

Ranking Policy Learning via Marketplace Expected Value Estimation From Observational Data*

Ehsan Ebrahimzadeh, Nikhil Monga, Hang Gao, Alex Cozzi, Abraham Bagherjeiran
eBay Search Ranking and Monetization

Abstract

We develop a decision making framework to cast the problem of learning a ranking policy for search or recommendation engines in a two-sided e-commerce marketplace as an expected reward optimization problem using observational data. As a value allocation mechanism, the ranking policy allocates retrieved items to the designated slots so as to maximize the user utility from the slotted items, at any given stage of the shopping journey. The objective of this allocation can in turn be defined with respect to the underlying probabilistic user browsing model as the expected number of interaction events on presented items matching the user intent, given the ranking context. Through recognizing the effect of ranking as an intervention action to inform users' interactions with slotted items and the corresponding economic value of the interaction events for the marketplace, we formulate the expected reward of the marketplace as the *collective* value from all presented ranking actions. The key element in this formulation is a notion of *context value distribution*, which signifies not only the attribution of value to ranking interventions within a session but also the distribution of marketplace reward across user sessions. We build empirical estimates for the expected reward of the marketplace from observational data that account for the heterogeneity of economic value across session contexts as well as the distribution shifts in learning from observational user activity data. The ranking policy can then be trained by optimizing the empirical expected reward estimates via standard Bayesian inference techniques. We discuss the connections and distinctions between our proposed perspective and the standard supervised approach to learning to rank via empirical risk minimization with respect to standard information retrieval metrics. The specific focus of this paper is to highlight the significance of the empirical context value distribution in shaping the properties of the corresponding ranking policies by contrasting various empirical importance sampling distributions. We report empirical results from online randomized controlled experiments on a product search ranking task in a major e-commerce platform demonstrating the fundamental trade-offs governed by ranking policies trained on empirical reward estimates with respect to extreme choices of the context value distribution.

Keywords

Learning to Rank, Counterfactual Inference, Expected Reward Estimation, E-commerce Marketplace

*Copyright 2024 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Presented at the SURE workshop held in conjunction with the 18th ACM Conference on Recommender Systems (RecSys), 2024, in Bari, Italy.

1 Introduction

1.1 Motivation

Two sided e-commerce marketplaces are intermediary economic platforms that connect buyers and sellers, usually providing a wide selection of products for the buyers from a diverse array of sellers. The primary *buyer focused* objective of the marketplace is to guide buyers through their browse and discovery journeys to identify and purchase the items that fulfill their *shopping intent*. Users' browsing and purchase journeys in the marketplace are impacted by an ecosystem of decision making systems, most notably via the ranking policies in various stages of their shopping journeys from discovery pages to the Search Engine Result Pages(SERP). An effective ranking policy aims to showcase a set of results matching the user intent at any given ranking context along the shopping journey with rewards realized as interaction events on the slotted items on the page. Collectively, user journeys are not equally likely to produce value for the marketplace and the goal is to expand the set of successful user sessions, optimizing a suitable notion of long term value for the users and the marketplace. It is therefore essential for the ranking policy to account for the utility of all stakeholders in this economic setting. In standard formulations of learning to rank in the information retrieval literature, however, there is usually no clear connection between the training objective for the ranking policy, the long term value for the collective of the users and the key performance metrics of the marketplace. In this paper, focusing primarily on the search ranking policy invoked in response to users' search queries, we formulate the ranking policy learning as an optimization problem based on a (counterfactual) estimate of the marketplace reward from observational data.

1.2 Contributions and Related Work

Contribution 1. *We propose a decision making framework establishing explicit connections between learning a ranking policy for a search/recommendation engines and building effective empirical estimates for a suitable notion of marketplace expected reward.*

The problem of developing merit scores for ranking items, post a selection stage from a large pool of candidates, is widely studies in the context of recommendation systems[18], display advertising[3], sponsored search[35], and search ranking[34], where the sequential and hierarchical nature of the user interaction events and sparsity of success events[15, 18, 21, 30] in user journeys are taken into account. Value-aware policies in the context of advertising[35, 35], and economic recommender systems[6, 7, 18] account for business objectives, primarily through manipulations of the merit scores based on conversion likelihood estimates and the price of the candidate items to develop a point-wise notion of expected value for a given candidate item. In contrast, our approach is user focused in that the goal of the ranking policy is to optimize for the user utility in the sense of maximizing the expected number of engagements

on desirable items at every stage of the search journey. An alternative formulation is to frame the search ranking policy learning as a multi-stakeholder multi-objective optimization problem[2, 19], with potentially conflicting objectives[28] that account either for business constraints[27] or group exposure constraints[20]. The notion of value for the marketplace is introduced in the ranking policy objective via an importance weighting distribution that signifies both the economic value and the likelihood of realizing some reward from an interaction event with an item that satisfies the user intent.

Contribution 2. *We characterize the key elements in building effective (policy-dependent) expected reward estimates from observational data, controlling for (1) the heterogeneity of the session value distribution, (2) the contribution of interventions within a user journey via the reward attribution scheme, and (3) the distribution shifts incurred by selection biases in observational data.*

Reinforcement learning(RL) is a powerful framework to account for sequential interventions within the session by formulating the problem of recommending new items[26] or search ranking[12] as a Markov Decision Process (MDP). By expanding the planning horizon and adopting intermediary reward shaping techniques, RL-based approaches account for delayed rewards in the session, via suitable representations of the dynamic session context(state) in session trajectories. Recognizing the selection biases in the observed user behavior data, offline reinforcement learning techniques, including inverse propensity weighting[4], and actor-critic methods[5], are adopted to account for distribution shifts in learning from logged data. Similar counterfactual training techniques based on propensity weighting and potential outcome modeling are developed in the context of counterfactual learning to rank for search ranking problems[11, 16, 22]. There is, however, no clear account of the heterogeneity of marketplace reward across session trajectories, neither in the standard counterfactual supervised learning perspective nor in offline reinforcement learning approaches.

Contribution 3. *We highlight the significance of the empirical session-context value distribution in building effective marketplace expected reward estimates by demonstrating fundamental performance trade-offs governed by the search ranking policies trained on extreme choices of the context value distribution via rigorous counterfactual evaluations as well as online randomized controlled experiments in a major e-commerce platform.*

The definition of success events and the associated reward to the user events is flexible in our framework and is informed by the strategic choices of the marketplace. Specifically, an early-stage marketplace may focus on maximizing the collective number of engagements, while an acquisition-oriented marketplace targets the collective number of purchases, while a revenue-driven marketplace chooses to maximize the long-term gross merchandise value.

1.3 Notation

Here is a list of notation adopted throughout the paper. Sets and ordered sets(lists) are represented with upper-case calligraphic symbols; such as \mathcal{X} . Random quantities are shown in bold such as \mathbf{x} with realization x . The expected value of random variable \mathbf{x} is denoted by $\mathbb{E}[\mathbf{x}]$ and the conditional expectation of a random

variable $\mathbf{z} = f(\mathbf{x}, \mathbf{y})$ given \mathbf{y} is denoted by $\mathbb{E}[\mathbf{z}|\mathbf{y}]$ or $\mathbb{E}_{\mathbf{x} \sim P(x)}[\mathbf{z}]$. For a function $f : \mathbb{X} \rightarrow \mathbb{R}$, the $|\mathcal{X}|$ dimensional array $[f(x)]_{x \in \mathcal{X}}$ is denoted by $f(\mathcal{X})$.

2 Problem Setup

2.1 Decision Making Framework

The marketplace is interested in maximizing the average total reward across all user session trajectories over a long time horizon

$$\frac{1}{T} \sum_{t \leq T} v_{s_t}, \quad (1)$$

where v_s is the economic value from a successfully served search session s . Our framework is flexible in the choice of the reward function and we discuss the fundamental trade-offs between multiple strategic marketplace long term reward choices, namely revenue-based, value per engagement and value per acquisition marketplaces. The reward from a session trajectory is assumed to be non-negative. Although our framework, can be extended to account for negative rewards, we ignore it in our formalization. We assume that the reward over search journeys is a stationary ergodic stochastic process. By invoking Birkhoff's ergodic theorem[9], with probability 1, the long term temporal average is same as the expected reward, i.e.

$$\mathbb{E}_{s,v}[v_s], \quad (2)$$

where the expectation is taken with respect to the randomness in the session context and reward distribution.

While our formalization can naturally be extended to search journeys with complex goals, we focus on a typical e-commerce purchase decision making scenario of session trajectory a with a single product intent, ignoring sessions with multi-product purchase intent, as well as informational and navigational search sessions. We recognize that users' decisions are impacted by multiple independently optimized decision making systems, but we are oblivious to potential interactions of the ranking policy with these systems, specifically to the closely related query understanding and candidate retrieval policies. We only focus on policy learning for search result pages with a single layer presentation semantic where the action of the ranking policy is the permutation/ranking of a largely homogeneous set of *comparable items* for a flat single-layered presentation of the results, ignoring the multiplicity of user's search intent and diversity considerations for the result set.

We cast the problem into a Bayesian decision making framework with a user-focused perspective on the definition of success upon a ranking action. The ranking policy aims to increase the expected number of engagements on items that meet the user intent, and the reward is proportional to the likelihood of a success event(non-zero reward) from the user interactions on the search results page produced by the ranking policy. A crucial aspect of this framework is to account for distribution shifts in observational data, i.e. the distinction between the distribution of the logged search activity data that the policy is trained and the inference time distribution of user events.

2.2 Success From a Ranking Intervention

Given a search query q within a session context s , the ranking policy $\pi : \mathcal{D}_q \rightarrow \{1, \dots, N\}$ maps a candidate item d from the retrieved set \mathcal{D}_q to a slot $\pi(d)$. The notion of success with respect to a slotting $\pi(\mathcal{D}_q)$ of the items on the SERP q is defined based on the effectiveness of the policy in driving user interaction events(Click). Specifically, the objective of the ranking policy on a given ranked SERP is to increase the expected number of engagements on desirable items(suitably defined) $c_{\pi(\mathcal{D}_q)}$ given the session context upon issuing the query $s_{< q}$; i.e.

$$\mathbb{E}[c_{\pi(\mathcal{D}_q)}|s_{< q}], \quad (3)$$

where the expectation is with respect to the randomness in user preferences and browsing behaviors in the given query context and possibly the randomness in the ranking policy(if stochastic). Note that $s_{< q}$ subsumes all the relevant contextual information upon issuing the query q ; including all the queries and the corresponding surfaced items, as well as the engaged items prior to the current query context. In the subsequent sections, we will make probabilistic assumptions on users' browsing and click behaviors and build effective policy dependent empirical estimators.

2.3 Success From a User Session

The notion of success with respect to a user session is defined based on interaction events on desirable items across all interventions by the marketplace within a user journey. Given the per query ranking objective $\mathbb{E}[c_{\pi(\mathcal{D}_q)}|s_{< q}]$ (i.e. the expected number of desirable engagements from the SERP given the session context upon issuing the query), the success from the overall user session is shaped by the distribution $\mathbb{P}(q|s)$ that signifies the contribution of the interactions on the ranked SERP q to the overall success of the user search session s .

$$\mathbb{E}_{q \sim \mathbb{P}(q|s_{< q})} [\mathbb{E}[c_{\pi(\mathcal{D}_q)}|s_{< q}]], \quad (4)$$

This distribution is usually referred to as the *success attribution distribution*, which is primarily studied in the context of online advertising[14]. The key difference is that in online advertising the unit of value attribution is item impression, while in this work we emphasize the attribution of value to the *ranked list* shown for the query given the prior context. This is also related to the *credit assignment problem* in reinforcement learning on how to attribute success to the intermediate actions of the agent.

2.4 Marketplace Expected Reward

The expected reward of the marketplace from the presented ranking is then shaped by the distribution of the value across search contexts, which signifies the economic value of the user-interaction events in the sessions for the marketplace. The random variable v_s captures the strategic notion of the value of the session for the marketplace, which corresponds to the value of the interactions events on the item(s) that satisfy the user's intent. The expected reward of the marketplace can then be written as

$$\mathbb{E}_{s, v} [v_s \mathbb{E}_{q \sim \mathbb{P}(q|s_{< q})} [\mathbb{E}[c_{\pi(\mathcal{D}_q)}|s_{< q}]]], \quad (5)$$

We can also consider an alternative formulation where we assume that session value distribution $\mathbb{P}_v(s)$ subsumes both the *likelihood*

of the user session to lead to some reward for the search engine as well as the *reward value* attributed to the user session s :

$$\mathbb{E}_{s \sim \mathbb{P}_v(s)} [\mathbb{E}_{q \sim \mathbb{P}(q|s_{< q})} [\mathbb{E}[c_{\pi(\mathcal{D}_q)}|s_{< q}]]]. \quad (6)$$

For a value-aware search engine with marketplace revenue objective, economic value is realized only in the event of a transaction as the success event from a search session and the reward is proportional to the price of the sold item(s). For a search engine that aims to optimize for the volume of transactions, it is more suitable to adopt a value per acquisition notion of reward oblivious to the price of the sold items. For a search engine with strategic goal of maximizing user engagements for increased user retention and minimizing abandonment, it is more suitable to adopt a value per click notion of reward oblivious to the post click transaction events.

In the next section, we discuss empirical modeling techniques to build effective empirical reward estimates from observational data, which effectively frame the problem as a standard counterfactual empirical risk minimization with a value-aware context distribution.

3 Expected Reward Estimation from Observational Data

3.1 Estimating the Per Query Success

We are interested in maximizing $\mathbb{E}[c_{\pi(\mathcal{D}_q)}|s_{< q}]$, the expected number of *desirable engagements* across all the slots on the page, with a suitably parameterized ranking policy π . By hypothesizing an explanatory click model based on causal constructs that govern user browsing and engagement behaviors on search result pages, we can build effective likelihood models from which we can estimate the parameters of the ranking policy via maximum likelihood estimation using logged observational data. We instantiate this process with the simple widely adopted click models in information retrieval. Assuming a vanilla Sequential Browsing Model along with the standard Position-Dependent Examination Model, we can write the expected number of desirable engagements as

$$\mathbb{E}[c_{\pi(\mathcal{D}_q)}|s_{< q}] = \sum_{r=1}^N \mathbb{P}(c_{\pi^{-1}(r)} = 1 | s_{< q}, \mathcal{D}_q^{< r}) \quad (7)$$

$$= \sum_{r=1}^N \mathbb{P}(c_{\pi^{-1}(r)} = 1 | s_{< q}, r) \quad (8)$$

$$= \sum_{r=1}^N \mathbb{P}(o_{\pi^{-1}(r)} = 1 | q) \times \mathbb{P}(R_{\pi^{-1}(r)} = 1 | s_{< q}) \quad (9)$$

$$= \sum_{r=1}^N \mathbb{P}(o_{\pi^{-1}(r)} = 1) \times \mathbb{P}(R_{\pi^{-1}(r)} = 1 | s_{< q}) \quad (10)$$

where the first line follows from sequential browsing assumption with $c_{\pi^{-1}(r)} = 1$ representing the click event on the item ranked at position r and $\mathcal{D}_q^{< r}$ representing the slotted items prior to the item ranked in position r . The second line follows from assuming that the user interaction event on a given slot is independent of the placed items in the previous slots. The third line follows from the standard examination-based click model that posits that a click event can be expressed as the intersection of a query specific examination

event $\mathbf{o}_{\pi^{-1}(r)} = 1$ and a presentation-independent contextual relevance event $\mathbf{R}_{\pi^{-1}(r)} = 1$; and the last line follows from assuming a query-independent global rank discount function on the examination probabilities. The examination probabilities, a.k.a. propensity scores, are context-specific and can be estimated via explicit online interventions or from observational data. By considering a simple uni-variate model fit on the estimated propensities as a function of rank, one can build data-driven rank-discount functions to estimate users' examination effort. However, the standard approach is to adopt vanilla log-based context-oblivious rank discounts as generic estimates of the examination probability $\hat{\mathbb{P}}(\mathbf{o}_{\pi^{-1}(r)} = 1)$, i.e.

$$\ell(r) = \frac{1}{\log(1+r)}.$$

Given an empirical estimate \hat{r} of Bayes contextual relevance probabilities $\mathbb{P}(\mathbf{R}_{\pi^{-1}(r)} = 1 | s_{\prec q})$, one can derive the standard discounted cumulative gain (DCG) estimate $\hat{\mathbb{E}}_{\text{DCG}}[\mathbf{c}_{\pi(\mathcal{D}_q)} | s_{\prec q}]$ of the expected reward per query for the policy $\pi(\cdot)$ as

$$\ell(\pi(\mathcal{D}_q))^T \hat{r}(\mathcal{D}_q). \quad (11)$$

Upon building all the elements of the expected reward estimates, specifically the policy-dependent per query expected reward estimates, we can train the ranking policy by maximizing this empirical reward estimate, as elaborated in the next section.

3.2 Estimating the Session Expected Reward

3.2.1 In-session success attribution. Several techniques can be adopted to estimate the contribution of a ranked SERP q and the corresponding *observed or potential* interactions on that page to the overall success of the user search session s . A simple yet popular solution in the context of online advertising is to adopt an attribution distribution that assigns all the probability mass to the immediate query context preceding the post-click conversion event, which is referred to as Last Touch Attribution scheme. In contrast to this tight attribution scheme, one can assume a uniform distribution across all queries in the session in which the item with the attributed interaction event of interest was retrieved as a candidate item, oblivious to whether it was even impressed on the search result page. This approach is referred to as All Touch attribution scheme. Alternatively, One can assume a (Markovian) probabilistic graphical model on user's touch points within a session journey and infer a probabilistic multi-touch attribution distribution

$$\hat{\mathbb{P}}(q | s_{\prec q})$$

from observational data. Similarly, one can adopt an attention based sequence modeling approach and infer the contribution weights for interaction events along the user journey with a conversion prediction model. Lose attribution schemes, like the all touch attribution scheme, signify the powerful idea of *counterfactual training context generation* for ranking policy learning, where in contrast to predictive perspectives, the policy can collect reward from a ranking context where the item of interest was not observed by the user. As discussed in the empirical results section, such attribution schemes are particularly effective for capturing the user behavior in search sessions with longer feedback loops, e.g. sessions with high purchase value user intent.

3.2.2 Session Value estimation. In order to highlight the importance of the session value distribution,

$$\hat{\mathbb{P}}(s)$$

let us focus on a search engine with a value per acquisition objective. A straightforward empirical session value distribution is to adopt a uniform distribution on sessions that lead to a transaction event. Such session value selection distribution leads to survivorship bias in training context selection in that traffic segments where transaction events are rare, e.g. user sessions with luxury intent, will be under-represented in training. A simple approach is to expand the definition of success events and estimate the likelihood of session success with a content-oblivious estimate based on the aggregate conversion likelihood of the *richest engagement event* attributed to the element(s) engaged. This perspective on building mixture distributions based on the richest post-click engagement event was shown to be effective in capturing potential conversions from browse-heavy user journeys[25].

For a revenue-focused marketplace, as discussed in Section 2.2, the value of a search session is proportional to the price of the item that matches the user intent. In the presence of an observed success event in the logged data, the purchase price of the item to which the success event is attributed is the realization of the session value; otherwise, in the absence of an interaction event, the value of the session has to be estimated from the content of the asked intent in user queries, or a Canonical set of actual or synthesized items that match the user intent.

3.3 Selection Bias Correction

One of the main challenges in learning from observational data is the distribution shift between the training data collected from the logging policies and the inference data distribution. We therefore have to introduce another set of techniques, e.g. importance weighting distributions, to account for this mismatch between the (population) expected reward in (6) and the estimated expected reward from the estimated quantities in the previous sections; that is,

$$\mathbb{E}_{s \sim \hat{\mathbb{P}}(s)} [\mathbb{E}_{q \sim \hat{\mathbb{P}}(q | s_{\prec q})} [\hat{\mathbb{E}}[\mathbf{c}_{\pi(\mathcal{D}_q)} | s_{\prec q}]]]. \quad (12)$$

An important source of distribution shift in observational search activity data is the the selection bias due to presentation of the items on the page and the sequential browsing of the users, implying that we only observe relevancy of the items to the user only in the event of an explicit user engagement and it is more likely to observe engagements on SERPs from higher ranking slots.

A key technique to account for this effect is to define a suitable notion of propensity, which is developed in the context of studying the effect of a treatment(an intervention) on a population by taking into account attributes of the treatment unit in the way the treatment is assigned. In the context of ranking, the treatment is defined in correspondence to the examination of a slotted item by the user, but the key difference with the standard applications of this concept is that the examination variable is not fully observable. An alternative approach based on potential outcome modeling, similar to actor-critic networks in the context of offline reinforcement learning, is proposed in [11], where distilled knowledge from a

teacher model is used in the form of soft predicted relevance labels to account for unobserved user feedback to achieve variance reduction and improved generalization.

3.4 Variance Reduction and Generalization

Having discussed an array of importance weighting schemes to build empirical expected reward estimates, it is essential to develop variance reduction techniques to control the generalization behavior of expected reward estimators. For brevity of presentation, we briefly discuss the various reduction techniques adopted and ignore developing generalization bounds on the bias and variance of the estimation error of the proposed empirical reward estimation techniques.

3.4.1 Truncation and Bucketing. Clipping and truncated importance sampling techniques[1, 13] are popular techniques to control the variance and generalization behavior of inverse propensity weighting estimators when there is high variance in the estimated propensities. Since we combine multiple importance sampling techniques to account for selection bias, success likelihood, and context value distribution across highly heterogeneous user trajectories, we adopt this simple variance reduction technique off the shelf.

In building empirical session value distributions for a revenue focused marketplace reward, relying on the purchase price of the success items leads to a very high variance estimator, particularly in the presence of high heterogeneity in price intent across user trajectories. Instead, we can use a stratification technique by bucketing user sessions based on value buckets defined according to the empirical revenue distribution. Specifically, we can build a session value distribution based on the empirical revenue share of the bucket corresponding to the price of the purchased item.

3.4.2 Potential Outcome Modeling. One of the primary challenges of counterfactual learning to rank from logged search activity data is that the relevancy of the items is observed only in the event of explicit user engagements. A popular idea in the context of contextual bandits and recommendation systems to circumvent the challenges in this partial information setting is to use predictive models for reward estimates as potential outcome models in conjunction with inverse propensity weighting[8, 24, 33]. There are a number of recent works in the context of unbiased response prediction that leverage and analyze the doubly robust technique[22, 23, 32, 36]. In [11], a generalized form of potential outcome modeling is proposed where the distilled knowledge from a relevance teacher is used in the form of soft predicted relevance labels to help the student with more effective list-wise comparisons, variance reduction, and improved generalization behavior. This is similar to the idea of actor-critic networks in the context of offline reinforcement learning[5], and augmentation policy in the context of contextual bandits[29]. Using knowledge distillation helps build training contexts from logged search contexts without user interaction events leveraging complex models. To simplify the discussions, we ignore discussing any details about the teacher models used in our experimental setup.

3.4.3 Stratification and Normalization. Effective stratification is a key technique in the context of importance weighting estimators, e.g. the context value binning idea discussed in sub-section 3.4.1

or training context stratification based on characteristics of logged training contexts[10]. We adopt Self-Normalizing propensity based estimators, recently analyzed in [17], where we use engagement ranks as yet another stratification dimension in our proposed estimators. Yet another standard variance reduction technique that we adopt to control the contribution of the search sessions with many success events in the observational data is to adopt normalization techniques; e.g. the standard Ideal cumulative gain normalization for the per query loss. We note that under this cumulative reward normalization technique, per item propensity weights should be reformulated as context weights.

Having equipped our empirical reward estimates with variance reduction techniques, from this point on, we can assume that the effect of all importance weighting schemes discussed so far are reflected in importance weights $\hat{v}_{q,s}$.

3.5 Optimization Objective for the Ranking Policy

We consider deterministic policies parameterized by a scoring function f , such that $\pi_f = \text{argSort}(f)$, oblivious to the representation of the items and the ranking context. An appealing approach, particularly in the context of online advertising and sponsored search, is to directly estimate the Bayes contextual relevance probabilities $\mathbb{P}(\mathbf{R}_d = 1|s_{\leq q})$, or equivalently the counterfactual probability of click had the item been examined $\mathbb{P}(\mathbf{c}_d = 1|s_{\leq q}, \text{do}(\mathbf{o}_d = 1))$ via a standard supervised predictive models, i.e.

$$\sum_{s,q} \hat{v}_{q,s} \sum_{d \in D_q} D(f(d) || \hat{r}_d), \quad (13)$$

where $\hat{v}_{q,s}$ is the empirical importance weights based on discussion in section 3.2, D is a distance measure, e.g. cross entropy, between the predicted distributions $f(d)$ and the properly debiased empirical label distribution \hat{r}_d . For estimating counterfactual probability of click that is *Contextually Well-Calibrated and Discriminative* for ranking, we need very complex models with rich feature representations, with careful data stratification and selection bias correction. Since absolute merit estimation is usually a harder problem than difference in merit estimation, we resort to alternative techniques for empirical expected reward optimization.

The standard alternative approach is to adopt the LambdaLoss framework[31] and optimize a pairwise upper bound on the (list-wise) empirical estimates for the expected number of engagements, $\ell_q(\pi_f, \hat{r})$, to circumvent the challenges of dealing with highly non-smooth rank-dependent policy function, which can be written as

$$\ell_q(\pi_f, \hat{r}) = \sum_{d,d' \in \mathcal{D}_q} \Delta \hat{\mathbb{E}}_{\pi_f} (\text{swap}_{\hat{r}}(d, d')) \sigma(f(d) - f(d')), \quad (14)$$

where $\hat{\mathbb{E}}_{\pi_f} (\text{swap}_{\hat{r}}(d, d'))$ is the difference in the estimated expected number of engagements had the ranked slots of the item pairs (d, d') been swapped and $\sigma(\cdot)$ is some inverse link function, e.g. softmax. The approximate surrogate objective, suitably weighted with the empirical reward estimates $\hat{v}_{q,s}$, expressed as

$$\sum_{s,q} \hat{v}_{q,s} \ell_q(\pi_f, \hat{r}) \quad (15)$$

can then be optimized using iterative optimization techniques, like Expectation-maximization; that is given an estimate $f^{(t)}$ at iteration t , in order to build $f^{(t+1)}$ from the gradient updates from the objective function, the difference in estimated objective $\hat{\mathbb{E}}_{\pi_{f(t)}}$ from the swap operation is computed based on ranking order produced by $f^{(t)}$.

4 Evaluations and Discussions

In Section 3, we discussed essential elements of building empirical expected reward estimates for training effective search ranking policies. Since conducting thorough ablation studies for characterizing the effect of each element in building empirical expected reward estimates is not possible given the space constraints, we focus primarily on the rather under-explored element in the literature, which is the effect of *context value distribution* discussed in section 3.2 in shaping the properties and the generalization performance of the ranking policy.

We focus on a product search ranking scenario in a major E-commerce platform and evaluate candidate policies via online randomized control experiments, as well as rigorous counterfactual evaluations on user session data collected from the online traffic. Since all experiment are performed on proprietary data, we only report lifts compared to a simple clearly-specified baseline, with a focus on the relevant choices for controlling the *estimation error* with respect to the research question of interest, oblivious to the optimization framework, the feature representations, and the hypothesis class. Specifically, we only discuss the choice of the ranking objective and the relevant importance sampling and attribution techniques for building our estimators of interest, without discussing the details of the models.

4.1 Online Evaluation Framework

Since the main goal of the proposed decision making framework is to build search ranking policies that generalize with respect to a given notion of marketplace expected reward, we primarily evaluate the performance of the candidate policies in online randomized controlled experiments. Specifically, we adopt an experiment design and primary success metric defined with respect to lifts in cumulative reward in treated user sessions. This cumulative reward driven design is in contrast to the standard experiment design practices for incremental ranking changes, where the primary success metric is set to be the standard (immediate) ranking efficiency metrics that measure concentration of success events in Top slots, through simple attribution and aggregation schemes across search result pages. In fact, top slot engagement concentration metrics, e.g. per query DCG with respect to SERP interactions aggregated uniformly across all queries, which are usually tightly correlated with the marketplace reward, should only be treated as secondary metrics in the presence of a measurement of cumulative reward in online experiments. We do recognize, however, that DCG-type metrics are particularly crucial for counterfactual off-policy evaluations, as approximations to the per query expected reward using logged data, because all we can do is to measure concentration of logged success events in top slots upon the shuffling action of the new target policy.

We establish the fundamental trade-offs between ranking policies trained on different empirical expected reward objectives primary based on session level cumulative reward metrics, including Number of Engagements, Number of Purchases, and Revenue, as measured in online AB tests. For metrics that attribute the observed effect to search events, we use a simple attribution schemes based on the immediate Search Result Page that precedes the user event of interest.

4.2 Training Objectives and Offline Evaluation Metrics

We adopt the standard supervised counterfactual training and evaluation framework based on logged search activity data collected from the online traffic of a major E-commerce platform. We are oblivious to the logging policy and collect datasets with importance sampling and reward attribution semantics based on the corresponding notions of expected reward of interest. Specifically, given a target notion of expected reward, the context value distribution remains the same for training and evaluation datasets. For candidate item selection per SERP, however, we sample three negative samples at random from impressed unengaged items within each training context, but keep all the candidate items to be re-ranked by the candidate ranker for the evaluation datasets.

For all empirical expected reward metrics, we use the same, suitably debiased and normalized, DCG approximation for the per query expected reward according to (11). Unless explicitly stated otherwise, we use the following vanilla empirical context value distribution for building expected reward estimates as training objectives and the counterfactual metrics.

Expected number of engagements $\hat{\mathbb{E}}[C]$: The session value distribution $\hat{\mathbb{P}}_C(s)$ is a uniform distribution across logged sessions with at least one click event. We consider a simple last touch attribution scheme $\hat{\mathbb{P}}_C(q|s_{<q})$ for the distribution of reward among queries within the session.

Expected number of purchases $\hat{\mathbb{E}}[P]$: The session value distribution $\hat{\mathbb{P}}_P(s)$ is uniform across logged sessions with at least one purchase event. We use a simple multi-touch attribution scheme $\hat{\mathbb{P}}_P(q|s_{<q})$ with uniform distribution across all queries in the converting session, where the purchased item appeared as a candidate.

Expected revenue $\hat{\mathbb{E}}[\text{Rev}]$: The session value distribution $\hat{\mathbb{P}}_R(s)$ is defined on the sessions with a transaction event according to the empirical revenue share of the bucket corresponding to the price of the purchased item. The same multi-touch attribution from above is adopted for this reward estimate as well.

To best highlight the heterogeneity of user behavior with respect to the underlying shopping intent and the associated fundamental trade-offs between different notions of marketplace expected reward, we also stratify our evaluations across traffic segments defined based on purchase price intent of the users, as realized in the price of the purchased item. The price intent bins are defined in such a way so that the empirical revenue distribution is roughly uniform across value buckets.

4.3 Research Questions

4.3.1 Marketplace Reward Trade offs. The primary insight that we would like to highlight in our evaluations is the heterogeneity of

users' browsing and shopping intents, as reflected in different notions of marketplace reward from user sessions. These observations signify the crucial importance of the choice of the empirical session value distribution in shaping the properties and the generalization behavior of the search ranking policy.

We do this by contrasting the performance of ranking policies trained on expected reward estimates corresponding to extreme choices of the empirical session value distribution. Specifically, we compare a policy π_C , corresponding to a scoring function f_C , trained on a simple engagement-driven expected reward estimate based on the session value distribution $\hat{\mathbb{P}}_C(s)$ against a policy π_P , corresponding to a scoring function f_P , trained on a simple acquisition-focused expected reward estimate based on the session value distribution $\hat{\mathbb{P}}_P(s)$. We observe meaningfully different performance trade-offs between these extreme policies with respect to the primary notions of marketplace reward in an online randomized controlled experiment. Table 1 summarizes the key observations on the average effect size $\Delta_S(m_{\pi_C}, m_{\pi_P})$ between engagement focused policy π_C and acquisition focused policy π_P , with respect to different cumulative metrics m , over the global session traffic S . The main takeaway from these observations is that the engagement focused policy π_C , on the one hand, drives significantly higher share of search sessions with at least one click ($> +3\%$), and on the other hand, leads to a significant drop in the share of search sessions with at least one purchase ($< -2\%$).

Online Metric m	$\Delta_S(m_{\pi_C}, m_{\pi_P})$
Sessions With Any Engagement	$>+3\%$
SERPs With Any Engagement	$>+5\%$
Sessions With A Purchase Event	$<-2\%$
Total Revenue	neutral

Table 1: Online AB test results contrasting policies with extreme session value distributions

It is interesting to note, however, that this drop is largely due to a significant loss in the number of bought items in search sessions with lower price intent, which usually take less exploration and browsing to identify and pinpoint the desirable item to purchase. Since the engagement-driven policy is more effective in driving success events with higher economic value in sessions that require more browsing effort, it can compensate for the revenue loss due to lower purchases in lower price intent segments, leading to an overall neutral effect size in total revenue.

In order to explore the fundamental trade-offs, highlighted in our online experiment, between different notions of marketplace expected reward across heterogeneous price intents in more depth, we build simple hybrid policies corresponding to a mixture of the engagement-based and acquisition based objectives. Specifically, we build a simple policy via a simple convex combination of the extreme policies

$$\pi_\alpha = \text{argSort}((1 - \alpha) f_P + \alpha f_C), \quad (16)$$

where π_α refers to the balanced ranking policy obtained via a linear combination of the scoring functions of the engagement focused policy, π_C and acquisition focused policy π_P , for some $\alpha \in [0, 1]$. The parameterized policies π_α behave similarly to a policy trained on a corresponding mixture session value distribution $(1 - \alpha)\hat{\mathbb{P}}_P(s) + \alpha\hat{\mathbb{P}}_C(s)$.

Due to the scarcity of online experimentation traffic, we only conduct counterfactual off-policy evaluations for these parameterized policies. While our counterfactual estimates are largely aligned, at least directionally, with the measured effect sizes in online experiments, we point out that all counterfactual off-policy evaluations are fundamentally limited having access only to snapshots of the users' behavior in the logged sessions. In particular, if the logging policy is substantially different from the target policy to be evaluated, the offline evaluation metrics could be very biased.

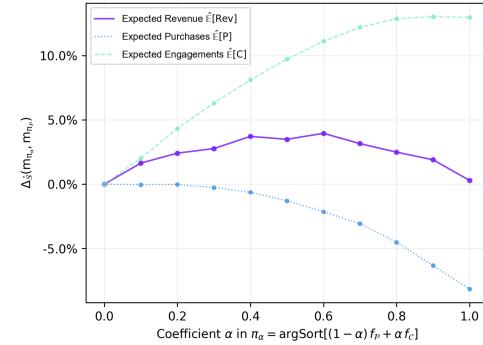


Figure 1: Counterfactual expected reward estimates $\Delta_{\hat{S}}(m_{\pi_\alpha}, m_{\pi_P})$ as a function of parameter α in .

Figure 1 highlights the essential trade offs between different expected reward estimates m from logged data \hat{S} , with the acquisition focused policy as the baseline

$$\Delta_{\hat{S}}(m_{\pi_\alpha}, m_{\pi_P}). \quad (17)$$

Biassing the training objective heavily on one extreme, leads to significant drops in the estimated reward corresponding to the other extreme. As the contribution of the engagement-focused policy increases, by increasing $\alpha > 0$, we estimate higher expected number of engagements, with a saturation point of diminishing return, after which a sharp drop in the expected number of purchases is observed. Interestingly the estimated expected revenue is convex as a function of α , which we will discuss in our subsequent research focused on value-aware objectives.

Next, we explore the observed trade-offs in the global analysis above across heterogeneous segments \hat{S}_p corresponding to different price intent segments, where the attribution of a session to a value bucket is done with respect to the price of the purchased items. Figure 2 and 3 show the lift in estimated expected number of engagements $\hat{E}[C]$ and estimated expected number of engagements $\hat{E}[P]$, respectively, for the hybrid policy π_α across value segments \hat{S}_p with the acquisition focused policy as the baseline.

We clearly see that the extreme acquisition focused policy performs poorly in terms of the expected number of engagements, across all segments, with particularly larger effects sizes in high value price intents that require more exploration. We also observe that, as α increases, the lift in expected clicks $\Delta_{\hat{S}_p}(\hat{E}_{\pi_\alpha}[C], \hat{E}_{\pi_P}[C])$ increases, with a saturation point in lower price segments (which is in fact an inflation point for low price intent segments). On the contrary, biasing the policy towards the engagement-focused policy, by setting α close to 1, leads to a meaningful drop in the expected number of purchases, $\Delta_{\hat{S}_p}(\hat{E}_{\pi_\alpha}[P], \hat{E}_{\pi_P}[P])$, particularly in low

value price segments, which constitute a high proportion of the overall number of purchases. An interesting observation, however, is that focusing more on an engagement-based objective is helpful for driving even higher expected number of purchases in higher price segments. We leave deeper dives on the observed trade offs for future work.

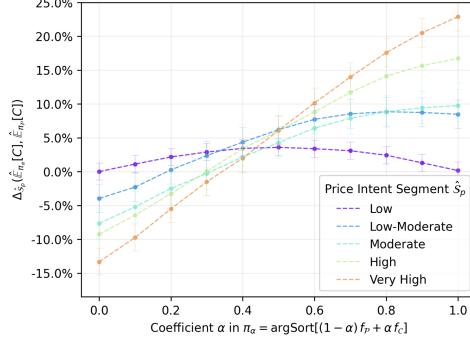


Figure 2: Counterfactual estimate for lift in the expected clicks

$\Delta_{\hat{S}_p}(\hat{E}_{\pi_\alpha}[C], \hat{E}_{\pi_p}[C])$ across price segments as a function of α .

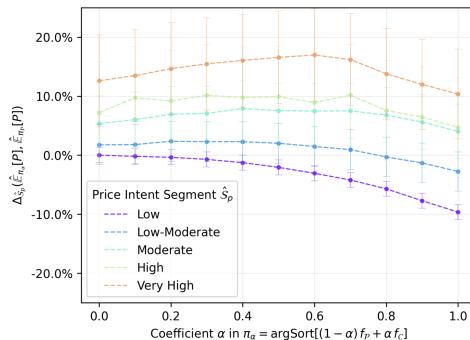


Figure 3: Counterfactual estimate for lift in the expected purchases

$\Delta_{\hat{S}_p}(\hat{E}_{\pi_\alpha}[P], \hat{E}_{\pi_p}[P])$ across price segments as a function of α .

4.3.2 Tight Attribution of Purchase Events. In order to highlight the significance of the reward attribution scheme within a user session, we contrast the generalization performance of policies trained with respect to extreme choices of the query contribution distribution $\hat{P}_P(q|s_{<q})$. Specifically, we contrast the performance of a policy π_t trained on a session value distribution with a tight attribution of success events to search events, similar to the last touch scheme discussed earlier, to a policy π_l trained with respect to a loose multi-touch attribution of success events to search events, similar to the all touch scheme discussed earlier.

$$\Delta_{\hat{S}_p}(m_{\pi_l}, m_{\pi_t})$$

Price Intent Buckets p	Purchases	Engagements	Revenue
Low	-0.95%	-0.66%	-0.68%
Low-Moderate	+0.61%	-0.13%	-0.14%
Moderate	+0.53%	+0.44%	-0.41%
High	+4.10%	+0.94%	+1.35%
Very High	+3.89%	+0.74%	+2.11%

Table 2: Online AB Test Results contrasting policies trained on extreme success attribution schemes

While the overall cumulative rewards do not show sizable performance trade-offs between the two extreme policies, we highlight substantially different effect sizes across different purchase price intents. Table 2 clearly demonstrates that a loose attribution scheme for the empirical query contribution distribution helps with a significant improved generalization in higher price intent sessions, which tend to be more exploratory and involve multiple ranking intervention touch points.

4.3.3 Purchase Price in Marketplace Reward. Finally, in order to highlight the significance of incorporating the purchase price in the session value distribution for a revenue focused marketplace reward, we highlight the results from an online AB test on a simple value-aware policy π_v in contrast to a value oblivious acquisition driven policy π_u . The primary difference between the two policies is the empirical session value distribution $\hat{P}(s)$ in the corresponding expected reward estimate for the training objective, which depends also on the price of the sold item in the case of π_v , and all the other importance weighting distributions and per query reward estimates are the same.

Price Intent Buckets p	$\Delta_{\hat{S}_p}(\text{Rev}_{\pi_v}, \text{Rev}_{\pi_u})$
Low	-1%
Low-Moderate	-0.1%
Moderate	+0.5%
High	+1.3%
Very High	+1.9%

Table 3: Online AB test results contrasting a purchase value-aware policy against a value oblivious policy

In Table 3, we clearly see a significant shift in the distribution of the accumulated revenue across price intent segments, which signifies the importance of taking into account the sparsity of purchase events from higher price intent sessions to avoid the selection bias incurred by the value oblivious policy, which is biased towards lower price intent segments, where sessions with purchase events are abundant.

5 Concluding Remarks

We established an explicit connection between the training objective for the search ranking policy and the key performance metrics of a two-sided commerce marketplace by building effective empirical estimates of the marketplace reward from observation data. Specifically, we highlighted the significance of the *search context value distribution* in building effective empirical estimates of the marketplace expected reward to inform the training and evaluation of the search ranking policy. We showcased empirical results from online randomized controlled experiments and counterfactual evaluations in a major e-commerce platform demonstrating the fundamental trade-offs governed by extreme choices of the context value distribution.

References

- [1] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research* 14, 11 (2013).
- [2] David Carmel, Elad Haramaty, Arnon Lazerson, and Liane Lewin-Eytan. 2020. Multi-Objective Ranking Optimization for Product Search Using Stochastic Label Aggregation. In *Proceedings of The Web Conference 2020*. 373–383.

- [3] Sougata Chaudhuri, Abraham Bagherjeiran, and James Liu. 2017. Ranking and calibrating click-attributed purchases in performance display advertising. In *Proceedings of the ADKDD'17*. 1–6.
- [4] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. 2019. Top-k off-policy correction for a REINFORCE recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 456–464.
- [5] Minmin Chen, Can Xu, Vince Gatto, Devanshu Jain, Aviral Kumar, and Ed Chi. 2022. Off-policy actor-critic for recommender systems. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 338–349.
- [6] Alvise De Biasio, Andrea Montagna, Fabio Aiolfi, and Nicolò Navarin. 2023. A systematic review of value-aware recommender systems. *Expert Systems with Applications* (2023), 120131.
- [7] Alvise De Biasio, Nicolò Navarin, and Dietmar Jannach. 2023. Economic recommender systems—a systematic review. *Electronic Commerce Research and Applications* (2023), 101352.
- [8] Miroslav Dudík, John Langford, and Lihong Li. 2011. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601* (2011).
- [9] Rick Durrett. 2019. *Probability: theory and examples*. Vol. 49. Cambridge university press.
- [10] Ehsan Ebrahizadeh, Alex Cozzi, and Abraham Bagherjeiran. 2023. Intent-Aware Propensity Estimation via Click Pattern Stratification. In *Companion Proceedings of the ACM Web Conference 2023*. 751–755.
- [11] Ehsan Ebrahizadeh, Alex Cozzi, and Abraham Bagherjeiran. 2024. Counterfactual Learning to Rank via Knowledge Distillation. In *Proceedings of ACM SIGIR Workshop on eCommerce (SIGIR eCom'24)*.
- [12] Yujing Hu, Qing Da, Anxiang Zeng, Yang Yu, and Yinghui Xu. 2018. Reinforcement learning to rank in e-commerce search engine: Formalization, analysis, and application. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 368–377.
- [13] Edward L Ionides. 2008. Truncated importance sampling. *Journal of Computational and Graphical Statistics* 17, 2 (2008), 295–311.
- [14] Wendi Ji, Xiaoling Wang, and Dell Zhang. 2016. A probabilistic multi-touch attribution model for online advertising. In *Proceedings of the 25th ACM international conference on information and knowledge management*. 1373–1382.
- [15] Jiarui Jin, Xianyu Chen, Weinan Zhang, Yuanbo Chen, Zaifan Jiang, Zekun Zhu, Zhewen Su, and Yong Yu. 2022. Multi-scale user behavior network for entire space multi-task learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 874–883.
- [16] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 781–789.
- [17] Ben London, Alexander Buchholz, Giuseppe Di Benedetto, Jan Malte Lichtenberg, Yannik Stein, and Thorsten Joachims. 2023. Self-normalized off-policy estimators for ranking. (2023).
- [18] Xiao Ma, Liqin Zhao, Guan Huang, Zhi Wang, Zelin Hu, Xiaoqiang Zhu, and Kun Gai. 2018. Entire space multi-task model: An effective approach for estimating post-click conversion rate. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1137–1140.
- [19] Debabrata Mahapatra, Chaosheng Dong, Yetian Chen, and Michinari Momma. 2023. Multi-label learning to rank through multi-objective optimization. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4605–4616.
- [20] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 2243–2251.
- [21] Conor O'Brien, Kin Sum Liu, James Neufeld, Rafael Barreto, and Jonathan J Hunt. 2021. An analysis of entire space multi-task models for post-click conversion prediction. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 613–619.
- [22] 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.
- [23] Yuta Saito. 2020. Doubly robust estimator for ranking metrics with post-click conversions. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 92–100.
- [24] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*. PMLR, 1670–1679.
- [25] Dominic Seyler, Ehsan Ebrahizadeh, Alex Cozzi, and Abraham Bagherjeiran. 2023. Aligning Ranking Objectives with E-commerce Search Intent. In *Proceedings of ACM SIGIR Workshop on eCommerce (SIGIR eCom'23)*.
- [26] Guy Shani, David Heckerman, Ronen I Brafman, and Craig Boutilier. 2005. An MDP-based recommender system. *Journal of machine Learning research* 6, 9 (2005).
- [27] Jie Tang, Huiji Gao, Liwei He, and Sanjeev Katayama. 2024. Multi-objective Learning to Rank by Model Distillation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5783–5792.
- [28] Manos Tsagkias, Tracy Holloway King, Surya Kallumadi, Vanessa Murdock, and Maarten de Rijke. 2021. Challenges and research opportunities in ecommerce search and recommendations. In *ACM Sigir Forum*, Vol. 54. ACM New York, NY, USA, 1–23.
- [29] Aaron D Tucker and Thorsten Joachims. 2023. Variance-Minimizing Augmentation Logging for Counterfactual Evaluation in Contextual Bandits. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 967–975.
- [30] Hao Wang, Tai-Wei Chang, Tianqiao Liu, Jianmin Huang, Zhichao Chen, Chao Yu, Ruopeng Li, and Wei Chu. 2022. ESCM2: entire space counterfactual multi-task model for post-click conversion rate estimation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 363–372.
- [31] 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.
- [32] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2019. Doubly robust joint learning for recommendation on data missing not at random. In *International Conference on Machine Learning*. PMLR, 6638–6647.
- [33] Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. 2018. The deconfounded recommender: A causal inference approach to recommendation. *arXiv preprint arXiv:1808.06581* (2018).
- [34] Liang Wu, Diane Hu, Liangjie Hong, and Huan Liu. 2018. Turning clicks into purchases: Revenue optimization for product search in e-commerce. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 365–374.
- [35] Yuyu Zhang, Hanjun Dai, Chang Xu, Jun Feng, Taifeng Wang, Jiang Bian, Bin Wang, and Tie-Yan Liu. 2014. Sequential click prediction for sponsored search with recurrent neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 28.
- [36] Lixin Zou, Changying Hao, Hengyi Cai, Shuaiqiang Wang, Suqi Cheng, Zhicong Cheng, Wenwen Ye, Simiu Gu, and Dawei Yin. 2022. Approximated doubly robust search relevance estimation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 3756–3765.