Towards Bridging Algorithm and Theory for Unbiased Recommendation

Teng Xiao¹, Zhengyu Chen², Suhang Wang¹
¹The Pennsylvania State University, ²Zhejiang University
{tengxiao,szw494}@psu.edu, chenzhengyu@zju.edu.cn,

Abstract

This work studies the problem of learning unbiased algorithms from biased feedback for recommender systems. We address this problem from both theoretical and algorithmic perspectives. Recent works in unbiased learning have advanced the state-of-the-art with various techniques such as meta-learning, knowledge distillation, and information bottleneck. Despite their empirical successes, most of them lack theoretical guarantee, forming non-negligible gaps between the theories and recent algorithms. To this end, we first view the unbiased recommendation problem from a distribution shift perspective. We theoretically analyze the generalization bounds of unbiased learning and suggest their close relations with recent unbiased learning objectives. Based on the theoretical analysis, we further propose a principled framework, Adversarial Self-Training (AST), for unbiased recommendation. Empirical evaluation on real-world and semi-synthetic datasets demonstrate the effectiveness of the proposed AST.

1 Introduction

Recommender systems have been widely used in many applications such as e-commerce platforms, social networks, and healthcare. However, recent studies show that it suffers from the selection bias issue [48, 33, 15]. The ideal feedback is collected by randomly and uniformly exposing items to users. However, the exposures are affected by the past recommendation policy, and this is called 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 these items than others [48, 33]. This system selection bias creates the rich gets richer phenomenon where head contents are getting more and more exposure while tail contents are not discovered. Selection bias also comes from user self-selection: users usually interact and rate items they like and rarely rate items they do not like [26, 33]. Previous studies [33, 41] have theoretically and empirically shown that directly learning from the biased feedback cannot reflect user true preferences.

Remarkable theoretical advances have been proposed in the unbiased recommendation. Specifically, [Schnabel et al., 33] and [Wang et al., 41] provided rigorous unbiased generalization bounds under the selection bias. On par with their theoretical findings, there are rich advances in unbiased recommendation [33, 41, 32, 48] based on inverse propensity score (IPS) [28] and doubly robust (DR) [1] in causal inference. Although IPS and DR can address the selection bias in theory, these solutions typically make the unconfoundedness assumption [46], i.e., user preference over items is independent of getting exposed given the feature [32, 33, 46], which is impractical and cannot be examined in many real-world RS. Moreover, they need to estimate propensity score for re-weighting and suffer from huge variance [37, 31] when the propensity score is small. Thus IPS and DR empirically perform badly compared to many recent works [22, 7, 42, 43] for unbiased learning.

Many unbiased recommendation algorithms thus have been introduced to conduct de-biasing learning with various machine learning techniques such as casual embedding [6], meta-learning [7, 42], knowledge distillation [22], and information bottleneck [43], which achieve promising empirical performance. However, rigorous theoretical analysis of these algorithms are severely lacking in the literature, which led to a disconnect between current theory and many strong empirical methods. Specifically, most of these methods [22, 32, 7, 6] solve the bias issue by introducing unbiased uniform data in the training, which is collected by a random logging policy. Nonetheless, to the best of our knowledge, no clear connection

Preprint. Under review.
between current theory and these algorithms has been made, i.e., unbiased learning generalization bounds for them have not been derived. More importantly, there is no solid theoretical justification of why utilizing unbiased uniform data can improve learning performance. Table 1 shows an overview of the discussed methods and suggests that most of them do not have theoretical guarantees. Hence, the above problem poses the following challenge: How to bridge the gap between theories and recent state-of-the-art unbiased learning objectives? Furthermore, Could we propose a more effective unbiased learning objectives for recommendation guided by rigorous theoretical justification?

1.1 Our Contributions

In this paper, we provide answers to the above research questions. We first revisit the unbiased recommendation from the distribution shift perspective, and then present a theoretical analysis of unbiased learning towards explicit guidance and explanation for current algorithm design. Our analysis shows that many unbiased learning objectives essentially optimize different terms in our bound. Unlike existing bounds [33, 41], our bounds explicitly suggest accounting for the unconfoundedness assumption, which is important as this assumption may not hold in the real-world. We provide important insights that our theoretical generalization bounds allows us to understand why and how unbiased uniform data helps to improve unbiased learning performance. Inspired by our theoretical analysis and insights, 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 demonstrate the effectiveness of AST.

Due to page limits, all proofs and a detailed discussion of related work, are given in the Appendix.

2 Preliminaries

2.1 Notations and Selection Bias

Let \( x_u \in \mathcal{X}_U \) be the feature vector for user \( u \in \{1, \ldots, |U|\} \), \( x_i \in \mathcal{X}_I \) be the feature vector for item \( i \in \{1, \ldots, |I|\} \). The feature vectors can be user/item one-hot encoding or embedding. \( \mathcal{X}_U \) and \( \mathcal{X}_I \) are the feature spaces, respectively. Following previous works, 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 \( R_{u,i} \in \{0, 1\} \) 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 \( R_{u,i} \). Let \( D_P = \{x_u, y_i, O_{u,i} | O_{u,i} = 1\} \) be the logged feedback and the number of samples is \( N \). The task of unbiased recommendation is to infer unobserved preference \( R_{u,i} \). Typically, the collected feedback follows a generative process:

\[
\begin{align*}
p(x_u, x_i, Y_{u,i}) &= p(x_u)p(x_i)p(R_{u,i}, O_{u,i} = 1|x_u, x_i) = \\
p(x_u)p(x_i)p(O_{u,i} = 1|x_u, x_i)p(R_{u,i}|O_{u,i} = 1, x_u, x_i) \cdot Y_{u,i} = O_{u,i} \cdot R_{u,i},
\end{align*}
\]

where the exposure distribution \( p(O_{u,i} = 1|x_u, x_i) \) makes the feedback be MNAR (we will drop \( 1 \) for all \( O_{u,i} \) in the remainder of the paper for conciseness). This exposure distribution \( p(O_{u,i}|x_u, x_i) \) 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., \( O_{u,i} \perp (R_{u,i}, x_u, x_i) \),

\[
\mathcal{L}_Q(f) \triangleq \mathbb{E}_Q[\ell(f(x_u, x_i), g(x_u, x_i))]
\]

where \( Q \triangleq p(x_u)p(x_i)p(O_{u,i}) \) with \( p(O_{u,i}) = 1 \) for all user-item pairs \( 33, 46, 47 \). \( f(x_u, x_i) \) is the estimated hypothesis. \( g(x_u, x_i) = p(R_{u,i}|x_u, x_i) \) is the optimal labeling function, depending on the true preference distribution \( p(R_{u,i}|x_u, x_i) \). \( Q \) is called as the marginal distribution over features. Typically, \( \ell(f(x_u, x_i), g(x_u, x_i)) \) is the 0-1 loss, which is the probability that \( f \) disagrees with \( g \) under \( Q \): \( \mathbb{E}_Q[\mathbb{I}(f(x_u, x_i) \neq g(x_u, x_i))] \).

In this paper, we conduct theoretical analysis based on 0-1 loss. But, in practice, we can use 0-1 log loss \( \ell(x, y) = -\log \sigma(y) - (1 - x) \log(1 - \sigma(y)) \) with \( \sigma(y) = 1/(1 + e^{-y}) \) which serves as a effective convex proxy for 0-1 loss.

---

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 theoretical framework |
|---------------------|-------------------------------|--------------------------|-------------------|-----------------------------|
| Re-weighting [14, 44, 19] | X | | X | |
| Information bottleneck [4, 13] | | | | |
| Multi-task learning [22, 26] | | | | |
| Meta-learning [4, 14] | | | | |
| Adversarial self-training | | | | |

---

**2 Preliminaries**

**2.1 Notations and Selection Bias**

Let \( x_u \in \mathcal{X}_U \) be the feature vector for user \( u \in \{1, \ldots, |U|\} \), \( x_i \in \mathcal{X}_I \) be the feature vector for item \( i \in \{1, \ldots, |I|\} \). The feature vectors can be user/item one-hot encoding or embedding. \( \mathcal{X}_U \) and \( \mathcal{X}_I \) are the feature spaces, respectively. Following previous works, 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 \( R_{u,i} \in \{0, 1\} \) 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 \( R_{u,i} \). Let \( D_P = \{x_u, y_i, O_{u,i} | O_{u,i} = 1\} \) be the logged feedback and the number of samples is \( N \). The task of unbiased recommendation is to infer unobserved preference \( R_{u,i} \). Typically, the collected feedback follows a generative process:

\[
\begin{align*}
p(x_u, x_i, Y_{u,i}) &= p(x_u)p(x_i)p(R_{u,i}, O_{u,i} = 1|x_u, x_i) = \\
p(x_u)p(x_i)p(O_{u,i} = 1|x_u, x_i)p(R_{u,i}|O_{u,i} = 1, x_u, x_i) \cdot Y_{u,i} = O_{u,i} \cdot R_{u,i},
\end{align*}
\]

where the exposure distribution \( p(O_{u,i} = 1|x_u, x_i) \) makes the feedback be MNAR (we will drop \( 1 \) for all \( O_{u,i} \) in the remainder of the paper for conciseness). This exposure distribution \( p(O_{u,i}|x_u, x_i) \) 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., \( O_{u,i} \perp (R_{u,i}, x_u, x_i) \),

\[
\mathcal{L}_Q(f) \triangleq \mathbb{E}_Q[\ell(f(x_u, x_i), g(x_u, x_i))]
\]

where \( Q \triangleq p(x_u)p(x_i)p(O_{u,i}) \) with \( p(O_{u,i}) = 1 \) for all user-item pairs \( 33, 46, 47 \). \( f(x_u, x_i) \) is the estimated hypothesis. \( g(x_u, x_i) = p(R_{u,i}|x_u, x_i) \) is the optimal labeling function, depending on the true preference distribution \( p(R_{u,i}|x_u, x_i) \). \( Q \) is called as the marginal distribution over features. Typically, \( \ell(f(x_u, x_i), g(x_u, x_i)) \) is the 0-1 loss, which is the probability that \( f \) disagrees with \( g \) under \( Q \): \( \mathbb{E}_Q[\mathbb{I}(f(x_u, x_i) \neq g(x_u, x_i))] \).

In this paper, we conduct theoretical analysis based on 0-1 loss. But, in practice, we can use 0-1 log loss \( \ell(x, y) = -\log \sigma(y) - (1 - x) \log(1 - \sigma(y)) \) with \( \sigma(y) = 1/(1 + e^{-y}) \) which serves as an effective convex proxy for 0-1 loss.
We observe that this assumption is actually referred to as unconfoundedness assumption \[29\] in causal inference. We can notice that the unbiased risk function in Eq. (2) is independent of the exposure distribution of all user-item pairs, \(P(O_{ui}) = 1\), rather than the exposure distribution \(p(O_{ui}|x_u, x_i)\). This uniform exposure scenario is ideal because the preference will not be affected by the previous exposure, thus leading to a 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 selection 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:

\[
\text{Training}: p(x_u)p(x_i)p(O_{ui}|x_u, x_i)p(R_{ui}|O_{ui}, x_u, x_i) \quad \text{Testing}: p(x_u)p(x_i)p(O_{ui})p(R_{ui}|x_u, x_i). \quad (3)
\]

Thus, the empirical risk \(\hat{L}_P(f)\) over logged feedback \(\mathcal{D}_P\) is an biased estimate of the ideal risk:

\[
\hat{L}_P(f) = \frac{1}{N} \sum_{(x_u, x_i, y_{ui}) \in \mathcal{D}_P} \ell(f(x_u, x_i), y_{ui}) \approx L_P(f) \neq L_Q(f),
\]

where \(L_P(f) \triangleq L_P(f, k) = \mathbb{E}_P[\ell(f(x_u, x_i), k(x_u, x_i))]\), \(P = p(x_u)p(x_i)p(O_{ui} = 1|x_u, x_i)\) is marginal distribution. \(k(x_u, x_i) = p(R_{ui} = R_{ui}|O_{ui}, x_u, x_i)\), is the optimal labeling function depending on preference distribution \(p(R_{ui}|x_u, x_i, O_{ui})\) on training. Thus, the learned \(f\) will not be approximately optimal even having sufficiently large training data \[33\].

### 2.2 The Unconfoundedness Assumption

To deal with this selection bias, many de-biasing methods \[33, 32, 41\] inspired by causal inference algorithms such as IPS and DR have been proposed. As mentioned by previous works \[32, 46\], these algorithms assume that being relevant is independent of getting exposed given the feature, i.e. \(R_{ui} \perp O_{ui}|x_u, x_i\):

\[
p(R_{ui}|O_{ui}, x_u, x_i) = p(R_{ui}|x_u, x_i). \quad (5)
\]

We observe that this assumption is actually referred to as unconfoundedness assumption \[29\] 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 (propensity) probability (see Eq. (3)) and the conditional distribution shift between \(p(R_{ui}|x_u, x_i)\) and \(p(R_{ui}|O_{ui}, x_u, x_i)\) vanishes (i.e., labeling function \(g(x_u, x_i) = k(x_u, x_i)\)). Thus, these methods \[33, 32, 41, 51\] can conduct unbiased estimation by inversely re-weighting each sample:

\[
\hat{L}_w(f) = \frac{1}{N} \sum_{(x_u, x_i, y_{ui}) \in \mathcal{D}_P} \frac{1}{p(O_{ui}|x_u, x_i)} \ell(f(x_u, x_i), y_{ui}). \quad (6)
\]

It is straightforward to verify that \(\hat{L}_w(f)\) is an unbiased estimation of ideal risk: \(\mathbb{E}[\hat{L}_w(f)] = L_Q(f)\). Clearly, this re-weighting objective can theoretically correct for the distribution shift caused by the exposure if \(p(O_{ui} = 1|x_u, x_i)\) is known in advance. Note that DR \[41\] is also built on this re-weighting objective although it has an additional imputation model. We just focus on the re-weighting part of it here. While this objective has theoretical guarantee \[41, 32, 33\], there are three crucial directions for improvement:

1. The unconfoundedness assumption may not be true and cannot be examined in the real recommendation scenarios \[46\], unless we can collect every single factor that may affect the users’ decision making process into feature. However, there always be other unobserved confounders such as user social influence, item popularity effect, public opinions that are not attainable from features. For instance, as shown in \[20\], user ratings follow different distributions when users rate items before or after reading the public opinions. Nevertheless, current methods \[41, 32, 33\] do not account for this unobserved confounders.

2. The theoretical analysis on this re-weighting objective \[33, 41, 32\] can not explain and generalize to many unbiased algorithms, especially for those \[6, 7, 42, 22\] utilizing unbiased uniform data.

3. This objective needs to accurately estimate exposure probability which is usually difficult \[31, 46\] and suffers from the huge variance. Thus, it empirically perform badly compared to recent algorithms \[7, 43, 22\].

### 3 Theoretical Analysis

In this section, we first present our framework on the unbiased recommendation from the distribution shift perspective and derive the two finite-sample generalization bounds. We provide the key insight that our theoretical framework is able to unify a series of recent unbiased learning objectives \[6, 22, 43, 7, 42, 23\].
3.1 Unbiased Learning via Feature Adaptation

In this section, we show how the feature adaptation in domain adaptation [12] related to the unbiased recommendation. Recall that we have logged feedback $D_P$ from distribution $P(x_u, x_i, O_u)$, where $P = P(x_u, x_i, O_u)$ is the training marginal distribution over features. Similarly, we have the testing marginal distribution $Q = p(x_u)p(x_i)p(O_u) = 1/|\mathcal{O}|$, meaning $(x_u, x_i)$ is sampled i.i.d. from uniform exposure distribution. Our goal is to learn a function $f(x_u, x_i)$ which can approximate the optimal function $g(x_u, x_i)$ which depends on preference distribution $p(R_{ui}|x_u, x_i)$.

To show how recent unbiased algorithms [22, 7, 43] related to the feature adaptation, without loss of generality, we further consider the hypothesis $f(x_u, x_i)$ is composed of two parts: $f = h \circ \phi$ where $\phi \in \Phi \subset \{\phi: \mathcal{X}_u \times \mathcal{X}_i \rightarrow \mathcal{Z}\}$ is the feature mapping and $h \in \mathcal{H} \subset \{h: \mathcal{Z} \rightarrow \mathcal{Y}\}$ is the hypothesis of the classification head. In general, $h$ is the linear or feed-forward network predictor. Given this, we notice that Ben-David et al. [2] 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 $\hat{\mathcal{L}}_P(h \circ \phi)$ and the discrepancy between the training and testing distributions:

**Theorem 1.** [2, 5] Let $\mathcal{H}$ be a hypothesis space with VC-dimension $d$. $P(z_{ui})$ (resp. $Q(z_{ui})$) is the distribution over $Z$ induced by marginal distribution $P(x_u, x_i, O_u)$ (resp. $Q(x_u, x_i, O_u)$) and $\phi$. Then, with probability (w.p) at least $1 - \delta$ over the natural exponential $e, \forall h \in \mathcal{H}$:

$$\mathcal{L}_Q(h \circ \phi) \leq \hat{\mathcal{L}}_P(h \circ \phi) + \frac{1}{2}d_{H\Delta H}(P(z_{ui}), Q(z_{ui}))+\lambda(\phi)+\sqrt{\frac{4}{N}(d \log \frac{2eN}{d}+\log \frac{4}{\delta})},$$

where $d_{H\Delta H}(P(z_{ui}), Q(z_{ui})) = 2\sup_{h_1, h_2 \in \mathcal{H}}|E_P[z_{ui}][\ell(h(z_{ui}), h'(z_{ui}))] - E_Q[z_{ui}][\ell(h(z_{ui}), h'(z_{ui}))]|$ is the $\mathcal{H} \setminus \mathcal{H}$-divergence [25] 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 1** shows that the ideal risk $\mathcal{L}_Q(h \circ \phi)$ depends on three terms, which include the empirical risk $\hat{\mathcal{L}}_P(h \circ \phi)$, the divergence between $P(z_{ui})$ and $Q(z_{ui})$, and the combined risk $\lambda(\phi)$. This bound serves as the theoretical foundation and has inspired the feature adaptation methods [12, 44], which simultaneously minimizes the divergence between $P(z_{ui})$ and $Q(z_{ui})$, and loss $\hat{\mathcal{L}}_P(h \circ \phi)$.

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 $P(x_u, x_i, O_u)$ and $Q(x_u, x_i, O_u)$ by using latent feature adaptation, however it does not theoretically reflect the unjustifiable unconfoundedness assumption as shown in §2.2, namely the conditional distribution shift. This will make the upper bound loose when the unconfoundedness assumption is violated in the real-world. (2) This bound still cannot give the guidance and explanation for unbiased learning objectives [6, 22, 42] that utilize unbiased uniform data. In what follows, we will introduce two generalization bounds to provide interpretations for these learning objectives based on multi-task learning and meta-learning.

3.2 Unbiased Multi-Task Learning Bound

In this section, we give a unbiased multi-task learning bound which measures the unconfoundedness assumption. We also demonstrate that a series of existing unbiased recommendation algorithms [6, 22, 43, 25, 32, 33] including those using unbiased uniform data can be interpreted by this bound.

Specifically, some recent algorithms [6, 22] conduct the de-biasing learning via the unbiased uniform data which is collected by a random exposure probability $Q$ and can reflect user preference in an unbiased way. Thus, besides the biased data $D_P$, we assume that we have some unbiased uniform data $D_Q = \{x_u, x_i, Y_{ui}|O_{ui} = 1\}$ and the number of samples is $M$. Given the combined biased and unbiased data, these algorithms [6, 22, 43, 23, 32, 33] generally have the following empirical multi-task learning objective:

$$\rho\hat{\mathcal{L}}_P(h \circ \phi) + (1 - \rho)\hat{\mathcal{L}}_Q(h \circ \phi) + \alpha R(\hat{P}(z_{ui}), \hat{Q}(z_{ui})), \rho \in [0, 1],$$

where $\rho = 1$ means that we do not have unbiased uniform data $D_Q$, thus this formulation can unify those algorithms [43, 23, 33] without using unbiased data. $R$ is the regularization function, and $\hat{P}(z_{ui})$, and $\hat{Q}(z_{ui})$ are empirical distributions of latent features over $P$ and $Q$, respectively. $\alpha$ is the hyper-parameters. $\hat{\mathcal{L}}_Q(h \circ \phi) = \frac{1}{M} \sum_{(x_u, x_i, Y_{ui}) \in D_Q} \ell(h \circ \phi(x_u, x_i), Y_{ui})$ is the empirical objective under unbiased uniform data $D_Q$. Based on this, we provide the following generalization bound:

**Theorem 2.** Let $\mathcal{H}$ be a hypothesis space with VC-dimension $d$, and $P(z_{ui})$ (resp. $Q(z_{ui})$) is the probability density functions over $Z$ induced by $P(x_u, x_i, O_{ui})$ (resp. $Q(x_u, x_i, O_{ui})$) and $\phi$, $\tilde{g}$ (resp. $\tilde{k}$) is the labeling
function over $Z$ induced by $g$ (resp. $k$) and $\phi$. Then, w.p. at least $1 - \delta$ over the exponential $e$, $\forall h \in \mathcal{H}$:

$$
\mathcal{L}_Q(h \circ \phi) \leq \rho \mathcal{L}_P(h \circ \phi) + (1 - \rho) \mathcal{L}_Q(h \circ \phi) + P \frac{d_{\Delta \mathcal{H}}(P(\mathcal{Z}_{ui}), Q(\mathcal{Z}_{ui}))}{2} + \rho \min \{E_P(\mathcal{Z}_{ui}) \| \hat{g}(\mathcal{Z}_{ui}) - \hat{k}(\mathcal{Z}_{ui}) \|, E_Q(\mathcal{Z}_{ui}) \| \hat{g}(\mathcal{Z}_{ui}) - \hat{k}(\mathcal{Z}_{ui}) \| \} + (1 - \rho) \sqrt{\frac{4}{M} (d \log 2 e M + \log \frac{4}{\delta})}.
$$

(9)

Remark. The proofs are provided in Appendix [13]. This bound suggests that the ideal risk depends on the empirical multi-task learning error, the divergence of feature distributions, and the distance $\min \{E_P(\mathcal{Z}_{ui}) \| \hat{g}(\mathcal{Z}_{ui}) - \hat{k}(\mathcal{Z}_{ui}) \|, E_Q(\mathcal{Z}_{ui}) \| \hat{g}(\mathcal{Z}_{ui}) - \hat{k}(\mathcal{Z}_{ui}) \| \}$ of labeling functions which is essentially the divergence between conditional distributions $p(R_{ui}|O_{ui}, x_i, \mathbf{x}_i)$ and $p(\mathcal{Z}_{ui}|x_i, \mathbf{x}_i)$. This bound explicitly considers this assumption and suggests that if it is violated, i.e., $p(\mathcal{Z}_{ui}|x_i, \mathbf{x}_i) \neq p(\mathcal{Z}_{ui}|O_{ui}, x_i, \mathbf{x}_i)$, $\delta$ will be loose. Thus, we should guarantee that the conditional distributions are not too far away from each other for successful unbiased recommendation.

This bound enables us to interpret many de-biasing objectives [33, 43, 22, 6, 32, 41] in a unified perspective. Specifically, we show that they all fall into the multi-task learning objective in Eq. (8) and approximately minimize the regularization term in the generalization bound.

Re-weighting objectives [32, 33]. These methods fall into the multi-task learning objective in Eq. (8) with $\rho = 1$ and $\alpha = 0$ since they do not utilize unbiased data and regularization. They re-weight the distribution $P$ via $w(x_i, \mathbf{x}_i) = 1/p(O_{ui}|x_i, \mathbf{x}_i)$. With setting $\phi$ as the identity function, it is easy to verify that the first term in the bound becomes the re-weighting objective in Eq. (5) and the third becomes $d_{\Delta \mathcal{H}}(w(x_i, \mathbf{x}_i)P(\mathcal{Z}_{ui}|x_i, O_{ui}), Q(\mathcal{Z}_{ui}|x_i, O_{ui}))$ which equals to zero. Thus, they essentially minimize the first and third term in this generalization bound with $\rho = 1$ and $\phi$ being identity function.

Information bottleneck objectives [43, 23]. This kind of algorithm also falls into the multi-learning objective with $\rho = 1$ and $\alpha \neq 0$. The regularization term in Eq. (5) is instantiated as the information bottleneck to regularize the model to learn an invariant representation across training and testing distributions [43, 23], which makes the $d_{\Delta \mathcal{H}}(P(\mathcal{Z}_{ui}), Q(\mathcal{Z}_{ui}))$ smaller. Thus, similar to the re-weighting objectives, the information bottleneck objectives also essentially minimize the first and third term in our bound.

Multi-task objectives [22, 6]. These algorithms utilize unbiased uniform data and have the regularization term for approximately reducing the divergence between $P(\mathcal{Z}_{ui})$ and $Q(\mathcal{Z}_{ui})$, thus $\rho \neq 1$ and $\alpha \neq 0$ in Eq. (5). Specifically, $R$ is $\|z_{ui} - \hat{z}_{ui}\|_2$ in [6] where $z_{ui}$ and $\hat{z}_{ui}$ are sampled from $P(\mathcal{Z}_{ui})$ and $Q(\mathcal{Z}_{ui})$, respectively. [22] 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.

3.3 Unbiased Meta-Learning Bound

As an alternative, there are some algorithms such as Learning to Debias [42] and AutoDebias [7] utilizing the unbiased uniform data via a meta-learning process. Typically, their objectives are still re-weighting objectives. However, unlike vanilla IPS or DR, they utilize the unbiased uniform data to train a weight function $w \in \mathcal{H}' \subset \{ w : \mathcal{X}_d \times \mathcal{X}_d \rightarrow \mathcal{W} \}$ such that the hypothesis $h$ trained on the biased data performs well on the unbiased data. Specifically, the meta-learning can be formulated as a bi-level optimization:

$$
\min_w \mathcal{L}_Q(h(w)) \text{s.t. } h(w) = \arg \min_h \mathcal{L}_{\mathcal{P}_w}(h),
$$

(10)

where $P_w = w(x_i, \mathbf{x}_i)P(\mathcal{Z}_{ui}|O_{ui})$ stands for a new re-weighted distribution. $\mathcal{L}_Q(h(w) \circ \phi)$ is the upper-level objective under unbiased uniform data. $w$ and $h$ are optimized alternately until convergence. Note that, in these methods, the $h(w)$ is the function of weight and its hypothesis space is $\mathcal{H}'$ depending on biased training data due to the bi-level optimization [35]. Empirically, this meta-learning objective empirical perform well [7, 32]. To theoretically understand them, we provide the following bound:

**Theorem 3.** Let $\mathcal{H}$ (resp. $\mathcal{H}'$) be hypothesis spaces with VC-dimension $d$ (resp. $d'$). $P(\mathcal{Z}_{ui})$ (resp. $Q(\mathcal{Z}_{ui})$) is the density functions over $Z$ induced by $P(\mathcal{Z}_{ui}, x_i, O_{ui})$ (resp. $Q(\mathcal{Z}_{ui}, x_i, O_{ui})$) and $\phi$. $\hat{g}$ (resp. $\hat{k}$) is the
We have shown how our framework allows us to reinterpret many learning objectives in unbiased recommen-
dation. In this paper, we consider the Fenchel-dual form of the KL-divergence \cite{27}.

Let

\begin{equation}
    \rho \min_h (\mathbb{E}_P(h \circ \phi) + (1 - \rho) \mathbb{E}_Q(h \circ \phi) + \frac{\rho}{2} d_{\mathcal{H} \mathcal{H}}(w(x_u, x_i)P(z_{ui}), Q(z_{ui}))
\end{equation}

+ \rho \min_h (\mathbb{E}_P(h \circ \phi) + (1 - \rho) \mathbb{E}_Q(h \circ \phi) + \frac{\rho}{2} d_{\mathcal{H} \mathcal{H}}(w(x_u, x_i)P(z_{ui}), Q(z_{ui}))
\end{equation}

\[+
(1 - \rho) \left( \frac{d \log M - \log \delta}{3M} + \sqrt{\frac{2(d \log M - \log \delta)}{M}} \right)\].

Remark. We provide the proofs in Appendix [3]. 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. Unlike Theorem [2], this empirical
error on the unbiased data is obtained via a meta validation process.

Meta-learning objectives \cite{7, 42}. Given this bound, we can understand why recent meta-learning ap-
proaches can achieve good performance and provide interpretations for them. It is worth noting that the
bi-level meta-learning objective in Eq. (10) exactly minimizes the first and second terms when \( \phi \) is the
identity function. Unlike re-weighting objectives, the \( w(x_u, x_i) \) may not be the optimal sample weight,
i.e., \( 1/p(O_{ui}|x_u, x_i) \). Thus, we can not theoretically guarantee that the third divergence term is small.

Meta-learning objectives also make the unconfoundedness assumptions and neglect the fourth term in the
bound.

4 Adversarial Self-Training

We have shown how our framework allows us to reinterpret many learning objectives in unbiased recom-
dendation. 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 [2] and [3]. Nevertheless, the unconfoundedness assumption is
rarely true and can not be examined in the real-world \cite{46}. (2) Some algorithms try to minimize the \( \mathcal{H} \mathcal{H} \) diver-
gence between marginal feature distributions via re-weighting \cite{33, 32} or different regularizers \cite{43, 22, 6}.

However re-weighting suffers from the variance issue \cite{36}. As for the regularizers \cite{43, 22, 6}, they are only
an approximation of the empirical \( \mathcal{H} \mathcal{H} \)-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 \cite{42, 7, 35}.

To address these issues, we exploit the theoretic analysis introduced in §3 to derive a novel and practical
unbiased learning algorithm. We propose the adversarial self-training (AST) algorithm, which can simultane-
ously alleviate the divergence of feature distributions and approximately account for unobserved confounders
by minimizing a new generalization bounds with KL-divergences in an end-end process.

4.1 Adversarial Training for Adaptation

Motivated by the discussion in §3, we need to design a mechanism to minimize the \( \mathcal{H} \mathcal{H} \)-divergence. How-
ever it is difficult to optimize it. Thus, we give a new generalization bound to guide the design of AST:

**Theorem 4.** Let \( \mathcal{H} \) be a space with VC-dimension \( d \). \( P(z_{ui}) \) (resp. \( Q(z_{ui}) \)) is the distribution over \( \mathcal{Z} \) induced by marginal distribution \( P(x_u, x_i, O_{ui}) \) (resp. \( Q(x_u, x_i, O_{ui}) \)) and \( \phi \), \( \hat{g} \) (resp. \( \hat{k} \)) is the latent
labeling function induced by \( g \) (resp. \( k \)) and \( \phi \). Then, w.p. at least \( 1 - \delta \) over the exponential \( \rho \), \( \forall \ h \in \mathcal{H}: 

\begin{align}
\mathcal{L}_Q(h \circ \phi) & \leq \rho \mathcal{L}_P(h \circ \phi) + (1 - \rho) \mathcal{L}_Q(h \circ \phi) + \frac{\rho}{2} \sqrt{2KL(P(z_{ui})||Q(z_{ui}))}
\end{align}

\[+
(1 - \rho) \left[ \frac{4}{M} \left( d \log \frac{2eM}{d} + \log \frac{4}{\delta} \right) + \rho \sqrt{\frac{4}{N} \left( d \log \frac{2eN}{d} + \log \frac{4}{\delta} \right)} \right].
\end{align}

Remark. The proofs are provided in Appendix [4]. This bounds provides theoretical justification for the use
of KL (Kullback–Leibler)-divergence to conduct the feature adaptation in unbiased recommendation. While
the explicit marginal densities of \( P(z_{ui}) \) and \( Q(z_{ui}) \) are intractable, we have data samples of them. This
motivates leveraging adversarial distribution matching strategies \cite{27} to minimize KL-divergence through a
mini-max game with samples. In particular, we minimize \( D_{KL}(P(z_{ui})||Q(z_{ui})) \) via the use of a critic
function (the max-step), and then update the feature mapping \( \phi \) accordingly to reduce the KL-divergence
the (min-step). In this paper, we consider the Fenchel-dual form of the KL-divergence \cite{27}:

\begin{align}
KL(P||Q) = \mathbb{E}_P[\log P - \log Q] = \max_{\nu > 0} \{ \mathbb{E}_P[\log \nu] - \mathbb{E}_Q[\nu] + 1 \}.
\end{align}

6
To optimize this Fenchel-dual form in practice, we model \( \log \nu \) as using another function \( \theta(z_{ui}) \) as our critic function. This results in the following adversarial neural estimator of \( \text{KL}(P(z_{ui}) || Q(z_{ui})) \):

\[
\hat{L}_A(\phi, \theta) = \min_{\phi} \max_{\theta} \mathbb{E}_{z_{ui}=\phi(x_u, x_i), (x_u, x_i) \sim P(x_u, x_i, O_{ui})} [\theta(z_{ui})] - \mathbb{E}_{z_{ui}=\phi(x_u, x_i), (x_u, x_i) \sim Q(x_u, x_i, O_{ui})} [\exp(\theta(z_{ui}))].
\]

(14)

Compared to \( \mathcal{H} \Delta \mathcal{H} \)-divergence, this is a much easy implementation to bound the ideal risk in Theorem[4]

4.2 Supervised Learning and Self-Training

As suggested by the generalization bound in Theorem[4] besides the KL-divergence, 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 \( h \) function:

\[
\hat{L}_D(\phi, \psi) = \mathbb{E}_{z_{ui}=\phi(x_u, x_i), (x_u, x_i) \sim D} [\ell(\psi(z_{ui}), Y_{ui})],
\]

(15)

where \( D = \mathcal{D}_P \cup \mathcal{D}_Q \) is the whole set of data, including the biased data the unbiased uniform data. Thus, our algorithm conduct de-biasing learning without unbiased uniform data when \( D = \mathcal{D}_P \). To further minimize the distance between labeling functions, we need to search for a feature mapping \( \phi \) such that the conditional distribution is invariant to training and testing: \( \mathbb{E}_P[Y_{ui} \mid \phi(x_u, x_i)] = \mathbb{E}_Q[Y_{ui} \mid \phi(x_u, x_i)] \) for every user-item pairs. If we also have a small amount of unbiased uniform data, then we can directly minimize this term. However, in some scenarios, collecting them is extraordinarily expensive [43, 22]. Thus directly optimizing this term becomes inaccessible. To account for this scenarios, in this paper, we propose to approximate evaluate and minimize this term by using pseudo-labels. Specifically, we adopt the principle of self-training, which has shown to be effective in semi-supervised learning [13, 4, 44]. Self-training first trains the model via \( \hat{L}_D(\phi, \psi) \) in Eq. (15), and model generates pseudo-labels for the unlabeled data sampled from \( Q(x_u, x_i, O_{ui}) \). Then self-training trains the feature mapping with pseudo-labels:

\[
\hat{L}_S(\phi) = \mathbb{E}_{z_{ui}=\phi(x_u, x_i), (x_u, x_i) \sim Q(x_u, x_i, O_{ui})} [\ell(\psi(z_{ui}), Y_{ui}^\text{bar})],
\]

(16)

where \( Y_{ui}^\text{bar} = \hat{\psi}(\hat{\phi}(x_u, x_i)) \) is the generated soft pseudo-label (it can be the ground-truth label if we have a small amount of unbiased uniform data). \( \psi \) and \( \hat{\phi} \) indicate that we do not propagate gradients through computing the pseudo labels. We empirically found that this self-training can effectively brings conditional distribution closer on our datasets even we do not have any unbiased uniform data. In addition, inspired by the recent work [8] which proves the entropy minimization has the similar effect as self-training algorithm. We also explicitly minimize the following entropy on unlabeled uniform data:

\[
\hat{L}_E(\phi, \psi) = \mathbb{E}_{z_{ui}=\phi(x_u, x_i), (x_u, x_i) \sim Q(x_u, x_i, O_{ui})} [H(\sigma(\psi(z_{ui})))].
\]

(17)

where \( H(X) = -\sum_i p(x_i) \log p(x_i) \) 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 [31].

In summary, the overall objective function of AST could be formulated as follows:

\[
\mathcal{L} = \min_{\phi, \psi, \theta} \max_{\phi, \theta} \hat{L}_D(\phi, \psi) + \alpha \hat{L}_A(\phi, \theta) + \beta \hat{L}_S(\phi) + \gamma \hat{L}_E(\phi, \psi),
\]

(18)

where \( \alpha, \beta \) and \( \gamma \) are hyper-parameters controlling the contributions of different losses. Our full algorithm, Adversarial Self-Training (AST), is given in Algorithm[1] and illustrated in Figure[5] in Appendix[8]

5 EXPERIMENT

In this section, we evaluate the performance of AST on both real-world and semi-synthetic datasets. Due the page limit, we show key results here and include additional experimental results in Appendix[1]

5.1 Experimental Settings

Datasets. Following [33, 32, 42, 43, 22], we use two real-world datasets: Yahoo [26] and Coat [33]. These two datasets are suitable for varying our theoretical analysis and evaluating our AST since (1) they contain unbiased data, where items are assigned randomly to users. Thus they can be used to measure the unbiased generalization performance with selection bias; (2) we also empirical find there are some conditional shift in Yahoo which might be caused by extra random perturbations of unobserved confounders (see Appendix[6]). To facilitate oracle evaluation against a fully known relevance, exposure and unobserved confounders, we also generate a semi-synthetic dataset based on the relatively large dataset Goodreads [40]. We provide the details, prepossessing and splitting of these datasets in Appendix[6].
Table 2: Unbiased learning performance. The best and second best performance are marked with boldface and underline, respectively. All results are averaged over five runs, and the improvement are statistically significant.

| Algorithms | Yahoo MF | Yahoo NCF | Coat MF | Coat NCF | Goodreads MF | Goodreads NCF |
|------------|----------|-----------|--------|----------|--------------|--------------|
| Biased     | 0.6533   | 0.6714    | 0.6205 | 0.6330   | 0.6914       | 0.6233       |
| IPS        | 0.6661   | 0.6756    | 0.6147 | 0.6440   | 0.7011       | 0.6484       |
| DRJL       | 0.6673   | 0.6789    | 0.6433 | 0.6376   | 0.7025       | 0.6517       |
| ATT        | 0.6778   | 0.6878    | 0.6332 | 0.6472   | 0.7097       | 0.6855       |
| CVIB       | 0.6717   | 0.6906    | 0.6529 | 0.6519   | 0.7152       | 0.6758       |
| AutoDebias | 0.6898   | 0.7004    | 0.6712 | 0.6589   | 0.7248       | 0.7026       |

5.2 Unbiased Learning Performance

Table 2 gives the unbiased learning results of AST and the baselines with NCF and MF backbone, respectively. From the table, we have the following observations: (1) As suggested by our theoretical analysis, our AST significantly outperforms other algorithms, which demonstrates its good generalization ability. This is because that our AST can effectively minimize the generalization bound of the ideal risk; (2) Overall, AST consistently outperforms other baselines on all datasets with both MF and NCF, which implies the effectiveness of AST and also shows that AST is flexible and robust to various recommendation models; (3) Although IPS and DR come with strong theoretical insights, they empirically perform not very good. In contrast, AST is empirically effective while maintaining the rigorous theoretical justification; (4) AST outperforms baselines on GOODREADS which has both selection bias and unobserved exposure factors, demonstrating that AST can simultaneously account for both selection bias and the unconfoundedness assumption, which makes the generalization bound smaller; (5) As suggested by our theoretic analysis, KD and AutoDebias can improve performance by utilizing the unbiased uniform data. However, as shown in Table 2, our AST still outperforms them by a large margin. We also evaluate the performance of AST and baselines under different percentage of unbiased uniform data, which is presented in Figure 1. We can find that AST significantly outperforms other methods designed for using unbiased data under all the data ration performed and the improvement increase when larger amount of uniform data is available. This is consistent with our theoretical results, since more unbiased data can make the generalization bound smaller.

5.3 Performance on Challenging Scenarios

In this section, we consider two more challenging scenarios. The first is de-biasing without any unbiased data. This scenarios is realistic since collecting unbiased data is expensive. We compare AST with the
baselines, which also do no need the unbiased data. Following [22, 7], we use some unbiased data for validation. We regard the biased data as training set and randomly sample 5% ratings from the unbiased test data as the validation set. The results are reported in Table 3. We can find that AST can still achieve the best performance compared to baselines. The results that our AST outperforms ATT and CVIB further shows the effectiveness of adversarial feature adaptation via the KL-divergence (i.e., Theorem 4).

The second scenario is the implicit feedback. Compared to explicit feedback, the implicit feedback is much more challenging since we do not have any negative evidence in the learning [32, 46]. Thus, we further evaluate AST on this scenario. To generate implicit feedback, we use the Yahoo and Coat datasets but remove the negative feedback in the training data. Table 4 gives the results. We find that AST outperforms baselines, which shows that AST can also effectively alleviate the selection bias issue for implicit feedback data. This matches our theoretical analysis since the ideal risk still can be bounded under the setting of implicit feedback.

5.4 Ablation Study

To take a deeper examination on how different components affect AST performance, we conduct ablation study. We follow the same setup in § 5.2 and build the following ablations of AST: (i) Removing adversarial matching component (AST w/o A); (ii) Removing self-training (AST w/o S); and (iii) removing entropy minimization (AST w/o E). Table 5 shows the ablation study results. We can find that our all designed components can contribute to the performance gain, and their contributions are complementary to each other.

We also investigate the sensitivity of hyperparameters $\alpha$ and $\beta$, which is presented in Appendix I.

5.5 Deeper Understanding of AST

**Distribution Shifts.** One of the most important properties of AST is to alleviate both marginal and conditional shifts. Thus, we investigate if the proposed AST has this ability and which component of AST plays the most important role. Figure 2 shows the empirically calculated $A$-distance [25] and MDD [21] by using the learned embeddings of AST. Note that $A$-distance and MDD can quantify covariate (marginal) and concept (conditional) shifts, respectively. From Figure 2, we can find (i) $A$-distance and MDD on AST embeddings are much smaller than those on vanilla NCF, suggesting that AST can reduce both covariate and concept distribution shifts more effectively. (ii) AST w/o D has a smaller $A$-distance than AST w/o A while AST w/o A has smaller MDD than AST w/o D. This observation matches our key idea that adversarial matching can minimize covariate shift while self-training can alleviate concept shift. We also visualize the t-SNE embeddings sampled from training $P(z_{ui})$ and testing $Q(z_{ui})$ in Appendix I.

**Convergence.** To understand the generalization and convergence of AST, Figure 2 shows the curves of the training losses of different components and the testing NDCG on two datasets. We observe that: (i) AST is generally training-stable and can consistently converge and improve the unbiased testing performance with iterations. (ii) NDCG almost monotonically increases with iterations. This suggests that minimizing our loss, i.e., the upper bound of the ideal loss is a valid approach to improve the unbiased ranking performance.

6 Conclusions

In this paper, we studied the problem of unbiased recommendation. We provided a novel perspective from the distribution shift for the unbiased recommendation problem. We derived several generalization bounds and presented both theoretical and algorithmic analyses of current unbiased learning objectives. Based on the theoretical analyses, we further proposed the AST framework, which can effectively alleviate the selection
bias and unobserved confounder issues. Extensive experiments on three datasets with various settings show that AST yields the best performance compared with state-of-the-art baselines. Ablation studies confirm the effectiveness of minimizing our proposed generalization bound in the unbiased recommendation.

References
[1] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. 2005.
[2] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. 2007.
[3] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. 2010.
[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. 2019.
[5] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. 2007.
[6] Stephen Bonner and Flavian Vasile. Causal embeddings for recommendation. 2018.
[7] Jiawei Chen, Hande Dong, Yang Qiu, Xiangnan He, Xin Xin, Liang Chen, Guli Lin, and Keping Yang. Autodebias: Learning to debias for recommendation. 2021.
[8] Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. Self-training avoids using spurious features under domain shift. 2020.
[9] Zhihong Chen, Rong Xiao, Chenliang Li, Gangfeng Ye, Haochuan Sun, and Hongbo Deng. Esam: Discriminative domain adaptation with non-displayed items to improve long-tail performance. 2020.
[10] Imre Csiszár and János Körner. Information theory: coding theorems for discrete memoryless systems. 2011.
[11] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. 2011.
[12] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. 2015.
[13] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. 2004.
[14] Siyuan Guo, Lixin Zou, Yiding Liu, Wenwen Ye, Suqi Cheng, Shuaiqiang Wang, Hechang Chen, Dawei Yin, and Yi Chang. Enhanced doubly robust learning for debiasing post-click conversion rate estimation. 2021.
[15] Shantanu Gupta, Hao Wang, Zachary Lipton, and Yuyang Wang. Correcting exposure bias for link recommendation. 2021.
[16] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. 2017.
[17] Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In AISTATS, 2019.
[18] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. 2009.
[19] Adit Krishnan, Ashish Sharma, Aravind Sankar, and Hari Sundaram. An adversarial approach to improve long-tail performance in neural collaborative filtering. 2018.
[20] Sanjay Krishnan, Jay Patel, Michael J Franklin, and Ken Goldberg. A methodology for learning, analyzing, and mitigating social influence bias in recommender systems. 2014.
[21] Jingjing Li, Erpeng Chen, Zhengming Ding, Lei Zhu, Ke Lu, and Heng Tao Shen. Maximum density divergence for domain adaptation. 2020.
[22] Dugang Liu, Pengxiang Cheng, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. A general knowledge distillation framework for counterfactual recommendation via uniform data. 2020.

[23] Dugang Liu, Pengxiang Cheng, Hong Zhu, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. Mitigating confounding bias in recommendation via information bottleneck. 2021.

[24] Weiming Liu, Jiajie Su, Chaochao Chen, and Xiaolin Zheng. Leveraging distribution alignment via stein path for cross-domain cold-start recommendation. 2021.

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

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

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

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

[29] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. 1974.

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

[31] Yuta Saito. Asymmetric tri-training for debiasing missing-not-at-random explicit feedback. 2020.

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

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

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

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

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

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

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

[39] Vladimir N Vapnik. An overview of statistical learning theory. 1999.

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

[41] Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. 2020.
[45] Teng Xiao and Suhang Wang. Towards unbiased and robust causal ranking for recommender systems. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, pages 1158–1167, 2022.

[46] Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Adversarial counterfactual learning and evaluation for recommender system. 2020.

[47] Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Rethinking neural vs. matrix-factorization collaborative filtering: the theoretical perspectives. 2021.

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

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

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

[51] Ziwei Zhu, Yun He, Yin Zhang, and James Caverlee. Unbiased implicit recommendation and propensity estimation via combinational joint learning. 2020.
A Related Work

A.1 Selection Bias in Recommendation

Unbiased learning algorithms such as IPS [33, 32, 51, 32] and DR [41, 45, 42] were proposed to deal with the bias issue theoretically. The DR combines the propensity score estimation and the error imputation model in a theoretically sophisticated manner. Although these methods can get theoretically unbiased models, it is impossible to know the true propensity score in the real world and the performance of these methods highly depends on the propensity score estimation. In addition, previous works [37, 11] have shown that these methods suffer from the high variance problem [30]. More importantly, these causal inference methods typically make an implicit unconfoundedness assumption, assuming that relevance is independent of exposure given the features [32, 46, 33]. Xu et al. [46] make similar observations about deficiencies of the unconfoundedness assumption and provide the theoretical analysis to show an inconsistent issue of supervised learning caused by the unknown exposure mechanism. However, they do not provide a theoretical analysis of the difference between the existing counterfactual learning on unbiased recommendation.

Recently, many empirical algorithms have been introduced to avoid estimating the propensity score with various techniques such as casual embedding [6], meta learning [7, 42], knowledge distillation [22] and information bottleneck [43]. Despite the promising performance in practice, most of them require additional unbiased uniform data collected at the expense of degrading user experiences and lack sufficient theoretical guarantees, which has led to a disconnect between theory and current algorithms. This paper mainly focuses on selection bias, aims to bridge the gaps between the theories and algorithms for the unbiased recommendation and proposes a novel theoretically motivated framework for unbiased recommendation.

A.2 Domain Adaptation and Self-Training

Our unbiased recommendation problem setting can be treated as a special instantiation of out-of-distribution generalization and is related to domain adaptation [2, 3, 12]. We discuss the relationships of our problem setting and our model with domain adaption. The goal of domain adaptation is to train a predictor that works well on a target domain by using only labeled source samples and unlabeled target samples during training. The adversarial feature adaptation methods [12] inspired by the theoretical analysis of [2] are most similar to ours. Specifically, Ganin and Lempitsky [12] proposed DANN to simultaneously minimize source empirical errors and approximate the divergence between source and target domains [2]. Our approach further develops this idea to the unbiased learning in recommendation, but our works are different from domain adaptation in three aspects: (1) Our work focuses on the unbiased recommendation scenario where there exist both selection bias and unobserved confounders simultaneously as shown in § 2.2 (2) We are the first to develop a generalization analysis with unbiased uniform data that take into account both selection bias and unobserved confounders. (3) We derive two novel generalization bounds for both multi-task and meta-learning strategies of using unbiased uniform data proposed by recent unbiased recommendation algorithms [22, 42, 7, 6].

Our work is also related to self-training [13, 4, 44], 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 the semi-supervised learning, in this paper, we address the unbiased recommendation problem with the self-training. There are also some works [49, 24, 19, 9] applying the self-training or domain adaptation for long-tail and cross-domain recommendation. Different from them, in this paper, we focus on providing a theoretical analysis of the unbiased recommendation and address the selection bias issue.

B The Training Algorithm

Algorithm. We summarize the training method of AST in Algorithm 1 and illustrate AST in Figure 3. At each iteration, we sample the 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 biased data and unbiased data together with biased ground-truths and unbiased pseudo-labels. The critic is optimized adversarially for the better distribution matching.
**Algorithm 1.** Adversarial Self-Training (AST)

**input** The collected biased data $\mathcal{D}_\rho$, unbiased data $\mathcal{D}_Q$ and hyperparameters $\alpha, \beta, \gamma$. Learning rate $\eta$. Maximum steps $T$.

1. **if** $\mathcal{D}_Q \neq \emptyset$ **then**
2.   $\mathcal{D} = \mathcal{D}_\rho \cup \mathcal{D}_Q$
3. **else**
4.   $\mathcal{D} = \mathcal{D}_\rho$
5. **for** $n = 1, \ldots, T$ **do**
6.   Sample batches of $(x_u, x_i) \in Q(x_u, x_i, O_u)$
7.   Generate pseudo-labels $Y_u'$ for each sample: $(x_u, x_i, Y_u')$
8.   $(\phi_n, \psi_n) \rightarrow (\phi_{n-1}, \psi_{n-1}) - \eta(\nabla_\phi \mathcal{L} - \nabla_\psi \mathcal{L}), \theta_n \rightarrow \theta_{n-1} + \eta \nabla \theta$
9. **end**

**Complexity.** Due to the online mini-batch updating, the time complexity of AST is at the same level as vanilla supervised learning and other unbiased learning algorithms [31, 43, 22, 6]. As shown in Figure 3 compared with other unbiased learning algorithms [31, 43, 22], 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 low.

### C The Lemmas

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

**Lemma C.1.** [3]. Let $\mathcal{H}$ be a hypothesis space of VC-dimension $d$, and for any distribution $P$ and $Q$ over $X_u \times X_i$, then $\forall h, h' \in \mathcal{H}$:

$$|\mathcal{L}_P(h) - \mathcal{L}_Q(h')| \leq \frac{1}{2} d_H(P, Q),$$

where $d_H(P, Q) = 2 \sup_{h, h' \in \mathcal{H}} |E_P[\ell(h(x_u, x_i), h'(x_u, x_i))] - E_Q[\ell(h(x_u, x_i), h'(x_u, x_i))]|$.

**Lemma C.2.** [39]. Let $S$ is a arbitrary 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 $e$:

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

### D Proof of Theorem 2

**Proof.** Following the definitions in §2 we have:

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

which has the following upper bound:

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

$$\leq |\mathcal{L}_P(f, k) - \mathcal{L}_P(f, g)| + |\mathcal{L}_P(f, g) - \mathcal{L}_Q(f, g)|$$

$$= |E_P[|f(x_u, x_i) - k(x_u, x_i)|] - |f(x_u, x_i) - g(x_u, x_i)||,$$

$$+ |\mathcal{L}_P(f, g) - \mathcal{L}_Q(f, g)| \leq E_P[|k(x_u, x_i) - g(x_u, x_i)|] + \frac{1}{2} d_H(P, Q),$$
where we utilize the triangular inequality and Lemma C.1. Similarly, due to the symmetric property, the following inequality for $Q$ holds:

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

$$\leq \mathbb{E}_Q[|k(x_u, x_i) - g(x_u, x_i)|] + \frac{1}{2}d_{\mathcal{H}}(P, Q). \quad (23)$$

Combine the inequalities (22) and (23) above, we have:

$$\mathcal{L}_Q(f) \leq \mathcal{L}_P(f) + \frac{1}{2}d_{\mathcal{H}}(P, Q)$$

$$+ \min \{\mathbb{E}_P[|k(x_u, x_i) - g(x_u, x_i)|], \mathbb{E}_Q[|k(x_u, x_i) - g(x_u, x_i)|]\}. \quad (24)$$

Combining Eqs. (23), (24) and Lemma C.2, and considering the hypothesis $f(x_u, x_i)$ is composed of two parts: $f = h \circ \phi$ where $h$ is the hypothesis and $\phi$ maps $(x_u, x_i)$ to $z_{ui}$. W.p. at least $1 - \delta$:

$$\mathbb{L}_Q(h \circ \phi) \leq \mathbb{L}_P(h \circ \phi) + \frac{1}{2}d_{\mathcal{H}}(P, Q)$$

$$+ \min \{\mathbb{E}_P[|\hat{k}(z_{ui}) - \hat{g}(z_{ui})|], \mathbb{E}_Q[|\hat{k}(z_{ui}) - \hat{g}(z_{ui})|]\}$$

$$+ \sqrt{\frac{4}{N} (d \log \frac{2cN}{d} + \log \frac{4}{\delta})}.$$  

We can complete the whole proof by combining Eq. (26) with Eq. (25) over coefficients $\rho$ and $1 - \rho$, respectively.

### E. Proof of Theorem 3

**Proof.** We firstly derive the upper bound between the expected unbiased error $\mathcal{L}_Q(h)$ and the empirical unbiased error $\widehat{\mathcal{L}}_Q(h(w))$ via the meta validation. Specifically, we define that $\varepsilon_i(h(w)) = \mathcal{L}_Q(h) - \ell(h(w)(x_u, x_i), y)$ for $h(w) \in \mathcal{H}'$ and every data sample in $(x_u, x_i, y) \in D_Q$. Then, we have:

$$\mathcal{L}_Q(h) - \widehat{\mathbb{L}}_Q(h(w)) = \frac{1}{M} \sum_{m=1}^{M} \varepsilon_m(h(w)). \quad (27)$$

Since $\mathcal{L}_Q(h) \in [0, 1]$ and $\ell(h(w)(x_u, x_i), y) \in [0, 1]$, we have $\mathcal{L}_Q(h) - \ell(h(w)(x_u, x_i), y) \in [-1, 1]$, $\mathbb{E}[\varepsilon_m(h(w))] \leq 1$, and $|\mathbb{E}[\varepsilon_m(h(w))]| \leq 1$. Based on the Bernstein inequality [38], we have:

$$p\left(\frac{1}{M} \sum_{m=1}^{M} \varepsilon_m(h(w)) > \xi\right) \leq \exp\left(-\frac{\xi^2M/2}{1+\xi^2/3}\right). \quad (28)$$

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

$$p(\cup_{h(w) \in \mathcal{H}'} \frac{1}{M} \sum_{m=1}^{M} \varepsilon_m(h(w)) > \xi) \leq M'' \exp\left(-\frac{\xi^2M/2}{1+\xi^2/3}\right). \quad (29)$$

Let $\delta = M'' \exp\left(-\frac{\xi^2M/2}{1+\xi^2/3}\right)$. Solving the above Inequality (29) for $\xi$ yields the following result (note that $\xi \geq 0$):

$$\xi = \frac{d' \log M - \log \delta}{3M} \pm \sqrt{\left(\frac{d' \log M - \log \delta}{3M}\right)^2 + \frac{2(d' \log M - \log \delta)}{M} \leq \frac{d' \log M - \log \delta}{3M} + \sqrt{\frac{2(d' \log M - \log \delta)}{M}}}.$$  

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

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

Similar to Eq. (26), by furthering considering the above bound in the latent feature space via $\phi$ and combine it with Eq. (25) over coefficients $\rho$ and $1 - \rho$ respectively, we complete the proof.
biased and unbiased datasets. We take the Yahoo as an example and plot the marginal distributions of
ratings, and is asked to rate 10 randomly displayed songs.

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

\[
\mathcal{L}_{Q_z}(h) \leq \mathcal{L}_{P_z}(h) = \mathcal{L}_{P_z}(h) - |\mathcal{L}_{P_z}(h, \tilde{h}) + |\mathcal{L}_{Q_z}(h) - \mathcal{L}_{P_z}(h, \tilde{h})| \\
\leq \mathcal{L}_{P_z}(h) + |\mathcal{L}_{P_z}(h, \tilde{h})| + |\mathcal{L}_{Q_z}(h) - \mathcal{L}_{P_z}(h, \tilde{h})| \\
= \mathcal{L}_{P_z}(h) + \mathcal{L}_{P_z}(|\mathcal{g}(z) - \tilde{k}(z)|) + \mathcal{L}_{Q_z}(h) - \mathcal{L}_{P_z}(h, \tilde{h}) \\
\leq \mathcal{L}_{P_z}(h) + \mathcal{L}_{P_z}(|\mathcal{g}(z) - \tilde{k}(z)|) + \mathcal{L}_{Q_z}(h) - \mathcal{L}_{P_z}(h, \tilde{h}) \\
\leq \mathcal{L}_{P_z}(h) + \mathcal{L}_{P_z}(|\mathcal{g}(z) - \tilde{k}(z)|) + TV(P_z||Q_z). \\
\leq \mathcal{L}_{P_z}(h) + \mathcal{L}_{P_z}(|\mathcal{g}(z) - \tilde{k}(z)|) + 2KL(P_z||Q_z),
\]

(32)

where we used triangular inequality multi-times and the Pinsker’s inequality 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:

\[
\mathcal{L}_{Q_z}(h) \leq \mathcal{L}_{P_z}(h) + \mathcal{L}_{Q_z}(|\mathcal{g}(z) - \tilde{k}(z)|) + 2KL(P_z||Q_z),
\]

(33)

Combining Eqs. (32), (33) and Lemmas C.2 we have (Note that):

\[
\mathcal{L}_Q(h \circ \phi) \leq \mathcal{L}_P(h \circ \phi) + \frac{1}{2} 2KL(P(z_{ui})||Q(z_{ui})) \\
+ \rho \min\{E_{P(z_{ui})}(|\mathcal{g}(z_{ui}) - \tilde{k}(z_{ui})|), E_{Q(z_{ui})}(|\mathcal{g}(z_{ui}) - \tilde{k}(z_{ui})|)\}
\]

(34)

We can complete the proof by summing this bound with Eq. (26) over coefficients \( \rho \) and \( 1 - \rho \), respectively.

\[
\text{\Box}
\]


\section*{G Datasets, processing and splitting}

\textbf{YAHOO.} \url{https://webscope.sandbox.yahoo.com/}. Its biased training set has approximately 300,000 five-star ratings of 1,000 songs from 15,400 users. It collects a test set by asking 5,400 users to rate 10 randomly displayed songs. \textbf{COAT.} \url{https://www.cs.cornell.edu/~schnabts/mnar/}

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. To better understand the biased and unbiased datasets. We take the Yahoo as an example and plot the marginal distributions of \( P(Y_{ui}) \) and \( Q(Y_{ui}) \), and the difference of rating distribution between training \( p(R_{ui}|x_{ui}, x_u, O_{ui}) \) and testing \( p(R_{ui}|x_{ui}, x_u) \) in Fig. 4 which shows that there exists some conditional distribution shifts which might be caused by extra random perturbations of unobserved confounders. \textbf{GOODREADS.} \url{https://sites.google.com/eng.ucsd.edu/ucsdbookgraph} This is book recommendation dataset and we use the book review subset in history and biography. It contains 238,450 users, 302,346 unique items, and 2,066,193 five-star ratings.
Prepossessing. Following [43] [22] [7], for COAT and YAHOO, 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.

Simulation process. The exposure mechanism depends on the underlying recommender model as well as unobserved confounders in the real-world. To mimic the real-world scenarios, we design a simulation approach based on GOODREADS. First, we train a MF model to approximate the rating matrix by minimizing the mean-squared loss. Then ground-truth preference probability is \( p(Y_{u,i} = 1|O_{u,i}) := \sigma(\hat{E}[R_{u,i}|O_{u,i}]) + \epsilon_R \) where \( \hat{E}[R_{u,i}|O_{u,i}] \) is the model output and \( \epsilon_R \) is Gaussian noisy controlling randomness of preference caused by unobserved factor. Then, similar to [32] [46], we utilize another logistic MF predicting if the rating is observed as the exposure \( \hat{p}(O_{u,i}) \). The final log-exposure probability \( \log p(O_{u,i}) = \log \hat{p}(O_{u,i}) + \epsilon_O \), where \( \epsilon_O \) measures the extra randomness of exposure by unobserved factor. Following the generative model in § 3, we generate the click as \( p(Y_{u,i} = 1) = p(Y_{u,i} = 1|O_{u,i})p(O_{u,i}) \). With this process, one can evaluate on the true relevance, exposure parameters and unobserved factors.

Splitting. For real-world datasets, following [22] [7], we regard all biased \( D_P \) data as training set, while split the unbiased data into three parts: 5% as additional training set \( D_O \) to help training, 5% as validation set, and the remaining 90% as test set. We split GOODREADS based on the order of interactions. We hold out the last interacted item of each user as the test data, and the item before the last item as the validation set. The rest items are treated as training dataset (5% is true preference and the rest is feedbacks).

H The Experimental Settings

Datasets. The datasets used in experiments can be public available. YAHOO: [43] [22] [7].

Implementations. For most of baselines, we use the official implementations publicly released by the authors on Github:

- IPS: [43]
- CVIB: [22]
- KD: [7]
- ACL: [43]
- ATT: [7]
- AutoDebias: [43]

For DRJL, we implement it by ourselves using Pytorch.

Hyper-parameters. Following previous works [43] [22] [7], 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 searches \( \alpha \), \( \beta \), and \( \gamma \) all from space \{0.2, 0.4, 0.6, 0.8\}. For a rigorous and fair comparison, we tune hyperparameters for all methods on the validation set individually.

I Additional Experimental Results

We also investigate the sensitivity of hyperparameters \( \alpha \) and \( \beta \), where \( \alpha \) and \( \beta \) control the contribution of adversarial matching and self-training, respectively. The \( \gamma \) shares the same trend as \( \beta \), thus we omit it to save space. We vary \( \alpha \) and \( \beta \) as \{0.2, 0.4, 0.6, 0.8\} and report the results in Fig. 5. We find that: (i) The performance of AST is generally better and stable when \( \alpha \in \{0.6, 0.8\} \) and \( \beta \in \{0.2, 0.6\} \), which eas the
hyperparameter selection.  \( \textbf{(ii)} \) We can balance the adversarial matching and self-training by varying \( \alpha \) and \( \beta \), leading to better generalization performance. This confirms the motivation of jointly alleviating selection bias and unobserved confounders.

To further intuitively understand the feature adaptation, we visualize the t-SNE embeddings sampled from training \( P(z_{ui}) \) and testing \( Q(z_{ui}) \) in a mini batch. We provide the results in Fig. 6 in Appendix. From Fig. 6 we can find AST can effectively bridge the feature gap across biased data and unbiased data, but Biased training can not since the embeddings are separated and have a certain distance.

### J Broader Impact and Discussion

While our framework can theoretically explain many unbiased learning objectives, some objectives can not currently fall into this framework. For instance, we do not consider the imputation term in the doubly robust learning objective [41]. Therefore, the extension of this framework to other learning objectives is an interesting theme. Also, while our results strongly advocate for considering the unobserved confounders in unbiased recommendation in the bound, it is challenging to optimize it in the real-world recommendation directly. We leave to future work to study how to optimize it more effectively.

As machine learning finds widespread applications in recommender systems, there is notable interest in understanding its societal impacts. Recommender systems could be crucial tools for a broad range of applications, including social networks, healthcare, and electronic commerce. Any of these applications may have a different social effect. While recommendation decisions can counteract existing biases by preventing human error and implicit bias, algorithms may also create new avenues for introducing unintended bias due to the closed feedback loop. This purpose introduces a novel framework to alleviate the model selection bias in this feedback loop process. Overcoming the selection bias can promote the recommender system to infer real user preference to facilitate a better user experience and people’s daily lives.

An major problem is that while our unbiased algorithm can have the ability to improve the unbiased ranking performance, security concerns such as fairness, inequality, and privacy often hinder their practical implementation. Specifically, for the privacy issue, although our method only utilizes the logged feedback and does not require other sensitive information. Some malicious adversaries could try to infer sensitive information such as gender and race from the user feedback. Thus how we could protect users’ sensitive information in the recommendation from malicious inference attacks while maintaining unbiasedness is an important challenge.