Counterfactual Learning to Rank via Knowledge Distillation

Ehsan Ebrahimzadeh[1], Alex Cozzi[1] and Abraham Bagherjeiran[1,*,†]

[1] *eBay Search Ranking and Monetization*

Abstract

Knowledge distillation is a transfer learning technique to improve the performance of a student model trained on a *Distilled Empirical Risk*, formed by a label distribution defined by some teacher model, which is typically trained on the *same task* and belongs to a hypothesis class with richer *representational capacity*. In this work, we study knowledge distillation in the context of counterfactual Learning To Rank(LTR) from implicit user feedback. We consider a generic partial information search ranking scenario, where the relevancy of the items in the logged search context is observed only in the event of an explicit user engagement. The key idea behind using knowledge distillation in this *counterfactual setup* is to leverage the teacher's distilled knowledge in the form of soft predicted relevance labels to help the student with more effective list-wise comparisons, variance reduction, and improved generalization behavior. We build empirical risk estimates that rely not only on the de-biased observed user feedback via standard *Inverse Propensity Weighting*, but also on the teacher's distilled knowledge via *potential outcome modeling*. We analyze the generalization performance of the proposed empirical risk estimators from a theoretical perspective by establishing bounds on their *estimation error*. We also conduct rigorous counterfactual offline evaluations as well as online controlled randomized experiments for a product search ranking task in a major E-commerce platform. The primary distinction of the proposed distillation-based perspective in contrast to the standard counterfactual inference based on potential outcome modeling is that we leverage teachers trained on different, yet related, tasks to improve the generalization power of a student model for a ranking task. Specifically, we report strong empirical results showing that the distilled knowledge from a teacher trained on *expert judgments* can significantly improve the generalization performance of a student ranker. We also show how *explanatory click models*, trained for a click prediction task with *privileged* encoding of the presentation context in observational data, can *explain away* the effect of presentation-related confounding for the LTR model.

Keywords

Learning to Rank, Counterfactual Inference, Knowledge Distillation, Potential Outcome Modeling, Doubly Robust Estimation

1. Introduction

1.1. Motivation

Learning To Rank(LTR) from implicit user feedback is the predominant approach in large scale information retrieval systems, where search context information and user engagement events are constantly logged and are available at desirable scale, granularity and recency. User engagement events on the Search Engine Result Pages(SERP) are driven not only by the underlying search

eCom'24: ACM SIGIR Workshop on eCommerce, July 18, 2024, Washington, DC, USA

eebrahimzadeh@ebay.com (E. Ebrahimzadeh); acozzi@ebay.com (A. Cozzi); abagherjeiran@ebay.com (A. Bagherjeiran)

0000-0002-0877-7063 (E. Ebrahimzadeh); 0000-0002-0877-7063 (A. Cozzi); 0000-0002-0877-7063 (A. Bagherjeiran)

intent, but also through the *presentation context*, governed by the slotting and organizational policies of the search engine. Specifically, in a list-wise presentation of the search results, users are more likely to engage with higher ranked items, due to their inherent sequential browsing of the page. This signifies the so called position bias in SERP-level user engagement events, which in turn implies that it is more likely to observe the relevance of the top ranked items in the logged SERPs than the items ranked in lower slots.

The standard approach to account for the inherent presentation biases in the observational search activity data is to adopt the propensity matching-based counterfactual inference framework [1]. The key idea behind this perspective is to hypothesize an explanatory click model based on *causal constructs* that govern users' browsing behavior[2] and estimate a notion of observation propensity as the probability that the user examines the relevance of an item in a given search context. The estimated propensity scores are then used to build a de-biased empirical risk estimate using a technique called Inverse Propensity Weighting(IPW). Note that, unlike the standard causal inference settings for estimating average treatment effect estimation, the examination event, which signifies the treatment in this formulation, is not fully observable and we need to make strong assumptions to build reliable propnesity score estimates.

On the other hand, predicting the counterfactual probability of click in the absence of presentation context confounding is a hard problem. While there is a rich literature on developing *predictive and descriptive click models* based on distributed representations of the search context, e.g. [3], or de-confounding click models that disentangle relevance and bias modeling [4] or causal probabilistic graphical models to explain users' browsing behavior, e.g. [2], the overwhelming body of research on unbiased LTR rely on slight variations of the IPW technique based on the so-called position-based examination model[5].

1.2. Contributions

Building reliable causal click models from observational data is challenging, particularly in the presence of heterogeneity in search context representation and user activity data. This is because for a predictive model to be contextually discriminative and well-calibrated, one needs complex models with rich representations, while LTR models trained on a listwise objectives can achieve similar ranking performance with cheaper representations. Despite recent attempts for building causal click models with richer encoding of the search context for disentangling the relevance from presentation-related confounding from observational data[4, 6], there is still fundamental theoretical gap to incorporate an arbitrarily complex click model in a counterfactual learning to rank setting with rank-discounted Information Retrieval(IR) objectives.

- In this work, we address this theoretical gap by adopting knowledge distillation in a counterfactual LTR framework, where a relevance teacher trained on a relevance prediction task is used as a potential outcome model that helps improve the generalization behavior of the student model for the LTR task.

Specifically, we study multiple empirical LTR risk estimates that are built via the *distilled* knowledge from a teacher model as well as the de-biased observed user feedback through IPW. To improve the generalization behavior of the vanilla IPW-based and distilled empirical risk estimates, we adopt an array of diverse approaches, including the doubly robust technique[7], to build estimators that are robust to inaccuracies of either the teacher model or the propensity model. The doubly robust estimator for LTR is also developed in the excellent concurrent work[8].

- We present a theoretical analysis of the *estimation error* of the proposed estimators by characterizing a bias-variance trade-off, which shows that well-calibrated teachers that approximate the Bayes relevance classifier generalize better.

By introducing knowledge distillation in the partial information LTR setting, we contribute to the growing literature on understanding the underlying statistical benefits of knowledge distillation. Our approach to employing knowledge distillation in counterfactual LTR is, however, different from the standard application of the technique; in that the relevance teacher is trained on a predictive/explanatory task, while the student uses the de-biased relevance probability estimates of the teacher as soft training labels for the ranking task. Essentially, through teacher's distilled knowledge on un-engaged items, the student can learn from more informative list-wise comparisons among all pairs of training samples in a query context rather than contrasting only the engaged items with the un-engaged ones, as in IPW-based techniques.

While our generalization analysis is generic, we discuss specific techniques for developing effective teacher models. In particular, we first discuss how explanatory click models that encode the page context confounding independently from relevance covariates help explain away the confounding presentation context in the logged search session[9, 4] and propose simple de-confounding techniques to use observational data for training the relevance teacher. This is very similar in nature to the IPW-based technique discussed in [8]. Beyond counterfactual click prediction teachers, we also introduce relevance teachers trained on a different, yet highly related, task.

- We report strong empirical results on a ranking task in a major e-commerce platform showing that a teacher model trained on expert relevance judgments can significantly improve the generalization power of the student model.

The significance of this result is multi fold: first, it offers a new perspective on knowledge distillation as a technique to incorporate heterogeneous sources of search utility to the user in training the ranker. Second, it offers an effective way to incorporate a broader set of queries, which may not have rich historic user engagements in training the ranker, thus alleviating the inherent selection bias of the rankers relying only on search s contexts with rich historic engagements.

1.3. Summary of Prior Art

Propensity Matching The standard prevalent technique in counterfactual LTR is the Inverse Propensity Weighting(IPW)[1], where the observed biased clicks are appropriately weighted in the empirical estimate based on estimated probabilities for observing the item's relevancy in a given search context. In [10], it is shown that the IPW estimator fails to account for trust bias[11], which is the gap between perceived relevance and true relevance that is often modeled as a position-dependent confounding, and new estimators based on affine corrections are proposed, which is adopted in multiple subsequent works[12, 13].

Explanatory Click Models The simplest, yet most popular, explanatory model is the examination-based click model[14, 5], defined by operationalizing the causal construct of observation as a latent *Examination* random variable. The examination-based models posit that an item is clicked if it is examined by the user and in addition is relevant to the user's intent. It is often further assumed that the examination event depends on the search context only through the position of the item in the SERP. Examination-based models are enriched in various ways to incorporate more granular presentation biases given the search context Most notably, [2] proposes a Dynamic Bayesian Network model (DBN) where new hidden variables are introduced to explain clicks and search abandonment events. More recently, two tower models for click prediction[15] with independent relevance and observation towers have gained popularity. The relevance-observation factorization assumption is revisited in [6] and deconfounding methods are proposed in [4, 16] to disentangle relevance from observation.

Propensity Estimation Standard click propensity models are often only based on the position bias [14] and there is a variety of techniques to estimate the parameters of such models based either on online interventions, through randomization and interleaving[17], or likelihood optimization from logged data[17, 18]. Intent aware examination models are proposed in [19, 20, 18], where different click propensities are learned for different intent classes. A dual learning algorithm is proposed in [21] to solve the problems of unbiased learning to rank and unbiased propensity estimation simultaneously.

Potential Outcome Modeling The idea of using potential outcome models for counterfactual inference is popular in the context of contextual bandits and recommendations systems where predictive models for the reward are developed instead of or in conjunction with inverse propensity weighting scheme to build empirical reward estimates[22, 23, 24]. There are a number of recent works in the context of unbiased response prediction that leverage and analyze the doubly robust technique[25, 12, 26, 8].

Counterfactual Evaluation Click models are widely used for counterfactual evaluation of IR metrics, either as data-driven rank-discount functions in *utility-based* metrics or as estimates to the user satisfaction gain in *effort-based* metrics[27]. The bias of the vanilla Average-Over-All evaluation metrics is studied in [28] and propensity-based correction is proposed for debiasing. Using predicted relevance probabilities as training labels has in fact been considered in [2, 19], but it is only limited to settings where the adopted measure of ranking loss only depends on the relative order of the labels than the actual values.

Knowledge Distillation There has been a growing interest in understanding the statistical underpinnings of knowledge distillation to explain its empirical success[29, 30, 31]. Most notably, the closest relevant work to our study is [29], where distillation is studied from a statistical learning theory perspective by establishing a bias-variance trade-off for the student based on the quality of the teacher in estimating the true Bayes class probabilities in a multi-label classification and retrieval. More recently, [32] extends knowledge distillation to Top K ranking problem, where ranking at the top is preserved by matching the order of items of student and teacher, while penalizing items ranked low by the teacher. In [33] multiple distillation techniques are proposed to improve the generalization power of the trained recommender model using data from a random logging policy. A similar uniform data augmentation technique is used in [34] to guard against feature distribution shifts. A recent work that offers a similar perspective on knowledge distillation in LTR is [35], where the teacher has access to some privileged features[36], that the student does not have access to. The other closely related works are the ranking distillation setting in [37, 38], where the teacher model is trained on the same ranking task. The main distinction between this study and prior adoptions of knowledge distillation for the ranking tasks is that we employ teacher models that are trained on a predictive task and are meant to provide predicted relevance probability estimates to the student.

1.4. Notation

Here is a list of notations adopted throughout the paper. Sets and ordered sets, also referred to as lists, are represented with upper-case calligraphic symbols; such as $\mathcal{D}$. Random quantities are shown in bold such as $\mathbf{d}$ with realization d. The expected value of random variable $\mathbf{c}$ is denoted by $\mathbb{E}[\mathbf{c}]$ and the conditional expectation of a random variable $\mathbf{c} = f(\mathbf{r}, \mathbf{o})$ given $\mathbf{r}$ is denoted by $\mathbb{E}[\mathbf{c}|\mathbf{r}]$. For a measurable event $\mathcal{A}$, we denote by $\mathbf{c}_{\mathcal{A}}$ a random variable defined in the measurable space equipped with a regular conditional probability measure $\mathbb{P}(.|\mathcal{A})$ with respect to the conditioning sigma algebra $\sigma(\mathcal{A})$. For a function $f : \mathbb{D} \to \mathbb{R}$, the $|\mathcal{D}|$ dimensional array $[f(d)]_{d\in\mathcal{D}}$ is denoted by $f(\mathcal{D})$.

2. Background and Problem Setup

2.1. Empirical Risk Minimization for LTR

We consider a generic supervised setup for learning to rank from implicit user feedback in the logged search activity data. A logged search $s \in \mathcal{S}$ is characterized by the query q, and the ordered collection of items $\mathcal{D}_s \in \mathcal{D}^N$ slotted by the search engine's ranking policy at the logging time, as well as the click events attributed to the items on the SERP, denoted by $c(\mathcal{D}_s) \in \{0, 1\}^N$. It is standard to assume a preference model with a context-item desirability distribution $\mathbb{P}_{\mathbf{r}_d|\mathbf{s}}(r_d|s)$ that produce oracle ground truth relevance labels r_d. We adopt a simple explanatory model on the sequential browsing and examination behavior of the user that explains a click event on an item d in the search context s, $\mathbf{c}_d = 1$, based on whether the item on the SERP is examined by the user, $\mathbf{o}_d = 1$ and that it matches the user's underlying search intent, $\mathbf{r}_d = 1$; asserting

$$\mathbb{P}(\mathbf{c}_d = 1|s) = \mathbb{P}(\mathbf{o}_d = 1|s)\mathbb{P}(\mathbf{r}_d = 1|s). \tag{1}$$

The conditional probability $\mathbb{P}(\mathbf{o}_d = 1|s)$ is referred to as the propensity of observing the relevancy of the item d in the search context s. We further simplify the adopted examination model by assuming that the click propensities depend on the presentation context only through the position of the slotted item on the SERP by the ranking policy; that is

$$\mathbb{P}(\mathbf{o}_d = 1|s) = \mathbb{P}(\mathbf{o}_d = 1|\pi(d)), \tag{2}$$

where $\pi(d)$ is the rank attributed to item d by the ranking policy $\pi(\cdot)$ served in the search context s. There is a large body of literature on estimating position-based user click propensities either through online interventions or offline estimation techniques based on latent variable models from logged search data. The position-based observation probabilities $\mathbb{P}(\mathbf{o} = 1|\mathsf{rank})$ define a rank-discount function $\ell(\mathsf{rank})$ that explains the effect of the presentation context on the user engagement in the search context primarily based on the position of the slotted items, which one can approximate via fitting a uni-variate model on the estimated propensities as a function of the ranking slot.

Given the adopted user behavior model and logged search activity data, our objective is to train a ranking policy by minimizing the statistical risk, corresponding to a search efficiency metric function, which approximates the expected number of user behavior events. Specifically, we define the statistical risk for a deterministic ranking policy $\pi_f(\cdot)$, corresponding to a scoring function $f(d; s)$ that produces a score for each individual document $d \in \mathscr{D}_s$ given the search context s, as

$$R(\pi_f) = \mathbb{E}[\mathscr{L}(\pi_f(\mathscr{D}_s), r(\mathscr{D}_s))], \tag{3}$$

where $\mathscr{L}(\pi_f(\mathscr{D}_s), r(\mathscr{D}_s)) = -\ell(\pi_f(\mathscr{D}_s))^T r(\mathscr{D}_s)$ is defined based on an estimate of the expected number of clicks in the search context s via a sequential browsing examination model in the form of Discounted Cumulative Gain(DCG) from the ranked list $\pi_f(\mathscr{D}_s)$ produced by the policy with respect to oracle labels $r(\mathscr{D}_s)$ and the rank-discount function $\ell(\cdot)$. We note that DCG-type information retrieval metrics are often defined as measures of gain, while with a simple sign tweak, we consider them as measures of loss for risk minization. While our discussions of the search efficiency-based loss function are primarily from the perspective of a likelihood model on the data, we note that our framework generalize to any standard ranking loss function that is linear as a function of relevance labels.

Since the underlying joint distribution of the search contexts and relevance labels is not known to the learner, the standard approach is to build an empirical risk estimate

$$\hat{R}(\pi_f) = \frac{1}{|\mathcal{S}^\tau|} \sum_{s \in \mathcal{S}^\tau} \mathscr{L}(\pi_f(\mathscr{D}_s), \hat{r}(\mathscr{D}_s)), \tag{4}$$

based on a sampled set $\mathcal{S}^\tau$ of search contexts with suitably defined empirical labels $\hat{r}(\mathscr{D}_s)$ to approximate the ground truth label distribution $\mathbb{P}_{\mathbf{r}_d|\mathbf{s}}(r_d|s)$. In order to characterize key statistical properties of an empirical risk estimate, we establish bounds on moments of the *estimation error*, which is the divergence of the empirical risk estimates from the Bayes statistical risk $\hat{R}(\pi_f) - R(\pi_f)$, based on the divergence of the empirical relevance labels from the underlying relevance probabilities.

Lemma 2.1. *For any ranking policy $\pi : \mathscr{D}^N \to [N]$ and non-negative integer n,*

$$\mathbb{E}\left[\left(\hat{R}(\pi) - R(\pi)\right)^n\right] \leq C_n \mathbb{E}\left[\left\|\mathbb{E}\left[(\hat{r}(\mathscr{D}_\mathbf{s}) - \mathbf{r}(\mathscr{D}_\mathbf{s}))^n \,|\mathbf{s}\right]\right\|_2\right],$$

where $C_n = \|[\ell^n(i)]_{i=1}^N\|_2$.

The proof follows by invoking the towering rule and a straightforward application of Cauchy-Schwarz inequality. Generalization bounds for empirical risk estimates can be obtained by invoking concentration techniques like Bernstein's inequality, which characterize the generalization behavior of a policy based on the bias and variance of the empirical risk estimator.

There is a remarkable body of work around developing efficient algorithms for Empirical Risk Minimization(ERM) on DCG-based loss functions over a variety of different hypothesis classes. In this study, we are oblivious to the choice of the hypothesis class and the particular optimization techniques adopted for training and focus on the generalization power of the estimators as it relates to the estimation error, oblivious to *approximation and optimization errors.*

2.2. Counterfactual LTR: Debiasing via Inverse Propensity Weighting

Given the examination model (1) on the user browsing behavior, we assume that we have an oracle algorithm to estimate the propensity to engage with item d in the search context s. Once click propensities, $\hat{\mathbb{P}}(\mathbf{o}_d = 1|s)$, are estimated, an empirical risk estimate can be built using de-biased labels formed via the Inverse Propensity Weighting(IPW) technique by simply setting

$$\hat{r}^{\mathsf{IPW}}(d) = \frac{c_d}{\hat{\mathbb{P}}(\mathbf{o}_d = 1|s)}, \tag{5}$$

We note that the underlying assumptions on the user behavior data for the propensity estimation task can in general be different/richer than the position-based model (2), adopted to define the ranking efficiency task. While our discussion can be applied to arbitrarily complex propensity estimation techniques, to simplify the presentation, we assume that the propensities are estimated based on a vanilla position bias model; that is $\hat{\mathbb{P}}(\mathbf{o}_d = 1|s) = \hat{p}_{\pi_0(d)}$, where $\pi_0(d)$ is the rank of the item in logged data $\mathscr{D}_s$. Owing to the linearity of the loss function $\mathscr{L}$ as function of relevance labels, it is straightforward to show that the IPW empirical risk estimate, denoted as

$$\hat{R}^{\mathsf{IPW}}_{\hat{p}}(\pi) = \frac{1}{|\mathcal{S}^\tau|} \sum_{s,c(\mathscr{D}_s)\in\mathcal{S}^\tau} \mathscr{L}(\pi(\mathscr{D}_s), \hat{r}^{\mathsf{IPW}}(\mathscr{D}_s)), \tag{6}$$

is an unbiased estimator of the Bayes statistical risk given the examination model, if the estimated propensities $\hat{p}$ are unbiased. The IPW estimator is known to have high variance, especially when the estimated propensity values are small, and there is a significant body of literature around clipping and normalization techniques to reduce the variance of this estimator. In this work, to improve the generalization of the LTR estimators, we introduce a number of techniques based on potential outcome modeling, and characterize their bias-variance trade-offs.

2.3. Knowledge Distillation

Knowledge distillation is a transfer learning technique where a student model is trained on a distilled empirical risk built based on a label distribution provided by a teacher model. Given a teacher model $T : \mathscr{D} \times \mathcal{S} \to \mathbb{R}$ that produces an estimated score for the relevance probability of each individual document $d \in \mathscr{D}_s$ given the search context s, a generic distilled empirical estimate for the risk of the student can be defined as

$$\hat{R}^{\mathsf{KD}}_{T}(\pi) = \frac{1}{|\mathcal{S}^\tau|} \sum_{s\in\mathcal{S}^\tau} \mathscr{L}(\pi(\mathscr{D}_s), T(\mathscr{D}_s; s)). \tag{7}$$

As shown in the next section, distilled empirical risk estimates has desirable variance-reduction properties compared to the vanilla empirical risk estimates based on observed feedback and inverse propensity weighting. As such, a straightforward, yet fundamental, observation to make about the variance reduction properties of distilled risk estimator is that

Remark 2.2. The variance of the distilled risk estimator built based on the true Bayes relevance probabilities is no greater than the variance of the vanilla empirical risk estimate formed by realizations from the relevance distribution.

The proof is trivial and can be found in [29, 23]. The improved variance in this estimator often comes at the cost of some bias, which leads us to explore hybrid estimators that take advantage of the observed user feedback as well as the teacher's distilled knowledge.

3. Debiasing Empirical Risk with Relevance Distillation

In this section, building upon the ideas discussed in the previous section, we propose empirical risk estimates that leverage the observed user feedback as well as the distilled knowledge from a suitably trained teacher. The core idea is to use the observed user feedback to improve the bias and at the same time use soft labels provided by the teacher to improve the variance of the estimator. We first discuss our proposed distilled empirical risk estimators oblivious to the choice of the teacher model and characterize their generalization behavior by developing bounds on their bias and variance and then discuss a number of techniques to develop effective teacher models.

3.1. Distilled Empirical Risk Estimates

3.1.1. Hybrid Distilled Risk Estimators

The main idea for building hybrid distilled risk estimators is to define a simple trade-off for bias/variance behavior of the estimator by a convex combination of the IPW and distilled empirical risk estimates. Formally, for a teacher $T : \mathscr{D} \times \mathscr{S} \to \mathbb{R}$, some estimate $\hat{p}$ of the click propensities, and any $0 \leq \alpha \leq 1$, a hybrid distilled estimator of the risk $\hat{R}^{\mathsf{HD}}_{T,\hat{p},\alpha}$ can be obtained by setting

$$\hat{r}^{\mathsf{HD}}_{T,\hat{p},\alpha}(d) = \alpha \frac{c_d}{\hat{p}_{\pi_0(d)}} + (1-\alpha)\, T(d; s) \tag{8}$$

in (4). Next, we analyze the bias/variance of this hybrid estimator. Note that the bias/variance analysis of the IPW and vanilla distilled emprirical risk estimates can be derived as corollary by setting α to 1 and 0, respectively. For a given search context s, and a document $d \in \mathscr{D}_s$ at rank $\pi_0(d)$, let $p^*_{r_d}$ denote the actual relevance probability $\mathbb{P}(\mathbf{r}_d = 1|s)$ and $p^*_{o_d}$ denote the actual propensity scores $\mathbb{P}(\mathbf{o}_d = 1|s)$. Let us define the deviation of the estimated relevance probabilities and the estimated propensity scores from the actual quantities as $\Delta_d = T(d; s) - p^*_{r_d}$, and $\delta_d = \frac{p^*_{o_d}}{\hat{p}_{\pi_0(d)}} - 1$, respectively.

Theorem 3.1. *For any student policy π,*

$$\mathbb{E}[\hat{R}^{\mathsf{HD}}(\pi) - R(\pi)] \leq C_1 \mathbb{E}\left[\left\|\left[\ \alpha p^*_{r_{\mathbf{d}}} \delta_{\mathbf{d}} + (1-\alpha)\Delta_{\mathbf{d}}\right]_{d \in \mathscr{D}_s}\right\|_2\right],$$

where C_1 is the same constant as in Lemma 2.1.

The proof is straightforward and is followed by characterizing the bias of the empirical labels and invoking Lemma 2.1. Please refer to the Appendix for a complete proof. Theorem 3.1 provides insight on how to control the bias of the estimator by adjusting the parameter α by trading off the bias terms due to the IPW and distillation-based components. The variance analysis in the Appendix shows that the variance due to the IPW component is sensitive to small propensity score values and the propensity estimation errors. In this case, the variance reduction that we achieve by including the distillation-based component helps with better generalization of the hybrid estimator.

3.1.2. Doubly Robust Risk Estimators

Although hybrid empirical risk estimates provide a simple and flexible way to control the bias-variance trade-off, they can incur significant estimation error if either of the risk components are inaccurate. To alleviate this, we adopt the doubly robust technique[23] to make the empirical risk estimator more robust when either of the risk components are accurate. Formally, for a suitable operationalization assumption for the observation event $\mathbf{o}_d$, a teacher $T : \mathscr{D} \times \mathscr{S} \to \mathbb{R}$ and some estimate $\hat{p}$ of click propensities, a doubly robust estimator of the risk $\hat{R}^{\mathsf{DR}}_{T,\hat{p}}$ can be obtained by setting

$$\hat{r}^{\mathsf{DR}}_{T,\hat{p}}(d) = T(d;s) + \frac{o_d}{\hat{p}_{\pi_0(d)}}(c_d - T(d;s)) \tag{9}$$

in (4). The next theorem shows the double robustness property of the doubly robust estimator.

Theorem 3.2. *For any student estimator π,*

$$\mathbb{E}[\hat{R}^{\mathsf{DR}}(\pi) - R(\pi)] \leq C_1 \mathbb{E}\left[\left\|[\ \Delta_{\mathbf{d}} \delta_{\mathbf{d}}]_{d \in \mathscr{D}_s}\right\|_2\right],$$

where C_1 is the same constant as in Lemma 2.1.

The proof is followed using similar techniques as in the proof of Theorem 3.1 and is included in the Supplemental Materials section along with a variance analysis. Theorem 3.2 shows the desirable double robustness of this estimator, where the bias of the individual terms will be small if we have good estimates for either the actual click propensities or the true relevance probabilities by the teacher.

3.2. Teacher Models

Our discussion of the distilled risk estimators were focused on their theoretical generalization properties in terms of the estimation error oblivious to the choice the teacher model. The key benefit of using knowledge distillation in counterfactual LTR is the improved generalization properties of the student, achieved through label smoothing via the teacher's distilled knowledge. While the student often has to satisfy stringent inference time constraints, the teacher model can be arbitrarily complex, not only in terms of the representational capacity of the hypothesis class, but also through leveraging features/representations that the student does not have access to at the inference time. Such features are referred to as privileged features in [35]. In this section, we discuss generic techniques to develop effective teacher models, which we will exemplify in the empirical results section.

3.2.1. De-confounding teacher

In order to train a relevance teacher using logged search data, we adopt the counterfactual inference framework based on potential outcome modeling. The fundamental inference challenge in this setting is that we are interested in counterfactual relevance outcome had the item was observed, denoted by $\mathbf{r}_d^{(\mathbf{o}_d=1)}$, while we only have observed relevancy under explicit user engagement events. In order to estimate expectations on counterfactual quantities using observational data, we invoke regression adjustment techniques from potential outcome modeling framework[39]. The core component in this technique is to control for the presentation-based confounders, e.g. the rank of the items, which affect both the observed outcome and the observation event.

Let $x_{d,s}$ be the (causal) covariates to predict the relevancy of d in the search context s and let $z_{d,s}$ be the presentation-based confounders. Controlling for confounders and making inference from observational data amounts to justifying a few technical conditions. Specifically, we have to ensure (a) *ignorability*, that there is no unmeasured confounder left out from the covariates $z_{d,s}$, as well as (b) *Overlap*, that for any feasible value of covariates, the probability of observing the relevancy of an item given the covariates, i.e., the propensity score, is bounded away from zero, $\mathbb{P}(\mathbf{o}_d = 1|x_{d,s}, z_{d,s}) > 0$. Moreover, we have to make sure that the definition of the observation event is (c) *stable* in the sense that the potential relevance outcome of an item does not depend on whether the other items are observed.

We can then compute the expectation of the outcome of counterfactual relevance using iterative expectations on observational data[39] using

$$\mathbb{E}[\mathbf{r}_\mathbf{d}^{(\mathbf{o}_\mathbf{d}=1)}] = \mathbb{E}[\mathbb{E}[\mathbf{c}_\mathbf{d}|\mathbf{o}_\mathbf{d} = 1, \mathbf{x}_{\mathbf{d},\mathbf{s}}, \mathbf{z}_{\mathbf{d},\mathbf{s}}]], \tag{10}$$

where the observational expectations on the right hand side can be estimated via a machine learned classifier trained on a dataset of items with explicit engagement events using a cross-entropy loss[4, 9, 3]. Assuming that the presentation bias is primarily signified in the rank of the items in the page, a vanilla instantiation of this approach is adopted in [3], where the relevance probability is estimated as the observational click probability in rank 1; that is, $\hat{\mathbb{P}}(\mathbf{r}_d = 1|s) = \hat{\mathbb{P}}(\mathbf{c}_d = 1|x_{d,s}, \pi_d = 1)$. Similarly, in the GBDT-based approach in [17], the model is trained with an interaction depth of 1 to avoid interactions between the rank and the relevance feature and at the inference time all the trees that use the rank feature are removed. In the two-tower approach in [4], a counterfactual model is developed via independent relevance and observation towers and only the relevance tower is used for inference. A similar approach[8], which we adopt in our empirical experiments, follows this perspective by training a model on a cross-entropy loss between the IPW-debiased empirical (Bernoulli) distribution of clicks and the relevance distribution defined by the predictor; that is

$$\frac{1}{|\mathcal{S}^{\tau}|}\sum_{s\in\mathcal{S}^{\tau}}\sum_{d\in\mathcal{D}^{s}} CE[\sigma(f(x_{d,s}))||\hat{r}^{\mathrm{IPW}}(d)]. \tag{11}$$

3.2.2. Relevance Teacher Trained on Expert Judgements

In contrast to the relevance teachers trained on the observed biased feedback, discussed in the previous sub-section, we can train teachers based on relevance labels provided by expert judgements. The fundamental premise of such relevance teachers is to approximate the marginal query-based relevance probabilities via expert-annotated labels that are assumed to capture the canonical intent of a given query. Knowledge distillation is crucial in this case, because the student does not have access to such judgement-based labels in its own training data, and a relevance teacher trained on these labels is not strong enough as a standalone ranker, because it cannot take advantage of the more granular preferences in user engagement data.

Another important advantage of using relevance teachers trained on expert judgments is that we can use a broad range of pooling techniques, including active learning, to collect judgements on queries for which we do not have rich user engagements, alleviating the selection bias inherent to all the methods discussed so far, which rely strongly upon implicit user feedback. In the next section, we provide strong empirical results by adopting such relevance teachers.

4. Empirical Results and Discussions

We evaluate the performance of the proposed empirical risk estimator by adopting a standard supervised counterfactual training and evaluation framework on real-world user behavior data collected from online traffic of a major E-commerce platform. We also report the results from an online randomized control experiment on a model trained on a distilled empirical risk estimate to showcase the generalization power of the proposed techniques from a rigorous causal evaluation perspective. For completeness, we also conduct offline counterfactual experiments on standard academic datasets generated via synthetic user behavior data generation on human-judged web-search data, following the simulation setup and code from [8].

Since the focus of the paper is on the incremental value from distilled empirical risk minimization techniques to control the *Estimation error* in LTR, we are oblivious to the experiment design choices related to *Approximation error*, *Optimization error*, and *parameter estimation* techniques for the observation model. As such, we fix the hypothesis class, model hyper parameters, training optimization techniques, and estimated/simulated propensity scores across all estimators and limit the set of baseline models accordingly. For the synthetic datasets, we describe the details of the experiment setup only briefly and refer the reader to [8] for extended discussions and alternative choices data synthesis choices and propensity estimation techniques. For experiments on the proprietary e-commerce setting, we only report lifts compared to a standard simple baseline, with a focus on the choices relevant to the estimation error, oblivious to model training and propensity estimation techniques.

4.1. Experiment Setup for Synthetic Datasets

Dataset Semantics: We use publicly available datsets with manual relevance labels Yahoo Webscope[40], MSLR-WEB30k[41], and Istella-S LETOR[42] and synthesize user engagement labels with standard logging policy and user behavior generation semantics, following [8, 1].
User behavior model: We use a vanilla position-bias based examination model with the observation probability $\mathbb{P}(\mathbf{o} = 1|\mathsf{rank}) = (1 + (\mathsf{rank} - 1)/5)^{-2}$, ignoring the trust bias. We also use the same clipping function used in [8] to control the variance of the simulated click propensities. Moreover, we assume that the parameters of the propensity model are known and need not be estimated. Since user behavior modeling and propensity estimation are not the focus of this paper, we avoid unnecessary comparisons with more complex techniques in the literature.
Estimators: The logging policy is trained based on a standard supervised training on 1% of training data using an empirical label distribution based on the available relevance judgments. The logging policy is assumed to be deterministic. The Naive estimator is trained on the biased empirical click distribution without any corrections, and the IPW estimator is trained via (6) using the actual propensity scores used for data synthesis. For all distilled risk estimators, we use the same pre-trained deconfounding teacher model, T_1, trained based on the IPW-debiased empirical distribution via the point-wise cross-entropy loss, discussed in (11). The KD_{T_1} is trained via (7), and the hybrid estimator $\mathsf{HD}_{T_1,\alpha=\frac{1}{2}}$ and the doubly robust estimator are trained using the same teacher model T_1 and the actual synthesized propensity scores, via label distributions (8) and (9) respectively.
Training: Following [8], for the hypothesis class, we use MLPs with two 32-unit hidden layers and adopt the standard LambdaLoss optimization framework[43] to train the model on the proposed empirical risk estimates with $N = 10^6$ training samples randomly sampled from the synthesized data.
Evaluation Metrics: We use the Normalized Expected Number of Clicks as the main evaluation metric, which is aligned with the training objective. Specifically, for a given search context s, with candidate set D_s, the expected number of clicks for a policy π is

$$\ell(\pi_f(\mathscr{D}_s))^T r(\mathscr{D}_s).$$

By adopting the synthesized observation probabilities $\mathbb{P}(\mathbf{o} = 1|\mathsf{rank})$ from our user behavior model as the discount function $\ell(\cdot)$ and IPW-debiased clicks as the relevance labels $r(\mathscr{D}_s)$, and normalizing the contribution of individual context by the maximum attainable value, we get our primary metric, NCTR. Moreover, by adopting the original Judgment-based relevance labels as relevance labels $r(\mathscr{D}_s)$ and data-agnostic standard log-rank discounts for $\ell(\cdot)$, we calculate NDCG_{rel} with respect to the using data-agnostic standard log-rank discount.

4.2. Experiment Setup for Real World User Data

Dataset Semantics: We collect a dataset of around 1M queries from a major e-commerce platform in 2023 with a single click event attributed to the logged SERP, oblivious to the logging policy and the post-click engagement events attributed to the clicked items. For training efficiency, we sample three negative samples at random from impressed items within each

training SERP. For the test data, we use a similar query sampling strategy, but keep all the candidate items to be re-ranked by the candidate ranker.
Estimators: The baseline model for all the reported lifts is trained on the Naive estimator corresponding to the observed clicks, without any debiasing. The IPW estimator, and all other estimators that rely on propensity correction, we use estimated propensities using the standard regression-based Expectation Maximization techniques on a vanilla examination based position bias model[17]. The teacher models are trained on a predictive task with a cross entropy loss over datasets with different distributions. The deconfounding teacher, T_1, is trained based on the IPW-debiased empirical distribution of labels for (query,item) pairs *with the same representation as the ranking task* sampled with more stringent conditions to ensure that the relevancy of both the positive and negative items are observed, satisfying the conditions explained in section 3.2.1 for training counterfactual models using observational data. Specifically, for positive examples, we require that there has to be an engagement event with shopping intent associated to the item, and for negative items we require that they appear above the last engaged item in the SERP. The relevance teacher, T_2, is trained via expert judgements trained on (query,item) pairs with binarized relevance annotations, which satisfies stringent calibration properties. We are oblivious to the data pooling and training techniques used to train this teacher. The hyper-parameter α of the the hybrid estimators is optimized via grid search on the NCTR objective.
Evaluations: For offline experiments, we use NCTR with respect to IPW-debiased clicks as relevance labels and a simple rank fit on the estimated propensities. We also use NDCG with respect to naive clicks as a vanilla evaluation metric. For online experiments, we report ranking efficiency metrics in terms of concentration of engagements in top slots(DCG) as well as the cumulative reward metrics with respect to the share of search result pages with an engagement event(CTR) from a randomized controlled experiment on the online search traffic of a major E-commerce platform.

4.3. Empirical Research Questions

Results and Discussions on Synthetic Data Table 1 provides an overview on the performance of the estimators of interest $M(\cdot, \mathcal{S}_E)$ for evaluation metrics M, specified as meta-columns, over the evaluation datasets $\mathcal{S}_E$, specified as columns. We observe across all datasets, that the IPW outperforms the Naive estimator and all distilled estimators outperform IPW. the An extended discussion on the performance of the IPW and DR_{T_1} can be found in [8] with interesting variance analysis.

The most interesting observation is that the pointwise counterfactual click prediction teacher T_1 exhibits a strong ranking performance when it is used as the only target component, which is aligned with the results from [8] where it is used as a standalone ranker. This observation signifies that there is no meaningful heterogeneity in training data distribution across search contexts in these public datasets and strong ranking performance can be obtained without having to resort to listwise modeling. There is barely any performance gain in using the doubly robust estimator and we omitted doing any parameter tuning for the hybrid estimator.

Counterfactual evaluations on Product Search Data Table 2 provides relative performance lift of the candidate estimators against a Naive Click model baseline. In stark contrast

Table 1
Performance evaluation of the candidate estimators in full ranking setting(there is no display rank cutoff) with known propensities using $N = 10^6$ simulated training queries, evaluated over 20 runs over the test dataset, following [8]. The Minimum Detectable Effect size given the size of the 90% confidence intervals from this experiment is around 0.005.

	NDCG_{rel}			NCTR		
	Yahoo!	MSLR	Istella	Yahoo!	MSLR	Istella
Logging	0.858	0.748	0.815	0.758	0.498	0.703
Naive	0.860	0.748	0.816	0.759	0.498	0.706
IPW	0.872	0.759	0.836	0.781	0.524	0.735
KD_{T_1}	0.881	0.772	0.846	0.796	0.548	0.749
$\text{HD}_{T_1,\alpha=\frac{1}{2}}$	0.879	0.767	0.843	0.793	0.538	0.746
DR_{T_1}	0.879	0.772	0.845	0.793	0.549	0.749

to the results on synthetic datasets, we observe that the point-wise teachers are not very effective for the ranking task due to heterogeneity in the training data distribution. This is because pointwise models focus their predictive power also on learning that some contexts are inherently easier/harder for relevance or engagement prediction tasks. Although the Bayes optimal relevance predictor is also Bayes optimal for list-wise ranking loss[44], when learning a contextually discriminative and well-calibrated relevance model is complex, such models fail to perform well in the ranking task. The performance gap(not reported in the table) is even higher for the relevance teacher trained on expert judgments, particularly because the model is primarily focused on the context affinity of the candidate items rather than historic performance.

We observe, however, that the hybrid risk estimators based the relevance teacher outperform the IPW estimator. It is interesting that the relevance teacher trained on expert judgements significantly helps with the performance of HD_{T_2,α^*}, signifying the importance of soft relevance labels for all candidate items and the synergy between the relevance annotations and the implicit user feedback. We note that the deconfounding teacher T_1 helps only marginally with the performance of the HD_{T_1,α^*} estimator compared to the IPW baseline. Despite the theoretical appeal, the empirical performance of the doubly robust estimator relies on effective assumptions on observing the relevance of the unengaged items. In fact, as demonstrated in the table, the doubly robust estimator based on the naive observation assumption that relevance is observed only in the event of explicit user engagements, fails to outperform even the IPW estimator in terms of ranking efficiency metrics.

Online Experiments on Product Search Ranking In section 3.2.2, we argued that a relevance teacher can help not only with better generalization, but also with alleviating selection bias of the ranker by further incorporating queries with poor engagement history in training. We also showed strong empirical results from offline counterfactual evaluations that incorporating relevance signals from a relevance teacher, trained on expert judgements, in a Hybrid distilled risk estimator can significantly improve offline SERP efficiency metrics. To support our claims on the generalization power of our proposed estimators, we report the results from a controlled randomized online experiment in a major e-commerce platform, where the relevance ranking

Table 2
Relative performance evaluation of the proposed risk estimators compared to a Naive click-based estimator as baseline, with 90% confidence intervals based on bootsraping on a test dataset of size $N = 10^5$. The effect size of a lift is deemed meaningful if greater than 0.3%.

	$\Delta(\mathrm{NDCG}_c(\cdot), \mathrm{NDCG}_c(\mathrm{Naive}))$	$\Delta(\mathrm{NCTR}(\cdot), \mathrm{NCTR}(\mathrm{Naive}))$
IPW	+0.15%	+0.6%
KD_{T_1}	-3%>	-3%>
KD_{T_2}	-3%>	-3%>
$\mathrm{HD}_{T_1,\alpha^*}$	+0.3%	+0.7%
$\mathrm{HD}_{T_2,\alpha^*}$	+0.7%	+1.3%
DR_{T_1}	-0.1%	+0.5%
DR_{T_2}	+0.2%	+0.7%

model is replaced with a model trained on a Hybrid distilled risk that relies both on debiased user engagement events and a relevance teacher trained on expert judgements. Beyond significant lifts($> +2\%$) in search efficiency metrics as measured by rank-discounted measures, we observed a remarkable lift($> +1\%$) in Click through rate, confirming generalization in terms of converting more novel queries, as well as $> 0.5\%$ reduction in search abandonment rate, which is mostly affected by *hard* queries with poor historic engagements.

5. Concluding Remarks

We proposed an effective debiasing technique for counterfactual learning to rank from observational search activity data by using distilled knowledge from a relevance teacher to inform the label distribution for a listwise ranking task. We established the generalization power of the proposed estimators through rigorous empirical results in offline counterfactual evaluations as well as online randomized controlled experiments on a ranking task in a major e-commerce platform. We also presented a theoretical analysis of the estimation error of the proposed estimators to justify the improved generalizations from a theoretical perspective. Our contributions highlight important insights into using potential outcome modeling from the more generic perspective of knowledge distillation.

References

[1] T. Joachims, A. Swaminathan, T. Schnabel, Unbiased learning-to-rank with biased feedback, in: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, ACM, 2017, pp. 781–789.

[2] O. Chapelle, Y. Zhang, A dynamic bayesian network click model for web search ranking, in: Proceedings of the 18th international conference on World wide web, 2009, pp. 1–10.

[3] A. Borisov, I. Markov, M. De Rijke, P. Serdyukov, A neural click model for web search, in: Proceedings of the 25th International Conference on World Wide Web, 2016, pp. 531–541.

[4] Y. Zhang, L. Yan, Z. Qin, H. Zhuang, J. Shen, X. Wang, M. Bendersky, M. Najork, To-

wards disentangling relevance and bias in unbiased learning to rank, arXiv preprint arXiv:2212.13937 (2022).

[5] N. Craswell, O. Zoeter, M. Taylor, B. Ramsey, An experimental comparison of click position-bias models, in: Proceedings of the 2008 international conference on web search and data mining, 2008, pp. 87–94.

[6] L. Yan, Z. Qin, H. Zhuang, X. Wang, M. Bendersky, M. Najork, Revisiting two tower models for unbiased learning to rank (2022).

[7] J. M. Robins, A. Rotnitzky, Semiparametric efficiency in multivariate regression models with missing data, Journal of the American Statistical Association 90 (1995) 122–129.

[8] H. Oosterhuis, Doubly robust estimation for correcting position bias in click feedback for unbiased learning to rank, ACM Transactions on Information Systems 41 (2023) 1–33.

[9] Z. Ovaisi, K. Vasilaky, E. Zheleva, Propensity-independent bias recovery in offline learning-to-rank systems, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1763–1767.

[10] A. Vardasbi, H. Oosterhuis, M. de Rijke, 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, 2020, pp. 1475–1484.

[11] A. Agarwal, X. Wang, C. Li, M. Bendersky, M. Najork, Addressing trust bias for unbiased learning-to-rank, in: The World Wide Web Conference, 2019, pp. 4–14.

[12] X. Wang, R. Zhang, Y. Sun, J. Qi, Doubly robust joint learning for recommendation on data missing not at random, in: International Conference on Machine Learning, PMLR, 2019, pp. 6638–6647.

[13] T. Yang, C. Luo, H. Lu, P. Gupta, B. Yin, Q. Ai, Can clicks be both labels and features? unbiased behavior feature collection and uncertainty-aware learning to rank, in: Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval, 2022, pp. 6–17.

[14] M. Richardson, E. Dominowska, R. Ragno, Predicting clicks: estimating the click-through rate for new ads, in: Proceedings of the 16th international conference on World Wide Web, 2007, pp. 521–530.

[15] H. Guo, J. Yu, Q. Liu, R. Tang, Y. Zhang, Pal: a position-bias aware learning framework for ctr prediction in live recommender systems, in: Proceedings of the 13th ACM Conference on Recommender Systems, 2019, pp. 452–456.

[16] M. Chen, C. Liu, Z. Liu, J. Sun, Scalar is not enough: Vectorization-based unbiased learning to rank, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 136–145.

[17] X. Wang, N. Golbandi, M. Bendersky, D. Metzler, M. Najork, 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, ACM, 2018, pp. 610–618.

[18] Z. Fang, A. Agarwal, T. Joachims, Intervention harvesting for context-dependent examination-bias estimation, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 825–834.

[19] X. Wang, M. Bendersky, D. Metzler, M. Najork, 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, 2016, pp. 115–124.

[20] E. Ebrahimzadeh, A. Cozzi, A. Bagherjeiran, Intent-aware propensity estimation via click pattern stratification, in: Companion Proceedings of the ACM Web Conference 2023, WWW '23 Companion, Association for Computing Machinery, New York, NY, USA, 2023, p. 751–755. URL: https://doi.org/10.1145/3543873.3587610. doi:10.1145/3543873.3587610.

[21] Q. Ai, K. Bi, C. Luo, J. Guo, W. B. Croft, Unbiased learning to rank with unbiased propensity estimation, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018, pp. 385–394.

[22] Y. Wang, D. Liang, L. Charlin, D. M. Blei, The deconfounded recommender: A causal inference approach to recommendation, arXiv preprint arXiv:1808.06581 (2018).

[23] M. Dudík, J. Langford, L. Li, Doubly robust policy evaluation and learning, arXiv preprint arXiv:1103.4601 (2011).

[24] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, T. Joachims, Recommendations as treatments: Debiasing learning and evaluation, in: international conference on machine learning, PMLR, 2016, pp. 1670–1679.

[25] L. Zou, C. Hao, H. Cai, S. Wang, S. Cheng, Z. Cheng, W. Ye, S. Gu, D. Yin, Approximated doubly robust search relevance estimation, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 3756–3765.

[26] Y. Saito, Doubly robust estimator for ranking metrics with post-click conversions, in: Proceedings of the 14th ACM Conference on Recommender Systems, 2020, pp. 92–100.

[27] A. Chuklin, P. Serdyukov, M. De Rijke, Click model-based information retrieval metrics, in: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, 2013, pp. 493–502.

[28] L. Yang, Y. Cui, Y. Xuan, C. Wang, S. Belongie, D. Estrin, Unbiased offline recommender evaluation for missing-not-at-random implicit feedback, in: Proceedings of the 12th ACM conference on recommender systems, 2018, pp. 279–287.

[29] A. K. Menon, A. S. Rawat, S. J. Reddi, S. Kim, S. Kumar, Why distillation helps: a statistical perspective, arXiv preprint arXiv:2005.10419 (2020).

[30] J. Tang, R. Shivanna, Z. Zhao, D. Lin, A. Singh, E. H. Chi, S. Jain, Understanding and improving knowledge distillation, arXiv preprint arXiv:2002.03532 (2020).

[31] H. Mobahi, M. Farajtabar, P. L. Bartlett, Self-distillation amplifies regularization in hilbert space, arXiv preprint arXiv:2002.05715 (2020).

[32] S. Reddi, R. K. Pasumarthi, A. Menon, A. S. Rawat, F. Yu, S. Kim, A. Veit, S. Kumar, Rankdistil: Knowledge distillation for ranking, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 2368–2376.

[33] D. Liu, P. Cheng, Z. Dong, X. He, W. Pan, Z. Ming, A general knowledge distillation framework for counterfactual recommendation via uniform data, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 831–840.

[34] S. Zeng, M. A. Bayir, J. J. Pfeiffer III, D. Charles, E. Kiciman, Causal transfer random forest: Combining logged data and randomized experiments for robust prediction, in: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 2021, pp. 211–219.

[35] S. Yang, S. Sanghavi, H. Rahmanian, J. Bakus, S. Vishwanathan, Toward understanding privileged features distillation in learning-to-rank, arXiv preprint arXiv:2209.08754 (2022).

[36] C. Xu, Q. Li, J. Ge, J. Gao, X. Yang, C. Pei, F. Sun, J. Wu, H. Sun, W. Ou, Privileged features distillation at taobao recommendations, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 2590–2598.

[37] J. Tang, K. Wang, Ranking distillation: Learning compact ranking models with high performance for recommender system, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 2289–2298.

[38] Z. Qin, L. Yan, Y. Tay, H. Zhuang, X. Wang, M. Bendersky, M. Najork, Improving neural ranking via lossless knowledge distillation, arXiv preprint arXiv:2109.15285 (2021).

[39] D. B. Rubin, Causal inference using potential outcomes: Design, modeling, decisions, Journal of the American Statistical Association 100 (2005) 322–331.

[40] O. Chapelle, Y. Chang, Yahoo! learning to rank challenge overview, in: Proceedings of the learning to rank challenge, PMLR, 2011, pp. 1–24.

[41] T. Qin, T.-Y. Liu, Introducing letor 4.0 datasets, arXiv preprint arXiv:1306.2597 (2013).

[42] C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, F. Silvestri, S. Trani, Post-learning optimization of tree ensembles for efficient ranking, in: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, 2016, pp. 949–952.

[43] X. Wang, C. Li, N. Golbandi, M. Bendersky, M. Najork, The lambdaloss framework for ranking metric optimization, in: Proceedings of the 27th ACM international conference on information and knowledge management, 2018, pp. 1313–1322.

[44] D. Cossock, T. Zhang, Statistical analysis of bayes optimal subset ranking, IEEE Transactions on Information Theory 54 (2008) 5140–5154.

A. Proof of Theorem 3.1

Note that we only need to focus on the bias of the empirical label of a single sample point and the proof follows by applying Lemma 2.1.

$$\begin{aligned}
\mathbb{E}\left[\hat{r}^{\mathsf{HD}}(\mathbf{d}) - \mathbf{r}_{\mathbf{d}}\middle|\mathbf{s}\right] &= \mathbb{E}\left[\alpha(\frac{\mathbf{c}_{\mathbf{d}}}{\hat{p}_{\pi_0(\mathbf{d})}} - \mathbf{r}_{\mathbf{d}})\middle|\mathbf{s}\right] + (1-\alpha)\Delta_{\mathbf{d}} \\
&= \mathbb{E}\left[\mathbb{E}\left[\alpha(\frac{\mathbf{o}_{\mathbf{d}}}{\hat{p}_{\pi_0(\mathbf{d})}} - 1)\mathbf{r}_{\mathbf{d}}\middle|\mathbf{s},\mathbf{r}\right]\right] + (1-\alpha)\Delta_{\mathbf{d}} \\
&= \mathbb{E}\left[\alpha(\frac{\mathbb{P}(\mathbf{o}_{\mathbf{d}} = 1|s)}{\hat{p}_{\pi_0(\mathbf{d})}} - 1)\mathbf{r}_{\mathbf{d}}\middle|\mathbf{s}\right] + (1-\alpha)\Delta_{\mathbf{d}} \\
&= \alpha\, p^*_{r_{\mathbf{d}}}\delta_{\pi_0(\mathbf{d})} + (1-\alpha)\Delta_{\mathbf{d}}.
\end{aligned}$$

B. Proof of Theorem 3.2

Note that we only need to derive a bound for the bias of the empirical label of a single sample point and the result follows by applying Lemma 2.1.

$$\begin{aligned}
\mathbb{E}\left[\hat{r}^{\mathsf{DR}}(\mathbf{d}) - \mathbf{r}_{\mathbf{d}}\middle|\mathbf{s}\right] &= \Delta_{\mathbf{d}} + \mathbb{E}\left[\frac{\mathbf{o}_{\mathbf{d}}}{\hat{p}_{\pi_)(\mathbf{d})}}(\mathbf{c}_{\mathbf{d}} - T(\mathbf{d};\mathbf{s}))\middle|\mathbf{s}\right] \\
&= \Delta_{\mathbf{d}} + \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbf{o}_{\mathbf{d}}}{\hat{p}_{\pi_0(\mathbf{d})}}(\mathbf{o}_{\mathbf{d}}\mathbf{r}_{\mathbf{d}} - T(\mathbf{d};\mathbf{s}))\middle|\mathbf{r},\mathbf{s}\right]\right] \\
&= \Delta_{\mathbf{d}} + \mathbb{E}\left[\frac{\mathbb{P}(\mathbf{o}_{\mathbf{d}} = 1|s)}{\hat{p}_{\pi_0(\mathbf{d})}}(\mathbf{r}_{\mathbf{d}} - T(\mathbf{d};\mathbf{s}))\middle|\mathbf{s}\right] \\
&= -\Delta_{\mathbf{d}}\delta_{\pi_0(\mathbf{d})}
\end{aligned}$$

C. Variance analysis for Hybrid Estimator

By invoking the law of total variance and using the fact that $R(f)$ is non-stochastic, we can write

$$\begin{aligned}
\mathbb{V}[\hat{R}^{\mathsf{HD}}(f)] &= \mathbb{E}\left[\mathbb{V}\left[\hat{R}^{\mathsf{HD}}(f) - \alpha R(f)\middle|\mathbf{s}\right]\right] \\
&\quad + \mathbb{V}\left[\mathbb{E}\left[\hat{R}^{\mathsf{HD}}(f) - \alpha R(f)\middle|\mathbf{s}\right]\right] \\
&= \mathbb{E}\left[\ell^2(f(\mathscr{D}_{\mathbf{s}}))^T\left[\mathbb{V}\left[\hat{r}^{\mathsf{HD}}(\mathbf{d}) - \alpha\mathbf{r}_{\mathbf{d}}\middle|\mathbf{s}\right]\right]_{\mathscr{D}_{\mathbf{s}}}\right] \\
&\quad + \mathbb{V}\left[\ell(f(\mathscr{D}_{\mathbf{s}}))^T\left[\mathbb{E}\left[\hat{r}^{\mathsf{HD}}(\mathbf{d}) - \alpha\mathbf{r}_{\mathbf{d}}\middle|\mathbf{s}\right]\right]_{\mathscr{D}_{\mathbf{s}}}\right] \\
&= \mathbb{E}\left[\ell^2(f(\mathscr{D}_{\mathbf{s}}))^T\left[\mathbb{V}\left[\alpha\mathbf{r}_{\mathbf{d}}(1 - \frac{\mathbf{o}_{\mathbf{d}}}{\hat{p}_{\pi_0(\mathbf{d})}})\middle|\mathbf{s}\right]\right]_{\mathscr{D}_{\mathbf{s}}}\right] \\
&\quad + \mathbb{V}\left[\ell(f(\mathscr{D}_{\mathbf{s}}))^T\left[\alpha\, p^*_{r_{\mathbf{d}}}\delta_{\pi_0(\mathbf{d})} + (1-\alpha)T(\mathbf{d};\mathbf{s})\right]_{\mathscr{D}_{\mathbf{s}}}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\ell^2(f(\mathscr{D}_{\mathbf{s}}))^T\left[\alpha^2 p^*_{r_{\mathbf{d}}}\left((1-p^*_{r_{\mathbf{d}}})\delta^2_{\pi_0(\mathbf{d})}+\frac{1-p^*_{o_{\mathbf{d}}}}{p^*_{o_{\mathbf{d}}}}(1+\delta_{\pi_0(\mathbf{d})})^2\right)\right]_{\mathscr{D}_{\mathbf{s}}}\right] \\
&\quad + \mathbb{V}\left[\ell(f(\mathscr{D}_{\mathbf{s}}))^T\left[\alpha\, p^*_{r_{\mathbf{d}}}\delta_{\pi_0(\mathbf{d})} + (1-\alpha)T(\mathbf{d};\mathbf{s})\right]_{\mathscr{D}_{\mathbf{s}}}\right]
\end{aligned}
$$

where the last two lines follows by a similar argument as in the proof of Theorem 3.1.

Our bias-variance analysis provides insight into the fundamental components of the generalization error of hybrid distilled empirical risk estimates. The first component of the variance in the last line is due to the IPW component and the variance penalty due to the distillation component is

$$
(1-\alpha)^2\mathbb{V}\left[\ell(f(\mathscr{D}_{\mathbf{s}}))^T\left[T(\mathbf{d};\mathbf{s})\right]_{\mathscr{D}_{\mathbf{s}}}\right]. \tag{12}
$$

Also, We have already observed that the bias penalty we incur by incorporating the distillation component is

$$
(1-\alpha)\mathbb{E}\left[\ell(f(\mathscr{D}_{\mathbf{s}}))^T\left[\Delta_{\mathbf{d};\mathbf{s}}\right]_{\mathscr{D}_{\mathbf{s}}}\right]. \tag{13}
$$

This analysis shows the variance reduction benefits of the Hybrid Distilled Risk estimator, which is further enhanced by increasing α, in the expense of potentially more bias, due to inaccuracy of the teacher's estimate of the actual relevance probability.

D. Variance analysis for Doubly Robust Estimator

By invoking the law of total variance and using the fact that $R(f)$ is non-stochastic, we can write

$$
\begin{aligned}
\mathbb{V}[\hat{R}^{\mathrm{DR}}(f)] &= \mathbb{E}\left[\mathbb{V}\left[\hat{R}^{\mathrm{DR}}(f)-R(f)\middle|\mathbf{s}\right]\right] \\
&\quad + \mathbb{V}\left[\mathbb{E}\left[\hat{R}^{\mathrm{DR}}(f)-R(f)\middle|\mathbf{s}\right]\right] \\
&= \mathbb{E}\left[\ell^2(f(\mathscr{D}_{\mathbf{s}}))^T\left[\mathbb{V}\left[\hat{r}^{\mathrm{DR}}(\mathbf{d})-\mathbf{r}_{\mathbf{d}}\middle|\mathbf{s}\right]\right]_{\mathscr{D}_{\mathbf{s}}}\right] \\
&\quad + \mathbb{V}\left[\ell(f(\mathscr{D}_{\mathbf{s}}))^T\left[\mathbb{E}\left[\hat{r}^{\mathrm{DR}}(\mathbf{d})-\mathbf{r}_{\mathbf{d}}\middle|\mathbf{s}\right]\right]_{\mathscr{D}_{\mathbf{s}}}\right] \\
&= \mathbb{E}\left[\ell^2(f(\mathscr{D}_{\mathbf{s}}))^T\left[\mathbb{E}\left[(\hat{r}^{\mathrm{DR}}(\mathbf{d})-\mathbf{r}_{\mathbf{d}})^2\middle|\mathbf{s}\right]\right]_{\mathscr{D}_{\mathbf{s}}}\right] \\
&\quad - \mathbb{E}\left[\ell(f(\mathscr{D}_{\mathbf{s}}))^T\left[\Delta_{\mathbf{d}}\delta_{\pi_0(\mathbf{d})}\right]_{\mathscr{D}_{\mathbf{s}}}\right]^2 \\
&= \mathbb{E}\left[\ell^2(f(\mathscr{D}_{\mathbf{s}}))^T\left[\Delta^2_{\mathbf{d}}\left(p^*_{o_{\mathbf{d}}}(\frac{1}{\hat{p}_{\pi_0(\mathbf{d})}}-1)^2+(1-p^*_{o_{\mathbf{d}}})\right)\right]_{\mathscr{D}_{\mathbf{s}}}\right] \\
&\quad - \mathbb{E}\left[\ell(f(\mathscr{D}_{\mathbf{s}}))^T\left[\Delta_{\mathbf{d}}\delta_{\pi_0(\mathbf{d})}\right]_{\mathscr{D}_{\mathbf{s}}}\right]^2
\end{aligned}
$$