Who Should Be Given Incentives? Counterfactual Optimal Treatment Regimes Learning for Recommendation

Haoxuan Li
Peking University
hxli@stu.pku.edu.cn

Chunyuan Zheng
University of California, San Diego
czheng@ucsd.edu

Peng Wu
Beijing Technology and Business University, pengwu@btbu.edu.cn

Kun Kuang
Zhejiang University
kunkuang@zju.edu.cn

Yue Liu*
Renmin University of China
liuyue_stats@ruc.edu.cn

Peng Cui
Tsinghua University
cuip@tsinghua.edu.cn

ABSTRACT

Effective personalized incentives can improve user experience and increase platform revenue, resulting in a win-win situation between users and e-commerce companies. Previous studies have used uplift modeling methods to estimate the conditional average treatment effects of users' incentives, and then placed the incentives by maximizing the sum of estimated treatment effects under a limited budget. However, some users will always buy whether incentives are given or not, and they will actively collect and use incentives if provided, named "Always Buyers". Identifying and predicting these "Always Buyers" and reducing incentive delivery to them can lead to a more rational incentive allocation. In this paper, we first divide users into five strata from an individual counterfactual perspective, and reveal the failure of previous uplift modeling methods to identify and predict the "Always Buyers". Then, we propose principled counterfactual identification and estimation methods and prove their unbiasedness. We further propose a counterfactual entire-space multi-task learning approach to accurately perform personalized incentive policy learning with a limited budget. We also theoretically derive a lower bound on the reward of the learned policy. Extensive experiments are conducted on three real-world datasets with two common incentive scenarios, and the results demonstrate the effectiveness of the proposed approaches.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

Counterfactual, Optimal treatment regime, Recommender system.

ACM Reference Format:
Haoxuan Li, Chunyuan Zheng, Peng Wu, Kun Kuang, Yue Liu, and Peng Cui. 2023. Who Should Be Given Incentives? Counterfactual Optimal Treatment Regimes Learning for Recommendation. 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, 13 pages. https://doi.org/10.1145/3580305.3599550

*Corresponding author.

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

ACM ISBN 979-8-4007-0103-0/23/08…$15.00
https://doi.org/10.1145/3580305.3599550

1 INTRODUCTION

Conversion feedback reflects a strong signal of user preference and is directly linked to Gross Commodity Volume (GMV) [42, 49, 68]. To attract user interest and increase platform revenue, many e-commerce companies offer personalized incentives to users (e.g., sending coupons, or giving cash bonuses) to increase conversions, which are widely adopted in many application scenarios, such as e-commerce transactions and music websites [8, 16]. Effective incentive regimes for specific consumers can increase user stickiness and achieve user growth, resulting in a win-win situation between users and e-commerce companies.

In general, personalized incentive policies give incentives to specific subgroups using observed user and item features. As a result, some users will accept the incentive and others will ignore it, and eventually these incentives would contribute to the conversions, as shown in the causal diagram in Figure 1. In order to effectively identify users who need to be incentivized, an important question to be answered is, "If an incentive is given to a specific user, will the user purchase the item?". However, in real-world scenarios, we can never simultaneously observe the conversion outcomes with and without incentives for the same user, which is also known as the fundamental problem of causal inference [21].

To tackle the above issues, recent studies have proposed modeling conditional average causal effects (CATEs, also known as uplift modeling) to identify individuals who should be given incentives [50–52, 77]. Specifically, as shown in Table 1, the CATE-based methods are able to identify "Coupon Buyers", i.e., users who would actively collect incentives and convert when incentives are given, but would not result in conversion when incentives are not given. Given the features of users and items, the personalized incentive algorithms first estimate the CATEs as the probability that each user belongs to "Coupon Buyers", and then place incentives by maximizing the sum of the CATEs with a limited budget.

However, in this paper, as shown in Table 2, we argue that these CATE-based methods *cannot* further identify "Always Buyers" from the remaining users, i.e., these users will buy with or without incentives, but they will actively collect incentives if given, which leads to unnecessary incentives. Another category of users that cannot be identified is the "Coupon Takers", i.e., they actively receive incentives without converting, which also results in wasted incentives

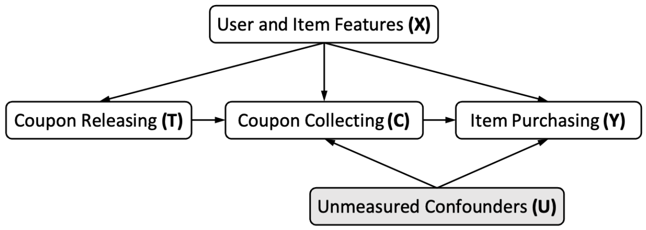


Figure 1: The causal diagram of *Coupon Releasing→Coupon Collecting→Item Purchasing* in e-commerce.

Table 1: The user-item pairs are divided into five strata from a counterfactual perspective, i.e., $(C(0), C(1), Y(0), Y(1))$**, named "never buyer", "never taker", "coupon taker", "coupon buyer", and "always buyer", respectively.**

Strata	Description	C(0)	C(1)	Y(0)	Y(1)	Reward
Y_{0000}	Never Buyer	0	0	0	0	0
Y_{0011}	Never Taker	0	0	1	1	0
Y_{0100}	Coupon Taker	0	1	0	0	0 or $-c(x)$
Y_{0101}	Coupon Buyer	0	1	0	1	$s(x)$
Y_{0111}	Always Buyer	0	1	1	1	$-c(x)$

Note: The "Coupon Takers" has different rewards for different forms of incentives: 0 if the incentive is a coupon that is only available when purchasing, and $-c(x)$ if the incentive is a cash bonus. $s(x)$ is the net profit from the incentive placement.

when the incentives are cash bonuses. In contrast to these two groups, "Never Buyers" never actively collect incentives and never convert, while "Never Takers" always convert, but never actively collect incentives. It can be summarized that both "Always Buyers" and "Coupon Takers" may cause unnecessary waste of incentives, while "Never Buyers" and "Never Takers" do not actively collect incentives (note that the causal effects of all these four subgroups of incentives on conversion are zero). Therefore, in addition to incentivizing "Coupon Buyers" using causal effects, *identifying and reducing incentives for "Always Buyers" and "Coupon Takers" can also help reduce costs and increase platform revenue.*

Towards this end, we first formalize the personalized incentive scenarios using the widely adopted potential outcome framework in causal inference, and then divide the population into five strata from a counterfactual perspective, i.e., based on the joint potential outcomes of the same individuals. Next, we formally reveal the limitations of the previous CATE-based methods that can only identify and consider "Coupon Buyers". Then, we propose **C**ounter**F**actual-**O**utcome **R**egression/**I**nverse **P**ropensity **S**coring/**D**oubly **R**obust estimators, named **CF-OR**, **CF-IPS**, and **CF-DR**, respectively, which can further identify and estimate all five strata. Through theoretical analysis, we demonstrate the double robustness property of the proposed CF-DR estimator, i.e., it is unbiased when either of the imputed outcomes or learned propensities are accurate.

Based on the proposed counterfactual identification methods, we further propose a **C**ounter**F**actual entire-space **M**ulti-**T**ask **L**earning approach with a limited budget, named **CF-MTL**, in which propensity model and counterfactual strata prediction models of individuals are jointly trained. Compared with training multiple regression and propensity models independently for policy learning, CF-MTL can alleviate the data sparsity and bias amplification problems,

Table 2: Comparison of the identifiability of the Naive methods, the conditional average treatment effect (CATE)-based methods, and the proposed counterfactual methods.

Method	Identifiable	Unidentifiable
Naive	N/A	Y_{0000}, Y_{0011}, Y_{0100}, Y_{0101}, Y_{0111}
CATE	Y_{0101}	Y_{0000}, Y_{0011}, Y_{0100}, Y_{0111}
Ours	Y_{0000}, Y_{0011}, Y_{0100}, Y_{0101}, Y_{0111}	N/A

which leads to more accurate strata prediction. We also theoretically derive lower bounds for the reward of learned personalized incentive policy. Extensive experiments are conducted on three real-world datasets with two common incentive scenarios, and the results demonstrate that the proposed learning approach can accurately achieve individual counterfactual predictions, thus leads to significantly more profitable incentive policies.

The main contributions of this paper are summarized as follows.

- We reformulate the personalized incentive policy learning problem from an individualized counterfactual perspective, reveal the limitations of previous uplift modeling methods, and propose principled counterfactual estimators to identify and estimate the probability that an individual belongs to a specific counterfactual strata.
- Based on the proposed counterfactual identification methods, we further propose a counterfactual entire-space multi-task learning approach to accurately perform individualized incentive policy learning. We also theoretically derive a lower bound on the reward of the learned policy.
- We conduct experiments on three real-world datasets with two common personalized incentive scenarios, and the results show the effectiveness of our approaches for counterfactual prediction and personalized incentive policy learning.

2 PRELIMINARIES AND DISCUSSIONS

2.1 Problem Setup

Let $\mathcal{U} = \{u_1, u_2, \dots, u_m\}$ be the set of m users, $\mathcal{I} = \{i_1, i_2, \dots, i_n\}$ be the set of n items, and $\mathcal{D} = \mathcal{U} \times \mathcal{I}$ be the set of all user-item pairs. Denote the observed features and binary treatments of user u and item i as $X_{u,i}$ and $T_{u,i}$, respectively, where $T_{u,i}$ is the indicator of whether the personalized policy released incentives (e.g., coupons or cash bonuses) to user u about item i or not. To study individuals' subsequent incentive collection and conversion behavior from a counterfactual perspective, we adopt the potential outcome framework in causal inference [21, 23]. Specifically, let $C_{u,i}(0)$ and $C_{u,i}(1)$ be the outcomes of whether the user actively collects incentives about items (e.g., actively collected coupons, or actively withdraws cash bonuses) had the platform does not release incentives $T_{u,i} = 0$ and release incentives $T_{u,i} = 1$, respectively. Similarly, let $Y_{u,i}(0)$ and $Y_{u,i}(1)$ be the outcomes of whether the user converts to the item had $T_{u,i} = 0$ and $T_{u,i} = 1$, respectively. Since each user-item pair can only be assigned one treatment, we always observe either of the corresponding outcomes $C_{u,i}(0)$ or $C_{u,i}(1)$, but never both, and similar conclusions hold for $Y_{u,i}(0)$ or $Y_{u,i}(1)$. This is also known as the fundamental problem of causal inference [21].

We assume that the observations for user u and item i are $C_{u,i} = (1-T_{u,i})C_{u,i}(0)+T_{u,i}C_{u,i}(1)$ and $Y_{u,i} = (1-T_{u,i})Y_{u,i}(0)+T_{u,i}Y_{u,i}(1)$.

In other words, the observed outcomes are the potential outcomes corresponding to the assigned treatment, which also known as the consistency assumption in the causal literature. We assume that the stable unit treatment value assumption (STUVA) holds, i.e., there should not be alternative forms of the treatment (i.e., incentives) and interference between units. In addition, we follow previous studies to assume that the unconfoundedness assumption holds, i.e., $(C_{u,i}(0), C_{u,i}(1), Y_{u,i}(0), Y_{u,i}(1)) \perp\!\!\!\perp T_{u,i} | X_{u,i}$ and let $\eta < \mathbb{P}(T_{u,i} = 1 | X_{u,i} = x) < 1 - \eta$, where $\perp\!\!\!\perp$ means independent and η is a constant between 0 and $1/2$. In the personalized incentive scenarios, the rationality of the unconfoundedness assumption is due to the fact that the recommendation system gives incentives only based on the *observed* user and item features. When no incentives are released, it is obvious that users cannot actively collect incentives, i.e., $C_{u,i}(0) = 0$. Next, we assume that incentive collection $C_{u,i}$ has a non-negative effect on conversion $Y_{u,i}$, i.e., there is no individual with $(C_{u,i}(0) = 0, C_{u,i}(1) = 1, Y_{u,i}(0) = 1, Y_{u,i}(1) = 0)$, who does not convert when the incentive is actively collected, but converts when there is no incentive. In addition, because of the limitations of the collected information, some unmeasured confounders (e.g., user's income, etc.) may also affect coupon collecting and item purchasing, which poses additional challenges. In Figure 1, we summarize the causal diagram of *Coupon Releasing→Coupon Collecting→Item Purchasing* in e-commerce.

From a counterfactual perspective, as shown in Table 1, we divide all user-item pairs into five strata based on the joint potential outcomes $(C(0), C(1), Y(0), Y(1))$ of individuals, and named as "never buyer", "never taker", "coupon taker", "coupon buyer", and " always buyer", respectively. For simplification, we denote the labels of the five groups as Y_{0000}, Y_{0011}, Y_{0100}, Y_{0101}, and Y_{0111}, correspondingly. Previous studies have modeled the conditional average treatment effect (CATE, also known as uplift modeling) to determine which users should be given incentives. Formally, CATE is defined as

$$\tau(x_{u,i}) = \mathbb{E}(Y(1) - Y(0) \mid X = x_{u,i}) = \mathbb{E}(Y_{0101} \mid X = x_{u,i}),$$

which is equivalent to the probability that a unit with feature $x_{u,i}$ belongs to "Coupon Buyer".

2.2 Uplift Modeling

Many methods were developed for the estimation of CATEs. Let $\mu_0(x) = \mathbb{E}[Y \mid T = 0, X = x]$ and $\mu_1(x) = \mathbb{E}[Y \mid T = 1, X = x]$, the outcome regression (OR) estimator builds regression models for $Y(0)$ and $Y(1)$ respectively to fill in the missing potential outcomes

$$\hat{p}^{OR}(x_{u,i}) = \hat{\mu}_1(x_{u,i}) - \hat{\mu}_0(x_{u,i}),$$

where $\hat{\mu}_0(x_{u,i})$ and $\hat{\mu}_1(x_{u,i})$ are estimates of $\mu_0(x_{u,i})$ and $\mu_1(x_{u,i})$, respectively. The OR estimator is unbiased when the imputed outcomes $\mu_0(x_{u,i})$ and $\mu_1(x_{u,i})$ are accurate, i.e., $\hat{\mu}_0(x_{u,i}) = \mu_0(x_{u,i})$ and $\hat{\mu}_1(x_{u,i}) = \mu_1(x_{u,i})$.

The alternative methods of estimating CATE are weighting-based estimators. Let $e(x) = \mathbb{P}(T = 1 \mid X = x)$ be the probability that the personalized incentive policy of the recommender system places an incentive on a user-item pair with feature x, called propensity. The inverse propensity scoring (IPS) estimator uses the inverse of the treatment probability to weight the observed potential outcomes

$$\hat{p}^{IPS}(x_{u,i}) = \frac{T_{u,i}Y_{u,i}(1)}{\hat{e}_{u,i}} - \frac{(1 - T_{u,i})Y_{u,i}(0)}{1 - \hat{e}_{u,i}},$$

where $\hat{e}_{u,i}$ is an estimate of $e(x_{u,i})$. The IPS estimator is unbiased when the learned propensities $\hat{e}_{u,i}$ are accurate, i.e., $\hat{e}_{u,i} = e(x_{u,i})$.

The doubly robust (DR) estimator uses both the outcome regression models and the propensity model to relax the conditions for unbiasedness

$$\hat{p}^{DR}(x_{u,i}) = \hat{\mu}_1(x_{u,i}) - \hat{\mu}_0(x_{u,i}) + \frac{T_{u,i}(Y_{u,i}(1) - \hat{\mu}_1(x_{u,i}))}{\hat{e}_{u,i}} - \frac{(1 - T_{u,i})(Y_{u,i}(0) - \hat{\mu}_0(x_{u,i}))}{1 - \hat{e}_{u,i}},$$

which has double robustness, i.e., it is unbiased if either the imputed outcomes or learned propensities are accurate.

2.3 Personalized Incentive Policy Learning

Let $\pi : \{x_{u,i} \mid (u, i) \in \mathcal{D}\} \to [0, 1]$ be a policy that maps from the individual context $X = x$ to the probability of the incentives $T = 1$ to be placed. Suppose the net profit from the incentive placement to "Coupon Buyer" is $s(x_{u,i})$. For symbolic simplicity, we consider the case $s(x_{u,i}) = 1$ hereafter. Similar results hold for the case of heterogeneous net profits, provided $s(x_{u,i})$ are bounded. Given the CATEs, the personalized incentive policy is trained to maximize the weighted sum of CATEs within a finite budget ϵ to place incentives

$$\max_{\pi \in \Pi} R(\pi) = \frac{1}{|\mathcal{D}|} \sum_{(u,i) \in \mathcal{D}} \pi(x_{u,i}) r_{u,i}$$

$$\text{s.t. } B(\pi) = \frac{1}{|\mathcal{D}|} \sum_{(u,i) \in \mathcal{D}} \pi(x_{u,i}) \le \epsilon,$$

where $r_{u,i} = \mathbb{P}(Y_{0101} \mid x_{u,i})$. The optimal policy $\pi^*(x_{u,i})$ is

$$\pi^*(x_{u,i}) = \begin{cases} 1, \mathbb{P}(Y_{0101} \mid x_{u,i}) > \gamma(\epsilon) \\ d, \mathbb{P}(Y_{0101} \mid x_{u,i}) = \gamma(\epsilon), \\ 0, \mathbb{P}(Y_{0101} \mid x_{u,i}) < \gamma(\epsilon) \end{cases}$$

where d is a value between 0 and 1, and $\gamma(\epsilon) \ge 0$ decreases monotonically as the budget ϵ increases. It can be seen that the optimal policy for uplift modeling with finite budget ϵ selects units with CATEs above a threshold $\gamma(\epsilon)$ to give incentives. Empirically, we use estimates of CATEs, e.g., $\hat{p}^{OR}(x_{u,i})$, $\hat{p}^{IPS}(x_{u,i})$, and $\hat{p}^{DR}(x_{u,i})$, to replace $r_{u,i}$ to perform personalized incentive policy learning.

3 PROPOSED METHODS

3.1 Limitations of Uplift Modeling

Despite the CATE-based incentive policy learning can effectively identify and estimate "Coupon Buyers", as shown in Table 2, it fails to identify and estimate users in other strata. In fact, identifying "Always Buyer" and "Coupon Taker" is meaningful for e-commerce platforms, and by reducing the incentive allocation to these two strata, personalized incentives can be more rationally allocated.

Specifically, from Table 1, "Always Buyer" always buys regardless of the coupon given, i.e., the causal effect of the coupon on the purchase is always zero, but they would actively use the coupon if it was offered. In addition, if the incentive is a cash bonus, then both the "Always Buyer" and "Coupon Taker" will actively collect and obtain the cash bonus, even if the latter will never purchase. This can result in unnecessary placement of incentives and additional costs. In contrast, "Never Buyer" and "Never Taker" never collect the incentive, so incentive placement for them leads to no additional incentive costs and **zero rewards.** Therefore, when the cost of the incentive is $c(x)$, incentive placement for "Always Buyer" will have $\boldsymbol{-c(x)}$ **reward** *due to the zero causal effect on purchases and the additional incentive cost of* $c(x)$. Last, incentive placement for "Coupon Taker" has different rewards for various incentives: **zero reward** if the incentive is a coupon that is only available when purchasing, and $\boldsymbol{-c(x)}$ **reward** if the incentive is a cash bonus.

3.2 Counterfactual Identification Methods

In order to accurately identify and estimate the probability of a unit belonging to the counterfactual strata Y_{0000}, Y_{0011}, Y_{0100}, Y_{0101}, and Y_{0111} based on the observed features $x_{u,i}$, we propose the following counterfactual identification method. For units in the observed data that receive an incentive $T = 1$, from the $C(1)$ and $Y(1)$ columns of Table 1, it can be found that only the Y_{0000} stratum results in the observed outcomes $(T = 1, C = 0, Y = 0)$. Similarly, only Y_{0011} stratum results in the observed outcomes $(T = 1, C = 0, Y = 1)$, and only Y_{0100} stratum results in the observed outcomes $(T = 1, C = 1, Y = 0)$. However, both Y_{0101} and Y_{0111} strata can lead to the observed outcomes $(T = 1, C = 1, Y = 1)$. Formally, we have

$$\mathbb{P}(C = 0, Y = 0 \mid T = 1, X) = \mathbb{P}(Y_{0000} \mid X),$$
$$\mathbb{P}(C = 0, Y = 1 \mid T = 1, X) = \mathbb{P}(Y_{0011} \mid X),$$
$$\mathbb{P}(C = 1, Y = 0 \mid T = 1, X) = \mathbb{P}(Y_{0100} \mid X),$$
$$\mathbb{P}(C = 1, Y = 1 \mid T = 1, X) = \mathbb{P}(Y_{0101} \mid X) + \mathbb{P}(Y_{0111} \mid X).$$

The above formulas cannot identify Y_{0101} as a "Coupon Buyer" and Y_{0111} as an "Always Buyer". Fortunately, by an similar argument for the units that are not released with incentive $T = 0$ in the observed data, the following formulas hold

$$\mathbb{P}(C = 1, Y = 1 \mid T = 0, X) = 0,$$
$$\mathbb{P}(C = 1, Y = 0 \mid T = 0, X) = 0,$$
$$\mathbb{P}(C = 0, Y = 1 \mid T = 0, X) = \mathbb{P}(Y_{0011} \mid X) + \mathbb{P}(Y_{0111} \mid X),$$
$$\mathbb{P}(C = 0, Y = 0 \mid T = 0, X) = \mathbb{P}(Y_{0000} \mid X) + \mathbb{P}(Y_{0100} \mid X) + \mathbb{P}(Y_{0101} \mid X).$$

Now, associating the above eight formulas, it is sufficient to identify the probability that a unit with feature X belong to each counterfactual strata. Solving these equations for $\mathbb{P}(Y_{0000} \mid X)$, $\mathbb{P}(Y_{0011} \mid X)$, $\mathbb{P}(Y_{0100} \mid X)$, $\mathbb{P}(Y_{0101} \mid X)$, and $\mathbb{P}(Y_{0111} \mid X)$ gives

$$\mathbb{P}(Y_{0000} \mid X) = \mathbb{P}(C = 0, Y = 0 \mid T = 1, X),$$
$$\mathbb{P}(Y_{0011} \mid X) = \mathbb{P}(C = 0, Y = 1 \mid T = 1, X),$$
$$\mathbb{P}(Y_{0100} \mid X) = \mathbb{P}(C = 1, Y = 0 \mid T = 1, X),$$
$$\mathbb{P}(Y_{0101} \mid X) = \mathbb{P}(Y = 1 \mid T = 1, X) - \mathbb{P}(Y = 1 \mid T = 0, X),$$
$$\mathbb{P}(Y_{0111} \mid X) = \mathbb{P}(Y = 1 \mid T = 0, X) - \mathbb{P}(C = 0, Y = 1 \mid T = 1, X).$$

3.3 Counterfactual Estimation Methods

We extend the previous OR, IPS, and DR estimators from uplift modeling to unbiasedly estimate the probability that the unit belongs to each counterfactual strata. Without loss of generality, we next discuss the estimation methods for the probability that a unit with feature X belongs to "Always Buyer", i.e., $\mathbb{P}(Y_{0111} \mid X)$. Other counterfactual strata probabilities can be derived from a similar view. By noting that the second term on the right hand side of $\mathbb{P}(Y_{0111} \mid X)$ in Section 3.2 is equivalent to

$$\begin{aligned}\mathbb{P}(C = 0, Y = 1 \mid T = 1, X) &= \mathbb{P}(C(1) = 0, Y(1) = 1 \mid T = 1, X)\\ &= \mathbb{P}(C(1) = 0, Y(1) = 1 \mid X),\end{aligned}$$

where $(C(1), Y(1))$ can be viewed as a whole as the joint potential outcomes under $T = 1$. Let $\mu_{01|1}(x) = \mathbb{E}[\mathbb{I}(C = 0, Y = 1) \mid T = 1, X = x]$, by noting that

$$\mathbb{P}(C(1) = 0, Y(1) = 1 \mid X) = \mu_{01|1}(X),$$

the proposed counterfactual-outcome regression (CF-OR) estimator for estimating $\mathbb{P}(Y_{0111} \mid X)$ is given as

$$\hat{p}^{OR}_{0111}(x_{u,i}) = \hat{\mu}_0(x_{u,i}) - \hat{\mu}_{01|1}(x_{u,i}),$$

where $\hat{\mu}_0(x_{u,i})$ and $\hat{\mu}_{01|1}(x_{u,i})$ are estimates of $\mu_0(x_{u,i})$ and $\mu_{01|1}(x_{u,i})$, respectively. The proposed CF-OR estimator is unbiased when the imputed outcomes $\hat{\mu}_0(x_{u,i})$ and $\hat{\mu}_{01|1}(x_{u,i})$ are accurate, i.e., $\hat{\mu}_0(x_{u,i}) = \mu_0(x_{u,i})$ and $\hat{\mu}_{01|1}(x_{u,i}) = \mu_{01|1}(x_{u,i})$.

Next, recall that $e(x) = \mathbb{P}(T = 1 \mid X = x)$, by noting that

$$\mathbb{P}(C(1) = 0, Y(1) = 1 \mid X) = \mathbb{E}\Big[\frac{\mathbb{I}(T = 1)\mathbb{I}(C(1) = 0, Y(1) = 1)}{e(X)} \mid X\Big],$$

the proposed counterfactual-inverse propensity scoring (CF-IPS) estimator for estimating $\mathbb{P}(Y_{0111} \mid X)$ is given as

$$\hat{p}^{IPS}_{0111}(x_{u,i}) = \frac{(1 - T_{u,i})Y_{u,i}(0)}{1 - \hat{e}_{u,i}} - \frac{T_{u,i}(1 - C_{u,i}(1))Y_{u,i}(1)}{\hat{e}_{u,i}},$$

where $\hat{e}_{u,i}$ is an estimate of $e(x_{u,i})$. The proposed CF-IPS estimator is unbiased when the learned propensities $\hat{e}_{u,i}$ are accurate, i.e., $\hat{e}_{u,i} = e(x_{u,i})$.

For estimators with doubly robust forms, by noting that

$$\begin{aligned}&\mathbb{P}(C(1) = 0, Y(1) = 1 \mid X)\\ &= \mathbb{E}\Big[\mu_{01|1}(X) + \frac{\mathbb{I}(T = 1)[\mathbb{I}(C(1) = 0, Y(1) = 1) - \mu_{01|1}(X)]}{e(X)} \mid X\Big],\end{aligned}$$

the proposed counterfactual-doubly robust (CF-DR) estimator for estimating $\mathbb{P}(Y_{0111} \mid X)$ is given as

$$\begin{aligned}\hat{p}^{DR}_{0111}(x_{u,i}) = \hat{\mu}_0(x_{u,i}) &+ \frac{(1 - T_{u,i})(Y_{u,i}(0) - \hat{\mu}_0(x_{u,i}))}{1 - \hat{e}_{u,i}}\\ &- \hat{\mu}_{01|1}(x_{u,i}) - \frac{T_{u,i}[(1 - C_{u,i}(1))Y_{u,i}(1) - \hat{\mu}_{01|1}(x_{u,i})]}{\hat{e}_{u,i}}.\end{aligned}$$

Theorem 3.1 derives the bias of the proposed CF-DR estimator.

Theorem 3.1 (Bias of CF-DR Estimator). *Given imputed outcomes $\hat{\mu}_0(x_{u,i})$, $\hat{\mu}_{01|1}(x_{u,i})$, and learned propensities $\hat{e}_{u,i} > 0$ for all user-item pairs, the bias of the CF-DR estimator is*

$$\begin{aligned}\mathrm{Bias}(\hat{p}^{DR}_{0111}(x_{u,i})) = \Big| &\frac{(e_{u,i} - \hat{e}_{u,i})(Y_{u,i}(0) - \hat{\mu}_0(x_{u,i}))}{1 - \hat{e}_{u,i}}\\ &+ \frac{(e_{u,i} - \hat{e}_{u,i})[(1 - C_{u,i}(1))Y_{u,i}(1) - \hat{\mu}_{01|1}(x_{u,i})]}{\hat{e}_{u,i}}\Big|.\end{aligned}$$

From Theorem 3.1, the proposed CF-DR estimator effectively takes advantage of the outcome regression models and the propensity model to reduce the bias of the estimation. We formally describe the double robustness property that CF-DR has as in Corollary 3.2.

Corollary 3.2 (Double Robustness). *The CF-DR estimator is unbiased when either imputed outcomes $\hat{\mu}_0(x_{u,i})$ and $\hat{\mu}_{01|1}(x_{u,i})$ or learned propensities $\hat{e}_{u,i} > 0$ are accurate for all user-item pairs.*

Next, we derive the variance of CF-DR estimator in Theorem 3.3.

Theorem 3.3 (Variance of CF-DR Estimator). *Given imputed outcomes $\hat{\mu}_0(x_{u,i})$, $\hat{\mu}_{01|1}(x_{u,i})$, and learned propensities $\hat{e}_{u,i} > 0$ for all user-item pairs, the variance of the CF-DR estimator is*

$$\mathrm{Var}(\hat{p}^{DR}_{0111}(x_{u,i})) = e_{u,i}(1-e_{u,i})\Big[\frac{Y_{u,i}(0)-\hat{\mu}_0(x_{u,i})}{1-\hat{e}_{u,i}} + \frac{(1-C_{u,i}(1))Y_{u,i}(1)-\hat{\mu}_{01|1}(x_{u,i})}{\hat{e}_{u,i}}\Big]^2.$$

Notably, when the imputed outcomes in the CF-DR estimator are zero, i.e., $\hat{\mu}_0(x_{u,i}) = 0$ and $\hat{\mu}_{01|1}(x_{u,i}) = 0$, the CF-DR estimator degenerates to the CF-IPS estimator, and therefore Theorem 3.3 degenerates to the variance of CF-IPS estimator. It can be seen that when the outcome regression models are approximately accurate, i.e., $\hat{\mu}_0(x_{u,i}) \approx Y_{u,i}(0)$ and $\hat{\mu}_{01|1}(x_{u,i}) \approx (1-C_{u,i}(1))Y_{u,i}(1)$, the CF-DR estimator would have a lower variance than the CF-IPS.

Let $\hat{p}^{DR}_{0111}$ be the average predicted probability of "Always Buyers" using CF-DR estimator over all user-item pairs

$$\hat{p}^{DR}_{0111} = \frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}} \hat{p}^{DR}_{0111}(x_{u,i}),$$

we further show the tail bound of CF-DR estimator in Theorem 3.4.

Theorem 3.4 (Tail Bound of CF-DR Estimator). *Given imputed outcomes $\hat{\mu}_0(x_{u,i})$, $\hat{\mu}_{01|1}(x_{u,i})$, and learned propensities $\hat{e}_{u,i} > 0$, with probability $1-\eta$, the deviation of the CF-DR estimator from its expectation has the following tail bound*

$$\left|\hat{p}^{DR}_{0111} - \mathbb{E}_T(\hat{p}^{DR}_{0111})\right| \le \sqrt{\frac{\log\left(\frac{2}{\eta}\right)}{2|\mathcal{D}|^2}\sum_{(u,i)\in\mathcal{D}}\left\{\frac{\tilde{\mu}_0(x_{u,i})}{1-\hat{e}_{u,i}} + \frac{\tilde{\mu}_{01|1}(x_{u,i})}{\hat{e}_{u,i}}\right\}^2},$$

where $\tilde{\mu}_0 = Y(0) - \hat{\mu}_0$ and $\tilde{\mu}_{01|1} = (1-C(1))Y(1) - \hat{\mu}_{01|1}$.

From Corollary 3.2 and Theorem 3.4, when either imputed outcomes or learned propensities are approximately accurate, the predicted amount of "Always Buyers" using the CF-DR estimator will tend to the true amount as the sample size increases. This further illustrates the effectiveness of the proposed CF-DR estimator for identifying and estimating these "Always Buyers".

4 MULTI-TASK LEARNING APPROACH

In this section, based on the proposed counterfactual identification method, we further propose a counterfactual entire-space multi-task learning approach, named CF-MTL, to accurately predict the probabilities of units belonging to different strata, which is then used to perform individualized incentive policy learning. Different from CF-OR, CF-IPS, and CF-DR estimators that first build outcome regression models and propensity model to predict potential outcomes and then estimate the probability of counterfactual strata by a plug-in manner, the proposed CF-MTL simultaneously trains a propensity model and a counterfactual strata prediction model, and the overview of the architecture is shown in Figure 2.

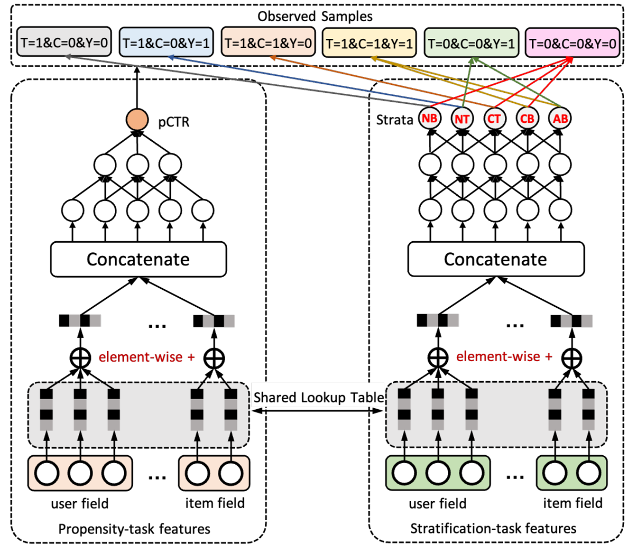


Figure 2: Proposed counterfactual entire-space multi-task learning architecture, which contains (i) a propensity model and (ii) an individual counterfactual strata prediction model.

Recall that in Section 3.2, we proved the following formula

$$\mathbb{P}(C=0, Y=0 \mid T=1, X) = \mathbb{P}(Y_{0000} \mid X).$$

By multiplying the propensity $\mathbb{P}(T=1 \mid X)$ on both sides, we have

$$\begin{aligned}\mathbb{P}(T=1, C=0, Y=0 \mid X) &= \mathbb{P}(C=0, Y=0 \mid T=1, X)\mathbb{P}(T=1 \mid X)\\ &= \mathbb{P}(Y_{0000} \mid X)\mathbb{P}(T=1 \mid X),\end{aligned}$$

where the left hand side is the joint distribution of the samples with observations $(T=1, C=0, Y=0)$ and the right hand side contains the counterfactual strata probability $\mathbb{P}(Y_{0000} \mid X)$ and propensity $\mathbb{P}(T=1 \mid X)$. Let $f_{0000}(X)$ and $g(X)$ be the models for predicting $\mathbb{P}(Y_{0000} \mid X)$ and the propensity model for predicting $\mathbb{P}(T=1 \mid X)$, respectively. Then both models can be trained jointly by minimizing the following losses using all user-item pairs in the entire-space

$$L\left(f_{0000}(X)g(X), T=1\&C=0\&Y=0\right),$$

where $L(\cdot,\cdot)$ is the average of a pre-specified loss (e.g., cross-entropy) over all user-item pair, and $T=1\&C=0\&Y=0$ equals to 1 only when the sample has observation $(T=1, C=0, Y=0)$, and equals to 0 otherwise. By adopting a similar view to the remaining identification formulas in Section 3.2, both models are jointly trained by minimizing the counterfactual stratification-task loss

$$\begin{aligned}&\mathcal{L}_s(f_{0000}, f_{0011}, f_{0100}, f_{0101}, f_{0111}; g)\\ &= L\left(f_{0000}(X)g(X), T=1\&C=0\&Y=0\right)\\ &+ L\left(f_{0011}(X)g(X), T=1\&C=0\&Y=1\right)\\ &+ L\left(f_{0100}(X)g(X), T=1\&C=1\&Y=0\right)\\ &+ L\left((f_{0101}(X)+f_{0111}(X))g(X), T=1\&C=1\&Y=1\right)\\ &+ L\left((f_{0011}(X)+f_{0111}(X))(1-g(X)), T=0\&C=0\&Y=1\right)\\ &+ L\left((f_{0000}(X)+f_{0100}(X)+f_{0101}(X))(1-g(X)), T=0\&C=0\&Y=0\right).\end{aligned}$$

In addition, to make the propensity model $g(X)$ accurately predict the incentive delivery probability $\mathbb{P}(T=1 \mid X)$, it is also trained by minimizing a binary classification loss

$$\mathcal{L}_p(g) = L(g(X), T = 1).$$

In summary, the proposed CF-MTL jointly trains $f(X)$ and $g(X)$ by minimizing the following loss in the entire-space

$$\mathcal{L} = \mathcal{L}_p(g) + \lambda \cdot \mathcal{L}_s(f_{0000}, f_{0011}, f_{0100}, f_{0101}, f_{0111}; g),$$

which λ is a hyper-parameter, and the training loss has the same training and inference space, results in more accurate and robust counterfactual strata predictions. We also empirically demonstrate the advantages of CF-MTL in Section 5 over the plug-in estimators in Section 3.3.

Given the probabilities of user-item pairs (u, i) belonging to each counterfactual strata, a more reasonable reward $r_{u,i}$ for incentive policy learning in Section 3.2 should be $r_{u,i} = \mathbb{P}(Y_{0101} \mid x_{u,i}) - c(x_{u,i}) \cdot (\mathbb{P}(Y_{0100} \mid x_{u,i}) + \mathbb{P}(Y_{0111} \mid x_{u,i}))$ when cash bonuses are used as incentives, and $r_{u,i} = \mathbb{P}(Y_{0101} \mid x_{u,i}) - c(x_{u,i}) \cdot \mathbb{P}(Y_{0111} \mid x_{u,i})$ when coupons are used as incentives. Take the coupon incentives as an example, the optimal policy $\pi^*(x_{u,i}; c)$ is

$$\pi^*(x_{u,i}; c) = \begin{cases} 1, \mathbb{P}(Y_{0101} \mid x_{u,i}) > c(x_{u,i}) \cdot \mathbb{P}(Y_{0111} \mid x_{u,i}) + \gamma(\epsilon; c) \\ d, \mathbb{P}(Y_{0101} \mid x_{u,i}) = c(x_{u,i}) \cdot \mathbb{P}(Y_{0111} \mid x_{u,i}) + \gamma(\epsilon; c), \\ 0, \mathbb{P}(Y_{0101} \mid x_{u,i}) < c(x_{u,i}) \cdot \mathbb{P}(Y_{0111} \mid x_{u,i}) + \gamma(\epsilon; c) \end{cases}$$

where d is a value between 0 and 1, and $\gamma(\epsilon; c) \geq 0$ decreases monotonically as the budget ϵ increases.

Compared with the incentive placement policy learning based on uplift modeling, the proposed counterfactual policy learning further takes into account the additional incentive cost $c(x_{u,i})$ arising from "Always Buyers". Given the predicted probabilities $f_{0101}(x_{u,i})$ and $f_{0111}(x_{u,i})$ of the counterfactual strata, the estimated reward is $\hat{r}_{u,i} = f_{0101}(x_{u,i}) - c(x_{u,i}) \cdot f_{0111}(x_{u,i})$, then the counterfactual personalized incentive policy $\pi^\dagger$ is learned by maximizing the estimated policy reward $\hat{R}(\pi)$ with the budget ϵ as constraint

$$\max_{\pi \in \Pi} \hat{R}(\pi) = \frac{1}{|\mathcal{D}|} \sum_{(u,i) \in \mathcal{D}} \pi(x_{u,i}) \hat{r}_{u,i}$$

$$\text{s.t. } B(\pi) = \frac{1}{|\mathcal{D}|} \sum_{(u,i) \in \mathcal{D}} \pi(x_{u,i}) \leq \epsilon.$$

Theorem 4.1 (Policy Reward Lower Bound). *Given coupon costs $0 < c(x_{u,i}) < 1$ for all user-item pairs, when either imputed outcomes or learned propensities are accurate, for any finite[1] policy hypothesis space Π, with probability $1 - \eta$, the true reward of the learned optimal policy using estimated strata probabilities has the lower bound*

$$R(\pi^\dagger) \geq \hat{R}(\pi^\dagger) - \sqrt{\frac{\log\left(\frac{2|\Pi|}{\eta}\right)}{2|\mathcal{D}|^2} \sum_{(u,i) \in \mathcal{D}} \left\{\pi^\S(x_{u,i})[1 + c(x_{u,i})]\right\}^2},$$

where $\pi^\S = \arg\max_{\pi \in \Pi} \sum_{(u,i) \in \mathcal{D}} \left\{\pi(x_{u,i})[1 + c(x_{u,i})]\right\}^2$.

Given the estimated reward $\hat{R}(\pi^\dagger)$ for the learned policy $\pi^\dagger$, we further derive a lower bound on the true reward $R(\pi^\dagger)$ in Theorem 4.1, and the result shows that the discrepancy between the estimated and the true reward will decrease as the sample size increases.

[1]For infinite hypothesis spaces, a similar policy reward lower bound can be derived using Rademacher complexity or VC-dimension of the policy class.

Table 3: Summary of the datasets.

	Yelp	ML-1M	KuaiRec
#Users	25,677	6,040	1,411
#Items	25,815	3,952	3,327
#Interactions	731,671	1,000,209	4,676,570

5 EXPERIMENTS

5.1 Experimental Setup

Dataset and Preprocessing. To evaluate the proposed counterfactual estimation and policy learning methods[2], we conducted extensive experiments on three real-world datasets Yelp [2], ML-1M[3] and KuaiRec[4] [11]. All of these datasets are publicly available and vary in domain, size, and sparsity, with the statistics being summarized in Table 3. Following the previous studies [38, 39, 70, 71], for both Yelp and ML-1M, we binarize the observed ratings to 1 for ratings greater than three, otherwise to 0, as the outcome variable Y, while the rating observation indicators are as the outcome variable C. KuaiRec is a fully exposed dataset from a short video sharing platform, we thus randomly select 20% as observations $C = 1$ and binarize to 1 for *video watching ratio* over 0.6, otherwise to 0.

Next we perform counterfactual strata labeling for each unit. Notably, for all datasets, the units with $(C = 1, Y = 0)$ observations must be "Coupon Taker" from Table 1. To label the remaining units, we pre-trained an Neural Collaborative Filtering (NCF) [19] model to generate the predicted ratings. Specifically, for units with $(C = 1, Y = 1)$ observations, we treat half of the units with the highest predicted ratings as "Always Buyer" and the remaining half as "Coupon Buyer". Since the former will always result in $Y(0) = Y(1) = 1$, which has a relatively higher observed rating on average. By a similar argument, for the units with missing ratings, i.e., $C = 0$, we label the half with highest predicted ratings as "Never Taker" and the remaining half as "Never Buyer".

Baselines and Experimental Details. We compare the proposed methods to the association-based Naive method, which gives incentives based on rating predictions, and also to the widely used uplift modeling methods, which determine the incentive assignment using OR, IPS, and DR estimators. In our experiments, the Neural Collaborative Filtering (NCF) are used as the base model for both regression models and propensity model. The default values of both user and item embedding size are set to 64. All the experiments are implemented on Pytorch with Adam as the optimizer[5]. For all three datasets, we tune the batch size in $\{2048, 4096, 8192\}$, the learning rate in $\{0.001, 0.005, 0.01, 0.05\}$, and the weight decay in $\{1e-7, 1e-6, 1e-5, 1e-4, 1e-3\}$. To evaluate various personalized incentive policy learning approaches, we conduct experiments in two separate incentive scenarios using three datasets, i.e., coupons as incentives and cash bonuses as incentives. Both scenarios yield a reward of 1 for giving the "Coupon Buyer" incentive and a reward of $-c$ for giving the "Always Buyer" incentive. In addition, in the case of cash bonuses as incentives, giving a "Coupon Taker" incentive will also receive a reward of $-c$. Therefore, the learned policy with larger total rewards should be considered more effective.

[2]Code is available at https://github.com/haoxuanli-pku/KDD23-Counterfactual
[3]https://grouplens.org/datasets/movielens/1M/
[4]https://github.com/chongminggao/KuaiRec
[5]For all experiments, we use the GeForce RTX 3090 as the computing resource.

Table 4: Performance comparison of naive, uplift modeling, and proposed counterfactual learning methods, with coupon as incentive and cash as incentive on Yelp, ML-1M, and KuaiRec. We bold the best results within OR, IPS, and DR methods.

Coupon	Yelp					ML-1M					KuaiRec				
Methods	Positive	Negative	Neutral	Reward	RI	Positive	Negative	Neutral	Reward	RI	Positive	Negative	Neutral	Reward	RI
Naive	35,829	31,919	90,220	17,549	-	50,903	34,618	130,524	37,055	-	3,332	57,461	141,235	-19,652	-
OR	58,593	27,389	71,986	47,637	-	76,906	45,425	93,714	58,736	-	67,052	24,988	109,988	57,056	-
CF-OR	**58,635**	**22,557**	76,776	**49,612**	**4.14%**	**78,674**	**40,673**	96,698	**62,404**	**6.24%**	**70,366**	**23,016**	108,646	**61,159**	**7.19%**
IPS	**56,549**	26,282	75,137	46,036	-	80,035	42,770	93,240	62,927	-	82,398	17,775	101,855	75,288	-
CF-IPS	56,470	**22,933**	78,565	**47,296**	**2.73%**	**80,782**	**36,054**	99,209	**66,360**	**5.45%**	**83,694**	**16,857**	101,477	**76,951**	**2.20%**
DR	58,534	27,232	72,202	47,641	-	78,830	44,789	92,426	60,914	-	76,529	19,219	106,280	68,841	-
CF-DR	**58,757**	**22,387**	76,824	**49,802**	**4.53%**	**80,621**	**39,002**	96,422	**65,020**	**6.74%**	**78,506**	**17,346**	106,176	**71,567**	**3.95%**
CF-MTL	**67,686**	**13,397**	76,885	**62,327**	**30.82%**	**85,653**	**30,069**	100,323	**73,625**	**17.00%**	**90,538**	**11,751**	99,739	**85,837**	**14.01%**

Cash	Yelp					ML-1M					KuaiRec				
Methods	Positive	Negative	Neutral	Reward	RI	Positive	Negative	Neutral	Reward	RI	Positive	Negative	Neutral	Reward	RI
Naive	35,829	68,173	53,966	8,559	-	50,903	90,130	75,012	14,851	-	3,332	108,762	89,934	-40,172	-
OR	**58,593**	51,611	47,764	37,948	-	**76,906**	106,324	32,815	34,376	-	67,052	45,011	89,965	49,047	-
CF-OR	56,797	**40,950**	60,221	**40,417**	**6.50%**	76,747	**90,196**	49,102	**40,668**	**18.30%**	**67,171**	**44,917**	89,940	**49,204**	**0.32%**
IPS	**56,549**	49,931	51,488	36,576	-	**80,035**	93,198	42,812	42,755	-	82,398	29,816	89,814	70,471	-
CF-IPS	57,050	**39,374**	61,544	**41,300**	**12.91%**	78,636	**71,076**	66,333	**50,205**	**17.42%**	**82,451**	**29,723**	89,854	**70,561**	**0.12%**
DR	**58,534**	51,162	48,272	38,069	-	**78,830**	100,835	36,380	38,496	-	76,529	35,650	89,849	62,269	-
CF-DR	56,963	**39,120**	61,885	**41,315**	**8.52%**	78,424	**79,109**	57,512	**46,780**	**21.51%**	**76,626**	**35,499**	89,903	**62,426**	**0.25%**
CF-MTL	**67,686**	**25,549**	64,733	**57,466**	**50.95%**	**85,548**	**52,218**	78,279	**64,660**	**51.23%**	**90,187**	**21,608**	89,813	**81,963**	**16.30%**

Note: (a) For coupons as incentives: "Positive" is # "Coupon Buyer" with incentives, "Negative" is # "Alwayer Buyer" with incentives, and "Neutral" is # ("Never Buyer"+"Never Taker"+"Coupon Taker") with incentives; (b) For cash as incentives: "Positive" is #"Coupon Buyer" with incentives, "Negative" is # ("Alwayer Buyer"+"Coupon Taker") with incentives, and "Neutral" is #("Never Buyer"+"Never Taker") with incentives. RI means the relative improvement.

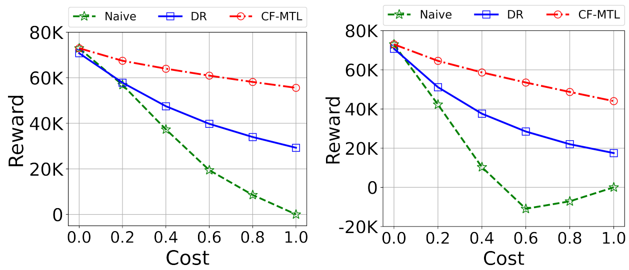
5.2 Performance Comparison

Overall Performance. We compare the proposed counterfactual estimators and multi-task learning approach using coupons as incentives and cash bonuses as incentives scenarios, respectively, and the results are shown in Table 4. We have the following findings. First, the association-based Naive method performs the worst under all scenarios and datasets, while all the uplift modeling-based methods, i.e., OR, IPS, and DR, show significant improvement compared to the Naive method. This validates the importance of estimating causal effects for personalized incentives allocation. Next, all the proposed counterfactual estimators, i.e., CF-OR, CF-IPS, and CF-DR, show significant improvement compared with the uplift modeling-based methods. This is because the proposed counterfactual estimators can identify and estimate the probability that an individual belongs to each of the five counterfactual strata, whereas uplift modeling can only identify and estimate the probability that the individual belongs to the "Coupon Buyer" stratum. Then, the proposed counterfactual multi-task learning approach, i.e., CF-MTL, demonstrates the best performance on all scenarios and datasets. Notably, CF-MTL has a total reward improvement of more than 50% over the optimal uplift model on both Yelp and ML-1M. This is attributed to the CF-MTL simultaneously learning probabilities of individuals belonging to each counterfactual strata and propensities, leading to higher estimation efficiency.

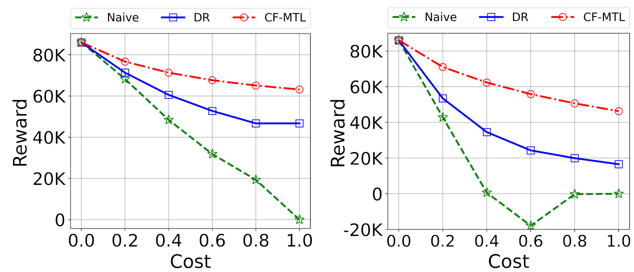
Effects of Varying Cost and Budget. We also study the effect of cost on the performance of various methods as shown in Figure 3, where the Naive method simply uses the ranking of predicted conversion rates for personalized incentives allocation at the training time, while the DR method uses the predicted probability of an individual belonging to the uplift stratum (i.e., "Coupon Buyer") as the reward, and the proposed CF-MTL uses the probability of an individual belonging to the uplift stratum minus the cost corresponding to the incentive scenario as the reward. For a fair comparison, all methods are evaluated using the same reward function in Section 3.1. First, when the cost is zero, all methods perform similarly. As the cost gradually increases, the identification and prediction of causal effects and counterfactual strata are more emphasized and desired, and the proposed CF-MTL demonstrates the optimal performance compared with the Naive and DR methods for all scenarios and datasets. Interestingly, the Naive method has negative total rewards at a cost of 0.6 in the cash as an incentive scenario, which is explained by the gap between correlation and causal effect, and Naive method incorrectly predicts relatively high probabilities for some individuals with negative rewards. Figure 4 further shows the effect of varying budgets on the rewards. As the budget increases, more units with positive estimated rewards are given incentives. The proposed CF-MTL demonstrates the optimal performance, while some methods show a decrease in reward when the budget exceeded 0.4, which is explained by the estimated rewards are positive from its 40th to 50th percentile, while the true rewards are negative.

5.3 Abalation Studies

We conduct ablation studies to validate the effectiveness of the proposed CF-MTL with varying training losses, taking the last two losses in $\mathcal{L}_s$ with respect to the observations $T = 0\&C = 0\&Y = 1$ and $T = 0\&C = 0\&Y = 0$ in Section 4, denoted as L_{001} and L_{000}, respectively. Tables 5 and 6 show the rewards in Yelp and ML-1M for the two incentive scenarios. Theoretically, CF-MTL cannot identify "Coupon Buyers" and "Always Buyers" when both L_{000} and L_{001} are removed. As a result, such a method leads to the significantly worst performance in all scenarios and datasets. When

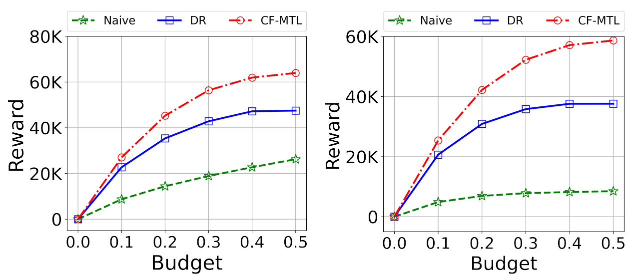


(a) Reward on Yelp with coupon incentives. (b) Reward on Yelp with cash incentives.

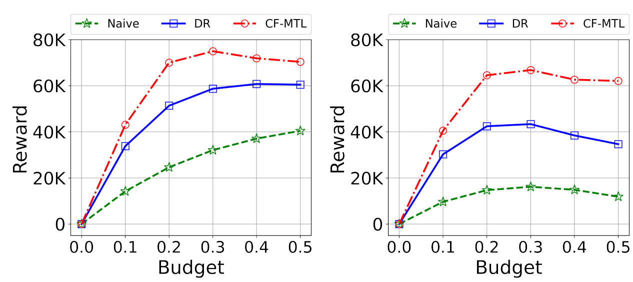


(c) Reward on ML-1M with coupon incentives. (d) Reward on ML-1M with cash incentives.

Figure 3: Effects of varying cost on Yelp and ML-1M.



(a) Reward on Yelp with coupon incentives. (b) Reward on Yelp with cash incentives.



(c) Reward on ML-1M with coupon incentives. (d) Reward on ML-1M with cash incentives.

Figure 4: Effects of varying budget on Yelp and ML-1M.

one of the losses is removed, despite the all counterfactual strata are theoretically identifiable, the total rewards of the learned incentive policies are lower than that of the CF-MTL trained with both losses.

5.4 In-depth Analysis

Now that it is clear that CF-MTL can lead to accurate counterfactual strata predictions, we further investigate the effect of different attributes of users and items on the counterfactual strata to which they are subjected. We show the group-wise strata prediction results

Table 5: Ablation study of training loss on Yelp.

Yelp	Training loss		Incentive scenario	
Methods	$\mathcal{L}_{001}$	$\mathcal{L}_{000}$	$\text{Reward}_{\text{Coupon}}$	$\text{Reward}_{\text{Cash}}$
CF-MTL w/o $\mathcal{L}_{001}$ $\mathcal{L}_{000}$	×	×	20,948	13,991
CF-MTL w/o $\mathcal{L}_{000}$	✓	×	58,849	54,064
CF-MTL w/o $\mathcal{L}_{001}$	×	✓	52,244	48,033
CF-MTL	✓	✓	**62,327**	**57,466**

Table 6: Ablation study of training loss on ML-1M.

ML-1M	Training loss		Incentive scenario	
Methods	$\mathcal{L}_{001}$	$\mathcal{L}_{000}$	$\text{Reward}_{\text{Coupon}}$	$\text{Reward}_{\text{Cash}}$
CF-MTL w/o $\mathcal{L}_{001}$ $\mathcal{L}_{000}$	×	×	39,120	29,815
CF-MTL w/o $\mathcal{L}_{000}$	✓	×	70,923	61,235
CF-MTL w/o $\mathcal{L}_{001}$	×	✓	68,883	59,439
CF-MTL	✓	✓	**73,625**	**64,660**

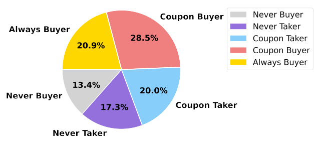
of CF-MTL on Yelp and ML-1M in Figures 5 and 6, respectively. We label a user as a "Frequent buyer" if he buys more items than the median of all users, and as a "Non-frequent buyer" otherwise. Similarly, we label an item as "Popular item" if it is sold above the median, and "Non-popular item" otherwise. We find that users with more frequent purchases and items with higher popularity tend to be "Always buyers". This is consistent with our intuition that users are more likely to buy highly popular items, regardless of whether they are incentivized or not. In addition, high popularity items are more likely to lead to "Coupon buyers", while low popularity items are more likely to lead to "Never buyers" and "Never takers", which is also explained by the presence of item popularity bias.

In addition, we present the confusion matrices of CF-MTL for the five counterfactual strata in Yelp and ML-1M in Tables 7 and 8, respectively. For each counterfactual stratum in the rows, we bold the highest predicted probability and underline the second highest predicted probability. On both datasets, CF-MTL predicted "Coupon Buyer" with approximately 70% accuracy and predicted "Coupon Taker" and "Always Buyer" both have a recall rate of more than 30%. The superior prediction performance of CF-MTL is further demonstrated by the diagonals with high probabilities from the two confusion matrices, which is attributed to the effectiveness of CF-MTL's joint training of the five counterfactual strata prediction models and propensity model in the entire-space.

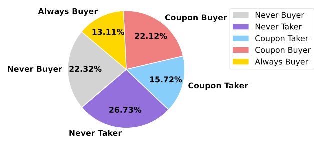
6 RELATED WORK

In this section, we review the previous related works, including uplift modeling methods and causal learning for recommendation.

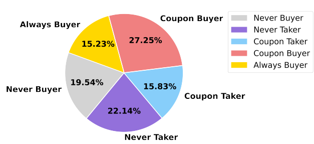
Uplift Modeling. Uplift modeling estimates the conditional average treatment effect (CATE) for individuals with specific features and is widely adopted in economics [3, 48, 58], precision medicine [25], decision making [14], and advertising placement [12, 17]. Many methods have been proposed for estimating the CATE, such as outcome regression (OR) [18, 24], inverse propensity scoring (IPS) [20, 22, 46], and doubly robust (DR) [27, 47] methods. Incorporating machine learning algorithms can further enhance the estimation accuracy of CATE, such as tree-based methods, including Bayesian Additive Regression Trees (BART) [7], Causal



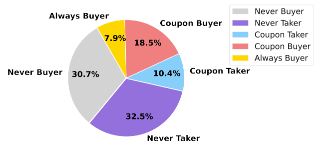
(a) Frequent buyers, popular items.



(b) Frequent buyers, non-popular items.

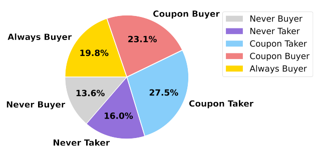


(c) Non-frequent buyers, popular items.

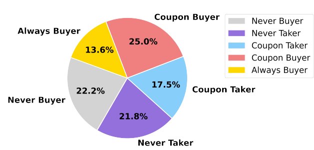


(d) Non-frequent buyers, non-popular items.

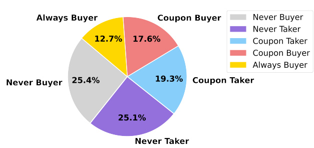
Figure 5: Group-wise strata prediction results on Yelp.



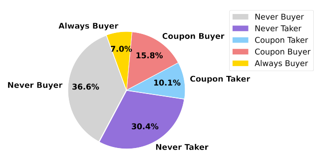
(a) Frequent buyers, popular items.



(b) Frequent buyers, non-popular items.



(c) Non-frequent buyers, popular items.



(d) Non-frequent buyers, non-popular items.

Figure 6: Group-wise strata prediction results on ML-1M.

Forest (CF) [59], and neural network-based methods, including Balancing Neural Network (BNN) [26], CounterFactual Regression (CFR) [55], Perfect Match (PM) [54], X-learner [29], DragonNet [56], and DESCN [77]. However, these uplift modeling methods can only estimate CATEs on subgroups, instead of the individual treatment effects [34]. In this paper, we extend the above methods to identify and estimate the probability of an individual belonging to each counterfactual stratum.

Causal Recommendation. Personalized incentives (e.g., sending coupons, or giving cash rewards) to users can be effective in increasing conversion rates [13] and Gross Commodity Volume (GMV) [42, 49, 68]. Previous data-driven recommendation methods use associations to predict conversion rates [4, 15, 61]. However, they fail to combat various biases and confounding in the collected data [63], such as popularity bias [76], selection bias [53], exposure bias [28], conformity bias [40], and position bias [1]. To tackle the above issues, many causal intervention-inspired methods have been developed [6, 37, 66, 69], such as outcome regression methods [43, 57, 74], propensity-based weighting methods [33, 53, 73], doubly robust learning methods [8, 16, 31, 35, 49, 64], and multiple robust learning method [30]. The causal prediction accuracy can be further improved by introducing a few unbiased ratings [3, 5, 32, 36, 65]. In addition, many entire-space multi-task learning methods have the same training space and inference space, and joint training of models empirically leads to better performance, such as Entire Space Multi-Task Model (ESMM) [42], Multi-gate

Table 7: Confusion matrix of CF-MTL on Yelp.

Yelp	Strata	$\hat{p}_{0000}$	$\hat{p}_{0011}$	$\hat{p}_{0100}$	$\hat{p}_{0101}$	$\hat{p}_{0111}$
Never Buyer	Y_{0000}	**0.553**	0.075	0.024	0.307	0.040
Never Taker	Y_{0011}	0.058	**0.596**	0.195	0.029	0.122
Coupon Taker	Y_{0100}	0.046	0.225	**0.325**	0.131	0.273
Coupon Buyer	Y_{0101}	0.229	0.012	0.019	**0.695**	0.045
Always Buyer	Y_{0111}	0.061	0.193	0.288	0.126	**0.333**

Table 8: Confusion matrix of CF-MTL on ML-1M.

ML-1M	Strata	$\hat{p}_{0000}$	$\hat{p}_{0011}$	$\hat{p}_{0100}$	$\hat{p}_{0101}$	$\hat{p}_{0111}$
Never Buyer	Y_{0000}	**0.644**	0.070	0.058	0.163	0.064
Never Taker	Y_{0011}	0.073	**0.685**	0.170	0.007	0.065
Coupon Taker	Y_{0100}	0.069	0.151	**0.405**	0.136	0.239
Coupon Buyer	Y_{0101}	0.200	0.003	0.028	**0.707**	0.061
Always Buyer	Y_{0111}	0.092	0.093	**0.327**	0.170	0.319

Mixture-of-Experts (MMoE) [41], Multi_IPW [75], and ESCM^2 [60]. However, these methods are unable to make counterfactual predictions for individuals, which uses the observed outcomes of individuals in a more fine-grained way for counterfactual outcome prediction [9]. Despite being rarely discussed, some recent counterfactual learning studies have been devoted to making Top-N recommendations [72], mitigating click-bait issues [62], estimating post-click conversions [13, 44], eliminating the popularity bias [67], and bursting filter bubbles [10]. In this paper, we explore novel personalized incentive scenarios, and extend the above causal and multi-task learning methods to perform individual counterfactual strata predictions for more rational incentive allocation.

7 CONCLUSION

This paper studies the personalized incentive policy learning from an individualized counterfactual perspective. First, we reformulate the personalized incentive policy learning problem based on the joint potential outcomes for individuals, and reveal the limitations of previous uplift modeling methods. We formally discuss the extra incentive costs incurred by "Always Buyers" and "Coupon Takers" in two incentive scenarios, and uplift modeling fails to identify and predict these two strata. Next, we propose counterfactual estimators, i.e., CF-OR, CF-IPS, and CF-DR, to identify and estimate the probability that an individual belongs to various counterfactual strata. By theoretical analysis, we prove the double robustness property of the CF-DR estimator. Then, based on the proposed counterfactual identification methods, we further propose a counterfactual entire-space multi-task learning approach, named CF-MTL, to accurately predict the counterfactual strata probabilities and perform individualized incentive policy learning. We also theoretically derive a lower bound on the reward of the learned policy. Extensive experiments are conducted on three real-world datasets with two common personalized incentive scenarios, and the results show the effectiveness of the proposed approaches for counterfactual prediction and personalized incentive policy learning.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (No. 12201629) and National Key R&D Program of China (No. 2020AAA0106300).

REFERENCES

[1] Qingyao Ai, Keping Bi, Cheng Luo, Jiafeng Guo, and W Bruce Croft. 2018. Unbiased learning to rank with unbiased propensity estimation. In *SIGIR*.

[2] Nabiha Asghar. 2016. Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362* (2016).

[3] Stephen Bonner and Flavian Vasile. 2018. Causal embeddings for recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 104–112.

[4] Chih-Yao Chang, Xing Tang, Bo-Wen Yuan, Jui-Yang Hsia, Zhirong Liu, Zhenhua Dong, Xiuqiang He, and Chih-Jen Lin. 2020. AutoConjunction: Adaptive Model-based Feature Conjunction for CTR Prediction. In *MDM*. IEEE, 202–209.

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

[6] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems* 41, 3 (2023), 1–39.

[7] Hugh A Chipman, Edward I George, and Robert E McCulloch. 2010. BART: Bayesian additive regression trees. *The Annals of Applied Statistics* 4, 1 (2010), 266–298.

[8] Quanyu Dai, Haoxuan Li, Peng Wu, Zhenhua Dong, Xiao-Hua Zhou, Rui Zhang, Xiuqiang He, Rui Zhang, and Jie Sun. 2022. A Generalized Doubly Robust Learning Framework for Debiasing Post-Click Conversion Rate Prediction. In *KDD*.

[9] Zhenhua Dong, Hong Zhu, Pengxiang Cheng, Xinhua Feng, Guohao Cai, Xiuqiang He, Jun Xu, and Jirong Wen. 2020. Counterfactual learning for recommender system. In *Fourteenth ACM Conference on Recommender Systems*. 568–569.

[10] Chongming Gao, Wenqiang Lei, Jiawei Chen, Shiqi Wang, Xiangnan He, Shijun Li, Biao Li, Yuan Zhang, and Peng Jiang. 2022. Cirs: Bursting filter bubbles by counterfactual interactive recommender system. *arXiv preprint arXiv:2204.01266* (2022).

[11] Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. 2022. KuaiRec: A Fully-observed Dataset and Insights for Evaluating Recommender Systems. In *CIKM* (Atlanta, GA, USA). 11 pages.

[12] Brett R Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. 2019. A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science* 38, 2 (2019), 193–225.

[13] Tiankai Gu, Kun Kuang, Hong Zhu, Jingjie Li, Zhenhua Dong, Wenjie Hu, Zhenguo Li, Xiuqiang He, and Yue Liu. 2021. Estimating True Post-Click Conversion via Group-stratified Counterfactual Inference. In *ADKDD*.

[14] Leo Guelman, Montserrat Guillén, and Ana M Pérez-Marín. 2015. A decision support framework to implement optimal personalized marketing interventions. *Decision Support Systems* 72 (2015), 24–32.

[15] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, Xiuqiang He, and Zhenhua Dong. 2018. Deepfm: An end-to-end wide & deep learning framework for CTR prediction. *arXiv preprint arXiv:1804.04950* (2018).

[16] 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. In *SIGIR*.

[17] Pierre Gutierrez and Jean-Yves Gérardy. 2017. Causal inference and uplift modelling: A review of the literature. In *Predictive Applications and APIs*. 1–13.

[18] Behram Hansotia and Brad Rukstales. 2002. Incremental value modeling. *Journal of Interactive Marketing* 16 (2002), 35–46.

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

[20] M.A. Hernán and J. M. Robins. 2020. *Causal Inference: What If.* Boca Raton: Chapman and Hall/CRC.

[21] Paul W Holland. 1986. Statistics and causal inference. *Journal of the American statistical Association* (1986), 945–960.

[22] D. G. Horvitz and D. J. Thompson. 1952. A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association* 47 (1952), 663–685.

[23] G. W. Imbens and D. B. Rubin. 2015. *Causal Inference For Statistics Social and Biomedical Science.* Cambridge University Press.

[24] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press.

[25] Maciej Jaskowski and Szymon Jaroszewicz. 2012. Uplift modeling for clinical trial data. In *ICML Workshop on Clinical Data Analysis*, Vol. 46. 79–95.

[26] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*. PMLR, 3020–3029.

[27] Edward H. Kennedy. 2020. Optimal doubly robust estimation of heterogeneous causal effects. *https://arxiv.org/abs/2004.14497v1* (2020).

[28] Sami Khenissi and Olfa Nasraoui. 2020. Modeling and Counteracting Exposure Bias in Recommender Systems. *arXiv:2001.04832* (2020).

[29] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences* 116, 10 (2019), 4156–4165.

[30] Haoxuan Li, Quanyu Dai, Yuru Li, Yan Lyu, Zhenhua Dong, Xiao-Hua Zhou, and Peng Wu. 2023. Multiple Robust Learning for Recommendation. In *AAAI*.

[31] Haoxuan Li, Yan Lyu, Chunyuan Zheng, and Peng Wu. 2023. TDR-CL: Targeted Doubly Robust Collaborative Learning for Debiased Recommendations. In *ICLR*.

[32] Haoxuan Li, Yanghao Xiao, Chunyuan Zheng, and Peng Wu. 2023. Balancing Unobserved Confounding with a Few Unbiased Ratings in Debiased Recommendations. In *WWW*.

[33] Haoxuan Li, Yanghao Xiao, Chunyuan Zheng, Peng Wu, and Peng Cui. 2023. Propensity Matters: Measuring and Enhancing Balancing for Recommendation. In *ICML*.

[34] Haoxuan Li, Chunyuan Zheng, Yixiao Cao, Zhi Geng, Yue Liu, and Peng Wu. 2023. Trustworthy Policy Learning under the Counterfactual No-Harm Criterion. In *ICML*.

[35] Haoxuan Li, Chunyuan Zheng, and Peng Wu. 2023. StableDR: Stabilized Doubly Robust Learning for Recommendation on Data Missing Not at Random. In *ICLR*.

[36] 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. In *SIGIR*. 831–840.

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

[38] Huafeng Liu, Liping Jing, Jingxuan Wen, Zhicheng Wu, Xiaoyi Sun, Jiaqi Wang, Lin Xiao, and Jian Yu. 2020. Deep global and local generative model for recommendation. In *Proceedings of The Web Conference 2020*. 551–561.

[39] Huafeng Liu, Jingxuan Wen, Liping Jing, and Jian Yu. 2019. Deep generative ranking for personalized recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 34–42.

[40] Yiming Liu, Xuezhi Cao, and Yong Yu. 2016. Are You Influenced by Others When Rating?: Improve Rating Prediction by Conformity Modeling. In *RecSys*. 269–272.

[41] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *KDD*. 1930–1939.

[42] Xiao Ma, Liqin Zhao, Guan Huang, Zhi Wang, Zelin Hu, Xiaoqiang Zhu, and Kun Gai. 2018. Entire Space Multi-Task Model: An Effective Approach for Estimating Post-Click Conversion Rate. In *SIGIR*.

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

[44] Rishabh Mehrotra, Prasanta Bhattacharya, and Mounia Lalmas. 2020. Inferring the Causal Impact of New Track Releases on Music Recommendation Platforms through Counterfactual Predictions. In *RecSys*. 687–691.

[45] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. 2018. *Foundations of Machine Learning.* MIT Press.

[46] X Nie and S Wager. 2021. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 108 (2021), 4156–4165.

[47] J.M. Robins, A. Rotnitzky, and L.P. Zhao. 1994. Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* 89 (1994), 846–866.

[48] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.

[49] Yuta Saito. 2020. Doubly robust estimator for ranking metrics with post-click conversions. In *Fourteenth ACM Conference on Recommender Systems*. 92–100.

[50] Yuta Saito, Hayato Sakata, and Kazuhide Nakata. 2019. Doubly Robust Prediction and Evaluation Methods Improve Uplift Modeling for Observational Data. In *SIAM*.

[51] Yuta Saito, Hayato Sakata, and Kazuhide Nakata. 2020. Cost-effective and stable policy optimization algorithm for uplift modeling with multiple treatments. In *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, 406–414.

[52] Masahiro Sato, Janmajay Singh, Sho Takemori, Takashi Sonoda, Qian Zhang, and Tomoko Ohkuma. 2019. Uplift-based evaluation and optimization of recommenders. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 296–304.

[53] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In *ICML*.

[54] Patrick Schwab, Lorenz Linhardt, and Walter Karlen. 2018. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656* (2018).

[55] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*. PMLR, 3076–3085.

[56] Claudia Shi, David Blei, and Victor Veitch. 2019. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems* 32 (2019).

[57] Harald Steck. 2010. Training and testing of recommender systems on data missing not at random. In *KDD*.

[58] H. Varian. 2016. Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences* 113 (2016), 7310 – 7315.

[59] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242.

[60] Hao Wang, Tai-Wei Chang, Tianqiao Liu, Jianmin Huang, Zhichao Chen, Chao Yu, Ruopeng Li, and Wei Chu. 2022. ESCM^2: Entire Space Counterfactual Multi-Task Model for Post-Click Conversion Rate Estimation. In *SIGIR*.

[61] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. 2017. Deep & cross network for ad click predictions. In *ADKDD*. 1–7.

[62] Wenjie Wang, Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Clicks can be cheating: Counterfactual recommendation for mitigating clickbait issue. In *SIGIR*. 1288–1297.

[63] Wenjie Wang, Yang Zhang, Haoxuan Li, Peng Wu, Fuli Feng, and Xiangnan He. 2023. Causal Recommendation: Progresses and Future Directions. In *Tutorial on SIGIR*.

[64] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2019. Doubly Robust Joint Learning for Recommendation on Data Missing Not at Random. In *ICML*.

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

[66] 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. *NeurIPS* (2020), 1854–1864.

[67] Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. 2021. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *SIGKDD*. 1791–1800.

[68] Liang Wu, Diane Hu, Liangjie Hong, and Huan Liu. 2018. Turning Clicks into Purchases: Revenue Optimization for Product Search in E-Commerce. In *SIGIR*.

[69] Peng Wu, Haoxuan Li, Yuhao Deng, Wenjie Hu, Quanyu Dai, Zhenhua Dong, Jie Sun, Rui Zhang, and Xiao-Hua Zhou. 2022. On the Opportunity of Causal Learning in Recommendation Systems: Foundation, Estimation, Prediction and Challenges. In *IJCAI*.

[70] Yao Wu, Christopher DuBois, Alice X Zheng, and Martin Ester. 2016. Collaborative denoising auto-encoders for top-n recommender systems. In *Proceedings of the ninth ACM international conference on web search and data mining*. 153–162.

[71] Yao Wu, Xudong Liu, Min Xie, Martin Ester, and Qing Yang. 2016. CCCF: Improving collaborative filtering via scalable user-item co-clustering. In *Proceedings of the ninth ACM international conference on web search and data mining*. 73–82.

[72] Mengyue Yang, Quanyu Dai, Zhenhua Dong, Xu Chen, Xiuqiang He, and Jun Wang. 2021. Top-N Recommendation with Counterfactual User Preference Simulation. In *CIKM*. 2342–2351.

[73] Bowen Yuan, Jui-Yang Hsia, Meng-Yuan Yang, Hong Zhu, Chih-Yao Chang, Zhenhua Dong, and Chih-Jen Lin. 2019. Improving ad click prediction by considering non-displayed events. In *CIKM*. 329–338.

[74] Bowen Yuan, Yaxu Liu, Jui-Yang Hsia, Zhenhua Dong, and Chih-Jen Lin. 2020. Unbiased Ad Click Prediction for Position-aware Advertising Systems. In *Fourteenth ACM Conference on Recommender Systems*. 368–377.

[75] Wenhao Zhang, Wentian Bao, Xiao-Yang Liu, Keping Yang, Quan Lin, Hong Wen, and Ramin Ramezani. 2020. Large-scale Causal Approaches to Debiasing Post-click Conversion Rate Estimation with Multi-task Learning. In *WWW*. 2775–2781.

[76] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal intervention for leveraging popularity bias in recommendation. In *SIGIR*. 11–20.

[77] Kailiang Zhong, Fengtong Xiao, Yan Ren, Yaorong Liang, Wenqing Yao, Xiaofeng Yang, and Ling Cen. 2022. DESCN: Deep Entire Space Cross Networks for Individual Treatment Effect Estimation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4612–4620.

A PROOF

Lemma A.1 (Hoeffding's inequality). *Let $X_1, \ldots, X_m$ be independent random variables with X_i taking values in $[a_i, b_i]$ for all $i \in [m]$. Then, for any $\epsilon > 0$, the following inequalities hold for $S_m = \sum_{i=1}^m X_i$:*

$$\mathbb{P}\left[S_m - \mathbb{E}\left[S_m\right] \geq \epsilon\right] \leq e^{-2\epsilon^2/\sum_{i=1}^m (b_i - a_i)^2}$$

$$\mathbb{P}\left[S_m - \mathbb{E}\left[S_m\right] \leq -\epsilon\right] \leq e^{-2\epsilon^2/\sum_{i=1}^m (b_i - a_i)^2}$$

Proof. The proof can be found in Appendix D.1 of [45]. □

Theorem 3.1 (Bias of CF-DR Estimator). *Given imputed outcomes $\hat{\mu}_0(x_{u,i})$, $\hat{\mu}_{01|1}(x_{u,i})$, and learned propensities $\hat{e}_{u,i} > 0$ for all user-item pairs, the bias of the CF-DR estimator is*

$$\begin{aligned}\operatorname{Bias}(\hat{p}_{0111}^{DR}(x_{u,i})) = \Bigg| & \frac{(e_{u,i} - \hat{e}_{u,i})(Y_{u,i}(0) - \hat{\mu}_0(x_{u,i}))}{1 - \hat{e}_{u,i}} \\ & + \frac{(e_{u,i} - \hat{e}_{u,i})[(1 - C_{u,i}(1))Y_{u,i}(1) - \hat{\mu}_{01|1}(x_{u,i})]}{\hat{e}_{u,i}} \Bigg|.\end{aligned}$$

Proof. The CF-DR estimator is given as

$$\begin{aligned}\hat{p}_{0111}^{DR}(x_{u,i}) = \hat{\mu}_0(x_{u,i}) & + \frac{(1 - T_{u,i})(Y_{u,i}(0) - \hat{\mu}_0(x_{u,i}))}{1 - \hat{e}_{u,i}} \\ & - \hat{\mu}_{01|1} - \frac{T_{u,i}[(1 - C_{u,i}(1))Y_{u,i}(1) - \hat{\mu}_{01|1}(x_{u,i})]}{\hat{e}_{u,i}}.\end{aligned}$$

By definition, the bias of the CF-DR estimator is

$$\operatorname{Bias}(\hat{p}_{0111}^{DR}(x_{u,i})) = \left| Y_{u,i}(0) - (1 - C_{u,i}(1))Y_{u,i}(1) - \mathbb{E}_T[\hat{p}_{0111}^{DR}(x_{u,i})] \right|.$$

The second term on the right hand side can be expanded as

$$\begin{aligned}\mathbb{E}_T[\hat{p}_{0111}^{DR}(x_{u,i})] = & \hat{\mu}_0(x_{u,i}) + \frac{(1 - e_{u,i})(Y_{u,i}(0) - \hat{\mu}_0(x_{u,i}))}{1 - \hat{e}_{u,i}} \\ & - \hat{\mu}_{01|1} - \frac{e_{u,i}[(1 - C_{u,i}(1))Y_{u,i}(1) - \hat{\mu}_{01|1}(x_{u,i})]}{\hat{e}_{u,i}}.\end{aligned}$$

By substituting the above formula into the bias of CF-DR, we have

$$\begin{aligned}\operatorname{Bias}(\hat{p}_{0111}^{DR}(x_{u,i})) = & \Bigg| Y_{u,i}(0) - \hat{\mu}_0(x_{u,i}) - \frac{(1 - e_{u,i})(Y_{u,i}(0) - \hat{\mu}_0(x_{u,i}))}{1 - \hat{e}_{u,i}} \\ & - (1 - C_{u,i}(1))Y_{u,i}(1) + \hat{\mu}_{01|1} + \frac{e_{u,i}[(1 - C_{u,i}(1))Y_{u,i}(1) - \hat{\mu}_{01|1}(x_{u,i})]}{\hat{e}_{u,i}} \Bigg| \\ = & \Bigg| \frac{(e_{u,i} - \hat{e}_{u,i})(Y_{u,i}(0) - \hat{\mu}_0(x_{u,i}))}{1 - \hat{e}_{u,i}} \\ & + \frac{(e_{u,i} - \hat{e}_{u,i})[(1 - C_{u,i}(1))Y_{u,i}(1) - \hat{\mu}_{01|1}(x_{u,i})]}{\hat{e}_{u,i}} \Bigg|\end{aligned}$$

□

Corollary 3.2 (Double Robustness). *The CF-DR estimator is unbiased when either imputed outcomes $\hat{\mu}_0(x_{u,i})$ and $\hat{\mu}_{01|1}(x_{u,i})$ or learned propensities $\hat{e}_{u,i} > 0$ are accurate for all user-item pairs.*

Proof. The bias of CF-DR estimator is equivalent to

$$\operatorname{Bias}(\hat{p}_{0111}^{DR}(x_{u,i})) = \left| e_{u,i} - \hat{e}_{u,i} \right| \cdot \left| \frac{Y_{u,i}(0) - \hat{\mu}_0(x_{u,i})}{1 - \hat{e}_{u,i}} + \frac{(1 - C_{u,i}(1))Y_{u,i}(1) - \hat{\mu}_{01|1}(x_{u,i})}{\hat{e}_{u,i}} \right|,$$

where the first term on the right hind side equals to 0, when the learned propensities $\hat{e}_{u,i} > 0$ are accurate for all user-item pairs, i.e., $\hat{e}_{u,i} = e_{u,i}$. The second term on the right hind side equals to 0, when both of the imputed outcomes $\mu_0(x_{u,i})$ and $\hat{\mu}_{01|1}(x_{u,i})$ are accurate for all user-item pairs, i.e., $\hat{\mu}_0(x_{u,i}) = Y_{u,i}(0)$ and $\hat{\mu}_{01|1}(x_{u,i}) = \mathbb{I}(C_{u,i}(1) = 0, Y_{u,i}(1) = 1)$. □

Theorem 3.3 (Variance of CF-DR Estimator). *Given imputed outcomes $\hat{\mu}_0(x_{u,i})$, $\hat{\mu}_{01|1}(x_{u,i})$, and learned propensities $\hat{e}_{u,i} > 0$ for all user-item pairs, the variance of the CF-DR estimator is*

$$\begin{aligned}\operatorname{Var}(\hat{p}_{0111}^{DR}(x_{u,i})) = e_{u,i}(1 - e_{u,i}) \Bigg[& \frac{Y_{u,i}(0) - \hat{\mu}_0(x_{u,i})}{1 - \hat{e}_{u,i}} \\ & + \frac{(1 - C_{u,i}(1))Y_{u,i}(1) - \hat{\mu}_{01|1}(x_{u,i})}{\hat{e}_{u,i}} \Bigg]^2.\end{aligned}$$

Proof. The CF-DR estimator is given as

$$\begin{aligned}\hat{p}_{0111}^{DR}(x_{u,i}) = \hat{\mu}_0(x_{u,i}) & + \frac{(1 - T_{u,i})(Y_{u,i}(0) - \hat{\mu}_0(x_{u,i}))}{1 - \hat{e}_{u,i}} \\ & - \hat{\mu}_{01|1} - \frac{T_{u,i}[(1 - C_{u,i}(1))Y_{u,i}(1) - \hat{\mu}_{01|1}(x_{u,i})]}{\hat{e}_{u,i}}.\end{aligned}$$

The variance of the CF-DR estimator on the treatment assignment $t_{u,i}$ is

$$\begin{aligned}\operatorname{Var}(\hat{p}_{0111}^{DR}(x_{u,i})) = & \operatorname{Var}(T_{u,i}) \Bigg[\frac{Y_{u,i}(0) - \hat{\mu}_0(x_{u,i})}{1 - \hat{e}_{u,i}} \\ & + \frac{(1 - C_{u,i}(1))Y_{u,i}(1) - \hat{\mu}_{01|1}(x_{u,i})}{\hat{e}_{u,i}} \Bigg]^2 \\ = & e_{u,i}(1 - e_{u,i}) \Bigg[\frac{Y_{u,i}(0) - \hat{\mu}_0(x_{u,i})}{1 - \hat{e}_{u,i}} + \frac{(1 - C_{u,i}(1))Y_{u,i}(1) - \hat{\mu}_{01|1}(x_{u,i})}{\hat{e}_{u,i}} \Bigg]^2.\end{aligned}$$

□

Theorem 3.4 (Tail Bound of CF-DR Estimator). *Given imputed outcomes $\hat{\mu}_0(x_{u,i})$, $\hat{\mu}_{01|1}(x_{u,i})$, and learned propensities $\hat{e}_{u,i} > 0$, with probability $1 - \eta$, the deviation of the CF-DR estimator from its expectation has the following tail bound*

$$\left| \hat{p}_{0111}^{DR} - \mathbb{E}_T(\hat{p}_{0111}^{DR}) \right| \leq \sqrt{\frac{\log\left(\frac{2}{\eta}\right)}{2|\mathcal{D}|^2} \sum_{(u,i) \in \mathcal{D}} \left\{ \frac{\tilde{\mu}_0(x_{u,i})}{1 - \hat{e}_{u,i}} + \frac{\tilde{\mu}_{01|1}(x_{u,i})}{\hat{e}_{u,i}} \right\}^2},$$

where $\tilde{\mu}_0 = Y(0) - \hat{\mu}_0$ and $\tilde{\mu}_{01|1} = (1 - C(1))Y(1) - \hat{\mu}_{01|1}$.

Proof. Since we assume that each treatment $T_{u,i}$ follows a Bernoulli distribution with probability $e_{u,i}$, we can rewrite the random variable $\hat{p}_{0111}^{DR}(x_{u,i})$ as follows

$$\begin{cases} \mathbb{P}\left(\hat{p}_{0111}^{DR}(x_{u,i}) = \hat{\mu}_0(x_{u,i}) - \hat{\mu}_{01|1} - \tilde{\mu}_{01|1}(x_{u,i})/\hat{e}_{u,i}\right) = e_{u,i}, \\ \mathbb{P}\left(\hat{p}_{0111}^{DR}(x_{u,i}) = \hat{\mu}_0(x_{u,i}) - \hat{\mu}_{01|1} + \tilde{\mu}_0(x_{u,i})/(1 - \hat{e}_{u,i})\right) = 1 - e_{u,i}, \end{cases}$$

where takes its value in

$$\left[\hat{\mu}_0(x_{u,i}) - \hat{\mu}_{01|1} - \tilde{\mu}_{01|1}(x_{u,i})/\hat{e}_{u,i}, \hat{\mu}_0(x_{u,i}) - \hat{\mu}_{01|1} + \tilde{\mu}_0(x_{u,i})/1 - \hat{e}_{u,i}\right]$$

of size $\tilde{\mu}_{01|1}(x_{u,i})/\hat{e}_{u,i} + \tilde{\mu}_0(x_{u,i})/(1 - \hat{e}_{u,i})$ with probability 1.

The independence of $\left\{\hat{p}^{DR}_{0111}(x_{u,i}) \mid u,i \in \mathcal{D}\right\}$ can be directly deduced from the independence of $\left\{T_{u,i} \mid u,i \in \mathcal{D}\right\}$. Therefore, according to the Hoeffding's inequality in Lemma A.1, for any $\epsilon > 0$, we have the following inequality

$$\mathbb{P}\left(\left|\frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}}\hat{p}^{DR}_{0111}(x_{u,i}) - \mathbb{E}_T\left[\frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}}\hat{p}^{DR}_{0111}(x_{u,i})\right]\right| \geq \epsilon\right)$$
$$\leq 2\exp\left(\frac{-2\epsilon^2|\mathcal{D}|^2}{\sum_{(u,i)\in\mathcal{D}}\{\tilde{\mu}_{01|1}(x_{u,i})/\hat{e}_{u,i} + \tilde{\mu}_0(x_{u,i})/(1-\hat{e}_{u,i})\}^2}\right).$$

Setting the right hand side of the inequality to η and solving for ϵ complete the proof. □

Theorem 4.1 (Policy Reward Lower Bound). *Given coupon costs $0 < c(x_{u,i}) < 1$ for all user-item pairs, when either imputed outcomes or learned propensities are accurate, for any finite policy hypothesis space Π, with probability $1-\eta$, the true reward of the learned optimal policy using estimated strata probabilities has the lower bound*

$$R(\pi^\dagger) \geq \hat{R}(\pi^\dagger) - \sqrt{\frac{\log\left(\frac{2|\Pi|}{\eta}\right)}{2|\mathcal{D}|^2}\sum_{(u,i)\in\mathcal{D}}\left\{\pi^{\S}(x_{u,i})[1+c(x_{u,i})]\right\}^2},$$

where $\pi^{\S} = \arg\max_{\pi\in\Pi}\sum_{(u,i)\in\mathcal{D}}\left\{\pi(x_{u,i})[1+c(x_{u,i})]\right\}^2$.

Proof. We first show that for any given policy $\pi \in \Pi$, with probability $1-\eta$, the deviation of the CF-DR estimator from its expectation has the following tail bound

$$\left|\hat{R}(\pi) - \mathbb{E}_T\left[\hat{R}(\pi)\right]\right| \leq \sqrt{\frac{\log\left(\frac{2}{\eta}\right)}{2|\mathcal{D}|^2}\sum_{(u,i)\in\mathcal{D}}\left\{\pi(x_{u,i})[1+c(x_{u,i})]\right\}^2}.$$

Note that $\hat{r}_{u,i} = \hat{p}_{0101}(x_{u,i}) - c(x_{u,i})\cdot(\hat{p}_{0100}(x_{u,i}) + \hat{p}_{0111}(x_{u,i}))$ takes its value in $[-c(x_{u,i}), 1]$ with probability 1, thus $\pi(x_{u,i})\hat{r}_{u,i}$ takes its value in $[-\pi(x_{u,i})c(x_{u,i}), \pi(x_{u,i})]$ with probability 1. For a given policy $\pi \in \Pi$, the independence of $\left\{\pi(x_{u,i})\hat{r}_{u,i} \mid u,i \in \mathcal{D}\right\}$ can be directly deduced from the independence of $\left\{\hat{p}^{DR}_{0111}(x_{u,i}) \mid u,i \in \mathcal{D}\right\}$ as shown in the proof of Theorem 4.1. Therefore, according to the Hoeffding's inequality in Lemma A.1, for any $\tilde{\epsilon} > 0$, we have the following inequality

$$\mathbb{P}\left(\left|\frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}}\pi(x_{u,i})\hat{r}_{u,i} - \mathbb{E}_T\left[\frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}}\pi(x_{u,i})\hat{r}_{u,i}\right]\right| \geq \tilde{\epsilon}\right)$$
$$\leq 2\exp\left(\frac{-2\tilde{\epsilon}^2|\mathcal{D}|^2}{\sum_{(u,i)\in\mathcal{D}}\left\{\pi(x_{u,i})[1+c(x_{u,i})]\right\}^2}\right)$$

Setting the right hand side of the inequality to η and solving for $\tilde{\epsilon}$ complete the proof.

Let $\pi^\dagger$ be the learned policy derived by optimizing the empirical form that

$$\max_{\pi\in\Pi}\ \hat{R}(\pi) = \frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}}\pi(x_{u,i})\hat{r}_{u,i}$$
$$\text{s.t.}\ \frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}}\pi(x_{u,i}) \leq \epsilon,$$

where $\hat{r}_{u,i} = \hat{p}_{0101}(x_{u,i}) - c(x_{u,i})\cdot(\hat{p}_{0100}(x_{u,i}) + \hat{p}_{0111}(x_{u,i}))$.

By making the arguments of uniform convergence and union bound, for any $\tilde{\epsilon} > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{R}(\pi^\dagger) - \mathbb{E}_T\left[\hat{R}(\pi^\dagger)\right]\right| \leq \tilde{\epsilon}\right) \geq 1-\eta,\\
\Leftarrow\ &\mathbb{P}\left(\max_{\pi\in\Pi}\left|\hat{R}(\pi) - \mathbb{E}_T\left[\hat{R}(\pi)\right]\right| \leq \tilde{\epsilon}\right) \geq 1-\eta,\\
\Leftrightarrow\ &\mathbb{P}\left(\bigcup_{\pi\in\Pi}\left|\hat{R}(\pi) - \mathbb{E}_T\left[\hat{R}(\pi)\right]\right| \geq \tilde{\epsilon}\right) < \eta,\\
\Leftarrow\ &\sum_{\pi\in\Pi}\mathbb{P}\left(\left|\hat{R}(\pi) - \mathbb{E}_T\left[\hat{R}(\pi)\right]\right| \geq \tilde{\epsilon}\right) < \eta,\\
\Leftarrow\ &\sum_{\pi\in\Pi}2\exp\left(\frac{-2\tilde{\epsilon}^2|\mathcal{D}|^2}{\sum_{(u,i)\in\mathcal{D}}\left\{\pi(x_{u,i})[1+c(x_{u,i})]\right\}^2}\right) < \eta,\\
\Leftarrow\ &2|\Pi|\exp\left(\frac{-2\tilde{\epsilon}^2|\mathcal{D}|^2}{\sum_{(u,i)\in\mathcal{D}}\left\{\pi^{\S}(x_{u,i})[1+c(x_{u,i})]\right\}^2}\right) < \eta,
\end{aligned}$$

where $\pi^{\S} = \arg\max_{\pi\in\Pi}\sum_{(u,i)\in\mathcal{D}}\left\{\pi(x_{u,i})[1+c(x_{u,i})]\right\}^2$. We solve the inequality in the last line for $\tilde{\epsilon}$ and obtain, with probability $1-\eta$, the following inequality

$$\hat{R}(\pi^\dagger) - \mathbb{E}_T[\hat{R}(\pi^\dagger)] \leq \sqrt{\frac{\log\left(\frac{2|\Pi|}{\eta}\right)}{2|\mathcal{D}|^2}\sum_{(u,i)\in\mathcal{D}}\left\{\pi^{\S}(x_{u,i})[1+c(x_{u,i})]\right\}^2}.$$

Then, when either imputed outcomes and or learned propensities $\hat{e}_{u,i} > 0$ are accurate for all user-item pairs, the unbiasedness of $\hat{R}(\pi^\dagger)$ directly follows from the unbiasedness of $p^{DR}_{0000}(x_{u,i})$, $p^{DR}_{0011}(x_{u,i})$, $p^{DR}_{0100}(x_{u,i})$, $p^{DR}_{0101}(x_{u,i})$, and $p^{DR}_{0111}(x_{u,i})$. Thus with probability $1-\eta$, the true reward of the learned optimal policy using estimated strata probabilities has the lower bound

$$R(\pi^\dagger) \geq \hat{R}(\pi^\dagger) - \sqrt{\frac{\log\left(\frac{2|\Pi|}{\eta}\right)}{2|\mathcal{D}|^2}\sum_{(u,i)\in\mathcal{D}}\left\{\pi^{\S}(x_{u,i})[1+c(x_{u,i})]\right\}^2}.$$

□