STABLEDR: STABILIZED DOUBLY ROBUST LEARNING FOR RECOMMENDATION ON DATA MISSING NOT AT RANDOM

Anonymous authors
Paper under double-blind review

ABSTRACT

In recommender systems, users always choose the favorite items to rate, which leads to data missing not at random and poses a great challenge for unbiased evaluation and learning of prediction models. Currently, the doubly robust (DR) methods have been widely studied and demonstrate superior performance. However, in this paper, we show that DR methods are unstable and have unbounded bias, variance, and generalization bounds to extremely small propensities. Moreover, the fact that DR relies more on extrapolation will lead to suboptimal performance. To address the above limitations while retaining double robustness, we propose a stabilized doubly robust (StableDR) learning approach with a weaker reliance on extrapolation. Theoretical analysis shows that StableDR has bounded bias, variance, and generalization error bound simultaneously under inaccurate imputed errors and arbitrarily small propensities. In addition, we propose a novel learning approach for StableDR that updates the imputation, propensity, and prediction models cyclically, achieving more stable and accurate predictions. Extensive experiments show that our approaches significantly outperform the existing methods.

1 INTRODUCTION

Modern recommender systems (RS) are rapidly evolving with the adoption of sophisticated deep learning models (Zhang et al., 2019). However, it is well documented that directly using advanced deep models usually achieves sub-optimal performance due to the existence of various biases in RS (Chen et al., 2020; Wu et al., 2022), and the biases would be amplified over time (Wen et al., 2022; Mansoury et al., 2020). A large number of debiasing methods have emerged and gradually become a trend. For many practical tasks in RS, such as rating prediction (Schnabel et al., 2016; Wang et al., 2020a; 2019), post-view click-through rate prediction (Guo et al., 2021), post-click conversion rate prediction (Zhang et al., 2020; Dai et al., 2022), and uplift modeling (Saito et al., 2019; Sato et al., 2019; 2020), a critical challenge is to combat the selection bias and confounding bias that leading to significantly difference between the trained sample and the targeted population (Hernán & Robins, 2020). Various methods were designed to address this problem and among them, doubly robust (DR) methods (Wang et al., 2019; Zhang et al., 2020; Chen et al., 2021; Dai et al., 2022; Ding et al., 2022) play the dominant role due to their better performance and theoretical properties.

The success of DR is attributed to its double robustness and joint-learning technique. However, the DR methods still have many limitations. Theoretical analysis shows that inverse probability scoring (IPS) and DR methods may have infinite bias, variance, and generalization error bounds, in the presence of extremely small propensity scores (Schnabel et al., 2016; Wang et al., 2019; Guo et al., 2021). In addition, due to the fact that users are more inclined to evaluate the preferred items, the problem of data missing not at random (MNAR) often occurs in RS. This would cause selection bias and results in inaccuracy for methods that more rely on extrapolation, such as error imputation based (EIB) (Marlin et al., 2007; Steck, 2013) and DR methods.

To overcome the above limitations while maintaining double robustness, we propose a stabilized doubly robust (SDR) estimator with a weaker reliance on extrapolation, which reduces the negative impact of extrapolation and MNAR effect on the imputation model. Through theoretical analysis,

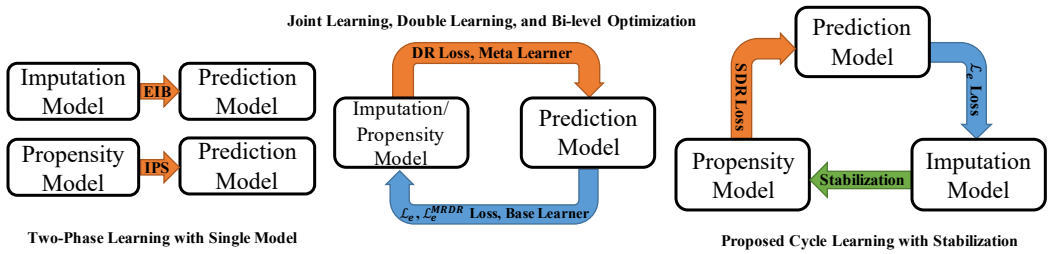

Figure 1: During the training of updating a prediction model, two-phase learning (Marlin et al., 2007; Steck, 2013; Schnabel et al., 2016) uses a fixed imputation/propensity model (Left), whereas DR-JL (Wang et al., 2019), MRDR-DL (Guo et al., 2021), and AutoDebias (Chen et al., 2021) uses alternative learning between the imputation/propensity and the prediction model (Middle). The proposed learning approach updates the three models cyclically with stabilization (Right).

we demonstrate that the SDR has bounded bias and generalization error bound for arbitrarily small propensities, which further indicates that the SDR can achieve more stable predictions.

Furthermore, we propose a novel cycle learning approach for SDR. Figure 1 shows the differences between the proposed cycle learning of SDR and the existing unbiased learning approaches. Two-phase learning (Marlin et al., 2007; Steck, 2013; Schnabel et al., 2016) first obtains an imputation/propensity model to estimate the ideal loss and then updates the prediction model by minimizing the estimated loss. DR-JL (Wang et al., 2019), MRDR-DL (Guo et al., 2021), and AutoDebias (Chen et al., 2021) alternatively update the model used to estimate the ideal loss and the prediction model. The proposed learning method cyclically uses different losses to update the three models with the aim of achieving more stable and accurate prediction results. We have conducted extensive experiments on two real-world datasets, and the results show that the proposed approach significantly improves debiasing and convergence performance compared to existing methods.

2 Preliminaries

2.1 Problem Setting

In RS, due to the fact that users are more inclined to evaluate the preferred items, the collected ratings are always missing not at random (MNAR). We formulate the data MNAR problem using the widely adopted potential outcome framework (Neyman, 1990; Imbens & Rubin, 2015). Let $\mathcal{U} = \{1, 2, ..., U\}$, $\mathcal{I} = \{1, 2, ..., I\}$ and $\mathcal{D} = \mathcal{U} \times \mathcal{I}$ be the index sets of users, items, all user-item pairs. For each $(u, i) \in \mathcal{D}$, we have a treatment $o_{u,i} \in \{0, 1\}$, a feature vector $x_{u,i}$, and an observed rating $r_{u,i}$, where $o_{u,i} = 1$ if user u rated the item i in the logging data, $o_{u,i} = 0$ if the rating is missing. Let $r_{u,i}(1)$ is defined as the be the rating that would be observed if item i had been rated by user u, which is observable only for $\mathcal{O} = \{(u, i) \mid (u, i) \in \mathcal{D}, o_{u,i} = 1\}$. Many tasks in RS can be formulated by predicting the potential outcome $r_{u,i}(1)$ using feature $x_{u,i}$ for each (u, i).

Let $\hat{r}_{u,i}(1) = f(x_{u,i}; \phi)$ be a prediction model with parameters ϕ. If all the potential outcomes $\{r_{u,i}(1) : (u, i) \in \mathcal{D}\}$ were observed, the ideal loss function for solving parameters ϕ is given as

$$\mathcal{L}_{ideal}(\phi) = |\mathcal{D}|^{-1} \sum_{(u,i)\in\mathcal{D}} e_{u,i}. \tag{1}$$

where $e_{u,i}$ is the prediction error, such as the squared loss $e_{u,i} = (\hat{r}_{u,i}(1) - r_{u,i}(1))^2$. $\mathcal{L}_{ideal}(\phi)$ can be regarded as a benchmark of unbiased loss function, even though it is infeasible due to the missingness of $\{r_{u,i}(1) : o_{u,i} = 0\}$. As such, a variety of methods are developed through approximating $\mathcal{L}_{ideal}(\phi)$ to address the selection bias, in which the propensity-based estimators show the relatively superior performance (Schnabel et al., 2016; Wang et al., 2019; Guo et al., 2021; Chen et al., 2021; Dai et al., 2022; Ding et al., 2022), and the IPS and DR estimator is given as

$$\mathcal{E}_{IPS} = |\mathcal{D}|^{-1} \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i} e_{u,i}}{\hat{p}_{u,i}} \quad \text{and} \quad \mathcal{E}_{DR} = |\mathcal{D}|^{-1} \sum_{(u,i)\in\mathcal{D}} \left[\hat{e}_{u,i} + \frac{o_{u,i}(e_{u,i} - \hat{e}_{u,i})}{\hat{p}_{u,i}}\right], \tag{2}$$

where $\hat{p}_{u,i}$ is an estimate of propensity score $p_{u,i} := P(o_{u,i} = 1|x_{u,i})$, $\hat{e}_{u,i}$ is an estimate of $e_{u,i}$.

2.2 RELATED WORK

Debiased learning in recommendation. The data collected in RS suffers from various biases (Chen et al., 2020; Wu et al., 2022), which are entangled with the true preferences of users and pose a great challenge to unbiased learning. There is increasing interest in coping with different biases in recent years (Zhang et al., 2021; Ai et al., 2018; Liu et al., 2016; Liu et al., 2021). Schnabel et al. (2016) proposed using inverse propensity score (IPS) and self-normalized IPS (SNIPS) methods to address the selection bias on data missing not at random, and Saito et al. (2020) and Saito (2019) extended them to implicit feedback data. Marlin et al. (2007) and Steck (2013) derived an error imputation-based (EIB) unbiased learning method. These three approaches adopt two-phase learning (Wang et al., 2021), which first learns a propensity/imputation model and then applies it to construct an unbiased estimator of the ideal loss to train the recommendation model. A doubly robust joint learning (DR-JL) method (Wang et al., 2019) was proposed by combining the IPS and EIB approaches. Subsequently, strands of enhanced joint learning methods were developed, including MRDR (Guo et al., 2021), Multi-task DR (Zhang et al., 2020), DR-MSE (Dai et al., 2022), BRD-DR (Ding et al., 2022), and uniform data-aware methods (Bonner & Vasile, 2018; Liu et al., 2020; Chen et al., 2021; Wang et al., 2021) that aimed to seek better recommendation strategies by leveraging a small uniform dataset. Chen et al. (2020) reviewed various biases in RS and discussed the recent progress on debiasing tasks. Wu et al. (2022) established the connections between the biases in causal inference and the biases in RS, thereby presenting the formal causal definitions of the biases in RS.

Stabilized causal effect estimation. The proposed method builds on the stabilized average caual effect estimation approaches in causal inference. Molenberghs et al. (2015) summarized the limitations of doubly robust methods, including unstable to small propensities, unboundedness (van der Laan & Rose, 2011), and large variance (Tan, 2007). These issues inspired a series of stabilized causal effect estimation methods in statistics (Kang & Schafer, 2007; Bang & Robins, 2005; van der Laan & Rose, 2011; Molenberghs et al., 2015). Unlike previous works that focused only on achieving learning with unbiasedness in RS, this paper provides a new perspective to develop doubly robust estimators with much more stable statistical properties.

3 STABILIZED DOUBLY ROBUST ESTIMATOR

In this section, we elaborate the limitations of DR methods and propose a stabilized DR (SDR) estimator with a weaker reliance on extrapolation. Theoretical analysis shows that SDR has bounded bias and generalization error bound for arbitrarily small propensities, while IPS and DR doesn't.

3.1 MOTIVATION

Even though DR estimator has double robustness property, its performance could be significantly improved if the following three aspects can be enhanced.

More stable to small propensities. As shown in Schnabel et al. (2016), Wang et al. (2019) and Guo et al. (2021), if there exist some extremely small estimated propensity scores, the IPS/DR estimator and its tail bound are unbounded, deteriorating the prediction accuracy. However, the problem is difficult to be addressed, given the fact that there are many long-tailed users and items in RS, resulting in the presence of extreme propensities.

More stable through weakening extrapolation. DR relies more on extrapolation because the imputation model in DR is learned from the exposed events $\mathcal{O}$ and extrapolated to the unexposed events. If the distributional disparity of $e_{u,i}$ on $o_{u,i} = 0$ and $o_{u,i} = 1$ is large, the imputed errors are likely to be inaccurate on the unexposed events and incur bias of DR. Therefore, it is beneficial to reduce bias if we can develop an enhanced DR method with weaker reliance on extrapolation.

More stable training process of updating a prediction model. In general, alternating training between models results in better performance. From Figure 1, Wang et al. (2019) proposes joint learning for DR, alternatively updating the imputation and prediction models with given estimated propensities. Double learning (Guo et al., 2021) further incorporates parameter sharing between the imputation and prediction models. Bi-level optimization (Wang et al., 2021; Chen et al., 2021) can be viewed as alternately updating the prediction model and the other parameters used to estimate the loss. To the best of our knowledge, this is the first paper that proposes a algorithm to update the

three models separately using different optimizers, which may resulting in more stable and accurate rating predictions.

3.2 STABILIZED DOUBLY ROBUST ESTIMATOR

We propose a stabilized doubly robust (SDR) estimator that has a weaker dependence on extrapolation and is robust to small propensities. The SDR estimator consists of the following three steps.

Step 1 (Initialize imputed errors). Pre-train an imputation model $\hat{e}_{u,i}$ and $\hat{\mathcal{E}} \triangleq |\mathcal{D}|^{-1} \sum_{(u,i)\in\mathcal{D}} \hat{e}_{u,i}$.

Step 2 (Learn constrained propensities). Learn a propensity model $\hat{p}_{u,i}$ satisfying

$$\frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}}{\hat{p}_{u,i}} \Big(\hat{e}_{u,i} - \hat{\mathcal{E}} \Big) = 0. \tag{3}$$

Step 3 (SDR estimator). The SDR estimator is given as

$$\mathcal{E}_{SDR} = \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i} e_{u,i}}{\hat{p}_{u,i}} \Big/ \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}}{\hat{p}_{u,i}} \triangleq \sum_{(u,i)\in\mathcal{D}} w_{u,i} e_{u,i},$$

where $w_{u,i} = \frac{o_{u,i}}{\hat{p}_{u,i}} \Big/ \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}}{\hat{p}_{u,i}}$. It can be seen that SDR estimator has the same form as SNIPS estimator, but the propensities are learned differently. In SDR, the estimation of propensity model relies on the imputed errors, whereas not in SNIPS.

Each step in the construction of SDR estimator plays a different role. Specifically, the Step 2 is designed to enable double robustness property as shown in Theorem 1 (see Appendix A.1 for proofs).

Theorem 1 (Double Robustness)**.** *$\mathcal{E}_{SDR}$ is an asymptotically unbiased[1] estimator of $\mathcal{L}_{ideal}$, when either the learned propensities $\hat{p}_{u,i}$ or the imputed errors $\hat{e}_{u,i}$ are accurate for all user-item pairs.*

We provide an intuitive way to illustrate the rationale of SDR. **On the one hand**, if the propensities can be accurately estimated (i.e., $\hat{p}_{u,i} = p_{u,i}$) by using a common model (e.g., logistic regression) without imposing constraint (3). Then the expectation of the left side of equation (3) becomes

$$\mathbb{E}_{\mathcal{O}} \Big[\frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}}{\hat{p}_{u,i}} \Big(\hat{e}_{u,i} - \hat{\mathcal{E}} \Big) \Big] = \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \Big(\hat{e}_{u,i} - \hat{\mathcal{E}} \Big) \equiv 0,$$

which indicates the equation (3) always holds as the sample size goes to infinity by the strong law of large numbers[2], irrespective of the accuracy of the imputed errors $\hat{e}_{u,i}$. This implies that the constraint (3) *imposes almost no restriction* on the estimation of propensities. In this case, the SDR estimator is almost equivalent to the original SNIPS estimator. **On the other hand**, if the propensities *cannot* be accurately estimated by using a common model, but the imputed errors are accurate (i.e., $\hat{e}_{u,i} = e_{u,i}$). In this case, $\hat{\mathcal{E}}$ is an unbiased estimator of $\mathcal{L}_{ideal}$. Specifically, $\mathcal{E}_{SDR}$ satisfies

$$\frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}(e_{u,i} - \mathcal{E}_{SDR})}{\hat{p}_{u,i}} = 0. \tag{4}$$

Combining the constraint (3) and equation (4) gives

$$\frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \left[\frac{o_{u,i}(e_{u,i} - \hat{e}_{u,i})}{\hat{p}_{u,i}} + \frac{o_{u,i}(\hat{\mathcal{E}} - \mathcal{E}_{SDR})}{\hat{p}_{u,i}} \right] = 0, \tag{5}$$

where the first term equals to 0 if $\hat{e}_{u,i} = e_{u,i}$, it implies that $\mathcal{E}_{SDR} = \hat{\mathcal{E}}$, then the unbiasedness of $\mathcal{E}_{SDR}$ follows immediately from the unbiasedness of $\hat{\mathcal{E}}$.

In addition, Step 3 is designed for two main reasons to achieve stability. First, $\mathcal{E}_{SDR}$ is more robust to extrapolation compared with DR. This is because the propensities are learned from the entire data

[1] Asymptotically unbiased means unbiasedness as the sample size goes to infinity.

[2] This is the reason why we adopt the notation of "asymptotically unbiased".

and thus have less requirement on extrapolation. Second, $\mathcal{E}_{SDR}$ is more stable to small propensities, since the self-normalization imposes the weight $w_{u,i}$ to fall on the interval [0,1].

In summary, forcing the propensities to satisfy the equation (3) makes the SDR estimator not only doubly robust, but also captures the advantages of both SNIPS and DR estimators. The design of SDR enables the constrained propensities to adaptively find the direction of debiasing if either the learned propensities without imposing constraint (3) or the imputed errors are accurate.

3.3 Theoretical Analysis of Stableness

Through theoretical analysis, we note that previous debiasing estimators such as IPS (Schnabel et al., 2016) and DR-based methods (Wang et al., 2019; Guo et al., 2021) tend to have infinite biases, variances, tail bound, and corresponding generalization error bounds, in the presence of extremely small estimated propensities. Remarkably, the proposed SDR estimator doesn't suffer from such problems and is stable to arbitrarily small propensities, as shown in the following Theorems (see Appendixes A.2, A.3 and A.4 for proofs).

Theorem 2 (Bias of SDR)**.** *Given imputed errors $\hat{e}_{u,i}$ and learned propensities $\hat{p}_{u,i}$ satisfying the stabilization constraint (3), with $\hat{p}_{u,i} > 0$ for all user-item pairs, the bias of $\mathcal{E}_{SDR}$ is*

$$\text{Bias}(\mathcal{E}_{SDR}) = \left| \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \left(\delta_{u,i} - \frac{\sum_{(u,i)\in\mathcal{D}} \delta_{u,i} p_{u,i}/\hat{p}_{u,i}}{\sum_{(u,i)\in\mathcal{D}} p_{u,i}/\hat{p}_{u,i}} \right) \right| + O(|\mathcal{D}|^{-1}),$$

where $\delta_{u,i} = e_{u,i} - \hat{e}_{u,i}$ is the error deviation.

Theorem 2 shows the bias of the SDR estimator consisting of a dominant term given by the difference between $\delta_{u,i}$ and its weighted average, and a negligible term of order $O(|\mathcal{D}|^{-1})$. The fact that the $\delta_{u,i}$ and its convex combinations are bounded, shows that the bias is bounded for arbitrarily small $\hat{p}_{u,i}$. Compared to the $\text{Bias}(\mathcal{E}_{\text{IPS}}) = |\mathcal{D}|^{-1} |\sum_{u,i\in\mathcal{D}} (\hat{p}_{u,i} - p_{u,i}) e_{u,i}/\hat{p}_{u,i}|$ and $\text{Bias}(\mathcal{E}_{\text{DR}}) = |\mathcal{D}|^{-1} |\sum_{u,i\in\mathcal{D}} (\hat{p}_{u,i} - p_{u,i}) \delta_{u,i}/\hat{p}_{u,i}|$, it indicates that IPS and DR will have extremely large bias when there exist an extremely small $\hat{p}_{u,i}$.

Theorem 3 (Variance of SDR)**.** *Under the conditions of Theorem 2, the variance of $\mathcal{E}_{SDR}$ is*

$$\text{Var}(\mathcal{E}_{SDR}) = \frac{\sum_{(u,i)\in\mathcal{D}} p_{u,i}(1-p_{u,i}) h_{u,i}^2/\hat{p}_{u,i}^2}{\left(\sum_{(u,i)\in\mathcal{D}} p_{u,i}/\hat{p}_{u,i}\right)^2} + O(|\mathcal{D}|^{-2}),$$

where $h_{u,i} = (e_{u,i} - \hat{e}_{u,i}) - \sum_{(u,i)\in\mathcal{D}} \{p_{u,i}(e_{u,i} - \hat{e}_{u,i})/\hat{p}_{u,i}\} / \sum_{(u,i)\in\mathcal{D}} \{p_{u,i}/\hat{p}_{u,i}\}$ is a bounded difference between $e_{u,i} - \hat{e}_{u,i}$ and its weighted average.

Theorem 3 shows the variance of the SDR estimator consisting of a dominant term and a negligible term of order $O(|\mathcal{D}|^{-2})$. The boundedness of the variance for arbitrarily small $\hat{p}_{u,i}$ is given directly from the fact that SDR has a bounded range given by the self-normalized form. Compared to the $\text{Var}(\mathcal{E}_{\text{IPS}}) = |\mathcal{D}|^{-2} \sum_{u,i\in\mathcal{D}} p_{u,i}(1-p_{u,i}) e_{u,i}^2/\hat{p}_{u,i}^2$ and $\text{Var}(\mathcal{E}_{\text{DR}}) = |\mathcal{D}|^{-2} \sum_{u,i\in\mathcal{D}} p_{u,i}(1-p_{u,i})(e_{u,i} - \hat{e}_{u,i})^2/\hat{p}_{u,i}^2$, it indicates that IPS and DR will have extremely large variance (tend to infinity) when there exist an extremely small $\hat{p}_{u,i}$ (tends to 0).

Theorem 4 (Tail Bound of SDR)**.** *Under the conditions of Theorem 2, for any prediction model, with probability $1-\eta$, the deviation of the SDR estimator from its expectation has the following tail bound*

$$|\mathcal{E}_{SDR} - \mathbb{E}_{\mathcal{O}}(\mathcal{E}_{SDR})| \leq \sqrt{\frac{1}{2}\log\left(\frac{4}{\eta}\right) \sum_{(u,i)\in\mathcal{D}} \frac{(\delta_{\max} - \delta_{u,i})^2 + (\delta_{u,i} - \delta_{\min})^2}{\{1 + \hat{p}_{u,i}(\sum_{\mathcal{D}\setminus(u,i)} p_{u,i}/\hat{p}_{u,i} - \epsilon')\}^2}}$$

where $\delta_{\min} = \min_{(u,i)\in\mathcal{D}} \delta_{u,i}$, $\delta_{\max} = \max_{(u,i)\in\mathcal{D}} \delta_{u,i}$, $\epsilon' = \sqrt{\log(4/\eta)/2 \cdot \sum_{\mathcal{D}\setminus(u,i)} 1/\hat{p}_{u,i}^2}$, and $\mathcal{D} \setminus (u,i)$ is the set of $\mathcal{D}$ excluding the element (u,i).

Note that $\sum_{\mathcal{D}\setminus(u,i)} p_{u,i}/\hat{p}_{u,i} = O(|\mathcal{D}|)$ and $\epsilon' = O(|\mathcal{D}|^{1/2})$ in Theorem 4, it follows that the tail bound of the SDR estimator converges to 0 for large samples. In addition, the tail bound is bounded

for arbitrarily small $\hat{p}_{u,i}$. Compared to the tail bound of IPS and DR, with probability $1-\eta$, we have

$$|\mathcal{E}_{\text{IPS}} - \mathbb{E}_{\mathcal{O}}\left[\mathcal{E}_{\text{IPS}}\right]| \leq \sqrt{\frac{\log\left(\frac{2}{\eta}\right)}{2|\mathcal{D}|^2} \sum_{(u,i)\in\mathcal{D}} \left(\frac{e_{u,i}}{\hat{p}_{u,i}}\right)^2}, |\mathcal{E}_{\text{DR}} - \mathbb{E}_{\mathcal{O}}\left[\mathcal{E}_{\text{DR}}\right]| \leq \sqrt{\frac{\log\left(\frac{2}{\eta}\right)}{2|\mathcal{D}|^2} \sum_{(u,i)\in\mathcal{D}} \left(\frac{\delta_{u,i}}{\hat{p}_{u,i}}\right)^2}$$

which are both unbounded when $\hat{p}_{u,i} \to 0$. For SDR in the prediction model training phase, the boundedness of the generalization error bound (see Theorem 5 in Appendix A.5) follows immediately from the boundedness of the bias and tail bound. The above analysis demonstrates that SDR can comprehensively mitigate the negative effects caused by extreme propensities and results in more stable predictions. Theorems 2-5 are stated under the constraint (3). If we estimate the propensities with constraint (3), but finally constraint (3) somehow doesn't hold exactly, the associated bias, variance, and generalization error bound of SDR are presented in Appendix B.

4 Cycle Learning with Stabilization

In this section, we propose a novel SDR-based cycle learning approach, that not only exploits the stable statistical properties of the SDR estimator itself, but also carefully designs the updating process among various models to achieve higher stability. In general, inspired by the idea of value iteration in reinforcement learning (Sutton & Barto, 2018), alternatively updating the model tends to achieve better predictive performance, as existing debiasing training approaches suggested (Wang et al., 2019; Guo et al., 2021; Chen et al., 2021). As shown in Figure 1, the proposed approach dynamically interacts with three models, utilizing the propensity model and imputation model simultaneously in a differentiated way, which can be regarded as an extension of these methods. In cycle learning, given pre-trained propensities, the inverse propensity weighted imputation error loss is used to first obtain an imputation model, and then take the equation (3) as the regularization term to train a stabilized propensity model and ensure the double robustness of SDR. Finally, the prediction model is updated by minimizing the SDR loss and used to readjust the imputed errors. By repeating the above update processes cyclically, the cycle learning approach can fully utilize and combine the advantages of the three models to achieve more accurate rating predictions.

Specifically, the data MNAR leads to the presence of missing $r_{u,i}(1)$, so that all $e_{u,i}$ cannot be used directly. Therefore, we obtain imputed errors by learning a pseudo-labeling model $\tilde{r}_{u,i}(1)$ parameterized by β, and the imputed errors $\hat{e}_{u,i} = CE(\tilde{r}_{u,i}(1), \hat{r}_{u,i}(1))$ are updated by minimizing

$$\mathcal{L}_e\left(\phi,\alpha,\beta\right) = |\mathcal{D}|^{-1} \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}(\hat{e}_{u,i} - e_{u,i})^2}{\pi(x_{u,i};\alpha)} + \lambda_e \|\beta\|_F^2,$$

where $e_{u,i} = CE(r_{u,i}(1), \hat{r}_{u,i}(1))$, $\lambda_e \geq 0$, $\hat{p}_{u,i} = \pi(x_{u,i};\alpha)$ is the propensity model, $\|\cdot\|_F^2$ is the Frobenius norm. For each observed ratings, the inverse of the estimated propensities are used for weighting to account for MNAR effects. Next, we consider two methods for estimating propensity scores, which are Naive Bayes with Laplace smoothing and logistic regression. The former provides a wide range of opportunities for achieving stability constraints (3) through the selection of smoothing coefficients. The latter requires user and item embeddings, which are obtained by employing MF before performing cycle learning. The learned propensities need to both satisfy the accuracy, which is evaluated with cross entropy, and meet the constraint (3) for stabilization and double robustness. The propensity model $\pi(x_{u,i};\alpha)$ is updated by using the loss $\mathcal{L}_{ce}\left(\phi,\alpha,\beta\right) + \eta \cdot \mathcal{L}_{stable}\left(\phi,\alpha,\beta\right)$, where $\mathcal{L}_{ce}\left(\phi,\alpha,\beta\right)$ is cross entropy loss of propensity model and

$$\mathcal{L}_{stable}\left(\phi,\alpha,\beta\right) = |\mathcal{D}|^{-1}\Big\{\sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}}{\pi(x_{u,i};\alpha)}\Big(\hat{e}_{u,i} - \hat{\mathcal{E}}\Big)\Big\}^2 + \lambda_{stable}\|\alpha\|_F^2,$$

where $\lambda_{stable} \geq 0$, and η is a hyper-parameter for trade-off. Finally, the prediction model $f(x_{u,i};\phi)$ is updated by minimizing the SDR loss

$$\mathcal{L}_{sdr}\left(\phi,\alpha,\beta\right) = \Big[\sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}e_{u,i}}{\pi(x_{u,i};\alpha)}\Big] \Big/ \Big[\sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}}{\pi(x_{u,i};\alpha)}\Big] + \lambda_{sdr}\|\phi\|_F^2,$$

where the first term is equivalent to (5), and $\lambda_{sdr} \geq 0$. In cycle learning, the updated prediction model will be used for re-update the imputation model using the next sample batch. We summarized the cycle learning approach in Alg. 1.

Algorithm 1: The Proposed Stable DR (MRDR) Cycle Learning, Stable-DR (MRDR)

Input: observed ratings $\mathbf{Y}^o$, and $\eta, \lambda_e, \lambda_{stable}, \lambda_{sdr} \geq 0$
while *stopping criteria is not satisfied* **do**
  for *number of steps for training the imputation model* **do**
    Sample a batch of user-item pairs $\{(u_j, i_j)\}_{j=1}^{J}$ from $\mathcal{O}$;
    Update β by descending along the gradient $\nabla_\beta \mathcal{L}_e(\phi, \alpha, \beta)$;
  end
  for *number of steps for training the propensity model* **do**
    Sample a batch of user-item pairs $\{(u_k, i_k)\}_{k=1}^{K}$ from $\mathcal{D}$;
    Calculate the gradient of propensity cross entropy error $\nabla_\alpha \mathcal{L}_{ce}(\phi, \alpha, \beta)$;
    Calculate the gradient of propensity stable constraint (3) $\nabla_\alpha \mathcal{L}_{stable}(\phi, \alpha, \beta)$;
    Update α by descending along the gradient $\nabla_\alpha \mathcal{L}_{ce}(\phi, \alpha, \beta) + \eta \cdot \nabla_\alpha \mathcal{L}_{stable}(\phi, \alpha, \beta)$
  end
  for *number of steps for training the prediction model* **do**
    Sample a batch of user-item pairs $\{(u_l, i_l)\}_{l=1}^{L}$ from $\mathcal{O}$;
    Update ϕ by descending along the gradient $\nabla_\phi \mathcal{L}_{sdr}(\phi, \alpha, \beta)$;
  end
end

5 Real-world Experiments

In this section, several experiments are conducted to evaluate the proposed methods on two real-world benchmark datasets. We conduct experiments to answer the following questions:

RQ1. Do the proposed Stable-DR and Stable-MRDR approaches improve in debiasing performance compared to the existing studies?

RQ2. Do the proposed methods increase the runtime empirically?

RQ3. Do our methods stably perform well under the various propensity models? How does the performance of our method change under different strengths of the stabilization constraint?

5.1 Experimental Setup

Dataset and preprocessing. To answer the above RQs, we need to use the datasets that contain both MNAR ratings and missing-at-random (MAR) ratings. Following the previous studies (Schnabel et al., 2016; Wang et al., 2019; Guo et al., 2021; Chen et al., 2021), we conduct our experiments on the two commonly used datasets: **Coat Shopping**[3] contains ratings from 290 users to 300 items. Each user evaluates 24 items, containing 6,960 MNAR ratings in total. Meanwhile, each user evaluates 16 items randomly, which generates 4,640 MAR ratings. **Yahoo! R3**[4] contains totally 54,000 MAR and 311,704 MNAR ratings from 15,400 users to 1,000 items.

Baselines. In our experiments, we take **Matrix Factorization (MF)** (Koren et al., 2009), **Sparse LInear Method (SLIM) (Ning & Karypis, 2011)** and **Neural Collaborative Filtering (NCF)** (He et al., 2017) as the base model respectively, and compare against the proposed methods with the following baselines: **Base Model**, **IPS** (Saito et al., 2020; Schnabel et al., 2016), **SNIPS** (Swaminathan & Joachims, 2015), **IPS-AT** (Saito, 2020), **CVIB** (Wang et al., 2020b), **DR** (Saito, 2020), **DR-JL** (Wang et al., 2019), and **MRDR-JL** (Guo et al., 2021). In addition, **Naive Bayes with Laplace smoothing** and **logistic regression** are used to establish the propensity model respectively.

Experimental protocols and details. The following four metrics are used simultaneously in the evaluation of debiasing performance: *MSE, AUC, NDCG@5,* and *NDCG@10*. All the experiments are implemented on PyTorch with Adam as the optimizer[5]. We tune the learning rate in $\{0.005, 0.01, 0.05, 0.1\}$, weight decay in $\{1e-6, 5e-6, \ldots, 5e-3, 1e-2\}$, constrain parameter

[3] https://www.cs.cornell.edu/~schnabts/mnar/

[4] http://webscope.sandbox.yahoo.com/

[5] For all experiments, we use NVIDIA GeForce RTX 2060 as the computing resource.

Table 1: Evaluation metrics of MSE, AUC, NDCG@5, and NDCG@10 on the MAR test set of Coat and Yahoo!R3 using Naive Bayes with Laplace smoothing as the propensity model and taking MF (Top) and SLIM (Bottom) as the base model respectively. The best two results are bolded.

Datasets	Coat				Yahoo!R3			
Methods	MSE	AUC	N@5	N@10	MSE	AUC	N@5	N@10
MF	0.2428	0.7063	0.6025	0.6774	0.2500	0.6722	0.6374	0.7634
+ IPS	0.2316	0.7166	0.6184	0.6897	0.2194	0.6742	0.6304	0.7556
+ SNIPS	0.2333	0.7070	0.6222	0.6851	**0.1931**	0.6831	0.6348	0.7608
+ AS-IPS	**0.2121**	0.7180	0.6160	0.6824	0.2391	0.6770	0.6364	0.7601
+ CVIB	0.2195	0.7239	0.6285	0.6947	0.2625	0.6853	0.6513	0.7729
+ DR	0.2298	0.7132	0.6243	0.6918	0.2093	0.6873	0.6574	0.7741
+ DR-JL	0.2254	0.7209	0.6252	0.6961	0.2194	0.6863	0.6525	0.7701
+ **Stable-DR**	0.2159	**0.7508**	**0.6511**	**0.7073**	0.2090	**0.6946**	**0.6620**	**0.7786**
+ MRDR-JL	0.2252	0.7318	0.6375	0.6989	0.2173	0.6830	0.6437	0.7652
+ **Stable-MRDR**	**0.2076**	**0.7548**	**0.6532**	**0.7105**	**0.2081**	**0.6915**	**0.6585**	**0.7757**
SLIM	0.2419	0.7074	0.7064	0.7650	0.2126	0.6636	0.7190	0.8134
+ IPS	0.2411	0.7058	0.7235	0.7644	0.2046	0.6583	0.7285	0.8244
+ SNIPS	0.2420	0.7071	0.7369	0.7672	0.2155	0.6720	0.7303	0.8227
+ AS-IPS	**0.2133**	0.7105	0.6238	0.6975	**0.1946**	0.6769	0.6508	0.7702
+ CVIB	0.2413	0.7108	0.7214	0.7638	**0.2024**	0.6790	0.7335	0.8221
+ DR	**0.2334**	0.7064	0.7267	0.7649	0.2054	0.6771	0.7344	0.8248
+ DR-JL	0.2407	0.7090	0.7279	0.7655	0.2044	0.6792	0.7360	0.8260
+ **Stable-DR**	0.2356	**0.7201**	**0.7389**	**0.7724**	0.2080	**0.6874**	**0.7473**	**0.8349**
+ MRDR-JL	0.2409	0.7074	0.7329	0.7679	0.2016	0.6791	0.7338	0.8239
+ **Stable-MRDR**	0.2369	**0.7148**	**0.7378**	**0.7711**	0.2086	**0.6842**	**0.7435**	**0.8308**

eta in $\{50, 100, 150, 200\}$ for **Coat** and $\{500, 1000, 1500, 2000\}$ for **Yahoo! R3**, and batch size in $\{128, 256, 512, 1024, 2048\}$ for **Coat** and $\{1024, 2048, 4096, 8192, 16384\}$ for **Yahoo! R3**. In addition, for the Laplacian smooth parameter in Naive Bayes model, the initial value is set to 0 and learning rate is tuned in $\{5, 10, 15, 20\}$ for **Coat** and in $\{50, 100, 150, 200\}$ for **Yahoo! R3**.

5.2 Performance Comparison (RQ1, RQ2)

Table 1 summarizes the performance of the proposed Stable-DR and Stable-MRDR methods compared with previous methods. First, the causally-inspired methods perform better than the base model, verifying the necessity of handling the selection bias in rating prediction. For previous methods, SNIPS, CVIB and DR demonstrate competitive performance. Second, the proposed Stable-DR and Stable-MRDR have the best performance in all four metrics. On the one hand, our methods outperform SNIPS, attributed to the inclusion of the propensity model in the training process, as well as the boundedness and double robustness of SDR. On the other hand, our methods outperform DR-JL and MRDR-JL, attributed to the stabilization constraint introduced in the training of the propensity model. This further demonstrates the benefit of cycle learning, in which the propensity model is acted as the mediation between the imputation and prediction model during the training process, rather than updating the prediction model from the imputation model directly. We also adopt NCF as the base model to verify the validity of the proposed Stable-DR and Stable-MRDR, and similar conclusions are found. The detailed experimental results are presented in Appendix C.1.

In addition, we conduct experiments to compare the empirical running times of different debiased prediction methods on the Coat and Yahoo!R3 datasets. As shown in Table 2, although the proposed Stable-DR and Stable-MRDR alternately updates the imputation, propensity, and prediction models, there is no significant increase in their empirical running time compared to the existing methods.

Table 2: Empirical runtime comparison on Coat and Yahoo!R3 datasets.

Datasets	MF	IPS	AS-IPS	SNIPS	DR	DR-JL	**Stable-DR**	MRDR-JL	**Stable-MRDR**
Coat	$33.17_{\pm 1.85}$	$35.04_{\pm 2.99}$	$138.74_{\pm 3.38}$	$30.52_{\pm 7.47}$	$58.35_{\pm 4.40}$	$136.56_{\pm 8.91}$	$131.22_{\pm 11.77}$	$126.28_{\pm 11.12}$	$154.71_{\pm 16.18}$
Yahoo!R3	$58.13_{\pm 3.58}$	$68.20_{\pm 2.39}$	$295.17_{\pm 14.56}$	$41.97_{\pm 7.60}$	$183.86_{\pm 13.11}$	$611.45_{\pm 36.33}$	$701.20_{\pm 45.10}$	$631.90_{\pm 30.03}$	$679.41_{\pm 60.55}$

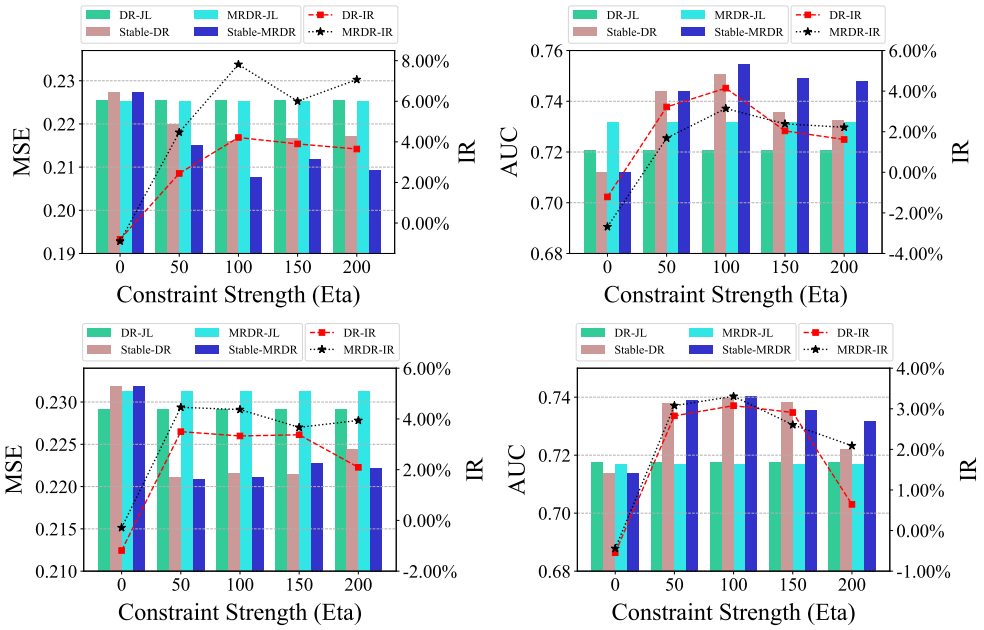

Figure 2: MSE, AUC and Increasing Ratio (IR) of Stable-DR and Stable-MRDR comparing with two baseline algorithms DR-JL and MRDR-JL in two different propensity model setting: Naive Bayes with Laplace smoothing (Top) and Logistic Regression (Bottom) respectively.

5.3 Ablation and Parameter Sensitivity Study (RQ3)

The debiasing performance under different stabilization constraint strength and propensity models is shown in Figure 2. First, the proposed Stable-DR and Stable-MRDR outperform DR-JL and MRDR-JL, when both Naive Bayes with Laplace smoothing and logistic regression are used as propensity models. It indicates that our methods have better debiasing ability in the feature containing and collaborative filtering scenarios simultaneously. Second, when the strength of the stabilization constraint is zero, our method performs similarly to SNIPS and slightly worse than the DR-JL and MRDR-JL, which indicates that simply using cross-entropy loss to update the propensity model is not effective in improving the model performance. However, as the strength of the stabilization constraint increases, Stable-DR and Stable-MRDR using cycle learning have a stable and significant improvement compared to DR-JL and MRDR-JL. Our methods achieve the best performance at the appropriate constraint strength, which can be interpreted as simultaneous consideration of accuracy and stability to ensure boundedness and double robustness of SDR.

6 Conclusion

In this paper, we propose an SDR estimator for data MNAR that maintains double robustness and improves the stability of DR in the following three aspects: first, we show that SDR has a weaker extrapolation dependence than DR and can result in more stable and accurate predictions in the presence of MNAR effects. Next, through theoretical analysis, we show that the proposed SDR has bounded bias, variance, and generalization error bounds under inaccurate imputed errors and arbitrarily small estimated propensities, while DR does not. Finally, we propose a novel learning approach for SDR that updates the imputation, propensity, and prediction models cyclically, achieving more stable and accurate predictions. Extensive experiments show that our approach significantly outperforms the existing methods in terms of both convergence and prediction accuracy.

ETHICS STATEMENT

This work is mostly theoretical and experiments are based on synthetic and public datasets. We claim that this work does not present any foreseeable negative social impact.

REPRODUCIBILITY STATEMENT

Code is provided in Supplementary Materials to reproduce the experimental results.

REFERENCES

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

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972, 2005.

Stephen Bonner and Flavian Vasile. Causal embeddings for recommendation. In *RecSys*, 2018.

Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *arXiv:2010.03240*, 2020.

Jiawei Chen, Hande Dong, Yang Qiu, Xiangnan He, Xin Xin, Liang Chen, Guli Lin, and Keping Yang. Autodebias: Learning to debias for recommendation. In *SIGIR*, 2021.

Quanyu Dai, Haoxuan Li, Peng Wu, Zhenhua Dong, Xiao-Hua Zhou, Rui Zhang, Xiuqiang He, Rui Zhang, and Jie Sun. A generalized doubly robust learning framework for debiasing post-click conversion rate prediction. In *KDD*, 2022.

Sihao Ding, Peng Wu, Fuli Feng, Xiangnan He, Yitong Wang, Yong Liao, and Yongdong Zhang. Addressing unmeasured confounder for recommendation with sensitivity analysis. In *KDD*, 2022.

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. In *SIGIR*, 2021.

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

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Boca Raton: Chapman and Hall/CRC, 2020.

Guido W. Imbens and Donald B. Rubin. *Causal Inference For Statistics Social and Biomedical Science*. Cambridge University Press, 2015.

Joseph D.Y. Kang and Joseph L. Schafer. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22: 523–539, 2007.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

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

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

Yiming Liu, Xuezhi Cao, and Yong Yu. Are you influenced by others when rating?: Improve rating prediction by conformity modeling. In *RecSys*, 2016.

Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, and Bamshad Mobasher. Feedback loop and bias amplification in recommender systems. In *WWW*, 2020.

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

Geert Molenberghs, Garrett Fitzmaurice, Michael G. Kenward, Anastasios Tsiatis, and Geert Verbeke. *Handbook of Missing Data Methodology*. Chapman & Hall/CRC, 2015.

Jerzy Splawa Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5:465–472, 1990.

Xia Ning and George Karypis. Slim: Sparse linear methods for top-n recommender systems. In *2011 IEEE 11th international conference on data mining*, pp. 497–506. IEEE, 2011.

Yuta Saito. Unbiased pairwise learning from implicit feedback. In *NeurIPS Workshop*, 2019.

Yuta Saito. Asymmetric tri-training for debiasing missing-not-at-random explicit feedback. In *SIGIR*, 2020.

Yuta Saito. Doubly robust estimator for ranking metrics with post-click conversions. In *RecSys*, 2020.

Yuta Saito, Hayato Sakata, and Kazuhide Nakata. Doubly robust prediction and evaluation methods improve uplift modeling for observational data. In *SIAM*, 2019.

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

Masahiro Sato, Janmajay Singh, Sho Takemori, Takashi Sonoda, Qian Zhang, and Tomoko Ohkuma. Uplif-based evaluation and optimization of recommenders. In *RecSys*, 2019.

Masahiro Sato, Sho Takemori, Janmajay Singh, and Tomoko Ohkuma. Unbiased learning for the causal effect of recommendation. In *RecSys*, 2020.

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

Harald Steck. Evaluation of recommendations: rating-prediction and ranking. In *RecSys*, 2013.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *NeurIPS*, 2015.

Zhiqiang Tan. Comment: understanding or, ps and dr. *Statistical Science*, 22:560–568, 2007.

Mark J. van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011.

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

Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Combating selection biases in recommender systems with a few unbiased ratings. In *WSDM*, 2021.

Yixin Wang, Dawen Liang, Laurent Charlin, and David M. Blei. Causal inference for recommender systems. In *RecSys*, 2020a.

Zifeng Wang, Xi Chen, Rui Wen, Shao-Lun Huang, Ercan Kuruoglu, and Yefeng Zheng. Information theoretic counterfactual learning from missing-not-at-random feedback. *NeurIPS*, 2020b.

Hongyi Wen, Xinyang Yi, Tiansheng Yao, Jiaxi Tang, Lichan Hong, and Ed H. Chi. Distributionally-robust recommendations for improving worst-case user experience. In *WWW*, 2022.

Peng Wu, Haoxuan Li, Yuhao Deng, Wenjie Hu, Quanyu Dai, Zhenhua Dong, Jie Sun, Rui Zhang, and Xiao-Hua Zhou. On the opportunity of causal learning in recommendation systems: Foundation, estimation, prediction and challenges. In *IJCAI*, 2022.

Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52(1):1–38, 2019.

Wenhao Zhang, Wentian Bao, Xiao-Yang Liu, Keping Yang, Quan Lin, Hong Wen, and Ramin Ramezani. Large-scale causal approaches to debiasing post-click conversion rate estimation with multi-task learning. In *WWW*, 2020.

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

APPENDIX

Throughout, following existing work (Schnabel et al., 2016; Wang et al., 2019; Guo et al., 2021; Dai et al., 2022), we assume that the indicator matrix $\mathcal{O}$ contains independent random variables and each $o_{u,i}$ follows a Bernoulli distribution with probability $p_{u,i}$.

A PROOF OF THEOREMS

A.1 PROOF OF THEOREM 1

Proof of Theorem 1. To demonstrate the double robustness of the SDR, first note that $P\left(\lim_{|\mathcal{D}|\to\infty} \mathcal{E}_{SDR} = \mathcal{L}_{ideal}\right) = 1$ if the learned propensities are accurate (Swaminathan & Joachims, 2015), since $|\mathcal{D}|^{-1}\sum_{(u,i)\in\mathcal{D}} o_{u,i}/\hat{p}_{u,i}$ converges to 1 almost surely as $|\mathcal{D}|$ goes to infinity and IPS is unbiased. Besides, the constraint (3) is constructed to ensure the unbiasedness of $\mathcal{E}_{SDR}$ if the error imputation model is correctly specified. In fact, $\mathcal{E}_{SDR}$ satisfies

$$\frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}(e_{u,i} - \mathcal{E}_{SDR})}{\hat{p}_{u,i}} = 0. \tag{6}$$

Combining the constraint (3) and equation (6) gives

$$\frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}} \left[\frac{o_{u,i}(e_{u,i} - \hat{e}_{u,i})}{\hat{p}_{u,i}} + \frac{o_{u,i}(\hat{\mathcal{E}} - \mathcal{E}_{SDR})}{\hat{p}_{u,i}}\right] = 0, \tag{7}$$

where the first term equals to 0 when the imputation model is correctly specified, it implies that $\mathcal{E}_{SDR} = \hat{\mathcal{E}}$, then the unbiasedness of $\mathcal{E}_{SDR}$ follows immediately from the unbiasedness of $\hat{\mathcal{E}}$.

□

A.2 PROOF OF THEOREM 2

Proof of Theorem 2. Equation (5) implies that $\mathcal{E}_{SDR}$ can be expressed as

$$\mathcal{E}_{SDR} = \Big[\frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}(e_{u,i} - \hat{e}_{u,i} + \hat{\mathcal{E}})}{\hat{p}_{u,i}}\Big] \Big/ \Big[\frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}}{\hat{p}_{u,i}}\Big] \tag{8}$$

For notational simplicity, let

$$w_{u,i} \triangleq o_{u,i}/\hat{p}_{u,i} \quad \text{and} \quad v_{u,i} \triangleq o_{u,i}(e_{u,i} - \hat{e}_{u,i} + \hat{\mathcal{E}})/\hat{p}_{u,i},$$

then $\mathcal{E}_{SDR}$ can be written as a ratio statistic

$$\mathcal{E}_{SDR} = \frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}} v_{u,i} \Big/ \frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}} w_{u,i} \triangleq f(\bar{v}, \bar{w}),$$

where $f(v, w) = v/w$, $\bar{v} = |\mathcal{D}|^{-1}\sum_{(u,i)\in\mathcal{D}} v_{u,i}$, and $\bar{w} = |\mathcal{D}|^{-1}\sum_{(u,i)\in\mathcal{D}} w_{u,i}$.

Applying the Taylor expansion around $(\mu_v, \mu_w) \triangleq (\mathbb{E}[\bar{v}], \mathbb{E}[\bar{w}])$ yields that

$$\begin{aligned} f(\bar{v}, \bar{w}) &= f(\mu_v, \mu_w) + f'_v(\mu_v, \mu_w)\,(\bar{v} - \mu_v) + f'_w(\mu_v, \mu_w)\,(\bar{w} - \mu_w) \\ &\quad + \frac{1}{2}\left\{f''_{vv}(\mu_v, \mu_w)\,(\bar{v} - \mu_v)^2 + 2f''_{vw}(\mu_v, \mu_w)\,(\bar{v} - \mu_v)\,(\bar{w} - \mu_w) + f''_{ww}\,(\bar{w} - \mu_w)^2\right\} \\ &\quad + R(\tilde{v}, \tilde{w}), \end{aligned}$$

where $R(\tilde{v}, \tilde{w})$ is the remainder term. Note that $f''_{vv}(\mu_v, \mu_w) = 0$, $f''_{vw}(\mu_v, \mu_w) = -1/\mu_w^2$, and $f''_{ww}(\mu_v, \mu_w) = 2\mu_v/\mu_w^3$, then taking an expectation on both sides of the Taylor expansion leads to

$$\mathbb{E}(\bar{v}/\bar{w}) = \frac{\mu_v}{\mu_w} - \frac{\mathrm{Cov}(\bar{v}, \bar{w})}{(\mu_w)^2} + \frac{\mathrm{Var}(\bar{w})\mu_v}{(\mu_w)^3} + \mathbb{E}[R(\tilde{v}, \tilde{w})].$$

By some calculations, we have $\mathrm{Cov}(\bar{v}, \bar{w}) = O(|\mathcal{D}|^{-1}), \mathrm{Var}(\bar{w}) = O(|\mathcal{D}|^{-1}), \mathbb{E}[R(\tilde{v}, \tilde{w})] = o(|\mathcal{D}|^{-1})$. Thus, the bias of $\mathcal{E}_{SDR}$ is given as

$$\mathrm{Bias}(\mathcal{E}_{SDR}) = \left| \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \left(\delta_{u,i} - \frac{\sum_{(u,i)\in\mathcal{D}} \delta_{u,i} p_{u,i}/\hat{p}_{u,i}}{\sum_{(u,i)\in\mathcal{D}} p_{u,i}/\hat{p}_{u,i}} \right) \right| + O(|\mathcal{D}|^{-1}).$$

□

A.3 Proof of Theorem 3

Proof of Theorem 3. According to the proof of Theorem 2, we have

$$\mathbb{E}[\bar{v}/\bar{w}] - \mu_v/\mu_w = O(|\mathcal{D}|^{-1}). \tag{9}$$

Then the variance of $\mathcal{E}_{SDR}$ can be decomposed into as

$$\begin{aligned}
\mathrm{Var}(\mathcal{E}_{SDR}) &= \mathrm{Var}(\bar{v}/\bar{w}) = \mathbb{E}\left[\{\bar{v}/\bar{w} - \mathbb{E}(\bar{v}/\bar{w})\}^2\right] \\
&= \mathbb{E}\left[\{\bar{v}/\bar{w} - \mu_v/\mu_w\}^2 - 2O(|\mathcal{D}|^{-1}) \cdot \{\bar{v}/\bar{w} - \mu_v/\mu_w\} + O(|\mathcal{D}|^{-2})\right], \\
&= \mathcal{V}_1 + \mathcal{V}_2 + O(|\mathcal{D}|^{-2}),
\end{aligned}$$

where $\mathcal{V}_1 \triangleq \mathbb{E}[\{\bar{v}/\bar{w} - \mu_v/\mu_w\}^2]$, $\mathcal{V}_2 \triangleq -2O(|\mathcal{D}|^{-1}) \cdot [\mathbb{E}(\bar{v}/\bar{w}) - \mu_v/\mu_w]$. Equation (9) implies that $\mathcal{V}_2 = O(|\mathcal{D}|^{-2})$.

Denote $f(v, w) = v/w$, and apply delta method around $(\mu_v, \mu_w) \triangleq (\mathbb{E}[\bar{v}], \mathbb{E}[\bar{w}])$ to calculate $\mathcal{V}_1$ yields that

$$\begin{aligned}
\mathcal{V}_1 &= \mathbb{E}\left\{\left[f(\mu_v, \mu_w) + f'_v(\mu_v, \mu_w)(\bar{v} - \mu_v) + f'_w(\mu_v, \mu_w)(\bar{w} - \mu_w) + O_p(|\mathcal{D}|^{-1}) - f(\mu_v, \mu_w)\right]^2\right\} \\
&= \mathbb{E}\left\{\left[f'_v(\mu_v, \mu_w)(\bar{v} - \mu_v) + f'_w(\mu_v, \mu_w)(\bar{w} - \mu_w) + O_p(|\mathcal{D}|^{-1})\right]^2\right\} \\
&= \mathbb{E}\left\{f'^2_v(\mu_v, \mu_w)(\bar{v} - \mu_v)^2 + 2f'_v(\mu_v, \mu_w)(\bar{v} - \mu_v) f'_w(\mu_v, \mu_w)(\bar{w} - \mu_w)\right. \\
&\quad \left. + f'^2_w(\mu_v, \mu_w)(\bar{w} - \mu_w)^2\right\} + O(|\mathcal{D}|^{-2}) \\
&= f'^2_v(\mu_v, \mu_w)\,\mathrm{Var}(\bar{v}) + 2f'_v(\mu_v, \mu_w) f'_w(\mu_v, \mu_w)\,\mathrm{Cov}(\bar{v}, \bar{w}) + f'^2_w(\mu_v, \mu_w)\,\mathrm{Var}(\bar{w}) + O(|\mathcal{D}|^{-2})
\end{aligned}$$

Note that $f'_v(\mu_v, \mu_w) = 1/\mu_w$ and $f'_w(\mu_v, \mu_w) = -\mu_v/\mu_w^2$. Then we have

$$\begin{aligned}
\mathcal{V}_1 &= \frac{1}{(\mu_w)^2}\mathrm{Var}(\bar{v}) + 2\frac{-\mu_v}{(\mu_w)^3}\mathrm{Cov}(\bar{v}, \bar{w}) + \frac{(\mu_v)^2}{(\mu_w)^4}\mathrm{Var}(\bar{w}) + O(|\mathcal{D}|^{-2}) \\
&= \frac{(\mu_v)^2}{(\mu_w)^2}\left[\frac{\mathrm{Var}(\bar{v})}{(\mu_v)^2} - 2\frac{\mathrm{Cov}(\bar{v}, \bar{w})}{\mu_v \mu_w} + \frac{\mathrm{Var}(\bar{w})}{(\mu_w)^2}\right] + O(|\mathcal{D}|^{-2}) \\
&= \frac{\mathbb{E}\left(\bar{v} - \frac{\mu_v}{\mu_w}\bar{w}\right)^2}{\mu_w^2} + O(|\mathcal{D}|^{-2}) \\
&= \frac{\sum_{(u,i)} p_{u,i}(1 - p_{u,i}) h^2_{u,i}/\hat{p}^2_{u,i}}{\left(\sum_{(u,i)} p_{u,i}/\hat{p}_{u,i}\right)^2} + O(|\mathcal{D}|^{-2}),
\end{aligned}$$

where $h_{u,i} = (e_{u,i} - \hat{e}_{u,i}) - \sum_{(u,i)}\{p_{u,i}(e_{u,i} - \hat{e}_{u,i})/\hat{p}_{u,i}\}/\sum_{(u,i)}\{p_{u,i}/\hat{p}_{u,i}\}$ is a bounded difference between $e_{u,i} - \hat{e}_{u,i}$ and its weighted average. The conclusion that the SDR variance is bounded for any propensities is given directly by the self-normalized form of SDR, i.e., the bounded range of SDR is $[\delta_{\min}, \delta_{\max}]$.

□

A.4 PROOF OF THEOREM 4

Proof of Theorem 4. The McDiarmid's inequality states that for independent bounded random variables $X_1, X_2, \dots X_n$, where $X_i \in \mathcal{X}_i$ for all i and a mapping $f : \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$. Assume there exist constant $c_1, c_2, \dots, c_n$ such that for all i,

$$\sup_{x_1,\cdots,x_{i-1},x_i,x_i',x_{i+1},\cdots,x_n} |f(x_1,\dots,x_{i-1},x_i,x_{i+1},\cdots,x_n) - f(x_1,\dots,x_{i-1},x_i',x_{i+1},\cdots,x_n)| \le c_i.$$

Then, for any $\epsilon > 0$,

$$P\left(|f(X_1,X_2,\cdots,X_n) - \mathbb{E}[f(X_1,X_2,\cdots,X_n)]| \ge \epsilon\right) \le 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

In fact, equation (8) implies that the SDR estimator can be written as

$$\mathcal{E}_{SDR} = \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}(e_{u,i} - \hat{e}_{u,i})}{\hat{p}_{u,i}} \Big/ \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}}{\hat{p}_{u,i}} + \hat{\mathcal{E}},$$

denoted as $f(o_{1,1},\dots,o_{u,i},\dots,o_{U,I})$. Note that

$$\begin{aligned}
&\sup_{o_{u,i},o_{u,i}'} \left|f(o_{1,1},\dots,o_{u,i},\dots,o_{U,I}) - f\left(o_{1,1},\dots,o_{u,i}'\dots,o_{U,I}\right)\right| \\
&\le \begin{cases} \delta_{\max} - \dfrac{\delta_{u,i}/\hat{p}_{u,i} + \sum_{\mathcal{D}\setminus(u,i)} o_{u,i}/\hat{p}_{u,i}\delta_{\max}}{1/\hat{p}_{u,i} + \sum_{\mathcal{D}\setminus(u,i)} o_{u,i}/\hat{p}_{u,i}}, & \text{if} \quad \delta_{u,i} \le (\delta_{\min}+\delta_{\max})/2, \\ \dfrac{\sum_{\mathcal{D}\setminus(u,i)} o_{u,i}/\hat{p}_{u,i}\delta_{\min} + \delta_{u,i}/\hat{p}_{u,i}}{\sum_{\mathcal{D}\setminus(u,i)} o_{u,i}/\hat{p}_{u,i} + 1/\hat{p}_{u,i}} - \delta_{\min}, & \text{if} \quad \delta_{u,i} > (\delta_{\min}+\delta_{\max})/2, \end{cases}
\end{aligned} \tag{10}$$

where $\mathcal{D}\setminus(u,i)$ is the set of $\mathcal{D}$ excluding the element (u,i).

Next, we focus on analyzing the $\sum_{\mathcal{D}\setminus(u,i)} o_{u,i}/\hat{p}_{u,i}$. The Hoeffding's inequality states that for independent bounded random variables $X_1,\dots,X_n$ that take values in intervals of sizes $\rho_1,\dots,\rho_n$ with probability 1 and for any $\epsilon > 0$,

$$P\Big(\Big|\sum_k X_k - \mathbb{E}(\sum_k X_k)\Big| \ge \epsilon\Big) \le 2\exp\left(\frac{-2\epsilon^2}{\sum_k \rho_k^2}\right).$$

For $\sum_{\mathcal{D}\setminus(u,i)} o_{u,i}/\hat{p}_{u,i}$, we have

$$P\Big(\Big|\sum_{\mathcal{D}\setminus(u,i)} o_{u,i}/\hat{p}_{u,i} - \sum_{\mathcal{D}\setminus(u,i)} p_{u,i}/\hat{p}_{u,i}\Big| \ge \epsilon) \le 2\exp\left(\frac{-2\epsilon^2}{\sum_{\mathcal{D}\setminus(u,i)} 1/\hat{p}_{u,i}^2}\right),$$

Setting the last term equals to $\eta/2$, and solving for ϵ gives that with probability at least $1-\eta/2$, the following inequality holds

$$\left|\sum_{\mathcal{D}\setminus(u,i)} o_{u,i}/\hat{p}_{u,i} - \sum_{\mathcal{D}\setminus(u,i)} p_{u,i}/\hat{p}_{u,i}\right| \le \sqrt{\frac{1}{2}\log\frac{4}{\eta}\sum_{\mathcal{D}\setminus(u,i)}\frac{1}{\hat{p}_{u,i}^2}} \triangleq \epsilon'. \tag{11}$$

Therefore, combining (10) and (11) yields that with probability at least $1-\eta/2$,

$$\begin{aligned}
&\sup_{o_{1,1},\dots,o_{u,i},o_{u,i}',\dots,o_{U,I}} \left|f(o_{1,1},\dots,o_{u,i},\dots,o_{U,I}) - f\left(o_{1,1},\dots,o_{u,i}'\dots,o_{U,I}\right)\right| \\
&\le \begin{cases} \delta_{\max} - \dfrac{\delta_{u,i}/\hat{p}_{u,i} + (\sum_{\mathcal{D}\setminus(u,i)} p_{u,i}/\hat{p}_{u,i} - \epsilon')\delta_{\max}}{1/\hat{p}_{u,i} + (\sum_{\mathcal{D}\setminus(u,i)} p_{u,i}/\hat{p}_{u,i} - \epsilon')}, & \text{if} \quad \delta_{u,i} \le (\delta_{\min}+\delta_{\max})/2, \\ \dfrac{(\sum_{\mathcal{D}\setminus(u,i)} p_{u,i}/\hat{p}_{u,i} - \epsilon')\delta_{\min} + \delta_{u,i}/\hat{p}_{u,i}}{(\sum_{\mathcal{D}\setminus(u,i)} p_{u,i}/\hat{p}_{u,i} - \epsilon') + 1/\hat{p}_{u,i}} - \delta_{\min}, & \text{if} \quad \delta_{u,i} > (\delta_{\min}+\delta_{\max})/2, \end{cases} \\
&\le \begin{cases} (\delta_{\max} - \delta_{u,i})/\{1 + \hat{p}_{u,i}(\sum_{\mathcal{D}\setminus(u,i)} p_{u,i}/\hat{p}_{u,i} - \epsilon')\}, & \text{if} \quad \delta_{u,i} \le (\delta_{\min}+\delta_{\max})/2, \\ (\delta_{u,i} - \delta_{\min})/\{1 + \hat{p}_{u,i}(\sum_{\mathcal{D}\setminus(u,i)} p_{u,i}/\hat{p}_{u,i} - \epsilon')\}, & \text{if} \quad \delta_{u,i} > (\delta_{\min}+\delta_{\max})/2, \end{cases}
\end{aligned}$$

where $\delta_{u,i} = e_{u,i} - \hat{e}_{u,i}$ is the error deviation, $\delta_{\min} = \min_{(u,i)\in\mathcal{D}} \delta_{u,i}$, and $\delta_{\max} = \max_{(u,i)\in\mathcal{D}} \delta_{u,i}$. Invoking McDiarmid's inequality leads to that

$$
\begin{aligned}
& P\left(|\mathcal{E}_{SDR} - \mathbb{E}_{\mathcal{O}}(\mathcal{E}_{SDR})| \geq \epsilon\right) \\
\leq\ & 2\exp\Bigg\{-2\epsilon^2 \Bigg/ \Bigg(\sum_{(u,i):\delta_{u,i} \leq \frac{\delta_{\min}+\delta_{\max}}{2}} \frac{(\delta_{\max} - \delta_{u,i})^2}{\{1 + \hat{p}_{u,i}(\sum_{\mathcal{D}-(u,i)} p_{u,i}/\hat{p}_{u,i} - \epsilon')\}^2} \\
& + \sum_{(u,i):\delta_{u,i} > \frac{\delta_{\min}+\delta_{\max}}{2}} \frac{(\delta_{u,i} - \delta_{\min})^2}{\{1 + \hat{p}_{u,i}(\sum_{\mathcal{D}-(u,i)} p_{u,i}/\hat{p}_{u,i} - \epsilon')\}^2} \Bigg)\Bigg\} \\
\leq\ & 2\exp\left(\frac{-2\epsilon^2}{\sum_{(u,i)} \{(\delta_{\max} - \delta_{u,i})^2 + (\delta_{u,i} - \delta_{\min})^2\} / \{1 + \hat{p}_{u,i}(\sum_{\mathcal{D}-(u,i)} p_{u,i}/\hat{p}_{u,i} - \epsilon')\}^2} \right)
\end{aligned}
$$

Setting the last term equals to $\eta/2$, and solving for ϵ complete the proof.

□

A.5 Generalization Bound under Inaccurate Models

Theorem 5 (Generalization Bound under Inaccurate Models)**.** *For any finite hypothesis space of predictions $\mathcal{H} = \{\hat{\mathbf{Y}}^1, \ldots, \hat{\mathbf{Y}}^{|\mathcal{H}|}\}$, with probability $1 - \eta$, the true risk $R(\hat{\mathbf{Y}}^\dagger)$ deviates from the SDR estimator with imputed errors $\hat{e}_{u,i}$ and learned propensities $\hat{p}_{u,i}$ satisfying the stabilization constraint 3 is bounded by*

$$
\begin{aligned}
R(\hat{\mathbf{Y}}^\dagger) \leq \hat{\mathcal{E}}_{SDR}(\hat{\mathbf{Y}}^\dagger) &+ \underbrace{\left| \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \delta^\dagger_{u,i} - \frac{\sum_{(u,i)\in\mathcal{D}} \delta^\dagger_{u,i} p_{u,i}/\hat{p}_{u,i}}{\sum_{(u,i)\in\mathcal{D}} p_{u,i}/\hat{p}_{u,i}} \right|}_{\textit{Bias Term}} \\
&+ \underbrace{\sqrt{\frac{1}{2}\log\left(\frac{4|\mathcal{H}|}{\eta}\right) \sum_{(u,i)\in\mathcal{D}} \frac{(\delta_{\max} - \delta^\S_{u,i})^2 + (\delta^\S_{u,i} - \delta_{\min})^2}{\{1 + \hat{p}_{u,i}(\sum_{\mathcal{D}\setminus(u,i)} p_{u,i}/\hat{p}_{u,i} - \epsilon')\}^2}}}_{\textit{Variance Term}}
\end{aligned}
$$

where $\delta^\S_{u,i}$ is the error deviation corresponding to the prediction model

$$
\hat{\mathbf{Y}}^\S = \operatorname{argmax}_{\hat{\mathbf{Y}}^h \in \mathcal{H}} \sum_{(u,i)\in\mathcal{D}} \frac{(\delta_{\max} - \delta^\S_{u,i})^2 + (\delta^\S_{u,i} - \delta_{\min})^2}{\{1 + \hat{p}_{u,i}(\sum_{\mathcal{D}\setminus(u,i)} p_{u,i}/\hat{p}_{u,i} - \epsilon')\}^2}.
$$

Proof of Theorem 5. Proof. Theorem 4 shows that for all predictions $\hat{\mathbf{Y}}^h \in \mathcal{H}$, we have

$$
\begin{aligned}
& P\left(\left| \mathcal{E}_{SDR}(\hat{\mathbf{Y}}^h) - \mathbb{E}[\mathcal{E}_{SDR}(\hat{\mathbf{Y}}^h)] \right| \geq \epsilon \right) \\
\leq\ & 2\exp\left(\frac{-2\epsilon^2}{\sum_{(u,i)} \{(\delta_{\max} - \delta^h_{u,i})^2 + (\delta^h_{u,i} - \delta_{\min})^2\} / \{1 + \hat{p}_{u,i}\}^2} \right)
\end{aligned}
$$

McDiarmid's inequality and union bound ensures the following uniform convergence results:

$$
\begin{aligned}
&P\left(\left|\mathcal{E}_{SDR}(\hat{\mathbf{Y}}^{\dagger}) - \mathbb{E}[\mathcal{E}_{SDR}(\hat{\mathbf{Y}}^{\dagger})]\right| \leq \epsilon\right) \geq 1-\eta \\
&\Leftarrow P\left(\max_{\hat{\mathbf{Y}}^h \in \mathcal{H}} \left|\mathcal{E}_{SDR}(\hat{\mathbf{Y}}^{h}) - \mathbb{E}[\mathcal{E}_{SDR}(\hat{\mathbf{Y}}^{h})]\right| \leq \epsilon\right) \geq 1-\eta \\
&\Leftrightarrow P\left(\bigvee_{\hat{\mathbf{Y}}_i \in \mathcal{H}} \left|\mathcal{E}_{SDR}(\hat{\mathbf{Y}}^{h}) - \mathbb{E}[\mathcal{E}_{SDR}(\hat{\mathbf{Y}}^{h})]\right| \geq \epsilon\right) < \eta \\
&\Leftarrow \sum_{h=1}^{|\mathcal{H}|} P\left(\left|\mathcal{E}_{SDR}(\hat{\mathbf{Y}}^{h}) - \mathbb{E}[\mathcal{E}_{SDR}(\hat{\mathbf{Y}}^{h})]\right| \geq \epsilon\right) < \eta \\
&\Leftarrow \sum_{h=1}^{|\mathcal{H}|} 2\exp\left(\frac{-2\epsilon^2}{\sum_{(u,i)}\{(\delta_{\max} - \delta^h_{u,i})^2 + (\delta^h_{u,i} - \delta_{\min})^2\}/\{1+\hat{p}_{u,i}(\sum_{\mathcal{D}-(u,i)} p_{u,i}/\hat{p}_{u,i} - \epsilon')\}^2}\right) < \eta \\
&\Leftarrow |\mathcal{H}| \cdot 2\exp\left(\frac{-2\epsilon^2}{\sum_{(u,i)}\{(\delta_{\max} - \delta^{\S}_{u,i})^2 + (\delta^{\S}_{u,i} - \delta_{\min})^2\}/\{1+\hat{p}_{u,i}(\sum_{\mathcal{D}-(u,i)} p_{u,i}/\hat{p}_{u,i} - \epsilon')\}^2}\right) < \eta
\end{aligned}
$$

Solving the last inequality for ϵ, it is concluded that, with probability $1-\eta$, the following inequality holds

$$
\mathbb{E}[\mathcal{E}_{SDR}(\hat{\mathbf{Y}}^{\dagger})] - \mathcal{E}_{SDR}(\hat{\mathbf{Y}}^{\dagger}) \leq \sqrt{\frac{1}{2}\log\left(\frac{4|\mathcal{H}|}{\eta}\right) \sum_{(u,i)\in\mathcal{D}} \frac{(\delta_{\max} - \delta^{\S}_{u,i})^2 + (\delta^{\S}_{u,i} - \delta_{\min})^2}{\{1+\hat{p}_{u,i}(\sum_{\mathcal{D}\setminus(u,i)} p_{u,i}/\hat{p}_{u,i} - \epsilon')\}^2}}.
$$

Theorem 2 shows that for the optimal prediction model $\hat{\mathbf{Y}}^{\dagger}$, the following inequality holds

$$
R(\hat{\mathbf{Y}}^{\dagger}) - \mathbb{E}[\mathcal{E}_{SDR}(\hat{\mathbf{Y}}^{\dagger})] \leq \left|\frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}} \delta^{\dagger}_{u,i} - \frac{\sum_{(u,i)\in\mathcal{D}} \delta^{\dagger}_{u,i} p_{u,i}/\hat{p}_{u,i}}{\sum_{(u,i)\in\mathcal{D}} p_{u,i}/\hat{p}_{u,i}}\right|.
$$

The stated results can be obtained by adding the two inequalities above.

□

B Further Theoretical Analysis of SDR

Without loss of generality, we assume

$$
\frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}}{\hat{p}_{u,i}}(\hat{e}_{u,i} - \hat{\mathcal{E}}) = \lambda, \quad \lambda \neq 0.
$$

In this case, the learned propensities must be inaccurate; otherwise, the constraint (3 holds naturally as the same size increases. Thus, if the imputed errors are accurate, then $\mathcal{L}_{ideal} = \hat{\mathcal{E}}$. By a exactly same arguments of equation (5), we have

$$
\frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}}{\hat{p}_{u,i}}(\hat{\mathcal{E}} - \mathcal{E}_{SDR}) = \lambda,
$$

which implies that

$$
\mathcal{E}_{SDR} = \mathcal{L}_{ideal} - \lambda \Big/ \frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}}{\hat{p}_{u,i}}.
$$

This means that the degree of violation of constraint (3) determines the size of the bias of SDR.

Furthermore, we can compute the bias, variance, tail bound, and generalization error bound of SDR. Specifically, if both the learned propensities and imputed errors are inaccurate, constraint (3) does not hold either. Then the bias of SDR is

$$
\text{Bias}(\mathcal{E}_{SDR}) = \left|\frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}} \left(e_{u,i} - \frac{\sum_{(u,i)\in\mathcal{D}} e_{u,i} p_{u,i}/\hat{p}_{u,i}}{\sum_{(u,i)\in\mathcal{D}} p_{u,i}/\hat{p}_{u,i}}\right)\right| + O(|\mathcal{D}|^{-1}),
$$

the variance of SDR becomes

$$\text{Var}\left(\mathcal{E}_{SDR}\right) = \frac{\sum_{(u,i)} p_{u,i}(1-p_{u,i})\tilde{h}_{u,i}^2/\hat{p}_{u,i}^2}{\left(\sum_{(u,i)} p_{u,i}/\hat{p}_{u,i}\right)^2} + O(|\mathcal{D}|^{-2}),$$

where $\tilde{h}_{u,i} = e_{u,i} - \sum_{(u,i)\in\mathcal{D}}\{p_{u,i}e_{u,i}/\hat{p}_{u,i}\} \Big/ \sum_{(u,i)\in\mathcal{D}}\{p_{u,i}/\hat{p}_{u,i}\}$. The tail bound of SDR is given as

$$|\mathcal{E}_{SDR} - \mathbb{E}_{\mathcal{O}}(\mathcal{E}_{SDR})| \leq \sqrt{\frac{1}{2}\log\left(\frac{4}{\eta}\right)\sum_{(u,i)\in\mathcal{D}} \frac{(e_{\text{max}} - e_{u,i})^2 + (e_{u,i} - e_{\text{min}})^2}{\{1 + \hat{p}_{u,i}(\sum_{\mathcal{D}\setminus(u,i)} p_{u,i}/\hat{p}_{u,i} - \epsilon')\}^2}},$$

where $\delta_{\text{min}} = \min_{(u,i)\in\mathcal{D}} e_{u,i}$, $\delta_{\text{max}} = \max_{(u,i)\in\mathcal{D}} e_{u,i}$, $\epsilon' = \sqrt{\log(4/\eta)/2 \cdot \sum_{\mathcal{D}\setminus(u,i)} 1/\hat{p}_{u,i}^2}$, and $\mathcal{D}\setminus(u,i)$ is the set of $\mathcal{D}$ excluding the element (u,i). In addition, we can derive the generation error bound of SDR. Given a finite hypothesis space $\mathcal{H}$ of the prediction model, then for any a prediction model $h \in \mathcal{H}$, with probability $1-\eta$, the true risk $R(h)$ deviates from the SDR estimator is bounded by

$$R(h) \leq \hat{\mathcal{E}}_{SDR}(h) + \text{Bias}(\mathcal{E}_{SDR}) + \sqrt{\frac{1}{2}\log(\frac{4|\mathcal{H}|}{\eta})\sum_{(u,i)\in D} \frac{(e_{\text{max}} - e_{u,i}^{\S})^2 + (e_{u,i}^{\S} - e_{\text{min}})^2}{\{1 + \hat{p}_{u,i}(\sum_{\mathcal{D}\setminus(u,i)} p_{u,i}/\hat{p}_{u,i} - \epsilon')\}^2}},$$

where $e_{u,i}^{\S}$ is the error deviation corresponding to the prediction model

$$h^{\S} = \arg\max_{h\in\mathcal{H}} \sum_{(u,i)\in D} \frac{(e_{max} - e_{u,i}^{\S})^2 + (e_{u,i}^{\S} - e_{min})^2}{\{1 + \hat{p}_{u,i}(\sum_{D-(u,i)} p_{u,i}/\hat{p}_{u,i} - \epsilon')\}^2}.$$

C Further Real-world Experiments

C.1 Take NCF as the Base Model

Table 3: Evaluation metrics of MSE, AUC, NDCG@5, and NDCG@10 on the MAR test set of Coat and Yahoo!R3 using Naive Bayes with Laplace smoothing as the propensity model and taking NCF as the base model. The best two results are bolded.

Datasets	Coat				Yahoo!R3			
Methods	MSE	AUC	N@5	N@10	MSE	AUC	N@5	N@10
NCF	0.2050	0.7670	0.6228	0.6954	0.3215	0.6782	0.6501	0.7672
+ IPS	0.2042	0.7646	0.6327	0.7054	0.1777	0.6719	0.6548	0.7703
+ SNIPS	0.1904	0.7707	0.6271	0.7062	0.1694	0.6903	0.6693	0.7807
+ AS-IPS	0.2061	0.7630	0.6156	0.6983	0.1715	0.6879	0.6620	0.7769
+ CVIB	0.2042	0.7655	0.6176	0.6946	0.3088	0.6715	0.6669	0.7793
+ DR	0.2081	0.7578	0.6119	0.6900	0.1705	0.6886	0.6628	0.7768
+ DR-JL	0.2115	0.7600	0.6272	0.6967	0.2452	0.6818	0.6516	0.7678
+ Stable-DR (Ours)	**0.1896**	**0.7712**	**0.6337**	**0.7095**	**0.1664**	**0.6907**	**0.6756**	**0.7861**
+ MRDR-JL	0.2046	0.7609	0.6182	0.6992	0.2367	0.6778	0.6465	0.7664
+ Stable-MRDR (Ours)	**0.1899**	**0.7710**	**0.6380**	**0.7082**	**0.1671**	**0.6910**	**0.6734**	**0.7846**

C.2 Effects of Varying Alternation Training Order

C.2.1 Motivation for Alternatively Updating the Models

From a high-level perspective, learning a recommendation model consists of two closely related tasks:

- policy evaluation: find an efficient estimator that approximates the ideal loss $\mathcal{L}_{ideal}$, and then use the estimation as the evaluation metric of the recommendation model.
- policy improvement: design an efficient algorithm that optimizes the proposed estimator to update the prediction model.

For unbiased recommendations learning, previous works (e.g., DR (Wang et al., 2019) and MRDR (Guo et al., 2021)) have shown that alternately updating the imputation model (policy evaluation) and prediction model (policy improvement) results in better performance due to their interdependence. Further, recent works (e.g., Learning to Debias (Wang et al., 2021) and Autodebias (Chen et al., 2021)) have shown that alternately updating all three models (propensity, imputation, and prediction models) can further improve the debiasing performance. However, our concern is that the latter simply treats the updating of the propensity and imputation model as a whole, without designing the loss separately, which is still essentially joint learning.

In our work, cycle learning is proposed, as shown in Figure 1. We provide more design details, especially focusing on the alternate-update-orders, and demonstrate the advantages of cycle learning over previous work theoretically and empirically as follows (see Appendixes C.2.2 and C.2.3).

C.2.2 Theoretical Guarantees for Alternation Training Orders of the Proposed Stable-DR and Stable-MRDR

Theoretically, the designed algorithm **strictly follows the proposed SDR estimator** in Section 3.2. As shown in Figure 1 and Alg. 1 in Section 4, our algorithm **first updates imputed errors $\hat{e}$ by Step 1**, and **then learns a propensity $\hat{p}$ based on learned $\hat{e}$ to satisfy the constraint (3) in Step 2.** The main purpose of the first two steps is to ensure that the SDR estimator in step 3 has double robustness and has a lower extrapolation dependence compared to the previous DR methods. **Finally, from Step 3 we update the predicted rating $\hat{r}$ by minimizing the estimation of the ideal loss using the proposed SDR estimator.** For the next round, instead of re-initializing, Step 1 updates the imputed errors $\hat{e}$ according to the new prediction model, then Step 2 re-updates the constrained propensities $\hat{p}$, and then uses Step 3 to update the prediction model $\hat{r}$ again. And so on, each model is updated for each Step in our algorithm: Step 1–imputation model $\hat{e}$, Step 2–propensity model $\hat{p}$, Step 3–prediction model $\hat{r}$. Thus the alternate-update-order in our algorithm is acted as $\cdots \longrightarrow \hat{e} \longrightarrow \hat{p} \longrightarrow \hat{r} \longrightarrow \cdots$.

C.2.3 Empirical Performance Comparisons on Varying Alternation Training Orders of the Proposed Stable-DR and Stable-MRDR

We conduct extensive experiments to investigate the effect of alternate-update-order on prediction performance using MF, NCF, and SLIM as the base model, Coat and Yahoo as two datasets respectively. Specifically, as in Figure 1, the vanilla DR uses a two-phase learning to update the prediction model without updating alternately. We use DR-JL and MRDR-JL as baselines, which both alternately update $\hat{e}$ and $\hat{r}$, while keeping $\hat{p}$ static. For the proposed Stable-DR and Stable-MRDR, the default alternate-update-order is $\cdots \longrightarrow \hat{e} \longrightarrow \hat{p} \longrightarrow \hat{r} \longrightarrow \cdots$, as stated in Appendix C.2.2. For comparison purposes, we also study Stable-DR and Stable-MRDR performance using an inverse alternate-update-order $\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \hat{p} \longrightarrow \cdots$ with † marked.

The empirical results are presented in Tables 4–9. We have the following findings: First, the same with the previous findings in Appendix C.2.1, updating alternately always leads to better prediction performance, which is due to performing policy evaluation and policy improvement alternately. Second, the proposed Stable-DR and Stable-MRDR have significantly better performance than DR-JL and MRDR-JL. This is due to the theoretical guarantees of the SDR estimator and our algorithm designs, in particular that the alternate-update-order of the models strictly follows the SDR estimator in Section 3.2 (see the discussion in Appendix C.2.2). Compared to DR and MRDR, our method has lower bias and variance in the presence of small propensities. Third, for the proposed Stable-DR and Stable-MRDR, we find that the regular alternate-update-order $\cdots \longrightarrow \hat{e} \longrightarrow \hat{p} \longrightarrow \hat{r} \longrightarrow \cdots$ has a slightly better performance than using a inverse alternate-update-order $\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \hat{p} \longrightarrow \cdots$. This is because the inverse alternate-update-order uses the **outdated information** when updating the prediction model. Specifically, the inverse alternate-update-order is

$$\cdots \longrightarrow \hat{e}_{t-2} \longrightarrow \hat{r}_{t-2} \longrightarrow \hat{p}_{t-2} \longrightarrow \hat{e}_{t-1} \longrightarrow \hat{r}_{t-1} \longrightarrow \hat{p}_{t-1} \longrightarrow \hat{e}_{t} \longrightarrow \hat{r}_{t} \longrightarrow \cdots,$$

where it should be noted that when updating $\hat{r}_t$, we always use $\hat{p}_{t-1}$ as the learned propensities instead of $\hat{p}_t$. However, the learning of constrained $\hat{p}_{t-1}$ uses $\hat{e}_{t-1}$ as the imputed errors in constraint (3), which is the policy evaluation of $\hat{r}_{t-2}$. **This results in the update of r_t not using the information of r_{t-1} but the information of r_{t-2}.** Despite the fact that the inverse alternate-update-order does not perform as well as the regular alternate-update-order, the Stable-DR and Stable-MRDR based on alternate-update-order still stably outperforms the competing DR baselines, because the smaller bias and variance properties are maintained in the presence of small propensities, and the problem for the inverse update is only that it uses outdated information.

Table 4: Effects of varying alternate-update-order of proposed Stable-DR compared with baseline DR methods on MSE, AUC, NDCG@5, and NDCG@10 on **Coat**, using MF as the base model. † denotes the use of an inverse alternate-update-order of Stable-DR to train the prediction model. The bold value marks the best method in one column, while the underlined value corresponds to the second best method.

Methods	Alternate-update-order	MSE	AUC	N@5	N@10
MF-DR	$\hat{e} \longrightarrow \hat{r}$	0.2298	0.7132	0.6243	0.6918
MF-DR-JL	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \cdots$	0.2254	0.7209	0.6252	0.6961
MF-Stable-DR	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{p} \longrightarrow \hat{r} \longrightarrow \cdots$	<u>0.2159</u>	**0.7508**	**0.6511**	**0.7073**
MF-Stable-DR†	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \hat{p} \longrightarrow \cdots$	**0.2133**	<u>0.7475</u>	<u>0.6391</u>	<u>0.7021</u>
MF-MRDR-JL	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \cdots$	0.2252	0.7318	0.6375	0.6989
MF-Stable-MRDR	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{p} \longrightarrow \hat{r} \longrightarrow \cdots$	**0.2076**	**0.7548**	**0.6532**	**0.7105**
MF-Stable-MRDR†	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \hat{p} \longrightarrow \cdots$	<u>0.2103</u>	<u>0.7507</u>	<u>0.6473</u>	<u>0.7054</u>

Table 5: Effects of varying alternate-update-order of proposed Stable-DR compared with baseline DR methods on MSE, AUC, NDCG@5, and NDCG@10 on **Yahoo**, using MF as the base model. † denotes the use of an inverse alternate-update-order of Stable-DR to train the prediction model. The bold value marks the best method in one column, while the underlined value corresponds to the second best method.

Methods	Alternate-update-order	MSE	AUC	N@5	N@10
MF-DR	$\hat{e} \longrightarrow \hat{r}$	0.2093	0.6873	0.6574	0.7741
MF-DR-JL	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \cdots$	<u>0.2194</u>	0.6863	0.6525	0.7701
MF-Stable-DR	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{p} \longrightarrow \hat{r} \longrightarrow \cdots$	**0.2090**	**0.6946**	<u>0.6620</u>	<u>0.7786</u>
MF-Stable-DR†	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \hat{p} \longrightarrow \cdots$	0.2753	<u>0.6929</u>	**0.6655**	**0.7788**
MF-MRDR-JL	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \cdots$	<u>0.2173</u>	0.6830	0.6437	0.7652
MF-Stable-MRDR	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{p} \longrightarrow \hat{r} \longrightarrow \cdots$	**0.2081**	<u>0.6915</u>	**0.6585**	**0.7757**
MF-Stable-MRDR†	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \hat{p} \longrightarrow \cdots$	0.2523	**0.6945**	<u>0.6555</u>	<u>0.7743</u>

Table 6: Effects of varying alternate-update-order of proposed Stable-DR compared with baseline DR methods on MSE, AUC, NDCG@5, and NDCG@10 on **Coat**, using NCF as the base model. † denotes the use of an inverse alternate-update-order of Stable-DR to train the prediction model. The bold value marks the best method in one column, while the underlined value corresponds to the second best method.

Methods	Alternate-update-order	MSE	AUC	N@5	N@10
NCF-DR	$\hat{e} \longrightarrow \hat{r}$	0.2081	0.7578	0.6119	0.6900
NCF-DR-JL	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \cdots$	0.2115	0.7600	<u>0.6272</u>	0.6967
NCF-Stable-DR	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{p} \longrightarrow \hat{r} \longrightarrow \cdots$	**0.1896**	**0.7712**	**0.6337**	**0.7095**
NCF-Stable-DR†	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \hat{p} \longrightarrow \cdots$	<u>0.1921</u>	<u>0.7650</u>	0.6194	<u>0.7004</u>
NCF-MRDR-JL	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \cdots$	0.2046	0.7609	0.6182	0.6992
NCF-Stable-MRDR	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{p} \longrightarrow \hat{r} \longrightarrow \cdots$	**0.1899**	**0.7710**	<u>0.6380</u>	**0.7082**
NCF-Stable-MRDR†	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \hat{p} \longrightarrow \cdots$	<u>0.1900</u>	<u>0.7673</u>	**0.6400**	<u>0.7046</u>

Table 7: Effects of varying alternate-update-order of proposed Stable-DR compared with baseline DR methods on MSE, AUC, NDCG@5, and NDCG@10 on **Yahoo**, using NCF as the base model. † denotes the use of an inverse alternate-update-order of Stable-DR to train the prediction model. The bold value marks the best method in one column, while the underlined value corresponds to the second best method.

Methods	Alternate-update-order	MSE	AUC	N@5	N@10
NCF-DR	$\hat{e} \longrightarrow \hat{r}$	0.1705	0.6886	0.6628	0.7768
NCF-DR-JL	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \cdots$	0.2452	0.6818	0.6516	0.7678
NCF-Stable-DR	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{p} \longrightarrow \hat{r} \longrightarrow \cdots$	**0.1664**	**0.6907**	<u>0.6756</u>	<u>0.7861</u>
NCF-Stable-DR†	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \hat{p} \longrightarrow \cdots$	<u>0.1681</u>	<u>0.6863</u>	**0.6768**	**0.7864**
NCF-MRDR-JL	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \cdots$	0.2367	0.6778	0.6465	0.7664
NCF-Stable-MRDR	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{p} \longrightarrow \hat{r} \longrightarrow \cdots$	**0.1671**	**0.6910**	**0.6734**	**0.7846**
NCF-Stable-MRDR†	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \hat{p} \longrightarrow \cdots$	<u>0.1672</u>	<u>0.6909</u>	<u>0.6719</u>	<u>0.7836</u>

Table 8: Effects of varying alternate-update-order of proposed Stable-DR compared with baseline DR methods on MSE, AUC, NDCG@5, and NDCG@10 on **Coat**, using SLIM as the base model. † denotes the use of an inverse alternate-update-order of Stable-DR to train the prediction model. The bold value marks the best method in one column, while the underlined value corresponds to the second best method.

Methods	Alternate-update-order	MSE	AUC	N@5	N@10
SLIM-DR	$\hat{e} \longrightarrow \hat{r}$	0.2334	0.7064	0.7267	0.7649
SLIM-DR-JL	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \cdots$	<u>0.2407</u>	0.7090	0.7279	0.7655
SLIM-Stable-DR	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{p} \longrightarrow \hat{r} \longrightarrow \cdots$	**0.2356**	**0.7201**	**0.7389**	**0.7724**
SLIM-Stable-DR†	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \hat{p} \longrightarrow \cdots$	0.2472	<u>0.7134</u>	<u>0.7250</u>	<u>0.7696</u>
SLIM-MRDR-JL	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \cdots$	<u>0.2409</u>	0.7074	<u>0.7329</u>	0.7679
SLIM-Stable-MRDR	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{p} \longrightarrow \hat{r} \longrightarrow \cdots$	**0.2369**	**0.7148**	**0.7378**	**0.7711**
SLIM-Stable-MRDR†	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \hat{p} \longrightarrow \cdots$	0.2511	<u>0.7140</u>	0.7294	<u>0.7683</u>

D MONTE CARLO STUDIES ON ML-100K AND ML-1M

D.1 EXPERIMENTAL SETUP

Table 9: Effects of varying alternate-update-order of proposed Stable-DR compared with baseline DR methods on MSE, AUC, NDCG@5, and NDCG@10 on **Yahoo**, using SLIM as the base model. † denotes the use of an inverse alternate-update-order of Stable-DR to train the prediction model. The bold value marks the best method in one column, while the underlined value corresponds to the second best method.

Methods	Alternate-update-order	MSE	AUC	N@5	N@10
SLIM-DR	$\hat{e} \longrightarrow \hat{r}$	0.2054	0.6771	0.7344	0.8248
SLIM-DR-JL	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \cdots$	**0.2044**	0.6792	0.7360	0.8260
SLIM-Stable-DR	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{p} \longrightarrow \hat{r} \longrightarrow \cdots$	0.2080	**0.6874**	**0.7473**	**0.8349**
SLIM-Stable-DR†	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \hat{p} \longrightarrow \cdots$	0.2063	0.6859	0.7431	0.8305
SLIM-MRDR-JL	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \cdots$	0.2016	0.6791	0.7338	0.8239
SLIM-Stable-MRDR	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{p} \longrightarrow \hat{r} \longrightarrow \cdots$	0.2086	**0.6842**	**0.7435**	**0.8308**
SLIM-Stable-MRDR†	$\cdots \longrightarrow \hat{e} \longrightarrow \hat{r} \longrightarrow \hat{p} \longrightarrow \cdots$	**0.1994**	0.6819	0.7403	0.8289

We conduct several experiments based on **MovieLens 100K**[6] **(ML-100K)** dataset, which has 943 users, 1,682 items with 100,000 MNAR ratings, and **MovieLens 1M**[7] **(ML-1M)** dataset, which has 6,040 users, 3,952 items with 1,000,209 MNAR ratings. MAR ratings are no longer available for both datasets. Following the previous studies Schnabel et al. (2016); Wang et al. (2019); Saito (2020); Guo et al. (2021), the detailed pre-processing process for above two datasets is shown below.

(1) Conduct the full rating matrix by implementing Matrix Factorization (MF) Koren et al. (2009). Since the rating matrix will have a unrealistic high prediction rating values for almost all user-item pair, we first sorting all ratings in ascending order, then set ratings between the lowest rating and p_1 fraction to 1, set all ratings in next fraction p_2 to 2, and so on. The adjusted rating matrix contains $R_{u,i} \in \{1, 2, 3, 4, 5\}$ with corresponding proportion $[p_1, p_2, p_3, p_4, p_5]$ respectively, which has a more realistic distribution.

(2) Set the true propensity $p_{u,i} \in (0, 1)$ for each rating $R_{u,i}$, with $p_{u,i} = p\alpha^{1+\min(4,5-R_{u,i})\cdot 0.25*k}$, where $p = 1$, $\alpha = 0.5$, $R_{u,i} \in \{1, 2, 3, 4, 5\}$, and $k \in \{1, 2, 3, 4\}$ to generate 4 variants with different selection bias level and propensity scale for **ML-100K** and **ML-1M** respectively. In addition, we transfer the adjusted rating matrix to true Click Value Rate (CVR) matrix by replacing $R_{u,i} \in \{1, 2, 3, 4, 5\}$ with $r_{u,i}^{true} \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ correspondingly.

(3) Because only binary click indicators and conversion labels can be observed, we sample both click indicators and conversion labels according to the following formula:

$$o_{u,i} \sim \mathrm{Bern}(p_{u,i}), r_{u,i} \sim \mathrm{Bern}(r_{u,i}^{true}), \forall (u, i) \in \mathcal{D},$$

where $\mathrm{Bern}(\cdot)$ denotes the Bernoulli distribution. Then we obtain a fully observed click indicator matrix $\mathbf{O}$ and a conversion label matrix $\mathbf{R}$.

Experimental details. To evaluate the performance of the proposed estimator, several different predicted CVR matrices $\hat{\mathbf{R}}$ based on the true CVR matrix are conducted as follows:

• **ONE**: The predicted CVR matrix $\hat{\mathbf{R}}$ is identical to the true CVR matrix, except that $|\{(u, i) \mid r_{u,i}^{\text{true}} = 0.9\}|$ randomly selected true ratings of 0.1 are flipped to 0.9. This means half of the predicted 0.9 are true 0.9, and half are true 0.1.

• **THREE**: Same as ONE, but flipping true CVR of 0.3 instead.

• **FIVE**: Same as ONE, but flipping true CVR of 0.5 instead.

• **SKEW**: Predictions CVR $\hat{r}_{u,i}$ are sampled from the Gaussian distribution $\mathcal{N}(\mu = r_{u,i}^{\text{true}}, \sigma = (1 - r_{u,i}^{\text{true}})/2)$, and clipped to the interval $[0.1, 0.9]$.

• **CRS**: If the true CVR $r_{u,i}^{\text{true}} \leq 0.6$, then $\hat{r}_{u,i} = 0.2$. Otherwise, $\hat{r}_{u,i} = 0.6$.

After getting the predicted CVR matrix, we obtain the predicted propensity matrix by $1/\hat{p}_{u,i} = (1-\beta)/p_{u,i} + \beta/p_e$, where $p_e = |\mathcal{D}|^{-1} \sum_{(u,i)\in\mathcal{D}} o_{u,i}$, and β is randomly sampled from the uniform

[6] https://grouplens.org/datasets/movielens/100k/

[7] https://grouplens.org/datasets/movielens/1M/

distribution $U(0,\ 1)$ to introduce noises. We follow Guo et al. (2021); Dai et al. (2022) to use the absolute relative error (RE) for evaluation, which is defined as follows

$$\mathrm{RE}(\mathcal{E}_{estimator}) = \frac{|\mathcal{E}_{ideal}(\hat{\mathbf{R}}) - \mathcal{E}_{estimator}(\hat{\mathbf{R}})|}{\mathcal{E}_{ideal}(\hat{\mathbf{R}}),}$$

where $\mathcal{E}_{\text{estimator}}$ denotes the estimator to be compared. In our experiments, the baseline estimators are chosen as Naive, IPS, SNIPS and DR. RE evaluates the accuracy of the estimated loss, the smaller the RE, the more accurate the estimation of the ideal loss.

D.2 Experimental Result

Table 10 shows the performance of the baseline estimators and the proposed SDR under different selection bias levels. We have the following findings: first, overall, SDR performs significantly better than baselines and the variance is smaller compared to both DR and SNIPS, which empirically demonstrates the stability of our method. Second, there is no significant difference in the performance of IPS and SNIPS in all cases, but our proposed SDR estimator outperforms DR estimator at a statistically significant level of 0.001 in all cases. This suggests that our superior performance is not attributed to self-normalize form, but to carefully designed model dependency, which utilizes the propensity model and imputation model simultaneously in a differentiated way. Third, as the selection bias level increases from ML-100K/1M-1 to ML-100K/1M-4, it can be seen that the estimation error of all baseline estimators increases significantly, while SDR does not. This is attributed to the fact that the proposed SDR estimator is more stable to small propensities.

Table 10: Mean and standard deviation of the relative error on the Naive, IPS, SNIPS, DR and SDR.

Dataset	Methods	ONE	THREE	FIVE	SKEW	CRS
ML-100K-1	Naive	0.0355 ± 0.0013	0.0471 ± 0.0013	0.0519 ± 0.0013	0.0204 ± 0.0010	0.0533 ± 0.0012
	IPS	0.0180 ± 0.0016	0.0236 ± 0.0015	0.0260 ± 0.0015	0.0101 ± 0.0011	0.0265 ± 0.0014
	SNIPS	0.0180 ± 0.0014	0.0236 ± 0.0013	0.0261 ± 0.0013	0.0102 ± 0.0011	0.0266 ± 0.0012
	DR	0.0129 ± 0.0013	0.0149 ± 0.0013	0.0124 ± 0.0013	0.0080 ± 0.0010	0.0178 ± 0.0011
	SDR	**0.0101 ± 0.0013***	**0.0115 ± 0.0012***	**0.0087 ± 0.0013***	**0.0070 ± 0.0009***	**0.0174 ± 0.0011***
ML-1M-1	Naive	0.0347 ± 0.0003	0.0467 ± 0.0004	0.0514 ± 0.0004	0.0204 ± 0.0003	0.0532 ± 0.0002
	IPS	0.0173 ± 0.0004	0.0233 ± 0.0005	0.0257 ± 0.0005	0.0101 ± 0.0004	0.0265 ± 0.0003
	SNIPS	0.0174 ± 0.0003	0.0234 ± 0.0004	0.0258 ± 0.0004	0.0102 ± 0.0003	0.0266 ± 0.0002
	DR	0.0123 ± 0.0003	0.0145 ± 0.0004	0.0122 ± 0.0004	0.0080 ± 0.0003	0.0179 ± 0.0002
	SDR	**0.0095 ± 0.0003***	**0.0111 ± 0.0004***	**0.0084 ± 0.0005***	**0.0070 ± 0.0003***	**0.0175 ± 0.0002***
ML-100K-2	Naive	0.0695 ± 0.0020	0.0932 ± 0.0024	0.1036 ± 0.0021	0.0394 ± 0.0020	0.1080 ± 0.0024
	IPS	0.0344 ± 0.0022	0.0465 ± 0.0027	0.0516 ± 0.0023	0.0196 ± 0.0020	0.0540 ± 0.0023
	SNIPS	0.0344 ± 0.0024	0.0465 ± 0.0027	0.0517 ± 0.0024	0.0197 ± 0.0020	0.0541 ± 0.0024
	DR	0.0286 ± 0.0026	0.0346 ± 0.0027	0.0304 ± 0.0026	0.0172 ± 0.0021	0.0355 ± 0.0024
	SDR	**0.0167 ± 0.0023***	**0.0202 ± 0.0025***	**0.0146 ± 0.0023***	**0.0128 ± 0.0019***	**0.0338 ± 0.0024***
ML-1M-2	Naive	0.0699 ± 0.0006	0.0931 ± 0.0006	0.1038 ± 0.0006	0.0397 ± 0.0005	0.1077 ± 0.0005
	IPS	0.0349 ± 0.0007	0.0466 ± 0.0007	0.0519 ± 0.0007	0.0199 ± 0.0006	0.0539 ± 0.0006
	SNIPS	0.0349 ± 0.0006	0.0466 ± 0.0006	0.0519 ± 0.0006	0.0199 ± 0.0006	0.0538 ± 0.0005
	DR	0.0289 ± 0.0006	0.0346 ± 0.0007	0.0305 ± 0.0007	0.0174 ± 0.0005	0.0352 ± 0.0005
	SDR	**0.0170 ± 0.0006***	**0.0203 ± 0.0007***	**0.0147 ± 0.0007***	**0.0130 ± 0.0005***	**0.0335 ± 0.0005***
ML-100K-3	Naive	0.1057 ± 0.0022	0.1377 ± 0.0021	0.1555 ± 0.0021	0.0567 ± 0.0024	0.1600 ± 0.0020
	IPS	0.0533 ± 0.0032	0.0691 ± 0.0028	0.0781 ± 0.0028	0.0287 ± 0.0030	0.0802 ± 0.0021
	SNIPS	0.0530 ± 0.0031	0.0689 ± 0.0027	0.0779 ± 0.0027	0.0285 ± 0.0025	0.0800 ± 0.0023
	DR	0.0523 ± 0.0032	0.0616 ± 0.0030	0.0574 ± 0.0031	0.0289 ± 0.0023	0.0515 ± 0.0023
	SDR	**0.0230 ± 0.0029***	**0.0269 ± 0.0026***	**0.0190 ± 0.0027***	**0.0180 ± 0.0023***	**0.0469 ± 0.0022***
ML-1M-3	Naive	0.1046 ± 0.0004	0.1373 ± 0.0005	0.1549 ± 0.0004	0.0568 ± 0.0005	0.1601 ± 0.0004
	IPS	0.0521 ± 0.0007	0.0686 ± 0.0007	0.0774 ± 0.0006	0.0283 ± 0.0007	0.0800 ± 0.0005
	SNIPS	0.0522 ± 0.0005	0.0686 ± 0.0005	0.0774 ± 0.0005	0.0283 ± 0.0006	0.0801 ± 0.0004
	DR	0.0513 ± 0.0006	0.0613 ± 0.0006	0.0568 ± 0.0006	0.0285 ± 0.0006	0.0515 ± 0.0004
	SDR	**0.0218 ± 0.0006***	**0.0263 ± 0.0006***	**0.0181 ± 0.0006***	**0.0175 ± 0.0006***	**0.0468 ± 0.0004***
ML-100K-4	Naive	0.1366 ± 0.0024	0.1760 ± 0.0028	0.2023 ± 0.0022	0.0698 ± 0.0022	0.2061 ± 0.0030
	IPS	0.0673 ± 0.0043	0.0866 ± 0.0045	0.0998 ± 0.0039	0.0334 ± 0.0031	0.1018 ± 0.0032
	SNIPS	0.0680 ± 0.0043	0.0873 ± 0.0043	0.1005 ± 0.0037	0.0341 ± 0.0030	0.1025 ± 0.0034
	DR	0.0788 ± 0.0048	0.0929 ± 0.0049	0.0903 ± 0.0043	0.0407 ± 0.0034	0.0650 ± 0.0035
	SDR	**0.0208 ± 0.0041***	**0.0247 ± 0.0041***	**0.0151 ± 0.0035***	**0.0187 ± 0.0031***	**0.0548 ± 0.0034***
ML-1M-4	Naive	0.1369 ± 0.0007	0.1767 ± 0.0008	0.2020 ± 0.0008	0.0701 ± 0.0007	0.2070 ± 0.0007
	IPS	0.0686 ± 0.0009	0.0886 ± 0.0009	0.1012 ± 0.0008	0.0350 ± 0.0008	0.1037 ± 0.0005
	SNIPS	0.0686 ± 0.0010	0.0886 ± 0.0011	0.1012 ± 0.0010	0.0350 ± 0.0008	0.1037 ± 0.0007
	DR	0.0789 ± 0.0013	0.0936 ± 0.0014	0.0904 ± 0.0014	0.0411 ± 0.0010	0.0662 ± 0.0007
	SDR	**0.0215 ± 0.0011***	**0.0259 ± 0.0011***	**0.0157 ± 0.0010***	**0.0193 ± 0.0009***	**0.0561 ± 0.0007***

Note: * means statistically significant results (p-value ≤ 0.001) using the paired-t-test compared with the best baseline.