CBPL: A Unified Calibration and Balancing Propensity Learning Framework in Causal Recommendation for Debiasing

Shufeng Zhang[1] , **Tianyu Xia**[2*]
[1]University of North Carolina at Chapel Hill, North Carolina, USA
[2]Peking University, Beijing, China
shufengz@alumni.unc.edu, 2311110185@bjmu.edu.cn

Abstract

In recommender systems, observed data always suffer from Missing-Not-At-Random (MNAR) issue: users rate only a non-random subset of items, leading to a biased recommendation if the model is trained on such biased data directly. One type of popular debiasing method is to learn an accurate propensity score (the probability a rating is observed) and then reweight the observed sample to achieve unbiased rating prediction. While calibration metric and balancing metric are widely adopted as constraints to learn a high-quality propensity model, existing methods optimize these objectives in an isolated manner, neglecting their inherent connections. To bridge this gap, we first decompose the balancing constraint, making the balancing loss and the calibration loss have a similar form. Then we propose a unified Calibration and Balancing Propensity Learning (CBPL) framework that minimizes calibration loss and balancing loss simultaneously. In addition, we provide a theoretical analysis showing that our method has a variance reduction property. Experimental results on three real-world recommendation datasets demonstrate that our method can outperform the state-of-the-art baselines.

1 Introduction

In real-world applications, recommender systems use user feedback (such as ratings) to infer user preferences. However, since users tend to rate items they are interested in, the observed data is often affected by selection bias, exhibiting a Missing-Not-at-Random (MNAR) pattern [Chen *et al.*, 2020; Li *et al.*, 2023f; Yang *et al.*, 2018], learning to sub-optimal rating prediction. Thus, addressing selection bias has become a central challenge for training a high-quality rating prediction model [Steck, 2010; Li *et al.*, 2023d].

To address the MNAR problem, many propensity-based methods are proposed. Specifically, inverse propensity score (IPS) based methods use propensity (the probability a rating is observed) to reweight the observed samples [Swaminathan and Joachims, 2015a; Swaminathan and Joachims, 2015a]. Furthermore, doubly robust (DR) based estimators use both the imputation model (impute missing rating) and the propensity model together to further debias [Wang *et al.*, 2019; Li *et al.*, 2023d]. Despite significant progress in existing propensity-based methods, several challenges in propensity estimation remain. (1) Existing models will always overestimate or underestimate propensity scores, resulting in propensity scores that are unrealistically low or high; (2) Many approaches overlook the propensity balancing property, i.e., the covariate balance between treated and untreated groups.

- Overconfident estimates, where predicted propensity scores are overly close to 0 or 1, are a common issue in existing propensity models [Guo *et al.*, 2017; Bai *et al.*, 2021; Kweon *et al.*, 2021]. Such overconfidence leads to unreliable probability estimates and motivates the need for proper calibration, which evaluates how well predicted probabilities align with actual outcomes [Kull *et al.*, 2017; Deshpande and Kuleshov, 2023]. Formally, for a propensity model $h(\cdot)$ and an observation indicator $o \in \{0, 1\}$, calibration is satisfied if

$$\mathbb{E}(o \mid h(x) = \hat{p}) = \hat{p}.$$

  For example, if there are 100 samples with estimated propensity 0.2, then we expect exactly 20 observations.

- Even when propensity predictions are accurate on average, poor covariate balance can lead to biased or high-variance estimates the doubly robust methods [Wang *et al.*, 2019; Guo *et al.*, 2021; Chen *et al.*, 2021; Dai *et al.*, 2022]. Formally, for any measurable and integrable function $\phi : \mathcal{X} \to \mathbb{R}^m$, propensity balance is satisfied if

$$\mathbb{E}\left[\frac{o}{h(x)} \cdot \phi(x)\right] = \mathbb{E}\left[\frac{1-o}{1-h(x)} \cdot \phi(x)\right] = \mathbb{E}\left[\phi(x)\right].$$

In practice, based on Expected Calibration Error (ECE), prior works [Kweon *et al.*, 2024; Hu *et al.*, 2025] have been proposed to learn a well-calibrated propensity. In addition, to promote balance, [Li *et al.*, 2023d] propose to use the Balancing Mean Squared Error (BMSE) metric as a regularizer to penalize misaligned propensities during training. However, calibration and balancing are always difficult to satisfy simultaneously in practice. As shown in Table 1, we present

*Corresponding author.

Toy Example 1		Toy Example 2	
Estimated propensity	Observed Indicator	Estimated propensity	Observed Indicator
0.77	1	0.85	1
0.76	1	0.75	0
0.75	1	0.75	1
0.74	1	0.65	0
0.71	0	0.55	1
0.57	1	0.45	0
0.50	1	0.35	0
0.26	0	0.25	1
0.25	0	0.25	1
0.17	1	0.15	0
ECE $\approx$ 0.1520		ECE $\approx$ 0.2700	
BMSE $\approx$ 0.7749		BMSE $\approx$ 0.0088	

Table 1: Toy examples demonstrating inconsistency between balance (BMSE) and calibration (ECE) metrics. The detailed calculation and definition are shown in Section 2.3.

two illustrative examples on a toy dataset: in Example 1, the propensity estimates exhibit good calibration but poor covariate balance (low ECE but a high BMSE). In contrast, Example 2 achieves better balance but suffers from poor calibration. This inconsistency highlights a fundamental challenge: how to ensure both calibration and balancing?

In this paper, we bridge the gap between calibration and balancing in propensity estimation by proposing a Calibration Balancing Propensity Learning (CBPL) framework. Specifically, we first decompose the balancing constraint, making the balancing loss and the calibration loss have a similar form, then we adopt them as constraints in the proposed unified framework. Our contributions are summarized as follows:

- We adjust miscalibration metric and decompose the balancing constraint to better align the balancing loss and the calibration loss, revealing their inherent connection and enabling a unified optimization strategy.
- We develop the CBPL framework that simultaneously minimizes both calibration and balancing losses, ensuring that the propensity scores satisfy these essential conditions jointly rather than separately.
- We provide a theoretical analysis demonstrating the variance reduction property of our unified approach
- We conduct extensive empirical evaluations on three real-world recommendation datasets, showing the proposed method can outperform the baselines.

2 Preliminary

2.1 Problem Setup

We adopt the potential outcome framework to formulate the selection bias problem in recommender systems formally. Let $\mathcal{U} = \{u_1, \ldots, u_n\}$ and $\mathcal{I} = \{i_1, \ldots, i_m\}$ denote the sets of users and items, respectively. The complete user-item interaction space is represented by $\mathcal{D} = \mathcal{U} \times \mathcal{I}$, covering all possible interactions between user-item pairs. For each user-item pair $(u, i) \in \mathcal{D}$, let $x_{u,i} \in \mathbb{R}^d$ represent observed covariates, such as user demographics or item attributes. Define the treatment indicator $o_{u,i} \in \{0, 1\}$, which equals 1 if user u rates item i, and 0 otherwise. Let $r_{u,i}(o)$ be the potential outcome for treatment $o \in \{0, 1\}$, where the observed rating is $r_{u,i}(1)$ when $o_{u,i} = 1$, and the unobserved rating (counterfactual rating if rated) is $r_{u,i}(0)$.

Our objective is to estimate ratings for all user-item pairs. Ideally, with full observability, the prediction model $f(x_{u,i}; \theta)$ can be trained by minimizing the ideal loss:

$$\mathcal{L}_{\text{ideal}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} e_{u,i}, \tag{1}$$

where $e_{u,i} = \ell_{\text{CE}}(f(x_{u,i}; \theta), r_{u,i}(1))$ represents the prediction error, typically mean squared error (MSE) or cross entropy loss.

In practice, ratings are partially observed with non-random missingness. Let $\mathcal{O} = \{(u, i) \in \mathcal{D} \mid o_{u,i} = 1\}$ denotes the observed samples, minimizing the loss only within the observed data to train the prediction model will result in suboptimal prediciton, due to this loss is not a unbisased estimation of the ideal loss. To address this bias, we introduce the propensity score, defined as the probability of rating given covariates:

$$p_{u,i} = \mathbb{P}(o_{u,i} = 1 \mid x_{u,i}), \tag{2}$$

and is estimated by a propensity model, $h_\psi(x)$.

2.2 Propensity-Based Estimators

To obtain unbiased estimates of evaluation metrics from biased observations, we rely on propensity-weighting methods. Two widely used estimators are the *Inverse Propensity Score (IPS)* and the *Doubly Robust (DR)* approaches.

Inverse Propensity Scoring (IPS) directly reweights observed ratings:

$$\mathcal{L}_{\text{IPS}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i} e_{u,i}}{\hat{p}_{u,i}}, \tag{3}$$

which is an unbiased estimation of ideal loss when $\hat{p}_{u,i} = p_{u,i}$ for all user-item pairs. The Doubly Robust (DR) estimator integrates imputation and propensity weighting:

$$\mathcal{L}_{\text{DR}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \left[\hat{e}_{u,i} + \frac{o_{u,i}(e_{u,i} - \hat{e}_{u,i})}{\hat{p}_{u,i}} \right], \tag{4}$$

where $\hat{e}_{u,i} = g_\phi(x_{u,i})$ is the imputation of the prediction error. DR-based estimators are unbiased when $\hat{p}_{u,i} = p_{u,i}$ or $\hat{e}_{u,i} = e_{u,i}$ for all user-item pairs. These approaches address selection bias but rely heavily on accurate estimation of propensity scores and will suffer from high bias and variance when learned propensity scores are inaccurate, motivating us to learn robust, high-quality propensity scores.

2.3 Evaluation Metrics of Calibration & Balancing

In addition to unbiasedness, we are interested in the *calibration* of predicted propensities and the *balancing* of propensity-weighted distributions. We consider two metrics to quantify these properties: the Expected Calibration Error (ECE) and the Balanced Mean Squared Error (BMSE).

The **Expected Calibration Error (ECE)** measures how well the predicted probabilities $\hat{p}_{u,i}$ align with actual outcome frequencies $o_{u,i}$. We partition the prediction range $[0, 1]$ into M bins $B_1, B_2, \ldots, B_M$ (e.g. equal-width or by quantiles of $\hat{p}$). In Table 1, we partition three bins based on the equal-width criterion: [0, 1/3], [1/3, 2/3], and [2/3, 1]. Let $|\mathcal{D}|$ be

the total number of evaluated instances. ECE is defined as the weighted average absolute difference between observed incidence rate and predicted probability in each bin:

$$\mathrm{ECE}_M\left(h_\psi\right) = \sum_{m=1}^{M} \frac{|B_m|}{|D|} \left| \frac{\sum_{(u,i)\in B_m} o_{u,i}}{|B_m|} - \frac{\sum_{(u,i)\in B_m} \hat{p}_{u,i}}{|B_m|} \right|. \tag{5}$$

A perfectly calibrated model would have $\mathrm{ECE} = 0$, meaning in every score bin the fraction of observed samples equals the predicted probability.

The **Balanced Mean Squared Error (BMSE)** assesses the propensity balancing property of the model's scores. It evaluates whether the distribution of features (or samples) is balanced when weighting by inverse propensity, which is crucial for debiasing. Formally, for a given vector-valued balancing function $\boldsymbol{\phi}(x_{u,i})$, BMSE is defined as:

$$\mathrm{BMSE}(\boldsymbol{\phi}, \hat{p}) = \left\| \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \left[\frac{o_{u,i}}{\hat{p}_{u,i}} - \frac{1-o_{u,i}}{1-\hat{p}_{u,i}} \right] \boldsymbol{\phi}(x_{u,i}) \right\|_F^2. \tag{6}$$

In Table 1, for illustration purposes, we choose $\phi(x) = 1$, which is a constant function.

3 Methodology

3.1 Motivation

Calibration and balancing of propensity scores are both necessary conditions for a reliable propensity score model and correct estimation, with reflection on different aspects of model quality. However, as shown in Table 1 (left), it has excellent calibration but deficient balancing. In addition, as shown in Table 1 (right), the balancing performance is excellent but with poor calibration. Previous methods either optimized the model calibration or the balancing, but not both. Thus, a straightforward question is whether there is a way to achieve calibration and balancing simultaneously?

We find that calibration and balancing are closely related to each other. Specifically, BMSE can be expressed as:

$$\mathrm{BMSE}(\phi, p) = \left\| \sum_{u,i} \frac{o_{u,i} - p_{u,i}}{p_{u,i}(1-p_{u,i})} \phi(x_{u,i}) \right\|^2 \tag{7}$$

highlighting its interpretation as a weighted global calibration measure (if treat $\phi(x)/p_{u,i}(1-p_{u,i})$ as weights). Thus, we first propose to decompose the BMSE to ensure the adaptation with bin partition, then we regard the ECE loss and decomposed BMSE loss as constraints to ensure balancing and calibration.

3.2 Bin-Based Evaluation Metrics for Calibration and Balancing

In this section, we introduce bin-based metrics that jointly evaluate these aspects. In particular, we define the squared expected calibration error (SECE) (with its finite-sample approximation SECE_M) to quantify calibration, which is defined below:

$$\mathrm{SECE}_\infty = \mathbb{E}\left[\left(h(X) - E[O \mid h(X)] \right)^2 \right]. \tag{8}$$

Intuitively, $\mathrm{SECE}_\infty = 0$ if and only if $h(X) = E[O \mid h(X)]$ almost surely, i.e. the calibration condition. In practice, we use the differentiable version with M bins as shown below.

$$\mathrm{SECE}_M = \sum_{m=1}^{M} \frac{|B_m|}{|D|} \left(\bar{\hat{p}}_m - \bar{o}_m \right)^2. \tag{9}$$

The choice of binning affects the accuracy of SECE_M as an estimator of true calibration error. We assume a **proper binning scheme** that satisfies the following conditions as the sample size N grows:

- **Sufficient samples per bin**: Each bin contains a large number of samples, such that no bin remains sparsely populated in the limit.
- **Refinement/consistency**: The bin partition becomes finer (more granular) as N increases, so that the range of predicted probabilities within any single bin shrinks. Ideally, within each bin the model's prediction is nearly constant.
- **Coverage of support**: The bins together cover the range of predicted probabilities where the model places mass. This ensures any systematic calibration error at a given predicted value will eventually be detected by some bin.

Lemma 3.1 (Consistency of $SECE_M$)**.** *Under proper binning, the finite-sample bin-based $SECE_M \to SECE_\infty$, i.e. converges to the true calibration error in probability, as $N, M \to \infty$.*

Lemma 3.2 (Calibration Property)**.** *If the propensity estimator is well specified, that is, $\hat{p}(x) = p(x), \forall x$, then for any fixed binning strategy, $SECE_M \to 0$ almost surely as $N, M \to \infty$.*

The corresponding proofs are shown in the Appendix.

In SECE, we calibrate the propensity model in a bin-based manner, dividing the predicted propensity scores into bins to assess local calibration. To achieve a unified framework, we also aim to ensure propensity balance under the same bin-based architecture. However, the standard BMSE formulation does not naturally support bin-wise computation, which makes it difficult to assess or enforce balance at a more granular level.

Therefore, we first design a within-bin BMSE error, ensuring that within each bin, the propensity scores are not only well-calibrated but also balanced. Furthermore, because true balance should hold globally across the entire distribution, it is necessary to ensure that propensity balance is achieved both within each bin and between different bins. We constructed a between-bin BMSE error, reflecting systematic discrepancies across bins. This dual-level balance guarantees that the model achieves not only local (per-bin) fairness, but also global (distribution-wide) balance.

We choose a vector-valued feature map $\phi(x_{u,i})$, (like a vector of user and item embeddings or other sufficient statistics) and define the imbalance vector for bin B_m as the average weighted difference between exposed and non-exposed

feature sums in that bin:

$$\Delta_m = \frac{1}{|B_m|} \sum_{(u,i)\in B_m} \left(\frac{o_{u,i}}{\hat{p}_{u,i}} - \frac{1-o_{u,i}}{1-\hat{p}_{u,i}} \right) \phi(x_{u,i}) \in \mathbb{R}^d, \tag{10}$$

where $\phi(x_{u,i}) \in \mathbb{R}^d$. Let $||\cdot||_F$ denotes the Frobenius norm. By construction, $||\Delta_m||_F^2$ is the within-bin covariate discrepancy between the exposed and non-exposed subsamples. We now decompose this error in terms of within-bin and between-bin contributions:

- **Within-bin BMSE** is defined as

$$\text{BMSE}_{\text{within}} = \sum_{m=1}^{M} \frac{|B_m|^2}{|D|^2} \|\Delta_m\|_F^2 , \tag{11}$$

  which sums the squared norm of the imbalance vector in each bin. It is large if, within any propensity bin, the model fails to balance the covariate distribution between exposed and unexposed samples. Minimizing $\text{BMSE}_{\text{within}}$ directly promotes feature balance in each stratum of $\hat{p}$.

- While $\text{BMSE}_{\text{within}}$ considers each bin in isolation, there could remain an overall imbalance aggregated between bins (e.g. if all Δ_m point in a similar direction). We define **Between-bin BMSE** to include the interaction (covariance) of imbalance vectors between bins:

$$\text{BMSE}_{\text{between}} = \sum_{m\neq m'} \frac{|B_m|\,|B_{m'}|}{|D|^2} \langle \Delta_m, \Delta_{m'} \rangle , \tag{12}$$

  which captures cross-bin discrepancies.

To summarize, we have the following lemmas to show the BMSE decomposition results and the convergence property (proofs are in the Appendix).

Lemma 3.3 (BMSE Decomposition)**.** *The total balancing error can be decomposed into within-bin and between-bin components,*

$$\begin{aligned} \textit{BMSE}_{\textit{overall}} &= \left\| \frac{1}{|D|} \sum_{(u,i)\in D} \left(\frac{o_{u,i}}{\hat{p}_{u,i}} - \frac{1-o_{u,i}}{1-\hat{p}_{u,i}} \right) \phi(x_{u,i}) \right\|_F^2 \\ &= \sum_{m=1}^{M} \frac{|B_m|^2}{|D|^2} \|\Delta_m\|_F^2 + \\ &\quad \sum_{m\neq m'} \frac{|B_m|\,|B_{m'}|}{|D|^2} \langle \Delta_m, \Delta_{m'} \rangle , \end{aligned}$$

Lemma 3.4 (Balancing Property [Li *et al.*, 2023d])**.** *If propensity model is correctly specified, i.e $\hat{p}_{u,i} = p_{u,i}$ for all instances, then for any integrable vector-valued functions $\phi(x)$, $\textit{BMSE}_{\textit{overall}}(\phi, \hat{p}) \to 0$ almost surely.*

3.3 Unified Loss of Calibration and Balancing

Having defined metrics for calibration and balance, we now integrate them into a single learning objective. Our goal is to learn a propensity model that not only fits the observed exposure data (via a standard likelihood loss) but also produces predictions that are well-calibrated and balanced with respect to the true propensities. We achieve this by combining the conventional cross-entropy loss with regularization terms based on the SECE_M and BMSE metrics above.

Specifically, let $L_{\text{CE}}(\psi)$ be the binary cross-entropy (log-loss) for predicting the exposure $o_{u,i}$:

$$L_{\text{CE}} = -\frac{1}{|D|} \sum_{(u,i)\in D} \left[o_{u,i} \log \hat{p}_{u,i} + (1-o_{u,i}) \log(1-\hat{p}_{u,i}) \right]. \tag{13}$$

Overall, the proposed unified loss of Calibration and Balancing Propensity Learning framework (CBPL) combines three components: (i) standard cross-entropy fit of the propensity model, (ii) a miscalibration penalty, and (iii) an imbalancing penalty measured. The resulting Calibration and Balancing Propensity Learning (CBPL) loss is formulated as:

$$\begin{aligned} L_{\text{CBPL}} &= L_{\text{CE}} + \lambda_1 \text{SECE}_M \\ &\quad + \lambda_2 (\text{BMSE}_{\text{within}} + \text{BMSE}_{\text{between}}), \end{aligned} \tag{14}$$

where $\lambda_1, \lambda_2 > 0$ are non negative hyperparameters controlling the trade-off between prediction fit, calibration, and balance regularizers.

We now provide theoretical guarantees for the proposed CBPL approach. In particular, we show that optimizing L_{CBPL} leads to an unbiased propensity estimator in the limit of infinite data, and that it achieves variance reduction in downstream estimates compared to a standard propensity model. These properties formalize the intuitive benefits of jointly addressing calibration and balance. All proofs are deferred to the appendix.

Theorem 3.5 (Unbiasedness of CBPL Loss)**.** *When learned propensities are accurate, the unified loss constructed by CBPL is unbiased and $L_{\textit{CBPL}} \to 0$ almost surely.*

Algorithm 1 Joint Training Algorithm

Input:
  Set of user-item features X;
  Observation indicator matrix $\mathcal{O}$;
  All user-item pairs matrix $\mathcal{D}$;
  Observed outcomes for observed samples $\mathbf{R}^o$;
  Hyperparameters λ_1, λ_2;
while not converge **do**
  for number of training iterations **do**
    Sample (u,i) pairs $\{(u_k, i_k)\}_{k=1}^{K}$ from $\mathcal{D}$;
    update the h_ψ based on L_{CBPL};
  end for
end while
while not converge **do**
  for number of training iterations **do**
    Sample (u,i) pairs $\{(u_j, i_j)\}_{j=1}^{J}$ from $\mathcal{D}$;
    Update f_θ using DR loss (Equation 4);
    Sample (u,i) pairs $\{(u_m, i_m)\}_{b=1}^{B}$ from $\mathcal{O}$;
    Update g_ϕ using imputation loss (Equation 18);
  end for
end while

Table 2: Performance on AUC, NDCG@T, and F1@T on **Coat**, **Yahoo! R3** and **KuaiRec**. The best and the second best results are bolded and underlined.

	Coat			Yahoo! R3			KuaiRec		
Method	AUC	NDCG@5	F1@5	AUC	NDCG@5	F1@5	AUC	NDCG@20	F1@20
Naive	$0.703_{\pm 0.006}$	$0.605_{\pm 0.012}$	$0.467_{\pm 0.007}$	$0.673_{\pm 0.001}$	$0.635_{\pm 0.002}$	$0.306_{\pm 0.002}$	$0.753_{\pm 0.001}$	$0.449_{\pm 0.002}$	$0.124_{\pm 0.002}$
IPS	$0.717_{\pm 0.007}$	$0.617_{\pm 0.009}$	$0.473_{\pm 0.008}$	$0.678_{\pm 0.001}$	$0.638_{\pm 0.002}$	$0.318_{\pm 0.002}$	$0.755_{\pm 0.004}$	$0.452_{\pm 0.010}$	$0.131_{\pm 0.004}$
SNIPS	$0.714_{\pm 0.012}$	$0.614_{\pm 0.012}$	$0.474_{\pm 0.009}$	$0.683_{\pm 0.002}$	$0.639_{\pm 0.002}$	$0.316_{\pm 0.002}$	$0.754_{\pm 0.003}$	$0.453_{\pm 0.004}$	$0.126_{\pm 0.003}$
ASIPS	$0.719_{\pm 0.009}$	$0.618_{\pm 0.012}$	$0.476_{\pm 0.009}$	$0.679_{\pm 0.003}$	$0.640_{\pm 0.003}$	$0.319_{\pm 0.003}$	$0.757_{\pm 0.005}$	$0.474_{\pm 0.007}$	$0.130_{\pm 0.005}$
IPS-V2	$0.726_{\pm 0.005}$	$0.627_{\pm 0.009}$	$0.479_{\pm 0.008}$	$0.685_{\pm 0.002}$	$0.646_{\pm 0.003}$	$0.320_{\pm 0.002}$	$0.764_{\pm 0.001}$	$0.476_{\pm 0.003}$	$0.135_{\pm 0.003}$
KBIPS	$0.714_{\pm 0.003}$	$0.618_{\pm 0.010}$	$0.474_{\pm 0.007}$	$0.676_{\pm 0.002}$	$0.642_{\pm 0.003}$	$0.318_{\pm 0.002}$	$0.763_{\pm 0.001}$	$0.463_{\pm 0.007}$	$0.134_{\pm 0.002}$
AKBIPS	$0.732_{\pm 0.004}$	$0.636_{\pm 0.006}$	$0.483_{\pm 0.006}$	$0.689_{\pm 0.001}$	$0.658_{\pm 0.002}$	$0.324_{\pm 0.002}$	$0.766_{\pm 0.003}$	$0.478_{\pm 0.009}$	$0.138_{\pm 0.003}$
DR	$0.718_{\pm 0.008}$	$0.623_{\pm 0.009}$	$0.474_{\pm 0.007}$	$0.684_{\pm 0.002}$	$0.658_{\pm 0.003}$	$0.326_{\pm 0.002}$	$0.755_{\pm 0.008}$	$0.462_{\pm 0.010}$	$0.135_{\pm 0.005}$
DR-JL	$0.723_{\pm 0.005}$	$0.629_{\pm 0.007}$	$0.479_{\pm 0.005}$	$0.685_{\pm 0.002}$	$0.653_{\pm 0.002}$	$0.324_{\pm 0.002}$	$0.766_{\pm 0.002}$	$0.467_{\pm 0.005}$	$0.136_{\pm 0.003}$
MRDR-JL	$0.727_{\pm 0.005}$	$0.627_{\pm 0.008}$	$0.480_{\pm 0.008}$	$0.684_{\pm 0.002}$	$0.652_{\pm 0.003}$	$0.325_{\pm 0.002}$	$0.768_{\pm 0.005}$	$0.473_{\pm 0.007}$	$0.139_{\pm 0.004}$
DR-BIAS	$0.726_{\pm 0.004}$	$0.629_{\pm 0.009}$	$0.482_{\pm 0.007}$	$0.685_{\pm 0.002}$	$0.653_{\pm 0.002}$	$0.325_{\pm 0.003}$	$0.768_{\pm 0.003}$	$0.477_{\pm 0.006}$	$0.137_{\pm 0.004}$
DR-MSE	$0.727_{\pm 0.007}$	$0.631_{\pm 0.008}$	$0.484_{\pm 0.007}$	$0.687_{\pm 0.002}$	$0.657_{\pm 0.003}$	$0.327_{\pm 0.003}$	$0.770_{\pm 0.003}$	$0.480_{\pm 0.006}$	$0.140_{\pm 0.003}$
MR	$0.724_{\pm 0.004}$	$0.636_{\pm 0.006}$	$0.481_{\pm 0.006}$	$0.691_{\pm 0.002}$	$0.647_{\pm 0.002}$	$0.316_{\pm 0.003}$	$0.776_{\pm 0.005}$	$0.483_{\pm 0.006}$	$0.142_{\pm 0.003}$
TDR	$0.714_{\pm 0.006}$	$0.634_{\pm 0.011}$	$0.483_{\pm 0.008}$	$0.688_{\pm 0.003}$	$0.662_{\pm 0.002}$	$0.329_{\pm 0.002}$	$0.772_{\pm 0.003}$	$0.486_{\pm 0.005}$	$0.140_{\pm 0.003}$
TDR-JL	$0.731_{\pm 0.005}$	$0.639_{\pm 0.007}$	$0.484_{\pm 0.007}$	$0.689_{\pm 0.002}$	$0.656_{\pm 0.004}$	$0.327_{\pm 0.003}$	$0.772_{\pm 0.003}$	$0.489_{\pm 0.005}$	$0.142_{\pm 0.003}$
StableDR	$0.735_{\pm 0.005}$	$0.640_{\pm 0.007}$	$0.484_{\pm 0.006}$	$0.688_{\pm 0.002}$	$0.661_{\pm 0.003}$	$0.329_{\pm 0.002}$	$0.773_{\pm 0.001}$	$0.491_{\pm 0.003}$	$0.143_{\pm 0.003}$
DR-V2	$0.734_{\pm 0.007}$	$0.639_{\pm 0.009}$	$0.487_{\pm 0.006}$	$0.690_{\pm 0.002}$	$0.660_{\pm 0.005}$	$0.328_{\pm 0.002}$	$0.773_{\pm 0.003}$	$0.488_{\pm 0.006}$	$0.142_{\pm 0.004}$
KBDR	$0.730_{\pm 0.003}$	$0.631_{\pm 0.005}$	$0.482_{\pm 0.006}$	$0.682_{\pm 0.002}$	$0.648_{\pm 0.003}$	$0.323_{\pm 0.002}$	$0.765_{\pm 0.004}$	$0.460_{\pm 0.006}$	$0.138_{\pm 0.003}$
AKBDR	$\underline{0.745}_{\pm 0.004}$	$0.645_{\pm 0.008}$	$0.493_{\pm 0.007}$	$0.692_{\pm 0.002}$	$0.661_{\pm 0.002}$	$0.328_{\pm 0.002}$	$0.782_{\pm 0.003}$	$0.498_{\pm 0.008}$	$0.147_{\pm 0.003}$
DCE-DR	$0.736_{\pm 0.006}$	$0.648_{\pm 0.007}$	$0.489_{\pm 0.005}$	$0.698_{\pm 0.002}$	$0.670_{\pm 0.002}$	$0.333_{\pm 0.003}$	$0.795_{\pm 0.004}$	$0.512_{\pm 0.005}$	$0.153_{\pm 0.002}$
DCE-TDR	$0.740_{\pm 0.004}$	$0.651_{\pm 0.006}$	$0.489_{\pm 0.007}$	$0.701_{\pm 0.002}$	$0.672_{\pm 0.002}$	$0.331_{\pm 0.002}$	$0.798_{\pm 0.005}$	$0.514_{\pm 0.006}$	$0.155_{\pm 0.002}$
Cali-MR	$0.741_{\pm 0.002}$	$\underline{0.658}_{\pm 0.004}$	$\underline{0.495}_{\pm 0.004}$	$\underline{0.703}_{\pm 0.002}$	$\underline{0.678}_{\pm 0.002}$	$\mathbf{0.338}_{\pm 0.004}$	$\underline{0.798}_{\pm 0.003}$	$\underline{0.521}_{\pm 0.005}$	$\underline{0.158}_{\pm 0.002}$
CBPL-DR	$\mathbf{0.747}_{\pm 0.003}$	$\mathbf{0.672}_{\pm 0.007}$	$\mathbf{0.500}_{\pm 0.006}$	$\mathbf{0.705}_{\pm 0.002}$	$\mathbf{0.682}_{\pm 0.003}$	$\mathbf{0.338}_{\pm 0.003}$	$\mathbf{0.804}_{\pm 0.004}$	$\mathbf{0.529}_{\pm 0.004}$	$\mathbf{0.160}_{\pm 0.003}$

Intuitively, due to both SECE_M and BMSE are zero when there is no difference between predicted and actual propensities, adding calibration and balance penalties does not introduce additional bias. In addition, it will reduce the variance, as shown in the following theorem.

Theorem 3.6 (Variance Reduction of CBPL Loss)**.** *Given learned propensities, the variance of* $\mathbb{V}(L_{CBPL} \mid \boldsymbol{o})$ *is minimized at the optimal point,*

$$\lambda_1^* = \frac{C \cdot E - B \cdot D}{A \cdot B - C^2} \tag{15}$$

$$\lambda_2^* = \frac{C \cdot D - A \cdot E}{A \cdot B - C^2}, \tag{16}$$

where $A = \mathbb{V}[\text{BMSE}], B = \mathbb{V}[\text{SECE}], C = \text{Cov}(\text{BMSE}, \text{SECE}), D = \text{Cov}(\text{CE}, \text{BMSE}), E = \text{Cov}(\text{CE}, \text{SECE})$. *And* $\boldsymbol{o} = \{o_{u,i} \mid (u,i) \in \mathcal{D}\}$. *The smallest variance is*

$$\begin{aligned} &\mathbb{V}(L_{CB\_PL} \mid \boldsymbol{o}) \mid_{\lambda_1=\lambda_1^*, \lambda_2=\lambda_1^*} \\ &= \mathbb{V}[CE] - \frac{BD^2 - 2CDE + AE^2}{AB - C^2} \leq \mathbb{V}[CE], \end{aligned} \tag{17}$$

which is also smaller than considering calibration or balancing only.

In summary, the unified loss L_{CBPL} produces propensity models that are both calibrated to the data and balanced in distribution, leading to unbiased and low-variance estimators for downstream usage.

3.4 Joint Training Algorithm

In order to incorporate propensity calibration and balancing capabilities into the model training process, we adopt a joint training algorithm [Wang *et al.*, 2019]. Specifically, the propensity model $\hat{p}_{u,i} = h_\psi(x_{u,i})$ is trained using L_{CBPL}, the prediction model $\hat{r}_{u,i} = f_\theta(x_{u,i})$ is trained by the DR loss (Equation 4), and the imputation model $\hat{e}_{u,i} = g_\phi(x_{u,i})$ is trained using the following loss function:

$$\mathcal{L}_e = \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i} \left(\hat{e}_{u,i} - e_{u,i}\right)^2}{\hat{p}_{u,i}}, \tag{18}$$

The whole training process is shown in Algorithm 1.

4 Experiments

4.1 Experimental Settings

Datasets

We evaluate debiasing performance on three standard benchmarks: **Coat**, **Yahoo! R3**, and **KuaiRec**, including both MNAR and unbiased (Missing-At-Random) ratings.

Baselines

We compare our method against the following baselines: **Naive** [Koren *et al.*, 2009], **IPS** [Schnabel *et al.*, 2016], **SNIPS** [Swaminathan and Joachims, 2015b], **ASIPS** [Saito, 2020a], **DR** [Saito, 2020b], **DR-JL** [Wang *et al.*, 2019], **MRDR** [Guo *et al.*, 2021], **DR-BIAS**, **DR-MSE** [Dai *et al.*, 2022], **MR** [Li *et al.*, 2023a], **TDR**, **TDR-JL** [Li *et al.*, 2023b], **StableDR** [Li *et al.*, 2023e], **IPS-V2**, **DR-V2** [Li *et al.*, 2023d], **KBIPS**, **KBDR**, **AKBIPS**, **AKBDR** [Li *et al.*, 2024b], **DCE-DR**, **DCE-TDR** [Kweon and Yu, 2024], and **Cali-MR** [Gong and Ma, 2025].

Table 3: Ablation study of CBPL on **Coat**, **Yahoo! R3** and **KuaiRec** datasets.

	Coat			**Yahoo! R3**			**KuaiRec**		
Method	AUC	NDCG@5	F1@5	AUC	NDCG@5	F1@5	AUC	NDCG@20	F1@20
CBPL-DR w/o $\mathcal{L}_{DCE}$	0.737	0.668	0.487	0.701	0.672	0.333	0.797	0.520	0.153
CBPL-DR w/o $\mathcal{L}_{BMSE}$	0.739	0.670	0.490	0.701	0.677	0.331	0.796	0.522	0.153
CBPL-DR w/o $\text{BMSE}_{between}$	0.743	0.671	0.493	0.704	0.682	0.337	0.800	0.525	0.158
CBPL-DR w/o BMSE_{within}	0.744	0.673	0.497	0.704	0.682	0.337	0.801	0.527	0.156
CBPL-DR	0.747	0.673	0.500	0.705	0.682	0.338	0.804	0.529	0.160

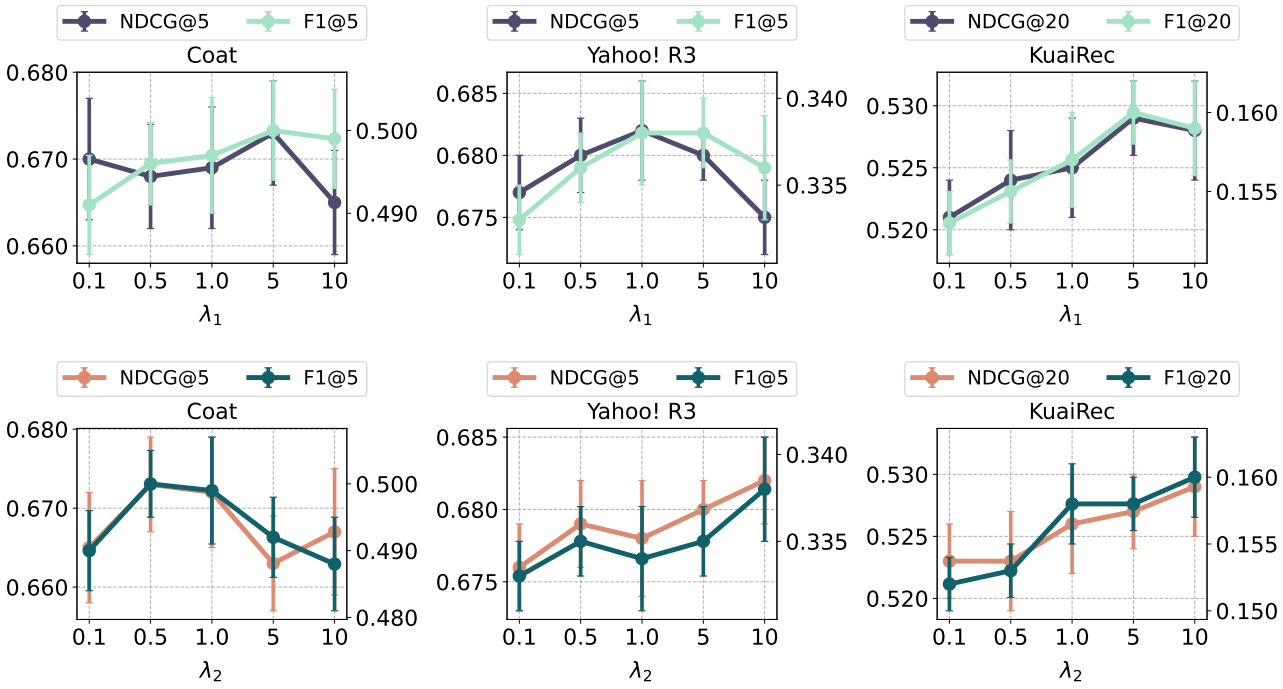

Figure 1: Parameter sensitivity analysis on the λ_1 and λ_2.

Evaluation Metrics

We evaluate prediction performance using three standard metrics: AUC, NDCG@T, and F1@T, where T is set to 5 on **Coat** and **Yahoo! R3**, and 20 on **KuaiRec**. More detailed descriptions of the experimental settings can be found in Appendix A.2.

4.2 Performance Analysis

Table 2 presents the comparative performance of various methods. As observed, all debiasing approaches, including both IPS-based and DR-based models, consistently surpass the Naive baseline, underscoring the necessity of addressing selection bias. TDR method leverages a targeted learning paradigm to improve imputation by incorporating propensity scores, effectively reducing both estimation bias and variance. Building on this, DCE-TDR introduces a calibration mechanism for the propensity model, which refines the targeted learning process and leads to further performance improvements. Meanwhile, AKBDR enhances propensity estimation by employing balancing kernels and selectively optimizing the most influential ones based on error feedback, constructing a robust propensity model that effectively mitigates bias and improves overall prediction quality.

Among the baselines, CBPL-DR achieves the best performance across all three datasets. This method simultaneously optimizes calibration and balancing objectives by reformulating their connection and introducing a unified loss. The calibrated and balanced propensity scores provide more accurate weighting for the observed samples and improve robustness across varying data distributions. As a result, CBPL-DR not only reduces estimation bias but also improves prediction variance, leading to substantial gains in multiple evaluation metrics, demonstrating the effectiveness of jointly optimizing calibration and balance for robust debiasing in MNAR.

4.3 Ablation Study

We conduct an ablation study of the proposed CBPL-DR method on three benchmark datasets, and the results are presented in Table 3. From the table, we observe that removing either the calibration loss ($\mathcal{L}$DCE) or the balancing loss ($\mathcal{L}$BMSE) leads to consistent performance degradation across all metrics and datasets, which confirms the necessity of

jointly optimizing both calibration and balancing objectives in CBPL-DR.

Specifically, when removing $\mathcal{L}_{\mathrm{DCE}}$, the performance drop is more noticeable on F1 scores, particularly on **Coat** and **KuaiRec**, indicating that miscalibrated propensities lead to reduced ranking precision. Similarly, removing the balancing term $\mathcal{L}_{\mathrm{BMSE}}$ also results in worse performance, which demonstrates the importance of ensuring covariate balance for robust estimation.

Furthermore, we decompose the $\mathcal{L}_{\mathrm{BMSE}}$ into within-bin and between-bin components to study their respective contributions. Removing either $\mathrm{BMSE}_{\mathrm{between}}$ or $\mathrm{BMSE}_{\mathrm{within}}$ causes noticeable performance declines. This suggests that both global and local balance are essential for effective debiasing. These results demonstrate the advantage of our unified CBPL-DR framework.

4.4 Parameter Sensitivity Analysis

We investigate the sensitivity of the CBPL framework to its two key regularization hyperparameters: the calibration weight λ_1 and the balancing weight λ_2. Figures 1 report the NDCG and F1 scores on **Coat**, **Yahoo! R3**, and **KuaiRec** datasets, under varying values of λ_1 and λ_2 in $\{0.1, 0.5, 1, 5, 10\}$.

As shown in the top row of Figure 1, the model achieves stable and strong performance across datasets when λ_1 falls within a moderate range. Excessively small values lead to a slight drop in performance, indicating under-regularization of miscalibration, while overly large values may over-penalize calibration error and hurt final predictions.

The bottom row of Figure 1 shows the impact of the balancing regularizer λ_2. Unlike calibration, performance consistently improves as λ_2 increases. This suggests that stronger balancing regularization is beneficial for aligning the observed and unobserved distributions, thereby reducing bias and variance in the learning process.

In summary, the CBPL-DR framework exhibits robustness to a range of regularization strengths, with best performance typically attained when λ_1 and λ_2 are both set to intermediate values. This reinforces the importance of calibration and balancing objectives for reliable propensity estimation and improved recommendation outcomes.

5 Conclusion

This paper addresses selection bias in recommender systems by unifying two key criteria, calibration and balancing, in propensity score learning. We propose the CBPL framework, which reformulates the balancing constraint to align structurally with the calibration objective, enabling joint optimization. Theoretical analysis demonstrates that CBPL consistently improves both bias and variance, and empirical results on three real-world datasets show that CBPL outperforms SOTA baselines in overall performance.

References

[Bai *et al.*, 2021] Yu Bai, Song Mei, Huan Wang, and Caiming Xiong. Don't just blame over-parametrization for over-confidence: Theoretical analysis of calibration in binary classification. *CoRR*, 2021.

[Bradley, 1997] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 1997.

[Chen *et al.*, 2020] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *CoRR*, 2020.

[Chen *et al.*, 2021] 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.

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

[Deshpande and Kuleshov, 2023] Shachi Deshpande and Volodymyr Kuleshov. Calibrated propensity scores for causal effect estimation. *CoRR*, 2023.

[Gong and Ma, 2025] Shuxia Gong and Chen Ma. Gradient-based multiple robust learning calibration on data missing-not-at-random via bi-level optimization. *Entropy*, 2025.

[Guo *et al.*, 2017] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

[Guo *et al.*, 2021] 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.

[Hu *et al.*, 2025] Wenbo Hu, Xin Sun, Qiang Liu, Le Wu, and Liang Wang. Uncertainty calibration for counterfactual propensity estimation in recommendation. *TKDE*, 2025.

[Järvelin and Kekäläinen, 2002] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *TOIS*, 2002.

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

[Kull *et al.*, 2017] Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *AISTATS*, 2017.

[Kweon and Yu, 2024] Wonbin Kweon and Hwanjo Yu. Doubly calibrated estimator for recommendation on data missing not at random. In *WWW*, 2024.

[Kweon *et al.*, 2021] Wonbin Kweon, SeongKu Kang, and Hwanjo Yu. Obtaining calibrated probabilities with personalized ranking models. *CoRR*, 2021.

[Kweon *et al.*, 2024] Wonbin Kweon, SeongKu Kang, Sanghwan Jang, and Hwanjo Yu. Top-personalized-k recommendation. In *WWW*, WWW '24, 2024.

[Li *et al.*, 2023a] Haoxuan Li, Quanyu Dai, Yuru Li, Yan Lyu, Zhenhua Dong, Xiao-Hua Zhou, and Peng Wu. Multiple robust learning for recommendation. In *AAAI*, 2023.

[Li *et al.*, 2023b] Haoxuan Li, Yan Lyu, Chunyuan Zheng, and Peng Wu. TDR-CL: Targeted doubly robust collaborative learning for debiased recommendations. In *ICLR*, 2023.

[Li *et al.*, 2023c] Haoxuan Li, Yanghao Xiao, Chunyuan Zheng, and Peng Wu. Balancing unobserved confounding with a few unbiased ratings in debiased recommendations. In *WWW*, 2023.

[Li *et al.*, 2023d] Haoxuan Li, Yanghao Xiao, Chunyuan Zheng, Peng Wu, and Peng Cui. Propensity matters: Measuring and enhancing balancing for recommendation. In *ICML*, 2023.

[Li *et al.*, 2023e] Haoxuan Li, Chunyuan Zheng, and Peng Wu. StableDR: Stabilized doubly robust learning for recommendation on data missing not at random. In *ICLR*, 2023.

[Li *et al.*, 2023f] Qian Li, Xiangmeng Wang, Zhichao Wang, and Guandong Xu. Be causal: De-biasing social network confounding in recommendation. *TKDD*, 2023.

[Li *et al.*, 2024a] Haoxuan Li, Kunhan Wu, Chunyuan Zheng, Yanghao Xiao, Hao Wang, Zhi Geng, Fuli Feng, Xiangnan He, and Peng Wu. Removing hidden confounding in recommendation: A unified multi-task learning approach. In *NeurIPS*, 2024.

[Li *et al.*, 2024b] Haoxuan Li, Chunyuan Zheng, Yanghao Xiao, Peng Wu, Zhi Geng, Xu Chen, and Peng Cui. Debiased collaborative filtering with kernel-based causal balancing. In *ICLR*, 2024.

[Lipton *et al.*, 2014] Zachary C Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. Optimal thresholding of classifiers to maximize f1 measure. In *ECML PKDD*, 2014.

[Murphy, 2012] Kevin P Murphy. *Machine Learning*. 2012.

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

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

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

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

[Swaminathan and Joachims, 2015a] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *ICML*, 2015.

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

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

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

A Appendix

A.1 Proof

Lemma 3.1 establishes that under appropriate binning (each bin having a nonzero probability mass as $N \to \infty$), the empirical squared expected calibration error SECE_M converges to the true calibration error. In summary, as the sample size grows, each bin's empirical frequency and average outcomes approach their expectations, making SECE_M a consistent estimator of the population calibration error.

Proof. Let the m-th bin be B_m (for $m = 1, \ldots, M$) and n_m be the number of samples falling into B_m. Define the empirical bin frequency $f_m = \frac{n_m}{N}$, the average predicted probability in the bin $\bar{p}_m = \frac{1}{n_m} \sum_{(u,i) \in B_m} \hat{p}(X_{u,i})$, and the average actual outcome in the bin $\bar{o}_m = \frac{1}{n_m} \sum_{(u,i) \in B_m} o_{u,i}$. Then the finite-sample squared calibration error is:

$$\mathrm{SECE}_M \;=\; \sum_{m=1}^{M} f_m \left(\bar{o}_m - \bar{p}_m\right)^2 .$$

Under the stated binning assumptions, each bin B_m with $\Pr(\hat{p}(X) \in B_m) > 0$ will contain $n_m \to \infty$ samples as $N \to \infty$. By the Law of Large Numbers (LLN), for each such bin we have:

$$f_m = \frac{n_m}{N} \xrightarrow{a.s.} \Pr(\hat{p}(X) \in B_m),$$

$$\bar{p}_m \xrightarrow{a.s.} \mathbb{E}[\hat{p}(X) \mid \hat{p}(X) \in B_m],\; \bar{o}_m \xrightarrow{a.s.} \mathbb{E}[O \mid \hat{p}(X) \in B_m]\,,$$

as $N \to \infty$. (If $\Pr(\hat{p}(X) \in B_m) = 0$ for some bin, that bin is eventually empty for large N and its contribution to SECE_M remains zero.) Therefore, almost surely as $N \to \infty$, each term $f_m(\bar{o}_m - \bar{p}_m)^2$ converges to:

$$\Pr(\hat{p}(X) \in B_m) \Big(\mathbb{E}[O \mid \hat{p}(X) \in B_m] - \mathbb{E}[\hat{p}(X) \mid \hat{p}(X) \in B_m]\Big)^2$$

which is the corresponding population-level calibration error contribution for bin m. Summing over $m = 1, \ldots, M$, we conclude that $\mathrm{SECE}_M \xrightarrow{a.s.} \sum_{m=1}^{M} \Pr(\hat{p}(X) \in B_m)\big(\mathbb{E}[O \mid \hat{p} \in B_m] - \mathbb{E}[\hat{p} \mid \hat{p} \in B_m]\big)^2$. In particular, this limit is the true (population) squared calibration error for the given proper binning scheme. Thus, SECE_M is a consistent (almost surely convergent) estimator of the population-level squared calibration error under the specified binning conditions.

Under the proper binning conditions, as N grows, each bin's statistics converge to their population values and the binned sum approaches the corresponding integral.

- Sufficient Samples per Bin: Each bin contains a large number of samples such that $|B_m| \to \infty$ for all bins (no bin remains sparsely populated in the limit). For example, one may use equal-frequency (quantile) bins or adaptive binning to ensure no bin is "too small." This condition lets us invoke the Law of Large Numbers within each bin.
- Refinement/Consistency: If the bin partition becomes finer (more granular) as N increases, so that the range of predicted probabilities within any single bin shrinks. In the limit of infinite data, bins can be taken to have infinitesimal width or even to group identical prediction values. Equivalently, for large N, $\bar{p}_m$ (average the predicted propensity score within one bin) represents a narrow band of predictions all close to some value $\hat{p}$ in $[0, 1]$. This guarantees that within each bin the model's prediction is nearly constant, and any calibration error inside the bin isn't being averaged out by wide disparities in $h(X)$.
- Coverage of Support: The bins together cover the range of predicted probabilities where the model places mass. No region of $h(X)$'s support is permanently ignored or combined in a pathological way. (Typically, one ensures the bins partition $[0, 1]$ or the range of $h(X)$.) This ensures any systematic calibration error at a given predicted value will eventually be detected by some bin.

□

Lemma 3.2 states the calibration property that if the predicted probability equals the true probability ($\hat{p}(X) = p(X)$ for all X), then the squared calibration error vanishes asymptotically. Intuitively, perfect calibration implies that within any bin, the empirical average outcome will coincide with the average predicted probability in the limit, so SECE_M converges to 0.

Proof. Assume $\hat{p}(X) = p(X)$ for every sample (almost surely). Then for each bin B_m, conditioned on the event $\{(u, i) : \hat{p}(X_{u,i}) \in B_m\}$, the outcome $O_{u,i}$ has expectation $\mathbb{E}[O_{u,i} \mid \hat{p}(X_{u,i}) \in B_m] = \mathbb{E}[p(X_{u,i}) \mid \hat{p}(X_{u,i}) \in B_m] = \mathbb{E}[\hat{p}(X_i) \mid \hat{p}(X_{u,i}) \in B_m]$. In other words, within each bin the true conditional probability of $O = 1$ equals the model's predicted probability. This implies that the difference $O_{u,i} - \hat{p}(X_{u,i})$ has expectation 0 for all (u, i). By the Law of Large Numbers, in each bin B_m with infinitely many samples we have:

$$\bar{O}_m - \bar{p}_m \;=\; \frac{1}{n_m} \sum_{(u,i):\hat{p}(X_{u,i}) \in B_m} \left(O_{u,i} - \hat{p}(X_i)\right) \xrightarrow{a.s.} 0,$$

since the summands are i.i.d. with mean 0. Therefore, $(\bar{O}_m - \bar{p}_m)^2 \to 0$ for each bin as $n_m \to \infty$. Moreover, the bin frequency $f_m = n_m/N$ approaches a finite limit (the bin's population probability) as shown in the previous proof. Thus the contribution $f_m(\bar{O}_m - \bar{p}_m)^2 \xrightarrow{a.s.} 0$ for each bin m. By summing over all bins, it follows that $\mathrm{SECE}_M = \sum_m f_m(\bar{O}_m - \bar{p}_m)^2 \xrightarrow{a.s.} 0$ as $N \to \infty$. In conclusion, when $\hat{p} = p$ the empirical squared calibration error converges to 0 (almost surely) for any fixed binning scheme, confirming the calibration property. □

For Lemma 3.3 gives a decomposition of the total BMSE (balanced MSE) into within-bin and between-bin components. We rewrite the total balanced MSE (BMSE) as the squared Frobenius norm of the average imbalance vector across all instances, then algebraically expand this expression

by grouping terms per calibration bin. This yields separate contributions from variability within each bin and interactions between different bins, corresponding to the within-bin and between-bin components of BMSE.

Proof. For each instance (u, i), define the per-instance imbalance vector as:

$$\Delta_{u,i} = \left(\frac{o_{u,i}}{\hat{p}_{u,i}} - \frac{1 - o_{u,i}}{1 - \hat{p}_{u,i}} \right) \phi(x_{u,i}) \in \mathbb{R}^d,$$

where $o_{u,i} \in 0, 1$ is the observed outcome and $\hat{p}u, i \in [0, 1]$ is the model's predicted propensity for that outcome. This vector $\Delta u, i$ represents the contribution of instance (u, i) to imbalance in the feature space $\phi(x_{u,i})$.

Now, let each calibration bin B_m collect a subset of instances. We define the bin-level imbalance vector for bin B_m as the sum of all instance vectors in that bin:

$$\Delta_m = \frac{1}{|B_m|} \sum_{(u,i) \in B_m} \left(\frac{o_{u,i}}{\hat{p}_{u,i}} - \frac{1 - o_{u,i}}{1 - \hat{p}_{u,i}} \right) \phi(x_{u,i}),$$

This Δ_m aggregates (and takes averages of) the imbalance contributions of bin B_m. By construction, the sum of all bin-level vectors recovers the total imbalance over the dataset D, i.e. $\sum_{m=1}^{M} |B_m| \Delta_m = \sum_{(u,i) \in D} \Delta_{u,i}$.

The total BMSE (overall balanced mean squared error) is defined as the squared Frobenius norm of the average imbalance vector over all instances. Using the above definitions, we can express this as:

$$\begin{aligned} \text{BMSE}_{\text{overall}} &= \left\| \frac{1}{|D|} \sum_{(u,i) \in D} \Delta_{u,i} \right\|_F^2 \\ &= \left\| \frac{1}{|D|} \sum_{m=1}^{M} |B_m| \Delta_m \right\|_F^2 \end{aligned}$$

where $|D|$ is the total number of instances and M is the number of bins. In other words, $\frac{1}{|D|} \sum_m \Delta_m$ is the overall average imbalance vector (aggregating all bins), and $\text{BMSE}_{\text{overall}}$ is the squared magnitude of this vector.

We now expand the squared norm and rearrange terms to separate within-bin and between-bin contributions. Starting from the definition above:

$$\begin{aligned} \text{BMSE}_{\text{overall}} &= \left\| \frac{1}{|D|} \sum_{m=1}^{M} |B_m| \Delta_m \right\|_F^2 \\ &= \frac{1}{|D|^2} ; \left\langle \sum_{m=1}^{M} |B_m| \Delta_m , ; \sum_{m'=1}^{M} |B_{m'}| \Delta_{m'} \right\rangle \\ &= \frac{1}{|D|^2} \sum_{m=1}^{M} \sum_{m'=1}^{M} |B_m| |B_{m'}| \langle \Delta_m , ; \Delta_{m'} \rangle \\ &= \frac{1}{|D|^2} \left(\sum_{m=1}^{M} |B_m|^2 \langle \Delta_m, \Delta_m \rangle + \right. \\ &\left. \sum_{m \neq m'}^{M} |B_m| |B_{m'}| \langle \Delta_m, \Delta_{m'} \rangle \right), . \end{aligned}$$

where $\langle \cdot \rangle$ by definition is the norm squared. In the last line, we separated the double sum into two parts: the diagonal terms where $m = m'$ and the off-diagonal terms where $m \neq m'$. We interpret these two parts as the within-bin and between-bin components, respectively.

□

Proof. For Lemma 3.4, according to the Proposition 3.1 proposed by Li *et al.* [Li *et al.*, 2023d], $\hat{p} = p$ is one of the necessary condition for $BMSE(\phi, \hat{p}) \to 0$. In addition, given Lemma statement is naturally followed by Lemma 3.3, when $\hat{p}(x) = p(x), \quad \forall x$, then the balancing property is valid across covariate distribution, and $\text{BMSE}_{\text{within}} + \text{BMSE}_{\text{between}} \to 0$ as well. □

Theorem 3.5. Recall the definition of the unified CBPL loss from Eq. (14):

$$L_{\text{CBPL}} = L_{\text{CE}} + \lambda_1 \text{SECE}_M + \lambda_2 (\text{BMSE}_{\text{within}} + \text{BMSE}_{\text{between}}).$$

We will analyze each term individually under the condition that the propensity model is correctly specified, meaning $\hat{p}_{u,i} = p_{u,i}$ almost surely.

When the predicted propensity equals the true propensity $p_{u,i}$, the cross-entropy loss is known to be minimized. Specifically, it is well-known from properties of the binary cross-entropy that, at the true probabilities, its expected value is minimal and equals the negative entropy of the true distribution (cf. standard results in statistical learning theory [Murphy, 2012]). Empirically, for a correctly specified model as sample size grows, the cross-entropy loss satisfies:

$$\begin{aligned} L_{\text{CE}} &= -\frac{1}{|D|} \sum_{(u,i) \in D} [o_{u,i} \log(p_{u,i}) + (1 - o_{u,i}) \log(1 - p_{u,i})] \\ &\xrightarrow{\text{a.s.}} \mathbb{E}[L_{\text{CE}}(p_{u,i}, o_{u,i})], \end{aligned}$$

and by definition, this expectation is minimized at the true probabilities. Thus, no additional deviation term remains, and at the true propensity, the empirical cross-entropy loss accurately reflects its theoretical minimal value.

Following from the Lemma 3.2 and 3.4, if the model is correctly specified, i.e. $\hat{p} = p, \quad \forall x$, then we have that $SECE_M \to 0$, $E[BMSE_{within} + BMSE_{between}] \to 0$ almost surely as $N \to \infty$. □

Theorem 3.6. The variance of the combined loss can be expanded to include all variance and covariance components:

$$\begin{aligned} \text{Var}(L_{\text{CB\_PL}}) &= \text{Var}(L_{CE}) + \lambda_1^2 \text{Var}(BMSE) \\ &+ \lambda_2^2 \text{Var}(SECE) + 2\lambda_1 \text{Cov}(CE, BMSE) \\ &+ 2\lambda_2 \text{Cov}(CE, SECE) \end{aligned}$$

To find the weight values that minimize $\text{Var}(L_{\text{CB\_PL}})$, we take partial derivatives with respect to:

$$\begin{cases} \frac{\partial \text{Var}(L_{\text{CB\_PL}})}{\partial \lambda_1} = 0 \\ \frac{\partial \text{Var}(L_{\text{CB\_PL}})}{\partial \lambda_2} = 0 \end{cases}$$

and get the optimal regularization weights as:

$$\lambda_1^* = \frac{C \cdot E - B \cdot D}{A \cdot B - C^2},$$

$$\lambda_2^* = \frac{C \cdot D - A \cdot E}{A \cdot B - C^2}.$$

By substituting λ_1^*, λ_2^* back to $\mathrm{Var}(L_{\mathrm{CB\_PL}})$, we can derive the minimal variance:

$$\mathbb{V}(L_{\mathrm{CB\_PL}})_{min} = \mathbb{V}[L_{CE}] - \frac{BD^2 - 2CDE + AE^2}{AB - C^2} \leq \mathbb{V}[L_{CE}].$$

where $A = \mathbb{V}[\mathrm{BMSE}]$, $B = \mathbb{V}[\mathrm{SECE}]$, $C = \mathrm{Cov}(\mathrm{BMSE}, \mathrm{SECE})$, $D = \mathrm{Cov}(\mathrm{CE}, \mathrm{BMSE})$, and $E = \mathrm{Cov}(\mathrm{CE}, \mathrm{SECE})$.

Now let us compared to the previous with similar setting but only include BMSE as one of the regularization. Even though BMSE is used to penalize overall loss of DR estimator in previous paper, here we use to directly penalize propensity model training just for comparison, i.e. $L'_{\mathrm{total}} = CE + \lambda_1 BMSE$. The the minimal variance at the optimal hyperparameter is:

$$\begin{aligned}\mathrm{Var}(L'_{\mathrm{onlybmse}})_{\mathrm{min}} &= \mathrm{Var}(L_{CE}) - \frac{[\mathrm{Cov}(CE, BMSE)]^2}{\mathrm{Var}(BMSE)} \\ &= \mathrm{Var}(L_{CE}) - \frac{D^2}{A}.\end{aligned}$$

If the covariance matrix of BMSE and SECE is positive-definite, then $AB - C^2 > 0$. If we times $A(AB - C^2) > 0$ to both $\frac{D^2}{A}$, $\frac{BD^2 - 2CDE + AE^2}{AB - C^2}$, then we have:

$$A^2E^2 - 2ACDE + C^2D^2 = (AE - CD)^2 \geq 0,$$

leading to the conclusion that:

$$\mathrm{Var}(L'_{\mathrm{onlybmse}})_{\mathrm{min}} \geq \mathbb{V}(L_{\mathrm{CB\_PL}})_{min}.$$

Generally, using both BMSE and SECE regularizers will weakly outperform using only BMSE. □

A.2 Experimental Details

Datasets

To evaluate the debiasing performance, we conduct experiments on three benchmark datasets **Coat**[1] and **Yahoo! R3**[2], and **KuaiRec**[3], which are widely used in debiased RS with both missing not at random (MNAR) and missing at random (MAR) data. **Coat** dataset consists of 6,960 MNAR training samples and 4,640 MAR test samples derived from 290 users rating on 300 items. The **Yahoo! R3** dataset includes 311,704 MNAR training samples and 54,000 MAR test samples derived from 15,400 users rating on 1,000 items. Both datasets are five-scale, and following previous works [Chen *et al.*, 2021; Li *et al.*, 2024a; Li *et al.*, 2023c], we binarize the ratings greater than three to 1, and others to 0. The **KuaiRec** dataset is collected from a video-sharing platform and contains 4,676,570 video watching ratios derived from 1,411 users evaluating 3,327 videos. We binarize the continuous ratios greater than two to 1, otherwise to 0.

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

[2]https://webscope.sandbox.yahoo.com

[3]https://github.com/chongminggao/KuaiRec

Evaluation Metrics

We evaluate the prediction performance using three widely adopted evaluation metrics: AUC (Area Under the ROC Curve), NDCG@T (Normalized Discounted Cumulative Gain), and F1@T.

- **AUC** [Bradley, 1997] is a performance metric for classifiers that measures the probability of a randomly chosen positive example being ranked higher than a randomly chosen negative one. A higher AUC score reflects better ranking performance in differentiating positive instances from negative ones.
- **NDCG@T** [Järvelin and Kekäläinen, 2002] evaluates ranking performance by comparing the Discounted Cumulative Gain (DCG) of the top-T results to the Ideal DCG (IDCG), producing a normalized score between 0 and 1. A higher NDCG@T implies that more relevant items are ranked towards the top.

Let r_i be the relevance of the item at rank i. We first compute the DCG at rank T as well as the IDCG@T by placing the most relevant items in the optimal (ideal) order:

$$\mathrm{DCG@}T = \sum_{i=1}^{T} \frac{2^{r_i} - 1}{\log_2(i+1)} \quad , \quad \mathrm{IDCG@}T = \sum_{i=1}^{T} \frac{2^{r_i^*} - 1}{\log_2(i+1)}.$$

where r_i is the relevance of the i-th item at rank i, and r_i^* denotes the relevance of the i-th item in the ideal ranking. NDCG@T is then defined as:

$$\mathrm{NDCG@}T = \frac{\mathrm{DCG@}T}{\mathrm{IDCG@}T}.$$

- **F1@T** [Lipton *et al.*, 2014] is the harmonic mean of precision and recall computed over the top-T predictions returned by a model. A higher F1@T indicates a better trade-off between precision and recall in the top-T results.

We set T = 5 on **Coat** and **Yahoo! R3**, and T = 20 on **KuaiRec**.

Experimental Protocols

We tune calibration weight λ_1 in $\{0.1, 0.5, 1, 5, 10\}$, balancing weight λ_2 in $\{0.1, 0.5, 1, 5, 10\}$, learning rate in $\{0.01, 0.05\}$, and weight decay in $\{1e-6, 5e-6, 1e-5, \ldots, 1e-3, 5e-3\}$. We use the same hyperparameter search space and follow the results in [Li *et al.*, 2024b].