Practical and Robust Safety Guarantees for Advanced Counterfactual Learning to Rank

Shashank Gupta
University of Amsterdam
Amsterdam, The Netherlands
s.gupta2@uva.nl

Harrie Oosterhuis
Radboud University
Nijmegen, The Netherlands
harrie.oosterhuis@ru.nl

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

Abstract

Counterfactual learning to rank (CLTR) can be risky and, in various circumstances, can produce sub-optimal models that hurt performance when deployed. Safe CLTR was introduced to mitigate these risks when using inverse propensity scoring to correct for position bias. However, the existing safety measure for CLTR is not applicable to state-of-the-art CLTR methods, cannot handle trust bias, and relies on specific assumptions about user behavior.

Our contributions are two-fold. First, we generalize the existing safe CLTR approach to make it applicable to state-of-the-art doubly robust CLTR and trust bias. Second, we propose a novel approach, *proximal ranking policy optimization* (PRPO), that provides safety in deployment without assumptions about user behavior. PRPO removes incentives for learning ranking behavior that is too dissimilar to a safe ranking model. Thereby, PRPO imposes a limit on how much learned models can degrade performance metrics, *without* relying on any specific user assumptions. Our experiments show that both our novel safe doubly robust method and PRPO provide higher performance than the existing safe inverse propensity scoring approach. However, in unexpected circumstances, the safe doubly robust approach can become unsafe and bring detrimental performance. In contrast, PRPO always maintains safety, even in maximally adversarial situations. By avoiding assumptions, PRPO is the first method with *unconditional* safety in deployment that translates to robust safety for real-world applications.

CCS Concepts

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

Keywords

Learning to Rank; Counterfactual Learning to Rank; Safety

ACM Reference Format:
Shashank Gupta, Harrie Oosterhuis, and Maarten de Rijke. 2024. Practical and Robust Safety Guarantees for Advanced Counterfactual Learning to Rank. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3627673.3679531

1 Introduction

Counterfactual learning to rank (CLTR) [11, 20, 24, 46] concerns the optimization of ranking systems based on user interaction data using learning to rank (LTR) methods [22]. A main advantage of CLTR is that it does not require manual relevance labels, which are costly to produce [6, 34] and often do not align with actual user preferences [39]. Nevertheless, CLTR also brings significant challenges since user interactions only provide a heavily biased form of implicit feedback [11]. User clicks are affected by many different factors, for example, the position at which an item is displayed in a ranking [8, 47]. Thus, click frequencies provide a biased indication of relevance, that is often more representative of how an item was displayed than actual user preferences [2, 46].

To correct for this bias, early CLTR applied inverse propensity scoring (IPS), which weights clicks inversely to the estimated effect of position bias [20, 46]. Later work expanded this approach to correct for other forms of bias, e.g., item-selection bias [29, 32] and trust bias [2, 45], and more advanced doubly robust (DR) estimation [28]. Using these methods, standard CLTR aims to create an unbiased estimate of relevance (or user preference) from click frequencies. In other words, their goal is to output an estimate per document with an expected value that is equal to their relevance.

However, unbiased estimates of CLTR have their limitations. Firstly, they assume a model of user behavior and require an accurate estimate of this model. If the assumed model is incorrect [29, 45] or its estimated parameters are inaccurate [20, 28], then their unbiasedness is not guaranteed. Secondly, even when unbiased, the estimates are subject to variance [27]. As a result, the actual estimated values are often erroneous, especially when the available data is sparse [14, 28]. Accordingly, unbiased CLTR does not guarantee that the ranking models it produces have optimal performance [27].

Safe counterfactual learning to rank. There are risks involved in applying CLTR in practice. In particular, there is a substantial risk that a learned ranking model is deployed that degrades performance compared to the previous production system [14, 16, 31]. This can have negative consequences to important business metrics, making CLTR less attractive to practitioners. To remedy this issue, a *safe* CLTR approach was proposed by Gupta et al. [14]. Their approach builds on IPS-based CLTR and adds exposure-based risk regularization, which keeps the learned model from deviating too much from a given safe model. Thereby, under the assumption of a position-biased user model, the safe CLTR approach can guarantee an upper bound on the probability of the model being worse than the safe model.

Limitations of the current safe CLTR method. Whilst safe CLTR is an important contribution to the field, it has two severe limitations – both are addressed by this work. Firstly, the existing approach is only applicable to IPS estimation, which is no longer

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679531

the state-of-the-art in the field [11], and it assumes a rank-based position bias model [8, 46], the most basic user behavior model in the field. Secondly, because its guarantees rely on assumptions about user behavior, it can only provide a conditional notion of safety. Moreover, since user behavior can be extremely heterogeneous, it is unclear whether a practitioner could even determine whether the safety guarantees would apply to their application.

Main contributions. Our first contribution addresses the mismatch between the existing safe CLTR approach and recent advances in CLTR. We propose a novel generalization of the exposure-based regularization term that provides safety guarantees for both IPS and DR estimation, also under more complex models of user behavior that cover both position and trust bias. Our experimental results show that our novel method reaches higher levels of performance significantly faster, while avoiding any notable decreases of performance. This is especially beneficial since DR is known to have detrimental performance when very little data is available [28].

Our second contribution provides an unconditional notion of safety. We take inspiration from advances in reinforcement learning (RL) [21, 35, 41, 49, 50] and propose the novel *proximal ranking policy optimization* (PRPO) method. PRPO removes incentives for LTR methods to rank documents too much higher than a given safe ranking model would. Thereby, PRPO imposes a limit on the performance difference between a learned model and a safe model, in terms of standard ranking metrics. Importantly, PRPO is easily applicable to *any* gradient-descent-based LTR method, and makes *no assumptions* about user behavior. In our experiments, PRPO prevents any notable decrease in performance even under extremely adversarial circumstances, where other methods fail. Therefore, we believe PRPO is the first *unconditionally* safe LTR method.

Together, our contributions bring important advances to the theory of safe CLTR, by proposing a significant generalization of the existing approach with theoretical guarantees, and the practical appeal of CLTR, with the first robustly safe LTR method: PRPO. All source code to reproduce our experimental results is available at: https://github.com/shashankg7/cikm-safeultr.

2 Related Work

Counterfactual learning to rank. Joachims et al. [20] introduced the first method for CLTR, a LTR specific adaptation of IPS from the bandit literature [12, 13, 15, 19, 38, 42] to correct for position bias. They weight each user interaction according to the inverse of its examination probability, i.e., its inverse propensity, during learning to correct for the position bias in the logged data. This weighting will remove the effect of position bias from the final ranking policy. Oosterhuis and de Rijke [29] extended this method for the top-K ranking setting with item-selection bias, where any item placed outside the top-K positions gets zero exposure probability, i.e., an extreme form of position bias. They proposed a policy-aware propensity estimator, where the propensity weights used in IPS are conditioned on the logging policy used to collect the data.

Agarwal et al. [2] introduced an extension of IPS, known as Bayes-IPS, to correct for *trust bias*, an extension of position-bias, with false-positive clicks at the higher ranks, because of the users' trust in the search engine. Vardasbi et al. [45] proved that Bayes-IPS cannot correct for trust bias and introduced an affine-correction method and unbiased estimator. Oosterhuis and de Rijke [30] combined the affine-correction with a policy-aware propensity estimator to correct for trust bias and item-selection bias simultaneously. Recently, Oosterhuis [28] introduced a DR-estimator for CLTR, which combines the existing IPS-estimator with a regression model to overcome some of the challenges with the IPS-estimator. The proposed DR-estimator corrects for item-selection and trust biases, with lower variance and improved sample complexity.

Safe policy learning from user interactions. In the context of offline evaluation for contextual bandits, Thomas et al. [43] introduced a high-confidence off-policy evaluation framework. A confidence interval is defined around the empirical off-policy estimates, and there is a high probability that the *true* utility can be found in the interval. Jagerman et al. [16] extended this framework for safe deployment in the contextual bandit learning setup. The authors introduce a safe exploration algorithm (SEA) method that selects with high confidence between a safe behavior policy and the newly learned policy. In the context of LTR, Oosterhuis and de Rijke [31] introduced the generalization and specialization (GENSPEC) method, which safely selects between a feature-based and tabular LTR model. For off-policy learning, Swaminathan and Joachims [42] introduced a counterfactual risk minimization (CRM) framework for the contextual bandit setup. They modify the IPS objective for bandits to include a regularization term, which explicitly controls for the variance of the IPS-estimator during learning, thereby overcoming some of the problems with the high-variance of IPS. Wu and Wang [52] extended the CRM framework by using a *risk* regularization, which penalizes mismatches in the action probabilities under the new policy and the behavior policy. Gupta et al. [14] made this general safe deployment framework effective in the LTR setting. They proposed an exposure-based risk regularization method where the difference in the document exposure distribution under the new and logging policies is penalized. When click data is limited, risk regularization ensures that the performance of the new policy is similar to the logging policy, ensuring safety.

To the best of our knowledge, the method proposed by Gupta et al. [14] is the only method for safe policy learning in the LTR setting. While it guarantees safe ranking policy optimization, it has two main limitations: (i) It is only applicable to the IPS estimator; and (ii) under the position-based click model assumption, the most basic click model in the CLTR literature [11, 20, 24].

Proximal policy optimization. In the broader context of RL, *proximal policy optimization* (PPO) was introduced as a policy gradient method for training RL agents to maximize long-term rewards [21, 35, 41, 49, 50]. PPO clips the importance sampling ratio of action probability under the new policy and the current behavior policy, and thereby, it prevents the new policy to deviate from the behavior policy by more than a certain margin. PPO is not directly applicable to LTR, for the same reasons that the CRM framework is not: the combinatorial action space of LTR leads to extremely small propensities that PPO cannot effectively manage [14].

3 Background

3.1 Learning to rank

The goal in LTR is to find a ranking policy (π) that optimizes a given ranking metric [22]. Formally, given a set of documents (D),

a distribution of queries Q, and the true relevance function ($P(R = 1 \mid d)$), LTR aims to maximize the following utility function:

$$U(\pi) = \sum_{q \in Q} P(q \mid Q) \sum_{d \in D} \omega(d \mid \pi)\ P(R = 1 \mid d), \tag{1}$$

where $\omega(d \mid \pi)$ is the weight of the document for a given policy π. The weight can be set accordingly to optimize for a given ranking objective, for example, setting the weight to:

$$\omega_{\text{DCG}}(d \mid q, \pi) = \mathbb{E}_{y \sim \pi(\cdot \mid q)} \left[(\log_2(\text{rank}(d \mid y) + 1))^{-1} \right], \tag{2}$$

optimizes discounted cumulative gain (DCG) [17]. For this paper, we aim to optimize the expected number of clicks, so we set the weight accordingly [14, 28, 53].

3.2 Assumptions about user click behavior

The optimization of the true utility function (Eq. 1) requires access to the document relevance ($P(R = 1 \mid d)$). In the CLTR setting, the relevances of documents are not available, and instead, click interaction data is used to estimate them [20, 24, 46]. However, naively using clicks to optimize a ranking system can lead to sub-optimal ranking policies, as clicks are a biased indicator of relevance [7, 8, 18, 20]. CLTR work with theoretical guarantees starts by assuming a model of user behavior. The earliest CLTR works [20, 46] assume a basic model originally proposed by Craswell et al. [8]:

Assumption 3.1 (*The rank-based position bias model*)**.** The probability of a click on document d at position k is the product of the rank-based examination probability and document relevance:

$$P(C = 1 \mid d, k) = P(E = 1 \mid k) P(R = 1 \mid d) = \alpha_k P(R = 1 \mid d). \tag{3}$$

Later work has proposed more complex user models to build on [11]. Relevant to our work is the model proposed by Agarwal et al. [2], and its re-formulation by Vardasbi et al. [45]; it is a generalization of the above model to include a form of trust bias:

Assumption 3.2 (*The trust bias model*)**.** The probability of a click on document d at position k is an affine transformation of the relevance probability of d in the form:

$$P(C = 1 \mid d, k) = \alpha_k P(R = 1 \mid d) + \beta_k, \tag{4}$$

where $\forall k, \alpha_k \in [0, 1] \wedge \beta_k \in [0, 1] \wedge (\alpha_k + \beta_k) \in [0, 1]$.

Whilst it is named after trust bias, this model actually captures three forms of bias that were traditionally categorized separately: rank-based position bias, item-selection bias, and trust bias. Position bias was originally approached as the probability that a user would examine an item, which would decrease at lower positions in the ranking [8, 20, 46, 47]. In the trust bias model, this effect can be captured by decreasing $\alpha_k + \beta_k$ as k increases. Additionally, with $\forall k, \beta_k = 0$, the trust bias model is equivalent to the rank-based position bias model. Item-selection bias refers to users being unable to see documents outside a top-K, where they receive zero probability of being examined or interacted with [29]. This can be captured by the trust bias model by setting $\alpha_k + \beta_k = 0$ when $k > K$. Lastly, the key characteristic of trust bias is that users are more likely to click on non-relevant items when they are near the top of the ranking [2]. This can be captured by the model by making β_k larger as k decreases [45]. Thereby, the trust bias model is in fact a generalization of most of the user models assumed by earlier work [11, 30]. The following works all assume models that fit Assumption 3.2: [1, 2, 14, 28–32, 45–47].

3.3 Counterfactual learning to rank

This section details the *policy-aware inverse propensity scoring* (IPS) estimator proposed by Oosterhuis and de Rijke [29] and the *doubly robust* (DR) estimator by Oosterhuis [28].

First, let $\mathcal{D}$ be a set of logged interaction data: $\mathcal{D} = \left\{q_i, y_i, c_i\right\}_{i=1}^{N}$, where each of the N interactions consists of a query q_i, a displayed ranking y_i, and click feedback $c_i(d) \in \{0, 1\}$ that indicates whether the user clicked on the document d or not. Both policies use propensities that are the expected α values for each document:

$$\rho_0(d \mid q_i, \pi_0) = \mathbb{E}_{y \sim \pi_0(q_i)} \left[\alpha_{k(d)} \right] = \rho_{i,0}(d). \tag{5}$$

Similarly, to keep our notation short, we also use $\omega(d \mid q_i, \pi) = \omega_i(d)$. Next, the policy-aware IPS estimator is defined as:

$$\hat{U}_{\text{IPS}}(\pi) = \frac{1}{N} \sum_{i=1}^{N} \sum_{d \in D} \frac{\omega_i(d)}{\rho_{i,0}(d)} c_i(d). \tag{6}$$

Oosterhuis and de Rijke [29] prove that under the rank-based position bias model (Assumption 3.1) and when $\forall (i, d), \rho_{i,0}(d) > 0$, this estimator is unbiased: $\mathbb{E}[\hat{U}_{\text{IPS}}(\pi)] = U(\pi)$.

The DR estimator improves over the policy-aware IPS estimator in terms of assuming the more general trust bias model (Assumption 3.2) and having lower variance. Oosterhuis [28] proposes the usage of the following ω values for the policy π:

$$\omega(d \mid q_i, \pi) = \mathbb{E}_{y \sim \pi(q_i)} \left[\alpha_{k(d)} + \beta_{k(d)} \right] = \omega_i(d), \tag{7}$$

since with these values U (Eq. 1) becomes the number of expected clicks on relevant items under the trust bias model; $U = (\alpha_k + \beta_k) P(R = 1 \mid d, q) = P(C = 1, R = 1 \mid k, d, q)$. We follow this approach and define the ω values for the logging policy π_0 as:

$$\omega_0(d \mid q_i, \pi_0) = \mathbb{E}_{y \sim \pi_0(q_i)} \left[\alpha_{k(d)} + \beta_{k(d)} \right] = \omega_{i,0}(d). \tag{8}$$

The DR estimator uses predicted relevances in its estimation, i.e., using predictions from a regression model. Let $\hat{R}_i(d) \approx P(R = 1 \mid d, q_i)$ indicate a predicted relevance; then the utility according to these predictions is:

$$\hat{U}_{\text{DM}}(\pi) = \frac{1}{N} \sum_{i=1}^{N} \sum_{d \in D} \omega_i(d) \hat{R}_i(d). \tag{9}$$

The DR estimator starts with this predicted utility and adds an IPS-based correction to remove its bias:

$$\hat{U}_{\text{DR}}(\pi) = \hat{U}_{\text{DM}}(\pi) + \frac{1}{N} \sum_{i=1}^{N} \sum_{d \in D} \frac{\omega_i(d)}{\rho_{i,0}(d)} \left(c_i(d) - \alpha_{k_i(d)} \hat{R}_i(d) - \beta_{k_i(d)} \right). \tag{10}$$

Thereby, the corrections of the IPS part of the DR estimator will be smaller if the predicted relevances are more accurate. Oosterhuis [28] proves that under the assumption of the trust bias model (Assumption 3.2), the DR estimator is unbiased when $\forall (i, d), \rho_{i,0}(d) > 0 \vee \hat{R}_i(d) = P(R = 1 \mid d, q_i)$ and has less variance if $0 \leq \hat{R}_i(d) \leq 2P(R = 1 \mid d, q_i)$. They also show that the DR estimator needs less data to reach the same level of ranking performance as IPS, with especially large improvements when applied to top-K rankings [28].

3.4 Safety in counterfactual learning to rank

IPS-based CLTR methods, despite their unbiasedness and consistency, suffer from the problem of high-variance [11, 20, 28]. Specifically, if the logged click data is limited, training an IPS-based method can lead to an unreliable and unsafe ranking policy [14]. The problem of *safe* policy learning is well-studied in the bandit literature [16, 42, 43, 52]. Swaminathan and Joachims [42] proposed the first risk-aware off-policy learning method for bandits, with their risk term quantified as the variance of the IPS-estimator. Wu and Wang [52] proposed an alternative method for risk-aware off-policy learning, where the risk is quantified using a Renyi divergence between the action distribution of the new policy and the logging policy [37]. Thus, both consider it a risk for the new policy to be too dissimilar to the logging policy, which is presumed safe. Whilst effective at standard bandit problems, these risk-aware methods are not effective for ranking tasks due to their enormous combinatorial action spaces and correspondingly small propensities.

As a solution for CLTR, Gupta et al. [14] introduced a risk-aware CLTR approach that uses divergence based on the exposure distributions of policies. They first introduce normalized propensities: $\rho'(d) = \rho/Z$, with a normalization factor Z based on K:

$$Z = \sum_{d \in D} \rho(d) = \sum_{d \in D} \mathbb{E}_{y\sim\pi}\left[\alpha_{k(d)}\right] = \mathbb{E}_{y\sim\pi}\left[\sum_{k=1}^{K} \alpha_{k(d)}\right] = \sum_{k=1}^{K} \alpha_k. \quad (11)$$

Since $\rho'(d) \in [0, 1]$ and $\sum_d \rho'(d) = 1$, they can be treated as a probability distribution that indicates how exposure is spread over documents. Gupta et al. [14] use Renyi divergence to quantify how dissimilar the new policy is from the logging policy:

$$d_2(\rho \,\|\, \rho_0) = \mathbb{E}_q\left[\sum_d \left(\frac{\rho'(d)}{\rho'_0(d)}\right)^2 \rho'_0(d)\right], \quad (12)$$

with the corresponding empirical estimate based on the log data ($\mathcal{D}$) defined as:

$$\hat{d}_2(\rho \,\|\, \rho_0) = \frac{1}{N}\sum_{i=1}^{N}\sum_d \left(\frac{\rho'_i(d)}{\rho'_{i,0}(d)}\right)^2 \rho'_{i,0}(d). \quad (13)$$

Based on this divergence term, they propose the following risk-aware CLTR objective, with parameter δ:

$$\max_\pi \hat{U}_{\text{IPS}}(\pi) - \sqrt{\frac{Z}{N}\left(\frac{1-\delta}{\delta}\right)\hat{d}_2(\rho \,\|\, \rho_0)}. \quad (14)$$

Thereby, the existing safe CLTR approach penalizes the optimization procedure from learning ranking behavior that is too dissimilar from the logging policy in terms of the distribution of exposure. The weight of this penalty decreases as the number of datapoints N increases, thus it maintains the same point of convergence as standard IPS. Yet, initially when little data is available and the effect of variance is the greatest, it forces the learned policy to be very similar to the safe logging policy. Gupta et al. [14] prove that their objective bounds the real utility with a probability of $1 - \delta$:

$$P\left(U(\pi) \geq \hat{U}_{\text{IPS}}(\pi) - \sqrt{\frac{Z}{N}\left(\frac{1-\delta}{\delta}\right) d_2(\rho \,\|\, \rho_0)}\right) \geq 1 - \delta. \quad (15)$$

However, their proof of safety relies on the rank-based position bias model (Assumption 3.1) and their approach is limited to the basic IPS estimator for CLTR.

3.5 Proximal policy optimization

In the more general reinforcement learning (RL) field, *proximal policy optimization* (PPO) was introduced as a method to restrict a new policy π from deviating too much from a previously rolled-out policy π_0 [40, 41]. In contrast with the earlier discussed methods, PPO does not make use of a divergence term but uses a simple clipping operation in its optimization objective. Let s indicate a state, a an action and R a reward function, the PPO loss is:

$$U^{PPO}(s, a, \pi, \pi_0) = \mathbb{E}\left[\min\left(\frac{\pi(a \mid s)}{\pi_0(a \mid s)} R(a \mid s), g\big(\epsilon, R(a \mid s)\big)\right)\right], \quad (16)$$

where g creates a clipping threshold based on the sign of $R(a \mid s)$:

$$g(\epsilon, R(a \mid s)) = \begin{cases} (1+\epsilon)\, R(a \mid s) & \text{if } R(a \mid s) \geq 0, \\ (1-\epsilon)\, R(a \mid s) & \text{otherwise.} \end{cases} \quad (17)$$

The clipping operation removes incentives for the optimization to let π deviate too much from π_0, since there are no further increases in U^{PPO} when $\pi(a \mid s) > (1+\epsilon)\pi_0(a \mid s)$ or $\pi(a \mid s) < (1-\epsilon)\pi_0(a \mid s)$, depending on the sign of $R(a \mid s)$. Similar to the previously discussed general methods, PPO is not effective when directly applied to the CLTR setting due to the combinatorial action space and corresponding extremely small propensities (for most a and s: $\pi_0(a \mid s) \simeq 0$).

4 Extending Safety to Advanced CLTR

In this section, we introduce our first contribution: our extension of the safe CLTR method to address trust bias and DR estimation.

4.1 Method: Safe doubly-robust CLTR

For the safe DR CLTR method, we extend the generalization bound from the existing IPS estimator and position bias [14, Eq. 26] to the DR estimator and trust bias.

Theorem 4.1. *Given the true utility $U(\pi)$ (Eq. 1) and its exposure-based DR estimate $\hat{U}_{DR}(\pi)$ (Eq. 10) of the ranking policy π with the logging policy π_0 and the metric weights ω and ω_0 (Eq. 7 and 8), assuming the trust bias click model (Assumption 3.2), the following generalization bound holds with probability $1 - \delta$:*

$$P\left(U(\pi) \geq \hat{U}_{DR}(\pi) - \left(1 + \max_k \frac{\beta_k}{\alpha_k}\right)\sqrt{\frac{2Z}{N}\left(\frac{1-\delta}{\delta}\right) d_2(\omega \| \omega_0)}\right) \geq 1 - \delta.$$

Proof. For a proof, we refer to the appendix (Theorem A.1). □

Given the novel generalization bound from Theorem 4.1, we define the safe DR CLTR objective as follows:

$$\max_\pi \hat{U}_{\text{DR}}(\pi) - \left(1 + \max_k \frac{\beta_k}{\alpha_k}\right)\sqrt{\frac{2Z}{N}\left(\frac{1-\delta}{\delta}\right)\hat{d}_2(\omega \,\|\, \omega_0)}, \quad (18)$$

where $\hat{d}_2(\omega \,\|\, \omega_0)$ is defined analogously to Eq. 13. The objective optimizes the lower-bound on the true utility function, through a linear combination of the empirical DR estimator ($\hat{U}_{\text{DR}}(\pi)$) and the empirical risk regularization term ($\hat{d}_2(\omega \,\|\, \omega_0)$). In a setting where click data is limited, our safe DR objective will weight the risk regularization term higher, and as a result, the objective ensures that the new policy stays close to the safe logging policy. When a sufficiently high volume of click data is collected, and thus we have higher confidence in the DR estimate, the objective falls back to its DR objective counterpart.

For the choice of the ranking policy (π), we propose to optimize a stochastic ranking policy π with a gradient descent-based method. For the gradient calculation, we refer to previous work [14, 28, 53].

Conditions for safe DR CLTR. Finally, we note that besides the explicit assumption that user behavior follows the trust bias model (Assumption 3.2), there is also an important implicit assumption in this approach. Namely, the approach assumes that the bias parameters (i.e., α and β) are known, a common assumption in the CLTR literature [24, 28]. However, in practice, either of these assumptions could not hold, i.e., user behavior could not follow the trust bias model, or a model's bias parameters could be wrongly estimated. Additionally, in adversarial settings where clicks are intentionally misleading or incorrectly logged [5, 23, 36], the user behavior assumptions do not hold, and, the generalization bound of our DR CLTR is not guaranteed to hold. Thus, whilst it is an important advancement over the existing safe CLTR method [14], our approach is limited to only providing a *conditional* form of safety.

5 Method: Proximal Ranking Policy Optimization (PRPO)

Inspired by the limitations of the method introduced in Section 4 and the PPO method from the RL field (Section 2), we propose the first *unconditionally* safe CLTR method: *proximal ranking policy optimization* (PRPO). Our novel PRPO method is designed for practical safety by making *no assumptions* about user behavior. Thereby, it provides the most robust safety guarantees for CLTR yet.

For safety, instead of relying on a high-confidence bound (e.g., Eq. 14 and 18), PRPO guarantees safety by removing the incentive for the new policy to rank documents too much higher than the safe logging policy. This is achieved by directly clipping the ratio of the metric weights for a given query q_i under the new policy $\omega_i(d)$, and the logging policy ($\omega_{i,0}(d)$), i.e., $\frac{\omega_i(d)}{\omega_{i,0}(d)}$ to be bounded in a fixed predefined range: $[\epsilon_-, \epsilon_+]$. As a result, the PRPO objective provides no incentive for the new policy to produce weights $\omega_i(d)$ outside of the range: $\epsilon_- \cdot \omega_{i,0}(d) \leq \omega_i(d) \leq \epsilon_+ \cdot \omega_{i,0}(d)$.

Before defining the PRPO objective, we first introduce a term $r(d|q)$ that represents an unbiased DR relevance estimate, weighted by ω_0, for a single document-query pair (cf. Eq. 10):

$$r(d|q) = \omega_0(d|q)\hat{R}(d|q) + \frac{\omega_0(d|q)}{\rho_0(d|q)} \sum_{i \in \mathcal{D}: q_i = q} \Big(c_i(d) - \alpha_{k_i(d)} \hat{R}(d|q) - \beta_{k_i(d)} \Big). \quad (19)$$

For the sake of brevity, we drop π and π_0 from the notation when their corresponding value is clear from the context. This enables us to reformulate the DR estimator around the ratios between the metric weights ω and ω_0 (cf. Eq. 10):

$$\hat{U}_{\text{DR}}(\pi) = \sum_{q,d \in \mathcal{D}} \frac{\omega(d \mid q)}{\omega_0(d \mid q)} r(d \mid q). \quad (20)$$

Before defining the proposed PRPO objective, we first define the following clipping function:

$$f(x, \epsilon_-, \epsilon_+, r) = \begin{cases} \min(x, \epsilon_+) \cdot r & r \geq 0, \\ \max(x, \epsilon_-) \cdot r & \text{otherwise.} \end{cases} \quad (21)$$

Given the reformulated DR estimator (Eq. 20), and the clipping function (Eq. 21), the PRPO objective can be defined as follows:

$$\hat{U}_{\text{PRPO}}(\pi) = \sum_{q,d \in \mathcal{D}} f\left(\frac{\omega(d \mid q)}{\omega_0(d \mid q)}, \epsilon_-, \epsilon_+, r(d \mid q) \right). \quad (22)$$

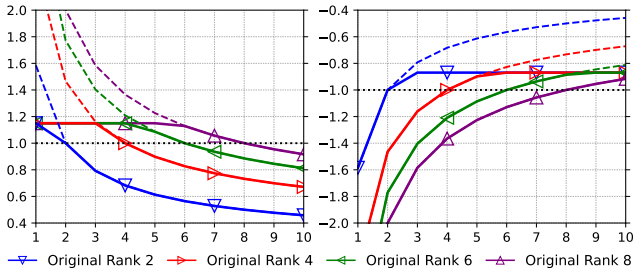

Figure 1: Weight ratios in the clipped PRPO objective (solid lines) and the unclipped counterparts (dashed lines), as documents are moved from four different original ranks. Left: positive relevance, $r = 1$; right: negative relevance, $r = -1$; x-axis: new rank for document; y-axis: unclipped weight ratios (dashed lines), $r \cdot \omega_i(d)/\omega_{i,0}(d)$; and clipped PRPO weight ratios (solid lines), $f\big(\omega_i(d)/\omega_{i,0}(d), \epsilon_- = 1.15^{-1}, \epsilon_+ = 1.15, r = \pm 1\big)$. DCG metric weights used: $\omega_i(d) = \log_2(\text{rank}(d \mid q_i, \pi) + 1)^{-1}$.

Fig. 1 visualizes the effect the clipping of PRPO has on the optimization incentives. We see how the clipped and unclipped weight ratios progress as documents are placed on different ranks. The unclipped weights keep increasing as documents are moved to the top of the ranking, when $r > 1$, or to the bottom, when $r < 1$. Consequently, optimization with unclipped weight ratios aims to place these documents at the absolute top or bottom positions. Conversely, the clipped weights do not increase beyond their clipping threshold, which for most document is reached before being placed at the very top or bottom position. As a result, optimization with clipped weight ratios will not push these documents beyond these points in the ranking. For example, when $r > 0$, we see that there is no incentive to place a document at higher than rank 6, if it was placed at rank 8 by the logging policy. Similarly, placement higher than rank 4 leads to no gain if the original rank was 6, and higher than rank 3 leads to no improvement gain from an original rank of 4. Vice versa, when $r < 0$, each document has a rank, where placing it lower than that rank brings no increase in clipped weight ratio. Importantly, this behavior only depends on the metric and the logging policy; PRPO makes *no further assumptions*.

Whilst the clipping of PRPO is intuitive, we can prove that it provides the following formal form of unconditional safety:

Theorem 5.1. *Let q be a query, ω be metric weights, y_0 be a logging policy ranking, and $y^*(\epsilon_-, \epsilon_+)$ be the ranking that optimizes the PRPO objective in Eq. 22. Assume that $\forall d, \in \mathcal{D}, r(d \mid q) \neq 0$. Then, for any $\Delta \in \mathbb{R}_{\geq 0}$, there exist values for ϵ_- and ϵ_+ that guarantee that the difference between the utility of y_0 and $y^*(\epsilon_-, \epsilon_+)$ is bounded by Δ:*

$$\forall \Delta \in \mathbb{R}_{\geq 0}, \exists \epsilon_- \in \mathbb{R}_{\geq 0}, \epsilon_+ \in \mathbb{R}_{\geq 0};\ |U(y_0) - U(y^*(\epsilon_-, \epsilon_+))| \leq \Delta. \quad (23)$$

Proof. A proof is given in Appendix A.2. □

Adaptive clipping. Theorem 5.1 describes a very robust sense of safety, as it shows PRPO can be used to prevent any given decrease in performance without assumptions. However, it also reveals that

this safety comes at a cost; PRPO prevents both decreases and increases of performance. This is very common in safety approaches, as there is a generally a tradeoff between risks and rewards [14]. Existing safety methods, such as the safe CLTR approach of Section 4, generally, loosen their safety measures as more data becomes available, and the risk is expected to have decreased [43].

We propose a similar strategy for PRPO through adaptive clipping, where the effect of clipping decreases as the number of datapoints N increases. Specifically, we suggest using a monotonically decreasing $\delta(N)$ function such that $\lim_{N\to\infty} \delta(N) = 0$. The ϵ parameters can then be obtained through the following transformation: $\epsilon_- = \delta(N)$ and $\epsilon_+ = \frac{1}{\delta(N)}$. This leads to a clipping range of $[\delta(N), \frac{1}{\delta(N)}]$, and in the limit: $\lim_{N\to\infty}$, it becomes: $[0, \infty]$. In other words, as more data is gathered, the effect of PRPO clipping eventually disappears, and the original objective is recovered. The exact choice of $\delta(N)$ determines how quickly this happens.

Gradient ascent with PRPO and possible extensions. Finally, we consider how the PRPO objective should be optimized. This turns out to be very straightforward when we look at its gradient. The clipping function f (Eq. 21) has a simpler gradient involving an indicator function on whether x is inside the bounded range:

$$\nabla_x f(x, \epsilon_-, \epsilon_+, r) = \mathbb{1}\left[(r > 0 \wedge x \leq \epsilon_+) \vee (r < 0 \wedge x \geq \epsilon_-)\right] r. \quad (24)$$

Applying the chain rule to the PRPO objective (Eq. 22) reveals:

$$\nabla_\pi \hat{U}_{\text{PRPO}}(\pi) = \sum_{q,d\in\mathcal{D}} \underbrace{\left[\nabla_\pi \frac{\omega(d|q)}{\omega_0(d|q)}\right]}_{\text{grad. for single doc.}} \underbrace{\nabla_\pi f\left(\frac{\omega(d|q)}{\omega_0(d|q)}, \epsilon_-, \epsilon_+, r(d|q)\right)}_{\text{indicator reward function}}.$$

Thus, the gradient of PRPO simply takes the importance weighted metric gradient per document, and multiplies it with the indicator function and reward. As a result, PRPO is simple to combine with existing LTR algorithms, especially ones that use policy-gradients [51], such as PL-Rank [25, 26] or StochasticRank [44]. For methods in the family of LambdaRank [3, 4, 48], it is a matter of replacing the $|\Delta DCG|$ term with an equivalent for the PRPO bounded metric.

Lastly, we note that whilst we introduced PRPO for DR estimation, it can be extended to virtually any relevance estimation by choosing a different r; e.g., one can easily adapt it for IPS [20, 30], or relevance estimates from a click model [7], etc. In this sense, we argue PRPO can be seen as a framework for robust safety in LTR.

6 Experimental Setup

For our experiments, we follow the semi-synthetic experimental setup that is prevalent in the CLTR literature [14, 28, 31, 45]. We make use of the three largest publicly available LTR datasets: Yahoo! Webscope [6], MSLR-WEB30k [33], and Istella [9]. The datasets consist of queries, a preselected list of documents per query, query-document feature vectors, and manually-graded relevance judgments for each query-document pair.

Following [14, 28, 45], we train a production ranker on a 3% fraction of the training queries and their corresponding relevance judgments. The goal is to simulate a real-world setting where a ranker trained on manual judgments is deployed in production and is used to collect click logs. The collected click logs can then be used for LTR. We assume the production ranker is safe, given that it would serve live traffic in a real-world setup.

We simulate a top-K ranking setup [29] where only $K = 5$ documents are displayed to the user for a given query, and any document beyond that gets zero exposure. To get the relevance probability, we apply the following transformation: $P(R = 1 \mid q, d) = 0.25 \cdot rel(q, d)$, where $rel(q, d) \in \{0, 1, 2, 3, 4\}$ is the relevance judgment for the given query-document pair. We generate clicks based on the trust bias click model (Assumption 3.2):

$$P(C = 1 \mid q, d, k) = \alpha_k P(R = 1 \mid q, d) + \beta_k. \quad (25)$$

The trust bias parameters are set based on the empirical observation in [2]: $\alpha = [0.35, 0.53, 0.55, 0.54, 0.52]$, and $\beta = [0.65, 0.26, 0.15, 0.11, 0.08]$. For CLTR training, we only use the training and validation clicks generated via the click simulation process (Eq. 25). To test the robustness of the safe CLTR methods in a setting where the click model assumptions do not hold, we simulate an *adversarial click model*, where the user clicks on the irrelevant document with a high probability and on a relevant document with a low click probability. We define the adversarial click model as:

$$P(C = 1 \mid q, d, k) = 1 - (\alpha_k P(R = 1 \mid q, d) + \beta_k). \quad (26)$$

Thereby, we simulate a maximally *adversarial* user who clicks on documents with a click probability that is inversely correlated with the assumed trust bias model (Assumption 3.2).

Further, we assume that the logging propensities have to be estimated. For the logging propensities ρ_0, and the logging metric weights (ω_0), we use a simple Monte-Carlo estimate [14]:

$$\hat{\rho}_0(d) = \frac{1}{N}\sum_{i=1:y_i\sim\pi_0}^{N} \alpha_{k_i(d)}, \quad \hat{\omega}_0(d) = \frac{1}{N}\sum_{i=1:y_i\sim\pi_0}^{N} \left(\alpha_{k_i(d)} + \beta_{k_i(d)}\right). \quad (27)$$

For the learned policies (π), we optimize Plackett-Luce (PL) ranking models [25] using the REINFORCE policy-gradient method [14, 53]. We perform clipping on the logging propensities (Eq. 5) only for the training clicks and not for the validation set. Following previous work, we set the clipping parameter to $10/\sqrt{N}$ [14, 30]. We do not apply the clipping operation for the logging metric weights (Eq. 8). To prevent overfitting, we apply early stopping based on the validation clicks. For variance reduction, we follow [14, 53] and use the average reward per query as a control-variate.

As our evaluation metric, we compute the NDCG@5 metric using the relevance judgments on the test split of each dataset [17]. Finally, the following methods are included in our comparisons: (i) *IPS*. The IPS estimator with affine correction [30, 45] for CLTR with trust bias (Eq. 6). (ii) *Doubly Robust*. The DR estimator for CLTR with trust bias (Eq. 10). This is the most important baseline for this work, given that the DR estimator is the state-of-the-art CLTR method [28]. (iii) *Safe DR*. Our proposed safe DR CLTR method (Eq. 18), which relies on the trust bias assumption (Assumption 3.2). (iv) *PRPO*. Our proposed *proximal ranking policy optimization* (PRPO) method for safe DR CLTR (Eq. 22). (v) *Skyline*. LTR method trained on the true relevance labels. Given that it is trained on the real relevance signal, the skyline performance is the upper bound on any CLTR methods performance.

7 Results and Discussion

Comparision with baseline methods. Fig. 2 presents the main results with different CLTR estimators with varying amounts of simulated click data. Amongst the baselines, we see that the DR

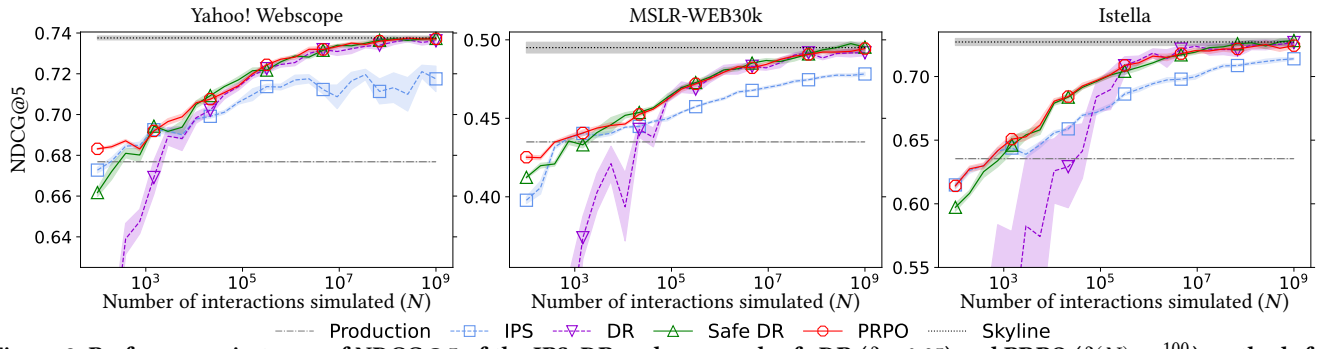

Figure 2: Performance in terms of NDCG@5 of the IPS, DR and proposed safe DR ($\delta = 0.95$) and PRPO ($\delta(N) = \frac{100}{N}$) methods for CLTR. The results are presented varying size of training data (N), with number of simulated queries varying from 10^2 to 10^9. Results are averaged over 10 runs; the shaded areas indicate 80% prediction intervals.

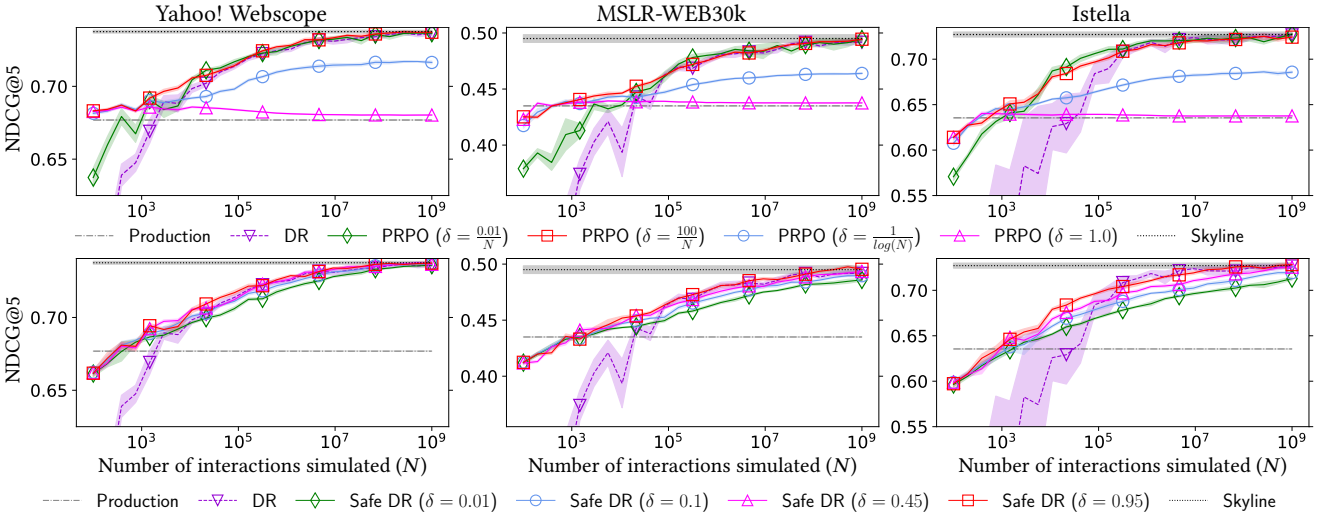

Figure 3: Performance of the safe DR and PRPO with varying safety parameter (δ). Top row: sensitivity analysis of PRPO with varying clipping parameter (δ) over varying dataset sizes N. Bottom row: sensitivity analysis for the safe DR method with varying safety confidence parameter (δ). Results are averaged over 10 runs; shaded areas indicate 80% prediction intervals.

estimator converges to the skyline much faster than the IPS estimator. The IPS estimator fails to reach the optimal performance even after training on 10^9 clicks, suggesting that it suffers from a high-variance problem. This aligns with the findings in [28]. As to safety, when the click data is limited ($N < 10^5$), the DR estimator performs much worse than the logging policy, i.e., it exhibits unsafe behavior, which can lead to a negative user experience if deployed online. A likely explanation is that when click data is limited, the regression estimates ($\hat{R}(d)$, Eq. 10) have high errors, resulting in a large performance degradation, compared to IPS.

Our proposed safety methods, safe DR and PRPO, reach the performance of the logging policy within ~500 queries on all datasets. For the safe DR method, we set the confidence parameter $\delta = 0.95$. For the PRPO method, we set $\delta(N) = \frac{100}{N}$. On the MSLR and the ISTELLA dataset, we see that PRPO reaches logging policy performance with almost 10^3 fewer queries than the DR method. Thus, our proposed methods, safe DR and PRPO, can be safely deployed, and avoid the initial period of bad performance of DR, whilst providing the same state-of-the-art performance at convergence.

Sensitivity analysis of the safety parameter. To understand the tradeoff between safety and utility, we performed a sensitivity analysis by varying the safety parameter (δ) for the safe DR method and PRPO. The top row of Fig. 3 shows us the performance of the PRPO method with different choices of the clipping parameter δ as a function of dataset size (N). We report results with the setting of the δ parameter, which results in different clipping widths. For the setting $\delta = \frac{0.01}{N}$ and $\delta = \frac{100}{N}$, the clipping range width grows linearly with the dataset size N. Hence, the resulting policy is safer at the start but converges to the DR estimator when N increases. With $\delta = \frac{0.01}{N}$, the clipping range is wider at the start. As a result, it is more unsafe than when $\delta = \frac{100}{N}$, which is the safest amongst all. For the case where the range grows logarithmically ($\delta = \frac{1}{\log(N)}$), the method is more conservative throughout, i.e., it is closer to the logging policy since the clipping window grows only logarithmically with N. For the extreme case where the clipping range is a constant ($\delta = 1$), PRPO avoids any change w.r.t. the logging policy, and as a result, it sticks closely to the logging policy.

The bottom row of Fig. 3 shows the performance of the safe DR method with varying confidence parameter values (δ). Due to the nature of the generalization bound (Eq. 18), the confidence parameter is restricted to: $0 \leq \delta \leq 1$. We vary the confidence parameters in the range $\delta \in \{0.01, 0.1, 0.45, 0.95\}$. We note that a

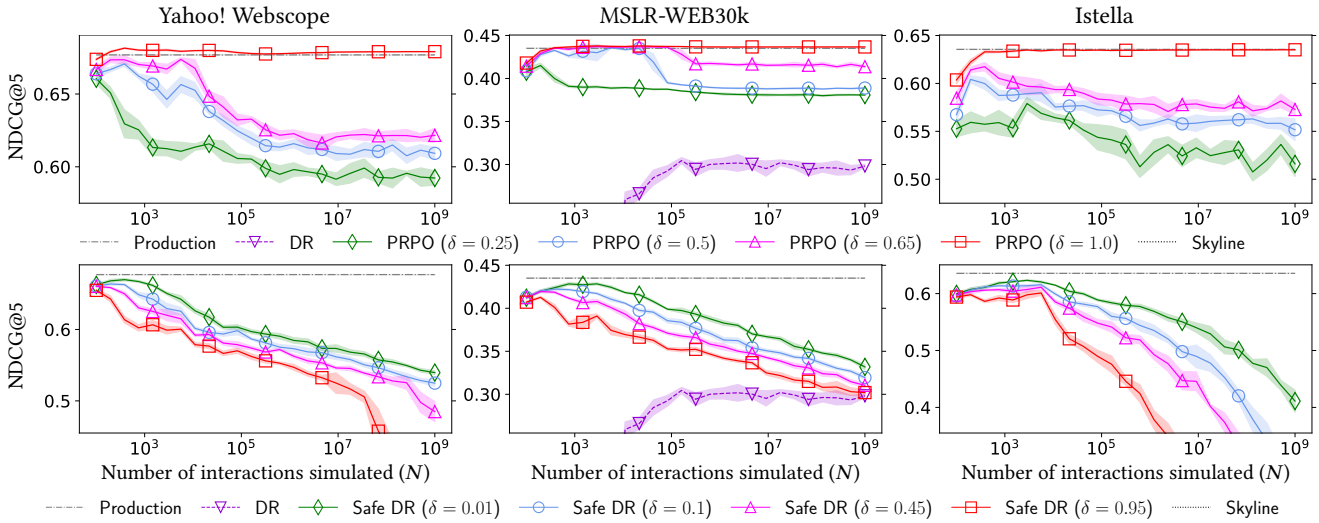

Figure 4: Performance of the proposed safe DR and PRPO with the adversarial click model. Top: sensitivity analysis results for the PRPO method with varying clipping parameter (δ). Bottom: sensitivity analysis for the safe DR method with varying safety confidence parameter (δ). Results are averaged over 10 independent runs; the shaded areas indicate 80% prediction intervals.

lower δ value results in higher safety, and vice-versa. Until $N < 10^5$, there is no noticeable difference in performance. For the Yahoo! Webscope dataset, almost all settings result in a similar performance. For the MSLR and ISTELLA datasets, when $N < 10^5$, a lower δ value results in a more conservative policy, i.e., a policy closer to the logging policy. However, the performance difference with different setups is less drastic than with the PRPO method. Thus, we note that the safe DR method is *less flexible* in comparison to PRPO.

Therefore, compared to our safe DR method, we conclude that our PRPO method provides practitioners with greater flexibility and control when deciding between safety and utility.

Robustness analysis using an adversarial click model. To verify our initial claim that our proposed PRPO method provides safety guarantees *unconditionally*, we report results with clicks simulated via the adversarial click model (Eq. 26). With the adversarial click setup, the initial user behavior assumptions (Assumption 3.2) *do not hold*. The top row of Fig. 4 shows the performance of the PRPO method with different safety parameters when applied to the data collected via the adversarial click model. We vary the δ parameter for PRPO in the range $\{0.25, 0.5, 0.65, 1.0\}$, e.g., $\delta = 0.5$ results in $\epsilon_- = 0.5$ and $\epsilon_+ = 2$. With the constant clipping range ($\delta = 1$), we notice that after ~400 queries, the PRPO methods performance never drops below the safe logging policy performance. For greater values of δ, there are drops in performance but they are all bounded. For the Yahoo! Webscope dataset, the maximum drop in the performance is ~12%; for the MSLR30K dataset, the maximum performance drop is ~10%; and finally, for the Istella dataset, the maximum drop is ~20%. Clearly, these observations show that PRPO provides robust safety guarantees, that are reliable even when user behavior assumptions are wrong.

In contrast, the generalization bound of our safe DR method (Theorem 4.1) holds only when the user behavior assumptions are true. This is not the case in the bottom row of Fig. 4, which shows the performance of the safe DR method under the adversarial click model. Even with the setting where the safety parameters have a high weight ($\delta = 0.01$), as the click data size increases, the performance drops drastically. Regardless of the exact choice of δ, the effect of the regularization of safe DR disappears as N grows, thus in this adversarial setting, it is only a matter of time before the performance of safe DR degrades dramatically.

8 Conclusion

In this paper, we have introduced the first safe CLTR method that uses state-of-the-art DR estimation and corrects trust bias. This is a significant extension of the existing safety method for CLTR that was restricted to position bias and IPS estimation. However, in spite of the importance of this extended safe CLTR approach, it heavily relies on user behavior assumptions. We argue that this means it only provides a *conditional* concept of safety, that may not apply to real-world settings. To address this limitation, we have made a second contribution: the *proximal ranking policy optimization* (PRPO) method. PRPO is the first LTR method that provides *unconditional* safety, that is applicable regardless of user behavior. It does so by removing incentives to stray too far away from a safe ranking policy. Our experimental results show that even in the extreme case of adversarial user behavior PRPO results in safe ranking behavior, unlike existing safe CLTR approaches.

PRPO easily works with existing LTR algorithms and relevance estimation techniques. We believe it provides a flexible and generic framework that enables practitioners to apply the state-of-the-art CLTR method with strong and robust safety guarantees. Future work may apply the proposed safety methods to exposure-based ranking fairness [25, 53] and to safe online LTR [30].

Acknowledgements

This research was supported by Huawei Finland, the Dutch Research Council (NWO), under project numbers VI.Veni.222.269, 024.004.022, NWA.1389.20.183, and KICH3.LTP.20.006, the EU's Horizon Europe program under grant No 101070212, and a SURF Cooperative using grant no. EINF-8200. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

A Appendix: Extended Safety Proof

Lemma A.1. *Under the trust bias click model (Assumption 3.2), and given the trust bias parameter α_k, β_k, the regression model estimates $\hat{R}_d$ and click indicator $c(d)$, the following holds:*

$$\mathrm{Cov}_{y,c}\big[c(d) - \beta_{k(d)}, \alpha_{k(d)}\hat{R}_d\big] \geq 0. \tag{28}$$

Proof. The covariance term can be rewritten as:

$$\begin{aligned}
&\mathrm{Cov}_{y,c}\big[c(d) - \beta_{k(d)}, \alpha_{k(d)}\hat{R}_d\big] \\
&= \mathbb{E}_{y,c}\big[(c(d) - \beta_{k(d)})\alpha_{k(d)}\hat{R}_d\big] - \mathbb{E}_{y,c}\big[c(d) - \beta_{k(d)}\big]\mathbb{E}_y\big[\alpha_{k(d)}\hat{R}_d\big] \\
&= \hat{R}_d\Big(\mathbb{E}_{y,c}\big[c(d)\alpha_{k(d)}\big] - \mathbb{E}_y\big[\beta_{k(d)}\alpha_{k(d)}\big] - R_d\ \rho_0(d)^2\Big),
\end{aligned} \tag{29}$$

where use $\rho_0(d) = \mathbb{E}_{y,c}\big[\alpha_{k(d)}\big]$ and $\mathbb{E}_{y,c}\big[(c_i(d) - \beta_{k_i(d)})/\rho_0(d)\big] = R_d$ [28]. Expanding the first expectation term in the expression:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{y,c}\big[c(d)\alpha_{k(d)}\big] = \sum_{y\in\pi_0} \pi_0(y)\alpha_{k(d)}P(C=1 \mid d, y) = \sum_{y\in\pi_0} \pi_0(y)\alpha_{k(d)} \\
&\cdot \Big(\alpha_{k(d)}R_d + \beta_{k(d)}\Big) = R_d\mathbb{E}_y\Big[\alpha^2_{k(d)}\Big] + \mathbb{E}_y\big[\alpha_{k(d)}\beta_{k(d)}\big],
\end{aligned} \tag{30}$$

where we substitute click model equation $P(C=1 \mid d, y)$ (Eq. 10). Substituting it back in Eq. 29, we get:

$$\begin{aligned}
&\mathrm{Cov}_{y,c}\big[c(d) - \beta_{k(d)}, \alpha_{k(d)}\hat{R}_d\big] = R_d\mathbb{E}_y\Big[\alpha^2_{k(d)}\Big] - R_d\ \mathbb{E}_y\big[\alpha_{k(d)}\big]^2 \\
&R_d\Big(\mathbb{E}_y\Big[\alpha^2_{k(d)}\Big] - \mathbb{E}_y\big[\alpha_{k(d)}\big]^2\Big) = R_d\mathrm{Var}_y[\alpha_k(d)] \geq 0.
\end{aligned}$$

□

A.1 Proof of Theorem 4.1

Proof. As per Cantelli's inequality [10], the following inequality must hold with probability $1-\delta$:

$$U(\pi) \geq \hat{U}_{\mathrm{DR}}(\pi) - \sqrt{\frac{1-\delta}{\delta}\mathrm{Var}_{q,y,c}\big[\hat{U}_{\mathrm{DR}}(\pi)\big]}. \tag{31}$$

Following a similar approach as previous works [14, 52], we look for an upper-bound on the variance of the DR estimator. From the definition of $\hat{U}_{\mathrm{DR}}(\pi)$ (Eq. 10), the variance of the DR estimator can be expressed as the variance of the second term:

$$\mathop{\mathrm{Var}}_{y,c}\big[\hat{U}_{\mathrm{DR}}(\pi)\big] = \frac{1}{N}\mathop{\mathrm{Var}}_{y,c}\left[\sum_{d\in D}\frac{\omega(d)}{\rho_0(d)}\big(c(d) - \alpha_{k(d)}\hat{R}(d) - \beta_{k(d)}\big)\right]. \tag{32}$$

Using Assumption 3.2 and assuming that document examinations are independent from each other [14], we rewrite further:

$$\begin{aligned}
N\cdot\mathop{\mathrm{Var}}_{y,c}\big[\hat{U}_{\mathrm{DR}}(\pi)\big] &= \sum_{d\in D_q}\mathop{\mathrm{Var}}_{y,c}\left[\frac{\omega(d)}{\rho_0(d)}\big(c(d) - \alpha_{k(d)}\hat{R}(d) - \beta_{k(d)}\big)\right] \\
&= \sum_{d\in D_q}\left(\frac{\omega(d)}{\rho_0(d)}\right)^2\mathop{\mathrm{Var}}_{y,c}\big[c(d) - \beta_{k(d)} - \alpha_{k(d)}\hat{R}(d)\big].
\end{aligned} \tag{33}$$

The total variance can be split into the following:

$$\begin{aligned}
&\mathrm{Var}_{y,c}\big[c(d) - \beta_{k(d)} - \alpha_{k(d)}\hat{R}_i(d)\big] = \mathrm{Var}_{y,c}\big[\alpha_{k(d)}\hat{R}(d)\big] \\
&\quad + \mathrm{Var}_{y,c}\big[c(d) - \beta_{k(d)}\big] - 2\mathrm{Cov}_{y,c}\big[c(d) - \beta_{k(d)}, \alpha_{k(d)}\hat{R}(d)\big].
\end{aligned} \tag{34}$$

Using Lemma A.1, we upper-bound the total variance term to:

$$\begin{aligned}
&\mathrm{Var}_{y,c}\big[c(d) - \beta_{k(d)} - \alpha_{k(d)}\hat{R}(d)\big] \\
&\leq \mathrm{Var}_{y,c}\big[\alpha_{k(d)}\hat{R}(d)\big] + \mathrm{Var}_{y,c}\big[c(d) - \beta_{k(d)}\big].
\end{aligned} \tag{35}$$

Next, we consider the two variance terms separately; with the variance of the first term following:

$$\mathop{\mathrm{Var}}_{y,c}\big[\alpha_{k(d)}\hat{R}(d)\big] = \mathop{\mathrm{Var}}_{y,c}\big[\alpha_{k(d)}\big]\hat{R}(d)^2 \leq \mathbb{E}_{y,c}\Big[\alpha^2_{k(d)}\Big] \leq \mathbb{E}_y\big[\alpha_{k(d)}\big].$$

where we make use of the fact that $\hat{R}^2_d \leq 1$, and $\alpha \in [0,1] \rightarrow \alpha^2_k \leq \alpha_k$. Next, we consider the second term:

$$\begin{aligned}
&\mathrm{Var}_{y,c}\big[c(d) - \beta_{k(d)}\big] \leq \mathbb{E}_{y,c}\Big[\big(c(d) - \beta_{k(d)}\big)^2\Big] \\
&\quad = \mathbb{E}_{y,c}\Big[c(d)^2 + \beta^2_{k(d)} - 2c(d)\beta_{k(d)}\Big] \leq \mathbb{E}_{y,c}[c(d)] + \mathbb{E}_y\big[\beta_{k(d)}\big],
\end{aligned} \tag{36}$$

since $c(d)^2 = c(d)$, $\beta^2_k \leq \beta_k$, and $\mathbb{E}_{y,c}\big[c(d)\beta_{k(d)}\big] \geq 0$. Substituting the click probabilities with Eq. 4, we get:

$$\begin{aligned}
&\mathbb{E}_{y,c}[c(d)] + \mathbb{E}_{y,c}[\beta_{k(d)}] = \mathbb{E}_{y,c}[\alpha_{k(d)}]P(R=1 \mid d) + 2\,\mathbb{E}_{y,c}[\beta_{k(d)}] \\
&\leq \mathbb{E}_y\big[\alpha_{k(d)}\big] + 2\,\mathbb{E}_y[\beta_{k(d)}],
\end{aligned} \tag{37}$$

where we use the fact that $P(R=1 \mid d) \leq 1$. Putting together the bounds on both parts of Eq. 35, we have:

$$\mathrm{Var}_{y,c}\big[c(d) - \beta_{k(d)} - \alpha_{k(d)}\hat{R}(d)\big] \leq 2\omega_0(d), \tag{38}$$

where $\omega_0(d) = \mathbb{E}_y\big[\alpha_{k(d)}\big] + \mathbb{E}_y\big[\beta_{k(d)}\big]$. Substituting the final variance upper bound in Eq. 33, we get:

$$\begin{aligned}
\mathop{\mathrm{Var}}_{y,c}\left[\sum_{d\in D}\frac{\omega(d)}{\rho_0(d)}\big(c(d) - \alpha_{k(d)}\hat{R}(d) - \beta_{k(d)}\big)\right] &\leq 2\sum_{d\in D_q}\left(\frac{\omega(d)}{\rho_0(d)}\right)^2\omega_0(d) \\
&= 2\sum_{d\in D_q}\left(\frac{\omega(d)}{\omega_0(d)}\right)^2\omega_0(d)\left(\frac{\omega_0(d)}{\rho_0(d)}\right)^2,
\end{aligned} \tag{39}$$

where we multiply and divide by $\omega_0(d)^2$ in the third step. Finally, we make use of the fact: $\frac{\omega_0(d)}{\rho_0(d)} \leq \max_{\pi_0}\frac{\omega_0(d)}{\rho_0(d)} \leq 1 + \max_k\frac{\beta_k}{\alpha_k}$, and put everything back together:

$$\begin{aligned}
N\cdot\mathrm{Var}_{y,c}\big[\hat{U}_{\mathrm{DR}}(\pi)\big] &\leq 2Z\left(1 + \max_k\frac{\beta_k}{\alpha_k}\right)^2\sum_{d\in D_q}\left(\frac{\omega'(d)}{\omega'_0(d)}\right)^2\omega'_0(d) \\
&= 2Z\left(1 + \max_k\frac{\beta_k}{\alpha_k}\right)^2 d_2(\omega \,\|\, \omega_0).
\end{aligned} \tag{40}$$

where $d_2(\omega \,\|\, \omega_0)$ is the Renyi divergence between the normalized expected exposure $\omega'(d)$ and $\omega'_0(d)$ (cf. Eq. 13). Substituting this into the upper-bound on variance in Eq. 31 completes the proof. □

A.2 Proof of Theorem 5.1

Proof. Given a logging policy ranking y_0, a user defined metric weight ω, and non-zero $r(d \mid q)$, for the choice of the clipping parameters $\epsilon_- = \epsilon_+ = 1$, the ranking $y^*(\epsilon_-, \epsilon_+)$ that maximizes the PRPO objective (Eq. 22) will be the same as the logging ranking y_0, i.e. $y^*(\epsilon_-, \epsilon_+) = y_0$. This is trivial to prove since any change in ranking can only lead in a decrease in the clipped ratio weights, and thus, a decrease in the PRPO objective. Therefore, $y^*(\epsilon_- = 1, \epsilon_+ = 1) = y_0$ when $\epsilon_- = \epsilon_+ = 1$. Accordingly: $|U(y_0) - U(y^*(\epsilon_- = 1, \epsilon_+ = 1))| = 0$ directly implies Eq. 23. This completes our proof. □

Whilst the above proof is performed through the extreme case where $\epsilon_- = \epsilon_+ = 1$ and the optimal ranking has the same utility as the logging policy ranking, other choices of ϵ_- and ϵ_+ bound the difference in utility to a lesser degree and allow for more deviation. As our experimental results show, the power of PRPO is that it gives practitioners direct control over this maximum deviation.

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*. 5–14.

[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*. 4–14.

[3] Christopher Burges, Robert Ragno, and Quoc Le. 2006. Learning to Rank with Nonsmooth Cost Functions. *Advances in Neural Information Processing Systems* 19 (2006).

[4] Christopher JC Burges. 2010. From RankNet to LambdaRank to LambdaMART. *Learning* 11, 23-581 (2010), 81.

[5] Carlos Castillo and Brian D. Davison. 2011. Adversarial Web Search. *Foundations and Trends in Information Retrieval* 4, 5 (2011), 377–486.

[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] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. *Click Models for Web Search*. Morgan & Claypool Publishers.

[8] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An Experimental Comparison of Click Position-bias Models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*. 87–94.

[9] Domenico Dato, Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, Nicola Tonellotto, and Rossano Venturini. 2016. Fast Ranking with Additive Ensembles of Oblivious and Non-Oblivious Regression Trees. *ACM Transactions on Information Systems (TOIS)* 35, 2 (2016), 1–31.

[10] Bhaskar Kumar Ghosh. 2002. Probability Inequalities Related to Markov's Theorem. *The American Statistician* 56, 3 (2002), 186–190.

[11] Shashank Gupta, Philipp Hager, Jin Huang, Ali Vardasbi, and Harrie Oosterhuis. 2024. Unbiased Learning to Rank: On Recent Advances and Practical Applications. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 1118–1121.

[12] Shashank Gupta, Olivier Jeunen, Harrie Oosterhuis, and Maarten de Rijke. 2024. Optimal Baseline Corrections for Off-Policy Contextual Bandits. *arXiv preprint arXiv:2405.05736* (2024).

[13] Shashank Gupta, Harrie Oosterhuis, and Maarten de Rijke. 2023. A Deep Generative Recommendation Method for Unbiased Learning from Implicit Feedback. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*. 87–93.

[14] Shashank Gupta, Harrie Oosterhuis, and Maarten de Rijke. 2023. Safe Deployment for Counterfactual Learning to Rank with Exposure-Based Risk Minimization. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[15] Shashank Gupta, Harrie Oosterhuis, and Maarten de Rijke. 2024. A First Look at Selection Bias in Preference Elicitation for Recommendation. *arXiv preprint arXiv:2405.00554* (2024).

[16] Rolf Jagerman, Ilya Markov, and Maarten de Rijke. 2020. Safe Exploration for Optimizing Contextual Bandits. *ACM Transactions on Information Systems (TOIS)* 38, 3 (2020), 1–23.

[17] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[18] Thorsten Joachims. 2002. Optimizing Search Engines Using Clickthrough Data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 133–142.

[19] Thorsten Joachims and Adith Swaminathan. 2016. Counterfactual Evaluation and Learning for Search, Recommendation and Ad Placement. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 1199–1201.

[20] 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.

[21] Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. 2019. Neural Trust Region/Proximal Policy Optimization Attains Globally Optimal Policy. *Advances in neural information processing systems* 32 (2019).

[22] Tie-Yan Liu. 2009. Learning to Rank for Information Retrieval. *Foundations and Trends in Information Retrieval* 3, 3 (2009), 225–331.

[23] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. 2023. Black-box Adversarial Attacks against Dense Retrieval Models: A Multi-view Contrastive Learning Method. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 1647–1656.

[24] Harrie Oosterhuis. 2020. *Learning from User Interactions with Rankings: A Unification of the Field*. Ph. D. Dissertation. Informatics Institute, University of Amsterdam.

[25] Harrie Oosterhuis. 2021. Computationally Efficient Optimization of Plackett-Luce Ranking Models for Relevance and Fairness. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1023–1032.

[26] Harrie Oosterhuis. 2022. Learning-to-Rank at the Speed of Sampling: Plackett-Luce Gradient Estimation With Minimal Computational Complexity. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2266–2271.

[27] Harrie Oosterhuis. 2022. Reaching the End of Unbiasedness: Uncovering Implicit Limitations of Click-Based Learning to Rank. In *Proceedings of the 2022 ACM SIGIR International Conference on the Theory of Information Retrieval*. ACM.

[28] Harrie Oosterhuis. 2023. Doubly Robust Estimation for Correcting Position Bias in Click Feedback for Unbiased Learning to Rank. *ACM Transactions on Information Systems* 41, 3 (2023), 1–33.

[29] 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*. 489–498.

[30] 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*. 463–471.

[31] Harrie Oosterhuis and Maarten de de Rijke. 2021. Robust Generalization and Safe Query-Specialization in Counterfactual Learning to Rank. In *Proceedings of the Web Conference 2021*. 158–170.

[32] 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*. 1863–1873.

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

[34] Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. 2010. LETOR: A Benchmark Collection for Research on Learning to Rank for Information Retrieval. *Information Retrieval* 13, 4 (2010), 346–374.

[35] James Queeney, Yannis Paschalidis, and Christos G Cassandras. 2021. Generalized Proximal Policy Optimization with Sample Reuse. In *Advances in Neural Information Processing Systems*, Vol. 34. 11909–11919.

[36] Filip Radlinski. 2007. Addressing Malicious Noise in Clickthrough Data. In *LR4IR 2007: Learning to Rank for Information Retrieval Workshop at SIGIR*, Vol. 2007.

[37] Alfréd Rényi. 1961. On Measures of Entropy and Information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, Vol. 1. Berkeley, California, USA.

[38] Yuta Saito and Thorsten Joachims. 2021. Counterfactual Learning and Evaluation for Recommender Systems: Foundations, Implementations, and Recent Advances. In *Fifteenth ACM Conference on Recommender Systems*. 828–830.

[39] Mark Sanderson. 2010. Test Collection based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval* 4, 4 (2010), 247–375.

[40] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2015. High-dimensional Continuous Control using Generalized Advantage Estimation. *arXiv preprint arXiv:1506.02438* (2015).

[41] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[42] Adith Swaminathan and Thorsten Joachims. 2015. Batch Learning from Logged Bandit Feedback through Counterfactual Risk Minimization. *The Journal of Machine Learning Research* 16, 1 (2015), 1731–1755.

[43] Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. 2015. High-Confidence Off-Policy Evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[44] Aleksei Ustimenko and Liudmila Prokhorenkova. 2020. StochasticRank: Global Optimization of Scale-Free Discrete Functions. In *International Conference on Machine Learning*. PMLR, 9669–9679.

[45] 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*. 1475–1484.

[46] 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, 115–124.

[47] Xuanhui Wang, Nadav Golbandi, Michael Bendersky, Donald Metzler, and Marc Najork. 2018. Position Bias Estimation for Unbiased Learning to Rank in Personal Search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 610–618.

[48] Xuanhui Wang, Cheng Li, Nadav Golbandi, Michael Bendersky, and Marc Najork. 2018. The LambdaLoss Framework for Ranking Metric Optimization. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 1313–1322.

[49] Yuhui Wang, Hao He, and Xiaoyang Tan. 2020. Truly Proximal Policy Optimization. In *Uncertainty in Artificial Intelligence*. PMLR, 113–122.

[50] Yuhui Wang, Hao He, Xiaoyang Tan, and Yaozhong Gan. 2019. Trust Region-guided Proximal Policy Optimization. In *Advances in Neural Information Processing Systems*, Vol. 32.

[51] Ronald J Williams. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine learning* 8, 3 (1992), 229–256.
[52] Hang Wu and May Wang. 2018. Variance Regularized Counterfactual Risk Minimization via Variational Divergence Minimization. In *International Conference on Machine Learning*. PMLR, 5353–5362.
[53] Himank Yadav, Zhengxiao Du, and Thorsten Joachims. 2021. Policy-Gradient Training of Fair and Unbiased Ranking Functions. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1044–1053.