Separating Examination and Trust Bias from Click Predictions for Unbiased Relevance Ranking

Haiyuan Zhao
School of Information
Renmin University of China
haiyuanzhao@ruc.edu.cn

Jun Xu*
Gaoling School of Artificial Intelligence
Renmin University of China
junxu@ruc.edu.cn

Xiao Zhang
Gaoling School of Artificial Intelligence
Renmin University of China
zhangx89@ruc.edu.cn

Guohao Cai
Noah's Ark Lab, Huawei
caiguohao1@huawei.com

Zhenhua Dong
Noah's Ark Lab, Huawei
dongzhenhua@huawei.com

Ji-Rong Wen
Gaoling School of Artificial Intelligence
Renmin University of China
jrwen@ruc.edu.cn

ABSTRACT

Alleviating the examination and trust bias in ranking systems is an important research line in unbiased learning-to-rank (ULTR). Current methods typically use the propensity to correct the biased user clicks and then learn ranking models based on the corrected clicks. Though successes have been achieved, directly modifying the clicks suffers from the inherent high variance because the propensities are usually involved in the denominators of corrected clicks. The problem gets even worse in the situation of mixed examination and trust bias. To address the issue, this paper proposes a novel ULTR method called *Decomposed Ranking Debiasing* (DRD). DRD is tailored for learning unbiased relevance models with low variance in the existence of examination and trust bias. Unlike existing methods that directly modify the original user clicks, DRD proposes to *decompose each click prediction as the combination of a relevance term outputted by the ranking model and other bias terms*. The unbiased relevance model, therefore, can be learned by fitting the overall click predictions to the biased user clicks. A joint learning algorithm is developed to learn the relevance and bias models' parameters alternatively. Theoretical analysis showed that, compared with existing methods, DRD has lower variance while retains unbiasedness. Empirical studies indicated that DRD can effectively reduce the variance and outperform the state-of-the-art ULTR baselines.

CCS CONCEPTS

• **Information systems → Learning to rank**.

KEYWORDS

Unbiased relevance ranking, decomposing ranking de-biasing

*Corresponding author

WSDM '23, February 27-March 3, 2023, Singapore, Singapore

ACM ISBN 978-1-4503-9407-9/23/02…$15.00
https://doi.org/10.1145/3539597.3570393

ACM Reference Format:
Haiyuan Zhao, Jun Xu, Xiao Zhang, Guohao Cai, Zhenhua Dong, and Ji-Rong Wen. 2023. Separating Examination and Trust Bias from Click Predictions for Unbiased Relevance Ranking. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23), February 27-March 3, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3539597.3570393

1 INTRODUCTION

User click log has been used as a source of supervision to learn the ranking models in modern search engines. However, the clicks are affected by user behaviors and thus suffer from various data biases. Among them, the examination bias [15, 16] and trust bias [2, 31] are two typical biases. Examination bias is caused by users' different examination probabilities on different ranking positions [17]. That is, a user may be less likely to examine the lower ranked results and then click it, resulting in the click signals no longer being indicators of true relevance; Trust bias is caused by users' trust in the effectiveness of the search engine to rank relevant documents higher [2]. With trust bias, a non-relevant result may be clicked by users as long as it is ranked at a higher position. It is worth noting that these two biases do not occur separately. They are mixed and collectively influence the user clicks.

In recent years, unbiased learning-to-rank (ULTR) models [4, 5] have been developed to address the examination bias and trust bias, including TrustPBM [2], Affine Correction (AC) [31] and Mixture-Based Correction (MBC) [30] etc. These methods focus on estimating the unbiased relevance labels, i.e., directly correcting the biased user clicks using the click propensities. Then, the unbiased relevance ranking models are learned by treating the corrected user clicks as the ground-truth relevance labels.

However, reducing bias inevitably comes at the cost of increasing variance for those methods that directly correcting the biased user clicks [18, 26–29, 32], especially when the examination bias and trust bias are mixed.[1] Hence, controlling the variance will reduce the generalization error of unbiased learning and improve the performance of the learned model. Furthermore, recent studies [25, 28] proved that reducing variance can also lead to a tighter error tail bound, thus improving the stability of the learned unbiased model.

[1] MBC [30] is an exception as it does not rely on propensity scores.

In this paper, we proposed a ULTR method called *Decomposed Ranking Debiasing* (DRD). Unlike existing methods that directly correct the biased user clicks, DRD proposes to learn the unbiased relevance ranking model by decomposing each click probability prediction. That is, decompose each click probability prediction as a combination of a relevance term (outputted by the unbiased relevance model) and other bias terms (outputted by the bias models). Then, DRD can learn the unbiased relevance model by fitting this decomposed prediction to the biased user clicks. In this way, DRD avoids involving the propensities in the denominators and can learn the models with lower variance. Theoretical analysis also shows that, compared to existing approaches, the proposed DRD has lower variance while still retaining unbiasedness, ensuring its superiority in the existence of both examination bias and trust bias.

To learn the model parameters, a joint learning algorithm is designed in which the bias estimation and relevance model learning are conducted alternatively and can be regularized by each other. Compared to current debiasing methods that separate bias estimation and model learning as two stages, our joint learning algorithm can avoid the problem of bias estimation error accumulating and amplifying into relevance model learning.

In summary, DRD provides an elegant and theoretical sound approach to alleviating the examination and trust bias in relevance ranking. It offers several advantages, including unbiased and low variance learning, joint bias estimation and relevance model learning, and high accuracy in relevance ranking. The major contributions of this study are:

(1) We proposed an ULTR model for addressing the high-variance problem in the existence of examination bias and trust bias.
(2) We designed an effective learning algorithm that benefits from the joint bias estimation and relevance model learning.
(3) We conducted groups of experiments on two public LTR datasets, and the experimental results verified the effectiveness of the proposed model and the theoretical conclusions.

2 RELATED WORK

Trust bias and examination bias are two typical data biases in learning to rank. To address the examination bias, [17] proposed propensity SVM-rank, which using IPS to address examination bias firstly. Besides this, there are more extensive methods. Agarwal et al. [1] extended the propensity SVM-rank and made it applicable to neural network ranker, and can optimize DCG directly. Fang et al. [9] and Wu et al. [36] extended traditional IPS method to the case where the examination is context dependent. Guo et al. [11] extend examination bias into grid-based web search. Chen et al. [7] further assumed that the examination bias in different ranking position is not isolated but will affect each other. For trust bias, it was first defined by Agarwal et al. [2], they considered trust bias as a click noise and proposed TrustPBM based on Bayesian rules. Vardasbi et al. [31] analyzed the error of TrustPBM, and further proposed Affine Correction to allievate the trust bias. However, above two methods for addressing trust bias rely on relEM [33] to estimate bias parameters. To overcome the estimated error brought by relEM, Vardasbi et al. [30] proposed Mixture-Based Correction, which employs a standard EM procedure to estimate relevance directly.

Current variance reduction techniques for ULTR usually directly control the variability of the propensity weights. For example, Swaminathan and Joachims [29] analyzed the problem of variability brought by propensity weighting and proposed to handle it via a self-normalized estimator. Schnabel et al. [28] employed the self-normalizing technique to achieve better performance in addressing selection bias. Propensity clipping is another method to reduce the variance, which limits the range of propensity weight by manually setting thresholds. This technique was widely used in many debiasing methods [12, 20, 27, 32]. Doubly robust estimator [8, 37] can also reduce the variance by integrating IPS with a direct method. However, current studies don't specifically address the variance in the existence of examination and trust bias.

Recent studies on recommender systems also mentioned a similar decomposition idea for alleviating bias. For example, Zheng et al. [39] proposed to decompose user interactions as interest and conformity and then model them respectively. Moreover, Zhang et al. [38] proposed to decompose click as the effect from user-item and item-popularity. Similarly, Wei et al. [34] proposed to decompose click as user effect, item effect, and matching effect, then model them as multi-task learning. However, the variance of the decompose-based methods has been less discussed.

3 PROBLEM FORMULATION AND ANALYSIS

3.1 Unbiased Learning to Rank

The problem of unbiased learning-to-rank can be described as follows. Given a user query q and K retrieved documents, each query-document pair (q, d) is described by an n-dimensional feature vector $\mathbf{x} = \phi(q, d) \in \mathbb{R}^n$. The relevance of d to q can be represented by an unobserved variable R. Without loss of generality, we assume that $R \in \{0, 1\}$ is a binary variable. The retrieved documents are ranked by an existing ranking model $\pi_0 : \mathbb{R}^n \to \{1, 2, \cdots, K\}$ where each document will be ranked at a position $P \in \{1, 2, \cdots, K\}$ by π_0. Following the practices in [1, 17, 31], suppose that random variable $E \in \{0, 1\}$ denotes whether a user has examined the presented document, and $C \in \{0, 1\}$ denotes whether a user clicks the document. Both E and C obey the Bernoulli distributions. The user clicks on a search engine can be recorded as the click log $\mathcal{D} = \{(\mathbf{x}_i, c_i, k_i)\}_{i=1}^N$, where $\mathbf{x}_i, c_i, k_i$ respectively denote the i-th query-document pair's feature vector, whether the document being clicked, and the ranking position of the document by π_0.

Ideally, we hope an unbiased relevance model $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ could be learned by maximizing the following ideal point-wise loss:

$$\mathcal{L}_{\text{ideal}} = \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}} -r \log\left[\sigma\left(f(\mathbf{x})\right)\right] - (1-r) \log\left[1 - \sigma\left(f(\mathbf{x})\right)\right], \quad (1)$$

where r is the unobserved true relevance of a query-document pair, σ is the sigmoid function. Equation (1) cannot be maximized because r cannot be observed directly. An alternative way is naively fitting the prediction to the observed clicks c in $\mathcal{D}$:

$$\mathcal{L}_{\text{naive}} = \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}} -c \log\left[\sigma\left(f(\mathbf{x})\right)\right] - (1-c) \log\left[1 - \sigma\left(f(\mathbf{x})\right)\right]. \quad (2)$$

As has been discussed, there exists a gap between the optimal solution of $\mathcal{L}_{\text{naive}}$ and that of $\mathcal{L}_{\text{ideal}}$, because the user clicks are biased. That is, the click c is no longer an indicator of unobserved true relevance r if bias exists, e.g., the examination bias and trust

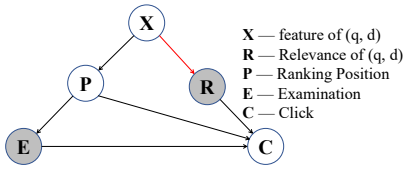


Figure 1: Causal graph of users' clicks in search ranking. Gray nodes denote unobserved variables. The red arrow indicates the effect that an unbiased relevance model needs to estimate.

bias. The goal of unbiased learning-to-rank is to alleviate the biases in user clicks and obtain an unbiased relevance model $f(\mathbf{x})$.

3.2 Examination Bias and Trust Bias

Next, we analyze how the examination and trust bias affect the user clicks based on the causal graph [10, 22, 35] shown in Figure 1. Given a query-document pair (q, d), its feature $\mathbf{x}$ affects the position P that d is placed through π_0. The position P affects E, which means whether the user can examine the document d. Also, $\mathbf{x}$ determines the relevance R between (q, d) trough the unbiased relevance model $f(\mathbf{x})$. Since relevance label R cannot be directly observed in $\mathcal{D}$, we have to leverage click signal C as a substitution of R. Unfortunately, besides the relevance R, C is also affected by the ranking position P and user examination E. Therefore, directly fitting clicks will result in a biased relevance model. According to Figure 1, the probability that a user clicks a document can be written as:

$$\begin{aligned}\Pr(C=1|\mathbf{x}) &= \sum_{E\in\{0,1\}}\sum_{R}\Pr(C=1|E,R,P)\Pr(R|\mathbf{x})\sum_{P}\Pr(E|P)\Pr(P|\mathbf{x})\\ &= \sum_{E\in\{0,1\}}\sum_{R}\Pr(C=1|E,R,P=k)\Pr(R|\mathbf{x})\Pr(E|P=k)\\ &= \underbrace{\Pr(E=1|P=k)}_{\text{exam. bias}}\sum_{R}\underbrace{\Pr(C=1|E=1,R,P=k)}_{\text{trust bias}}\underbrace{\Pr(R|\mathbf{x})}_{\text{unbiased rel.}},\end{aligned} \tag{3}$$

where $k = \pi_0(\mathbf{x})$ is the ranking position. The first equation is decomposition based on the Figure 1; the second equation is based on the fact that the ranking policy π_0 is deterministic, which means one (q, d) pair has only one ranking position in the ranking result; the last equation is based on the examination hypothesis [24]. That is, only the examined documents can be clicked by users, while users can never click those un-examined documents. Equation (3) indicates that the click probability can be decomposed as a combination of examination bias term $\Pr(E = 1|P = k)$, trust bias term $\Pr(C = 1|E = 1, R, P = k)$, and unbiased relevance term $\Pr(R|\mathbf{x})$.

To transform the biased clicks to the unbiased relevance labels, one popular approach is using the affine transformation. For example, the propensity weighting method of Affine Correction (AC) [31] defines its relevance label $\hat{r}_{\text{AC}}$ as:

$$\hat{r}_{\text{AC}} = (c - \theta p^-) \,/\theta(p^+ - p^-),$$

where the $c \in \{0, 1\}$ is the observed user click, $\theta = \Pr(E = 1|P = k)$ is the examination bias, $p^+ = \Pr(C = 1|E = 1, R = 1, P = k)$ and $p^- = \Pr(C = 1|E = 1, R = 0, P = k)$ are the trust bias on examined relevant and irrelevant documents ranked at k, respectively.

3.3 Variance of Propensity Weighting Methods

Although existing propensity weighting methods (e.g., AC) achieved good performance in mitigating both examination bias and trust bias, they face high variance challenges due to re-weighting propensities in the denominators of the estimated label. Moreover, the variance gets even larger in the existence of both examination and trust bias, as shown in the following Lemma 1 and Theorem 1.

Lemma 1. *Let* $\mathcal{L}_{\text{AC}} = \frac{1}{|\mathcal{D}|}\sum_{\mathcal{D}}\left(-\hat{r}_{\text{AC}}\log(\hat{p}_{\mathbf{x}}) - (1-\hat{r}_{\text{AC}})\log(1-\hat{p}_{\mathbf{x}})\right)$. *be a binary cross entropy loss function that uses the AC estimator. If* $\Delta_p = p^+ - p^- > 0$, *the variance of* $\mathcal{L}_{\text{AC}}$ *is*

$$\begin{aligned}\mathbb{V}_{\text{AC}} &= \frac{1}{|\mathcal{D}|^2}\sum_{\mathcal{D}}\mathbb{V}\left[-\hat{r}_{\text{AC}}\log(\hat{p}_{\mathbf{x}}) - (1-\hat{r}_{\text{AC}})\log(1-\hat{p}_{\mathbf{x}})\right]\\ &= \frac{1}{|\mathcal{D}|^2}\sum_{\mathcal{D}}\mathbb{V}\left[-\log(\frac{\hat{p}_{\mathbf{x}}}{1-\hat{p}_{\mathbf{x}}})\hat{r}_{\text{AC}} - \log(1-\hat{p}_{\mathbf{x}})\right]\\ &= \frac{1}{|\mathcal{D}|^2}\sum_{\mathcal{D}}\log^2(\frac{\hat{p}_{\mathbf{x}}}{1-\hat{p}_{\mathbf{x}}})\mathbb{V}\left[\frac{c-\theta p^-}{\theta\Delta_p}\right],\end{aligned}$$

where $\hat{p}_{\mathbf{x}} = \sigma(f(\mathbf{x}))$ *is the predicted relevance probability and* $p_r = \Pr(R = 1|\mathbf{x})$ *is the true relevance probability.*

Note that the condition $\Delta_p > 0$ is natural and generally holds because users trust the relevant documents more than the irrelevant ones. It is easy to know that if the user clicks are only affected by the examination bias, $\mathcal{L}_{AC}$ naturally degenerates to:

$$\mathcal{L}_{\text{exam}} = \frac{1}{|\mathcal{D}|}\sum_{\mathcal{D}} -\frac{c}{\theta}\log(\hat{p}_{\mathbf{x}}) - \left(1-\frac{c}{\theta}\right)\log(1-\hat{p}_{\mathbf{x}}).$$

We can prove that $\mathcal{L}_{\text{AC}}$ has larger variance than $\mathcal{L}_{\text{exam}}$.

Theorem 1 (Variance Comparison). $\mathbb{V}_{\text{AC}} \geq \mathbb{V}_{\text{exam}}$. *where* $\mathbb{V}_{\text{exam}}$ *is the variance of* $\mathcal{L}_{\text{exam}}$.

Proof of Theorem 1 can be found in Appendix A.1. Theorem 1 indicates that mixing the trust bias with the user clicks already affected by the examination bias will further increase the variance.

4 OUR APPROACH: DRD

This section proposes the Decomposed Ranking Debiasing (DRD) which simultaneously alleviates the examination and trust bias.

4.1 Decomposed Ranking Debiasing Method

Inspired by the decomposition in Equation (3), we propose to decompose the predicted click probability as a combination of a relevance term and other bias terms. Based on the decomposition, D4D can learn the unbiased relevance model by fitting the click predictions to the biased user clicks.

4.1.1 Decomposed Click Probability and Loss Function. Specifically, given a $\mathbf{x} = \phi(q, d)$, the overall predicted click probability, denoted as $\hat{p}_c$, can be decomposed as

$$\hat{p}_c = \theta p^+\hat{p}_{\text{x}} + \theta p^-(1-\hat{p}_{\text{x}}) = \theta\Delta_p\hat{p}_{\text{x}} + \theta p^-, \tag{4}$$

where $\hat{p}_{\mathbf{x}} = \sigma(f(\mathbf{x}))$ is the predicted relevance probability, the first term $\theta p^+\hat{p}_{\text{x}}$ is the probability that d is relevant to q and it is examined and clicked, the second term $\theta p^-(1-\hat{p}_{\text{x}})$ is the probability that d is irrelevant and it is examined and clicked.

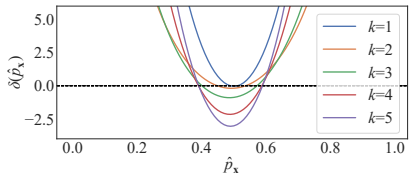


Figure 2: Curves of $\delta(\hat{p}_{\mathbf{x}})$. k denotes the ranking position.

Based on a set of click log $\mathcal{D}$, a new cross-entropy loss function $\mathcal{L}_{\text{DRD}}$ can be constructed:

$$\mathcal{L}_{\text{DRD}} = \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}} -c \log \hat{p}_c - (1-c) \log (1-\hat{p}_c) . \tag{5}$$

We can prove that, through optimizing $\mathcal{L}_{\text{DRD}}$, the learned relevance model $\hat{p}_{\mathbf{x}} = \sigma(f(\mathbf{x}))$ converges to predict the unbiased true relevance probability p_r:

Theorem 2. *Let $\hat{p}^*_{\mathbf{x}} = \arg\min_{p_{\mathbf{x}}} \mathcal{L}_{\text{DRD}}$, we have*

$$\forall \mathbf{x},\ \hat{p}^*_{\mathbf{x}} = p_r,$$

where $p_r = \Pr(R=1|\mathbf{x})$ is the true relevance probability of $\mathbf{x}$.

Proof of Theorem 2 can be found in Appendix A.2.

4.1.2 Variance Analysis of $\mathcal{L}_{DRD}$. We analyze the variance of $\mathcal{L}_{\text{DRD}}$ from Equation (5), denoted as $\mathbb{V}_{\text{DRD}}$, in the following Theorem 3.

Theorem 3 (Variance of $\mathcal{L}_{\text{DRD}}$). *The variance of $\mathcal{L}_{\text{DRD}}$ is*

$$\mathbb{V}_{\text{DRD}} = \frac{1}{|\mathcal{D}|^2} \sum_{\mathcal{D}} \log^2 \left(\frac{\theta \Delta_p \hat{p}_{\mathbf{x}} + \theta p^-}{1 - \theta \Delta_p \hat{p}_{\mathbf{x}} - \theta p^-} \right) \mathbb{V}_c .$$

Proof of this theorem can be found in Appendix A.3.

One advantage of the new loss $\mathcal{L}_{\text{DRD}}$ is its low variance in learning. Specifically, we compare $\mathcal{L}_{\text{DRD}}$ with AC's variance $\mathbb{V}_{\text{AC}}$ (defined in Lemma 1) in the following Remark 1:

Remark 1 (Lower Variance of $\mathcal{L}_{\text{DRD}}$). *Based on Lemma 1 and Theorem 3, the variances of $\mathcal{L}_{DRD}$ and $\mathcal{L}_{AC}$ can be compared:*

$$\mathbb{V}_{\text{AC}} - \mathbb{V}_{\text{DRD}} = \frac{1}{|\mathcal{D}|^2} \sum_{\mathcal{D}} \left(\frac{\log^2 (\frac{\hat{p}_{\mathbf{x}}}{1-\hat{p}_{\mathbf{x}}})}{\theta^2 \Delta_p^2} - \log^2 \left(\frac{\theta \Delta_p \hat{p}_{\mathbf{x}} + \theta p^-}{1 - \theta \Delta_p \hat{p}_{\mathbf{x}} - \theta p^-} \right) \right) \mathbb{V}_c .$$

Let $\delta(\hat{p}_{\text{x}}) = \log^2 \left(\frac{\hat{p}_{\text{x}}}{1-\hat{p}_{\text{x}}} \right) - \theta^2 \Delta_p^2 \log^2 \left(\frac{\theta \Delta_p \hat{p}_{\text{x}} + \theta p^-}{1 - \theta \Delta_p \hat{p}_{\text{x}} - \theta p^-} \right)$. It is obvious that $\mathbb{V}_{\text{AC}} \geq \mathbb{V}_{\text{DRD}}$ if $\delta(\hat{p}_{\mathbf{x}}) \geq 0$ holds for most $\hat{p}_{\mathbf{x}}$ ranges.

We empirically show that the condition $\delta(\hat{p}_{\mathbf{x}}) \geq 0$ holds in real-world unbiased learning-to-rank applications with examination bias and trust bias. Specifically, based on the bias terms calculated according to the equations in Section 5, we illustrate the curves of $\delta(\hat{p}_{\mathbf{x}})$ in Figure 2. The 5 curves respectively correspond to the ranking positions $k = \{1, 2, 3, 4, 5\}$. From the curves, we found that:

(1) **DRD reduced the variance in most cases:** We observed that $\delta(\hat{p}_{\mathbf{x}}) > 0$ for most of the $\hat{p}_{\mathbf{x}}$'s values, indicating $\mathbb{V}_{\text{AC}} \geq \mathbb{V}_{\text{DRD}}$ holds in most cases. We empirically observed that the condition is violated only near the point $\hat{p}_{\mathbf{x}} = 0.5$, e.g., around $\hat{p}_{\mathbf{x}} \in (0.4, 0.6)$.

(2) **DRD can dramatically reduces the variance for confident query-document pairs:** We also observed that $\delta(\hat{p}_{\mathbf{x}}) \gg 0$ when $\hat{p}_{\mathbf{x}}$ is near to 0 or 1. Note that $\hat{p}_{\mathbf{x}}$ is near 0 or 1 means that the relevance model is confident about the relevance of $\mathbf{x}$ (either relevant or irrelevant). The phenomenon indicates that $\mathbb{V}_{\text{AC}} > \mathbb{V}_{\text{DRD}}$ with a large margin if the relevance prediction is confident. It also means that DRD's learning variance can be steadily reduced as the training goes on because the relevance model usually becomes more confident with more training iterations.

(3) **DRD can reduce the estimation variance on high-ranked documents:** Comparing the 5 curves with k values, we found that the curves correspond to the high-ranked documents (e.g., $k = 1$) has much smaller $\hat{p}_{\mathbf{x}}$'s value ranges in which $\mathbb{V}_{\text{DRD}} > \mathbb{V}_{\text{AC}}$. Note that the higher ranked documents have more impact on the ranking accuracy. The phenomenon indicates that DRD has more advantages in alleviating the biases for high-ranked documents.

Algorithm 1: Joint Learning Algorithm for DRD

Input: User interactions $\mathcal{D} = \{(\mathbf{x}_i, c_i, k_i)\}_{i=1}^N$, learning rates η_1, η_2, number of iters T, number of batches B_1, B_2

```
Θ^r, Θ^+, Θ^- ← random values
for 1 ≤ t ≤ T do
    for 1 ≤ b ≤ B_1 do
        Randomly sample a batch of sessions D^b from D;
        Calculate L^rel_DRD on a batch D^b ⊆ D {Eq. (6)};
        Θ^r ← Θ^r − η_1 ∂L^rel_DRD/∂Θ^r;
    end
    for 1 ≤ b ≤ B_2 do
        Randomly sample a batch of sessions D^b from D;
        Calculate L^bias_DRD on a batch D^b ⊆ D {Eq. (7)};
        Θ^+ ← Θ^+ − η_2 ∂L^bias_DRD/∂Θ^+; Θ^- ← Θ^- − η_2 ∂L^bias_DRD/∂Θ^-;
    end
end
return Θ^r
```

4.2 Optimizing with Joint Learning

In practice, the true examination and trust bias parameters are unknown. Hence, we propose to parameterize the bias terms in Eq. (4) and employ a joint learning algorithm to learn the parameters in the relevance and bias models alternatively.

4.2.1 Estimating bias parameters with bias models. First, we estimate the bias terms θp^+ and θp^- in the predicted click probability $\hat{p}_c$ in Eq. (4), denote as $\hat{\theta}\hat{p}^+$ and $\hat{\theta}\hat{p}^-$:

$$\hat{\theta}\hat{p}^+ := \sigma(g^+(\mathbf{x}')) \quad \text{and} \quad \hat{\theta}\hat{p}^- := \sigma(g^-(\mathbf{x}')),$$

where $\mathbf{x}'$ is a one-hot representation of the position outputted by π_0, $g^-(\cdot)$ and $g^+(\cdot)$ are two bias models with scalar outputs for estimating the bias parameters on relevant and irrelevant documents, respectively[2]. Therefore, the click probability $\hat{p}_c$ in Eq. (4) becomes:

$$\hat{p}_c = \left(\sigma(g^+(\mathbf{x}')) - \sigma(g^-(\mathbf{x}'))\right) \sigma\left(f(\mathbf{x})\right) + \sigma(g^-(\mathbf{x}')).$$

The goal of the learning is to determine the parameters of the

[2]Both the examination bias and trust bias depend on the ranking position (see Eq. (3)). We use the one-hot ranking position representations as the inputs of the bias models.

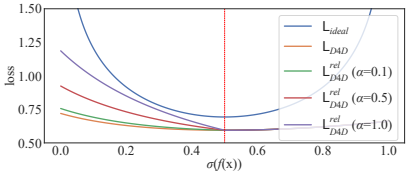


Figure 3: Curves of $\mathcal{L}_{\text{ideal}}$, $\mathcal{L}_{\text{DRD}}$ and $\mathcal{L}^{\text{rel}}_{\text{DRD}}$ w.r.t. $\sigma(f(\mathbf{x}))$. Other parameters: $p_r = 0.5, \sigma(g^+(\mathbf{x}')) = 0.46$, and $\sigma(g^-(\mathbf{x}')) = 0.10$. The red dashed line indicates when $\sigma(f(\mathbf{x})) = p_r$

three neural networks: Θ^+ and Θ^- respectively from the two bias models $g^+(\mathbf{x}')$ and $g^-(\mathbf{x}')$, and Θ^r from the relevance model $f(\mathbf{x})$.

4.2.2 The joint learning algorithm. Algorithm 1 shows the joint learning procedure. After initializing the parameters, the algorithm runs T iterations. Each iteration consists of two parts: the first part updates the relevance model parameters Θ^r, and the second part optimizes the bias models' parameters Θ^+ and Θ^-.

In the first part (lines 3-7) and at each of the batch b, a batch of data $\mathcal{D}^b$ is randomly sampled. Based on $\mathcal{D}^b$, the loss for training relevance model is constructed and optimized:

$$\mathcal{L}^{\text{rel}}_{\text{DRD}} = \xi \cdot \mathcal{L}_{\text{DRD}}(\mathcal{D}^b), \tag{6}$$

where $\mathcal{L}_{\text{DRD}}(\mathcal{D}^b)$ calculates the loss $\mathcal{L}_{\text{DRD}}$ defined in Equation (5) based on the sampled batch of click log $\mathcal{D}^b$, ξ is an adjustment term

$$\xi = \exp\left(\alpha \left(p^*_r - \sigma(f(\mathbf{x}))\right) \frac{\text{sgn}\left(p^*_r - \sigma(f(\mathbf{x}))\right) + 1}{2}\right),$$

where $\alpha \in (0, 1)$ controls the adjustment degree, $p^*_r = \frac{p_c - \sigma(g^-(\mathbf{x}'))}{\sigma(g^+(\mathbf{x}')) - \sigma(g^-(\mathbf{x}'))}$ is the global optima of $\mathcal{L}_{\text{DRD}}$, $\text{sgn}(\cdot) = 1$ if the input is positive otherwise -1. Intuitively, the adjustment loss avoids the gradient vanishing problem in directly optimizing $\mathcal{L}_{\text{DRD}}$ while still achieving the unbiased relevance model. More discussions on ξ are given in Section 4.2.3.

In the second part (lines 8-12) and at each of the batch b, a batch of data $\mathcal{D}^b$ is randomly sampled, and the loss for updating bias models based on $\mathcal{D}^b$ is constructed:

$$\mathcal{L}^{\text{bias}}_{\text{DRD}} = \mathcal{L}_{\text{DRD}}(\mathcal{D}^b) - \epsilon, \tag{7}$$

where $\mathcal{L}_{\text{DRD}}(\mathcal{D}^b)$ calculates the loss $\mathcal{L}_{\text{DRD}}$ based on the sampled batch of data $\mathcal{D}^b$, and the regularizer term

$$\epsilon = \beta \cdot \log \sigma(g^+(\mathbf{x}') - g^-(\mathbf{x}')),$$

where $\beta \in (0, 1)$ controls the regularization degree. Intuitively, ϵ introduces the prior knowledge that users trust the examined relevant documents more than the examined irrelevant ones.

Next, we explain the adjustment term ξ and regularizer ϵ.

4.2.3 Discussion on the joint learning algorithm. Instead of using the original loss $\mathcal{L}_{\text{DRD}}$ in Equation (5), Algorithm 1 resorts to the adjusted losses $\mathcal{L}^{\text{rel}}_{\text{DRD}}$ and $\mathcal{L}^{\text{bias}}_{\text{DRD}}$. We explain the reasons as follows.

First, the adjustment in $\mathcal{L}^{\text{rel}}_{\text{DRD}}$ avoids the problem of gradient vanishing while still achieving the unbiased relevance model. Figure 3 illustrates the curves of $\mathcal{L}_{\text{ideal}}$, $\mathcal{L}_{\text{DRD}}$, and $\mathcal{L}^{\text{rel}}_{\text{DRD}}$ (with different α values) w.r.t. $\sigma(f(\mathbf{x}))$. We can find that compared to $\mathcal{L}_{\text{ideal}}$, the curve of $\mathcal{L}_{\text{DRD}}$ is more flattened, resulting in gradient vanishing and further hurting the learning of the relevance model. Empirical studies also showed that it tends to collapse at the points when $\hat{p}_{\text{x}}$ is close to 0 or 1. The adjusted loss $\mathcal{L}^{\text{rel}}_{\text{DRD}}$, however, increased the losses (and also the gradients) when $\hat{p}_{\text{x}}$ is close to 0 or 1. Also, the adjustment is based on the global optima of $\mathcal{L}_{\text{DRD}}$ and does not change the optimal point. Optimizing $\mathcal{L}^{\text{rel}}_{\text{DRD}}$ can still achieve the same unbiased relevance model as that of optimizing $\mathcal{L}_{\text{DRD}}$.

Second, the regularizer in $\mathcal{L}^{\text{bias}}_{\text{DRD}}$ introduces the prior knowledge that an examined relevant document has higher click probability than an examined irrelevant one when they are ranked at the same position, i.e., $\sigma(g^+(\mathbf{x}')) > \sigma(g^-(\mathbf{x}'))$. The prior knowledge is implemented as the regularizer term $\log \sigma(g^+(\mathbf{x}') - g^-(\mathbf{x}'))$ that forces the logit $g^+(\mathbf{x}')$ greater than $g^-(\mathbf{x}')$. The regularizer helps the algorithm to learn more reasonable bias models $g^+(\mathbf{x}')$ and $g^-(\mathbf{x}')$. Also, forcing $\sigma(g^+(\mathbf{x}')) - \sigma(g^-(\mathbf{x}')) > 0$ during the bias model learning is beneficial to the learning of the relevance models.

5 EXPERIMENT SETUP

We conducted experiments to evaluate the proposed DRD and the baselines. The source code is available at the anonymous site https://anonymous.4open.science/r/DecomposedRankingDebiasing-DC82/.

Datasets: Two widely used public datasets, YahooC14B [6] and WEB30K [23] were used in the experiments. YaHooC14B contains around 30,000 queries, each associated with about 24 documents on average. Each query-document pair is depicted with a 700-dimension feature vector and five-grade relevance labels. WEB10K contains 30,000 queries and each associated with about 125 documents. Each query-document pair is depicted with a 136-dimension feature vector and a five-grade relevance label.

Following the practices in [3, 13], we converted the graded relevance label $y \in \{0, 1, 2, 3, 4\}$ in both two datasets into probability $\Pr(R = 1 \mid y) = \frac{2^y - 1}{2^{y_{max}} - 1}$, where $y_{max} = 4$ is the maximum of y.

Click simulation: Following the practices in [17], the user clicks on search engines were simulated. First, 1% of the labeled data were randomly sampled and used to train an SVM^{rank} [14] as the production ranker π_0. Then for each click session, a query was uniformly sampled, and the ranking result was generated by π_0. To simulate users' clicks, we set up the parameters of examination bias and trust bias according to existing studies: for each (q, d) pair, the examination probability is based on the displayed position: $\Pr(E = 1 \mid k) = (p^k_{eye})^\eta$, where $[p^k_{eye}]^{10}_{k=1} = [0.68, 0.61, 0.48, 0.34, 0.28, 0.20, 0.11, 0.10, 0.08, 0.06]$ are the examination probabilities via eye-tracking experiments [15], and η is the parameter to control the severity of examination bias. Note that p^k_{eye} has values only at the top-10 positions. Therefore, all of our experiments are based on the top-10 rankings. Although the cut-off will cause selection bias [19, 21], it doesn't affect the conclusion of the examination and trust bias. We follow the setting in [20, 30, 31] and set the trust bias parameters as follows:

$$\Pr(C = 1 \mid R = 1, E = 1, P = k) = (98 - k)/100,$$
$$\Pr(C = 1 \mid R = 0, E = 1, P = k) = \rho/(k + 1),$$

where ρ is the probability of the user clicking irrelevant results at the first position. This probability controls the severity of trust bias. In our experiments, we consider $\rho \in \{0.338, 0.638, 0.938\}$ and $\eta \in \{0.5, 1.0, 1.5\}$. These bias parameters are utilized to generate

Table 1: Ranking accuracy of debiasing methods with or without known bias parameters on YaHooC14B and WEB10K. For each group of methods, boldface means the best performed methods (excluding Oracle), while underline means the second best performed methods. Superscripts † means the significance compared to second best performed methods with $p < 0.05$. Experimental settings: $\eta = 1.0$, $\rho = 0.638$, 10^6 click sessions.

Method	YahooC14B					WEB30K				
	MAP	nDCG@k				MAP	nDCG@k			
		k=1	k=3	k=5	k=10		k=1	k=3	k=5	k=10
Naive	0.853	0.589	0.610	0.642	0.702	0.520	0.162	0.182	0.200	0.237
Debiasing methods that need true propensity										
IPS-exam [17]	0.867	0.645	0.653	0.677	0.730	0.557	0.261	0.277	0.292	0.327
Bayes-IPS [2]	0.866	0.662	0.666	0.690	0.740	0.569	0.299	0.304	0.318	0.349
AC [31]	0.872	0.673	0.679	0.702	0.750	0.600	0.366	0.364	0.372	0.398
DRD-ideal	**0.876**	**0.680**†	**0.686**†	**0.709**†	**0.755**	**0.609**†	**0.380**†	**0.379**†	**0.385**†	**0.406**
Debiasing methods that estimate propensity from interaction data										
DLA [3]	0.867	0.656	0.662	0.684	0.736	0.581	0.304	0.317	0.332	0.361
relEM-AC [31]	0.871	0.642	0.662	0.690	0.741	0.597	0.346	0.349	0.361	0.387
MBC [30]	0.858	0.647	0.653	0.677	0.730	0.602	0.360	0.359	0.370	0.395
DRD	**0.873**	**0.678**†	**0.683**†	**0.704**†	**0.753**†	**0.604**	**0.379**†	**0.375**†	**0.383**†	**0.405**†
Oracle	0.876	0.683	0.686	0.710	0.758	0.610	0.382	0.382	0.389	0.406

clicks for training. The annotated labels in the validation set and test sets were used to select models and evaluate the ranking accuracy.

Baselines: Several state-of-the-art ULTR models were chosen as the baselines. First, we choose several debiasing methods that need true propensity to correct biased interactions, including **AC** [31], a correction-based method that handles both trust bias and examination bias; **Bayes-IPS** [2], another correction-based method using Bayes rules; **IPS-exam** [17], a debiasing method adapted from Propensity SVM. Note that it is designed for only addressing the examination bias. In the experiments, we used the relevance estimator in [17] and trained the model with point-wise loss. To make a fair comparison with this group of methods, we also used the true bias parameters when learning the relevance model of DRD, denoted as **DRD-ideal**.

The baselines also include the debiasing methods that estimate propensity from the user clicks, including **relEM-AC** [31], a method that first estimate examination and trust bias parameters with regression EM, and then apply AC [31]; **DLA** [3], a joint learning debiasing method that only handles the examination bias; **MBC** [30], a method that employs a standard EM to estimate the relevance of (q, d) at each position. Then, it uses the estimated relevance label for unbiased learning. For a fair comparison with this group of methods, the proposed **DRD** employs its joint-learning algorithm for training the parameters in the bias models and relevance model.

We also report the results of the **Naive** method that optimizes $\mathcal{L}_{\text{naive}}$ in Equation (2), and the **Oracle** method that optimizes $\mathcal{L}_{\text{ideal}}$ in Equation (1). They are respectively used as the theoretical lower and upper bounds in the experiments.

Implementation details: Similar to existing studies [3, 30, 31], we used three 3-layer neural networks with *elu* activation function as the ranking model $f(\mathbf{x})$ and bias models $g^+(\mathbf{x})$ and $g^-(\mathbf{x})$. The hidden sizes were set to $\{256, 128, 64\}$. We utilized the dropout probability of 0.1 in the last two layers. The batch size was set to 128. The learning rates η_1 and η_2 were tuned among $\{2e{-}4, 5e{-}4, 1e{-}3, 2e{-}3, 5e{-}3\}$. The adjust degree α in Equation (6) was tuned between $[0.0, 0.6]$ for known bias parameters, and between $[0.4, 1.4]$ for joint learning with a strategies that decay with epochs. The regularize degree β in Equation (7) was tuned between $[0.4, 0.6]$. Note that in the setting of known bias parameters, we still use Equation (6) to improve our model learning. In all of the experiments, the reported values were the averaged results after training with 5 different random seeds.

6 EXPERIMENTAL RESULTS AND ANALYSIS

6.1 Overall Performance Comparison

Table 1 reports the ranking accuracy of the proposed DRD and the baselines on YahooC14B and WEB30K. The experimental results are grouped into methods that need true propensities and methods that estimate propensity from the click log. Please note that to make fair comparisons with the baselines that need true propensities, we degenerate the proposed DRD so that the relevance model is learned with true bias parameters, denoted as "DRD-ideal" in Table 1. The results indicate that the proposed DRD and DRD-ideal outperformed the corresponding baselines. We conducted significant tests, and the results showed significant improvements (t-test and p-value <0.05). The DRD-ideal performed slightly better than DRD because DRD-ideal knows the true propensity. The results indicate that DRD and DRD-ideal can learn unbiased relevance models with lower variance.

Among the baselines needing true propensity, AC performed best because it can address both examination and trust bias. Therefore, the performance gaps between AC and Oracle are mainly from the high learning variance. Though AC already achieved relatively high ranking accuracy (especially on YahooC14B), DRD-ideal can further outperform AC. The results verified the theoretical conclusion in Remark 1 that DRD can effectively reduce the variance.

Among the baselines that estimate propensity from the click, relEM-AC and MBC learn the bias and relevance models in two separate stages. In addition, DLA uses a joint learning algorithm while

Table 2: Ranking accuracy of AC equipped with propensity clip (denoted as '+clip_val' where val is the clip value) and self normalization (denoted as 'self norm').

Method	YahooC14B		WEB30K	
	MAP	nDCG@3	MAP	nDCG@3
AC	0.872	0.679	0.600	0.364
+ clip_0.05	0.872	0.679	0.600	0.365
+ clip_0.10	0.874	0.683	0.601	0.372
+ clip_0.30	0.875	0.678	0.601	0.366
+ self norm	0.873	0.680	0.603	0.362
DRD-ideal	**0.876**	**0.686**	**0.609**	**0.379**

only addressing the examination bias. We see that DRD significantly outperformed all these baselines, indicating the effectiveness of DRD's joint learning algorithm in bias parameters estimation.

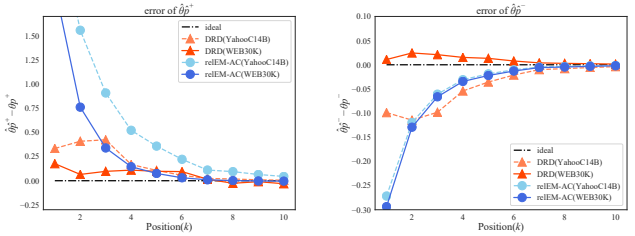
6.2 Effectiveness in Reducing Variance

We conducted experiments to test DRD's ability to reduce the estimation variance. More specifically, we equipped the best-performed baseline AC with the general variance techniques of propensity clip [20, 27, 32] and self normalization [28, 29], achieving new strong baselines of variance reduced AC [3]. From the results reported in Table 2, we found that DRD-ideal outperformed all the AC variations. The results are not surprising because propensity clip and self normalization are general-purpose variance reduction methods and cannot handle biased user clicks in relevance ranking well: (1) both propensity clipping and self normalization break the unbiasedness; (2) propensity clipping is very sensitive to the clipping value. The results verified the advantages of DRD in terms of reducing estimation variance while keeping unbiasedness.

6.3 Effectiveness on Estimating Propensities

Besides estimating the unbiased relevance model, DRD can also accurately estimate the bias models via the tailored joint learning algorithm. Based on YahooC14B and WEB30K, we conducted experiments to show the deviations $\hat{\theta}\hat{p}^+ - \theta p^+$ and $\hat{\theta}\hat{p}^- - \theta p^-$ (calculated according to the click simulation in Section 5). In the experiment, DRD was compared with the baseline relEM-AC. Note that MBC and DLA cannot be used here because MBC directly estimates the relevance label and has no bias terms, and DLA only addresses the examination bias. Figure 4(a) and (b) respectively shows the deviation curves of $\hat{\theta}\hat{p}^+ - \theta p^+$ and $\hat{\theta}\hat{p}^- - \theta p^-$ on different ranking positions. The curves close to the horizontal line mean accurate bias estimation. From the results, we found that the DRD curves, including DRD(YahooC14B) and DRD(WEB30K) are much closer to the true horizontal line than that of relEM-AC(YahooC14B) and relEM-AC(WEB30K). The phenomenon can be observed for both $\hat{\theta}\hat{p}^+$ and $\hat{\theta}\hat{p}^-$, indicating that DRD estimated the propensities more accurately than relEM-AC. Moreover, the figure indicate that $\hat{\theta}\hat{p}^+$ and $\hat{\theta}\hat{p}^-$ are more accurately estimated at the ranking positions of $k = 1$ and 2. Considering that high-ranked documents have more effects on the overall ranking accuracy, we conclude that DRD improves the ranking accuracy by accurately estimating the propensities for high-ranked documents.

[3]These two methods are applied to the propensities in the denominators in our study.



(a) Deviation of $\hat{\theta}\hat{p}^+$ from true values (b) Deviation of $\hat{\theta}\hat{p}^-$ from true values

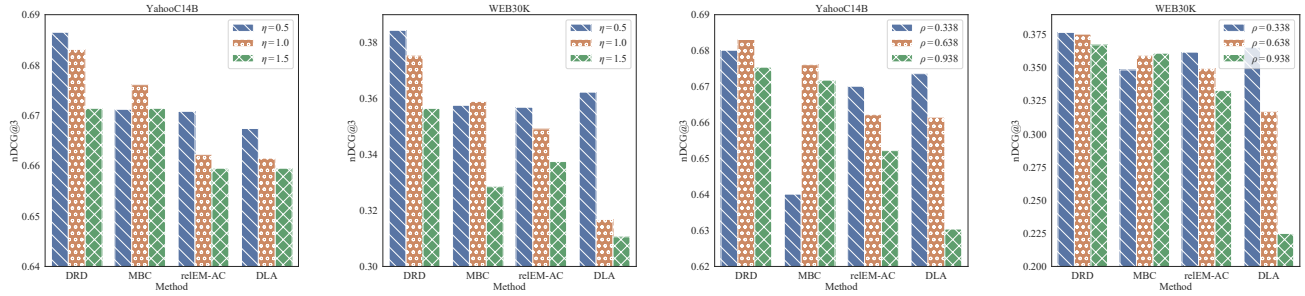
Figure 4: Deviation of the estimated bias terms from the their true values at different ranking positions. The dark horizontal line means the ideal estimation has zero deviation.

6.4 Effects at Varying Bias Severity Levels

We also conducted experiments to test DRD when facing varying levels of bias severity. Specifically, based on YahooC14B and WEB30K, we varied the examination and trust bias severity levels by changing the parameters η and ρ during the click simulation (details described in Section 5), resulting in datasets with different bias severity levels. Figure 5 shows the ranking accuracy (nDCG@3) of different methods respectively trained and tested on these datasets. For example, Figure 5(a) shows the nDCG@3 of different methods on three YahooC14B datasets whose clicks are respectively simulated with $\eta = 0.5, 1.5$, and 1.5 ($\rho = 0.638$ was kept as the default value). From Figure 5(a) and (b), we can see that increasing the level of the examination bias severity (i.e., larger η) degrades the performance of all methods. Compared to the baselines, DRD achieved the best performance in all levels of examination bias severity. From Figure 5(c) and (d), we can see that increasing the level of the trust bias severity (i.e., larger ρ while keeping $\eta = 1.0$) has different effects for different methods. DRD still performed the best for all levels of the trust bias severity. We also observed that the methods that address both examination bias and trust bias (i.e., DRD, relEM-AC, and MBC) are not sensitive to the change of trust bias severity. DLA only addresses the examination bias and is more sensitive to the trust bias severity. All the results indicate that DRD deals with the bias at different severity levels better.

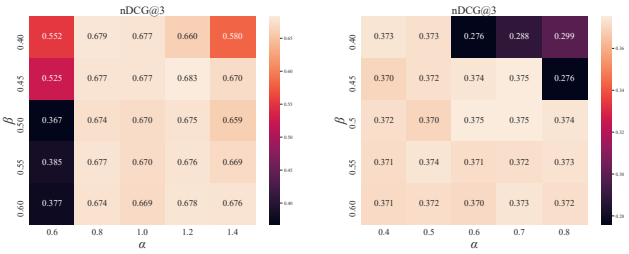
6.5 Sensitivity to the Hyper-parameters

Finally, we tested the effects of the hyper-parameters α and β, which are respectively used to control the adjustment term and regularizer term during the joint learning. Specifically, we trained and tested the performance of DRD on YahooC14B and WEB30K, with different α and β values. The rows and columns in Figure 6 indicate different combinations of (α, β) values, and the numbers in the cells are the corresponding test nDCG@3. From the results, can see best performed combinations are $\alpha = 1.2$, $\beta = 0.45$ on YahooC14B, and $\alpha = 0.6$, $\beta = 0.5$ on WEB30K. Also, the nDCG@3 values near the best-performed combinations did not change sharply, indicating that DRD is not very sensitive to the hyper-parameter settings.



(a) YahooC14B, varying η, fix $\rho = 0.638$ (b) WEB30K, varying η, fix $\rho = 0.638$ (c) YahooC14B, varying ρ, fix $\eta = 1.0$ (d) WEB30K, varying ρ, fix $\eta = 1.0$

Figure 5: (a) and (b): nDCG@3 w.r.t. different severity levels of examination bias (controlled by η); (c) and (d): nDCG@3 w.r.t. different severity levels of trust bias (controlled by ρ).



(a) YahooC14B (b) WEB30K

Figure 6: Performance of DRD on different degrees of adjustment α and regularization β.

7 CONCLUSION

In this paper, we proposed a novel ULTR method called Decomposed Ranking Debiasing (DRD) to address the examination bias and trust bias in relevance ranking. Compared to existing propensity weighted methods, DRD decomposes each click prediction as a combination of a relevance term and other bias terms, and thus avoids involving the propensities in the denominator of labels. A joint learning algorithm is proposed to estimate the model parameters. Theoretical analysis showed DRD has the ability to learn unbiased relevance models with lower variances than existing methods. Groups of empirical studies also verified that DRD improved the baselines through effectively reducing the learning variances and accurately estimating the bias terms.

ACKNOWLEDGMENTS

This work was funded by the National Key R&D Program of China (2019YFE0198200), National Natural Science Foundation of China (61872338, 62006234, 61832017, U2001212), Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098. Work partially done at Beijing Key Laboratory of Big Data Management and Analysis Methods.

A PROOF OF THEOREMS

This section shows the proofs of the theorems.

A.1 Proof of Theorem 1

Proof. According to Theorem 4.4 in [27], variance of binary cross entropy loss that only handles examination bias as propensity is:

$$\mathbb{V}_{\text{exam}} = \frac{1}{|\mathcal{D}|^2}\sum_{\mathcal{D}} \log^2\left(\frac{\hat{p}_{\mathbf{x}}}{1-\hat{p}_{\mathbf{x}}}\right)\mathbb{V}\left[\frac{c}{\theta}\right] = \frac{1}{|\mathcal{D}|^2}\sum_{\mathcal{D}} \log^2\left(\frac{\hat{p}_{\mathbf{x}}}{1-\hat{p}_{\mathbf{x}}}\right)\left(\frac{\mathbb{V}_c}{\theta^2}\right),$$

where $\mathbb{V}_c$ is the variance of clicks. According to Lemma 1

$$\mathbb{V}_{\text{AC}} = \frac{1}{|\mathcal{D}|^2}\sum_{\mathcal{D}} \log^2\left(\frac{\hat{p}_{\mathbf{x}}}{1-\hat{p}_{\mathbf{x}}}\right)\left(\frac{\mathbb{V}_c}{\theta^2\Delta_p^2}\right).$$

It is easy to know $p^+, p^- \in (0,1)$, then $\Delta_p^2 \in (0,1)$ and $\theta^2\Delta_p^2 < \theta^2$. Then, we have $\mathbb{V}_{\text{AC}} \geq \mathbb{V}_{\text{exam}}$ because $\frac{\mathbb{V}_c}{\theta^2\Delta_p^2} \geq \frac{\mathbb{V}_c}{\theta^2}$. □

A.2 Proof of Theorem 2

Proof. The expectation of $\mathcal{L}_{\text{DRD}}$ given a specific $\mathbf{x}$ is:

$$\mathbb{E}[\mathcal{L}_{\text{DRD}}|\mathbf{x}] = -p_c \log\left(\theta\Delta_p\hat{p}_{\mathbf{x}} - \theta p^-\right) + (1-p_c)\log\left(1-\theta\Delta_p\hat{p}_{\mathbf{x}} - \theta p^-\right).$$

$\mathbb{E}[\mathcal{L}_{\text{DRD}}|\mathbf{x}]$ is a convex function w.r.t. $\hat{p}_{\text{x}}$ because

$$\frac{\partial^2\,\mathbb{E}[\mathcal{L}_{\text{DRD}}|\mathbf{x}]}{\partial^2\,\hat{p}_{\mathbf{x}}} = \frac{p_c\theta^2\Delta_p^2}{\left(\theta\Delta_p\hat{p}_{\mathbf{x}} + \theta p^-\right)^2} + \frac{(1-p_c)\theta^2\Delta_p^2}{\left(1-\theta\Delta_p\hat{p}_{\mathbf{x}} - \theta p^-\right)^2} \geq 0.$$

Therefore, optimizing $\mathbb{E}[\mathcal{L}_{\text{DRD}}|\mathbf{x}]$ converges to the global optima:

$$\theta\left(p^+ - p^-\right)\hat{p}^*_{\mathbf{x}} + \theta p^- = p_c \Longrightarrow \hat{p}^*_{\mathbf{x}} = \frac{p_c - \theta p^-}{\theta(p^+ - p^-)} = p_r.$$

We conclude that DRD converges to the unbiased solution. □

A.3 Proof of Theorem 3

Proof.

$$\begin{aligned}\mathbb{V}_{\text{DRD}} &= \frac{1}{|\mathcal{D}|^2}\sum_{\mathcal{D}}\mathbb{V}\left[-c\log\hat{p}_c - (1-c)\log(1-\hat{p}_c)\right]\\ &= \frac{1}{|\mathcal{D}|^2}\sum_{\mathcal{D}}\mathbb{V}\left[-\log(\frac{\hat{p}_c}{1-\hat{p}_c})c - \log(1-\hat{p}_c)\right]\\ &= \frac{1}{|\mathcal{D}|^2}\sum_{\mathcal{D}}\log^2\left(\frac{\theta\Delta_p\hat{p}_{\mathbf{x}} + \theta p^-}{1-\theta\Delta_p\hat{p}_{\mathbf{x}} - \theta p^-}\right)\mathbb{V}_c,\end{aligned}$$

where the last line replaces $\hat{p}_c$ with $\theta\Delta_p\hat{p}_{\mathbf{x}} + \theta p^-$ (Eq. (4)). □

REFERENCES

[1] Aman Agarwal, Kenta Takatsu, Ivan Zaitsev, and Thorsten Joachims. 2019. A General Framework for Counterfactual Learning-to-Rank. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 5–14. https://doi.org/10.1145/3331184.3331202

[2] Aman Agarwal, Xuanhui Wang, Cheng Li, Michael Bendersky, and Marc Najork. 2019. Addressing Trust Bias for Unbiased Learning-to-Rank. In *The World Wide Web Conference*. ACM, New York, NY, USA, 4–14.

[3] Qingyao Ai, Keping Bi, Cheng Luo, Jiafeng Guo, and W. Bruce Croft. 2018. Unbiased Learning to Rank with Unbiased Propensity Estimation. In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*. ACM, New York, NY, USA, 385–394. https://doi.org/10.1145/3209978.3209986

[4] Qingyao Ai, Jiaxin Mao, Yiqun Liu, and W. Bruce Croft. 2018. Unbiased Learning to Rank: Theory and Practice. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 1–2. https://doi.org/10.1145/3234944.3234980

[5] Qingyao Ai, Tao Yang, Huazheng Wang, and Jiaxin Mao. 2021. Unbiased Learning to Rank: Online or Offline? *ACM Trans. Inf. Syst.* 39, 2, Article 21 (Feb. 2021), 29 pages. https://doi.org/10.1145/3439861

[6] Olivier Chapelle and Yi Chang. 2011. Yahoo! learning to rank challenge overview. In *Proceedings of the learning to rank challenge*. PMLR, 1–24.

[7] Mouxiang Chen, Chenghao Liu, Jianling Sun, and Steven C.H. Hoi. 2021. Adapting Interactional Observation Embedding for Counterfactual Learning to Rank. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 285–294. https://doi.org/10.1145/3404835.3462901

[8] Miroslav Dudík, John Langford, and Lihong Li. 2011. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*. 1097–1104.

[9] Zhichong Fang, Aman Agarwal, and Thorsten Joachims. 2019. Intervention Harvesting for Context-Dependent Examination-Bias Estimation. ACM, New York, NY, USA.

[10] Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. 2020. A Survey of Learning Causality with Data: Problems and Methods. 53, 4 (2020).

[11] Ruocheng Guo, Xiaoting Zhao, Adam Henderson, Liangjie Hong, and Huan Liu. 2020. Debiasing Grid-Based Product Search in E-Commerce. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*. Association for Computing Machinery, New York, NY, USA, 2852–2860. https://doi.org/10.1145/3394486.3403336

[12] Siyuan Guo, Lixin Zou, Yiding Liu, Wenwen Ye, Suqi Cheng, Shuaiqiang Wang, Hechang Chen, Dawei Yin, and Yi Chang. 2021. Enhanced Doubly Robust Learning for Debiasing Post-Click Conversion Rate Estimation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 275–284. https://doi.org/10.1145/3404835.3462917

[13] Ziniu Hu, Yang Wang, Qu Peng, and Hang Li. 2019. Unbiased LambdaMART: An Unbiased Pairwise Learning-to-Rank Algorithm. ACM, New York, NY, USA.

[14] Thorsten Joachims. 2002. Optimizing Search Engines Using Clickthrough Data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, USA, 133–142. https://doi.org/10.1145/775047.775067

[15] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately Interpreting Clickthrough Data as Implicit Feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 154–161. https://doi.org/10.1145/1076034.1076063

[16] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. 2007. Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search. 25, 2 (2007).

[17] 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*. ACM, New York, NY, USA, 781–789. https://doi.org/10.1145/3018661.3018699

[18] Jae-woong Lee, Seongmin Park, Joonseok Lee, and Jongwuk Lee. 2022. Bilateral Self-Unbiased Learning from Biased Implicit Feedback. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 29–39. https://doi.org/10.1145/3477495.3531946

[19] Harrie Oosterhuis and Maarten de Rijke. 2020. Policy-Aware Unbiased Learning to Rank for Top-k Rankings. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 489–498. https://doi.org/10.1145/3397271.3401102

[20] Harrie Oosterhuis and Maarten de Rijke. 2021. Unifying Online and Counterfactual Learning to Rank: A Novel Counterfactual Estimator That Effectively Utilizes Online Interventions. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. ACM, New York, NY, USA, 463–471. https://doi.org/10.1145/3437963.3441794

[21] Zohreh Ovaisi, Ragib Ahsan, Yifan Zhang, Kathryn Vasilaky, and Elena Zheleva. 2020. Correcting for Selection Bias in Learning-to-Rank Systems. In *Proceedings of The Web Conference 2020*. ACM, New York, NY, USA, 1863–1873. https://doi.org/10.1145/3366423.3380255

[22] Judea Pearl. 2009. *Causality*. Cambridge university press.

[23] Tao Qin and Tie-Yan Liu. 2013. Introducing LETOR 4.0 datasets. *arXiv preprint arXiv:1306.2597* (2013).

[24] Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting Clicks: Estimating the Click-through Rate for New Ads. In *Proceedings of the 16th International Conference on World Wide Web*. Association for Computing Machinery, New York, NY, USA, 521–530. https://doi.org/10.1145/1242572.1242643

[25] Yuta Saito. 2020. Doubly Robust Estimator for Ranking Metrics with Post-Click Conversions. In *Proceedings of the 14th ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 92–100. https://doi.org/10.1145/3383313.3412262

[26] Yuta Saito. 2020. Unbiased Pairwise Learning from Biased Implicit Feedback. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 5–12. https://doi.org/10.1145/3409256.3409812

[27] 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*. Association for Computing Machinery, New York, NY, USA, 501–509. https://doi.org/10.1145/3336191.3371783

[28] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*. PMLR, 1670–1679.

[29] Adith Swaminathan and Thorsten Joachims. 2015. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems* 28 (2015).

[30] Ali Vardasbi, Maarten de Rijke, and Ilya Markov. 2021. Mixture-Based Correction for Position and Trust Bias in Counterfactual Learning to Rank. In *Proceedings of the 30th ACM International Conference on Information Knowledge Management*. Association for Computing Machinery, New York, NY, USA, 1869–1878. https://doi.org/10.1145/3459637.3482275

[31] Ali Vardasbi, Harrie Oosterhuis, and Maarten de Rijke. 2020. When Inverse Propensity Scoring Does Not Work: Affine Corrections for Unbiased Learning to Rank. In *Proceedings of the 29th ACM International Conference on Information Knowledge Management*. ACM, New York, NY, USA, 1475–1484. https://doi.org/10.1145/3340531.3412031

[32] Xiao Zhang, Sunhao Dai, Jun Xu, Zhenhua Dong, Quanyu Dai and Ji-Rong Wen. 2022. Counteracting User Attention Bias in Music Streaming Recommendation via Reward Modification. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2504–2514.

[33] Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. 2016. Learning to Rank with Selection Bias in Personal Search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 115–124.

[34] Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. 2021. Model-Agnostic Counterfactual Reasoning for Eliminating Popularity Bias in Recommender System. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery Data Mining*. Association for Computing Machinery, New York, NY, USA, 1791–1800. https://doi.org/10.1145/3447548.3467289

[35] Peng Wu, Haoxuan Li, Yuhao Deng, Wenjie Hu, Quanyu Dai, Zhenhua Dong, Jie Sun, Rui Zhang, and Xiao-Hua Zhou. 2022. On the Opportunity of Causal Learning in Recommendation Systems: Foundation, Estimation, Prediction and Challenges. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. 5646–5653. Survey Track.

[36] Xinwei Wu, Hechang Chen, Jiashu Zhao, Li He, Dawei Yin, and Yi Chang. 2021. Unbiased Learning to Rank in Feeds Recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM '21)*. Association for Computing Machinery, New York, NY, USA, 490–498.

[37] Bowen Yuan, Yaxu Liu, Jui-Yang Hsia, Zhenhua Dong, and Chih-Jen Lin. 2020. Unbiased Ad Click Prediction for Position-Aware Advertising Systems. In *Fourteenth ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 368–377. https://doi.org/10.1145/3383313.3412241

[38] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal Intervention for Leveraging Popularity Bias in Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 11–20. https://doi.org/10.1145/3404835.3462875

[39] Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. 2021. Disentangling User Interest and Conformity for Recommendation with Causal Embedding. In *Proceedings of the Web Conference 2021*. ACM, New York, NY, USA, 2980–2991. https://doi.org/10.1145/3442381.3449788