



# Adversarial Collaborative Filtering for Free

Huiyuan Chen  
hchen@visa.com  
Visa Research  
Palo Alto, CA, USA

Xiaoting Li  
xiaotili@visa.com  
Visa Research  
Palo Alto, CA, USA

Vivian Lai  
viv.lai@visa.com  
Visa Research  
Palo Alto, CA, USA

Chin-Chia Michael Yeh  
miyeh@visa.com  
Visa Research  
Palo Alto, CA, USA

Yujie Fan  
Yan Zheng  
yazheng@visa.com  
Visa Research  
Palo Alto, CA, USA

Mahashweta Das  
Hao Yang  
haoyang@visa.com  
Visa Research  
Palo Alto, CA, USA

## ABSTRACT

Collaborative Filtering (CF) has been successfully used to help users discover the items of interest. Nevertheless, existing CF methods suffer from noisy data issue, which negatively impacts the quality of recommendation. To tackle this problem, many prior studies leverage adversarial learning to regularize the representations of users/items, which improves both generalizability and robustness. Those methods often learn adversarial perturbations and model parameters under min-max optimization framework. However, there still have two major drawbacks: 1) Existing methods lack theoretical guarantees of why adding perturbations improve the model generalizability and robustness; 2) Solving min-max optimization is time-consuming. In addition to updating the model parameters, each iteration requires additional computations to update the perturbations, making them not scalable for industry-scale datasets.

In this paper, we present Sharpness-aware Collaborative Filtering (SharpCF), a simple yet effective method that conducts adversarial training without extra computational cost over the base optimizer. To achieve this goal, we first revisit the existing adversarial collaborative filtering and discuss its connection with recent Sharpness-aware Minimization. This analysis shows that adversarial training actually seeks model parameters that lie in neighborhoods around the optimal model parameters having uniformly low loss values, resulting in better generalizability. To reduce the computational overhead, SharpCF introduces a novel trajectory loss to measure the alignment between current weights and past weights. Experimental results on real-world datasets demonstrate that our SharpCF achieves superior performance with almost zero additional computational cost comparing to adversarial training.

## CCS CONCEPTS

• Information systems → Collaborative filtering; • Computing methodologies → Machine learning.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

RecSys '23, September 18–22, 2023, Singapore, Singapore

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 979-8-4007-0241-9/23/09...\$15.00  
<https://doi.org/10.1145/3604915.3608771>

## KEYWORDS

Collaborative filtering, Adversarial Training, Sharpness-aware Minimization, Loss Landscape Visualization, Generalization

### ACM Reference Format:

Huiyuan Chen, Xiaoting Li, Vivian Lai, Chin-Chia Michael Yeh, Yujie Fan, Yan Zheng, Mahashweta Das, and Hao Yang. 2023. Adversarial Collaborative Filtering for Free. In *Seventeenth ACM Conference on Recommender Systems (RecSys '23)*, September 18–22, 2023, Singapore, Singapore. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3604915.3608771>

## 1 INTRODUCTION

Recommender systems have become increasingly popular due to their ability to filter information and provide personalized recommendations to users on the web [8, 19, 25]. One of the most prominent techniques used in recommender systems is Collaborative Filtering (CF) [16, 18]. CF considers users' historical interactions and assumes that users who have shared similar preferences in the past tend to make similar decisions in the near future. To achieve this goal, most CF methods learn an encoder to embed users and items into a shared space and then optimize an objective function to learn informative user and item representations. CF-based methods have been successfully deployed in industries due to their simplicity and effectiveness, such as Factorization Machines [13, 22] and Graph Neural Networks [14, 40].

However, many state-of-the-art collaborative filtering (CF) methods remain vulnerable to noisy data, and their performance can degrade significantly under unnoticeable perturbations [4, 5, 29, 34]. Indeed, real-world data is often noisy, where user preferences may not necessarily align with the interacted items. For instance, a significant portion of purchases may result in negative reviews or returns. Such false-positive interactions hinder a model from learning the actual user preferences, leading to low-quality recommendations [33, 36]. To tackle this problem, many prior studies leverage the adversarial learning principle to regularize the representations of users and items [10, 15, 28, 38, 41]. These methods employ a min-max optimization framework to alternatively learn adversarial perturbations and model parameters. For instance, in the APR model [15], adversarial training is applied to a Matrix Factorization model by directly adding adversarial perturbations to the embedding vectors of both users and items.

While existing adversarial collaborative filtering methods have shown promising results, there still have two major limitations: 1)

These methods lack theoretical guarantees on why the addition of adversarial perturbations improves the model's generalizability and robustness. Interpretability remains unclear as noisy data is inherently different from adversarial perturbations, which are carefully crafted modifications to the original data [12]; 2) The process of solving min-max optimization is time-consuming [37]. In each iteration, besides updating the model parameters (e.g., the outer minimization problem), additional computations are required to update the learnable perturbations (e.g., the inner maximization problem). This makes the existing approaches unsuitable for learning with large-scale datasets. Therefore, developing more efficient and interpretable adversarial collaborative filtering methods remains a significant challenge in recommender systems.

**Present Work.** To overcome the limitations outlined above, we propose Sharpness-aware Collaborative Filtering (SharpCF), a simple yet effective method that enables adversarial training without incurring extra computational costs over the base optimizer. To achieve this goal, we first revisit existing adversarial collaborative filtering techniques and discuss their connection with recent developed Sharpness-aware Minimization [2, 9, 11, 20]. Our analysis reveals that adversarial training actually seeks model parameters that lie in neighborhoods around the optimal model parameters with uniformly low loss values. In other words, the adversarial training favors *flat* minima rather than *sharp* minima, resulting in better model generalizability.

To reduce the computational overhead, our SharpCF introduces a novel trajectory loss that measures the alignment between current and past model states. We further demonstrate that this trajectory loss can prevent convergence to the sharp minima and thus tend to drive the model parameters to the flat region, similar to the goal of the adversarial training. Interestingly, unlike adversarial training methods that use a min-max optimization framework, our SharpCF can be trained using standard Stochastic Gradient Descent, which significantly reduces the time complexity. Experimental results on real-world datasets show that our SharpCF outperforms the state-of-the-art adversarial collaborative filtering methods with almost no additional computational cost.

We summarize our contributions as follows:

- We have revisited the existing adversarial collaborative filtering methods and established their connection to recent Sharpness-aware Minimization. Through this, we have unveiled that adversarial training tends to favor flat minima over sharp ones, which results in better generalizability.
- We propose Sharpness-aware Collaborative Filtering (SharpCF) to introduce a novel trajectory loss that measures the alignment between current and past model states. Therefore, our SharpCF is able to avoid the need to solve a min-max optimization problem, and enables adversarial training without adding extra computational costs.
- Experimental results demonstrate that SharpCF outperforms the existing collaborative filtering methods. Specifically, our SharpCF consistently outperforms the BPR with an average improvement of 18.1% and an average improvement of 6.82% over the APR. Additionally, in terms of time complexity, our SharpCF is comparable to the BPR and has the same training speed as the BPR and is 2× faster than the APR.

## 2 RELATED WORK

In this section, we briefly review the related work on Collaborative Filtering and Adversarial Training. We also highlight the differences between the existing efforts and our proposed method.

### 2.1 Collaborative Filtering

Collaborative filtering (CF) is a widely used technique in recommender systems, which plays an essential role in addressing the information overload problem for users [26]. The core idea of CF is that users tend to have similar preferences and hence, their opinions and behavior can be utilized to make recommendations in the near future. One of the primary methods for CF is the Matrix Factorization, which learns the latent user and item representations by factorizing the observed interaction matrix [19, 25]. The predicted score of an unobserved user-item pair can be then derived by the similarity between the user and item representations, typically measured by the dot product.

Inspired by the success of deep neural networks, neural CF models have been proposed to learn more powerful user/item representations. For example, The Wide&Deep recommender model [7] combines linear models and deep neural networks to capture both memorization and generalization. He et al. [16] propose a neural collaborative filtering model that replaces the dot product with a neural network architecture to model the user-item interactions. Another popular model is DeepFM [13], which combines factorization machines with neural networks to learn both low- and high-order feature interactions. XdeepFM [22] extends DeepFM by introducing a novel cross network to capture more complex feature interactions. Recently, xLightFM [17] greatly reduces the memory footprint of factorization machines via quantization techniques. In addition to factorization machines, graph neural networks have also attracted increasing attention recently, and a number of graph-based CF models have been proposed, such as LightGCN [14]. These models have shown great success in recommendation tasks and have been applied in various domains [8, 19, 25, 32, 35, 39].

However, noisy data (e.g., false-positive feedback) can have a significant impact on the performance of recommender systems. Generally, noisy data can cause a bias towards popular items, which is very challenging for accurate and diverse recommendations [4, 29, 33, 34, 36]. To collect the data, prior studies consider incorporating more user feedback (e.g., multi-type clicks, user textual comments) during training. Wen et al. [36] suggest that training the recommender models uses three kinds of scenarios: "click-complete", "click-skip", and "non-click" ones, where last two kinds of items are both treated as negative samples. Julian et al. [24] combine the user latent factors with textual review to justify users' ratings. Nevertheless, additional feedback might be expensive or unavailable in many scenarios. Researchers have alternatively attempted to explore the adversarial training to mitigate the negative impact of noisy feedback without additional information [3, 10, 15, 38, 41].

### 2.2 Adversarial Training

Adversarial training has emerged as a promising technique to mitigate the negative effects of noisy data in recommender systems [3, 6, 10, 15, 38, 41]. Adversarial training involves training the model on a combination of clean and adversarial data, where

the adversarial data is generated by adding the worst-case perturbations to the input data [6, 12, 15]. The goal is to encourage the model to be robust to these perturbations, which in turn improves its ability to generalize to unseen data. For example, In APR [15], adversarial training is used to perturb the embedding vectors of users and items in a Matrix Factorization model, resulting in a more robust and accurate recommendation system. NMRN [31] applies adversarial training to discover both long-term stable interests and short-term dynamic behaviors in streaming recommender models. ACAE [41] combines collaborative auto-encoder with adversarial training, which outperforms highly competitive state-of-the-art recommendation methods. ATF [3] reaps the benefits of adversarial training to improve the context-aware recommendations. Typically, these methods employ a min-max optimization framework to alternatively learn adversarial perturbations and model parameters. However, solving min-max frameworks incurs a two-fold computational overhead of the given base optimizer, making it not scalable to large datasets in practice.

Recently, several studies have made attempts to develop the *fast* version of adversarial training [1, 27, 37]. One example is FreeAT [27], which removes the overhead cost of generating adversarial examples by reusing the gradient information computed during model training. GradAlign [1] explicitly maximizes the gradient alignment inside the perturbation set. Despite their bleeding edge performance, those frameworks rely on a series of heuristics-based strategies, such as good initialization, large step size, and cyclic learning rate schedule. Recent studies indicate that these *free* strategies often lack of stability and are prone to catastrophic overfitting issue [42]. Furthermore, the majority of fast adversarial training algorithms are intended for generating continuous adversarial examples (e.g., image pixels), rendering them unsuitable for information retrieval purposes due to their discrete space [15, 30].

In contrast, we propose a novel trajectory loss function that measures the alignment between current and past model states to reduce the complexity, which avoids solving min-max optimization. And we show that our trajectory loss function is connected to the recent Sharpness-aware Minimization [2, 11, 20], which tends to seek flat minima, leading to better model generalization.

### 3 PRELIMINARIES

In this section, we first formulate the basic problem of collaborative filtering. Then, we briefly revisit the Bayesian Personalized Ranking (BPR) [25] and the Adversarial Personalized Ranking (APR) [15] for implicit recommendations.

#### 3.1 Collaborative Filtering

In this paper, we focus on the task of learning user preferences from implicit feedback. Specifically, the users' behavior data, e.g., click, view, comment, purchase, etc., consists of a set of users  $\mathcal{U} = \{u\}$  and items  $\mathcal{I} = \{i\}$ , such that the set  $\mathcal{I}_u^+$  represents the items that user  $u$  has interacted with before, whereas  $\mathcal{I}_u^- = \mathcal{I} - \mathcal{I}_u^+$  represents unobserved items. In general, the unobserved interactions are not necessarily negative, but rather, the user may simply be unaware of them. The goal of collaborative filtering is to estimate the user preference towards items.

The majority of CF methods learn each user and item into a low-dimensional latent space. Matrix Factorization is widely acknowledged as the fundamental and most effective model in recommendation. Specifically, the representations of a user  $u$  and an item  $i$  can be obtained via embedding lookup tables:

$$\mathbf{e}_u = \text{lookup}(u), \quad \mathbf{e}_i = \text{lookup}(i), \quad (1)$$

where  $u$  and  $i$  denote the IDs of user and item;  $\mathbf{e}_u \in \mathbb{R}^d$  and  $\mathbf{e}_i \in \mathbb{R}^d$  are the embeddings of user  $u$  and item  $i$ , respectively, and  $d$  is the embedding size. These embeddings are intended to capture and retain the initial characteristics of both items and users, which can be updated during training. Then, the predicted score is defined as the similarity between the user and item representations via dot product:

$$\hat{y}_{ui} = \mathbf{e}_u^T \mathbf{e}_i. \quad (2)$$

As for the learning objective, we next introduce the Bayesian Personalized Ranking (BPR) loss [25] and the Adversarial Personalized Ranking (APR) [15] loss to train the model.

#### 3.2 Bayesian Personalized Ranking

Bayesian Personalized Ranking (BPR) [25] is a popular and effective pairwise method used in learning-to-rank for recommender systems. Its primary goal is to optimize recommender models towards personalized ranking. BPR is particularly suitable for learning from implicit feedback, where the observed interactions are often incomplete and the unobserved ones are assumed to be ranked lower.

Unlike pointwise methods, which focus on optimizing each model prediction towards a predefined ground truth, BPR prioritizes maximizing the margin between an observed interaction and its unobserved counterparts. This margin-based approach allows BPR to perform well even when the number of negative samples is much larger than the number of positive samples. Formally, BPR aims to minimize the following objective function:

$$\mathcal{L}_{\text{BPR}}(\Theta) = - \sum_{(u,i,j) \in \mathbb{O}} \ln \sigma(\hat{y}_{ui} - \hat{y}_{uj}), \quad (3)$$

where  $\mathbb{O} = \{(u, i, j) \mid u \in \mathcal{U} \wedge i \in \mathcal{I}_u^+ \wedge j \in \mathcal{I}_u^-\}$  denotes the pairwise training data,  $\sigma(\cdot)$  is the sigmoid function, and  $\Theta$  denotes model parameters. Typically, standard Stochastic Gradient Descent (SGD) is used for its optimization. Once the parameters are obtained, a personalized ranking list for a user  $u$  can be generated by evaluating the value of  $\hat{y}_{ui}(\Theta)$  over all unobserved items  $i \in \mathcal{I}_u^-$ .

However, training Matrix Factorization with the BPR loss is not robust as it is vulnerable to adversarial perturbations on the model parameters. Therefore, the model can easily learn a complex function, which may result in overfitting on the training data and poor generalization on unseen data.

#### 3.3 Adversarial Personalized Ranking

To tackle above issue, Adversarial Personalized Ranking (APR) [15] intends to create an objective function that optimizes the recommender model for both personalized ranking and resistance to adversarial perturbations. To achieve this, APR injects adversarial perturbations  $\Delta$  on latent factors to quantify the loss of the model under perturbations on its parameters  $\Theta$ :

$$\hat{y}_{ui}(\Theta + \Delta) = (\mathbf{e}_u + \Delta_u)^T (\mathbf{e}_i + \Delta_i), \quad (4)$$

where the perturbation vectors  $\Delta$  are coupled with their corresponding latent factors, *i.e.*,  $\Delta_u \in \mathbb{R}^d$  denotes the perturbation vector for user latent vector  $\mathbf{e}_u$ . Adversarial perturbations aim to cause the largest influence on the model and are also known as worst-case perturbations. Therefore, APR maximizes the BPR loss to find the optimal adversarial perturbations:

$$\begin{aligned} \Delta_{\text{adv}} = \arg \max_{\Delta} \mathcal{L}_{\text{BPR}}(\hat{\Theta} + \Delta), \\ \text{s.t. } \|\Delta\|_2 \leq \rho, \end{aligned} \quad (5)$$

where  $\rho$  controls the magnitude of adversarial perturbations,  $\|\cdot\|_2$  denotes the  $L_2$  norm, and  $\hat{\Theta}$  is the intermediate model parameters. To this end, APR designs a new objective function that is both reasonable for personalized ranking and robust to adversarial perturbations. Formally, it minimizes the adversarial BPR loss by jointly integrating Eq. (3) and Eq. (5) as follow [10, 15]:

$$\begin{aligned} \mathcal{L}_{\text{APR}}(\Theta) = \mathcal{L}_{\text{BPR}}(\Theta) + \alpha \cdot \mathcal{L}_{\text{BPR}}(\Theta + \Delta_{\text{adv}}), \\ \text{where } \Delta_{\text{adv}} = \arg \max_{\Delta} \mathcal{L}_{\text{BPR}}(\hat{\Theta} + \Delta), \\ \|\Delta\|_2 \leq \rho \end{aligned} \quad (6)$$

where  $\alpha$  controls the impact of the adversarial perturbations on the model optimization. In the extreme case where  $\alpha = 0$ , the APR algorithm reduces to the original BPR framework as defined in Eq. (3). Therefore, APR can be considered a generalization of BPR that takes into account the robustness of the models.

The above min-max objective function in Eq. (6) can be expressed as playing a min-max game: the optimization of model parameters  $\Theta$  serves as the minimizing player, while adversarial perturbations  $\Delta$  act as the maximizing player. The two players alternate between playing this min-max game until convergence. Nevertheless, APR still lacks theoretical guarantees of why adding perturbations improves the model's generalizability. Additionally, solving min-max optimization is a time-consuming process. These shortcomings have motivated us to develop a new training method for personalized ranking. We next present our SharpCF that provides theoretical guarantees and requires nearly zero additional computational cost compared to the APR.

## 4 THE PROPOSED SHARPCF

In this section, we begin by simplifying the design of the APR loss function. Next, we establish a connection between the APR loss and recent research on Sharpness-aware Minimization [2, 11, 20]. Building on this insight, we put forward Sharpness-aware Collaborative Filtering (SharpCF), which includes a novel trajectory loss that measures the alignment between the current and past model states. Importantly, we demonstrate theoretically that our trajectory loss performs a similar role as adversarial training on improving model generalization.

### 4.1 Simplify APR

Since the intermediate variable  $\Delta$  maximizes the objective function that  $\Theta$  minimizes, the optimization problem in Eq. (6) can be expressed as:

$$\Theta^*, \Delta^* = \arg \min_{\Theta} \max_{\|\Delta\|_2 \leq \rho} \underbrace{\mathcal{L}_{\text{BPR}}(\Theta)}_{\text{the BPR loss}} + \alpha \cdot \underbrace{\mathcal{L}_{\text{BPR}}(\Theta + \Delta)}_{\text{the adversarial loss}}. \quad (7)$$

One can adopt the classical gradient descent-ascent algorithm for min-max optimization, which alternatively updates one variable while fixing the other one. That is

$$\begin{aligned} \Delta^{(t+1)} &\leftarrow \arg \max_{\|\Delta\|_2 \leq \rho} \mathcal{L}_{\text{BPR}}(\Theta^{(t)} + \Delta), \\ \Theta^{(t+1)} &\leftarrow \arg \min_{\Theta} \mathcal{L}_{\text{BPR}}(\Theta) + \alpha \cdot \mathcal{L}_{\text{BPR}}(\Theta + \Delta^{(t+1)}). \end{aligned} \quad (8)$$

Following [15], we adopt the Fast Gradient Sign Method to solve the inner maximization while apply the standard Stochastic Gradient Descent to optimize the outer minimization. Clearly, we have the following observations: 1) While updating  $\Delta$ , only the adversarial loss contributes to the gradient  $\nabla_{\Delta} \mathcal{L}_{\text{BPR}}(\Theta^{(t)} + \Delta)$ ; 2) While updating  $\Theta$ , the gradient of adversarial loss  $\nabla_{\Theta} \mathcal{L}_{\text{BPR}}(\Theta + \Delta^{(t+1)})$  already contains the gradient of BPR loss  $\nabla_{\Theta} \mathcal{L}_{\text{BPR}}(\Theta)$  as the  $\Delta^{(t+1)}$  is a constant. Given this information, a natural question arises here: is it necessary to include the original BPR loss  $\mathcal{L}_{\text{BPR}}(\Theta)$  in the min-max optimization?

To answer the above question, we conduct a series of experiments using four public benchmark datasets: MovieLens1M, Gowalla, Yelp2018, and Amazon-Book (The detailed data description can be found in Sec 5). In the experiments, we fix the embedding size  $d = 128$ , and the magnitude of adversarial perturbations  $\rho = 0.5$ . Then we vary the regularization parameter  $\alpha$  within the range of  $\{0, 1e-2, 1e-1, 1, 1e1, 1e2, 1e3, +\infty\}$ . We next compare the model training results using two different loss functions: Hybrid loss and Adversarial loss. The Hybrid loss is a combination of the BPR loss and the Adversarial loss with a range of  $\alpha$  values except for  $\{0, +\infty\}$ , while the Adversarial loss completely eliminates the impact of the BPR loss by setting  $\alpha = +\infty$ . Our comparison of these two methods provided valuable insights into their respective impact for adversarial training.

As depicted in Figure 1, choosing a non-zero value of  $\alpha$  consistently outperforms the baseline that sets  $\alpha = 0$ , *i.e.*, the model training only with the BPR loss. This means that the model gets benefits from the adversarial training. Gradual improvements are commonly observed as  $\alpha$  increases, particularly when  $\alpha$  is smaller than 100. However, future increases in  $\alpha$  do not significantly improve or decrease the performance when  $\alpha$  is larger than 100.

In the extreme case  $\alpha = +\infty$ , training with only the adversarial loss slightly improves performance, except for the Yelp dataset, but the difference between the Hybrid loss and the Adversarial loss is not statistically significant. Based on the above analysis, we can conclude that the model benefits from a larger value of  $\alpha$  and becomes rather insensitive when  $\alpha$  is sufficiently large. This implied that the Adversarial loss actually dominates the training, and the BPR loss can be safely removed without hurting the performance.

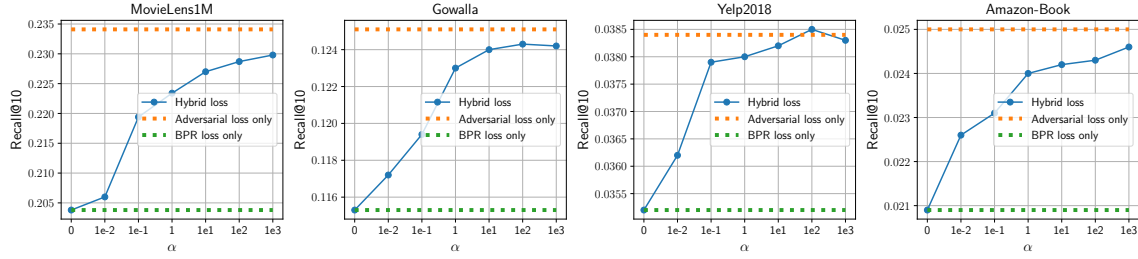
To this end, we can simplify the APR model in Eq. (6) into the following:

$$\min_{\Theta} \max_{\|\Delta\|_2 \leq \rho} \mathcal{L}_{\text{BPR}}(\Theta + \Delta). \quad (9)$$

We next connect the adversarial training to the recent Sharpness-aware Minimization [2, 11, 20].

### 4.2 Connect to Sharpness-aware Minimization

Understanding the generalization of adversarial training is critical as the training objective Eq. (9) has multiple local optima that can



**Figure 1: The performance of APR with different values of  $\alpha$ . The Hybrid loss is a combination of the BPR loss and the Adversarial loss with a range of  $\alpha$  values except for  $+\infty$ , while the Adversarial loss completely eliminates the impact of the BPR loss by setting  $\alpha = +\infty$ .**

perfectly fit the training data, but different optima lead to dramatically different generalization performance. To better understand the generalization of Eq. (9), we decompose it as:

$$\min_{\Theta} \mathcal{L}(\Theta) + \mathcal{R}(\Theta),$$

$$\text{where } \mathcal{R}(\Theta) = \max_{\|\Delta\|_2 \leq \rho} \mathcal{L}(\Theta + \Delta) - \mathcal{L}(\Theta), \quad (10)$$

where  $\mathcal{L}$  is short for  $\mathcal{L}_{\text{BPR}}$ . Interestingly, we observe that the Eq. (10) is the same as the Sharpness-aware Minimization [11], where the term  $\mathcal{R}(\Theta)$  captures the sharpness of  $\mathcal{L}$  at  $\Theta$  by measuring how quickly the training loss can be increased by moving  $\Theta$  to a nearby region:  $\{\Theta + \Delta \mid \|\Delta\|_2 \leq \rho\}$ .

Therefore, rather than seeking model parameters  $\Theta$  that have low training loss values in Eq. (3), adversarial training in Eq. (10) actually seeks out  $\Theta$  whose entire neighborhoods have uniformly low training loss values. In other words, adversarial training favors *flat* minima rather than *sharp* minima, resulting in better generalizability. One can further derive generalization bounds as:

**PROPOSITION 1.** *For any  $\rho > 0$ , let  $\mathcal{L}_{\mathcal{D}}$  be the expected loss and  $\mathcal{L}_{\mathcal{S}}$  be the training loss, where the training set  $\mathcal{S}$  is drawn from data distribution  $\mathcal{D}$  with i.i.d condition, then*

$$\mathcal{L}_{\mathcal{D}}(\Theta) \leq \max_{\|\Delta\|_2 \leq \rho} \mathcal{L}_{\mathcal{S}}(\Theta + \Delta) + h(\|\Theta\|_2^2 / \rho^2), \quad (11)$$

where  $h(\cdot) : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  is a strictly increasing function (e.g., the weight decay  $L_2$  norm).

The proof of the above inequality can be driven by using the PAC-Bayes theory [23] (more details can be seen in Theorem 2 in [11]). Above proposition provides a generalization upper bound to explain why the adversarial training can help improve generalization.

However, both Adversarial Training [15] and Sharpness-aware Minimization [11] adopt the gradient descent-ascent framework, which requires two forward and backward passes on each sample in a batch, namely a gradient ascent step to update the perturbation  $\Delta$  and a gradient descent step to update the current model  $\Theta$ . This doubling of computation time compared to the base optimizer makes them unsuitable for scaling to industry-scale datasets.

### 4.3 Our SharpCF

To overcome the computational bottleneck, we propose Sharpness-aware Collaborative Filtering (SharpCF), which includes a novel

trajectory loss that measures the alignment between the current and past model states. We begin by approximate the inner maximization problem in Eq. (10) via a first-order Taylor expansion of  $\Delta$  as [11]:

$$\begin{aligned} \hat{\Delta} &= \arg \max_{\|\Delta\|_2 \leq \rho} \mathcal{L}(\Theta + \Delta) \\ &\approx \arg \max_{\|\Delta\|_2 \leq \rho} \mathcal{L}(\Theta) + \Delta^T \nabla_{\Theta} \mathcal{L}(\Theta) \\ &= \arg \max_{\|\Delta\|_2 \leq \rho} \Delta^T \nabla_{\Theta} \mathcal{L}(\Theta) \\ &= \rho \frac{\nabla_{\Theta} \mathcal{L}(\Theta)}{\|\nabla_{\Theta} \mathcal{L}(\Theta)\|_2}. \end{aligned} \quad (12)$$

With the approximated  $\hat{\Delta}$ , for each batch size  $\mathbb{B}$ , we can rewrite its sharpness term  $\mathcal{R}_{\mathbb{B}}(\Theta)$  as:

$$\begin{aligned} \mathcal{R}_{\mathbb{B}}(\Theta) &= \max_{\|\Delta\|_2 \leq \rho} \mathcal{L}_{\mathbb{B}}(\Theta + \Delta) - \mathcal{L}_{\mathbb{B}}(\Theta) \\ &\approx \max_{\|\Delta\|_2 \leq \rho} \mathcal{L}_{\mathbb{B}}(\Theta) + \hat{\Delta}^T \nabla_{\Theta} \mathcal{L}_{\mathbb{B}}(\Theta) - \mathcal{L}_{\mathbb{B}}(\Theta) \\ &= \rho \frac{\nabla_{\Theta} \mathcal{L}_{\mathbb{B}}(\Theta)^T}{\|\nabla_{\Theta} \mathcal{L}_{\mathbb{B}}(\Theta)\|_2} \nabla_{\Theta} \mathcal{L}_{\mathbb{B}}(\Theta) \\ &= \rho \|\nabla_{\Theta} \mathcal{L}_{\mathbb{B}}(\Theta)\|_2. \end{aligned} \quad (13)$$

This remarks that minimizing the sharpness term  $\mathcal{R}_{\mathbb{B}}(\Theta)$  is equivalent to minimizing the  $l_2$ -norm of the gradient  $\nabla_{\Theta} \mathcal{L}_{\mathbb{B}}(\Theta)$ , which is the same gradient used to minimize the vanilla loss  $\mathcal{L}_{\mathbb{B}}(\Theta)$ . However, directly optimizing the gradient norm involves second-order derivative information (e.g., Hessian), which is computationally demanding. To overcome this challenge, we next introduce a novel trajectory loss that measures the alignment between current and past model states [9]. Thus, the trajectory loss provides a way to optimize the gradient norm implicitly without requiring the computation of second-order derivatives.

For current iteration  $t$ , we denote its pass trajectory of the model weights as  $\Theta = \{\Theta_1, \dots, \Theta_{t-1}\}$ , and  $\Theta_t$  represents the current weights in the  $t$ -th iteration. Recall that standard SGD updates the weights as  $\Theta_{t+1} = \Theta_t - \eta_t \nabla_{\Theta_t} \mathcal{L}_{\mathbb{B}_t}(\Theta_t)$ . For current batch  $\mathbb{B}_t$  and model state  $\Theta_t$ , as  $\mathcal{R}_{\mathbb{B}_t}(\Theta_t)$  is always non-negative, and thus we

have:

$$\begin{aligned}
\arg \min_{\Theta_t} \mathcal{R}_{\mathbb{B}_t}(\Theta_t) &\Leftrightarrow \arg \min_{\Theta_t} \eta_t \cos(\Phi_t) \mathcal{R}_{\mathbb{B}_t}(\Theta_t) \mathcal{R}_{\mathbb{B}_t}(\Theta_t) \\
&\Leftrightarrow \arg \min_{\Theta_t} \eta_t \cos(\Phi_t) \mathcal{R}_{\mathbb{B}_t}(\Theta_t) \mathcal{R}_{\mathbb{B}_t}(\Theta_t) \\
&+ \sum_{i < t} \eta_i \cos(\Phi_i) \mathcal{R}_{\mathbb{B}_t}(\Theta_i) \mathcal{R}_{\mathbb{B}_t}(\Theta_i) \\
&= \sum_{i=1}^t \arg \min_{\Theta_t} \eta_i \cos(\Phi_i) \mathcal{R}_{\mathbb{B}_t}(\Theta_i) \mathcal{R}_{\mathbb{B}_t}(\Theta_i) \\
&= \mathbb{E}_{\Theta_i \sim \text{Unif}(\Theta)} [\eta_i \cos(\Phi_i) \mathcal{R}_{\mathbb{B}_t}(\Theta_i) \mathcal{R}_{\mathbb{B}_t}(\Theta_i)], \tag{14}
\end{aligned}$$

where  $\Theta_i \sim \text{Unif}(\Theta)$  denotes that  $\Theta_i$  is uniformly distributed in the set  $\Theta$ , and  $\mathbb{E}[\cdot]$  denotes the expectation;  $\Phi_i$  is the angle between the gradients that are computed using the mini-batches  $\mathbb{B}_i$  and  $\mathbb{B}_t$  and  $\cos(\Phi_i) = 1$ . Note that  $\eta_i \cos(\Phi_i) \mathcal{R}_{\mathbb{B}_t}(\Theta_i) \mathcal{R}_{\mathbb{B}_t}(\Theta_i)$  becomes a constant with respect to variable  $\Theta_t$  for all  $i \neq t$ . By substituting Eq. (13) into Eq (14), we have following:

$$\begin{aligned}
\arg \min_{\Theta_t} \mathcal{R}_{\mathbb{B}_t}(\Theta_t) &\Leftrightarrow \mathbb{E}_{\Theta_i \sim \text{Unif}(\Theta)} [\eta_i \cos(\Phi_i) \mathcal{R}_{\mathbb{B}_t}(\Theta_i) \mathcal{R}_{\mathbb{B}_t}(\Theta_i)] \\
&= \mathbb{E}_{\Theta_i \sim \text{Unif}(\Theta)} [\rho^2 \eta_i \cos(\Phi_i) \|\nabla_{\Theta_i} \mathcal{L}_{\mathbb{B}_t}(\Theta_i)\|_2 \|\nabla_{\Theta_i} \mathcal{L}_{\mathbb{B}_t}(\Theta_i)\|_2] \\
&= \mathbb{E}_{\Theta_i \sim \text{Unif}(\Theta)} [\rho^2 \eta_i \nabla_{\Theta_i} \mathcal{L}_{\mathbb{B}_t}(\Theta_i)^T \nabla_{\Theta_i} \mathcal{L}_{\mathbb{B}_t}(\Theta_i)] \\
&\approx \rho^2 \mathbb{E}_{\Theta_i \sim \text{Unif}(\Theta)} [\mathcal{L}_{\mathbb{B}_t}(\Theta_i) - \mathcal{L}_{\mathbb{B}_t}(\Theta_{i+1})] \\
&= \frac{\rho^2}{t-1} [\mathcal{L}_{\mathbb{B}_t}(\Theta_1) - \mathcal{L}_{\mathbb{B}_t}(\Theta_2) + \dots + \mathcal{L}_{\mathbb{B}_t}(\Theta_{t-1}) - \mathcal{L}_{\mathbb{B}_t}(\Theta_t)] \\
&= \frac{\rho^2}{t-1} [\mathcal{L}_{\mathbb{B}_t}(\Theta_1) - \mathcal{L}_{\mathbb{B}_t}(\Theta_t)]. \tag{15}
\end{aligned}$$

We remark that minimizing the sharpness term  $\mathcal{R}_{\mathbb{B}_t}(\Theta_t)$  is equivalent to minimizing the loss difference  $\mathcal{L}_{\mathbb{B}_t}(\Theta_1) - \mathcal{L}_{\mathbb{B}_t}(\Theta_t)$ . To avoid the second term  $-\frac{\rho^2}{t-1} \mathcal{L}_{\mathbb{B}_t}(\Theta_t)$  cancel out a partial of the vanilla loss, we replace  $\mathcal{L}_{\mathbb{B}_t}(\Theta_1) - \mathcal{L}_{\mathbb{B}_t}(\Theta_t)$  with its  $l_2$  norm. In addition, as the model parameters  $\Theta_1$  contains less information, for current epoch  $e$ , we only track the loss trajectory in the past  $E$  epochs:  $\{\Theta_{e-E}, \dots, \Theta_e\}$ .

To this end, the mini-batch loss of our SharpCF at  $e$ -th epoch is defined as follow:

$$\mathcal{L}_{\text{SharpCF}}(\Theta_e) = \underbrace{\mathcal{L}_{\mathbb{B}}(\Theta_e)}_{\text{the BPR loss}} + \frac{\lambda}{|\mathbb{B}|} \underbrace{\|\mathcal{L}_{\mathbb{B}}(\Theta_e) - \mathcal{L}_{\mathbb{B}}(\Theta_{e-E})\|_2^2}_{\text{the trajectory loss}}, \tag{16}$$

where  $\lambda$  is a regularized parameter that is used to balance the vanilla loss with the trajectory loss. Meanwhile,  $\mathcal{L}_{\mathbb{B}}(\Theta_{e-E})$  represents the loss of the batch  $\mathbb{B}$  in  $E$  epochs ago, which is being recorded during training.

Nonetheless, both  $\mathcal{L}_{\mathbb{B}}(\Theta_{e-E})$  and  $\mathcal{L}_{\mathbb{B}}(\Theta_e)$  involve the utilization of the negative sampling technique to generate the negative samples as described in Eq. (3). In practice, the same batch of positive user-item pairs can be utilized for  $\mathcal{L}_{\mathbb{B}}(\Theta_{e-E})$  and  $\mathcal{L}_{\mathbb{B}}(\Theta_e)$  through an indexing method, but distinct negative pairs are generated due to the random negative sampling approach. To address this issue, we initially train the model with BPR, and subsequently monitor the trajectory after a predefined epoch  $E_{\text{start}}$ .

After the pretrained epoch  $E_{\text{start}}$ , the user/item representations become more reliable and stable, and the BPR loss enforces a large margin between positive pairs and negative pairs, i.e.,  $\hat{y}_{ui} \gg \hat{y}_{uj}$ , for  $i \in I_u^+ \wedge j \in I_u^-$ . As such, we empirically find that it is sufficient to only track the predicted scores of positive pairs in the mini-batch  $\hat{Y}_{\mathbb{B}}(\Theta_{e-E})$  and  $\hat{Y}_{\mathbb{B}}(\Theta_e)$  as described in Eq. (2). By doing so, our empirical loss function of Eq. (16), which is simple yet effective, can be expressed as follows:

$$\mathcal{L}_{\text{SharpCF}}^{\text{emp}}(\Theta_e) = \mathcal{L}_{\mathbb{B}}(\Theta_e) + \frac{\lambda}{|\mathbb{B}|} \|\hat{Y}_{\mathbb{B}}(\Theta_e) - \hat{Y}_{\mathbb{B}}(\Theta_{e-E})\|_2^2, \tag{17}$$

Essentially, our loss is applied to slow down the rate of change of the training loss to prevent convergence to sharp local minima. Algorithm 1 summarizes the overall training of SharpCF.

---

**Algorithm 1:** SharpCF

---

**Input:** The training data  $\mathbb{O}$ , the regularizer  $\lambda$ , the number of epoch  $\bar{E}$ , the started trajectory loss epoch  $E_{\text{start}}$ , the epoch window  $E$ .

- 1 Initialize model parameters  $\Theta$ ;
- 2 **for**  $e \leftarrow 1$  **to**  $\bar{E}$  **do**
- 3     **for** each mini-batch  $\mathbb{B} \subset \mathbb{O}$  **do**
- 4         Compute the predicted scores for positive pairs  $\hat{Y}_{\mathbb{B}}(\Theta_e)$ ;
- 5         Compute the predicted scores for negative pairs  $\hat{N}_{\mathbb{B}}(\Theta_e)$ ;
- 6         Compute the BPR loss  $\mathcal{L}_{\mathbb{B}}(\Theta_e)$  based on  $\hat{Y}_{\mathbb{B}}(\Theta_e)$  and  $\hat{N}_{\mathbb{B}}(\Theta_e)$ ;
- 7         Cache the loss  $\hat{Y}_{\mathbb{B}}(\Theta_{e-E})$  in  $E$  epochs ago for same batch  $\mathbb{B}$ ;
- 8         **if**  $e > E_{\text{start}}$  **then**
- 9              $\mathcal{L}_{\text{SharpCF}}^{\text{emp}}(\Theta_e) = \mathcal{L}_{\mathbb{B}}(\Theta_e) + \frac{\lambda}{|\mathbb{B}|} \|\hat{Y}_{\mathbb{B}}(\Theta_e) - \hat{Y}_{\mathbb{B}}(\Theta_{e-E})\|_2^2$ ;
- 10         **else**
- 11              $\mathcal{L}_{\text{SharpCF}}^{\text{emp}}(\Theta_e) = \mathcal{L}_{\mathbb{B}}(\Theta_e)$ ;
- 12         **end**
- 13         Update the model weights  $\Theta_e$  by using SGD;
- 14     **end**
- 15 **end**

**Output:** The optimal model parameters  $\Theta^*$ .

---

**Model Complexity.** For time complexity, Unlike adversarial training [11, 15] that requires two forward and backward passes to alternatively update the auxiliary variable  $\Delta$  and model parameters  $\Theta$ , Our SharpCF can update  $\Theta$  by using standard SGD. This makes the time complexity of SharpCF similar to that of the original BPR model [25], with almost no additional computational cost required to record the loss trajectory.

In terms of memory complexity, APR [15] requires  $O((|\mathcal{U}| + |\mathcal{I}|)d)$ , where  $|\mathcal{U}|$  and  $|\mathcal{I}|$  represent the number of users and items, respectively, and  $d$  is the embedding size. In contrast, our SharpCF needs extra memory to cache the loss trajectory  $O(E|\mathbb{O}|)$ , where  $E$  is the window size of the trajectory and  $|\mathbb{O}|$  is the number of training

**Table 1: Dataset statistics.**

| Dataset     | #User  | #Items | #Interactions | Density |
|-------------|--------|--------|---------------|---------|
| MovieLens1M | 6,040  | 3,900  | 1,000,209     | 4.190%  |
| Gowalla     | 29,858 | 40,981 | 1,027,370     | 0.084%  |
| Yelp2018    | 31,668 | 38,048 | 1,561,406     | 0.130%  |
| Amazon-Book | 52,643 | 91,599 | 2,984,108     | 0.062%  |

data. However, since the window size of the trajectory  $E$  is typically a small constant like 3 or 5, the memory complexity of SharpCF is negligible compared to BPR [25]. Overall, our proposed SharpCF method offers a computationally efficient and memory-friendly alternative for collaborative filtering tasks.

## 5 EXPERIMENTS

In this section, we present the results of our extensive experiments on four public datasets, aimed at validating the effectiveness of SharpCF. Firstly, we describe our experimental settings, and then we compare the overall top- $K$  recommendation performance of SharpCF with other CF methods. Next, we showcase the loss landscape of our SharpCF, which reveals that the adversarial training effectively smooths the loss landscape to achieve flat minima. Finally, we investigate the performance of sparse recommendations to verify the generalizability of different CF methods.

### 5.1 Experimental Settings

**5.1.1 Datasets.** We use four public benchmark datasets for evaluating recommendation performance:

- **MovieLens-1M**<sup>1</sup> is a widely used dataset for recommendations. This dataset contains 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 users who joined in 2000.
- **Gowalla**<sup>2</sup> is a location-based social networking website where users share their locations by checking-in. It mainly collects the check-ins of these users over the period from Feb. 2009 to Oct. 2010.
- **Yelp2018**<sup>3</sup> is released by the Yelp challenge that consists of a subset of the businesses, reviews, and user data. The Yelp2018 version is used in the experiments.
- **Amazon-Book**<sup>4</sup> comprises a vast corpus of user reviews, ratings, timestamps, and product metadata gathered from Amazon.com. For our experiments, we select the largest category available, namely Book.

For the MovieLens1M dataset, we consider all ratings as implicit feedback, where each rating score is converted to either 1 or 0 to indicate whether a user rated a movie. For sparser datasets such as Gowalla, Yelp2018, and Amazon-Book, we use the 10-core setting to ensure that all users and items have at least 10 interactions [14]. Table 1 provides a summary of the dataset statistics.

<sup>1</sup><https://grouplens.org/datasets/movielens/>

<sup>2</sup><https://github.com/kuandeng/LightGCN/tree/master/Data>

<sup>3</sup><https://www.yelp.com/dataset>

<sup>4</sup><https://jmcauley.ucsd.edu/data/amazon/>

**5.1.2 Baselines.** As our primary goal of this work is to give a deep insight about the adversarial training and accelerate its computational overhead, we mainly compare with the two popular baselines:

- **BPR** [25]: A classic model that seeks to optimize the Bayesian personalized ranking loss. We employ Matrix Factorization as its preference predictor.
- **APR** [15]: Similar to BPR, it also chooses the Matrix Factorization model and considers the adversarial perturbations during training.

For BPR, APR, and SharpCF, we choose the basic Matrix Factorization as their backbones due to its simplicity and effectiveness. However, it is worth mentioning that SharpCF, similar to BPR, is compatible with any encoders, such as Graph Neural Networks [5, 14]. We leave this extension for further work since our focus is not on developing a new encoder for recommendation in this study. In contrast, we pay more attention to explaining and accelerating the adversarial collaborative filtering.

**5.1.3 Implementation Details.** We implement our SharpCF model in PyTorch on NVIDIA Tesla V100 32GB machines. For all models, the embedding dimension  $d$  of users and items (e.g., in Eq. (1)) is set to 128. We initialize the hyper-parameters for APR as suggested in the original paper and then fine-tune them to optimize performance. For our SharpCF, we choose an epoch window  $E = 3$ . As suggested by APR [15], we set  $E_{\text{start}} = 200$ , which means we first train the model with BPR only for the first 200 epochs to warm up, and then continue training with APR or SharpCF in the experiments.

We adopt two popular top- $K$  metrics, Recall and Normalized Discounted Cumulative Gain (NDCG) [14], for evaluation purposes. The default value of  $K$  is set to [10, 20], and we report the average of Recall@10 and NDCG@10 over all users in the test set. During inference, we consider items that the user has never interacted with in the training set as candidate items. All models predict the users' preference scores over these candidate items and rank them based on the computed scores to calculate Recall@10 and NDCG@10. We independently repeated the experiments five times and report the averaged results.

### 5.2 Experimental Results

**5.2.1 Overall Performance.** In this study, we compare our proposed method SharpCF with BPR and APR on four real-world datasets. Table 2 summarizes the experimental results, including the running time per epoch. Based on our experiments, we have made the following two observations:

- Although the underlying recommender model remains the same (e.g., Matrix Factorization), both APR and SharpCF outperform BPR by leveraging adversarial training. APR achieves this by explicitly introducing worst-case perturbations, while our SharpCF implicitly smooths the learning trajectory to achieve the same effect. Both methods have the potential to smooth the loss landscape to reach out the flat minima, leading to better performance. These findings suggest that the way of training a recommender model is a critical factor in the recommendation process, and one can easily modify the vanilla loss function to enhance performance further.



**Table 2: The performance of different baselines in terms of Recall@K, NDCG@K, and the duration of each epoch in seconds. RI means Relative Improvement w.r.t. baselines.**

| Dataset      | Metrics        | BPR    | APR    | SharpCF       | RI w.r.t. BPR | RI w.r.t. APR |
|--------------|----------------|--------|--------|---------------|---------------|---------------|
| Movielens-1M | Recall@10      | 0.2038 | 0.2273 | <b>0.2354</b> | +15.5%        | +3.56%        |
|              | NDCG@10        | 0.3396 | 0.3705 | <b>0.4123</b> | +21.4%        | +11.3%        |
|              | Recall@20      | 0.3025 | 0.3126 | <b>0.3179</b> | +5.09%        | +1.70%        |
|              | NDCG@20        | 0.3387 | 0.3566 | <b>0.3949</b> | +16.6%        | +10.7%        |
|              | Time per epoch | 16.5s  | 30.5s  | 17.1s         | -             | -             |
| Gowalla      | Recall@10      | 0.1153 | 0.1242 | <b>0.1296</b> | +12.4%        | +4.35%        |
|              | NDCG@10        | 0.1258 | 0.1375 | <b>0.1456</b> | +15.7%        | +5.89%        |
|              | Recall@20      | 0.1674 | 0.1784 | <b>0.1838</b> | +9.80%        | 3.03%         |
|              | NDCG@20        | 0.1404 | 0.1524 | <b>0.1594</b> | +13.5%        | +4.59%        |
|              | Time per epoch | 28.3s  | 50.1s  | 28.7s         | -             | -             |
| Yelp2018     | Recall@10      | 0.0352 | 0.0397 | <b>0.0411</b> | +16.8%        | +3.53%        |
|              | NDCG@10        | 0.0446 | 0.0477 | <b>0.0526</b> | +17.9%        | +10.3%        |
|              | Recall@20      | 0.0591 | 0.0673 | <b>0.0686</b> | +16.1%        | +1.93%        |
|              | NDCG@20        | 0.0519 | 0.0571 | <b>0.0608</b> | +17.1%        | +6.48%        |
|              | Time per epoch | 45.8s  | 88.6s  | 46.1s         | -             | -             |
| Amazon-Book  | Recall@10      | 0.0207 | 0.0241 | <b>0.0267</b> | +29.0%        | +10.8%        |
|              | NDCG@10        | 0.0223 | 0.0259 | <b>0.0294</b> | +31.9%        | +13.5%        |
|              | Recall@20      | 0.0366 | 0.0421 | <b>0.0451</b> | +23.2%        | +7.13%        |
|              | NDCG@20        | 0.0286 | 0.0329 | <b>0.0363</b> | +26.9%        | +10.3%        |
|              | Time per epoch | 127.8s | 230.1s | 128.9s        | -             | -             |

- Our proposed SharpCF has shown improvements in all comparisons. It consistently outperforms the BPR by up to 30%, with an average improvement of 18.1% and a standard deviation of 7% across all datasets. While the improvements over the APR are not as significant as those over the BPR, our SharpCF still shows positive results, with an average improvement of 6.82% and a standard deviation of 3.81% over the APR. Presumably, this is because our SharpCF does not require solving the min-max minimization, which avoids the risk of getting stuck at saddle points that often exhibit large variances, as explained in the original APR paper.

Figure 2 depicts the training curves of various baselines for four datasets. After pretraining for 200 epochs, we observe that continued training of APR and SharpCF results in significant improvement, whereas further training of BPR leads to only marginal gains. For instance, in the case of Amazon-Book, BPR achieves a maximum Recall@10 score of around 0.0207, which is then boosted to 0.0267 by training with SharpCF, leading to a relative improvement of approximately 29%. Table 2 also shows running time elapsed for training per epoch of BPR, APR and SharpCF. In general, the training time for BPR and SharpCF is almost equal, while the computational time for APR is roughly twice as expensive as these two approaches. For example, for the largest dataset Amazon-Book, the training times per epoch for BPR, APR, and SharpCF are around 127.8s, 230.1s, and 128.9s, respectively.

Overall, the experimental results demonstrate the superiority of our proposed SharpCF. Specifically, it outperforms the BPR and APR across four datasets, while maintaining comparable complexity

to the BPR model. These appealing properties make our SharpCF practical for industry-scale applications.

**5.2.2 Loss Landscapes.** Prior works have demonstrated a strong correlation between the flatness of the loss landscape and the generalizability and robustness of a model. In Section 4.2, we establish a connection between Adversarial Training and Sharpness-aware Minimization to explain their generalizability. To verify this assumption, we visualize the loss landscapes of the baselines and our proposed method on four different datasets. Following the methodology outlined in [21], give a well-train model  $\Theta$ , we compute the loss values when moving the model parameters  $\Theta$  along a random direction  $t \in [-2, 2]$  to generate a 1D loss landscape.

From Figure 3, we observe that BPR has a sharper local minima in the loss landscape as compared to APR and SharpCF across all datasets. In other words, the adversarial training techniques, including APR and SharpCF, prefer to generate flatter loss landscapes. A flatter weight loss surface often leads to a smaller gap between training and testing performances, thereby improving the robustness of the model’s generalization capabilities [2, 11, 20, 21]. Rather than interpreting adversarial training as a min-max game aimed at enhancing robustness and generalizability, as done in APR, we provide an alternative perspective by showing that adversarial training tends to achieve a flatter weight loss landscape, thereby reducing the robust generalization gap.

### 5.3 Further Probe

In this section, we conducted a series of detailed analyses on the proposed SharpCF to confirm its effectiveness. We report the results



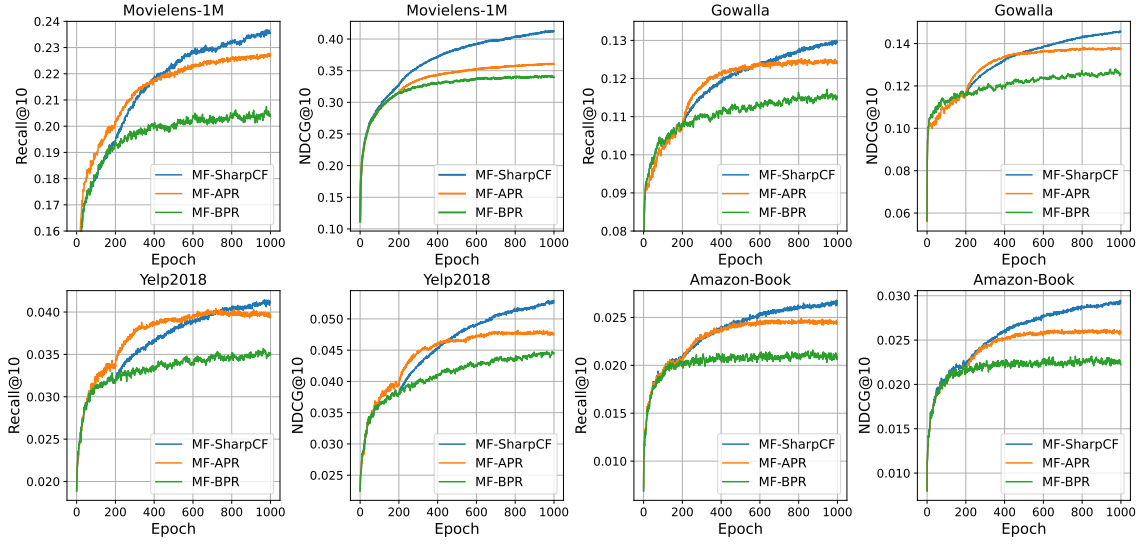


Figure 2: Training curves of BPR, APR, and SharpCF on different datasets.

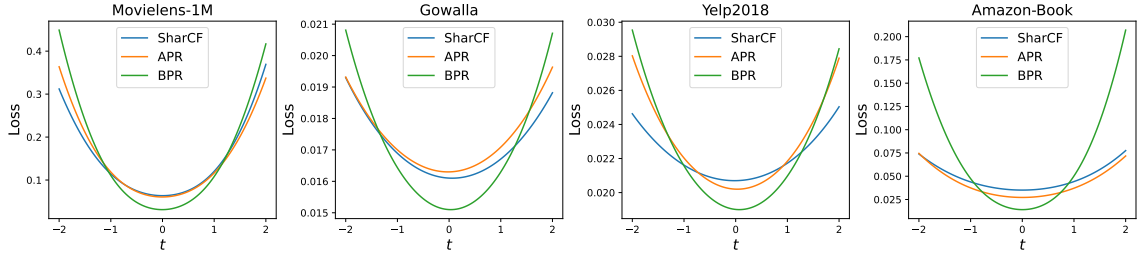


Figure 3: Training curves of BPR, APR, and SharpCF on different datasets.

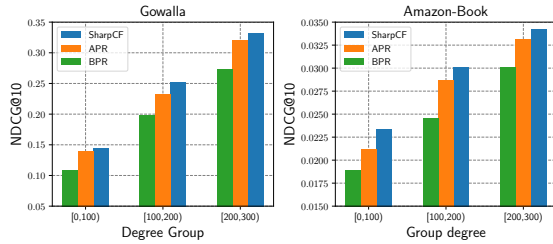


Figure 4: The performance of different models with respect to item degree on Gowalla and Amazon-Book.

only on the Gowalla and Amazon-Book datasets, as the observations are similar on the other two datasets and thus omitted here.

**5.3.1 Long-tail Recommendation.** To further investigate the model's generalization, we grouped items by degree and visualized the average performance of each group. We mainly focused on Gowalla and Amazon-Book since they are sparser than the other two datasets. Additionally, we explored the item degree within the ranges of

[0,100), [100, 200), and [200, 300) to focus on long-tail items. Figure 4 displays the performance of different models with respect to item degrees. As shown, SharpCF and APR achieved higher performance for low-degree items. This indicates that both APR and SharpCF are capable of providing high-quality recommendations even with sparse interaction data, thanks to smoothing the loss landscape and better generalization.

**5.3.2 The Impact of the Coefficient  $\lambda$ .** In the objective function of SharpCF defined in Eq. (17), the coefficient  $\lambda$  is used to balance the two losses: the original task loss and the trajectory loss. To analyze the influence of  $\lambda$ , we vary its value in the range of 0.001 to 10 and reported the experimental results in Figure 5. The results indicate that an appropriate value of  $\lambda$  can effectively improve the performance of SharpCF. Specifically, when the hyper-parameter  $\lambda$  is set to around 0.1 or 1.0, the performance becomes better on both datasets, revealing that tracking the learning trajectory is valuable for improving performance. However, when  $\lambda$  is set to around 10.0, the performance of our SharpCF drops quickly, suggesting that too strong regularization on the trajectory loss will negatively affect normal training of the model and is not encouraged in practice.

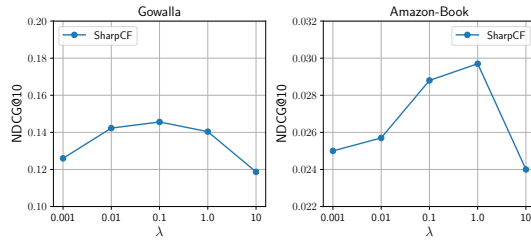


Figure 5: The performance of SharpCF for different settings of  $\lambda$ .

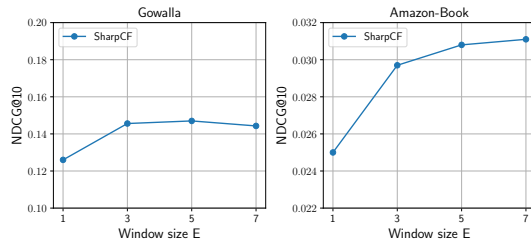


Figure 6: The performance of SharpCF for different settings of the epoch window size  $E$ .

**5.3.3 The Impact of the Epoch Window Size  $E$ .** In our proposed SharpCF model, the epoch window size  $E$  in Eq. (17) is a crucial hyper-parameter that affects its performance. A larger value of  $E$  means that the SharpCF will track the long-range trajectory to smooth the loss landscape. To examine the impact of  $E$  on SharpCF, we vary it from 1 to 7 and evaluate the model's performance on different datasets. From Figure 6, the results show that the performance of SharpCF is relatively stable across different settings of  $E$ . In some cases, increasing the value of  $E$  can gradually boost the performance on the Amazon-Book dataset. It should be noted that tracking longer trajectories in SharpCF requires more memory to cache the model states. While this is a cheaper alternative to the APR method, it still needs to be considered in real-world applications. In our experiments, we found that choosing an epoch window size of  $E = 3$  provides good recommendation performance with the reasonable memory requirement.

## 6 CONCLUSION

In this paper, we first revisit the existing adversarial collaborative filtering methods and demonstrate that adversarial training favors flat minima over sharp ones, which results in better generalizability. We then propose our Sharpness-aware Collaborative Filtering (SharpCF), a simple yet effective method that conducts adversarial training without extra computational cost over the base optimizer. Our proposed method has shown superior performance on four public real-world datasets compared to current adversarial collaborative filtering methods with reasonable time complexity.

In our future work, we aim to expand our assessment of the effectiveness of our adversarial training approach in diverse recommendation tasks, including fairness recommendation. Additionally,

we plan to extend the capabilities of our framework to establish it as a versatile training technique that improves model generalization.

## REFERENCES

- [1] Maksym Andriushchenko and Nicolas Flammarion. 2020. Understanding and improving fast adversarial training. In *Advances in Neural Information Processing Systems*.
- [2] Maksym Andriushchenko and Nicolas Flammarion. 2022. Towards understanding sharpness-aware minimization. In *International Conference on Machine Learning*.
- [3] Huiyuan Chen and Jing Li. 2019. Adversarial tensor factorization for context-aware recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 363–367.
- [4] Huiyuan Chen, Yusan Lin, Menghai Pan, Lan Wang, Chin-Chia Michael Yeh, Xiaoting Li, Yan Zheng, Fei Wang, and Hao Yang. 2022. Denoising self-attentive sequential recommendation. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 92–101.
- [5] Huiyuan Chen, Chin-Chia Michael Yeh, Yujie Fan, Yan Zheng, Junpeng Wang, Vivian Lai, Mahashweta Das, and Hao Yang. 2023. Sharpness-aware Graph Collaborative Filtering. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- [6] Huiyuan Chen, Kaixiong Zhou, Kwei-Heng Lai, Xia Hu, Fei Wang, and Hao Yang. 2022. Adversarial graph perturbations for recommendations at scale. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1854–1858.
- [7] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Isipir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*.
- [8] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*. 191–198.
- [9] Jiawei Du, Daquan Zhou, Jiashi Feng, Vincent Tan, and Joey Tianyi Zhou. 2022. Sharpness-aware training for free. *Advances in Neural Information Processing Systems* 35, 23439–23451.
- [10] Lun Du, Xu Chen, Fei Gao, Qiang Fu, Kunqing Xie, Shi Han, and Dongmei Zhang. 2022. Understanding and Improvement of Adversarial Training for Network Embedding from an Optimization Perspective. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 230–240.
- [11] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2021. Sharpness-aware Minimization for Efficiently Improving Generalization. In *International Conference on Learning Representations*.
- [12] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*.
- [13] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*.
- [14] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 639–648.
- [15] Xiangnan He, Zhankui He, Xiaoyu Du, and Tat-Seng Chua. 2018. Adversarial personalized ranking for recommendation. In *The 41st International ACM SIGIR conference on research & development in information retrieval*. 355–364.
- [16] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*. 173–182.
- [17] Gangwei Jiang, Hao Wang, Jin Chen, Haoyu Wang, Defu Lian, and Enhong Chen. 2021. xLightFM: Extremely memory-efficient factorization machine. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 337–346.
- [18] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 426–434.
- [19] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.
- [20] Jungmin Kwon, Jeongseop Kim, Hyunseop Park, and In Kwon Choi. 2021. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*. 5905–5914.
- [21] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. 2018. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*.
- [22] Jianxun Lian, Xiaohuan Zhou, Fuzheng Zhang, Zhongxia Chen, Xing Xie, and Guangzhong Sun. 2018. xdeepfm: Combining explicit and implicit feature interactions for recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1754–1763.

- [23] David A McAllester. 1999. PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*. 164–170.
- [24] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*. 165–172.
- [25] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th conference on uncertainty in artificial intelligence*. 452–461.
- [26] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. Collaborative filtering recommender systems. In *The adaptive web*. Springer, 291–324.
- [27] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free!. In *Advances in Neural Information Processing Systems*.
- [28] Susheel Suresh, Pan Li, Cong Hao, and Jennifer Neville. 2021. Adversarial graph augmentation to improve graph contrastive learning. In *Advances in Neural Information Processing Systems*.
- [29] Changxin Tian, Yuexiang Xie, Yaliang Li, Nan Yang, and Wayne Xin Zhao. 2022. Learning to Denoise Unreliable Interactions for Graph Collaborative Filtering. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 122–132.
- [30] Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang. 2017. Irgan: A minimax game for unifying generative and discriminative information retrieval models. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. 515–524.
- [31] Qinyong Wang, Hongzhi Yin, Zhiting Hu, Defu Lian, Hao Wang, and Zi Huang. 2018. Neural memory streaming recommender networks with adversarial training. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2467–2475.
- [32] Song Wang, Xingbo Fu, Kaize Ding, Chen Chen, Huiyuan Chen, and Jundong Li. 2023. Federated Few-shot Learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
- [33] Wenjie Wang, Fuli Feng, Xiangnan He, Liqiang Nie, and Tat-Seng Chua. 2021. Denoising implicit feedback for recommendation. In *Proceedings of the 14th ACM international conference on web search and data mining*. 373–381.
- [34] Wenjie Wang, Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Clicks can be cheating: Counterfactual recommendation for mitigating clickbait issue. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1288–1297.
- [35] Yu Wang, Yuying Zhao, Yushun Dong, Huiyuan Chen, Jundong Li, and Tyler Derr. 2022. Improving fairness in graph neural networks via mitigating sensitive attribute leakage. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1938–1948.
- [36] Hongyi Wen, Longqi Yang, and Deborah Estrin. 2019. Leveraging post-click feedback for content recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 278–286.
- [37] Eric Wong, Leslie Rice, and J. Zico Kolter. 2020. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*.
- [38] Chenwang Wu, Defu Lian, Yong Ge, Zhihao Zhu, Enhong Chen, and Senchao Yuan. 2021. Fight fire with fire: towards robust recommender systems via adversarial poisoning training. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1074–1083.
- [39] Chin-Chia Michael Yeh, Mengting Gu, Yan Zheng, Huiyuan Chen, Javid Ebrahimi, Zhongfang Zhuang, Junpeng Wang, Liang Wang, and Wei Zhang. 2022. Embedding Compression with Hashing for Efficient Representation Learning in Large-Scale Graph. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4391–4401.
- [40] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. 2018. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 974–983.
- [41] Feng Yuan, Lina Yao, and Boualem Benatallah. 2019. Adversarial collaborative neural network for robust recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1065–1068.
- [42] Yihua Zhang, Guanhua Zhang, Prashant Khanduri, Mingyi Hong, Shiyu Chang, and Sijia Liu. 2022. Revisiting and advancing fast adversarial training through the lens of bi-level optimization. In *International Conference on Machine Learning*. 26693–26712.