathbb{E}_{p(x)\pi_0(A|x)} \left[\sum_{a \in \mathcal{A}} w^2(x, a, \pi, \pi_0) \sigma^2(x, a, A) \right] \\
&\quad + \mathbb{V}_{p(x)\pi_0(A|x)} \left[\sum_{a \in \mathcal{A}} w(x, a, \pi, \pi_0) \cdot (p_c(x, a, A)q_r(x, a) - p_c(x, a, A)\hat{q}_r(x, a)) + \mathbb{E}_{\pi(A|x)} [p_c(x, a, A)\hat{q}_r(x, a)] \right] \\
&= \mathbb{E}_{p(x)\pi_0(A|x)} \left[\sum_{a \in \mathcal{A}} w^2(x, a, \pi, \pi_0) \sigma^2(x, a, A) \right] \\
&\quad + \mathbb{E}_{p(x)} \left[\mathbb{V}_{\pi_0(A|x)} \left[\sum_{a \in \mathcal{A}} w(x, a, \pi, \pi_0) \cdot (p_c(x, a, A)q_r(x, a) - p_c(x, a, A)\hat{q}_r(x, a)) + \mathbb{E}_{\pi(A|x)} [p_c(x, a, A)\hat{q}_r(x, a)] \right] \right] \\
&\quad + \mathbb{V}_{p(x)} \left[\mathbb{E}_{\pi_0(A|x)} \left[\sum_{a \in \mathcal{A}} w(x, a, \pi, \pi_0) \cdot (p_c(x, a, A)q_r(x, a) - p_c(x, a, A)\hat{q}_r(x, a)) + \mathbb{E}_{\pi(A|x)} [p_c(x, a, A)\hat{q}_r(x, a)] \right] \right] \\
&= \mathbb{E}_{p(x)\pi_0(A|x)} \left[\sum_{a \in \mathcal{A}} w^2(x, a, \pi, \pi_0) \sigma^2(x, a, A) \right] \\
&\quad + \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} w^2(x, a, \pi, \pi_0) \cdot (q_r(x, a) - \hat{q}_r(x, a))^2 \mathbb{V}_{\pi_0(A|x)} [p_c(x, a, A)] \right] \\
&\quad + \mathbb{V}_{p(x)} \left[\sum_{a \in \mathcal{A}} w(x, a, \pi, \pi_0) \cdot (q_r(x, a) - \hat{q}_r(x, a)) \sum_A \pi_0(A|x) p_c(x, a, A) + \mathbb{E}_{\pi(A|x)} [p_c(x, a, A)\hat{q}_r(x, a)] \right] \\
&= \mathbb{E}_{p(x)\pi_0(A|x)} \left[\sum_{a \in \mathcal{A}} w^2(x, a, \pi, \pi_0) \sigma^2(x, a, A) \right] \\
&\quad + \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} w^2(x, a, \pi, \pi_0) \cdot (q_r(x, a) - \hat{q}_r(x, a))^2 \mathbb{V}_{\pi_0(A|x)} [p_c(x, a, A)] \right] \\
&\quad + \mathbb{V}_{p(x)} \left[\sum_{a \in \mathcal{A}} w(x, a, \pi, \pi_0) \cdot (q_r(x, a) - \hat{q}_r(x, a)) p_c(x, a, \pi_0) + \mathbb{E}_{\pi(A|x)} [p_c(x, a, A)\hat{q}_r(x, a)] \right] \\
&= \mathbb{E}_{p(x)\pi_0(A|x)} \left[\sum_{a \in \mathcal{A}} w^2(x, a, \pi, \pi_0) \sigma^2(x, a, A) \right] \\
&\quad + \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} w^2(x, a, \pi, \pi_0) \cdot (q_r(x, a) - \hat{q}_r(x, a))^2 \mathbb{V}_{\pi_0(A|x)} [p_c(x, a, A)] \right] \\
&\quad + \mathbb{V}_{p(x)} \left[\sum_{a \in \mathcal{A}} p_c(x, a, \pi) \cdot (q_r(x, a) - \hat{q}_r(x, a)) + p_c(x, a, \pi) \hat{q}_r(x, a) \right] \\
&= \mathbb{E}_{p(x)\pi_0(A|x)} \left[\sum_{a \in \mathcal{A}} w^2(x, a, \pi, \pi_0) \sigma^2(x, a, A) \right] \\
&\quad + \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} w^2(x, a, \pi, \pi_0) \cdot (q_r(x, a) - \hat{q}_r(x, a))^2 \mathbb{V}_{\pi_0(A|x)} [p_c(x, a, A)] \right] \\
&\quad + \mathbb{V}_{p(x)} \left[\sum_{a \in \mathcal{A}} p_c(x, a, \pi) q_r(x, a) \right] \\
&= \sum_{a \in \mathcal{A}} \left\{ \mathbb{E}_{p(x)\pi_0(A|x)} \left[w^2(x, a, \pi, \pi_0) \sigma^2(x, a, A) \right] \right. \\
&\quad \left. + \mathbb{E}_{p(x)} \left[w^2(x, a, \pi, \pi_0) \cdot \Delta_r^2(x, a) \mathbb{V}_{\pi_0(A|x)} [p_c(x, a, A)] \right] + \mathbb{V}_{p(x)} [p_c(x, a, \pi) q_r(x, a)] \right\}
\end{aligned}$$

where $\Delta_r(x, a) = q_r(x, a) - \hat{q}_r(x, a)$.

E ADDITIONAL EXPERIMENTAL SETUPS AND RESULTS

This section describes the detailed experimental settings and reports additional results.

E.1 SYNTHETIC EXPERIMENTS

Detailed Setup. We first describe the synthetic experiment settings in detail, followed by how we define the expected reward function. In the synthetic experiments, the expected reward function is defined as follows:

$$q_c(x, A(k)) = \frac{1}{k} \cdot \text{sigmoid}(\hat{q}_c(x, A(k))) + \sum_{l \neq k} \frac{1}{|k - l|} \cdot \mathbb{W}_c(A(l), A(k))$$

$$q_r(x, A(k)) = \hat{q}_r(x, A(k)) + \lambda \cdot \sum_{l \neq k} \frac{1}{|k - l|} \cdot \mathbb{W}_r(A(l), A(k)),$$

where \mathbb{W}_c and \mathbb{W}_r are sampled from a uniform distribution with range $[-3.0, 3.0]$ and $[-1.0, 1.0]$, respectively. Additionally, $\hat{q}_c(x, A(k))$ and $\hat{q}_r(x, A(k))$ are defined as follows.

$$\hat{q}_c(x, A(k)) = x^T M_{x, x_a}^c x_a + (\theta_x^c)^T \cdot x + (\theta_a^c)^T \cdot x_a,$$

$$\hat{q}_r(x, A(k)) = x^T M_{x, x_a}^r x_a + (\theta_x^r)^T \cdot x + (\theta_a^r)^T \cdot x_a,$$

where x_a denotes action context for $A(k) = a$ represented by a one-hot vector. M_{x, x_a} , θ_x and θ_a are sampled from a uniform distribution with range $[-1, 1]$.

Additional Results. In Figure 7 we report additional results from the synthetic experiments to demonstrate that CIPS outperforms the baselines. We vary the new ranking policies. Note that the logged data size is set to 1000, the number of unique actions is 6, the ranking length is 6, $\alpha = \infty$, and $\lambda = 0.5$.

How does CIPS perform with varying new ranking policies? Figure 7 shows comparisons of estimators under varying new ranking policies ($\epsilon \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ in Eq. 12). A larger value of ϵ makes the new ranking policy closer to a random uniform distribution. $\epsilon = 0.0$ means the policy is completely deterministic. We observe that CIPS outperforms the baselines across all values of ϵ . As ϵ increases, the probability of the supported action increases. Consequently, the baselines reduce their bias. However, the baselines still exhibit severe bias, particularly when the new ranking policy is completely deterministic ($\epsilon = 0.0$).

E.2 ABLATION STUDY ON SYNTHETIC DATA

In addition to the main performance comparisons, we conduct two ablation studies.

Figure 9 compares CDR against CIPS as we vary the accuracy of the regression model \hat{q}_r used in CDR. Here, the x-axis represents the estimation error of the regression model; larger values correspond to less accurate estimates of q_r . The results show that, when q_r is estimated with reasonable accuracy, CDR outperforms CIPS by leveraging its variance reduction effect, consistent with our theoretical analysis. We also observe that as the estimation noise in the regression model increases, the variance of CDR grows steadily, eventually eliminating its advantage over CIPS.

Next, Figure 8 evaluates the policy selection accuracy of the estimators under varying logged data sizes (n) and deterministic user thresholds (α). We measure how accurately each estimator identifies the better policy between the new policy π and the logging policy π_0 across 100 independent trials. Note that, in our setup, the new policy consistently outperforms the logging policy in expectation. The left plot demonstrates that CIPS achieves the highest selection accuracy across all data sizes, while the baselines fail entirely with zero accuracy. This failure is expected because, under severe common support violations, as shown in Theorem 2.1, baseline methods substantially underestimate the value of the new policy π , leading to systematic errors in policy selection. The right plot shows that CIPS maintains high selection accuracy even at larger α values, where the logging policy becomes more deterministic and the baselines deteriorate substantially. These findings highlight the superiority of CIPS over the baselines in both policy selection and policy evaluation.

E.3 REAL-WORLD EXPERIMENTS

Detailed Setup. The detailed real-world experimental settings are described in the main text. We conduct OPE experiments on a real-world dataset called KuaiRec [Gao et al. \(2022\)](#), which consists of recommendation logs from the video-sharing app Kuaishou. KuaiRec contains fully observed user-item interactions with nearly 100% density for a subset of its users and items. By leveraging these user-item interactions, we can construct OPE experiments with minimal synthetic components [Gao et al. \(2022\)](#).

KuaiRec includes `user_feature.csv`, which we use as context vectors (x). We apply feature dimension reduction using PCA implemented in scikit-learn [\(Pedregosa et al., 2011\)](#). We then use the user-item interaction matrix recorded in the original data as the potential reward function $q_r(x, A(k))$. We construct $q_r(x, A(k))$ by randomly sampling rows and columns from the user-item matrix. We also define the expected click probability as

$$q_c(x, A(k)) = \begin{cases} 1 - \eta_{x, A(k)} & \text{if } q_r(x, A(k)) > 2.0 \\ \eta_{x, A(k)} & \text{otherwise} \end{cases}, \quad (17)$$

where η is a noise parameter sampled independently from the uniform distribution over $[0, 0.5]$. We set the threshold for binarization by following the example provided on the KuaiRec webpage.

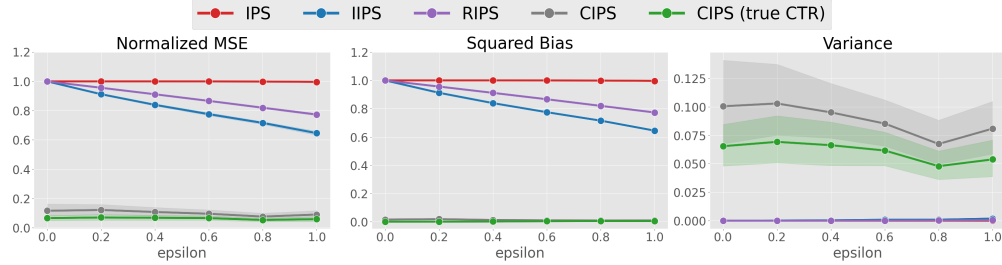


Figure 7: Comparison of CIPS and CDR' MSE, Squared Bias, and Variance with varying new policies (ϵ).

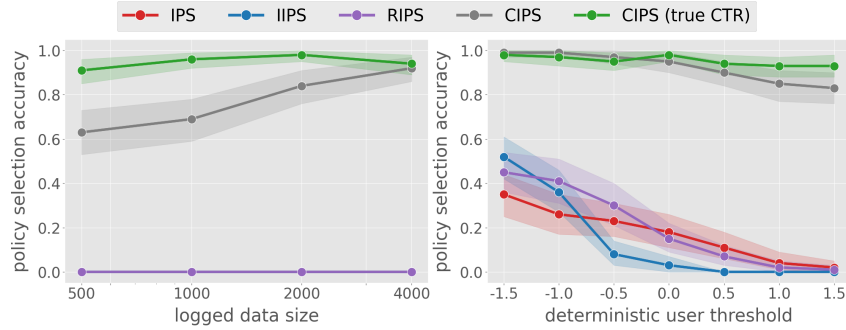


Figure 8: Comparison of the policy selection accuracy with varying logged data sizes (n) and deterministic user thresholds (α).

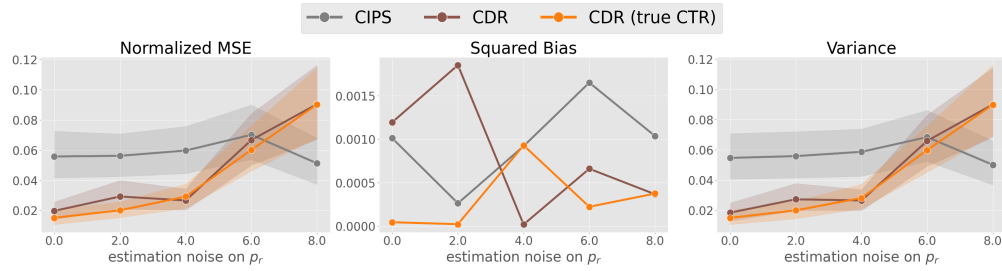


Figure 9: Comparison of CIPS and CDR' MSE, Squared Bias, and Variance with varying estimation noises on q_r .