Towards Unbiased and Robust Causal Ranking for Recommender Systems

Teng Xiao
The Pennsylvania State University
tengxiao@psu.edu

Suhang Wang
The Pennsylvania State University
szw494@psu.edu

ABSTRACT

We study the problem of optimizing ranking metrics with unbiased and robust causal estimation for recommender systems. A user may click/purchase an item regardless of whether the item is recommended or not. Thus, it is important to estimate the causal effect of recommendation and rank items higher with a larger causal effect. However, most existing works focused on improving the accuracy of recommendations, which usually have large bias and variance. Therefore, in this paper, we provide a general and theoretically rigorous framework for causal recommender systems, which enables unbiased evaluation and learning for the ranking metrics with confounding bias. We first propose a robust estimator for unbiased ranking evaluation and theoretically show that this estimator has a smaller bias and variance. We then propose a deep variational information bottleneck (IB) approach to exploit the sufficiency of the propensity score for estimation adjustment and better generalization. We also provide the learning bound and develop an unbiased learning algorithm to optimize the causal metric. Results on semi-synthetic and real-world datasets show that our evaluation and learning algorithms significantly outperform existing methods.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**.

KEYWORDS

Causal Inference; Recommender Systems; Counterfactual Learning

ACM Reference Format:
Teng Xiao and Suhang Wang . 2022. Towards Unbiased and Robust Causal Ranking for Recommender Systems. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM '22), February 21–25, 2022, Tempe, AZ, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3488560.3498521

1 INTRODUCTION

Recommender systems (RS) focus on modeling the probability of clicking on recommendations from the logged feedbacks, which have shown widespread success. However, they are oblivious to whether logged feedbacks were coming from recommendations or other factors irrelevant to recommendations. User logged feedbacks might be attributed to recommendations or other causes. For instance, people will actively search and buy popular or necessary items even they are not recommended. Thus if we neglect this scenario, the algorithm will be biased and overestimate those popular items. Similarly, students actively buy textbooks even they are not recommended. Sharma et al. (2015) analyzed the browsing logs containing anonymized activities for 2.1 million users on Amazon.com and revealed that at least 75% of activities would likely occur in the absence of recommendations [41]. If a model is trained to maximize logged feedbacks which are not the causal effects of recommendation, we can expect it would not increase positive interactions and cannot generate an optimal ranking. Hence, it is important to estimate the causal effect of the recommendation.

Formally, we first illustrate the causal problem in recommendation studied in this paper. Let $r_{ui} \in \{0, 1\}$ denote whether item i is recommended to user u based on user and item features $\mathbf{x}_{ui}$. The recommendation assignment (treatment) r_{ui} leads to two potential outcomes (e.g., clicks), i.e., $c_{ui}^{(1)}$ and $c_{ui}^{(0)}$, where $c_{ui}^{(1)}$ is the activity of user u when item i is recommended to u and $c_{ui}^{(0)}$ is the activity of u when i is not recommended to u. The causal effect of recommendation is defined as the difference $\tau_{ui} = c_{ui}^{(1)} - c_{ui}^{(0)}$ caused purely by the recommendation. Thus, instead of simply modeling users' probabilities of clicking items via the supervised learning, we are more interested in developing *causal recommender system that can assign higher ranks to items which have larger causal effects.*

However, it is non-trivial to directly optimize the causal effect of the recommendation from the observational logged feedback because of the following challenges: (*i*) Partial feedback. For $\mathbf{x}_{ui}$, and for each potential treatment $r_{ui} = \{0, 1\}$, $c_{ui}^{(r)}$ is the potential outcome for the intervention. We only known either $c_{ui}^{(1)}$ or $c_{ui}^{(0)}$ because i is either recommended or not to u at a given time as shown in Table 1, which makes it difficult to calculate the individual causal effect $\tau_{ui} = c_{ui}^{(1)} - c_{ui}^{(0)}$. (*ii*) Due to the presence of the confounding, the recommendation (treatment) assignment is not at random. As such, the treatment assignment mechanism will be causally affected by context variables $\mathbf{x}_{ui}$ that also causally influence the outcome. For example, students (context) are more likely to be recommended (treatment) some textbooks and click (outcome) them. Neglecting the confounding bias may overlook key limitations of recommendation algorithms, such as overestimating recommendation effect and exacerbating unhealthful user behavior [8]. Thus, it is important to optimize the causal effect and correct the confounding bias for RS.

In this paper, to address the challenges outlined above, we present a theoretically principled and empirically effective approach for optimizing the causal effect and correcting the confounding bias from observational logged feedbacks. Recently, several methods [25,

WSDM '22, February 21–25, 2022, Tempe, AZ, USA

ACM ISBN 978-1-4503-9132-0/22/02…$15.00
https://doi.org/10.1145/3488560.3498521

r_{ui}	$\mathbf{x}_{ui}$	$c_{ui}^{(0)}$	$c_{ui}^{(1)}$	$\tau_{ui} = c_{ui}^{(1)} - c_{ui}^{(0)}$
1	✓		✓	
1	✓		✓	
0	✓	✓		
0	✓	✓		

Table 1: An illustration of the observed (factual) (✓) and unobserved (counterfactual) variables, and our interested causal effect τ_{ui} of the recommendation.

29, 40] have been developed for the causal estimation. However, they mainly focus on the evaluation instead of learning a ranking algorithm, which is one of the most important characteristics of the recommendation. We note that a recent work [38] employed IPS method in causal inference fields, and developed causal ranking methods for RS. However, unlike our work, their work still has two limitations: 1) The IPS approach utilized by [38] suffers from a large variance in estimating the causal effect [7, 45]. (2) Their work assumes that the propensity score of IPS is known in advance, whereas our method not only focuses on optimizing the ranking algorithm, but also targets how to effectively learn the estimator from the observational data without knowing the propensity score.

We first analyze the bias and variance of existing IPS-based methods [38] with unknown but estimated propensity score. Based on the analysis, we then develop a provably unbiased and robust estimator for causal effect of recommendation. Our analysis is built on doubly robust estimators that was first developed in statistics [5, 9, 13, 22, 31] for causal inference from incomplete data. However, these works focus on regression *evaluation* with known propensity scores; while we explore the problem of *learning* a ranking policy with unknown propensity scores for RS. We also propose an adaptive information bottleneck (IB) approach to effectively learn this estimator, which can trade-off between outcome accuracy and the propensity-score representation, and improve the model generalization. Based on the learned estimator, we develop a differentiable learning algorithm for unbiased ranking algorithm by considering the connection between the estimation and learning steps, and provide the learning bound which is tighter than previous work [38]. In addition to the theoretical derivation and justification, we demonstrate the effectiveness of our methods through extensive experiments with both semi-synthetic and real-world datasets.

2 RELATED WORK

Unbiased Evaluation and Learning. For the unbiased evaluation, Gilotte et al. [14] utilize several counterfactual estimators to conduct unbiased evaluation of the new policy in RS. For the unbiased learning, previous works [28, 35, 36, 39, 53] mainly focus on the selection bias problem with missing not at random (MNAR) feedback data. Specifically, the work [28] introduces an exposure matrix to model the selection bias and Schnabel et al. [39] adopt the inverse propensity score (IPS) technique from causal inference to address the selection bias. [6] focuses on utilizing the uniform data to alleviate the selection biases in RS. To further address the implicit feedback problem, Yang et al. [53] and Saito et al. [36] both propose the unbiased IPS estimator for the ideal loss inspired by the estimation method of causal inference and positive-unlabeled learning. However, previous studies [16, 23, 45, 48, 51, 52] indicate that the variance of the IPS estimator can be significant. To further reduce the variance of IPS, prior works [34, 47] propose a doubly robust (DR) estimator to conduct the unbiased learning for RS.

However, the focus of the above works is not ranking with causal effect that we address, which makes the estimators proposed by them unsuitable in our scenario. Sato et al. [38] utilize IPS to correct the confounding bias in RS. However, it suffers from two issues: (i) The IPS approach in [38] has large variance for causal effect estimation [45] and can lead to a poor generalization ability [7]; and (ii) it assumes that the propensity score in IPS is known already, whereas our method target on how to effectively learn the estimator from the observational data without knowing propensity score.

Our approach is also built on the DR estimator that was first developed in statistics [5, 9, 22] for causal inference from incomplete data. It was then brought to batch bandit in the machine learning community [10, 43, 44, 46]. Different from them, in this paper, we study the problem of correcting confounding bias in casual ranking.

Information Bottleneck. Our work is also related to the representation learning with the information bottleneck (IB). Alemi et al. first propose DVIB which shows increased robustness of learned representations. Other methods apply IB to various domains such as reinforcement learning [15], graph neural networks [49] and natural language processing [27]. In this paper, we adopt IB to improve the generalization performance of the causal effect estimation.

3 NOTATIONS AND PROBLEM SETTINGS

Let $u \in \mathcal{U}$ be a user and $i \in \mathcal{I}$ be an item. $\mathbf{C} = \{0,1\}^{|\mathcal{U}|\times|\mathcal{I}|}$ denotes observed interactions, e.g., clicks. $c_{ui} = 1$ if the interaction (u, i) is observed; otherwise $c_{ui} = 0$. To formulate the causal effect of recommendation algorithm, we introduce a recommendation matrix $\mathbf{R} = \{0,1\}^{|\mathcal{U}|\times|\mathcal{I}|}$, where $r_{ui} = 1$ means that item i is recommended to user u, otherwise $r_{ui} = 0$. Thus, the binary treatment assignment r_{ui} leads to two potential outcomes, i.e., $c_{ui}^{(1)}$ if $r_{ui} = 1$ and $c_{ui}^{(0)}$ if $r_{ui} = 0$. Note that we can only observe one of the outcomes. Due to confounding factors $\mathbf{x}_u$ and $\mathbf{x}_i$ generally approximated by the features of users and items, even without recommendation, an item could be purchased by a user. For example, due to their properties, popular or necessary items would be purchased by users with little or no affects by recommendation. Thus, we consider the average treatment effect to measure the recommendation effect [32]:

$$\tau_{ui} = \mathbb{E}\left[c_{ui}^{(1)} - c_{ui}^{(0)} | \mathbf{x}_{ui}\right], \tag{1}$$

where $\mathbf{x}_{ui} = \{\mathbf{x}_u, \mathbf{x}_i\}$ is the set of features of u and i. Let c_{ui} be the observed outcome; then, $c_{ui} = c_{ui}^{(r)}$ when $r_{ui} = r$. Given the observed dataset $\mathcal{D} = \{(\mathbf{x}_{ui}, r_{ui}, c_{ui})\}_{u=1,i=1}^{|\mathcal{U}|\times|\mathcal{I}|}$, our goal is to develop a recommender system that can recommend items which maximize the true causal effect in Eq. (1), i.e., *recommend items users like but are not likely to purchase without recommendation.* Specifically, we optimize the a causal ranking metric which extends the traditional ranking metric $R(\hat{Z}) = \frac{1}{|\mathcal{U}|}\sum_u \sum_i \lambda(\hat{z}_{ui}) c_{ui}$ [2] to causal settings [38]. The ideal causal ranking metric is:

$$R^{\text{Ideal}}(\hat{Z}) = \frac{1}{|\mathcal{U}|}\sum_u \sum_i R^{\text{Ideal}}(\hat{z}_{ui}) = \frac{1}{|\mathcal{U}|}\sum_u \sum_i \lambda(\hat{z}_{ui})\tau_{ui}, \tag{2}$$

where $\hat{Z} = \{\hat{z}_{ui}\}_{(u,i)\in\mathcal{D}}$ is the predicted ranking and $\lambda(\cdot)$ is the top-N ranking metric such as Discounted Cumulative Gain (DCG) [2, 38]. For DCG, the function $\lambda(\cdot)$ is defined as $\lambda(\hat{z}_{u,i}) = 1/\log(\hat{z}_{u,i}+1)$.

4 EXISTING ESTIMATORS

To optimize Eq. (2), we need to first estimate the causal effect τ_{ui} and then learn the ranking $\hat{z}_{ui}$. Estimating causal effect τ_{ui} has two main challenges: (i) we cannot observe the causal effect τ_{ui} directly; and (ii) we need to remove confounding bias. We first summarize existing causal effect estimators, then introduce the DR estimator.

Naive Estimator. The naive estimator [38] is the most basic estimator for the causal effect for RS. Naively, the average causal effect over whole user-item pairs can be estimated as the difference between the averages of outcomes under treatment and control:

$$\hat{\tau}_{ui}^{\text{Naive}} = \frac{r_{ui}c_{ui}^{(1)}}{\sum r_{ui}/|\mathcal{U}||\mathcal{I}|} - \frac{(1-r_{ui})c_{ui}^{(0)}}{\sum(1-r_{ui})/|\mathcal{U}||\mathcal{I}|}. \tag{3}$$

This estimator is intuitive and assumes the treatment assignment is random such that the covariate distributions between treated and control are identical. However, we are interested in estimating the causal effects in observational data and cannot apply this since this would lead to biased estimates due to the confoundedness.

Inverse Propensity Score (IPS) Estimator. To address the confounding bias of the naive estimator, Sato et al. [38] adopt an unbiased estimator for the causal effect in RS by utilizing IPS, i.e.,

$$\hat{\tau}_{ui}^{\text{IPS}} = \mathbb{E}\left[\frac{r_{ui}c_{ui}^{(1)}}{e(\mathbf{x}_{ui})} - \frac{(1-r_{ui})c_{ui}^{(0)}}{1-e(\mathbf{x}_{ui})}|\mathbf{x}_{ui}\right], \tag{4}$$

where $e(\mathbf{x}_{ui}) = p(r_{ui}=1|\mathbf{x}_{ui})$ is the true propensity score that represents the probability of i being recommended to u. Under stable unit treatment value and unconfoundedness assumptions [21], i.e., $\{c_{ui}^{(1)}, c_{ui}^{(0)}\} \perp\!\!\!\perp r_{ui}|e_{ui}$ for all user-item pairs (u,i), the estimator τ_{ui} is unbiased. However, we generally do not know the true propensity score in the observational study. Here, we provide the bias (B) and variance (V) of the IPS estimator with estimated propensity score $\hat{e}(\mathbf{x}_{ui})$ (see Appendix A.1 and A.2 for derivations):

$$\text{B}^{\text{IPS}} = \left|\left(Q_{ui}^{(1)} + \frac{\hat{e}_{ui}}{1-\hat{e}_{ui}}Q_{ui}^{(0)}\right)\delta_{ui}^{(1)}\right|, \tag{5}$$

$$\begin{aligned}\text{V}^{\text{IPS}} = {} & \mathbb{E}\left[\left(\epsilon_{ui}^{(1)}\right)^2 \mid \mathbf{x}_{ui}\right] + \mathbb{E}\left[\left(\epsilon_{ui}^{(0)}\right)^2 \mid \mathbf{x}_{ui}\right] + \frac{e_{ui}}{1-e_{ui}}\left(Q_{ui}^{(0)}\right)^2\left(1-\delta_{ui}^{(0)}\right)^2 \\ & + \frac{2e_{ui}(1-e_{ui})}{\hat{e}_{ui}(1-\hat{e}_{ui})}Q_{ui}^{(1)}Q_{ui}^{(0)} + \frac{1-e_{ui}}{e_{ui}}\left(Q_{ui}^{(1)}\right)^2\left(1-\delta_{ui}^{(1)}\right)^2\end{aligned} \tag{6}$$

where $Q_{ui}^{(1)} = \mathbb{E}\left[c_{ui}^{(1)}|\mathbf{x}_{ui}\right]$, $Q_{ui}^{(0)} = \mathbb{E}\left[c_{ui}^{(0)}|\mathbf{x}_{ui}\right]$, $\delta_{ui}^{(1)} = 1 - \frac{e_{ui}}{\hat{e}_{ui}}$ and $\delta_{ui}^{(0)} = 1 - \frac{1-e_{ui}}{1-\hat{e}_{ui}}$. Note that $e(\mathbf{x}_{ui})$ and $\hat{e}(\mathbf{x}_{ui})$ are simplified by e_{ui} and $\hat{e}_{ui}$ for notation clarity. $\epsilon_{ui}^{(1)}$ and $\epsilon_{ui}^{(0)}$ denote $\left(c_{ui}^{(1)} - Q_{ui}^{(1)}\right)\frac{R_{ui}}{\hat{e}_{ui}}$ and $\left(c_{ui}^{(0)} - Q_{ui}^{(0)}\right)\frac{1-R_{ui}}{1-\hat{e}_{ui}}$, respectively. The derived bias and variance suggest that if e_{ui} is close to 1 or 0, the IPS suffers from a large variance and bias with inaccurate estimated propensity scores.

Direct Model (DM) Estimator. Instead of using non-parametric the IPS estimator, we can also directly use the parametric method to estimate the recommendation causal effect. Parametric methods directly model the relation between the confounding, treatment, and causal effect via two regression models [30] as follows:

$$\hat{\tau}_{ui}^{\text{DM}} = \hat{Q}_{ui}^{(1)} - \hat{Q}_{ui}^{(0)}, \tag{7}$$

where $\hat{Q}_{ui}^{(r)} = \mathbb{E}\left[\hat{c}_{ui}^{(r)}|\mathbf{x}_{ui}, r\right]$ is a direct estimation of the conditional outcome using samples from observation dataset. The bias (B) of the parametric model can be represented as follows:

$$\text{B}^{\text{DM}} = \left|\left(\hat{Q}_{ui}^{(1)} - \hat{Q}_{ui}^{(0)}\right) - \left(Q_{ui}^{(1)} - Q_{ui}^{(0)}\right)\right|. \tag{8}$$

Since $\hat{Q}_{ui}^{(1)} - \hat{Q}_{ui}^{(0)}$ is a constant given $\mathbf{x}_{ui}$, the variance $\text{V}^{\text{DM}} = 0$. Thus the parametric DM estimator has the smaller variance compared to the IPS estimator. However, as pointed out by [33], the DM estimator is very sensitive to model misspecification and will lead to large bias if the two groups differ considerably in covariates.

5 UNBIASED AND ROBUST ESTIMATOR

In this section, we first introduce the doubly robust estimator for casual effect in recommendation, and then propose a deep variational information bottleneck approach to learn the estimator effectively.

5.1 The Doubly Robust Estimator

Based on our analysis above, the DM estimator has zero variance but a large bias in practice due to the model misspecification; while the IPS estimator often suffers from high variance. This motivates us to use both parametric model and propensities to overcome the limitations of the DM and IPS approaches. A conceptually straightforward way is to combine the IPS and DM as a joint estimator: $\alpha\hat{\tau}_{ui}^{\text{IPS}} + (1-\alpha)\hat{\tau}_{ui}^{\text{DM}}$. However, such linear combination estimator is still biased even when the propensities are accurate but the model is not accurate. We observe that this weakness can be addressed by designing an estimator in a doubly robust (DR) [5, 13, 31] way such that the bias remains zero and variance is small even with inaccurate model as long as the propensities are accurate. The key idea of DR estimator [5, 13, 31] is to add a correction term obtained by importance weighting of the difference between observed outcomes and predicted outcomes. Following this idea, the DR causal effect estimator for recommendation is given as follows:

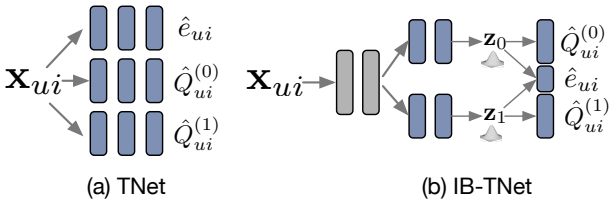
$$\hat{\tau}_{ui}^{\text{DR}} = \left(c_{ui}^{(1)} - \hat{Q}_{ui}^{(1)}\right)\frac{r_{ui}}{\hat{e}_{ui}} - \left(c_{ui}^{(0)} - \hat{Q}_{ui}^{(0)}\right)\frac{1-r_{ui}}{1-\hat{e}_{ui}} + \left(\hat{Q}_{ui}^{(1)} - \hat{Q}_{ui}^{(0)}\right). \tag{9}$$

The bias (B) and variance (V) of this DR estimator with estimated propensity score and model can be represented as follows (see Appendix A.1 and A.2 for derivations):

$$\text{B}^{\text{DR}} = \left|\delta_{ui}^{(1)}\left(q_{ui}^{(1)} + \frac{\hat{e}_{ui}}{1-\hat{e}_{ui}}q_{ui}^{(0)}\right)\right|, \tag{10}$$

$$\begin{aligned}\text{V}^{\text{DR}} = {} & \mathbb{E}\left[\left(\epsilon_{ui}^{(1)}\right)^2 |\mathbf{x}_{ui}\right] + \mathbb{E}\left[\left(\epsilon_{ui}^{(0)}\right)^2 |\mathbf{x}_{ui}\right] + \frac{2e_{ui}(1-e_{ui})}{\hat{e}_{ui}(1-\hat{e}_{ui})}q_{ui}^{(1)}q_{ui}^{(0)} \\ & + \frac{1-e_{ui}}{e_{ui}}\left(q_{ui}^{(1)}\right)^2\left(1-\delta_{ui}^{(1)}\right)^2 + \frac{e_{ui}}{1-e_{ui}}\left(q_{ui}^{(0)}\right)^2\left(1-\delta_{ui}^{(0)}\right)^2,\end{aligned} \tag{11}$$

where $q_{ui}^{(1)} = \hat{Q}_{ui}^{(1)} - Q_{ui}^{(1)}$ and $q_{ui}^{(0)} = \hat{Q}_{ui}^{(0)} - Q_{ui}^{(0)}$. The derived bias in Eq. (10) indicates that $\hat{\tau}_{ui}^{\text{DR}}$ is an unbiased estimator of causal effect if either propensity score is correct ($\hat{e}_{ui} = e_{ui} \rightarrow \delta_{ui}^{(1)} = 0$) or direct model is correct $q_{ui}^{(0)} = q_{ui}^{(1)} = 0$. The variance $\text{V}^{\text{DR}} < \text{V}^{\text{IPS}}$ under the condition of $\left|q_{ui}^{(0)}\right| < Q_{ui}^{(0)}$ or $\left|q_{ui}^{(1)}\right| < Q_{ui}^{(1)}$. This is easily satisfied since the error is usually small if universal function approximators such as neural networks are used. Thus, the DR estimator is robust on variance and instability issues. In other words, even if the DM estimator does not perform well here, the resulting DR estimator is expected to be more accurate than the IPS estimator.



(a) TNet (b) IB-TNet

Figure 1: Overview of architectures. (a) is Triple-head Net (TNet) (b) is our information bottleneck T-learner (IB-TNet). The encoders map covariates to latent representations. The decoder is used to predict propensity score and outcomes.

Note that the idea of correcting confounding bias via doubly robust estimators has been investigated in statistics in the context of regression with incomplete data [5, 13, 31]. The main differences between these studies and our work are two folds: (i) These works focus on evaluation instead of learning; while we explore the problem of learning a recommendation policy based on a dataset consisting of confounding bias and give a theoretical analysis on the learnability; and (ii) We focus on the ranking not the regression problem, which is one of the most important characteristics of the RS. In what follows, we show how to effectively learn the DR estimator with information bottleneck and a ranking algorithm based on the learned DR estimator.

5.2 Learning with Information Bottleneck

Typically, learning the DR estimator requires optimizing the propensity score model $\hat{e}_{ui}$, and two outcome models $\hat{Q}_{ui}^{(1)}$ and $\hat{Q}_{ui}^{(0)}$ given the observed dataset $\mathcal{D}$. Recently, Farrell et al. [11] give theoretical justification for the use of neural networks (NNs) to model propensity scores and conditional outcomes. Thus, in this paper, we adopt neural networks to parameterize the propensity score and conditional outcomes. The most straightforward implementation consists of using separate networks for $\hat{e}_{ui}$ and $\hat{Q}_{ui}$ which would be a good choice asymptotically (see Figure 1(a)). However, complex interactions within and across models need to be properly addressed for optimal predictive power. That is, it may be more efficient to share representations between different propensity score and conditional outcomes. In addition, the learned models are also prone to be overfitting [4] where models fit the training data very well but generalizes poorly to the testing data. It will suffer from the overfitting issue more severely, when we utilize the NNs to parameterize them with limited data and high-dimensional covariates.

Thus, in this paper, we adopt NNs to model the non-linear relationship and propose to regularize the propensity score and conditional outcome models based on deep information bottleneck (DIB) [3]. The DIB framework [3] has been recently studied to address multi-view problems [12] and learn disentangled representations as shown in beta-VAE [17]. In this paper, we frame our casual estimation framework under the information bottleneck principle based on the following considerations: (i) *we consider covariate $\mathbf{x}_{ui}$ as noisy proxies for the true unobservable confounders.* (ii) *The variational information bottleneck can capture the uncertainties and improve the model generalization by adaptively updating.*

Specifically, the bottleneck can be incorporated by introducing two encoders $q_\phi(\mathbf{z}_0, \mathbf{z}_1|\mathbf{x}_{ui})$ that map the $\mathbf{x}_{ui}$ to a latent distributions over $\mathbf{z}_0$ and $\mathbf{z}_1$, where $\mathbf{z}_0$ encodes information for estimating $Q_{ui}^{(0)}$ and $\mathbf{z}_1$ extracts information for estimating $Q_{ui}^{(1)}$. By assuming that there exists a common feature space underlying both propensity score and outcomes, we jointly utilize $\mathbf{z}_0$ and $\mathbf{z}_1$ to model the propensity score to take away their representation capacities. With the shared representation, ideally, the model itself can choose a tradeoff between outcome accuracy and the propensity-score representation. We also enforce an upper bound I_c on the mutual information between the encoding and the original features, which results in a regularized loss for each observed triple $(\mathbf{x}_{ui}, r_{ui}, c_{ui})$:

$$\mathcal{L}(\theta, \phi) = -\mathbb{E}_{\mathbf{z}_0,\mathbf{z}_1\sim q_\phi(\mathbf{z}_0,\mathbf{z}_1|\mathbf{x}_{ui})}\big[\log p_\theta(r_{ui}|\mathbf{z}_0, \mathbf{z}_1) + \log p_\theta(c_{ui}|\mathbf{x}_{ui}, r_{ui})\big]$$
$$\text{s.t.} \quad I(X_{ui}, Z_0) + I(X_{ui}, Z_1) \le I_c, \tag{12}$$

where probability $p_\theta(r_{ui}|\mathbf{z}_0, \mathbf{z}_1) = \text{Ber}(\hat{e}(\mathbf{z}_0, \mathbf{z}_1))$ is essentially the binary propensity score model. $p_\theta(c_{ui}|\mathbf{x}_{ui}, r_{ui} = 0) = \text{Ber}(\hat{Q}^{(0)}(\mathbf{z}_0))$ and $p_\theta(c_{ui}|\mathbf{x}_{ui}, r_{ui} = 1) = \text{Ber}(\hat{Q}^{(1)}(\mathbf{z}_1))$ are for two outcome models. The mutual information $I(X, Z_0)$ is defined according to:

$$I(X_{ui}, Z_0) = \int q(\mathbf{x}_{ui}) q_\phi(\mathbf{z}_0|\mathbf{x}_{ui}) \log \frac{q_\phi(\mathbf{z}_0|\mathbf{x}_{ui})}{q(\mathbf{z}_0)} d\mathbf{z}_0, \tag{13}$$

where $q(\mathbf{x}_{ui})$ is the empirical data distribution which can be represented by each sample $\mathbf{x}_{ui}$. Since computing the marginal distribution $q(\mathbf{z}_0) = \int q(\mathbf{x}_{ui}) q_\phi(\mathbf{z}_0|\mathbf{x}_{ui}) d\mathbf{z}_0$ can be challenging. Instead, we consider using the variational lower bound [1] of the mutual information by introducing a variational approximation $p(\mathbf{z}_0) = \mathcal{N}(0, I)$ modeled with standard Gaussian distribution to this marginal:

$$\begin{aligned} I(X_{ui}, Z_0) &\le \int q(\mathbf{x}_{ui}) q_\phi(\mathbf{z}_0|\mathbf{x}_{ui}) \log \frac{q(\mathbf{z}_0|\mathbf{x}_{ui})}{p(\mathbf{z}_0)} \\ &= \mathbb{E}_{q(\mathbf{x}_{ui})}\big[\text{KL}\left(q\left(\mathbf{z}_0|\mathbf{x}_{ui}\right)||p\left(\mathbf{z}_0\right)\right)\big]. \end{aligned} \tag{14}$$

Similarly, we can also obtain the variational bound for $\mathbf{z}_1$: $I(X_{ui}, Z_1) \le \mathbb{E}_{q(\mathbf{x}_{ui})}\big[\text{KL}\left(q\left(\mathbf{z}_1|\mathbf{x}_{ui}\right)||p\left(\mathbf{z}_1\right)\right)\big]$. The upper bound $\tilde{\mathcal{L}}(\theta, \phi)$ of the objective $\mathcal{L}(\theta, \phi)$ in Eq. (12) can be optimized as follows:

$$\tilde{\mathcal{L}}(\theta, \phi) = -\mathbb{E}_{\mathbf{z}_0,\mathbf{z}_1\sim q_\phi(\mathbf{z}_0,\mathbf{z}_1|\mathbf{x}_{ui})}\big[\log p_\theta(r_{ui}|\mathbf{z}_0, \mathbf{z}_1) + \log p_\theta(c_{ui}|\mathbf{x}_{ui}, r_{ui})\big]$$
$$\text{s.t. KL}\left(q\left(\mathbf{z}_0|\mathbf{x}_{ui}\right)||p\left(\mathbf{z}_0\right)\right) + \text{KL}\left(q\left(\mathbf{z}_1|\mathbf{x}_{ui}\right)||p\left(\mathbf{z}_1\right)\right) \le I_c. \tag{15}$$

To solve this problem, the constraint can be subsumed into the objective by introducing the Lagrange multiplier β:

$$\tilde{\mathcal{L}}(\theta, \phi, \beta) = \min_{\theta,\phi} \max_{\beta\ge 0} -\mathbb{E}_{\mathbf{z}_0,\mathbf{z}_1\sim q_\phi(\mathbf{z}_0,\mathbf{z}_1|\mathbf{x}_{ui})}\big[\log p_\theta(r_{ui}|\mathbf{z}_0, \mathbf{z}_1) \tag{16}$$
$$+ \log p_\theta(c_{ui}|\mathbf{x}_{ui}, r_{ui})\big] + \beta\left(\text{KL}[q(\mathbf{z}_1, \mathbf{z}_0|\mathbf{x}_{ui})\|p(\mathbf{z}_1)p(\mathbf{z}_0)] - I_c\right),$$

where $p(\mathbf{z}_1) = p(\mathbf{z}_0) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ are standard Gaussian distributions. $q_\phi(\mathbf{z}_1, \mathbf{z}_0|\mathbf{x}_{ui}) = q_\phi(\mathbf{z}_0|\mathbf{x}_{ui}) q_\phi(\mathbf{z}_1|\mathbf{x}_{ui})$, where $q_\phi(\mathbf{z}_1|\mathbf{x}_{ui})$ and $q_\phi(\mathbf{z}_0|\mathbf{x}_{ui})$ are modeled with two Gaussian distributions with parameterized mean and diagonal covariance matrix. Figure 1(b) shows our model architecture. As we will demonstrate in our experiments, enforcing a specific mutual information budget between $\mathbf{x}_{ui}$ and $(\mathbf{z}_0, \mathbf{z}_1)$ naturally regularizes for model generalization and is critical for good performance. Instead of fixing β [17], we adaptively update β via dual gradient descent to enforce a constraint I_c on the mutual information. This formulation can automate the value of β, as shown below:

$$\theta, \phi \leftarrow \arg\min_{\theta,\phi} \tilde{\mathcal{L}}(\theta, \phi, \beta), \tag{17}$$

$$\beta \leftarrow \max\Big(0, \beta + \alpha_\beta\left(\text{KL}[q(\mathbf{z}_1, \mathbf{z}_0|\mathbf{x}_{ui})\|p(\mathbf{z}_1)p(\mathbf{z}_0)] - I_c\right)\Big), \tag{18}$$

where α_β is the learning rate in dual gradient descent. Intuitively, our loss is a combination of outcome-loss and propensity score

loss, adaptively regularized by KL-divergence on the latent representations to encourage better generalization. By optimizing the triplet jointly, the proposed model can choose a tradeoff between predictive accuracy and the propensity-score representation.

After optimizing the model, we can get the causal estimation $\hat{\tau}_{ui}^{\mathrm{DR}}$ from the learned model: we calculate the propensity score $\hat{e}(\mathbf{z}_0, \mathbf{z}_1)$, and conditional outcomes $\hat{Q}^{(0)}$ and $\hat{Q}^{(1)}$ based on samplings from $q_\phi(z_0, z_1|\boldsymbol{x})$. With $\hat{e}(\mathbf{z}_0, \mathbf{z}_1)$, $\hat{Q}^{(0)}$ and $\hat{Q}^{(1)}$, we get $\hat{\tau}_{ui}^{\mathrm{DR}}$ with Eq. (9).

6 UNBIASED AND ROBUST LEARNING

Given the learned DR estimator, in this section, we propose Doubly Robust Unbiased Learning (DRUL) to optimize the causal ranking metric by connecting the estimation and learning steps. Our key insight in this section is that we can decrease the learning bound by slightly modifying the vanilla DR estimator discussed in section 5.1.

6.1 Learning Bound with DR Estimator

Recall that our goal is to optimize the ideal causal metric $R^{\mathrm{Ideal}}(\hat{Z})$ in Eq. (2). As already shown in section 5.1, the unobservable τ_{ui} in causal DCG (see Eq. (2)) can be estimated now based on the learned DR estimator. Then we can approximate the causal metric via $\hat{\tau}_{ui}^{\mathrm{DR}}$:

$$\hat{R}^{\mathrm{DR}}(\hat{Z}) = \frac{1}{|\mathcal{U}|}\sum_u \sum_i \hat{R}^{\mathrm{DR}}(\hat{z}_{ui}) = \frac{1}{|\mathcal{U}|}\sum_u \sum_i \lambda(\hat{z}_{ui})\hat{\tau}_{ui}^{\mathrm{DR}}, \quad (19)$$

where the DR estimator has double robustness: it is unbiased if either propensity score is correct ($\hat{e}_{ui} = e_{ui}$) or outcome models is correct $q_{ui}^{(0)} = q_{ui}^{(1)} = 0$. In general, the unbiased learning of the causal metric is a two-stage process, In the first stage, we infer the casual effect $\hat{\tau}_{ui}^{\mathrm{DR}}$ as shown in the last section, and in the second stage we can utilize Empirical Risk Minimization (ERM) framework to conduct the policy learning process based on the estimator $\hat{\tau}_{ui}^{\mathrm{DR}}$.

With the proposed $\hat{\tau}_{ui}^{\mathrm{DR}}$ and the IB training objective, $\hat{\tau}_{ui}^{\mathrm{DR}}$ yields unbiased and robust estimates of $\hat{\tau}_{ui}$. However, what is important for RS is the performance of the downstream learning not the estimation, which is limited by the independent two-stage process and the learned estimator may be suboptimal for the ranking model on given tasks. *Can we modify the design and training of casual effect estimation in the first stage in order to improve the learning performance in the second stage?* We address this by first giving the tail bound of the proposed DRUL with finite samples.

Proposition 1 (Tail Bound of Unbiased Learning). *Given the true propensity score e_{ui}, the estimated outcome model $\hat{Q}_{ui}^{(1)}$ and $\hat{Q}_{(ui)}^{(0)}$, for the given independent and identically distributed dataset $\{(\mathbf{x}_{ui}, r_{ui}, c_{ui})\}_{u=1,i=1}^{|\mathcal{U}|\times|\mathcal{I}|}$, with probability $1-\eta$, the $\hat{R}^{\mathrm{DR}}(\hat{z}_{ui})$ does not deviate from its expectation by more than (see Appendix B.1):*

$$\left|\hat{R}^{\mathrm{DR}}(\hat{z}_{ui}) - R^{\mathrm{Ideal}}(\hat{z}_{ui})\right| \le \frac{1}{|\mathcal{U}|}\sqrt{\frac{\log\frac{2}{\eta}}{2}}\sqrt{\sum_{u,i}\lambda(\hat{z}_{ui})^2 d_{ui}^2}, \quad (20)$$

where $d_{ui}^2 = (\frac{c_{ui}^{(1)} - \hat{Q}_{ui}^{(1)}}{e_{ui}} + \frac{c_{ui}^{(0)} - \hat{Q}_{ui}^{(0)}}{1-e_{ui}})^2$.

Given the tail bound, we further present the following corollary to compare our tail bound of unbiased learning and previous bound based on IPS for the causal effect of the recommendation.

Corollary 1.1 (Tighter Bound). *Given $0 < \hat{Q}_{ui}^{(1)} \le 2c_{ui}^{(1)}$ and $0 < \hat{Q}_{ui}^{(0)} \le 2c_{ui}^{(0)}$, the bound of the DR estimator will be tighter than that of the IPS estimator. (see Appendix B.2 for proofs).*

This corollary shows that our causal DCG estimator consistently improves the estimation error bound compared with previous work [38]. The bias-variance analysis in the last section and the Corollary 1.1 show that our proposed unbiased learning algorithm based on the DR estimator outperforms the previous method [38] and has better statistical properties compared to the IPS estimator.

Proposition 1 above shows that the ranking tail bound is positively correlated with the magnitude of $d_{ui} = \frac{c_{ui}^{(1)} - \hat{Q}_{ui}^{(1)}}{\hat{e}_{ui}} + \frac{c_{ui}^{(0)} - \hat{Q}_{ui}^{(0)}}{1-\hat{e}_{ui}}$ in which each outcome model is weighted by the propensity score. In the last section, we propose a information bottleneck to learn the propensity score the conditional outcomes. To further improve the downstream ranking learning, we modify the loss in Eq. (12) in order to decrease the magnitude of d_{ui} as follows:

$$\begin{aligned}\mathcal{L} = -\mathbb{E}_{z_0,z_1\sim q_\phi(z_0,z_1|\mathbf{x}_{ui})}\left[\log p_\theta(r_{ui}|\mathbf{z}_0,\mathbf{z}_1) + w_{ui}\log p_\theta(c_{ui}|\mathbf{x}_{ui}, r_{ui})\right] \\ \text{s.t.} \quad I(X, Z_0) + I(X, Z_1) \le I_c. \end{aligned} \quad (21)$$

where $w_{ui} = r_{ui}\cdot\frac{1}{e_{\bar{\theta}}(\mathbf{z}_0,\mathbf{z}_1)} + (1-r_{ui})\cdot\frac{1}{1-e_{\bar{\theta}}(\mathbf{z}_0,\mathbf{z}_1)}$ and $\bar{\theta}$ indicates that gradients for θ are not being computed through it. Compared with Eq. (12), we raise the log-likelihood $\log p_\theta(c_{ui}|\mathbf{x}_{ui}, r_{ui})$ by a weight w_{ui} to reduce the d_{ui}. With this objective, our model yields consistent estimates of causal effect and downstream performance. We can optimize this modified loss the same as we do in section 5.2.

6.2 Optimizing the Ranking Metric

In this section, we show how to conduct the second learning stage given the estimated $\hat{\tau}_{ui}^{DR}$ in the first stage. We adopt the ERM framework [39] to conduct the unbiased learning stage:

$$\hat{Z}^{ERM} = \underset{\hat{Z}\in\mathcal{H}_{\hat{z}}}{\operatorname{argmin}}(-\hat{R}^{DR}(\hat{Z})), \quad (22)$$

where $\mathcal{H}_{\hat{Z}}$ is a hypothesis space of prediction ranking $\hat{Z}$.

Upper Bound of Causal DCG. In practically, we would like to learn a scoring function $f_\theta(\mathbf{x}_{ui})$ to rank candidate items $\mathcal{I}$ for user u by the score. However the resulting rank position $\hat{z}_{ui}$ is a discrete step function of the score. We need to make it differentiable to optimize. Thanks to the hinge-loss upper bound [24], we have:

$$\begin{aligned}\hat{z}_{ui} - 1 &= \sum_{j\in\mathcal{I}, j\neq i}\mathbb{1}(f_\psi(\mathbf{x}_{uj}) > f_\psi(\mathbf{x}_{ui}))) \\ &\le \sum_{j\in\mathcal{I}, j\neq i}\max\left(1 - \left(f_\psi(\mathbf{x}_{ui}) - f_\psi(\mathbf{x}_{uj})\right), 0\right).\end{aligned} \quad (23)$$

With this upper bound and our learned DR estimator, instead of directly optimizing Eq. (22), we can optimize the following differentiable bound for the causal DCG:

$$\begin{aligned}\hat{\psi} = \underset{\psi}{\operatorname{argmin}}(-\hat{R}^{DR}(\hat{Z})) = \underset{\psi}{\operatorname{argmin}}\frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\sum_{i\in\mathcal{I}} \\ -\lambda\big(1 + \sum_{j\in\mathcal{I}, j\neq i}\max\left(1 - \left(f_\psi(\mathbf{x}_{ui}) - f_\psi(\mathbf{x}_{uj})\right), 0\right)\big)\hat{\tau}_{ui}^{DR}.\end{aligned} \quad (24)$$

7 EXPERIMENTS

In this section, we conduct experiments to evaluate the effectiveness of the our frameworks on the unbiased evaluation and learning tasks. Specifically, we aim to answer the following questions:

(RQ1) How does the proposed DR estimator perform in the unbiased evaluation task for the recommendation?
(RQ2) Does the proposed DR estimator work better in the causal effect estimation task compared to state-of-the-art estimators?
(RQ3) How does the proposed IB improve the performance?
(RQ4) Can the proposed unbiased learning method work better than other biased and unbiased methods in real-word dataset, and How does it affect the recommendation results?

7.1 Unbiased Evaluation Performance (RQ1)

Setups. To conduct the unbiased evaluation with confounding bias, we experiment with the publicly available dataset: *Dunnhumby* [38]. This dataset includes purchase and promotion logs at a retailer. It provides product category information. Other public datasets are either missing recommendation logs or recording user interactions only for recommended items. Since there is no ground truth because of the unobservable nature of causal effect, following existing work [38, 42], we construct a semi-synthetic dataset. Note that although the proposed estimator only uses observable variables, the ground truth is required for the evaluation. In addition, constructing a semi-synthetic dataset is a widely used procedure in causal inference literature [38, 42]. Strictly following to [38], we employ the following procedure to create a biased datasets. (1) We preprocess the *Dunnhumby* dataset to get the observation $\{r_{uit}, c_{uit}\}$, where t denotes the t-th week. (2) We then model the purchase probabilities $Q_{ui}^{(1)}$ and $Q_{ui}^{(0)}$ with and without recommendation for each user-item pair. (3) We model the propensities by simulating a common situation where a currently running recommender tends to select items that match the preference of the users with higher probabilities. (4) Sampling the observed data. $c_{ui}^{(1)} \sim \mathcal{B}\left(Q_{ui}^{(1)}\right), c_{ui}^{(0)} \sim \mathcal{B}\left(Q_{ui}^{(0)}\right), r_{ui} \sim \mathcal{B}(e_{ui})$, where $\mathcal{B}$ denotes the Bernoulli distribution. The causal effect τ_{ui} and observed outcomes c_{ui} can be obtained as follows: $\tau_{ui} = c_{ui}^{(1)} - c_{ui}^{(0)}$ and $c_{ui} = r_{ui}c_{ui}^{(1)} + (1 - r_{ui})c_{ui}^{(0)}$. We use the numbers of purchases and recommendations during previous four weeks as the proxy of covariates $\mathbf{x}_{ui}$ [37]. This sampling is repeated n times to generate dataset. Training, validation, and test sets are independently sampled for $n_{train} = 10$, $n_{val} = 1$, and $n_{test} = 10$ times, respectively. After preprocessing, 2,309 users and 11,331 items are left. Note that we only use $(c_{ui}, r_{ui}, \mathbf{x}_{ui})$ as the training observable samples, while τ_{ui} is just used for the evaluation purpose.
Compared Estimators: We compare the following estimators: *Naive*, *IPS*, *SNIPS*, *DM* and proposed DR estimator. For our *DR*, we denote TNet-based as *DRT* and IB-TNet-based as *DRIB*. We compare the unbiased evaluation performance by the mean absolute error (MAE) which can evaluate the sum of the bias and variance [43]:

$$MAE(\hat{\mathcal{R}}) = |\mathcal{R}_{GT}(\hat{Z}) - \hat{\mathcal{R}}(\hat{Z})|, \tag{25}$$

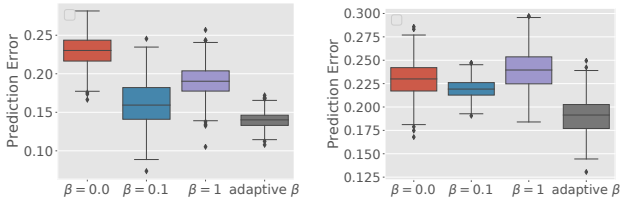
where $\hat{Z}$ is the set of outputs by the candidate recommender, and $\hat{\mathcal{R}}(\hat{Z})$ the candidate recommender with one of the compared estimators. The MAE evaluates an estimators' ability to accurately evaluate the performance of candidate recommenders. For the ground-truth ranking metrics $\mathcal{R}_{GT}(\hat{Z})$, we use causal DCG (cDCG) and casual Precision (cP)@10 in all experiments model. $cP@K = \frac{1}{|\mathcal{U}|}\sum_u \sum_i I(\hat{Z}_{ui} \le K)\tau_{ui}$ and $cDCG = \frac{1}{|\mathcal{U}|}\sum_u \sum_i \frac{I}{\log(1+\hat{Z}_{ui})}\tau_{ui}$.
Candidate Recommenders: To prove effectiveness of our estimator, we use several candidate recommenders for evaluation: Pop: Items are ranked by the global popularity. MF [19] is a basic baseline for recommendation. BPR: The MF method with pairwise objective for recommendation. CausE [6]: The joint training of two MFs with and without recommendations. CausE-Prod [6]: The variant of CausE, where the two MFs share the common user factors. ULMF [37]: A biased pointwise learning method for the causal effect of recommendation. ULBPR [37]: A biased pairwise learning method for the causal effect of recommendation. DLCE [38]: A unbiased learning method based on IPS for the casual effect of recommendation. DRUL: Our proposed unbiased learning method in the section 6 for the casual effect of recommendation.
Parameter Settings All the base recommendation models except Pop are implemented by matrix factorization models [26]. We set the latent dimensions to 200. The regularization coefficient $\lambda \in \{1e-4, 3e-4, \cdots, 1e-1, 3e-1\}$. BGD (batch gradient descent) was employed, and the initial learning rate was set to 0.0001. For *IPS*, *DM* and *DRT*, we use separate NN-based logistic regression model to estimate the propensity score and outcome, respectively. For *DRIB*, the hidden layer size is 200 for the means of $\mathbf{z}_1$ and $\mathbf{z}_0$ and 100 for the conditional outcome and propensity layers. For *DRIB*, we set the mutual information constraint I_c to 0.3 and the learning rate α_β in dual gradient descent to 0.001. We use the baseline implementations provided by the authors. Hyperparameters were tuned in the validation phase to maximize cP@10.
Results. Table 2 shows that the proposed DR estimator outperforms the other estimators in all cases. All reported results are averaged over five runs, and the improvement are statistically significant. From Table 2, we observe: (1) As suggested by our theoretical analysis, the *DR* estimators significantly outperform the other estimators in the model evaluation task. The results verify that it can help the estimation of the ranking performance of recommenders in a real-world offline setting. (2) The biased naive estimator has worse performance, which demonstrates that using unbiased estimator is important when evaluating recommenders under the confounding bias settings. (3) Though they are both DR-based estimator, our *DRIB* outperforms *DRT*, which shows that sharing representations and incorporating information bottleneck can make it more powerful to achieve better evaluation performance.

7.2 Causal Estimation Performance (RQ2)

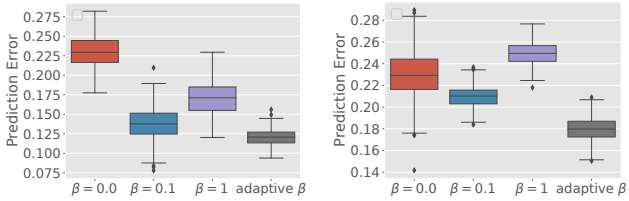
Setups. We have shown that our methods do help improve performance on unbiased ranking evaluation. In this subsection, we further study if our methods can achieve better performance on the causal effect estimation task. We conduct the experiment on IHDP [42, 54], a widely used semi-synthetic dataset in the causal inference literature. We randomly split the data into train/val/test with proportion as 10/27/63 and report the in sample (train and validation) and out of sample (test) estimation errors. We report mean absolute error between the estimated and the actual average treatment effect (ATE), i.e., $\Delta = |\hat{\psi} - (1/n\sum_i Q(1, x_i) - Q(0, x_i))|$, where $\hat{\psi}$ is the true ATE. We compare *DRIB* with state-of-the-art neural networks listed in Table 3 for causal effect estimation. For all models, the hidden layer size is 200 for the shared representation layers and 100 for the conditional outcome layers. We train them using stochastic gradient descent with momentum. For Dragonnet [42], we set the α and β to 1. For *DRIB*, we set the constraint

Table 2: The unbiased evaluation performance on the recommendation of different estimators.

	MAE of cP@10						MAE of cDCG					
	Naive	IPS	DM	SNIPS	DRT	DRIB	Naive	IPS	DM	SNIPS	DRT	DRIB
Pop	.0506	.0311	.0274	.0298	.0252	**.0241**	.366	.345	.312	.323	.293	**.273**
MF	.0312	.0289	.0256	.0271	.0229	**.0211**	.312	.287	.268	.271	.247	**.227**
BPR	.0324	.0227	.0243	.0225	.0211	**.0196**	.343	.318	.278	.299	.268	**.254**
CausE	.0268	.0241	.0219	.0233	.0193	**.0185**	.310	.299	.267	.281	.250	**.242**
Caus-Prod	.0177	.0158	.0162	.0151	.0138	**.0121**	.288	.259	.233	.239	.213	**.192**
ULBPR	.0168	.0114	.0093	.0086	.0071	**.0063**	.270	.249	.218	.222	.200	**.196**
BLCE	.0212	.0175	.0162	.0151	.0130	**.0123**	.252	.241	.198	.206	.154	**.141**
DLCE	.0318	.0234	.0211	.0217	.0168	**.0147**	.283	.257	.229	.238	.213	**.202**
DRUL	.0233	.0214	.0196	.0191	.0179	**.0166**	.256	.244	.228	.230	.205	**.197**



(a) Δ_{in} with $I_c = 0.1$ (b) Δ_{out} with $I_c = 0.1$



(c) Δ_{in} with $I_c = 0.3$ (d) Δ_{out} with $I_c = 0.3$

Figure 2: The estimation errors on IHDP with different β.

Table 3: The causal effect estimation performance on the IHDP dataset. Estimators are computed with the training and validation (Δ_{in}), test set (Δ_{out}).

Method	Δ_{in}	Δ_{out}
BNN [25]	0.37 ± .03	0.42 ± .03
TARNET [40]	0.26 ± .01	0.28 ± .01
CFR Wass [40]	0.25 ± .01	0.27 ± .01
CEVAEs [29]	0.34 ± .01	0.46 ± .02
GANITE [54]	0.43 ± .05	0.49 ± .05
Dragonnet [42]	0.14 ± .01	0.20 ± .01
DRIB ($\beta = 0.0$)	0.23 ± .02	0.31 ± .02
DRIB (adaptive β, I_c=0.0)	0.17 ± .02	0.22 ± .02
DRIB (adaptive β, I_c=0.1)	0.14 ± .01	0.19 ± .02
DRIB (adaptive β, I_c=0.3)	**0.12 ± .01**	**0.18 ± .01**
DRIB (adaptive β, I_c=1.0)	0.13 ± .02	0.19 ± .01
DRIB (adaptive β, I_c=3.0)	0.16 ± .03	0.21 ± .03
DRIB (adaptive β, I_c=10)	0.15 ± .01	0.21 ± .02

I_c to 0.3 and the learning rate α_β in dual gradient descent to 0.001.

Results. Table 3 reports the estimation error of a number of approaches on the IHDP dataset. (1) The results from Table 3 indicate that the proposed *DRIB* (adaptive β, I_c=0.3) estimators based on deep information bottleneck achieve the state-of-the-art performance. (2) The finding that other *DRIB* methods outperform *DRIB* (β=0.0) shows that incorporating information bottleneck does help improve estimation performance. (3) Table 3 also shows that tuning the KL constraint I_c, leading to better estimation performance.

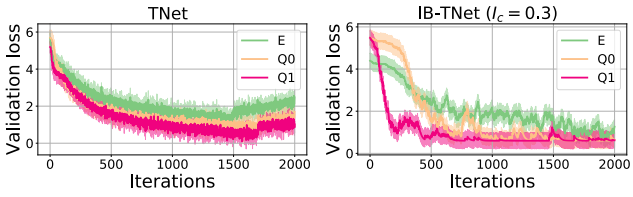


Figure 3: The validation losses of TNet and IB-TNet.

7.3 The Effects of Deep IB (RQ3)

Setups. In this section, we further study the effect of proposed information bottleneck. We conduct experiments on both *Dunnhumby* and IHDP datasets and follow the same settings in RQ1 and RQ2.

Results. Table 5 shows the performance of *DRIB* by varying I_c on *Dunnhumby*. As we can see, *DRIB* achieves the best recommendation performance when I_c=0.3. We also can see that the performance is worse when I_c=10. The reason is that I_c controls the mutual information between the encoding and the original features, a too large I_c will lead the latent representations to neglect the supervised signal. Table 5 shows that we can vary I_c to achieve better model evaluation performance. To evaluate the effects of the adaptive β updates, we compare *DRIB* trained with different fixed values of β and adaptive updated β using dual gradient in Eq. (18). Fig. 2 shows that the networks trained using dual descent to update β achieves better performance compared with fixed β. We plot the the learning curves of validation losses of propensity score (E) and two outcomes Q1 and Q2 in TNet and IB-TNet as shown in Figure 3. We can find TNet suffers from overfitting when the number of iterations > 1700. In contrast, IB-TNet does not suffer from this, which shows that representations learned by IB improve generalization by ignoring irrelevant parts present in noisy covariates.

7.4 Unbiased Learning Performance (RQ4)

In this section, we evaluate and compare our unbiased learning algorithm, i.e., the DRUL with baselines on the real-world dataset.

Setups. Since we need both recommendation and interaction logs for this experiment, we still evaluate methods on *Dunnhumby*. Instead of evaluating the estimators as in RQ1, we evaluate the proposed learning algorithm DRUL. We conduct chronological splitting of the datasets for training and evaluation. *Dunnhumby* has 93 discrete time periods. We train all models by periods from 1 to 77,

Table 4: Top-5 items frequently recommended in the Dunnhumby. The items recommended by PMF, the best unbiased baseline DLCE and our DRUL are presented in the columns, respectively. Numbers in parentheses are popularity ranks from logs.

PMF	DLCE	DRUL
FLUID MILK WHITE ONLY(1)	REFRIGERATED PASTA SAUCE(848)	INFANT FORMULA TODDLER(863)
SHREDDED CHEESE(5)	DRY &SPRAY STARCH(805)	REFRIGERATED PASTA SAUCE(848)
MAINSTREAM WHITE BREAD(3)	BEERALEMALT LIQUORS(11)	TORTILLA/NACHO CHIPS(15)
SOFT DRINKS PK CAN(4)	FLUID MILK WHITE ONLY(1)	DECOR BULBS(687)
TOILET TISSUE(10)	TEA UNSWEETENED(833)	JARRED FRUIT(889)

Table 5: The unbiased evaluation performance w.r.t I_c.

	MAE of cP@10					MAE of cDCG				
	0	0.1	0.3	1	10	0	0.1	0.3	1	10
Pop	.0264	.0252	**.0241**	.0269	.0278	.296	.288	**.273**	.291	.310
MF	.0247	.0234	**.0211**	.0238	.0255	.256	.243	**.227**	.254	.261
BPR	.0228	.0210	**.0196**	.0226	.0237	.270	.261	**.254**	.269	.277
CausE	.0204	.0196	**.0185**	.0201	.0215	.264	.254	**.242**	.261	.271
Caus-Prod	.0158	.0139	**.0121**	.0147	.0165	.227	.211	**.192**	.219	.238
ULBPR	.0095	.0078	**.0063**	.0082	.0103	.219	.208	**.196**	.213	.226
BLCE	.0148	.0137	**.0123**	.0143	.0156	.177	.153	**.141**	.162	.183
DLCE	.0168	.0156	**.0147**	.0173	.0182	.225	.214	**.202**	.217	.229
DRUL	.0189	.0175	**.0166**	.0182	.0193	.221	.208	**.197**	.219	.226

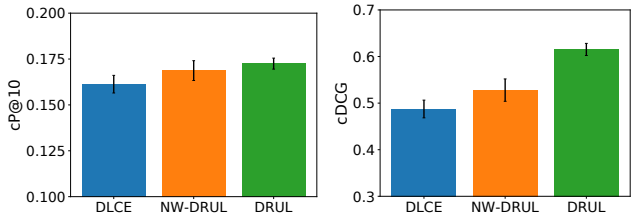


Figure 4: Performance comparisons between DLCE, (Non-Weighted) NW-DRUL (see Eq. (12)) and DRUL (see Eq. (21)).

validate them by 77 to 85 and test them by 85 to 93. The causal recommendation quality is measured with two metrics: cP@{10, 30, 100} and cDCG. Precision(P) was also measured, as a reference of a traditional metric to show that there is a large gap between casual metric (cP) and traditional metric (P). For all methods, we train our *DRIB* estimator based on the training dataset as we did in RQ1. Since the ground truth of causal effect τ_{ui} is unobservable in real world dataset, for fairness, we use the trained $\hat{\tau}_{ui}^{IPS}$ to measure the performances of every methods, in order to avoid the choice of estimator to be a confounding factor and make the experimental result more reliable.

Results. Table 6 shows the comparison results. From the table, we can observe: (1) Our propose DRUL achieves the best for all cases, which demonstrates the effectiveness of our method in learning from biased feedback. (2) The biased accuracy-based methods (PoP, MF and BPR) perform better in P@K; however, they perform worse in the causal metrics than other methods, which proves that there is gap between the tradition metric and the causal metric. (3) The methods (ULBPR, CausE, DLCE and our DRUL) that aim at optimizing the true causal effect outperform biased accuracy-based methods, which indicts that it's helpful to directly optimize the unbiased causal effect under the confounding bias. In Fig. 4, we take a closer look at the effect of weighted loss in Eq. (21). We compare DRUL using weighted loss in Eq. (21) and non-weighted loss in Eq. (12), and the baseline DLCE. We can find DRUL outperforms DLCE, and see a substantial improvement from the weighted loss.

Case Study. We take a deeper examination on our unbiased learning algorithm to understand how it affects the recommendation list

Table 6: Unbiased learning performance comparisons. Precision(P) is measured as a reference to show that there is gap between casual metric (cP) and traditional metric (P).

	cP			cDCG	P		
	@10	@30	@100	–	@10	@30	@100
Pop	.0231	.0198	.0181	.1021	.1301	.1205	.1104
MF	.0443	.0402	.0352	.1452	**.1719**	**.1598**	.1371
BPR	.0517	.0474	.0403	.1877	.1711	.1545	**.1426**
CausE	.0803	.0753	.0634	.2881	.0954	.0862	.0753
Caus-Prod	.1091	.0899	.0798	.3202	.0921	.0801	.0724
ULBPR	.1314	.1177	.0922	.4874	.0534	.0416	.0376
DLCE	.1613	.1502	.1301	.5279	.0502	.0404	.0341
DRUL	**.1725**	**.1631**	**.1472**	**.6152**	.0452	.0301	.0297

in production. As we mentioned in the introduction, confounding bias overestimates for popular items. One of the most important properties of correcting the confounding bias is that we can alleviate the popular bias and recommend items that user likes but will not buy if not recommended. To understand the differences between the biased and unbiased methods, we show the top five most frequently recommended items by the biased PMF, the best unbiased baseline DLCE and our DRUL on Dunnhumby dataset. From Table 4, we can find that the traditional method PMF tends to recommend more popular items than causal methods, i.e., DLCE and our DRUL. Compared to the baseline, DRUL is more effective to alleviate the popular bias and emphasis less on popular items.

8 CONCLUSIONS

In this paper, we study the problem of learning true causal effect from logged feedbacks under a confounding bias scenario, where recommendation and outcome are both affected by the confounding. To address this problem, we first propose a DR estimator for the causal effect of recommendation and showed its unbiasedness and desired statistical properties. We then propose a new method to estimate the propensity score and outcome based on deep information bottleneck. With the proposed DR estimator, we futher propose DRUL, an unbiased ranking algorithm to optimize the causal DCG via stochastic gradient descent. Experimental results show that our DRIB and DRUL significantly outperform existing methods in the unbiased evaluation and unbiased learning tasks, respectively. As to future work, we intend to extend our unbiased model to the dynamic or sequential settings, in which the confounding bias and user preferences are both dynamic over time [20, 50].

ACKNOWLEDGMENTS

This work is partially supported by the National Science Foundation (NSF) under grant number IIS-1909702 and IIS1955851, and Army Research Office (ARO) under grant number W911NF-21-1-0198.

REFERENCES

[1] David Barber Felix Agakov. 2004. The im algorithm: a variational approach to information maximization. *NIPS* 16 (2004), 201.

[2] Aman Agarwal, Kenta Takatsu, Ivan Zaitsev, and Thorsten Joachims. 2019. A general framework for counterfactual learning-to-rank. In *SIGIR*. 5–14.

[3] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2017. Deep variational information bottleneck. *International Conference on Learning Representations* (2017).

[4] Mohammad Taha Bahadori, Krzysztof Chalupka, Edward Choi, Robert Chen, Walter F Stewart, and Jimeng Sun. 2017. Causal regularization. *arXiv preprint arXiv:1702.02604* (2017).

[5] Heejung Bang and James M Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61, 4 (2005), 962–973.

[6] Stephen Bonner and Flavian Vasile. 2018. Causal embeddings for recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 104–112.

[7] David Brandfonbrener, William F Whitney, Rajesh Ranganath, and Joan Bruna. 2020. Overfitting and Optimization in Offline Policy Learning. *arXiv preprint arXiv:2006.15368* (2020).

[8] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. 2018. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In *SIGIR*. 35–44.

[9] Claes M Cassel, Carl E Särndal, and Jan H Wretman. 1976. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* 63, 3 (1976), 615–620.

[10] Miroslav Dudík, John Langford, and Lihong Li. 2011. Doubly robust policy evaluation and learning. *ICML* (2011).

[11] Max H Farrell, Tengyuan Liang, and Sanjog Misra. 2018. Deep neural networks for estimation and inference. *arXiv preprint arXiv:1809.09953* (2018).

[12] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. 2019. Learning Robust Representations via Multi-View Information Bottleneck. In *International Conference on Learning Representations*.

[13] Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. 2011. Doubly robust estimation of causal effects. *American journal of epidemiology* (2011), 761–767.

[14] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. 2018. Offline a/b testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 198–206.

[15] Anirudh Goyal, Riashat Islam, DJ Strouse, Zafarali Ahmed, Hugo Larochelle, Matthew Botvinick, Yoshua Bengio, and Sergey Levine. 2018. InfoBot: Transfer and Exploration via the Information Bottleneck. In *International Conference on Learning Representations*.

[16] Shantanu Gupta, Hao Wang, Zachary C Lipton, and Yuyang Wang. 2021. Correcting Exposure Bias for Link Recommendation. *arXiv preprint arXiv:2106.07041* (2021).

[17] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework. (2016).

[18] Wassily Hoeffding. 1994. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*. Springer, 409–426.

[19] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*. Ieee, 263–272.

[20] Jin Huang, Harrie Oosterhuis, and Maarten de Rijke. 2021. It Is Different When Items Are Older: Debiasing Recommendations When Selection Bias and User Preferences Are Dynamic. *arXiv preprint arXiv:2111.12481* (2021).

[21] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[22] Nan Jiang and Lihong Li. 2016. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*. PMLR, 652–661.

[23] Thorsten Joachims, Adith Swaminathan, and Maarten de Rijke. 2018. Deep learning with logged bandit feedback. In *International Conference on Learning Representations*.

[24] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 781–789.

[25] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*. 3020–3029.

[26] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[27] Xiang Lisa Li and Jason Eisner. 2019. Specializing Word Embeddings (for Parsing) by Information Bottleneck. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 2744–2754.

[28] Dawen Liang, Laurent Charlin, James McInerney, and David M Blei. 2016. Modeling user exposure in recommendation. In *Proceedings of the 25th international conference on World Wide Web*. 951–961.

[29] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*. 6446–6456.

[30] Ross Prentice. 1976. Use of the logistic model in retrospective studies. *Biometrics* (1976), 599–606.

[31] James M Robins and Andrea Rotnitzky. 1995. Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* 90, 429 (1995), 122–129.

[32] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[33] Donald B Rubin. 1979. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Amer. Statist. Assoc.* (1979), 318–328.

[34] Yuta Saito. 2020. Doubly Robust Estimator for Ranking Metrics with Post-Click Conversions. In *Fourteenth ACM Conference on Recommender Systems*. 92–100.

[35] Yuta Saito. 2020. Unbiased Pairwise Learning from Biased Implicit Feedback. In *ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14-17, 2020*. 5–12.

[36] Yuta Saito, Suguru Yaginuma, Yuta Nishino, Hayato Sakata, and Kazuhide Nakata. 2020. Unbiased Recommender Learning from Missing-Not-At-Random Implicit Feedback. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 501–509.

[37] Masahiro Sato, Janmajay Singh, Sho Takemori, Takashi Sonoda, Qian Zhang, and Tomoko Ohkuma. 2019. Uplift-based evaluation and optimization of recommenders. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 296–304.

[38] Masahiro Sato, Sho Takemori, Janmajay Singh, and Tomoko Ohkuma. 2020. Unbiased Learning for the Causal Effect of Recommendation. *Proceedings of the 14th ACM Conference on Recommender Systems* (2020).

[39] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. (2016), 1670–1679.

[40] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*. 3076–3085.

[41] Amit Sharma, Jake M Hofman, and Duncan J Watts. 2015. Estimating the causal impact of recommendation systems from observational data. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*. 453–470.

[42] Claudia Shi, David Blei, and Victor Veitch. 2019. Adapting neural networks for the estimation of treatment effects. In *NeurIPS*. 2507–2517.

[43] Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. 2020. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*. PMLR, 9167–9176.

[44] Yi Su, Lequn Wang, Michele Santacatterina, and Thorsten Joachims. 2019. Cab: Continuous adaptive blending for policy evaluation and learning. In *International Conference on Machine Learning*. PMLR, 6005–6014.

[45] Adith Swaminathan and Thorsten Joachims. 2015. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research* (2015), 1731–1755.

[46] Philip Thomas and Emma Brunskill. 2016. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*. PMLR, 2139–2148.

[47] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2019. Doubly robust joint learning for recommendation on data missing not at random. In *International Conference on Machine Learning*. 6638–6647.

[48] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2021. Combating Selection Biases in Recommender Systems with a Few Unbiased Ratings. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 427–435.

[49] Tailin Wu, Hongyu Ren, Pan Li, and Jure Leskovec. 2020. Graph Information Bottleneck. *Advances in Neural Information Processing Systems* 33 (2020), 20437–20448.

[50] Teng Xiao, Shangsong Liang, and Zaiqiao Meng. 2019. Hierarchical neural variational model for personalized sequential recommendation. In *The World Wide Web Conference*. 3377–3383.

[51] Teng Xiao and Donglin Wang. 2021. A general offline reinforcement learning framework for interactive recommendation. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*.

[52] Teng Xiao and Suhang Wang. 2022. Towards off-policy learning for ranking policies with logged feedback. In *The Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*.

[53] Longqi Yang, Yin Cui, Yuan Xuan, Chenyang Wang, Serge Belongie, and Deborah Estrin. 2018. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 279–287.

[54] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*.

A DERIVATIONS OF BIAS AND VARIANCE

For the completeness, we give the detailed derivations of the bias and variance of the DR estimator with estimated propensity score. All notations below are defined in the body of the paper.

A.1 Derivations of the Bias

We first rewrite the DR estimator in Eq. (9) as $\hat{\tau}_{ui}^{\text{DR}} = \hat{c}_{ui}^{(1)} - \hat{c}_{ui}^{(0)}$, where $(c_{ui}^{(1)} - \hat{Q}_{ui}^{(1)})\frac{R_{ui}}{\hat{e}_{ui}} + \hat{Q}_{ui}^{(1)} = \hat{c}_{ui}^{(1)}$ and $(c_{ui}^{(0)} - \hat{Q}_{ui}^{(0)})\frac{R_{ui}}{\hat{e}_{ui}} + \hat{Q}_{ui}^{(0)} = \hat{c}_{ui}^{(0)}$. Note that $\hat{\tau}_{ui}^{\text{DR}}$ is conditioned on confounding $\mathbf{x}_{ui}$. The expectation $\mathbb{E}[\hat{c}_{ui}^{(1)}|\mathbf{x}_{ui}]$ can be easily derived as follows:

$$\mathbb{E}\left[\hat{c}_{ui}^{(1)}|\mathbf{x}_{ui}\right] = (\hat{Q}_{ui}^{(1)} - Q_{ui}^{(1)})(1 - \frac{e_{ui}}{\hat{e}_{ui}}) + Q_{ui}^{(1)} = q_{ui}^{(1)}\delta_{ui}^{(1)} + Q_{ui}^{(1)}. \quad (26)$$

Similarly, the expectation $\mathbb{E}[\hat{c}_{ui}^{(0)}|\mathbf{x}_{ui}] = q_{ui}^{(0)}\delta_{ui}^{(0)} + Q_{ui}^{(0)}$ due to the symmetry. Thus, the bias of DR estimator can be derived as follows:

$$\text{Bias}(\hat{\tau}_{ui}^{\text{DR}}) = \left|\mathbb{E}\left[\hat{\tau}_{ui}^{\text{DR}} \mid \mathbf{x}_{ui}\right] - \mathbb{E}\left[c_{ui}^{(1)} - c_{ui}^{(0)} \mid \mathbf{x}_{ui}\right]\right| = \left|\delta_{ui}^{(1)}(q_{ui}^{(1)} + \frac{\hat{e}_{ui}}{1-\hat{e}_{ui}}q_{ui}^{(0)})\right|. \quad (27)$$

Recall that the IPS estimator is $\hat{\tau}_{ui}^{\text{IPW}} = \mathbb{E}[\frac{r_{ui}c_{ui}^{(1)}}{e(\mathbf{x}_{ui})} - \frac{(1-r_{ui})c_{ui}^{(0)}}{1-e(\mathbf{x}_{ui})}|\mathbf{x}_{ui}]$. Setting $\hat{Q}_{ui}^{(1)} = 0$ and $\hat{Q}_{ui}^{(0)} = 0$ in Eq. (9), the DR estimator reduces to IPS estimator. We can obtain the following bias of IPS by setting $\hat{Q}_{ui}^{(1)} = 0$ and $\hat{Q}_{ui}^{(0)} = 0$: $\text{Bias}(\hat{\tau}_{ui}^{\text{IPW}}) = |\delta_{ui}^{(1)}(Q_{ui}^{(1)} + \frac{\hat{e}_{ui}}{1-\hat{e}_{ui}}Q_{ui}^{(0)})|$.

A.2 Derivations of the Variance

We can first rewrite the variance as follows:

$$\text{Var}(\tau_{ui}^{\text{DR}}) = -2\mathbb{E}\left[\hat{c}_{ui}^{(1)}\hat{c}_{ui}^{(0)}|\mathbf{x}_{ui}\right] + 2\mathbb{E}\left[\hat{c}_{ui}^{(1)}|\mathbf{x}_{ui}\right]\mathbb{E}\left[\hat{c}_{ui}^{(0)}|\mathbf{x}_{ui}\right] + \mathbb{E}\left[(\hat{c}_{ui}^{(1)})^2|\mathbf{x}_{ui}\right] - (\mathbb{E}\left[\hat{c}_{ui}^{(1)}|\mathbf{x}_{ui}\right])^2 + \mathbb{E}\left[(\hat{c}_{ui}^{(0)})^2|\mathbf{x}_{ui}\right] - (\mathbb{E}\left[\hat{c}_{ui}^{(0)}|\mathbf{x}_{ui}\right])^2. \quad (28)$$

Here the term $\mathbb{E}[\hat{c}_{ui}^{(1)}\hat{c}_{ui}^{(0)}|\mathbf{x}_{ui}]$ is equal to:

$$(1-\delta_{ui}^{(1)})Q_{ui}^{(1)}\hat{Q}_{ui}^{(0)} - (1-\delta_{ui}^{(1)}-\delta_{ui}^{(0)})\hat{Q}_{ui}^{(1)}\hat{Q}_{ui}^{(0)} + (1-\delta_{ui}^{(0)})\hat{Q}_{ui}^{(1)}Q_{ui}^{(0)}. \quad (29)$$

The term $\mathbb{E}[\hat{c}_{ui}^{(1)}|\mathbf{x}_{ui}]\mathbb{E}[\hat{c}_{ui}^{(0)}|\mathbf{x}_{ui}]$ is equal to:

$$\begin{aligned}&(1-\delta_{ui}^{(1)})(1-\delta_{ui}^{(0)})Q_{ui}^{(1)}Q_{ui}^{(0)} + \delta_{ui}^{(1)}(1-\delta_{ui}^{(0)})\hat{Q}_{ui}^{(1)}Q_{ui}^{(0)} \\ &+ (1-\delta_{ui}^{(1)})\delta_{ui}^{(0)}Q_{ui}^{(1)}\hat{Q}_{ui}^{(0)} + \delta_{ui}^{(1)}\delta_{ui}^{(0)}\hat{Q}_{ui}^{(1)}\hat{Q}_{ui}^{(0)}. \end{aligned} \quad (30)$$

With some mathematical derivations, the term $\mathbb{E}[(\hat{c}_{ui}^{(1)})^2|\mathbf{x}_{ui}]$ is:

$$\mathbb{E}\left[(\epsilon_{ui}^{(1)})^2|\mathbf{x}_{ui}\right] + (Q_{ui}^{(1)} + q_{ui}^{(1)}\delta_{ui}^{(1)})^2 + \frac{1-e_{ui}}{e_{ui}}(q_{ui}^{(1)})^2(1-\delta_{ui}^{(1)})^2. \quad (31)$$

Similarly, $\mathbb{E}[(\hat{c}_{ui}^{(0)})^2|\mathbf{x}_{ui}] = \mathbb{E}[(\epsilon_{ui}^{(0)})^2|\mathbf{x}_{ui}] + (Q_{ui}^{(0)} + q_{ui}^{(0)}\delta_{ui}^{(0)})^2 + \frac{e_{ui}}{1-e_{ui}}(q_{ui}^{(0)})^2(1-\delta_{ui}^{(0)})^2$. Combining Eqs. (26), (29), (30), (31) and (28), we can obtain the final form of the variance of DR:

$$\begin{aligned}\text{Var}(\tau_{ui}^{\text{DR}}) &= \mathbb{E}\left[(\epsilon_{ui}^{(1)})^2|\mathbf{x}_{ui}\right] + \mathbb{E}\left[(\epsilon_{ui}^{(0)})^2|\mathbf{x}_{ui}\right] + \frac{1-e_{ui}}{e_{ui}}(q_{ui}^{(1)})^2(1-\delta_{ui}^{(1)})^2 \\ &+ \frac{e_{ui}}{1-e_{ui}}(q_{ui}^{(0)})^2(1-\delta_{ui}^{(0)})^2 + \frac{2e_{ui}(1-e_{ui})}{\hat{e}_{ui}(1-\hat{e}_{ui})}q_{ui}^{(1)}q_{ui}^{(0)}. \end{aligned} \quad (32)$$

We can obtain the variance of IPS by setting $\hat{Q}_{ui}^{(1)}$ and $\hat{Q}_{ui}^{(0)}$ as zero: $\text{Var}(\hat{\tau}_{ui}^{\text{IPS}}) = \mathbb{E}[(\epsilon_{ui}^{(1)})^2|\mathbf{x}_{ui}] + \mathbb{E}[(\epsilon_{ui}^{(0)})^2|\mathbf{x}_{ui}] + \frac{2e_{ui}(1-e_{ui})}{\hat{e}_{ui}(1-\hat{e}_{ui})}Q_{ui}^{(1)}Q_{ui}^{(0)} + \frac{e_{ui}}{1-e_{ui}}(Q_{ui}^{(0)})^2(1-\delta_{ui}^{(0)})^2 + \frac{1-e_{ui}}{e_{ui}}(Q_{ui}^{(1)})^2(1-\delta_{ui}^{(1)})^2$.

B PROOFS OF LEMMA AND PROPOSITION

B.1 Proof of Proposition 1

Proof. Recall the DR estimator with the true propensity score is:

$$\hat{\tau}_{ui}^{\text{DR}} = (c_{ui}^{(1)} - \hat{Q}_{ui}^{(1)})\frac{r_{ui}}{e_{ui}} - (c_{ui}^{(0)} - \hat{Q}_{ui}^{(0)})\frac{1-r_{ui}}{1-e_{ui}} + (\hat{Q}_{ui}^{(1)} - \hat{Q}_{ui}^{(0)}), \quad (33)$$

where the binary treatment r_{ui} (whether recommending or not) follows a unknown Bernoulli distribution with probability e_{ui}. We define the random variable $\gamma_{ui} = \lambda(\hat{z}_{ui})\hat{\tau}_{ui}^{DR}$ as follows:

$$p(\gamma_{ui} = \alpha_{ui}) = e_{ui}.\ p(\gamma_{ui} = \beta_{ui}) = 1 - e_{ui}, \quad (34)$$

where the probabilities $\alpha_{ui} = \lambda(\hat{z}_{ui})(\frac{c_{ui}^{(1)} - \hat{Q}_{ui}^{(1)}}{e_{ui}} + \hat{Q}_{ui}^{(1)} - \hat{Q}_{ui}^{(0)})$ and $\beta_{ui} = \lambda(\hat{z}_{ui})(-\frac{c_{ui}^{(0)} - \hat{Q}_{ui}^{(0)}}{1-e_{ui}} + \hat{Q}_{ui}^{(1)} - \hat{Q}_{ui}^{(0)})$. We can observe that the square of interval size of random variables γ_{ui} is:

$$(\alpha_{ui} - \beta_{ui})^2 = \lambda(\hat{z}_{ui})^2(\frac{c_{ui}^{(1)} - \hat{Q}_{ui}^{(1)}}{e_{ui}} + \frac{c_{ui}^{(0)} - \hat{Q}_{ui}^{(0)}}{1-e_{ui}})^2. \quad (35)$$

Recall that we assume that the treatment assignments $\{r_{ui}|(u,i) \in \mathcal{U} \times \mathcal{I}\}$ are independent random variables, thus the random variables $\{\gamma_{ui}|(u,i) \in \mathcal{U} \times \mathcal{I}\}$ are also independent. Based on Hoeffding's inequality (see Theorem 2 in the [18] for proof), we have:

$$\begin{aligned}&P(|\sum_u\sum_i \lambda(\hat{z}_{ui})\tau_{ui}^{DR} - \mathbb{E}[\sum_u\sum_i \lambda(\hat{z}_{ui})\tau_{ui}^{DR}]| \geq \epsilon) \\ &\leq 2\exp(\frac{-2\epsilon^2}{\sum_{u,i}(\alpha_{ui}-\beta_{ui})^2}) = 2\exp(\frac{-2\epsilon^2}{\sum_{u,i}\lambda(\hat{z}_{ui})^2 d_{u,i}^2}) \Leftrightarrow \\ &P(|\frac{1}{|\mathcal{U}|}\sum_u\sum_i(\lambda(\hat{z}_{ui})\tau_{ui}^{DR} - \lambda(\hat{z}_{ui})\tau_{ui})| \geq \frac{\epsilon}{|\mathcal{U}|}) \leq 2\exp(\frac{-2\epsilon^2}{\sum_{u,i}\lambda(\hat{z}_{ui})^2 d_{u,i}^2}) \\ &\Leftrightarrow P(|\hat{R}^{\text{DR}}(\hat{Z}) - R^{\text{Ideal}}(\hat{Z})| \geq \epsilon) \leq 2\exp(\frac{-2|\mathcal{U}|^2\epsilon^2}{\sum_{u,i}\lambda(\hat{z}_{ui})^2 d_{u,i}^2}), \end{aligned} \quad (36)$$

where $d_{ui} = (\frac{c_{ui}^{(1)} - \hat{Q}_{ui}^{(1)}}{e_{ui}} + \frac{c_{ui}^{(0)} - \hat{Q}_{ui}^{(0)}}{1-e_{ui}})$. Setting the right hand side above to η and solving ϵ yields:

$$P(\left|\hat{R}^{\text{DR}}(\hat{Z}) - R^{\text{Ideal}}(\hat{Z})\right| \leq \frac{1}{|\mathcal{U}|}\sqrt{\frac{\log\frac{2}{\eta}}{2}}\sqrt{\sum_{u,i}\lambda(\hat{z}_{ui})^2 d_{ui}^2}) \geq 1 - \eta. \quad (37)$$

This completes the proof. □

B.2 Proof of Corollary 1.1

Proof. Given conditions that $0 < \hat{Q}_{ui}^{(1)} \leq 2c_{ui}^{(1)}$ and $0 < \hat{Q}_{ui}^{(0)} \leq 2c_{ui}^{(0)}$, we can derive the following inequalities (Note that $\hat{e}_{ui} \in [0,1]$):

$$d_{ui} - \hat{d}_{ui} = -\frac{Q_{ui}^{(1)}}{\hat{e}_{ui}} - \frac{Q_{ui}^{(0)}}{(1-\hat{e}_{ui})} \leq 0 \Rightarrow d_{ui} \leq \hat{d}_{ui} \quad (38)$$

$$d_{ui} + \hat{d}_{ui} = \frac{2c_{ui}^{(1)} - Q_{ui}^{(0)}}{\hat{e}_{ui}} + \frac{2c_{ui}^{(0)} - Q_{ui}^{(0)}}{(1-\hat{e}_{ui})} \geq 0 \Rightarrow d_{ui} \geq -\hat{d}_{ui}, \quad (39)$$

where $\hat{d}_{ui}^2 = (\frac{c_{ui}^{(1)}}{e_{ui}} + \frac{c_{ui}^{(0)}}{1-e_{ui}})^2$. Thus, we have $d_{ui}^2 \leq \hat{d}_{ui}^2$, resulting in:

$$\frac{1}{|\mathcal{U}|}\sqrt{\frac{\log\frac{2}{\eta}}{2}}\sqrt{\sum_{u,i}d_{ui}^2} \leq \frac{1}{|\mathcal{U}|}\sqrt{\frac{\log\frac{2}{\eta}}{2}}\sqrt{\sum_{u,i}\hat{d}_{ui}^2}. \quad (40)$$

The left hand side is the tail bound of the DR and the right hand side is the tail bound of the IPS [38]. This completes the proof. □