

Learning Robust Recommender from Noisy Implicit Feedback

Wenjie Wang, Fuli Feng, Xiangnan He, Liqiang Nie, Tat-Seng Chua

Abstract—The ubiquity of implicit feedback makes it indispensable for building recommender systems. However, it does not actually reflect the actual satisfaction of users. For example, in E-commerce, a large portion of clicks do not translate to purchases, and many purchases end up with negative reviews. As such, it is of importance to account for the inevitable noises in implicit feedback. However, little work on recommendation has taken the noisy nature of implicit feedback into consideration. In this work, we explore the central theme of denoising implicit feedback for recommender learning, including training and inference. By observing the process of normal recommender training, we find that noisy feedback typically has large loss values in the early stages. Inspired by this observation, we propose a new training strategy named *Adaptive Denoising Training* (ADT), which adaptively prunes the noisy interactions by two paradigms (*i.e.*, *Truncated Loss* and *Reweighted Loss*). Furthermore, we consider extra feedback (*e.g.*, rating) as auxiliary signal and propose three strategies to incorporate extra feedback into ADT: *finetuning*, *warm-up training*, and *colliding inference*. We instantiate the two paradigms on the widely used binary cross-entropy loss and test them on three representative recommender models. Extensive experiments on three benchmarks demonstrate that ADT significantly improves the quality of recommendation over normal training without using extra feedback. Besides, the proposed three strategies for using extra feedback largely enhance the denoising ability of ADT.

Index Terms—Recommendation, Denoising Implicit Feedback, Adaptive Denoising Training, Colliding Inference

1 INTRODUCTION

RECOMMENDER systems have been widely used to mine user preference over various items in many online services such as news portals [2] and social networks [3]. As the clue to user choices, implicit feedback (*e.g.*, click and purchase) are typically the default choice to train a recommender model due to their large volume. However, prior work [2], [4], [5] points out the gap between implicit feedback and the actual user satisfaction due to the existence of *noisy interactions* (*a.k.a.* *false-positive interactions*) where the users dislike the interacted items. For instance, in E-commerce, a large portion of purchases end up with negative reviews or being returned. This is because implicit interactions are easily affected by the first impression of users and factors [6] like caption bias [7] and position bias [8]. Moreover, existing studies [5] have validated the detrimental effect of such noisy interactions on user

- This is an extension of our previous work [1] (WSDM 2021). We substantially extend it from the following aspects: 1) we propose to utilize sparse extra feedback (*e.g.*, ratings) to help denoising recommendation; 2) we develop two training strategies to incorporate extra feedback into ADT: finetuning and warm-up training. Besides, we propose one new causality-based inference strategy — colliding inference; 3) we conduct extensive experiments and analysis on the three strategies in Section 5; and 4) we add related work about causal inference and robust recommendation in Section 4.
- W. Wang, F. Feng and TS. Chua are with School of Computing, National University of Singapore, Computing 1, Computing Drive, 117417, Singapore. E-mail: {wenjiewang96, fulifeng93}@gmail.com, dcscts@nus.edu.sg. Corresponding author: Fuli Feng.
- X. He is with School of Information Science and Technology, University of Science and Technology of China, Hefei, China. E-mail: xiangnanhe@gmail.com.
- L. Nie is with the School of Computer Science and Technology, Shandong University, Qingdao, China. E-mail: nieliqiang@gmail.com.

experience. However, little work on recommendation has considered the noisy nature of implicit feedback.

In this work, we argue that false-positive interactions would prevent a recommender model from learning the actual user preference, leading to low-quality recommendations. Table 1 provides empirical evidence on the negative effects of false-positive interactions when we train a competitive recommender model, Neural Matrix Factorization (NeuMF) [3], on two real-world datasets. We construct a “clean” testing set by removing the false-positive interactions for recommender evaluation¹. As shown in Table 1, training NeuMF with false-positive interactions (*i.e.*, *normal training*) results in sharp performance drop as compared to the NeuMF trained without false-positive interactions (*i.e.*, *clean training*). Therefore, it is essential to denoise implicit feedback for recommender learning.

Existing studies [9]–[11] on denoising implicit feedback mainly fall into two categories: 1) negative experience identification [10]; and 2) the incorporation of various feedback [11], [12]. As illustrated in Figure 1(b), the former identifies the false-positive interactions with additional user behaviors (*e.g.*, dwell time and gaze pattern) and auxiliary item features (*e.g.*, length of the item description) [2] before recommender training. Besides, Figure 1(c) shows that the latter leverages extra feedback (*e.g.*, favorite and skip) to prune the effects of false-positive interactions [12], for example, multi-task learning [13]. A key limitation with these methods is that they heavily rely on extra feedback or auxiliary item features, which might be hard to collect.

1. Each false-positive interaction is identified by some reliable feedback, *e.g.*, rating score ([1, 5]) < 3, indicating that the user is dissatisfied with the item. Refer to Section 2 for more details.

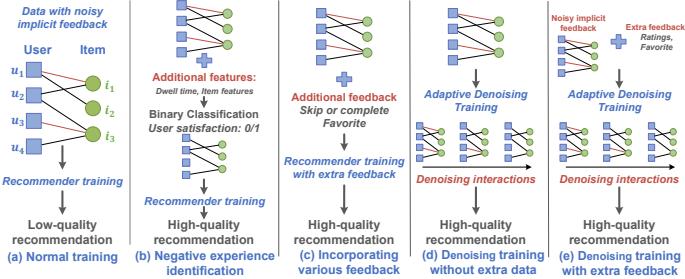


Fig. 1. The comparison between normal training (a); two prior solutions (b) and (c); and the proposed denoising training without additional data (d) and with extra feedback (e). Note that red lines in the user-item graph denote false-positive interactions, and extra feedback usually cannot identify all false-positive ones due to the sparsity issue.

Furthermore, extra feedback (*e.g.*, rating or favorite) is often of a smaller scale, which will suffer from the sparsity issue. For instance, many users do not give any extra feedback after watching a movie or purchasing a product [4].

This work aims to automatically reduce the negative effect of false-positive interactions, where we mainly count on the implicit interactions for denoising. Prior research on robust learning [14], [15] and curriculum learning [16] demonstrates that noisy samples are relatively harder to fit into models, exhibiting distinct patterns in the model training. We conduct preliminary experiments across different recommender models and datasets, obtaining a similar observation: *the loss values of false-positive interactions are relatively larger in the early stages of training (cf. Figure 2)*. Consequently, false-positive interactions will mislead the recommender training in the early stages. Worse still, the recommender model ultimately fits the false-positive interactions due to its high representation capacity, which could be overfitting and hurt the generalization. In this light, a potential solution of denoising is to prune the interactions with large loss values in the early training stage.

Towards this end, we propose Adaptive Denoising Training (ADT), which dynamically prunes the large-loss interactions along the training process (illustrated in Figure 1(d)). To avoid losing generality, we only modify the training loss, which can be applied to any differentiable models. In detail, we devise two paradigms to formulate the training loss: 1) *Truncated Loss*, which discards the large-loss interactions dynamically with a threshold function, and 2) *Reweighted Loss*, which adaptively assigns hard interactions (*i.e.*, large-loss ones) with smaller weights to reduce their influence on the recommender optimization.

To keep the flexibility of ADT in the scenarios with easily available extra feedback, we propose two training strategies to utilize extra feedback (Figure 1(e)): 1) finetuning, which refines the recommender model with extra feedback after ADT; and 2) warm-up training, which trains the recommender model with extra feedback before ADT. To tackle the sparsity issue of extra feedback, we design an inference strategy, called colliding inference, which leverages the colliding effect [17] of extra feedback on the predictions of sparse users for better recommendation. From a causal view, extra feedback and the predictions of sparse users are *d*-separated due to the sparsity issue. By conditioning on a collider between them, we make them *d*-

TABLE 1
Results of the clean training and normal training over NeuMF. #Drop denotes the relative performance drop of normal training as compared to clean training.

Dataset Metric	Adressa		Amazon-book	
	Recall@20	NDCG@20	Recall@20	NDCG@20
Clean training	0.4040	0.1963	0.0293	0.0159
Normal training	0.3159	0.1886	0.0265	0.0145
#Drop	21.81%	3.92%	9.56%	8.81%

connected and leverage the colliding effect to augment the predictions of sparse users [18].

We instantiate ADT on the widely used binary cross-entropy (CE) loss. On three benchmarks, we test the Truncated Loss and Reweighted Loss over three representative recommender models: Generalized Matrix Factorization (GMF) [3], NeuMF [3], and Collaborative Denoising Auto-Encoder (CDAE) [19]. The results show significant performance improvements of ADT over normal training. Besides, we conduct extensive experiments to validate the effectiveness of the strategies that incorporate extra feedback into ADT. They further improve the performance of the vanilla ADT, and outperform some competitive baselines that use extra feedback for training. Codes and data are publicly available at: <https://github.com/WenjieWWJ/DenoisingRec>.

Our main contributions are summarized as:

- We reveal the negative effect of false-positive interactions and propose Adaptive Denoising Training to prune the large-loss interactions dynamically, which introduces two paradigms to formulate the training loss: Truncated Loss and Reweighted Loss.
- We develop three strategies (*i.e.*, finetuning, warm-up training, and colliding inference) to incorporate extra feedback into ADT, which provides the potential to further improve the denoising ability of ADT.
- Extensive experiments on three benchmarks validate the effectiveness of the proposed ADT and three strategies in improving the recommendation quality.

2 STUDY ON FALSE-POSITIVE FEEDBACK

The influence of noisy training samples has been explored in many conventional machine learning tasks such as image classification [14], [15]. However, little work has considered denoising implicit feedback in recommendation, which is inherently different from conventional tasks due to the relations across training samples, *e.g.*, interactions on the same item. We investigate the effects of false-positive interactions on recommender training by comparing the performance under two settings: 1) normal training, which utilizes all observed user-item interactions, and 2) clean training, which discards false-positive interactions. An interaction is identified as false-positive or true-positive one according to the explicit feedback. For instance, a purchase is false-positive if the following rating score ($[1, 5]$) < 3 . Although such explicit feedback is sparse, the scale is sufficient to conduct a pilot experiment and construct a clean testing set. The recommender models are evaluated on a holdout clean testing set with only true-positive interactions kept, *i.e.*, the evaluation focuses on

recommending more satisfying items to users. More details can be seen in Section 5.

Results. From Table 1, we can observe that, as compared to clean training, the performance of normal training drops significantly (e.g., 21.81% *w.r.t.* Recall@20 on Adressa). It validates the negative effects of false-positive interactions on recommending satisfying items to users. Worse still, if the recommender models under normal training are deployed, they will have higher risk to produce more false-positive interactions, which further hurt the user experience [2].

3 METHOD

In this section, we detail the proposed ADT and three strategies to incorporate extra feedback into ADT. Prior to that, the task formulation of denoising recommender training and some observations that inspire the strategy design are introduced.

3.1 Task Formulation

Generally, the target of recommender training is to learn a scoring function $\hat{y}_{ui} = f(u, i | \Theta)$ to assess the preference of user u over item i with parameters Θ . Ideally, Θ is learned from reliable feedback between users (\mathcal{U}) and items (\mathcal{I}) by minimizing a recommendation loss \mathcal{L} (e.g., the binary CE loss). Formally, $\Theta^* = \arg \min_{\Theta} \mathcal{L}(D^*)$ where $D^* = \{(u, i, y_{ui}^*) | u \in \mathcal{U}, i \in \mathcal{I}\}$, and $y_{ui}^* \in \{0, 1\}$ represents whether the user u really prefers the item i . The recommender model with Θ^* would be reliable to generate high-quality recommendations. In practice, due to the lack of large-scale reliable feedback, recommender training is typically formalized as: $\bar{\Theta} = \arg \min_{\Theta} \mathcal{L}(\bar{D})$, where $\bar{D} = \{(u, i, \bar{y}_{ui}) | u \in \mathcal{U}, i \in \mathcal{I}\}$ is a set of implicit interactions. \bar{y}_{ui} denotes whether the user u has interacted with the item i , such as click and purchase.

However, because noisy interactions in \bar{D} would mislead the learning of Θ , the typical recommender training might lead to a poor model (*i.e.*, $\bar{\Theta}$) that lacks generalization ability on the clean testing set. As such, we formulate a *denoising recommender training* task as:

$$\Theta^* = \arg \min_{\Theta} \mathcal{L}(\text{denoise}(\bar{D})), \quad (1)$$

aiming to learn a reliable recommender model with parameters Θ^* by denoising implicit feedback \bar{D} . Formally, by assuming the existence of inconsistency between y_{ui}^* and \bar{y}_{ui} , we define noisy interactions (*a.k.a.* false-positive interactions) as $\{(u, i) | y_{ui}^* = 0 \wedge \bar{y}_{ui} = 1\}$ [1].

3.2 Observations

False-positive interactions are harder to fit in the early stages. According to the theory of robust learning [14], [15] and curriculum learning [16], easy samples are more likely to be the clean ones and fitting the hard samples may hurt the generalization. To verify the existence of this phenomenon in recommendation, we compare the loss of true- and false-positive interactions along the process of normal training. Figure 2 shows the results of NeuMF on the Adressa dataset. Similar trends are also found over other recommender models and datasets (see more details in Section 5.2). From Figure 2, we observe that:

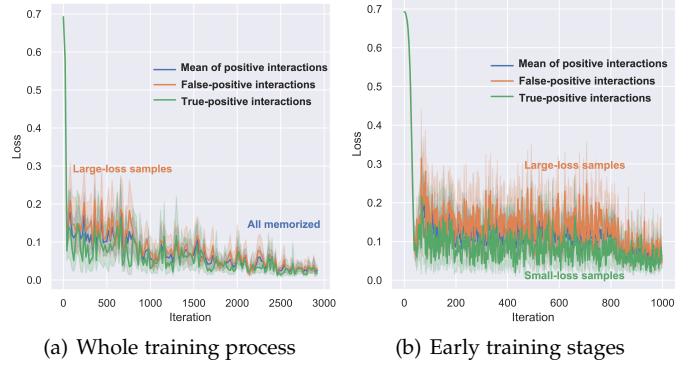


Fig. 2. The training loss of true- and false-positive interactions on Adressa in the normal training of NeuMF. The false-positive interactions have larger loss values in the early training stage.

- Ultimately, the loss of both of true- and false-positive interactions converges to a stable state with close values, which implies that NeuMF fits both of them well. It reflects that deep recommender models can “memorize” all the training interactions, including the noisy ones. As such, if the data is noisy, the memorization will lead to poor generalization performance.
- The loss values of true- and false-positive interactions decrease differently in the early stages of training. From Figure 2(b), we can see that the loss of false-positive interactions is clearly larger than that of the true-positive ones, which indicates that false-positive interactions are harder to memorize than the true-positive ones in the early stages. The reason might be that false-positive ones represent the items that users dislike, and they are more similar to the items the user didn’t interact with. The findings also support the prior theory in robust learning and curriculum learning [15], [16], [20].

Overall, the results are also consistent with the memorization effect [21]: deep recommender models will learn the easy and clean patterns in the early stage, and eventually memorize all training interactions [15] because of the strong representation ability.

3.3 Adaptive Denoising Training

According to the observations, we propose ADT strategies for recommender models, which estimate $P(y_{ui}^* = 0 | \bar{y}_{ui} = 1, u, i)$ according to the training loss. In particular, ADT either *discards* or *reweights* the interactions with large loss values to reduce their influences on the recommender training. Towards this end, we devise two paradigms to formulate loss functions for denoising training without using extra feedback:

- *Truncated Loss.* This is to truncate the loss values of hard interactions to 0 with a dynamic threshold function.
- *Reweighted Loss.* It adaptively assigns hard interactions with smaller weights during training.

Note that the two paradigms can be applied to various recommendation loss functions, *e.g.*, CE loss, square loss, and BPR loss [22]. In the work, we take CE loss as an example to elaborate them.

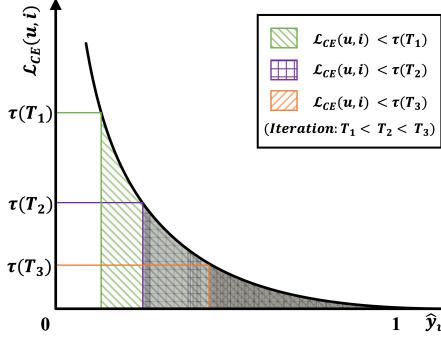


Fig. 3. Illustration of T-CE loss for the observed interactions (*i.e.*, $\bar{y}_{ui} = 1$). T_i is the iteration number and $\tau(T_i)$ refers to the threshold. Dash area indicates the effective loss and the loss values larger than $\tau(T_i)$ are truncated during training.

3.3.1 Truncated Cross-Entropy Loss

Functionally speaking, the Truncated Cross-Entropy (shorted as T-CE) loss discards positive interactions according to the values of CE loss. Formally, we define it as:

$$\mathcal{L}_{T\text{-CE}}(u, i) = \begin{cases} 0, & \mathcal{L}_{CE}(u, i) > \tau \wedge \bar{y}_{ui} = 1 \\ \mathcal{L}_{CE}(u, i), & \text{otherwise,} \end{cases} \quad (2)$$

where τ is a pre-defined threshold. The T-CE loss removes any positive interactions with CE loss (*i.e.*, \mathcal{L}_{CE}) larger than τ from the training. While the T-CE loss is easy to interpret and implement, the fixed threshold may not work properly. This is because the loss value is decreasing with the increase of training iterations. Inspired by the dynamic adjustment ideas [23], we replace the fixed threshold with a dynamic threshold function $\tau(T)$ *w.r.t.* the training iteration T , which changes the threshold value along the training process (Figure 3). Besides, since the loss values vary across different datasets, it would be more flexible to devise $\tau(T)$ as a function of the drop rate $\epsilon(T)$. There is a bijection between the drop rate and the threshold, *i.e.*, we can calculate the threshold based on the drop rate $\epsilon(T)$ at each iteration.

Based on prior observations, a proper drop rate function should satisfy: 1) $\epsilon(\cdot)$ should have an upper bound to limit the proportion of discarded interactions so as to prevent data missing; 2) $\epsilon(0) = 0$, *i.e.*, it should allow all the interactions to be fed into the models in the beginning; and 3) $\epsilon(\cdot)$ should increase smoothly from zero to its upper bound, so that the model can learn and distinguish the true-and false-positive interactions gradually.

To this end, we formulate the drop rate function as:

$$\epsilon(T) = \min(\alpha T, \epsilon_{max}), \quad (3)$$

where ϵ_{max} is an upper bound and α is a hyper-parameter to adjust the pace to reach the maximum drop rate. Here we increase the drop rate in a linear fashion rather than a more complex function such as a polynomial function or a logarithm function. Despite the expressiveness of these functions, they have more hyper-parameters, increasing cost of tuning a recommender model. The algorithm is formally explained in Algorithm 1.

Algorithm 1 Adaptive Denoising Training with T-CE loss

Input: trainable parameters Θ , training interactions \bar{D} , the iterations T_{max} , learning rate η , ϵ_{max} , α , \mathcal{L}_{CE} , parameter optimization function ∇ .

- 1: **for** $T = 1 \rightarrow T_{max}$ **do** ▷ shuffle interactions every epoch
- 2: Fetch mini-batch data \bar{D}_{pos} from \bar{D}
- 3: Sample unobserved interactions \bar{D}_{neg}
- 4: Define $\bar{D}_T = \bar{D}_{pos} \cup \bar{D}_{neg}$
- 5: Obtain $\hat{D} = \arg \max_{\hat{D} \subset \bar{D}_{pos}, |\hat{D}| = |\epsilon(T)*|\bar{D}_T|} \sum_{(u,i) \in \hat{D}} \mathcal{L}_{CE}(u, i | \Theta_{T,1})$
- 6: Obtain $D^* = \bar{D}_T - \hat{D}$
- 7: Update $\Theta_T = \nabla(\Theta_{T-1}, \eta, \mathcal{L}_{CE}, D^*)$
- 8: Update $\epsilon(T) = \min(\alpha T, \epsilon_{max})$
- 9: **end for**

Output: the optimized parameters $\Theta_{T_{max}}$ of the recommender model.

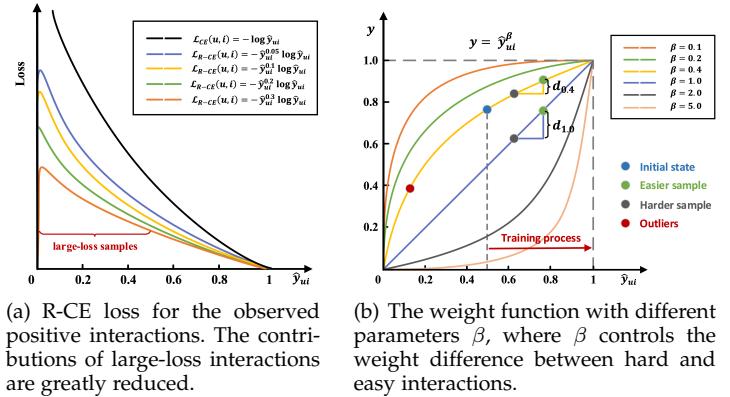


Fig. 4. Illustration and analysis of R-CE loss.

3.3.2 Reweighted Cross-Entropy Loss

Generally, the Reweighted Cross-Entropy (shorted as R-CE) loss down-weights the hard positive interactions. Formally,

$$\mathcal{L}_{R\text{-CE}}(u, i) = \omega(u, i) \mathcal{L}_{CE}(u, i), \quad (4)$$

where $\omega(u, i)$ is a weight function that adjusts the contribution of an interaction to the training objective. To properly down-weight the hard interactions, the weight function $\omega(u, i)$ is expected to have the following properties: 1) it dynamically adjusts the weights of interactions during training; 2) the function makes the influence of a hard interaction weaker than an easy interaction; and 3) the extent of weight reduction can be easily adjusted to fit different models and datasets.

Inspired by the success of Focal Loss [24], we estimate $\omega(u, i)$ according to the prediction score \hat{y}_{ui} . The prediction score is within $[0, 1]$ whereas the value of CE loss is in $[0, +\infty]$. And thus the prediction score is more accountable for further computation. Note that the smaller prediction score on the positive interaction leads to larger CE loss. Formally, we define the weight function as:

$$\omega(u, i) = \hat{y}_{ui}^\beta, \quad (5)$$

where $\beta \in [0, +\infty]$ is a hyper-parameter to control the range of weights. From Figure 4(a), we can see that R-CE loss can significantly reduce the loss of hard interactions (*e.g.*, $\hat{y}_{ui} \ll 0.5$) as compared to the original CE loss. Furthermore, the proposed weight function satisfies the aforementioned requirements:

- It generates dynamic weights during training since \hat{y}_{ui} changes along the training process.
- The interactions with extremely large CE loss (e.g., the “outlier” in Figure 4(b)) will be assigned with very small weights because \hat{y}_{ui} is close to 0. Therefore, the influence of such large-loss interactions is largely reduced. Besides, as shown in Figure 4(b), harder interactions always have smaller weights because the function \hat{y}_{ui}^β monotonically increases w.r.t. \hat{y}_{ui} when $\hat{y}_{ui} \in [0, 1]$ and $\beta \in [0, +\infty]$. As such, it can avoid that false-positive interactions dominate the optimization during training [25].
- The hyper-parameter β dynamically controls the gap between the weights of hard and easy interactions. By observing the examples in Figure 4(b), we can find that: 1) the gap between the weights of easy and hard interactions becomes larger as β increases from 0 to 1 (e.g., $d_{0.4} < d_{1.0}$) when $\hat{y}_{ui} > 0.5$; and 2) the R-CE loss will degrade to a standard CE loss if β is 0.

Lastly, in order to prevent negative interactions with large loss values from dominating the optimization, we revise the weight function as:

$$\omega(u, i) = \begin{cases} \hat{y}_{ui}^\beta, & \bar{y}_{ui} = 1, \\ (1 - \hat{y}_{ui})^\beta, & \text{otherwise.} \end{cases} \quad (6)$$

Indeed, it may provide a idea to alleviate the impact of false-negative interactions, which is left for future exploration.

3.3.3 In-depth Analysis

Due to depending totally on recommender models to identify false-positive interactions, the reliability of ADT needs to be discussed. Actually, existing work [14], [15] has pointed out the connection between the large loss and noisy interactions, and explained the underlying causality: the “memorization” effect of deep models [21]. We discuss the memorization effect of recommender models by experiments in Section 3.2 and 5.2.

Another concern is that discarding hard interactions would limit the model’s learning ability since some hard interactions may be more informative than the easy ones. Indeed, as discussed in the prior studies on curriculum learning [16], hard interactions in the noisy data probably confuse the classifier rather than help it to establish the right decision surface. It’s actually a trade-off between denoising and learning. In ADT, the $\epsilon(\cdot)$ of T-CE loss and β of R-CE loss are used to control the balance.

3.4 ADT with Extra Feedback

Although extra feedback (e.g., ratings) is usually sparse, it is reliable to reflect the actual user satisfaction, *i.e.*, indicating the true-positive interactions. We thus further utilize extra feedback for ADT when available².

3.4.1 Training with Extra Feedback

By considering the order of training with implicit feedback and extra feedback, we introduce two training strategies: finetuning and warm-up training.

2. This is not equal to clean training in Section 2 because available extra feedback is insufficient to identify all true-positive interactions in implicit feedback.

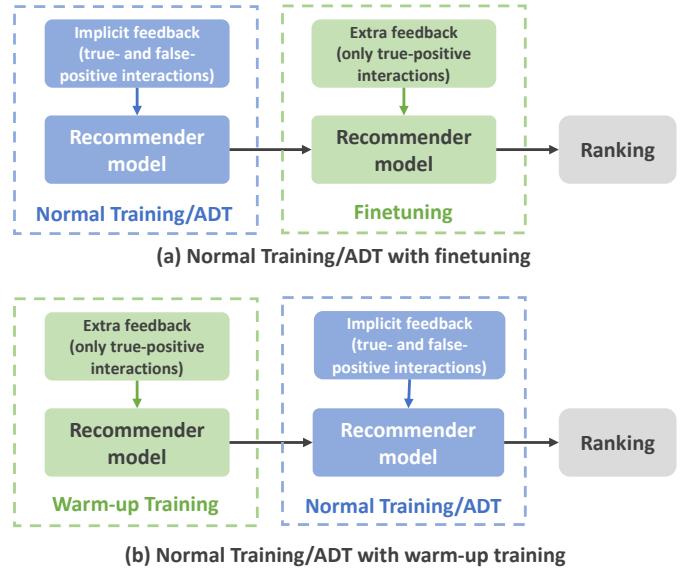


Fig. 5. Illustration of finetuning and warm-up training with extra feedback.

Finetuning. As shown in Figure 5(a), an intuitive idea is to retrain the recommender model by fitting the available true-positive interactions. Specifically, we first train the recommender model with ADT over the implicit feedback, and finetune the model over the extra feedback. The finetuning will refine the recommendation lists of users based only on the extra feedback. Formally, given the extra feedback D^* , the model parameter Θ^* obtained from ADT will be further optimized by:

$$\arg \min_{\Theta^*} \mathcal{L}_{CE}(u, i), \text{ where } (u, i, y_{ui}^*) \in D^*, \quad (7)$$

where the negative interactions in D^* are sampled with the help of implicit feedback \bar{D} . Only the interactions which are not in \bar{D} are sampled into D^* to partly avoid using the missing true-positive interactions as negative ones.

Warm-up Training. From Section 3.2, we know that the recommender model learns to distinguish the true-positive and false-positive interactions in the early stage of training. If extra feedback D^* is available, we can leverage it to “warm up” the recommender model before ADT (illustrated in Figure 5(b)). Warm-up training will prevent the recommender model from being misled by false-positive interactions in the beginning of training. The recommender model will thus be more reliable to identify false-positive interactions by the large loss values during ADT. Formally, warm-up training can be seen as optimizing the model initialization for ADT by Equation (7).

3.4.2 Colliding Inference

The key to the success of incorporating extra feedback lies in resolving its sparsity issue. For the users with sparse extra feedback, the influence of extra feedback D^* on their final prediction scores \hat{y} is weakened due to the small number of extra feedback as opposed to implicit feedback. Therefore, our target is to enhance the effect of reliable D^* on \hat{y} .

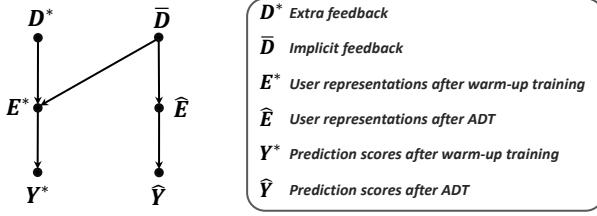


Fig. 6. Causal graph for ADT with warm-up training.

Implementation. We achieve the target during model inference by supplementing the model prediction for sparse user u with similar users. Formally,

$$\hat{y}'_u = \lambda * \hat{y}_u + (1 - \lambda) * \sum_{j \in \mathcal{N}_u} w_j \hat{y}_j, \quad (8)$$

where $\hat{y}_u = [\hat{y}_{u1}, \dots, \hat{y}_{u|\mathcal{I}|}]^T$ denotes the model prediction of user u over all items (*i.e.*, item set \mathcal{I}), \mathcal{N}_u is the neighbor set of user u , \hat{y}_j represents the prediction scores of neighbor j , and w_j denotes the weights to adjust the contribution of each neighbor. Specifically, we calculate the user similarity via inner product in the user representations space learned from extra feedback, and then select $|\mathcal{N}_u|$ similar neighbors for each user. The weighted sum of prediction scores of the neighbors summarizes the colliding effect of extra feedback on the final prediction score \hat{y}'_u (see the following section for the explanation). Besides, a hyper-parameter λ is employed to balance the original prediction scores \hat{y}_u and the colliding effect. \hat{y}'_u is the final ranking score for making recommendation.

Intuitively, the neighbors are obtained from a user representation space learned from true-positive interactions, where the interactions reflect the actual user preference of sparse users. Due to the sparsity of extra feedback, some interests of sparse users are not reflected by the representation, making the representation unsuitable for making prediction. Nevertheless, such representation reflects user similarity regarding their true interests. As such, we lookup neighbors according to the representation and augment the recommendations of the sparse user with neighbors' recommendations.

Rationality. The idea of Equation 8 comes from the colliding effect in causality theory [17]. In this work, we utilize warm-up training as an example to illustrate the underlying theory of colliding inference because of its superior performance to enhance ADT (see the empirical evidence in Section 5.3).

As shown in Figure 6, we introduce a causal graph [17], [26] to inspect the causal relations in ADT with warm-up training for sparse users, where nodes and edges represent the variables and causal relations, respectively [6]. Specifically, it presents extra feedback (D^*), implicit feedback (\bar{D}), user representations after warm-up training (E^*), user representations after ADT (\hat{E}), prediction scores over items after warm-up training (Y^*), and prediction scores over items after ADT (\hat{Y}). In particular,

- $D^* \rightarrow E^*$ or $\bar{D} \rightarrow \hat{E}$: the user representations are learned from the user feedback.
- $\bar{D} \rightarrow E^*$: the user representations after warm-up training are affected by implicit feedback because it is used to sample negative interactions during training.

- $E^* \rightarrow Y^*$ or $\hat{E} \rightarrow \hat{Y}$ ³: the prediction scores are based on the user representations.

According to the d -separation principle [17], D^* is independent from \bar{D} , \hat{E} , and \hat{Y} , which blocks the effect of extra feedback D^* on the final prediction score \hat{Y} of sparse users. Considering that E^* is a collider between D^* and \bar{D} , conditioning on E^* will make D^* and \hat{Y} conditionally d -connected [17], [18]. In this way, we can estimate the colliding effect [26] of D^* on \hat{Y} , which is formulated as:

$$\begin{aligned} & P(\hat{Y}|E^*, D^* = d_u) - P(\hat{Y}|E^*, D^* = \emptyset) \\ &= \sum_{\bar{D}} P(\hat{Y}|E^*, \bar{D})(P(\bar{D}|E^*, D^* = d_u) - P(\bar{D}|E^*, D^* = \emptyset)) \\ &= \sum_{\bar{D}} P(\hat{Y}|\bar{D})(P(\bar{D}|E^*, D^* = d_u) - P(\bar{D}|E^*, D^* = \emptyset)), \end{aligned} \quad (9)$$

where $D^* = \emptyset$ denotes the reference status without any extra feedback [6], [18] and d_u indicates the true-positive interactions of user u . $P(\hat{Y}|E^*, \bar{D}) = P(\hat{Y}|\bar{D})$ because E^* can only affect \hat{Y} via \bar{D} . Intuitively, the causal effect in Equation (9) quantifies the influence of extra feedback of user u on the prediction scores by comparing two statuses.

We utilize the technique of *Distilling Colliding Effect* [18] to calculate Equation (9). In particular, we will employ a weighting method to estimate the causal effect. Formally, $P(\bar{D}|E^*, D^* = d_u) - P(\bar{D}|E^*, D^* = \emptyset)$ indicates the probability of selecting users' implicit feedback \bar{D} given E^* and D^* , which can be denoted as the weight $W(\bar{D}, E^*, D^*)$ [18], corresponding to the w_j in Equation (8). Besides, $P(\hat{Y}|\bar{D})$ represents the prediction scores after ADT given the implicit feedback \bar{D} of one user. Because of the large sample space of \bar{D} , the sum over \bar{D} is approximated by only selecting the implicit feedback of neighbors \mathcal{N}_u with the highest weights $W(\cdot)$. In this way, Equation (9) is estimated by the weighted sum of $P(\hat{Y}|\bar{D})$ over the neighbors, *i.e.*, $\sum_{j \in \mathcal{N}_u} w_j \hat{y}_j$. Correspondingly, we search the neighbors of each sparse user according to similarity in the space of E^* , and then refine the user's prediction scores \hat{Y} by combining the prediction scores of the neighbors.

4 RELATED WORK

Denoising implicit feedback is highly related to the negative experience identification, incorporating various feedback, and the robustness of recommendation. Besides, we introduce causal recommendation that is related to colliding inference.

Negative Experience Identification. To mitigate the gap between implicit feedback and the actual user preference, many studies have focused on identifying negative experiences in implicit signals [9], [27], [28]. Prior work usually collects the various users' feedback (*e.g.*, dwell time [10], gaze patterns [29], and skip [9]) and the item characteristics [2], [30] to predict the user's satisfaction. However, they need additional feedback and extensive manual label work, *e.g.*, users have to tell if they are

3. We assume that E^* and \hat{E} are independent for the users with sparse extra feedback. While E^* is the initialization of \hat{E} , the influence of the initialization E^* will be easily overwhelmed by the extensive implicit feedback during ADT [18].

satisfied for each interaction. Besides, the quantification of item quality is non-trivial [2], which largely relies on the manually feature design and the labeling of domain experts [2]. The unaffordable labor cost hinders the practical usage of these methods, especially in the scenarios with constantly changing items.

Incorporating Various Feedback. To alleviate the impact of false-positive interactions, previous approaches [11], [31] also consider incorporating more feedback (*e.g.*, dwell time [12], skip [5], and adding to favorites) into training directly. For instance, Wen *et al.* [5] proposed to train the recommender using three kinds of items: “click-complete”, “click-skip”, and “non-click” ones. The last two kinds of items are both treated as negative samples but with different weights. However, extra feedback might be unavailable in complex scenarios. For example, we cannot acquire dwell time and skip patterns after users buy products or watch movies in a cinema. Most users even don’t give any informative feedback after clicks. In an orthogonal direction, this work explores denoising implicit feedback without additional data during training, *i.e.*, ADT. Furthermore, if some extra feedback (*e.g.*, ratings) is available, we also introduce three strategies to utilize it to enhance ADT.

Robustness of Recommender Systems. Gunawardana *et al.* [32] defined the robustness of recommender systems as “the stability of the recommendation in the presence of fake information”. Prior work [33], [34] has tried to evaluate the robustness of recommender systems under various external attacks, such as shilling attacks [33], fuzzing attacks [34], and fake feedback from fraudsters [35]. To build more robust recommender systems, some auto-encoder based models [19], [36], [37] introduce the denoising techniques. These approaches (*e.g.*, CDAE [19]) first corrupt the interactions of user by random noises, and then try to reconstruct the original one with auto-encoders. Besides, adversarial learning [38] might serve as a reasonable solution to defend the diverse attacks or detect fake users [39], [40]. And robust learning [15] will also be a promising choice to empower recommender models with the ability of denoising user feedback [41]. However, as these methods focus on external attacks or random noises, they ignore the natural false-positive interactions in data. This work highlights the negative impact of natural noisy interactions, and improve the robustness against them.

Causal Recommendation. Recently, causal inference [17], [26] has been widely used to liberate deep learning from blindly fitting the correlation [42]–[45]. Theoretically, it leverages causal graphs to describe the causal relationships between different variables and conducts three-level reasoning over the causal graph: observation, intervention, and counterfactual [17]. As to the information retrieval domain, early research [46]–[49] starts from using causal inference to remove various bias in user feedback [50], such as position bias [51] and popularity bias [52]. Recently, more techniques in causal inference have been applied to solve the issues in recommendation [53]–[55], for example, causal intervention [56] and counterfactual inference [6]. However, existing work on causal recommendation has never studied the colliding effect or the causality-based inference strategy.

TABLE 2
Statistics of the datasets. #FP interactions refer to the number of false-positive interactions.

Dataset	#User	#Item	#Interaction	#FP Interaction
Adressa	212,231	6,596	419,491	247,628
Amazon-book	80,464	98,663	2,714,021	199,475
Yelp	45,548	57,396	1,672,520	260,581

5 EXPERIMENT

Dataset. To evaluate the effectiveness of ADT, we conducted experiments on three benchmarks: Adressa, Amazon-book, and Yelp. Extra feedback (*e.g.*, ratings and dwell time) is used to identify the false-positive interactions.

- **Adressa:** This is a news recommendation dataset from Adressavisen⁴ [57]. It includes user clicks over news and the dwell time for each click, where the clicks with dwell time < 10s are treated as false-positive ones [10], [12].
- **Amazon-book:** It is from the Amazon-review datasets⁵ [58]. It records users’ purchases over books with rating scores. A rating score below 3 is regarded as a false-positive interaction.
- **Yelp⁶:** It covers the businesses in the catering industry (*e.g.*, restaurants and bars), which are reviewed by users. Similar to Amazon-book, the rating scores below 3 are false-positive interactions.

We split the dataset into training, validation, and testing sets, and explored two experimental settings:

- 1) Extra feedback is unavailable during training (Section 5.1 and 5.2). To evaluate the performance of denoising implicit feedback, we kept all interactions, including the false-positive ones, in training and validation, and tested the models only on true-positive interactions.
- 2) Sparse extra feedback is available during training (Section 5.3). We assume that partial true-positive interactions have already been known, which will be used to verify the performance of the proposed three strategies: finetuning, warm-up training, and colliding inference.

Evaluation Protocols. For each user in the testing set, we predicted the preference score over all the items except the positive ones used during training. Following existing studies [3], [59], we reported the performance *w.r.t.* two widely used metrics: Recall@K and NDCG@K, where higher scores indicate better performance. For both metrics, we set K as 50 and 100 for Amazon-book and Yelp, while 3 and 20 for Adressa due to its much smaller item space.

Testing recommender models. To demonstrate the effectiveness of our proposed ADT strategy on denoising implicit feedback, we compared the performance of recommender models trained with T-CE or R-CE and normal training with standard CE. We selected two representative user-based neural models, GMF and NeuMF [3], and one item-based model, CDAE [19], which is a representative recommender model which can defend random noises in implicit feedback. More details can be found in [1].

4. <https://www.adressa.no/>

5. <http://jmcauley.ucsd.edu/data/amazon/>

6. <https://www.yelp.com/dataset/challenge>

TABLE 3

Overall performance of three testing recommender models trained with ADT strategies and normal training over the three datasets. Note that Recall@K and NDCG@K are shorted as R@K and N@K to save space, respectively, and “RI” in the last column denotes the relative improvement of ADT over normal training on average. The best results are highlighted in bold.

Dataset Metric	Adressa				Amazon-book				Yelp				RI
	R@3	R@20	N@3	N@20	R@50	R@100	N@50	N@100	R@50	R@100	N@50	N@100	
GMF	0.0880	0.2141	0.0780	0.1237	0.0609	0.0949	0.0256	0.0331	0.0840	0.1339	0.0352	0.0465	-
GMF+T-CE	0.0892	0.2170	0.0790	0.1254	0.0707	0.1113	0.0292	0.0382	0.0871	0.1437	0.0359	0.0486	7.14%
GMF+R-CE	0.0891	0.2142	0.0765	0.1229	0.0682	0.1075	0.0275	0.0362	0.0861	0.1361	0.0366	0.0480	4.34%
NeuMF	0.1416	0.3159	0.1267	0.1886	0.0512	0.0829	0.0211	0.0282	0.0750	0.1226	0.0304	0.0411	-
NeuMF+T-CE	0.1418	0.3106	0.1227	0.1840	0.0725	0.1158	0.0289	0.0385	0.0825	0.1396	0.0323	0.0451	15.62%
NeuMF+R-CE	0.1414	0.3185	0.1266	0.1896	0.0628	0.1018	0.0248	0.0334	0.0788	0.1304	0.0320	0.0436	8.77%
CDAE	0.1394	0.3208	0.1168	0.1808	0.0989	0.1507	0.0414	0.0527	0.1112	0.1732	0.0471	0.0611	-
CDAE+T-CE	0.1406	0.3220	0.1176	0.1839	0.1088	0.1645	0.0454	0.0575	0.1165	0.1806	0.0504	0.0652	5.36%
CDAE+R-CE	0.1388	0.3164	0.1200	0.1827	0.1022	0.1560	0.0424	0.0542	0.1161	0.1801	0.0488	0.0632	2.46%

Parameter Settings. For the three testing recommender models, we followed their default settings, and verified the effectiveness of our methods under the same conditions. For GMF and NeuMF, the factor numbers of users and items are both 32. As to CDAE, the hidden size of MLP is set as 200. In addition, Adam [23] is applied to optimize all the parameters with the learning rate initialized as 0.001 and the batch size set as 1,024. As to the ADT strategies, they have three hyper-parameters in total: α and ϵ_{max} in the T-CE loss, and β in the R-CE loss. In detail, ϵ_{max} is searched in {0.05, 0.1, 0.2} and β is tuned in {0.05, 0.1, ..., 0.25, 0.5, 1.0}. As for α , we controlled its range by adjusting the iteration number ϵ_N to the maximum drop rate ϵ_{max} , and ϵ_N is adjusted in {1k, 5k, 10k, 20k, 30k}. In colliding inference, the number of neighbors N_u is tuned in {1, 3, 5, 10, 20, 50, 100}, w_j is set as $1/|N_u|$, and λ is searched in {0, 0.1, 0.2, ..., 1}. We used the validation set to tune the hyper-parameters and reported the performance on the testing set.

5.1 Performance Comparison

Overall Performance. Table 3 summarizes the performance of the three testing models trained with standard CE, T-CE, or R-CE. From Table 3, we could observe:

- In all cases, both the T-CE loss and R-CE loss effectively improve the performance, e.g., NeuMF+T-CE outperforms NeuMF by 15.62% on average over the three datasets. The significant performance gain indicates the better generalization ability of neural recommender models trained by the T-CE loss and R-CE loss. It validates the effectiveness of ADT, i.e., discarding or down-weighting hard interactions during training.
- By comparing T-CE and R-CE, we found that the T-CE loss performs better in most cases. We postulate that the recommender model trained with the R-CE loss still suffers from the false-positive interactions, even though they have smaller weights and contribute little to the overall training loss. In addition, the superior performance of the T-CE Loss might be attributed to the additional hyper-parameters in the dynamic threshold function which can be tuned more granularly. Further improvements might be achieved by a fine-grained user-specific or item-specific tuning of these hyper-parameters, which can be done automatically [60].
- Both T-CE and R-CE achieve the largest performance increase on NeuMF, which validates the effectiveness of

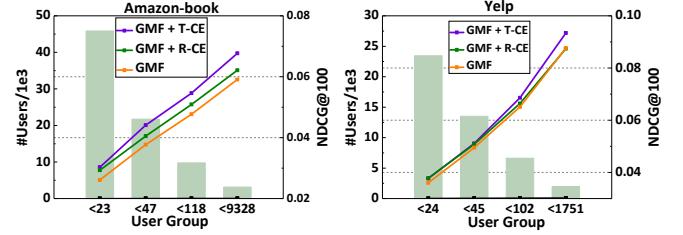


Fig. 7. Performance comparison of GMF over user groups with different sparsity levels. The histograms represent the user number and the lines denote the performance.

ADT to prevent vulnerable models from the disturbance of noisy data. On the contrary, the improvement over CDAE is relatively small, showing that the design of defending random noise can also improve the robustness against false-positive interactions to some extent. Nevertheless, applying T-CE or R-CE still leads to performance gain over CDAE, which further validates the rationality of denoising implicit feedback.

In the following, GMF is taken as an example to conduct thorough investigation to save computation cost.

Performance Comparison w.r.t. Interaction Sparsity. Since ADT prunes many interactions during training, we explored whether ADT hurts the preference learning of inactive users because their interacted items are sparse. Following the former studies [59], we split testing users into four groups according to the interaction number of each user where each group has the same number of interactions. Figure 7 shows the group-wise performance comparison where we could observe that the proposed ADT strategies achieve stable performance gain over normal training in all cases. It validates that ADT is also effective for the inactive users.

5.2 In-depth Analysis on ADT

Memorization of False-positive Interactions. Recall that false-positive interactions are memorized by recommender models eventually under normal training. We then investigated whether they are also fitted well when ADT is applied. From Figure 8(a), we could find the CE loss values of false-positive interactions eventually become similar to other interactions, indicating that GMF fits false-positive interactions well at last. On the contrary, as shown in Figure 8(b), by applying T-CE, the loss values of false-positive interactions increase while the overall training loss

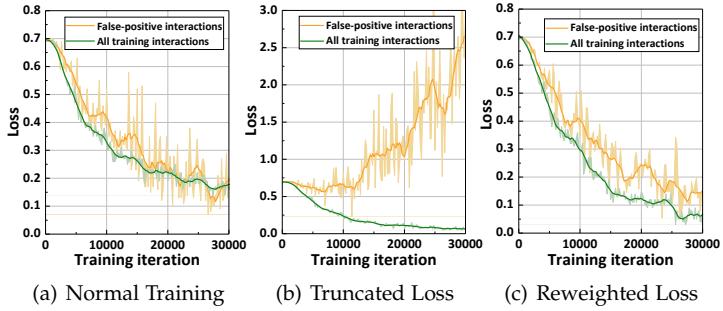


Fig. 8. Loss of GMF with normal training (a), T-CE (b) and R-CE (c).

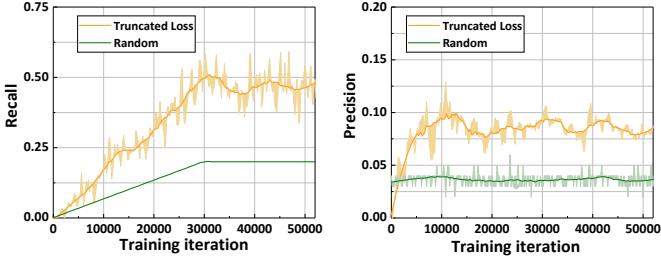


Fig. 9. Recall and precision of false-positive interactions over GMF trained with T-CE on Amazon-book.

stably decreases step by step. This indicates that the model parameters are not optimized over these false-positive interactions, validating the capability of T-CE to identify and discard such interactions. As to R-CE (Figure 8(c)), the loss of false-positive interactions also shows a decreasing trend, showing that the recommender model still fits such interactions. However, their loss values are still larger than the real training loss, indicating assigning false-positive interactions with smaller weights is effective. It prevents the model from fitting them quickly. Therefore, we could conclude that both paradigms reduce the effect of false-positive interactions on recommender training, which can explain their improvement over normal training.

Study of Truncated Loss. Since the Truncated Loss achieves promising performance, we studied how accurately it identifies and discards false-positive interactions. We first defined *recall* to represent what percentage of false-positive interactions in the training data are discarded, and *precision* as the ratio of discarded false-positive interactions to all discarded interactions. Figure 9 visualizes the changes of the recall and precision along the training process. We take random discarding as a reference, where the recall of random discarding equals to the drop rate $\epsilon(T)$ during training and the precision is the proportion of false-positive interactions within the training batch at each iteration. From Figure 9, we observed that:

- The Truncated Loss discards nearly half of false-positive interactions after the drop rate keeps stable, greatly reducing the impact of noisy interactions. This might explain the superior performance of T-CE in Table 3.
- A key limitation of the Truncated Loss is the low precision, *e.g.*, only 10% precision in Figure 9, which implies that it inevitably discards many true-positive interactions. However, given the overall superior performance of the

TABLE 4
Performance of using extra feedback *w.r.t.* GMF on Amazon-book. The best results in each method group are highlighted in bold.

Method	R@50	R@100	N@50	N@100
GMF	0.0609	0.0949	0.0256	0.0331
GMF+T-CE	0.0707	0.1113	0.0292	0.0382
GMF+R-CE	0.0682	0.1075	0.0275	0.0362
GMF+NMTR	0.0616	0.0967	0.0254	0.0332
GMF+NR	0.0615	0.0958	0.0254	0.0331
GMF+Finetuning	0.0608	0.0956	0.0253	0.0330
GMF+T-CE+Finetuning	0.0722	0.1140	0.0293	0.0386
GMF+R-CE+Finetuning	0.0684	0.1087	0.0276	0.0365
GMF+Warm-up	0.0611	0.0954	0.0255	0.0332
GMF+T-CE+Warm-up	0.0737	0.1159	0.0302	0.0395
GMF+R-CE+Warm-up	0.0691	0.1099	0.0278	0.0368

Truncated Loss, it seems worthwhile to prune noise interactions even at the cost of losing many true-positive interactions. Nevertheless, how to further improve the precision is a promising research direction in the future.

Hyper-parameter Sensitivity. ADT introduces three hyper-parameters into the two paradigms. Specifically, ϵ_{max} and ϵ_N are used to control the drop rate in T-CE loss, and β adjusts the weight function in R-CE loss. To facilitate the future tuning of the two paradigms, we studied how the hyper-parameters affect the performance. From Figure 10, we could have the following findings:

- The recommender model trained with the T-CE loss performs well when $\epsilon_{max} \in [0.1, 0.3]$. If ϵ_{max} exceeds 0.4, the performance drops significantly because a large proportion of interactions are discarded. Therefore, the upper bound ϵ_{max} should be restricted.
- The recommender model is relatively sensitive to ϵ_N , especially on Amazon-book, and the performance still increases when $\epsilon_N > 30k$. This shows that a limitation of the T-CE loss is the big search space of hyper-parameters. Therefore, it is promising to introduce some automatic hyper-parameter tuning methods [60], which could further improve denoising performance of the Truncated Loss.
- The adjustment of β in the Reweighted Loss is consistent across different datasets, and the best results happen when β ranges from 0.15 to 0.3. These observations provide insights on how to tune β if it's applied to other recommender models and datasets.

5.3 ADT with Extra Feedback

Performance Comparison. To avoid the detrimental effect of false-positive interactions, a popular idea is to incorporate the extra user feedback (*e.g.*, ratings and favorite) for training although they are usually sparse. Existing work either adopts the additional feedback by multi-task learning [13], or leverages it to identify the true-positive interactions [2], [5]. In this work, we introduced two classical models for comparison: Neural Multi-Task Recommendation (NMTR) [13] and Negative feedback Re-weighting (NR) [5]. In particular, NMTR considers both click and extra feedback with multi-task learning. NR uses extra feedback (*i.e.*, dwell time or rating) to identify true-positive interactions and re-weights the false-positive and non-interacted ones. We applied NMTR, NR, the

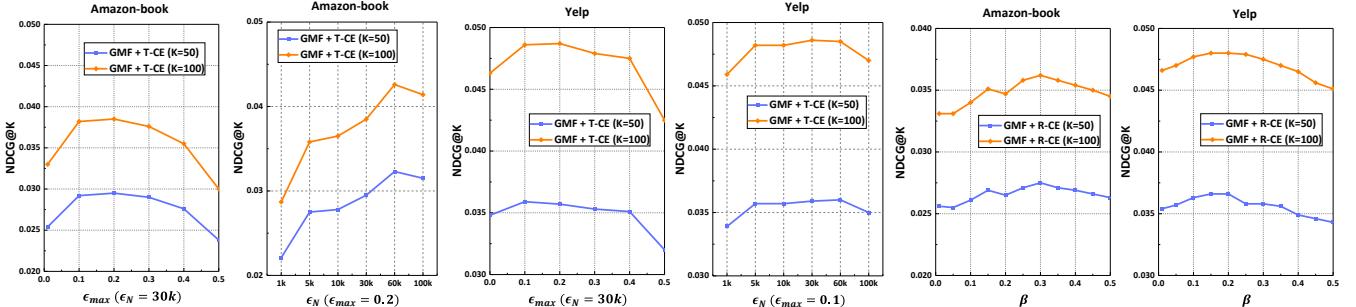


Fig. 10. Performance comparison of GMF trained with ADT on Yelp and Amazon-book *w.r.t.* different hyper-parameters, including ϵ_{max} , ϵ_N , and β .

TABLE 5

Performance of colliding inference over different user groups. The best results in each user group are highlighted in bold.

ADT	User ratio	Method	Recall50	Recall100	NDCG50	NDCG100
T-CE	Ratio < 0.1	GMF+Warm-up training	0.0720	0.0805	0.0201	0.0217
	Ratio < 0.1	GMF+Colliding inference	0.0763	0.0890	0.0257	0.0280
	Ratio < 0.2	GMF+Warm-up training	0.0650	0.0924	0.0204	0.0250
	Ratio < 0.2	GMF+Colliding inference	0.0626	0.0959	0.0211	0.0267
	Ratio < 0.3	GMF+Warm-up training	0.0673	0.0991	0.0231	0.0287
	Ratio < 0.3	GMF+Colliding inference	0.0674	0.0993	0.0232	0.0288
R-CE	Ratio < 0.1	GMF+Warm-up training	0.0466	0.0805	0.0109	0.0166
	Ratio < 0.1	GMF+Colliding inference	0.0678	0.0847	0.0165	0.0195
	Ratio < 0.2	GMF+Warm-up training	0.0571	0.0817	0.0183	0.0224
	Ratio < 0.2	GMF+Colliding inference	0.0625	0.0819	0.0201	0.0234
	Ratio < 0.3	GMF+Warm-up training	0.0619	0.0903	0.0207	0.0257
	Ratio < 0.3	GMF+Colliding inference	0.0625	0.0939	0.0212	0.0267

proposed finetuning, and warm-up training to the testing recommender models and reported the results of GMF in Table 4. Other results are omitted due to the similar trend. From Table 4, we could have the following observations:

- NMTR and NR achieve slightly better performance than GMF, which validates the effectiveness of using extra feedback. However, the results of NMTR and NR are inferior to those of the T-CE loss and R-CE loss, which might be attributed to the sparsity of extra feedback. Indeed, the clicks with satisfaction (*i.e.*, high ratings) is much fewer than the total number of true-positive interactions. Hence, NR will lose extensive positive training interactions while NMTR will inevitably use many false-positive interactions for training.
- Directly using extra feedback for finetuning or warm-up training outperforms the vanilla GMF but performs worse than NMTR or NR, showing the advantages of multi-task training [13] and the re-weighting technique [5].
- Finetuning and warm-up training significantly improve the performance of ADT over both R-CE and T-CE, further justifying the effectiveness of using extra feedback on denoising methods. Besides, warm-up training is more effective than finetuning. It validates the idea that warm-up training will expose the recommender model to true-positive interactions in the early training stage, and thus improve the model’s identification ability during ADT.

Effectiveness of Colliding Inference. Since colliding inference is designed for the users with sparse extra feedback, we leveraged the interaction numbers of extra feedback and implicit feedback to measure the sparsity and divided users into groups. Specifically, we defined *user ratio* as $\frac{\text{the number of extra feedback}}{\text{the number of implicit feedback}}$, and then selected the users with

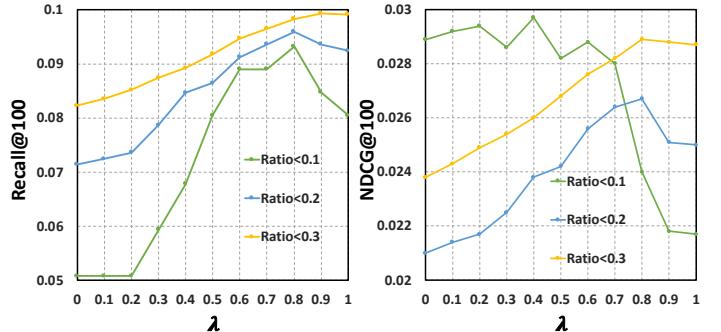


Fig. 11. Recall@100 and NDCG@100 of GMF+T-CE with colliding inference *w.r.t.* the hyper-parameter λ .

the ratios smaller than the threshold to conduct colliding inference. We chose warm-up training as an example to test colliding inference due to its superior performance, and the results of T-CE and R-CE equipped with warm-up training and colliding inference are summarized in Table 5.

From Table 5, we could observe that: 1) colliding inference significantly improves the performance of T-CE and R-CE with warm-up training over the users with low ratios. It validates that distilling colliding effect does utilize extra feedback to extract some useful knowledge from dense users, leading to the better rankings for sparse users; and 2) the performance margin is becoming smaller over the user groups from ratio < 0.1 to ratio < 0.3, implying that colliding inference is more effective for sparse users. For the users with high ratios, a larger proportion of extra feedback will prevent the causal effect of E^* on \hat{E} (in Figure 6) from being blocked during ADT.

Hyper-parameter Analysis. To study the effect of λ in Equation (8), we show the performance comparison over GMF+T-CE *w.r.t.* different λ values in Figure 11. The results of R-CE are omitted due to the similar trend. From the figure, we could see that: 1) utilizing the prediction scores of neighbors is effective because the best performance happens when $\lambda < 1$. It further validates the effectiveness of colliding inference. And 2) the peaks of the curves occur at $\lambda > 0.5$ in most cases, and move close to 1 as the ratio increases. This first shows that the personalization is essential although the neighbors are selected from the reliable user representation space. Besides, distilling the knowledge from neighbors is more important for the users with sparse extra feedback (*i.e.*, low ratio) because the causal effect of extra feedback on their prediction scores will be easily reduced during ADT with noisy implicit feedback.

6 CONCLUSION

In this work, we explored denoising implicit feedback for recommender learning. We found the negative effects of noisy implicit feedback, and proposed adaptive denoising training to reduce their impact. In particular, this work contributes two paradigms to formulate the loss functions: Truncated Loss and Reweighted Loss. Both paradigms are general and can be applied to different recommendation loss functions, neural recommender models, and optimizers. In this work, we applied the two paradigms on the widely used binary cross-entropy loss and conducted extensive experiments over three recommender models on three datasets, showing that the paradigms could effectively reduce the disturbance of noisy implicit feedback. In addition, we also developed three strategies: finetuning, warm-up training, and colliding inference, which can incorporate extra feedback to further improve the denoising performance of the two paradigms.

This work formulates the task of denoising implicit feedback for recommendation, which points to some new research directions. Specifically, adaptive denoising training is not specific to the recommendation task, and it can be widely used to denoise implicit interactions in other domains, such as Web search and question answering. Besides, it is interesting to explore how the proposed two paradigms perform on other loss functions. Lastly, we propose a novel causality-based inference strategy to improve the ranking performance over sparse users, which inspires the exploration of causal inference into the inference stage of more tasks.

REFERENCES

- [1] W. Wang, F. Feng, X. He, L. Nie, and T.-S. Chua, “Denoising implicit feedback for recommendation,” in *WSDM*. ACM, 2021, pp. 373–381.
- [2] H. Lu, M. Zhang, and S. Ma, “Between clicks and satisfaction: Study on multi-phase user preferences and satisfaction for online news reading,” in *SIGIR*. ACM, 2018, pp. 435–444.
- [3] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, “Neural collaborative filtering,” in *WWW*. IW3C2, 2017, pp. 173–182.
- [4] Y. Hu, Y. Koren, and C. Volinsky, “Collaborative filtering for implicit feedback datasets,” in *ICDM*. IEEE, 2008, pp. 263–272.
- [5] H. Wen, L. Yang, and D. Estrin, “Leveraging post-click feedback for content recommendations,” in *RecSys*. ACM, 2019, pp. 278–286.
- [6] W. Wang, F. Feng, X. He, H. Zhang, and T.-S. Chua, “Click can be cheating: Counterfactual recommendation for mitigating clickbait issue,” in *SIGIR*. ACM, 2021.
- [7] K. Hofmann, F. Behr, and F. Radlinski, “On caption bias in interleaving experiments,” in *CIKM*. ACM, 2012, p. 115–124.
- [8] R. Jagerman, H. Oosterhuis, and M. de Rijke, “To model or to intervene: A comparison of counterfactual and online learning to rank from user interactions,” in *SIGIR*. ACM, 2019, pp. 15–24.
- [9] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White, “Evaluating implicit measures to improve web search,” *TOIS*, vol. 23, no. 2, pp. 147–168, 2005.
- [10] Y. Kim, A. Hassan, R. W. White, and I. Zitouni, “Modeling dwell time to predict click-level satisfaction,” in *WSDM*. ACM, 2014, pp. 193–202.
- [11] B. Yang, S. Lee, S. Park, and S. goo Lee, “Exploiting various implicit feedback for collaborative filtering,” in *WWW*. IW3C2, 2012, p. 639–640.
- [12] X. Yi, L. Hong, E. Zhong, N. N. Liu, and S. Rajan, “Beyond clicks: dwell time for personalization,” in *RecSys*. ACM, 2014, p. 113–120.
- [13] C. Gao, X. He, D. Gan, X. Chen, F. Feng, Y. Li, T.-S. Chua, L. Yao, Y. Song, and D. Jin, “Learning to recommend with multiple cascading behaviors,” *TKDE*, 2019.
- [14] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, “Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels,” in *ICML*. PMLR, 2018, pp. 2304–2313.
- [15] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, “Co-teaching: Robust training of deep neural networks with extremely noisy labels,” in *NeurIPS*. MIT Press, 2018, pp. 8527–8537.
- [16] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” in *ICML*. ACM, 2009, p. 41–48.
- [17] J. Pearl, *Causality*. Cambridge university press, 2009.
- [18] X. Hu, K. Tang, C. Miao, X.-S. Hua, and H. Zhang, “Distilling causal effect of data in class-incremental learning,” in *CVPR*. IEEE, 2021.
- [19] Y. Wu, C. DuBois, A. X. Zheng, and M. Ester, “Collaborative denoising auto-encoders for top-n recommender systems,” in *WSDM*. ACM, 2016, pp. 153–162.
- [20] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang, and M. Sugiyama, “How does disagreement help generalization against label corruption?” in *ICML*. PMLR, 2019, pp. 7164–7173.
- [21] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio *et al.*, “A closer look at memorization in deep networks,” in *ICML*. PMLR, 2017, pp. 233–242.
- [22] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, “Bpr: Bayesian personalized ranking from implicit feedback,” in *UAI*. AUAI Press, 2009, pp. 452–461.
- [23] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *ICCV*. IEEE, 2017, pp. 2980–2988.
- [25] P. Yang, P. Zhao, V. W. Zheng, L. Ding, and X. Gao, “Robust asymmetric recommendation via min-max optimization,” in *SIGIR*. ACM, 2018, p. 1077–1080.
- [26] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*, 1st ed. Basic Books, Inc., 2018.
- [27] H. Jiang, W. Wang, Y. Wei, Z. Gao, Y. Wang, and L. Nie, “What aspect do you like: Multi-scale time-aware user interest modeling for micro-video recommendation,” in *MM*. ACM, 2020, pp. 3487–3495.
- [28] G. Lavee, N. Koenigstein, and O. Barkan, “When actions speak louder than clicks: A combined model of purchase probability and long-term customer satisfaction,” in *RecSys*. ACM, 2019, pp. 287–295.
- [29] Q. Zhao, S. Chang, F. M. Harper, and J. A. Konstan, “Gaze prediction for recommender systems,” in *RecSys*. ACM, 2016, pp. 131–138.
- [30] H. Lu, M. Zhang, W. Ma, C. Wang, F. xia, Y. Liu, L. Lin, and S. Ma, “Effects of user negative experience in mobile news streaming,” in *SIGIR*. ACM, 2019, pp. 705–714.
- [31] C. Liu, R. W. White, and S. Dumais, “Understanding web browsing behaviors through weibull analysis of dwell time,” in *SIGIR*. ACM, 2010, pp. 379–386.
- [32] A. Gunawardana and G. Shani, “Evaluating recommender systems,” in *Recommender Systems handbook*. Springer, 2015, pp. 265–308.

- [33] S. K. Lam and J. Riedl, "Shilling recommender systems for fun and profit," in *WWW*. IW3C2, 2004, pp. 393–402.
- [34] D. Shriver, S. G. Elbaum, M. B. Dwyer, and D. S. Rosenblum, "Evaluating recommender system stability with influence-guided fuzzing," in *AAAI*. AAAI Press, 2019, pp. 4934–4942.
- [35] S. Zhang, H. Yin, T. Chen, Q. V. N. Hung, Z. Huang, and L. Cui, "Gcn-based user representation learning for unifying robust recommendation and fraudster detection," in *SIGIR*. ACM, 2020, pp. 689–698.
- [36] F. Strub and J. Mary, "Collaborative filtering with stacked denoising autoencoders and sparse inputs," in *NeurIPS*. MIT Press, 2015.
- [37] D. Liang, R. G. Krishnan, M. D. Hoffman, and T. Jebara, "Variational autoencoders for collaborative filtering," in *WWW*. IW3C2, 2018, pp. 689–698.
- [38] J. Wang, L. Yu, W. Zhang, Y. Gong, Y. Xu, B. Wang, P. Zhang, and D. Zhang, "Irgan: A minimax game for unifying generative and discriminative information retrieval models," in *SIGIR*. ACM, 2017, pp. 515–524.
- [39] Y. Cao, X. Chen, L. Yao, X. Wang, and W. E. Zhang, "Adversarial attacks and detection on reinforcement learning-based interactive recommender systems," in *SIGIR*. ACM, 2020, pp. 1669–1672.
- [40] K. Christakopoulou and A. Banerjee, "Adversarial attacks on an oblivious recommender," in *RecSys*. ACM, 2019, pp. 322–330.
- [41] W. Yu and Z. Qin, "Sampler design for implicit feedback data by noisy-label robust learning," in *SIGIR*, 2020, pp. 861–870.
- [42] Y. Niu, K. Tang, H. Zhang, Z. Lu, X.-S. Hua, and J.-R. Wen, "Counterfactual vqa: A cause-effect look at language bias," in *CVPR*. IEEE, 2021, pp. 12700–12710.
- [43] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *CVPR*. IEEE, 2020, pp. 3716–3725.
- [44] F. Feng, J. Zhang, X. He, H. Zhang, and T. Chua, "Empowering language understanding with counterfactual reasoning," in *ACL/IJCNLP*, ser. Findings of ACL. ACL, 2021, pp. 2226–2236.
- [45] F. Feng, W. Huang, X. He, X. Xin, Q. Wang, and T.-S. Chua, "Should graph convolution trust neighbors? a simple causal inference method," in *SIGIR*. ACM, 2021, pp. 1208–1218.
- [46] Q. Ai, K. Bi, C. Luo, J. Guo, and W. B. Croft, "Unbiased learning to rank with unbiased propensity estimation," in *SIGIR*. ACM, 2018, pp. 385–394.
- [47] S. Bonner and F. Vasile, "Causal embeddings for recommendation," in *RecSys*. ACM, 2018, pp. 104–112.
- [48] D. Liang, L. Charlin, J. McInerney, and D. M. Blei, "Modeling user exposure in recommendation," in *WWW*. ACM, 2016, pp. 951–961.
- [49] X. Wang, R. Zhang, Y. Sun, and J. Qi, "Doubly robust joint learning for recommendation on data missing not at random," in *ICML*. PMLR, 2019, pp. 6638–6647.
- [50] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 04 1983.
- [51] M. Morik, A. Singh, J. Hong, and T. Joachims, "Controlling fairness and bias in dynamic learning-to-rank," in *SIGIR*. ACM, 2020, pp. 429–438.
- [52] H. Abdollahpouri, R. Burke, and B. Mobasher, "Managing popularity bias in recommender systems with personalized re-ranking," in *FLAIRS*. AAAI Press, 2019.
- [53] K. Christakopoulou, M. Traverse, T. Potter, E. Marriott, D. Li, C. Haulk, E. H. Chi, and M. Chen, "Deconfounding user satisfaction estimation from response rate bias," in *RecSys*. ACM, 2020, pp. 450–455.
- [54] Y. Saito, S. Yaginuma, Y. Nishino, H. Sakata, and K. Nakata, "Unbiased recommender learning from missing-not-at-random implicit feedback," in *WSDM*. ACM, 2020, pp. 501–509.
- [55] D. Liu, P. Cheng, Z. Dong, X. He, W. Pan, and Z. Ming, "A general knowledge distillation framework for counterfactual recommendation via uniform data," in *SIGIR*. ACM, 2020, pp. 831–840.
- [56] W. Wang, F. Feng, X. He, X. Wang, and T. Chua, "Deconfounded recommendation for alleviating bias amplification," in *KDD*. ACM, 2021, pp. 1717–1725.
- [57] J. A. Gulla, L. Zhang, P. Liu, O. Özgöbek, and X. Su, "The addressa dataset for news recommendation," in *WI*. ACM, 2017, p. 1042–1048.
- [58] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *WWW*. IW3C2, 2016, p. 507–517.
- [59] X. Wang, X. He, M. Wang, F. Feng, and T. Chua, "Neural graph collaborative filtering," in *SIGIR*. ACM, 2019, pp. 165–174.
- [60] Y. Chen, B. Chen, X. He, C. Gao, Y. Li, J.-G. Lou, and Y. Wang, "lambdaopt: Learn to regularize recommender models in finer levels," in *KDD*. ACM, 2019, pp. 978–986.