Reconsidering Learning Objectives in Unbiased Recommendation: A Distribution Shift Perspective

Teng Xiao
The Pennsylvania State University
tengxiao@psu.edu

Zhengyu Chen
Zhejiang University
chenzhengyu@zju.edu.cn

Suhang Wang
The Pennsylvania State University
szw494@psu.edu

ABSTRACT

This work studies the problem of learning unbiased algorithms from biased feedback for recommendation. We address this problem from a novel distribution shift perspective. Recent works in unbiased recommendation have advanced the state-of-the-art with various techniques such as re-weighting, multi-task learning, and meta-learning. Despite their empirical successes, most of them lack theoretical guarantees, forming non-negligible gaps between theories and recent algorithms. In this paper, we propose a theoretical understanding of why existing unbiased learning objectives work for unbiased recommendation. We establish a close connection between unbiased recommendation and distribution shift, which shows that existing unbiased learning objectives implicitly align biased training and unbiased test distributions. Built upon this connection, we develop two generalization bounds for existing unbiased learning methods and analyze their learning behavior. Besides, as a result of the distribution shift, we further propose a principled framework, Adversarial Self-Training (AST), for unbiased recommendation. Extensive experiments on real-world and semi-synthetic datasets demonstrate the effectiveness of AST.

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

Causal Inference; Unbiased Recommendation

ACM Reference Format:
Teng Xiao, Zhengyu Chen, and Suhang Wang . 2023. Reconsidering Learning Objectives in Unbiased Recommendation: A Distribution Shift Perspective. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3580305.3599487

1 INTRODUCTION

Recommender systems are widely used in many applications such as e-commerce platforms, social networks, and healthcare. However, recommender systems learn from logged user-item feedback data and are subject to selection bias as the training data collected by the logging policy is observational rather than experimental [17, 38, 56, 57, 61]. Ideally, the feedback should be collected by randomly and uniformly exposing items to users. However, in the real world, exposures are affected by the past recommendation policy, which is known as model selection bias. For example, users are more likely to interact with popular items than tail items, and recommender systems are also more likely to recommend popular items than others [38, 58, 61]. This model selection bias results in the "rich get richer" phenomenon, where head contents are getting more and more exposure while tail contents are rarely discovered. Selection bias also comes from user self-selection, i.e., users usually interact and rate items they like and rarely rate items they do not like [31, 38]. Previous studies [37, 38, 48, 55] have theoretically and empirically shown that directly learning from the biased feedback cannot reflect user true preferences on items.

Remarkable theoretical advances have been proposed for unbiased recommendation. Specifically, [38] and [48] provide rigorous generalization bounds under selection bias. On par with their theoretical findings, there have been rich advances in unbiased recommendation [37, 38, 48, 61] based on inverse propensity score (IPS)[33] and doubly robust (DR)[1] in causal inference. Although IPS and DR can address the selection bias in theory, these solutions typically assume unconfoundedness [59], i.e., the independence of user preference over items given the feature of getting exposed [37, 38, 59], which is impractical and cannot be examined in many real-world RS. Moreover, they need to estimate the propensity score for re-weighting and suffer from huge variance when the propensity score is small [36, 42]. Thus, IPS and DR empirically perform poorly compared to many recent works [7, 26, 36, 46, 49, 50].

Many unbiased recommendation algorithms have been introduced to conduct debiasing learning using various machine learning techniques, such as multi-task learning [6, 26], meta-learning [7, 49], and information bottleneck [50], which achieve promising empirical performance. However, there is a severe lack of rigorous theoretical analysis for these algorithms in the literature, creating a gap between current theory and many strong empirical methods. Specifically, most of these methods [6, 7, 26, 49] solve the bias issue by introducing unbiased uniform data in the training, which is collected by a random logging policy. Nevertheless, no clear and unified connection between current theory and these algorithms has been established. In other words, unbiased learning generalization bounds for them have not been derived. Furthermore, there is no solid theoretical justification for why utilizing unbiased uniform data can improve learning performance. Table 1 provides an overview of the discussed methods and suggests that most of them lack theoretical guarantees. This significant gap between theory and practice raises an important question: *How to bridge the gap between theories and recent unbiased learning objectives? Furthermore,*

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08…$15.00
https://doi.org/10.1145/3580305.3599487

Table 1: An overview of representative unbiased learning objectives we theoretically discuss in this paper, and how they relate to one another in terms of unconfoundedness assumption and whether they can work w/o unbiased uniform data, suffer from the variance issue, or whether the methods can theoretically unify other algorithms.

Learning objectives	w/o unconfoundedness assumption	w/o unbiased uniform data	w/o variance issue	unified framework
Re-weighting [15, 37, 38, 48]	✗	✓	✗	✗
Information bottleneck [28, 50]	✗	✓	✓	✗
Multi-task learning [6, 26]	✗	✗	✓	✗
Meta-learning [7, 49]	✗	✗	✓	✗
Adversarial self-training	✓	✓	✓	✓

could we propose a more effective unbiased learning objective guided by rigorous theoretical justification?

In this paper, we provide answers to the research question stated above. We first revisit unbiased recommendation from the perspective of distribution shift and then present a theoretical analysis of unbiased learning to provide explicit guidance and explanation for the current algorithm design. Our analysis shows that many unbiased learning objectives essentially optimize different terms in our bound. Unlike existing bounds [38, 48], our bounds explicitly suggest accounting for the unobserved confounders, which is important since the assumption of unobserved confounders may not hold in the real world (please see details in § 3.2). Our theoretical generalization bounds pave the way for us to understand why and how unbiased uniform data improves unbiased learning performance. We further provide insights into our theory analysis and propose a novel unbiased learning algorithm, Adversarial Self-Training (AST), which effectively minimizes the upper bound of the error and reduces the unbiased generalization gap. We evaluate AST on both real-world and semi-synthetic datasets and conduct ablation studies to analyze its behaviors. Extensive experimental results validate the effectiveness of AST. The main contributions of this work can be summarized as follows:

- We reconsider unbiased learning objectives proposed recently for recommendation from the perspective of distribution shift and provide a novel theoretical analysis towards explicit guidance and explanations for algorithm design.
- We provide important insights that our theoretical generalization bounds allow us to understand why and how unbiased uniform data helps to improve unbiased learning performance.
- Inspired by our theoretical analysis, we propose a novel unbiased algorithm, AST, which can maintain rigorous theoretical justification and address limitations of current algorithms. Extensive experiments on both semi-synthetic and real-world datasets also demonstrate the effectiveness of AST.

2 RELATED WORK

2.1 Selection Bias in Recommendation

Unbiased learning algorithms such as IPS [16, 37, 37, 38, 58, 64] and DR [48, 49] are proposed to theoretically address selection bias. For example, DR combines propensity score estimation and error imputation in a theoretically sophisticated manner. However, these methods heavily rely on accurately estimating the propensity score, which is often impossible to know in the real world. Furthermore, previous works [12, 42] have demonstrated that these methods suffer from high variance [35]. It is important to note that these causal inference methods typically assume unconfoundedness, where the relevance of user-item pairs is assumed to be independent of exposure given the user and item features [37, 38, 59]. Xu et al. [59] make similar observations regarding the limitations of the unconfoundedness assumption and highlight the inconsistent issues in supervised learning caused by unknown exposure mechanisms. However, they do not provide a theoretical framework to explain existing unbiased learning methods.

Recently, several empirical algorithms have been proposed to avoid the need for estimating the propensity score, utilizing techniques such as causal embedding [6], knowledge distillation [26, 27], and transfer learning [25]. These algorithms follow a multi-task learning scheme, where both unbiased uniform data and biased data are used, and the difference between the resulting user-item representations is regularized. Additionally, some algorithms adopt a meta-learning scheme [7, 49], where unbiased uniform data is used to supervise the learning of debiasing parameters within a bi-level optimization framework. Despite their promising performance in practice, most of these algorithms require additional unbiased uniform data, which can degrade user experiences, and they lack sufficient theoretical guarantees. As a result, there is currently a disconnect between theory and the existing algorithms. This work primarily focuses on addressing selection bias, with the aim of bridging the gap between theories and algorithms by proposing a theoretically motivated framework for unbiased recommendation.

2.2 Domain Adaptation and Self-Training

The unbiased recommendation problem setting can be treated as a special instantiation of out-of-distribution generalization and is related to domain adaptation [2, 3, 13, 30]. We discuss the relationships of our problem setting and our model with domain adaptation. The goal of domain adaptation is to train a predictor that performs well on a target domain using only labeled source samples and unlabeled target samples during training. The adversarial feature adaptation methods [13], inspired by the theoretical analysis of [3], are most similar to ours. Specifically, in [13], DANN is proposed to simultaneously minimize source empirical errors and approximate the divergence between source and target domains [3]. Our approach further develops this idea for unbiased learning in recommendation, but our work differs from domain adaptation in :(1) Our work focuses on the unbiased recommendation scenario where both selection bias and unobserved confounders exist simultaneously, as shown in § 3.2, and (2) we derive two novel generalization bounds for both multi-task and meta-learning strategies using unbiased uniform data proposed by recent unbiased recommendation algorithms [6, 7, 26, 49].

Our work is also related to self-training [4, 14, 51], which is a popular technique for semi-supervised learning. Self-training

assigns pseudo-labels to unlabeled samples by using a classifier's predictions and jointly re-trains the model with pseudo-labeled and labeled samples. Instead of focusing on semi-supervised learning, in this paper, we address the unbiased recommendation problem with the self-training. There are also some works [9, 22, 29, 62] applying the self-training for long-tail and cross-domain recommendation. Several previous works also have explored adversarial training to improve fairness [53], robustness [52], and accuracy [18, 47] of recommendation. Different from them, we focus on providing a theoretical analysis of existing unbiased learning objectives and addressing the selection bias issue via adversarial self-training.

3 PRELIMINARIES

In this section, we introduce basic notations and formulate the unbiased recommendation from the distribution shift perspective.

3.1 Notations and Selection Bias

Let $\mathbf{x}_u \in \mathcal{X}_\mathcal{U}$ be the feature vector for user $u \in \{1, \ldots, |\mathcal{U}|\}$, $\mathbf{x}_i \in \mathcal{X}_\mathcal{I}$ be the feature vector for item $i \in \{1, \ldots, |\mathcal{I}|\}$. Typically, the feature vectors can be user/item one-hot encoding, profile or embedding. $\mathcal{X}_\mathcal{U}$ and $\mathcal{X}_\mathcal{I}$ are the feature spaces, respectively. Following previous works [37, 59], we let $O_{u,i} \in \{0, 1\}$ be the exposure status, $Y_{u,i} \in \{0, 1\}$ be the feedback such as the click, and $R_{ui} \in \{0, 1\}$ be the true preference of user u on item i. $O_{u,i} = 1$ if the feedback $Y_{u,i}$ is observed and $Y_{u,i} = O_{u,i} \cdot R_{u,i}$ which means that, when item i has been exposed to u, the true preference should be equal to the feedback [37, 38]. Let $\mathcal{D}_P = \{\mathbf{x}_u, \mathbf{x}_i, Y_{ui} | O_{u,i} = 1\}$ be the logged feedback and the number of samples is N. The task of unbiased recommendation is to infer unobserved preference R_{ui}. Typically, the collected feedback follows a generative process [37, 38, 59]:

$$p(\mathbf{x}_u, \mathbf{x}_i, Y_{ui}) = p(\mathbf{x}_u)p(\mathbf{x}_i)p(R_{ui}, O_{ui} = 1 | \mathbf{x}_u, \mathbf{x}_i) = \quad (1)$$
$$p(\mathbf{x}_u)p(\mathbf{x}_i)p(O_{ui} = 1 | \mathbf{x}_u, \mathbf{x}_i)p(R_{ui} | O_{ui} = 1, \mathbf{x}_u, \mathbf{x}_i) \because Y_{ui} = O_{ui} \cdot R_{ui},$$

where the exposure distribution $p(O_{ui} = 1 | \mathbf{x}_u, \mathbf{x}_i)$ makes the observed feedback be missing-not-at-random (MNAR). We will drop $= 1$ for all O_{ui} in the remainder of the paper for conciseness. The exposure distribution $p(O_{ui} | \mathbf{x}_u, \mathbf{x}_i)$ is unknown and depends on user self-selection or the item exposure process by which past-recommendation policies match users and items. Since we want to eliminate the influence from the underlying exposure mechanism, ideally, we are interested in learning with the following unbiased risk function where the exposure is missing completely at random (MCAR), i.e., $O_{ui} \perp (R_{ui}, \mathbf{x}_u, \mathbf{x}_i)$:

$$\mathcal{L}_Q(f) \triangleq \mathcal{L}_Q(f, g) = \mathbb{E}_Q[\ell(f(\mathbf{x}_u, \mathbf{x}_i), g(\mathbf{x}_u, \mathbf{x}_i))] \quad (2)$$

where $Q \triangleq p(\mathbf{x}_u)p(\mathbf{x}_i)p(O_{ui})$ with $p(O_{ui}) = 1$ for all user-item pairs [38, 59, 60]. $f(\mathbf{x}_u, \mathbf{x}_i)$ is the estimated hypothesis. $g(\mathbf{x}_u, \mathbf{x}_i) = p(R_{ui} | \mathbf{x}_u, \mathbf{x}_i)$ is the optimal labeling function, depending on the true preference distribution $p(R_{ui} | \mathbf{x}_u, \mathbf{x}_i)$. Q is called as the marginal distribution over features. Typically, $\ell(f(\mathbf{x}_u, \mathbf{x}_i), g(\mathbf{x}_u, \mathbf{x}_i))$ is the 0-1 loss, which is the probability that f disagrees with g under Q: $\mathbb{E}_Q[\mathbf{I}(f(\mathbf{x}_u, \mathbf{x}_i) \neq g(\mathbf{x}_u, \mathbf{x}_i))]$. In this paper, we conduct theoretical analysis based on 0-1 loss. But, in practice, we can use 0-1 log loss $\ell(x, y) = -y \log \sigma(x) - (1 - y) \log(1 - \sigma(x))$ with $\sigma(x) = 1/(1 + e^{-x})$ which serves as a effective convex proxy for 0-1 loss.

We can notice that the unbiased risk function in Eq. (2) is independent of the exposure distribution of logged feedback, i.e., $p(O_{ui} | \mathbf{x}_u, \mathbf{x}_i)$. That is, we average the instance-wise loss over the uniform exposure distributions of all user-item pairs, $P(O_{ui}) = 1$, rather than the exposure distribution $p(O_{ui} | \mathbf{x}_u, \mathbf{x}_i)$. This uniform exposure scenario is ideal because the preference will not be affected by the previous exposure, thus leading to an unbiased estimation. In other words, unbiased recommendation wants to learn hypothesis f which generalizes well for all possible pairs of users and items, not just the pairs that are frequently exposed. The reason we suffer from the bias is because of the discrepancy between the exposure distribution of the logged feedback, and the testing distribution to which the model will be practically applied:

$$\textbf{Training} : p(\mathbf{x}_u)p(\mathbf{x}_i)p(O_{ui} | \mathbf{x}_u, \mathbf{x}_i)p(R_{ui} | O_{ui}, \mathbf{x}_u, \mathbf{x}_i) \quad (3)$$
$$\textbf{Testing} : p(\mathbf{x}_u)p(\mathbf{x}_i)p(O_{ui})p(R_{ui} | \mathbf{x}_u, \mathbf{x}_i). \quad (4)$$

Due the this discrepancy, the empirical risk $\widehat{\mathcal{L}}_P(f)$ over logged feedback $\mathcal{D}_P$ is a biased estimate of the ideal risk:

$$\hat{\mathcal{L}}_P(f) = \frac{1}{N} \sum\nolimits_{(\mathbf{x}_u, \mathbf{x}_i, Y_{ui}) \in \mathcal{D}_p} \ell(f(\mathbf{x}_u, \mathbf{x}_i), Y_{ui}) \simeq \mathcal{L}_P(f) \neq \mathcal{L}_Q(f),$$
$$\text{where } \mathcal{L}_P(f) \triangleq \mathcal{L}_P(f, k) = \mathbb{E}_P[\ell(f(\mathbf{x}_u, \mathbf{x}_i), k(\mathbf{x}_u, \mathbf{x}_i))], \quad (5)$$

$P = p(\mathbf{x}_u)p(\mathbf{x}_i)p(O_{ui} = 1 | \mathbf{x}_u, \mathbf{x}_i)$ and $k(\mathbf{x}_u, \mathbf{x}_i)$ is the optimal labeling function depending on distribution $p(R_{ui} | \mathbf{x}_u, \mathbf{x}_i, O_{ui})$ in the logged feedback. Thus, the learned f will not be approximately optimal even having sufficiently large training data [38].

3.2 The Unconfoundedness Assumption

To deal with this selection bias, many de-biasing methods [37, 38, 48] inspired by causal inference algorithms such as IPS and DR have been proposed. As mentioned by previous works [37, 59], these algorithms assume that being relevant is independent of getting exposed given the feature, i.e, $R_{ui} \perp O_{ui} | \mathbf{x}_u, \mathbf{x}_i$:

$$p(R_{ui} | O_{ui}, \mathbf{x}_u, \mathbf{x}_i) = p(R_{ui} | \mathbf{x}_u, \mathbf{x}_i). \quad (6)$$

We notice that this assumption is actually referred to as unconfoundedness assumption [34] in causal inference: assuming that there are no other latent variables except the features that affect both the outcome and the treatment assignment. With this assumption, we only have the distribution shift with respect to the exposure probability (see Eq. (4)) and the conditional distribution shift between $p(R_{ui} | \mathbf{x}_u, \mathbf{x}_i)$ and $p(R_{ui} | O_{ui}, \mathbf{x}_u, \mathbf{x}_i)$ vanishes (i.e., labeling function $g(\mathbf{x}_u, \mathbf{x}_i) = k(\mathbf{x}_u, \mathbf{x}_i)$). Thus, these methods [37, 38, 48, 64] conduct unbiased estimation by inversely re-weighting logged feedback:

$$\widehat{\mathcal{L}}_w(f) = \frac{1}{N} \sum\nolimits_{(\mathbf{x}_u, \mathbf{x}_i, Y_{ui}) \in \mathcal{D}_p} \frac{1}{p(O_{ui} | \mathbf{x}_u, \mathbf{x}_i)} \ell(f(\mathbf{x}_u, \mathbf{x}_i), Y_{ui}). \quad (7)$$

It is straightforward to verify that $\widehat{\mathcal{L}}_w$ is an unbiased estimation of ideal risk: $\mathbb{E}_P[\widehat{\mathcal{L}}_w(f)] = \mathcal{L}_Q(f)$ with the unconfoundedness assumption in Eq. (6). Thus, this objective can theoretically correct for the distribution shift caused by the exposure if $p(O_{ui} = 1 | \mathbf{x}_u, \mathbf{x}_i)$ is known in advance. While this objective has theoretical guarantee [37, 38, 48], there are three crucial directions for improvement:

(1) The unconfoundedness assumption may not be true and cannot be examined in real recommendation scenarios [11, 59], unless we can include every single factor that may affect users' decision-making process as a feature. However, there are other unobserved confounders, such as user social influence, item popularity effect, and public opinions, that cannot be captured through features. For example, as demonstrated in [23], user ratings exhibit different distributions when users rate items before or after reading public opinions. Additionally, due to privacy restrictions, recommender systems inevitably face unobserved confounders. For instance, user financial status directly affects feedback but is not measurable in

many recommender systems. Ignoring such confounders leads to an over-recommendation of inexpensive items. Nevertheless, current methods [37, 38, 48] do not consider these unobserved confounders.

(2) The theoretical analysis of this re-weighting objective [37, 38, 48] cannot explain and generalize well to many unbiased algorithms, especially those [6, 7, 26, 49] that utilize unbiased uniform data.

(3) This objective also requires accurate estimation of the exposure probability, which is usually challenging [36, 59] and suffers from significant variance. Consequently, it performs poorly in empirical comparison to recent algorithms [7, 26, 50].

4 THEORETICAL ANALYSIS

In this section, we first present our framework for unbiased recommendation from the distribution shift perspective with feature adaptation and derive two finite-sample generalization bounds. We provide a key insight that our theoretical framework is able to unify a series of recent unbiased learning objectives [6, 7, 26, 28, 49, 50].

4.1 Unbiased Learning via Feature Adaptation

In this subsection, we show how feature adaptation is related to unbiased recommendation. Recall that we have logged feedback $\mathcal{D}_P$ from distribution $P = p(O_{ui}|\mathbf{x}_u, \mathbf{x}_i,) p(\mathbf{x}_u)p(\mathbf{x}_i)$, where $P \triangleq P(\mathbf{x}_u, \mathbf{x}_i, O_{ui})$ is the training marginal distribution over features. Similarly, we have the testing marginal distribution $Q \triangleq p(\mathbf{x}_u)p(\mathbf{x}_i)p(O_{ui}) = 1/|\mathcal{U}||\mathcal{I}|$, meaning $(\mathbf{x}_u, \mathbf{x}_i)$ is sampled i.i.d. from uniform exposure distribution. Our goal is to learn a function $f(\mathbf{x}_u, \mathbf{x}_i)$ which can approximate the optimal function $g(\mathbf{x}_u, \mathbf{x}_i)$ which depends on preference distribution $p(R_{ui}|\mathbf{x}_u, \mathbf{x}_i)$.

To show how recent unbiased algorithms [6, 7, 26, 50] are related to feature adaptation, without loss of generality, we further consider the hypothesis $f(\mathbf{x}_u, \mathbf{x}_i)$, which is composed of a two parts: $f = h \circ \phi$ where $\phi \in \Phi \subset \{\phi : \mathcal{X}_u \times \mathcal{X}_i \to \mathcal{Z}\}$ is the feature mapping function and $h \in \mathcal{H} \subset \{h : \mathcal{Z} \to \mathcal{Y}\}$ is the hypothesis of the classification head. In general, h is a linear or feed-forward network predictor. Given this, we noticed that Ben-David et al. [3] and Blitzer et al. [5] proved the following bound on the unbiased risk $\mathcal{L}_Q(h \circ \phi)$ in terms of the empirical biased risk $\widehat{\mathcal{L}}_P(h \circ \phi)$ and the discrepancy between the training and testing distributions:

Theorem 4.1. *[3, 5] Let $\mathcal{H}$ be a hypothesis space with VC-dimension d. $P(\mathbf{z}_{ui})$ (resp. $Q(\mathbf{z}_{ui})$) is the distribution over $\mathcal{Z}$ induced by marginal distribution $P(\mathbf{x}_u, \mathbf{x}_i, O_{ui})$ (resp. $Q(\mathbf{x}_u, \mathbf{x}_i, O_{ui})$) and ϕ. Then, with probability (w.p) at least $1-\delta$ over the natural exponential e, $\forall h \in \mathcal{H}$:*

$$\mathcal{L}_Q(h \circ \phi) \le \widehat{\mathcal{L}}_P(h \circ \phi) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(P(\mathbf{z}_{ui}), Q(\mathbf{z}_{ui})) + \lambda(\phi) + \sqrt{\frac{4}{N}(d \log \frac{2eN}{d} + \log \frac{4}{\delta})}, \text{ where} \tag{8}$$

$d_{\mathcal{H}\Delta\mathcal{H}}(P(\mathbf{z}_{ui}), Q(\mathbf{z}_{ui})) = 2 \sup_{h,h' \in \mathcal{H}} \left|\mathbb{E}_{P(\mathbf{z}_{ui})}[\ell(h(\mathbf{z}_{ui}), h'(\mathbf{z}_{ui}))] - \mathbb{E}_{Q(\mathbf{z}_{ui})}[\ell(h(\mathbf{z}_{ui}), h'(\mathbf{z}_{ui}))]\right|$ is the $\mathcal{H}\Delta\mathcal{H}$-divergence [5] which measures the discrepancy between two distributions on symmetric difference hypothesis space and $\lambda(\phi) = \inf_{h \in \mathcal{H}}(\mathcal{L}_P(h \circ \phi) + \mathcal{L}_Q(h \circ \phi))$ is the combined risk of the ideal hypothesis.

Theorem 4.1 shows that the ideal risk $\mathcal{L}_Q(h \circ \phi)$ depends on three terms, which include the empirical risk $\widehat{\mathcal{L}}_P(h \circ \phi)$, the divergence between $P(\mathbf{z}_{ui})$ and $Q(\mathbf{z}_{ui})$, and the combined risk $\lambda(\phi)$. This bound serves as the theoretical foundation and has inspired the domain adaptation methods [13, 39], which simultaneously minimizes the divergence between $P(\mathbf{z}_{ui})$ and $Q(\mathbf{z}_{ui})$, and loss $\widehat{\mathcal{L}}_P(h \circ \phi)$.

This bound has made influential impacts in domain adaptation and we find there are two crucial directions to improve it for unbiased recommendation: (1) This bound considers aligning marginal distribution between $P(\mathbf{x}_u, \mathbf{x}_i, O_{ui})$ and $Q(\mathbf{x}_u, \mathbf{x}_i, O_{ui})$ by using latent feature adaptation, however it does not theoretically reflect the unjustifiable unconfoundedness assumption as shown in § 3.2. This will make the upper bound loose when the unconfoundedness assumption is violated in the real-world. (2) This bound still can not give the guidance and explanation for unbiased learning objectives [6, 7, 26, 49] that utilize unbiased uniform data. In what follows, we will introduce two generalization bounds based on this framework to address these two problems.

4.2 Unbiased Multi-Task Learning Bound

In this subsection, we give an unbiased multi-task learning bound which measures the unconfoundedness assumption. We also demonstrate that a a series of existing unbiased recommendation algorithms [6, 26, 28, 37, 38, 50] including those using unbiased uniform data can be partly interpreted by our new bound.

Specifically, some recent algorithms [6, 26] conduct de-biasing learning via unbiased uniform data, which is collected by a random exposure probability Q and can reflect user preferences in an unbiased way. Thus, besides the biased data $\mathcal{D}_P$, we assume that we have some unbiased uniform data $\mathcal{D}_Q = \{\mathbf{x}_u, \mathbf{x}_i, Y_{ui} | O_{u,i} = 1\}$ and the number of samples is M. Given the combined biased and unbiased data, these algorithms [6, 26, 28, 37, 38, 50] generally have the following empirical multi-task learning objective:

$$\rho \widehat{\mathcal{L}}_P(h \circ \phi) + (1-\rho)\widehat{\mathcal{L}}_Q(h \circ \phi) + \alpha R(\widehat{P}(\mathbf{z}_{ui}), \widehat{Q}(\mathbf{z}_{ui})), \tag{9}$$

where $\rho \in [0, 1]$. $\rho = 1$ means that we do not have unbiased uniform data $\mathcal{D}_Q$. Thus, this formulation can also unify those algorithms [28, 38, 50] without using unbiased data. $R(\cdot, \cdot)$ is the regularization function, and $\widehat{P}(\mathbf{z}_{ui})$ and $\widehat{Q}(\mathbf{z}_{ui})$ are empirical distributions of latent features over P and Q, respectively. α is the hyper-parameter. $\widehat{\mathcal{L}}_Q(h \circ \phi) = \frac{1}{M} \sum_{(\mathbf{x}_u, \mathbf{x}_i, Y_{ui}) \in \mathcal{D}_Q} \ell(h \circ \phi(\mathbf{x}_u, \mathbf{x}_i), Y_{ui})$ is the empirical objective under unbiased uniform data $\mathcal{D}_Q$. Based on this, we provide the following generalization bound:

Theorem 4.2. *Let $\mathcal{H}$ be a hypothesis space with VC-dimension d, and $P(\mathbf{z}_{ui})$ (resp. $Q(\mathbf{z}_{ui})$ is the probability density functions over $\mathcal{Z}$ induced by $P(\mathbf{x}_u, \mathbf{x}_i, O_{ui})$ (resp. $Q(\mathbf{x}_u, \mathbf{x}_i, O_{ui})$) and ϕ. $\tilde{g}$ (resp. $\tilde{k}$) is the labeling function over $\mathcal{Z}$ induced by g (resp. k) and ϕ. Then, w.p. at least $1-\delta$ over the exponential e, $\forall h \in \mathcal{H}$:*

$$\mathcal{L}_Q(h \circ \phi) \le \rho \widehat{\mathcal{L}}_P(h \circ \phi) + (1-\rho)\widehat{\mathcal{L}}_Q(h \circ \phi) + \frac{\rho}{2} d_{\mathcal{H}\Delta\mathcal{H}}(P(\mathbf{z}_{ui}), Q(\mathbf{z}_{ui})) + \rho \min\{\mathbb{E}_{P(\mathbf{z}_{ui})}[|\tilde{g}(\mathbf{z}_{ui}) - \tilde{k}(\mathbf{z}_{ui})|], \mathbb{E}_{Q(\mathbf{z}_{ui})}[|\tilde{g}(\mathbf{z}_{ui}) - \tilde{k}(\mathbf{z}_{ui})|]\} + \rho\sqrt{\frac{4}{N}(d \log \frac{2eN}{d} + \log \frac{4}{\delta})} + (1-\rho)\sqrt{\frac{4}{M}(d \log \frac{2eM}{d} + \log \frac{4}{\delta})}. \tag{10}$$

Remark. The proof is provided in Appendix B. This bound suggests that the ideal risk depends on the empirical multi-task learning error, the divergence of feature distributions, and the distance $\min\{\mathbb{E}_{P(\mathbf{z}_{ui})}[|\tilde{g}(\mathbf{z}_{ui}) - \tilde{k}(\mathbf{z}_{ui})|], \mathbb{E}_{Q(\mathbf{z}_{ui})}[|\tilde{g}(\mathbf{z}_{ui}) - \tilde{k}(\mathbf{z}_{ui})|]\}$ of labeling functions, which is essentially the divergence between conditional distributions $p(R_{ui}|O_{ui}, \mathbf{x}_u, \mathbf{x}_i)$ and $p(R_{ui}|\mathbf{x}_u, \mathbf{x}_i)$ [63].

Compared with the bound in Theorem 4.1 and other bounds in unbiased recommendation [38, 48], the bound in Theorem 4.2 has

two key differences: (1) it involves an empirical multi-task learning objective. When $\rho = 1$, the unbiased empirical error is not considered, with $\rho \in [0, 1)$, we introduce both biased and unbiased uniform data for de-biasing, and the generalizability of the model could be improved. This appeals to us since we can theoretically justify algorithms that employ unbiased data to conduct debiasing. (2) The term $\min\{\mathbb{E}_{P(\mathbf{z}_{ui})}[|\tilde{g}(\mathbf{z}_{ui}) - \tilde{k}(\mathbf{z}_{ui})|], \mathbb{E}_{Q(\mathbf{z}_{ui})}[|\tilde{g}(\mathbf{z}_{ui}) - \tilde{k}(\mathbf{z}_{ui})|]\}$ reflects the unconfoundedness assumption. This bound explicitly considers this assumption and suggests that if it is violated, i.e., $p(R_{ui}|O_{ui}, \mathbf{x}_u, \mathbf{x}_i) \neq p(R_{ui}|\mathbf{x}_u, \mathbf{x}_i)$, the bound will be loose. Thus, we should guarantee that the conditional distributions are not too far away from each other for successful unbiased recommendations.

The bound in Theorem 4.2 enables us to interpret many learning objectives [6, 26, 37, 38, 48, 50] in a unified perspective. Specifically, we show that they all fall into the multi-task objective in Eq. (9) and approximately minimize different terms in the bound.

Re-weighting Objectives [37, 38]. These methods fall into the multi-task learning objective in Eq. (9) with $\rho = 1$ and $\alpha = 0$ since they do not utilize unbiased uniform data and regularization. They re-weight the distribution P via $w(\mathbf{x}_u, \mathbf{x}_i) = 1/p(O_{ui}|\mathbf{x}_u, \mathbf{x}_i)$. By setting ϕ as the identity function, it is easy to verify that the first term in the bound becomes the re-weighting objective in Eq. (7) and the third becomes $d_{\mathcal{H}\Delta\mathcal{H}}\big(w(\mathbf{x}_u, \mathbf{x}_i)P(\mathbf{x}_u, \mathbf{x}_i, O_{ui}), Q(\mathbf{x}_u, \mathbf{x}_i, O_{ui})\big)$ which equals to zero. Thus, they essentially minimize the first and third term in this bound with $\rho = 1$ and ϕ being identity function.

Information Bottleneck Objectives [28, 50]. These algorithms also fall into the multi-learning objective with $\rho = 1$ and $\alpha \neq 0$. The regularization term in Eq. (9) is instantiated as the information bottleneck to regularize the model to learn a invariant representation across training and testing distributions, which makes the divergence $d_{\mathcal{H}\Delta\mathcal{H}}(P(\mathbf{z}_{ui}), Q(\mathbf{z}_{ui}))$ smaller. Thus, similar to re-weighting objectives, essentially, information bottleneck objectives also minimize the first and third term in this generalization bound.

Multi-task Objectives [6, 26]. These algorithms utilize unbiased uniform data and have the regularization term for approximately reducing the divergence between $P(\mathbf{z}_{ui})$ and $Q(\mathbf{z}_{ui})$, thus $\rho \neq 1$ and $\alpha \neq 0$ in Eq. (9). Specifically, R is $\|\mathbf{z}_{ui} - \hat{\mathbf{z}}_{ui}\|_2$ in [6] where $\mathbf{z}_{ui}$ and $\hat{\mathbf{z}}_{ui}$ are sampled from $\widehat{P}(\mathbf{z}_{ui})$ and $\widehat{Q}(\mathbf{z}_{ui})$, respectively. [26] designs other strategies for this regularization. Although the specific regularization may be different, the high-level motivation of them can be theoretically understood as minimizing the first, second and approximately reducing the third divergence terms in this bound.

4.3 Unbiased Meta-Learning Bound

As an alternative, there are some algorithms [7, 49] utilizing the unbiased uniform data via a meta-learning process [40, 54]. Their objectives are still based on the re-weighting method but they utilize the unbiased uniform data to train a weight function $w \in \mathcal{H}' \subset \{w : \mathcal{X}_u \times \mathcal{X}_i \to \mathcal{W}\}$ such that the hypothesis h trained on the biased data performs well on the unbiased uniform data. Specifically, the meta-learning can be formulated as a bi-level optimization as:

$$\min_{w} \widehat{\mathcal{L}}_Q(h(w) \circ \phi) \text{ s.t. } h(w) = \arg\min_{h,\phi} \widehat{\mathcal{L}}_{P_w}(h \circ \phi), \tag{11}$$

where $P_w = w(\mathbf{x}_u, \mathbf{x}_i)P(\mathbf{x}_u, \mathbf{x}_i, Y_{ui})$ stands for a new re-weighted distribution. $\widehat{\mathcal{L}}_Q(h(w) \circ \phi)$ is the upper-level objective under unbiased uniform data. Note that, $h(w)$ is the function of re-weighting and its new hypothesis space $\mathcal{H}'$ depends on biased training data due to the bi-level optimization [40]. Empirically, this objective perform well on unbiased recommendation as shown by [7, 49]. To theoretically understand this, we provide the following bound:

Theorem 4.3. *Let $\mathcal{H}$ and $\mathcal{H}'$ be hypothesis spaces with VC-dimension d and d', respectively. $P(\mathbf{z}_{ui})$ (resp. $Q(\mathbf{z}_{ui})$) is the density functions over $\mathcal{Z}$ induced by $P(\mathbf{x}_u, \mathbf{x}_i, O_{ui})$ (resp. $Q(\mathbf{x}_u, \mathbf{x}_i, O_{ui})$) and ϕ. $\tilde{g}$ (resp. $\tilde{k}$) is the latent labeling function induced by g (resp. k) and ϕ. Then w.p. at least $1 - \sigma$ and natural exponential e, $\forall h \in \mathcal{H}$, we have:*

$$\begin{aligned}
&\mathcal{L}_Q(h \circ \phi) \leq \rho\, \widehat{\mathcal{L}}_{P_w}(h \circ \phi) + (1 - \rho)\widehat{\mathcal{L}}_Q(h(w) \circ \phi) \\
&+ \frac{\rho}{2} d_{\mathcal{H}\Delta\mathcal{H}}(w(\mathbf{x}_u, \mathbf{x}_i)P(\mathbf{z}_{ui}), Q(\mathbf{z}_{ui})) + \rho\sqrt{\frac{4}{N}(d \log \frac{2eN}{d} + \log \frac{4}{\delta})} + \\
&+ \rho \min\{\mathbb{E}_{P(\mathbf{z}_{ui})}[|\tilde{g}(\mathbf{z}_{ui}) - \tilde{k}(\mathbf{z}_{ui})|], \mathbb{E}_{Q(\mathbf{z}_{ui})}[|\tilde{g}(\mathbf{z}_{ui}) - \tilde{k}(\mathbf{z}_{ui})|]\} \\
&+ (1 - \rho)(\frac{d' \log M - \log \delta}{3M} + \sqrt{\frac{2(d' \log M - \log \delta)}{M}}).
\end{aligned} \tag{12}$$

Remark. We provide the proof in Appendix C. This bound shows that the ideal risk depends on four non-constant terms: the empirical training errors on biased and unbiased data, the discrepancy between latent feature distributions, and the distance between the conditional distribution similar to Theorem 4.2. However, unlike Theorem 4.2, this empirical error on the unbiased uniform data is obtained via a meta validation process.

Meta-learning Objectives [7, 49]. With the bound in Theorem 4.3, we can understand why recent meta-learning approaches for unbiased recommendation can achieve good performance. It is worth noting that the bi-level meta-learning objectives in [7, 49] exactly minimizes the first and second terms with ϕ being the identity function. Unlike re-weighting objectives, $w(\mathbf{x}_u, \mathbf{x}_i)$ may not be the optimal sample weight, i.e., $1/p(O_{ui}|\mathbf{x}_u, \mathbf{x}_i)$. Thus, the meta learning objective can not theoretically guarantee that the third divergence term is small. Moreover, it also neglects the fifth term in the bound and essentially makes the unconfoundedness assumption.

5 ADVERSARIAL SELF-TRAINING

We have shown how our framework allows us to reinterpret many learning objectives in unbiased recommendation. With the above theoretical analysis and insights, we summarize the limitations of current learning objectives as follows: (1) They all make the unconfoundedness assumption, namely they do not account for the term about the conditional shifts in Theorems 4.2 and 4.3. Nevertheless, the unconfoundedness assumption is rarely true and can not be examined in the real-world [59]. (2) Some objectives try to minimize the $\mathcal{H}\Delta\mathcal{H}$ divergence between marginal feature distributions via re-weighting [37, 38] or different regularizers [6, 26, 50]. However re-weighting suffers from the variance issue [41]. As for the regularizers [6, 26, 50], they are only an approximation of the empirical $\mathcal{H}\Delta\mathcal{H}$-divergence which is hard to optimize. (3) Meta-learning objectives need to compute the second-order gradient is expensive in both computational cost and memory [7, 40, 49].

To address these issues, we exploit the theoretic analysis introduced in § 4 to derive a practical algorithm, adversarial self-training, which can simultaneously alleviate the divergence of feature distributions and approximately account for unobserved confounders. We optimize a feature mapping such that the conditional distribution is invariant to the biased training and unbiased testing data.

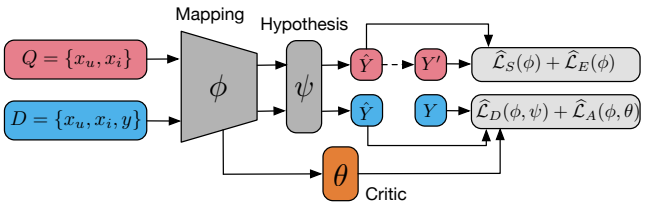

Figure 1: The architecture of AST. $\mathcal{D}$ is set of logged feedback, and Q is the set of randomly sampled user-item pairs.

5.1 Adversarial Training for Adaptation

Motivated by the discussion in § 4, we need to design a mechanism that enables feature adaptation for minimizing the $\mathcal{H}\Delta\mathcal{H}$-divergence. However it is difficult to optimize it. Thus, we give a new generalization bound to guide the design of AST.

Theorem 5.1. *Let $\mathcal{H}$ be a hypothesis space with VC-dimension d. $P(\mathbf{z}_{ui})$ (resp. $Q(\mathbf{z}_{ui})$) is the distribution over $\mathcal{Z}$ induced by marginal distribution $P(\mathbf{x}_u, \mathbf{x}_i, O_{ui})$ (resp. $Q(\mathbf{x}_u, \mathbf{x}_i, O_{ui})$) and ϕ. $\tilde{g}$ (resp. $\tilde{k}$) is the latent labeling function induced by g (resp. k) and ϕ Then, with probability at least $1-\delta$ over the natural exponential e, $\forall h \in \mathcal{H}$:*

$$\mathcal{L}_Q(h \circ \phi) \le \rho \widehat{\mathcal{L}}_P(h \circ \phi) + (1-\rho)\widehat{\mathcal{L}}_Q(h \circ \phi) + \frac{\rho\sqrt{2\mathrm{KL}(P(\mathbf{z}_{ui}) \| Q(\mathbf{z}_{ui}))}}{2}$$
$$+ \rho \min\{\mathbb{E}_{P(\mathbf{z}_{ui})}[|\tilde{g}(\mathbf{z}_{ui}) - \tilde{k}(\mathbf{z}_{ui})|], \mathbb{E}_{Q(\mathbf{z}_{ui})}[|\tilde{g}(\mathbf{z}_{ui}) - \tilde{k}(\mathbf{z}_{ui})|]\}$$
$$+ (1-\rho)\sqrt{\frac{4}{M}(d \log \frac{2eM}{d} + \log \frac{4}{\delta})} + \rho\sqrt{\frac{4}{N}(d \log \frac{2eN}{d} + \log \frac{4}{\delta})}. \quad (13)$$

Remark. The proof is provided in Appendix D. This bound provides theoretical justification for the use of KL (Kullback–Leibler)-divergence to conduct feature adaptation in unbiased recommendation. While the explicit marginal densities of $P(\mathbf{z}_{ui})$ and $Q(\mathbf{z}_{ui})$ are intractable, we have data samples of them. This motivates us to leverage adversarial distribution matching strategies [32] to minimize KL-divergence through a mini-max game with samples. In particular, we minimize $\mathrm{KL}(P(\mathbf{z}_{ui}) \| Q(\mathbf{z}_{ui}))$ via the use of a critic function (the max-step), and then update the feature mapping ϕ accordingly to reduce the KL-divergence (the min-step). In this paper, we consider the Fenchel-dual form of the KL-divergence [32], i.e.,

$$\mathrm{KL}(P \| Q) = \mathbb{E}_P[\log P - \log Q] = \max_{\nu > 0}\{\mathbb{E}_P[\log \nu] - \mathbb{E}_Q[\nu] + 1\}. \quad (14)$$

To optimize this Fenchel-dual form in practice, we model $\log \nu$ using another function $\theta(\mathbf{z}_{ui})$ as our critic function. This results in the following adversarial neural estimator of $\mathrm{KL}(P(\mathbf{z}_{ui}) \| Q(\mathbf{z}_{ui}))$:

$$\widehat{\mathcal{L}}_A(\phi, \theta) = \min_{\phi} \max_{\theta} \mathbb{E}_{\mathbf{z}_{ui}=\phi(\mathbf{x}_u,\mathbf{x}_i),(\mathbf{x}_u,\mathbf{x}_i)\sim P(\mathbf{x}_u,\mathbf{x}_i,O_{ui})}[\theta(\mathbf{z}_{ui})]$$
$$- \mathbb{E}_{\mathbf{z}_{ui}=\phi(\mathbf{x}_u,\mathbf{x}_i),(\mathbf{x}_u,\mathbf{x}_i)\sim Q(\mathbf{x}_u,\mathbf{x}_i,O_{ui})}[\exp(\theta(\mathbf{z}_{ui}))]. \quad (15)$$

Compared to $\mathcal{H}\Delta\mathcal{H}$-divergence, this objective is much easier to minimize and can theoretically bound the ideal unbiased risk as shown in Theorem 5.1.

5.2 Supervised Learning and Self-Training

As suggested by the generalization bound in Theorem 5.1, we also need to minimize the empirical learning error and the distance between the optimal labeling functions. For the empirical multi-task learning error, we can directly minimize it by parameterizing hypothesis h with function ψ:

$$\widehat{\mathcal{L}}_D(\phi, \psi) = \mathbb{E}_{\mathbf{z}_{ui}=\phi(\mathbf{x}_u,\mathbf{x}_i),(\mathbf{x}_u,\mathbf{x}_i,y)\sim\mathcal{D}}[\ell(\psi(\mathbf{z}_{ui}), Y_{ui})], \quad (16)$$

where $\mathcal{D} = \mathcal{D}_P \cup \mathcal{D}_Q$ is the whole set of data, including the biased data and the unbiased uniform data. Note that our algorithm can conduct de-biasing learning without unbiased uniform data when $\mathcal{D} = \mathcal{D}_P$. To further minimize the distance between conditional distributions (i.e., the regularizing term on conditional distributions), we need to search for a feature mapping ϕ such that the conditional distribution is invariant to training and testing: $\mathbb{E}_P[Y_{ui}|\phi(\mathbf{x}_u, \mathbf{x}_i)] = \mathbb{E}_Q[Y_{ui}|\phi(\mathbf{x}_u, \mathbf{x}_i)]$. If we have a small amount of unbiased uniform data from Q, we can directly minimize this regularizing term on conditional distributions by jointly minimizing $\widehat{\mathcal{L}}_D(\phi, \psi)$ on both biased data and unbiased uniform data.

However, in some scenarios, collecting unbiased uniform data is extraordinarily expensive [26, 50]. Thus, directly optimizing this term $\widehat{\mathcal{L}}_D(\phi, \psi)$ with unbiased uniform data becomes inaccessible. To account for this scenarios, in this paper, we propose to approximately evaluate and minimize this term by using self-training. Previous works [8, 51] have theoretically shown that self-training can learn the invariant predictive distribution, which can yield equally optimal performance across environments. This matches our goal of making conditional distribution invariant to the training and testing. Specifically, we adopt the principle of self-training, which has shown to be effective in semi-supervised learning [4, 14, 51]. Self-training first trains the feature mapping ϕ and prediction head ψ via $\widehat{\mathcal{L}}_D(\phi, \psi)$ in Eq. (16), and the trained model generates pseudo-labels for the unlabeled data sampled from $Q(\mathbf{x}_u, \mathbf{x}_i, O_{ui})$. Then self-training trains feature mapping with pseudo-labels as:

$$\widehat{\mathcal{L}}_S(\phi) = \mathbb{E}_{\mathbf{z}_{ui}=\phi(\mathbf{x}_u,\mathbf{x}_i),(\mathbf{x}_u,\mathbf{x}_i)\sim Q(\mathbf{x}_u,\mathbf{x}_i,O_{ui})}[\ell(\psi(\mathbf{z}_{ui}), Y'_{ui})], \quad (17)$$

where $Y'_{ui} = \bar{\psi}(\bar{\phi}(\mathbf{x}_u, \mathbf{x}_i))$ is the generated *soft* pseudo-label (it can be the ground-true label if we have a small amount of unbiased uniform data). $\bar{\psi}$ and $\bar{\phi}$ indict that we do not propagate gradients through computing the pseudo labels. We empirically found that this self-training can effectively brings conditional distributions closer even we do not have any unbiased uniform data. In addition, inspired by the recent work [8] which proves that entropy minimization has a similar effect as self-training algorithm, we also explicitly minimize the entropy on unlabeled uniform data:

$$\widehat{\mathcal{L}}_E(\phi) = \mathbb{E}_{\mathbf{z}_{ui}=\phi(x_u,x_i),(x_u,x_i)\sim Q(\mathbf{x}_u,\mathbf{x}_i,O_{ui})}[\mathrm{H}(\sigma(\psi(\mathbf{z}_{ui}))], \quad (18)$$

where $\mathrm{H}(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$ is the entropy of X. Intuitively, by minimizing this entropy, we can effectively encourage the prediction to be low-entropy (i.e., high-confidence) on unlabeled data and the classifier's decision boundary should not pass through high-density regions of the data distribution [4]. In summary, the overall objective function of AST could be formulated as follows:

$$\mathcal{L} = \min_{\phi,\psi} \max_{\theta} \widehat{\mathcal{L}}_D(\phi, \psi) + \alpha \widehat{\mathcal{L}}_A(\phi, \theta) + \beta \widehat{\mathcal{L}}_S(\phi) + \gamma \widehat{\mathcal{L}}_E(\phi, \psi), \quad (19)$$

where α, β and γ are trade-off hyper-parameters controlling the contributions of different losses.

Overall algorithm. Our full algorithm, Adversarial Self-Training (AST) is illustrated in Figure 1 and given in Algorithm 1. At each iteration, we sample mini-batches from biased labeled and unlabeled unbiased data. We generate the pseudo-labels for the unlabeled unbiased data by the current model. Then the model is further trained on the labeled biased data and pseudo-labeled unbiased data. The critic θ is optimized adversarially for minimizing the conditional shift between biased training and unbiased test distributions.

Algorithm 1: Adversarial Self-Training (AST)

Input: The collected biased data $\mathcal{D}_P$, unbiased data D_Q and parameters α, β, γ. Learning rate η. Maximum steps T.
if $\mathcal{D}_Q \neq \emptyset$: $\mathcal{D} = \mathcal{D}_P \cup \mathcal{D}_Q$ **else**: $\mathcal{D} = \mathcal{D}_P$
For n= $1, \cdots, T$ **do**
  Sample batches of $(\mathbf{x}_u, \mathbf{x}_i) \in Q(\mathbf{x}_u, \mathbf{x}_i, O_{ui})$
  Generate pseudo-labels Y'_{ui} for each sample: $(\mathbf{x}_u, \mathbf{x}_i, Y'_{ui})$
  $(\phi_n, \psi_n) \leftarrow (\phi_{n-1}, \psi_{n-1}) - \eta(\nabla_\phi \mathcal{L}, \nabla_\psi \mathcal{L})$
  $\theta_n \leftarrow \theta_{n-1} + \eta \nabla_\theta \mathcal{L}$
Return θ, ψ, ϕ

Complexity. As shown in Figure 1, compared with other unbiased learning algorithms [26, 36, 50], we introduce only one linear additional head for the critic which reuses embeddings obtained from the encoder. This suggests that our AST only introduces very few parameters and the model complexity is at the same level as other unbiased learning algorithms [6, 26, 36, 50].

6 EXPERIMENT

In this section, we empirically evaluate the performance of the proposed AST on both real-world and semi-synthetic datasets.

6.1 Experimental Settings

6.1.1 Datasets. Following previous works [26, 37, 38, 49, 50], we use two real-world datasets: **Yahoo** [31] and **Coat** [38]. These two datasets are suitable for verifying our theoretical analysis and evaluating our AST since they contain both biased and unbiased data, where unbiased data is formed by randomly assigning items to users for ratings. Thus they can be used to measure the unbiased generalization performance with selection bias. **Yahoo**[1]: Its biased training set has approximately 300,000 five-star ratings of 1,000 songs from 15,400 users. It collects an unbiased test set by asking 5,400 users to rate 10 randomly displayed songs. **Coat**[2]: It has 290 users and 300 items. Each user rates 24 items by themselves forming 6,500 biased five-star ratings, and is asked to rate 16 uniformly displayed items as the unbiased set. Since these two real-world datasets are relatively small, we also generate a relatively large semi-synthetic dataset based on **Goodreads**[3]. It is a book recommendation dataset [45] and we use the book review subset in history and biography, containing 238,450 users, 302,346 items, and 2,066,193 five-star ratings.

6.1.2 Prepossessing. Following [7, 26, 50]. For all datasets, we treat rating which is 3 or higher as positive feedback and the others as negative. For GoodReads, we remove those items and users that have less than 20 interactions.

6.1.3 Splitting and Simulation Settings. For Coat and Yahoo, following [7, 26], we treat all biased $\mathcal{D}_P$ data as training set, and split the unbiased data into three parts: 5% as additional training set $\mathcal{D}_Q$ to help training, 5% as validation set, and the remaining 90% as test set. Since Goodreads does not contain an unbiased testing set, we simulate a semi-synthetic dataset to facilitate ground-truth evaluation against a fully known relevance and exposure parameter. Strictly following previous works [37, 59], we first hold out the last feedback of all users in the last time slice as the test data and the feedback before the last is treated as the validation set. The remaining feedback serves as training set. We train a MF model to approximate the rating matrix by minimizing the mean-squared loss based on the training set. Then ground-truth preference probability for unbiased testing is $p(Y_{u,i} = 1|O_{u,i}) := \sigma(\hat{\mathbb{E}}[R_{u,i}|O_{u,i}] + \epsilon_R)$ where $\hat{\mathbb{E}}[R_{u,i}|O_{u,i}]$ is the model output and ϵ_R is Gaussian noisy controlling randomness of preference caused by unobserved confounders. Then, similar to [37, 59], we utilize another logistic MF predicting if the rating is observed as the exposure $\hat{p}(O_{u,i})$ model. The final log-exposure probability $\log p(O_{u,i}) = \log \hat{p}(O_{u,i}) + \epsilon_O$, where ϵ_O measures the extra randomness of exposure by unobserved confounders. In our experiments, we set ϵ_O and ϵ_R as five [37]. Following the generative model in § 3, we generate the biased training feedback as $p(Y_{u,i} = 1) = p(Y_{u,i} = 1|O_{u,i})p(O_{u,i})$. With this simulation process, we can obtain the true relevance, exposure parameters and unobserved confounders for the unbiased evaluation.

6.1.4 Setup. We compare AST with the following learning objectives: direct supervised training (Biased), IPS [37, 38], DRJL [48], CVIB [50], ACL [59], ATT [36], KD [26], and AutoDebias [7]. Since our AST is high-level learning approach that is compatible with almost all existing recommendation models, we consider two representative recommendation models: matrix factorization (MF) [21] and neural collaborative filtering (NCF) [19]. Following previous works [36, 48, 59], we utilize Hit Ratio (HR)@5 and NDCG@5 to evaluate the unbiased ranking performance. For all methods, the hyper-parameter search space is: dropout {0.2, 0.4, 0.6}, learning rate {0.001, 0.005, 0.01}, weight-decay {1e-4, 1e-5, 1e-6}, embedding dimension {64, 128, 256}. Specifically, for AST, we further search α, β, and γ from space {0.2, 0.4, 0.6, 0.8}. For a rigorous and fair comparison, we use the grid search to find the best hyperparameters of the baselines for all methods based on the validation performance.

6.2 Unbiased Learning Performance

Table 2 presents the unbiased learning results of AST and the baselines with NCF and MF as backbones, respectively. Observations from the table are as follows:

- Consistent with our theoretical analysis, AST significantly outperforms other algorithms, demonstrating its strong generalization ability. This is attributed to AST effectively minimizing the generalization bound of the ideal risk.
- Overall, AST consistently outperforms other baselines on all datasets using both MF and NCF backbones. This indicates the effectiveness of AST and showcases its flexibility and robustness across different backbones.
- Despite IPS and DR having strong theoretical insights, their empirical performance is poor. In contrast, AST demonstrates empirical effectiveness while maintaining rigorous justification.
- AST outperforms baselines on the Goodreads dataset, which exhibits both selection bias and unobserved exposure factors. This demonstrates AST's ability to simultaneously account for selection bias and the unconfoundedness assumption, resulting in tighter generalization bounds.

[1] https://webscope.sandbox.yahoo.com/
[2] https://www.cs.cornell.edu/~schnabts/mnar/
[3] https://sites.google.com/eng.ucsd.edu/ucsdbookgraph

Table 2: Unbiased learning performance of different algorithms with standard deviation over five runs. The best and second best performance are marked with boldface and underline, respectively.

Algorithms	Yahoo				Coat				Goodreads			
	MF		NCF		MF		NCF		MF		NCF	
	HR@5	NDCG@5	HR@5	NDCG@5	HR@5	NDCG@5	HR@5	NDCG@5	HR@5	NDCG@5	HR@5	NDCG@5
Biased	0.6471 (±0.0035)	0.6542 (±0.0037)	0.6352 (±0.0029)	0.6584 (±0.0017)	0.4338 (±0.0051)	0.6457 (±0.0072)	0.4281 (±0.0045)	0.6257 (±0.0048)	0.3071 (±0.0024)	0.1057 (±0.0011)	0.3214 (±0.0029)	0.1089 (±0.0008)
IPS	0.6598 (±0.0047)	0.6661 (±0.0052)	0.6415 (±0.0038)	0.6663 (±0.0029)	0.4131 (±0.0064)	0.6361 (±0.0079)	0.4255 (±0.0056)	0.6219 (±0.0050)	0.3156 (±0.0038)	0.1108 (±0.0027)	0.3462 (±0.0041)	0.1152 (±0.0018)
DRJL	0.6632 (±0.0038)	0.6732 (±0.0042)	0.6581 (±0.0033)	0.6716 (±0.0025)	0.4255 (±0.0040)	0.6378 (±0.0049)	0.4391 (±0.0023)	0.6381 (±0.0027)	0.3237 (±0.0034)	0.1255 (±0.0021)	0.3531 (±0.0034)	0.1265 (±0.0012)
CVIB	0.6756 (±0.0042)	0.6834 (±0.0047)	0.6635 (±0.0036)	0.6873 (±0.0027)	0.4531 (±0.0039)	0.6680 (±0.0034)	0.4487 (±0.0029)	0.6498 (±0.0033)	0.3467 (±0.0025)	0.1397 (±0.0026)	0.3687 (±0.0038)	0.1469 (±0.0015)
ATT	0.6635 (±0.0044)	0.6784 (±0.0049)	0.6497 (±0.0037)	0.6829 (±0.0023)	0.4371 (±0.0040)	0.6349 (±0.0037)	0.4357 (±0.0022)	0.6358 (±0.0024)	0.3307 (±0.0035)	0.1209 (±0.0026)	0.3562 (±0.0030)	0.1343 (±0.0014)
ACL	0.6801 (±0.0040)	0.6839 (±0.0045)	0.6522 (±0.0032)	0.6857 (±0.0022)	0.4529 (±0.0036)	0.6721 (±0.0033)	0.4631 (±0.0021)	0.6536 (±0.0029)	0.3587 (±0.0030)	0.1477 (±0.0025)	0.3714 (±0.0034)	0.1498 (±0.0015)
KD	0.6779 (±0.0043)	0.6781 (±0.0044)	0.6571 (±0.0031)	0.6814 (±0.0024)	0.4561 (±0.0038)	0.6584 (±0.0036)	0.4451 (±0.0024)	0.6471 (±0.0021)	0.3533 (±0.0029)	0.1368 (±0.0023)	0.3669 (±0.0035)	0.1405 (±0.0012)
AutoDebias	0.6835 (±0.0046)	0.6959 (±0.0051)	0.6609 (±0.0035)	0.6925 (±0.0028)	0.4628 (±0.0042)	0.6651 (±0.0037)	0.4568 (±0.0028)	0.6587 (±0.0035)	0.3608 (±0.0041)	0.1428 (±0.0026)	0.3751 (±0.0038)	0.1518 (±0.0016)
AST	**0.6985** (±0.0041)	**0.7147** (±0.0046)	**0.6813** (±0.0030)	**0.7094** (±0.0021)	**0.4775** (±0.0037)	**0.6819** (±0.0035)	**0.4728** (±0.0023)	**0.6630** (±0.0031)	**0.3712** (±0.0036)	**0.1678** (±0.0024)	**0.3834** (±0.0033)	**0.1655** (±0.0012)

Table 3: Ablation study (NDCG@5) with MF backbone.

Methods	Coat	Yahoo	Goodreads
AST w/o A	0.6628 (±0.0039)	0.7025 (±0.0043)	0.1498 (±0.0039)
AST w/o S	0.6693 (±0.0029)	0.7114 (±0.0038)	0.1615 (±0.0022)
AST w/o E	0.6733 (±0.0041)	0.7104 (±0.0052)	0.1545 (±0.0031)
AST w/o S & E	0.6587 (±0.0032)	0.6978 (±0.0039)	0.1317 (±0.0025)
Biased	0.6457 (±0.0072)	0.6542 (±0.0037)	0.1057 (±0.0011)
AutoDebias	0.6651 (±0.0037)	0.6959 (±0.0051)	0.1428 (±0.0026)
AST	**0.6819** (±0.0035)	**0.7147** (±0.0046)	**0.1678** (±0.0024)

Table 4: Performance (NDCG@5) without unbiased data.

	Yahoo		Coat	
	MF	NCF	MF	NCF
Biased	0.6533	0.6714	0.6205	0.6330
IPS	0.6661	0.6756	0.6147	0.6440
DRJL	0.6673	0.6789	0.6433	0.6376
ATT	0.6778	0.6788	0.6332	0.6472
CVIB	0.6717	0.6906	0.6529	0.6519
AST	**0.6898**	**0.7004**	**0.6712**	**0.6589**

- As per our theoretical analysis, KD and AutoDebias show performance improvements by utilizing unbiased uniform data. However, as shown in Table 2, Our AST consistently outperforms them by a significant margin.

6.3 Ablation Study and Parameter Sensitivity

Setup: To conduct a detailed analysis of how different components impact AST performance, we perform an ablation study and parameter sensitivity analysis. We follow the same setup as described in § 6.2 and construct the following variants of AST: (i) AST without the adversarial matching component (AST w/o A); (ii) AST without self-training (AST w/o S); (iii) AST without entropy minimization (AST w/o E) and (iv) AST without both self-training and entropy minimization (AST w/o S & E).

Table 5: Performance (NDCG@5) on implicit feedback.

	Yahoo		Coat	
	MF	NCF	MF	NCF
Biased	0.6914	0.6233	0.5514	0.6373
IPS	0.7011	0.6484	0.5458	0.6144
DRJL	0.7025	0.6517	0.5833	0.6181
ACL	0.7097	0.6885	0.5921	0.6348
KD	0.7152	0.6758	0.5692	0.6214
AutoDebias	0.7195	0.6742	0.5873	0.6388
AST	**0.7248**	**0.7026**	**0.6037**	**0.6631**

Results: The ablation study results are presented in Table 3. We observe that all the designed components contribute to performance improvements, and their contributions are complementary to each other. We also investigate the sensitivity of the hyperparameters α and β, where α and β control the contribution of adversarial matching and self-training, respectively. As the trend of γ is similar to β, we omit it for brevity. We vary α and β as $[0.2, 0.4, 0.6, 0.8]$ and report the results in Fig. 2. Key findings are as follows: (**i**) AST performs better and exhibits stability when $\alpha \in 0.6, 0.8$ and $\beta \in 0.2, 0.6$, simplifying the process of hyperparameter selection. (**ii**) By varying α and β, we can achieve a balanced trade-off between adversarial matching and self-training, leading to improved generalization performance. This confirms the motivation behind jointly mitigating selection bias and unobserved confounders, as finding a suitable trade-off enhances the transferability of biased embeddings for better unbiased performance.

6.4 Performance on Challenging Scenarios

In this subsection, we consider two more challenging scenarios. The first scenario is debiasing without any unbiased training data, which is realistic as collecting unbiased data is typically expensive. We compare AST with baselines that do not require unbiased training data. Following the methodology in [7, 26], we treat the biased data

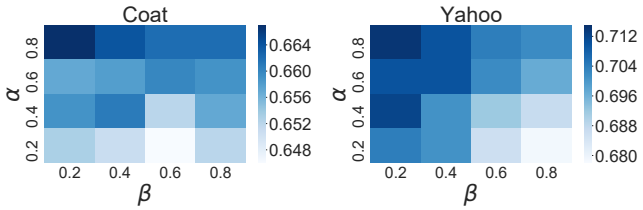

Figure 2: Sensitivity Analysis (NDCG@5) using MF.

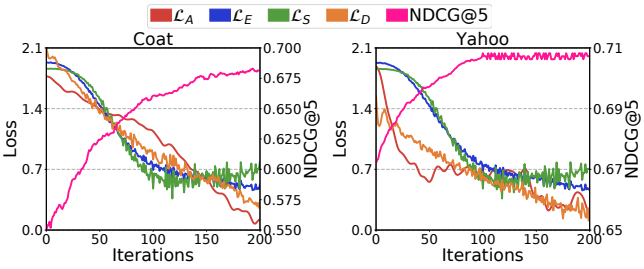

Figure 3: Generalization performance and training losses.

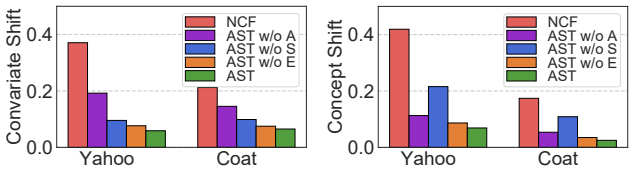

Figure 4: Marginal and conditional shifts on Yahoo and Coat.

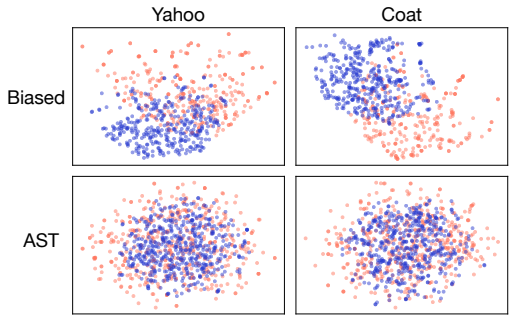

Figure 5: Visualization fo embeddings of AST and Biased method. The blue points correspond to the biased training data, while red ones correspond to unbiased testing data.

as the training set and randomly sample 5% of the ratings from the unbiased test data as the validation set. The results are reported in Table 4. We observe that AST achieves the best performance compared to the baselines. In particular, AST outperforms ATT and CVIB, further demonstrating the effectiveness of adversarial feature adaptation using the KL-divergence (Theorem 5.1).

The second scenario is implicit feedback. Implicit feedback is more challenging than explicit feedback since we do not have negative evidence in the learning process [37, 59]. We evaluate AST on this scenario as well. To generate implicit feedback, we use the Yahoo and Coat datasets but remove the negative feedback from the training data. The results are presented in Table 5. We observe that AST outperforms all baselines, indicating its ability to effectively mitigate selection bias in implicit feedback data. This aligns with our theoretical analysis, as the ideal risk can still be bounded under the setting of implicit feedback.

6.5 Deeper Understanding of AST

Setup. We conduct a detailed analysis of AST to gain insights into its behavior. We follow the same setup as described in Section 6.2.

6.5.1 Generalization and Convergence. . To examine the generalization and convergence of AST, we plot the training loss curves of different components and the testing NDCG on two datasets in Figure 3. We make the following observations: **(i)** AST exhibits training stability and consistently improves the unbiased testing performance as iterations progress. **(ii)** The NDCG metric shows a nearly monotonic increase with iterations, suggesting that minimizing our loss, which is an upper bound of the ideal loss, is an effective approach to improve accuracy based on biased feedback.

6.5.2 Unobserved Confounders. . One of the key strengths of AST is its ability to mitigate both marginal and conditional shifts caused by unobserved confounders. Therefore, we investigate whether AST possesses this capability and identify the most important component contributing to it. Figure 4 displays the empirically calculated $\mathcal{A}$-distance [30] and MDD [24] using the learned embeddings of AST. The $\mathcal{A}$-distance measures covariate shift, while MDD quantifies concept shift. Our findings from Figure 4 are as follows: **(i)** The $\mathcal{A}$-distance and MDD values of AST embeddings are significantly smaller than those of vanilla NCF, indicating that AST can more effectively reduce both covariate and concept distribution shifts. **(ii)** AST without the D component exhibits a smaller $\mathcal{A}$-distance than AST without the A component, while AST without the A component has a smaller MDD than AST without the D component. This observation aligns with our idea that adversarial matching minimizes covariate shift, while self-training alleviates concept shift. To gain further intuition about feature adaptation, we visualize the t-SNE embeddings sampled from $P(\boldsymbol{z}ui)$ and $Q(\boldsymbol{z}ui)$. From Figure 5, we observe that AST effectively bridges the feature gap between biased and unbiased data, whereas biased training fails as the embeddings are separated and have a certain distance.

7 CONCLUSIONS

In this paper, we studied the problem of unbiased recommendation. We provided a novel perspective on the distribution shift for the unbiased recommendation problem. We derived several generalization bounds and presented both theoretical and algorithmic analyses of current learning algorithms. We also proposed the AST algorithm, which effectively addresses the issues of selection bias and unobserved confounders. Extensive experiments on three datasets with various settings demonstrated the effectiveness of AST. While our results strongly advocate for considering unobserved confounders in unbiased recommendation, optimizing them directly in the real world poses a challenge. Exploring more effective optimization methods is an interesting topic that requires further investigation.

ACKNOWLEDGMENTS

This material is based upon work supported by, or in part by the National Science Foundation (NSF) under grants number IIS-1909702, Army Research Office (ARO) under grant number W911NF-21-1-0198, and Cisco Faculty Research Award.

REFERENCES

[1] Heejung Bang and James M Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics*.

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning*.

[3] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. 2007. Analysis of representations for domain adaptation. *NIPS*.

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019. MixMatch: A Holistic Approach to Semi-Supervised Learning. *NIPS*.

[5] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. 2007. Learning Bounds for Domain Adaptation. *NIPS*.

[6] Stephen Bonner and Flavian Vasile. 2018. Causal embeddings for recommendation. *RecSys*.

[7] Jiawei Chen, Hande Dong, Yang Qiu, Xiangnan He, Xin Xin, Liang Chen, Guli Lin, and Keping Yang. 2021. AutoDebias: Learning to Debias for Recommendation. *SIGIR*.

[8] Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. 2020. Self-training avoids using spurious features under domain shift. *NIPS* (2020), 21061–21071.

[9] Zhihong Chen, Rong Xiao, Chenliang Li, Gangfeng Ye, Haochuan Sun, and Hongbo Deng. 2020. Esam: Discriminative domain adaptation with non-displayed items to improve long-tail performance. *SIGIR*.

[10] Imre Csiszár and János Körner. 2011. *Information theory: coding theorems for discrete memoryless systems*.

[11] Sihao Ding, Peng Wu, Fuli Feng, Yitong Wang, Xiangnan He, Yong Liao, and Yongdong Zhang. 2022. Addressing Unmeasured Confounder for Recommendation with Sensitivity Analysis. In *KDD*.

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

[13] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. *ICML*.

[14] Yves Grandvalet and Yoshua Bengio. 2004. Semi-supervised learning by entropy minimization. *NIPS*.

[15] Siyuan Guo, Lixin Zou, Yiding Liu, Wenwen Ye, Suqi Cheng, Shuaiqiang Wang, Hechang Chen, Dawei Yin, and Yi Chang. 2021. Enhanced Doubly Robust Learning for Debiasing Post-Click Conversion Rate Estimation. *SIGIR*.

[16] Zhimeng Guo, Teng Xiao, Charu Aggarwal, Hui Liu, and Suhang Wang. 2023. Counterfactual Learning on Graphs: A Survey. *arXiv preprint arXiv:2304.01391* (2023).

[17] Shantanu Gupta, Hao Wang, Zachary Lipton, and Yuyang Wang. 2021. Correcting Exposure Bias for Link Recommendation. *ICML*.

[18] Xiangnan He, Zhankui He, Xiaoyu Du, and Tat-Seng Chua. 2018. Adversarial personalized ranking for recommendation. In *SIGIR*.

[19] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. *WWW*.

[20] Fredrik D Johansson, David Sontag, and Rajesh Ranganath. 2019. Support and invertibility in domain-invariant representations. In *AISTATS*.

[21] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer*.

[22] Adit Krishnan, Ashish Sharma, Aravind Sankar, and Hari Sundaram. 2018. An adversarial approach to improve long-tail performance in neural collaborative filtering. *CIKM*.

[23] Sanjay Krishnan, Jay Patel, Michael J Franklin, and Ken Goldberg. 2014. A methodology for learning, analyzing, and mitigating social influence bias in recommender systems. *RecSys*.

[24] Jingjing Li, Erpeng Chen, Zhengming Ding, Lei Zhu, Ke Lu, and Heng Tao Shen. 2020. Maximum density divergence for domain adaptation. *PAMI*.

[25] Zinan Lin, Dugang Liu, Weike Pan, and Zhong Ming. 2021. Transfer Learning in Collaborative Recommendation for Bias Reduction. *RecSys*.

[26] Dugang Liu, Pengxiang Cheng, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. 2020. A general knowledge distillation framework for counterfactual recommendation via uniform data. *SIGIR*.

[27] Dugang Liu, Pengxiang Cheng, Zinan Lin, Jinwei Luo, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. 2022. KDCRec: Knowledge Distillation for Counterfactual Recommendation Via Uniform Data. *TKDE* (2022).

[28] Dugang Liu, Pengxiang Cheng, Hong Zhu, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. 2021. Mitigating Confounding Bias in Recommendation via Information Bottleneck. *RecSys*.

[29] Weiming Liu, Jiajie Su, Chaochao Chen, and Xiaolin Zheng. 2021. Leveraging Distribution Alignment via Stein Path for Cross-Domain Cold-Start Recommendation. *NeurIPS*.

[30] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2018. Conditional adversarial domain adaptation. *NIPS*.

[31] Benjamin M Marlin, Richard S Zemel, Sam Roweis, and Malcolm Slaney. 2007. Collaborative filtering and the missing at random assumption. *UAI*.

[32] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. 2016. f-gan: Training generative neural samplers using variational divergence minimization. *NIPS*.

[33] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*.

[34] Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *journal of educational Psychology*.

[35] Noveen Sachdeva, Yi Su, and Thorsten Joachims. 2020. Off-policy bandits with deficient support. *KDD*.

[36] Yuta Saito. 2020. Asymmetric Tri-training for Debiasing Missing-Not-At-Random Explicit Feedback. *SIGIR*.

[37] Yuta Saito, Suguru Yaginuma, Yuta Nishino, Hayato Sakata, and Kazuhide Nakata. 2020. Unbiased recommender learning from missing-not-at-random implicit feedback. *WSDM*.

[38] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. *ICML*.

[39] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. 2017. Learning from simulated and unsupervised images through adversarial training. *CVPR*.

[40] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. 2019. Meta-weight-net: Learning an explicit mapping for sample weighting. *NIPS*.

[41] Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. 2020. Doubly robust off-policy evaluation with shrinkage. *ICML*.

[42] Adith Swaminathan and Thorsten Joachims. 2015. The self-normalized estimator for counterfactual learning. *NIPS*.

[43] James Victor Uspensky. 1937. Introduction to mathematical probability.

[44] Vladimir N Vapnik. 1999. An overview of statistical learning theory. *IEEE transactions on neural networks*.

[45] Mengting Wan and Julian McAuley. 2018. Item recommendation on monotonic behavior chains. In *RecSys*.

[46] Qi Wan, Xiangnan He, Xiang Wang, Jiancan Wu, Wei Guo, and Ruiming Tang. 2022. Cross Pairwise Ranking for Unbiased Item Recommendation. In *WWW*.

[47] Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang. 2017. Irgan: A minimax game for unifying generative and discriminative information retrieval models. In *SIGIR*.

[48] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2019. Doubly robust joint learning for recommendation on data missing not at random. *ICML*.

[49] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2021. Combating Selection Biases in Recommender Systems with a Few Unbiased Ratings. *WSDM*.

[50] Zifeng Wang, Xi Chen, Rui Wen, Shao-Lun Huang, Ercan Kuruoglu, and Yefeng Zheng. 2020. Information Theoretic Counterfactual Learning from Missing-Not-At-Random Feedback. *NIPS*.

[51] Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. 2020. Theoretical Analysis of Self-Training with Deep Networks on Unlabeled Data. In *ICLR*.

[52] Chenwang Wu, Defu Lian, Yong Ge, Zhihao Zhu, Enhong Chen, and Senchao Yuan. 2021. Fight fire with fire: towards robust recommender systems via adversarial poisoning training. In *SIGIR*. 1074–1083.

[53] Chuhan Wu, Fangzhao Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. 2021. Fairness-aware news recommendation with decomposed adversarial learning. In *AAAI*. 4462–4469.

[54] Teng Xiao, Zhengyu Chen, Donglin Wang, and Suhang Wang. 2021. Learning how to propagate messages in graph neural networks. In *KDD*. 1894–1903.

[55] Teng Xiao, Zhengyu Chen, and Suhang Wang. 2022. Representation Matters When Learning From Biased Feedback in Recommendation. In *CIKM*. 2220–2229.

[56] Teng Xiao, Shangsong Liang, and Zaiqiao Meng. 2019. Hierarchical neural variational model for personalized sequential recommendation. In *WWW*.

[57] Teng Xiao and Donglin Wang. 2021. A general offline reinforcement learning framework for interactive recommendation. In *AAAI*.

[58] Teng Xiao and Suhang Wang. 2022. Towards unbiased and robust causal ranking for recommender systems. In *WSDM*. 1158–1167.

[59] Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2020. Adversarial Counterfactual Learning and Evaluation for Recommender System. *NIPS*.

[60] Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2021. Rethinking Neural vs. Matrix-Factorization Collaborative Filtering: the Theoretical Perspectives. *ICML*.

[61] Longqi Yang, Yin Cui, Yuan Xuan, Chenyang Wang, Serge Belongie, and Deborah Estrin. 2018. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. *RecSys*.

[62] Feng Yuan, Lina Yao, and Boualem Benatallah. 2019. DARec: Deep domain adaptation for cross-domain recommendation via transferring rating patterns. *arXiv*.

[63] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. 2019. On learning invariant representations for domain adaptation. *ICML*.

[64] Ziwei Zhu, Yun He, Yin Zhang, and James Caverlee. 2020. Unbiased Implicit Recommendation and Propensity Estimation via Combinational Joint Learning. *RecSys*.

A THE LEMMAS

Before we conduct the proof, we first state the following Lemmas:

Lemma A.1. *[5]. Let $\mathcal{H}$ be a hypothesis space of VC-dimension d, and for any distribution P and Q over $\mathcal{X}_u \times \mathcal{X}_i$, then $\forall h, h' \in \mathcal{H}$:*

$$|\mathcal{L}_P(h, h') - \mathcal{L}_Q(h, h')| \leq \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(P, Q), \tag{20}$$

where $d_{\mathcal{H}\Delta\mathcal{H}}(P, Q) = 2 \sup_{h,h' \in \mathcal{H}} |\mathbb{E}_P[\ell(h(\mathbf{x}_u, \mathbf{x}_i), h'(\mathbf{x}_u, \mathbf{x}_i))] - \mathbb{Q}[\ell(h(\mathbf{x}_u, \mathbf{x}_i), h'(\mathbf{x}_u, \mathbf{x}_i))]$.

Lemma A.2. *[44]. Let S is a arbitrarily data distribution and $\mathcal{H}$ be a hypothesis space of VC-dimension d. Then $\forall h \in \mathcal{H}, \forall \delta > 0$, w.p. at least $1 - \delta$ over the a sample size N and natural exponential e:*

$$\mathcal{L}_S(h) \leq \widehat{\mathcal{L}}_S(h) + \sqrt{\frac{4}{N}(d \log \frac{2eN}{d} + \log \frac{4}{\delta})}. \tag{21}$$

B PROOF OF THEOREM 4.2

Proof. Following the definitions in § 3, we have:

$$|\mathcal{L}_P(f) - \mathcal{L}_Q(f)| = |\mathcal{L}_P(f, k) - \mathcal{L}_Q(f, g)|, \tag{22}$$

which has the following upper bound:

$$\begin{aligned} |\mathcal{L}_P(f, k) - \mathcal{L}_Q(f, g)| &= |\mathcal{L}_P(f, k) - \mathcal{L}_P(f, g) + \mathcal{L}_P(f, g) - \mathcal{L}_Q(f, g)| \\ &\leq |\mathcal{L}_P(f, k) - \mathcal{L}_P(f, g)| + |\mathcal{L}_P(f, g) - \mathcal{L}_Q(f, g)| \\ &= |\mathbb{E}_P[|f(\mathbf{x}_u, \mathbf{x}_i) - k(\mathbf{x}_u, \mathbf{x}_i)| - |f(\mathbf{x}_u, \mathbf{x}_i) - g(\mathbf{x}_u, \mathbf{x}_i)|]| \\ &+ |\mathcal{L}_P(f, g) - \mathcal{L}_Q(f, g)| \leq \mathbb{E}_P[|k(\mathbf{x}_u, \mathbf{x}_i) - g(\mathbf{x}_u, \mathbf{x}_i)|] + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(P, Q), \end{aligned} \tag{23}$$

where we utilize the triangular inequality and Lemma A.1. Similarly, due to the symmetric property, the following inequality for Q holds:

$$|\mathcal{L}_P(f) - \mathcal{L}_Q(f)| = |\mathcal{L}_P(f, k) - \mathcal{L}_Q(f, g)| \leq \mathbb{E}_Q[|k(\mathbf{x}_u, \mathbf{x}_i) - g(\mathbf{x}_u, \mathbf{x}_i)|] + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(P, Q). \tag{24}$$

Combine the inequalities (23) and (24) above, we have:

$$\mathcal{L}_Q(f) \leq \mathcal{L}_P(f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(P, Q) + \min\{\mathbb{E}_P[|k(\mathbf{x}_u, \mathbf{x}_i) - g(\mathbf{x}_u, \mathbf{x}_i)|], \mathbb{E}_Q[|k(\mathbf{x}_u, \mathbf{x}_i) - g(\mathbf{x}_u, \mathbf{x}_i)|]\}. \tag{25}$$

Combining Eqs. (24), (25) and Lemma A.2, and considering the hypothesis $f(\mathbf{x}_u, \mathbf{x}_i)$ is composed of a two parts: $f = h \circ \phi$ where h is the hypothesis and ϕ maps $(\mathbf{x}_u, \mathbf{x}_i)$ to $\mathbf{z}_{ui}$. W.p. at least $1 - \delta$:

$$\begin{aligned} \mathcal{L}_Q(h \circ \phi) &\leq \widehat{\mathcal{L}}_P(h \circ \phi) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(P(\mathbf{z}_{ui}), Q(\mathbf{z}_{ui})) \\ &+ \min\{\mathbb{E}_{P(\mathbf{z}_{ui})}[|\tilde{k}(\mathbf{z}_u) - \tilde{g}(\mathbf{z}_{ui})|], \mathbb{E}_{Q(\mathbf{z}_{ui})}[|\tilde{k}(\mathbf{z}_{ui}) - \tilde{g}(\mathbf{z}_{ui})|]\} + \sqrt{\frac{4}{N}\left(d \log \frac{2eN}{d} + \log \frac{4}{\delta}\right)}, \end{aligned} \tag{26}$$

where $P(\mathbf{z}_{ui})$ (resp. $Q(\mathbf{z}_{ui})$) is the probability density functions over $\mathcal{Z}$ induced by $P(\mathbf{x}_u, \mathbf{x}_i, O_{ui})$ (resp. $Q(\mathbf{x}_u, \mathbf{x}_i, O_{ui})$) and ϕ. The latent labeling function induced by g and ϕ: $\tilde{g} = \int_{\phi^{-1}_{(z)}} g(x)p(x)dx / \int_{\phi^{-1}_{(z)}} p(x)dx$ where x denotes the features [20]. Similarly, $\tilde{k} = \int_{\phi^{-1}_{(z)}} k(x)p(x)dx / \int_{\phi^{-1}_{(z)}} p(x)dx$. With Lemma A.2 and $f = h \circ \phi$, we also have, w.p. at least $1 - \delta$:

$$\mathcal{L}_Q(h \circ \phi) \leq \widehat{\mathcal{L}}_Q(h \circ \phi) + \sqrt{\frac{4}{M}(d \log \frac{2eM}{d} + \log \frac{4}{\delta})}. \tag{27}$$

By combining Eq. (27) with Eq. (26) over coefficients ρ and $1 - \rho$, respectively, we have:

$$\begin{aligned} &\mathcal{L}_Q(h \circ \phi) \leq \rho \widehat{\mathcal{L}}_P(h \circ \phi) + (1 - \rho)\widehat{\mathcal{L}}_Q(h \circ \phi) + \frac{\rho}{2} d_{\mathcal{H}\Delta\mathcal{H}}(P(\mathbf{z}_{ui}), Q(\mathbf{z}_{ui})) + \rho \min\{\mathbb{E}_{P(\mathbf{z}_{ui})}[|\tilde{g}(\mathbf{z}_{ui}) - \tilde{k}(\mathbf{z}_{ui})|], \mathbb{E}_{Q(\mathbf{z}_{ui})}[|\tilde{g}(\mathbf{z}_{ui}) - \tilde{k}(\mathbf{z}_{ui})|]\} + \\ &\rho\sqrt{\frac{4}{N}(d \log \frac{2eN}{d} + \log \frac{4}{\delta})} + (1 - \rho)\sqrt{\frac{4}{M}(d \log \frac{2eM}{d} + \log \frac{4}{\delta})}, \end{aligned} \tag{28}$$

which completes the whole proof. □

C PROOF OF THEOREM 4.3

Proof. We derive the upper bound between the expected error $\mathcal{L}_Q(h)$ and the empirical error $\widehat{\mathcal{L}}_Q(h(w))$ via the meta validation. Specifically, we define $\epsilon_i(h(w)) = \mathcal{L}_Q(h) - \ell(h(w)(\mathbf{x}_u, x_i), Y_{ui})$ for $h(w) \in \mathcal{H}'$ and every data sample in $(\mathbf{x}_u, \mathbf{x}_i, Y_{ui}) \in \mathcal{D}_Q$. Then, we have:

$$\mathcal{L}_Q(h) - \widehat{\mathcal{L}}_Q(h(w)) = \frac{1}{M}\sum_{m=1}^{M}\epsilon_m(h(w)). \tag{29}$$

Since $\mathcal{L}_Q(h) \in [0,1]$ and $\ell(h(w)(\mathbf{x}_u, x_i), y) \in [0,1]$, we have $\mathcal{L}_Q(h) - \ell(h(w)(\mathbf{x}_u, x_i), y) \in [-1,1]$, $\mathbb{E}[\epsilon_m(h(w))^2] \le 1$, and $|\epsilon_m(h(w))| \le 1$. Based on the Bernstein inequality [43], we have:

$$p(\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(h(w)) > \xi) \le \exp(-\frac{\xi^2 M/2}{1+\xi/3}). \tag{30}$$

Taking the union bound of this inequality over all $h(w) \in \mathcal{H}'$ has:

$$p(\cup_{h(w)\in\mathcal{H}'}\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(h(w)) > \xi) \le M^{d'}\exp(-\frac{\xi^2 M/2}{1+\xi/3}). \tag{31}$$

Let $\delta = M^{d'}\exp(-\frac{\xi^2 M/2}{1+\xi/3})$. Solving the above Inequality (31) for ξ yields the following result (note that $\xi \ge 0$):

$$\xi = \frac{d'\log M - \log\delta}{3M} \pm \sqrt{(\frac{d'\log M - \log\delta}{3M})^2 + \frac{2(d'\log M - \log\delta)}{M}} \le \frac{d'\log M - \log\delta}{3M} + \sqrt{\frac{2(d'\log M - \log\delta)}{M}} \because \sqrt{a+b} \le \sqrt{a} + \sqrt{b}. \tag{32}$$

Thus, for any $\delta > 0$, with probability at least $1-\delta$, for $h' \in \mathcal{H}'$,

$$\mathcal{L}_Q(h) \le \widehat{\mathcal{L}}_Q(h(w)) + \frac{d'\log M - \log\delta}{3M} + \sqrt{\frac{2(d'\log M - \log\delta)}{M}}. \tag{33}$$

Similar to Eq. (28), by furthering considering the above bound in the latent feature space via ϕ and combining it with Eq. (26) over coefficients $1-\rho$ ρ and ρ respectively, we complete the final proof. □

D PROOF OF THEOREM 5.1

Proof. We show that the ideal risk $\mathcal{L}_Q(f) = \mathcal{L}_{Q_z}(h)$ can bounded as (note that we denote $P(\mathbf{z}_{ui})$ $(Q(\mathbf{z}_{ui}))$ as P_z (Q_z) for brevity):

$$\begin{aligned}
\mathcal{L}_Q(f) = \mathcal{L}_{Q_z}(h) &= \mathcal{L}_{Q_z}(h) - \mathcal{L}_{P_z}(h,\tilde{k}) + \mathcal{L}_{P_z}(h,\tilde{k}) - \mathcal{L}_{P_z}(h) + \mathcal{L}_{P_z}(h)\\
&\le \mathcal{L}_{P_z}(h) + |\mathcal{L}_{P_z}(h) - \mathcal{L}_{P_z}(h,\tilde{k})| + |\mathcal{L}_{Q_z}(h) - \mathcal{L}_{P_z}(h,\tilde{k})|\\
&= \mathcal{L}_{P_z}(h) + |\mathbb{E}_{P_z}[|h(z) - \tilde{g}(z)| - |h(z) - \tilde{k}(z)|]| + |\mathcal{L}_{Q_z}(h) - \mathcal{L}_{P_z}(h,\tilde{k})|\\
&\le \mathcal{L}_{P_z}(h) + \mathbb{E}_{P_z}[|\tilde{g}(z) - \tilde{k}(z)|] + |\mathcal{L}_{Q_z}(h) - \mathcal{L}_{P_z}(h,\tilde{k})|\\
&\le \mathcal{L}_{P_z}(h) + \mathbb{E}_{P_z}[|h(z) - \tilde{k}(z)|] + \int |P_z - Q_z|\cdot|h(z) - \tilde{k}(z)|dz\\
&\le \mathcal{L}_{P_z}(h) + \mathbb{E}_{P_z}[|\tilde{g}(z) - \tilde{k}(z)|] + \int |P_z - Q_z|dz \because h(z) - \tilde{k}(z) \in [0,1]\\
&= \mathcal{L}_{P_z}(h) + \mathbb{E}_{P_z}[|\tilde{g}(z) - \tilde{k}(z)|] + \mathrm{TV}(P_z||Q_z) \le \mathcal{L}_{P_z}(h) + \mathbb{E}_{P_z}[|\tilde{g}(z) - \tilde{k}(z)|] + \sqrt{2\mathrm{KL}(P_z\|Q_z)},
\end{aligned} \tag{34}$$

where we used triangular inequality multi-times and the Pinsker's inequality [10] in the last line. $h(z) - \tilde{k}(z) \in [0,1]$ since our loss is 0-1 binary loss. Due to the the symmetric property, we also have:

$$\mathcal{L}_{Q_z}(h) \le \mathcal{L}_{P_z}(h) + \mathbb{E}_{Q_z}[|\tilde{g}(z) - \tilde{k}(z)|] + \sqrt{2\mathrm{KL}(P_z\|Q_z)}, \tag{35}$$

Combining Eqs. (34), (35) and Lemmas A.2, we have:

$$\mathcal{L}_Q(h\circ\phi) \le \widehat{\mathcal{L}}_P(h\circ\phi) + \frac{1}{2}\sqrt{2\mathrm{KL}(P(\mathbf{z}_{ui})\|Q(\mathbf{z}_{ui}))} + \min\{\mathbb{E}_{P(\mathbf{z}_{ui})}[|\tilde{g}(\mathbf{z}_{ui}) - \tilde{k}(\mathbf{z}_{ui})|], \mathbb{E}_{Q(\mathbf{z}_{ui})}[|\tilde{g}(\mathbf{z}_{ui}) - \tilde{k}(\mathbf{z}_{ui})|]\} + \sqrt{\frac{4}{N}(d\log\frac{2eN}{d} + \log\frac{4}{\delta})}$$

By summing this bound with Eq. (27) over coefficients ρ and $1-\rho$, respectively, we have:

$$\begin{aligned}
\mathcal{L}_Q(h\circ\phi) \le{}& \rho\widehat{\mathcal{L}}_P(h\circ\phi) + (1-\rho)\widehat{\mathcal{L}}_Q(h\circ\phi) + \frac{\rho\sqrt{2\mathrm{KL}(P(\mathbf{z}_{ui})\|Q(\mathbf{z}_{ui}))}}{2}\\
&+ \rho\min\{\mathbb{E}_{P(\mathbf{z}_{ui})}[|\tilde{g}(\mathbf{z}_{ui}) - \tilde{k}(\mathbf{z}_{ui})|], \mathbb{E}_{Q(\mathbf{z}_{ui})}[|\tilde{g}(\mathbf{z}_{ui}) - \tilde{k}(\mathbf{z}_{ui})|]\} + (1-\rho)\sqrt{\frac{4}{M}(d\log\frac{2eM}{d} + \log\frac{4}{\delta})} + \rho\sqrt{\frac{4}{N}(d\log\frac{2eN}{d} + \log\frac{4}{\delta})}.
\end{aligned} \tag{36}$$

Thus, the proof is completed. □