The Minority Matters: A Diversity-Promoting Collaborative Metric Learning Algorithm

Shilong Bao[1,2]   **Qianqian Xu**[3*]   **Zhiyong Yang**[4]   **Yuan He**[5]
Xiaochun Cao[6]   **Qingming Huang**[3,4,7,8*]

[1] State Key Laboratory of Information Security, Institute of Information Engineering, CAS
[2] School of Cyber Security, University of Chinese Academy of Sciences
[3] Key Lab. of Intelligent Information Processing, Institute of Computing Technology, CAS
[4] School of Computer Science and Tech., University of Chinese Academy of Sciences
[5] Alibaba Group
[6] School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University
[7] Key Laboratory of Big Data Mining and Knowledge Management, CAS
[8] Peng Cheng Laboratory

baoshilong@iie.ac.cn, xuqianqian@ict.ac.cn, heyuan.hy@alibaba-inc.com,
caoxiaochun@mail.sysu.edu.cn, {yangzhiyong21,qmhuang}@ucas.ac.cn

Abstract

Collaborative Metric Learning (CML) has recently emerged as a popular method in recommendation systems (RS), closing the gap between metric learning and Collaborative Filtering. Following the convention of RS, existing methods exploit unique user representation in their model design. This paper focuses on a challenging scenario where a user has multiple categories of interests. Under this setting, we argue that the unique user representation might induce preference bias, especially when the item category distribution is imbalanced. To address this issue, we propose a novel method called *Diversity-Promoting Collaborative Metric Learning* (DPCML), with the hope of considering the commonly ignored minority interest of the user. The key idea behind DPCML is to include a multiple set of representations for each user in the system. Based on this embedding paradigm, user preference toward an item is aggregated from different embeddings by taking the minimum item-user distance among the user embedding set. Furthermore, we observe that the diversity of the embeddings for the same user also plays an essential role in the model. To this end, we propose a *Diversity Control Regularization Scheme* (DCRS) to accommodate the multi-vector representation strategy better. Theoretically, we show that DPCML could generalize well to unseen test data by tackling the challenge of the annoying operation that comes from the minimum value. Experiments over a range of benchmark datasets speak to the efficacy of DPCML.

1   Introduction

Recommender system (RS) is a well-known building block in eCommerce, which can assist buyers to find products they wish to purchase by giving them the relevant recommendations. The key recipe behind RS is to learn from user-item interaction records [45, 30, 31, 29, 19]. In practice, since user preferences are hard to collect, such records often exist as implicit feedback [48, 1, 42] where only indirect actions are provided (say clicks, collections, reposts, and etc.). Such a property of implicit

---

*Corresponding authors.

feedback raises a great challenge to RS-targeted machine learning methods and thus stimulates a wave of relevant studies along this course [56, 63, 51].

Over the past two decades, most literature follows a typical paradigm known as the One-Class Collaborative Filtering (OCCF) [35], where the items not being observed are usually assumed to be of less interest for the user and labeled as negative instances. In the early days, the vast majority of studies in the OCCF community focus on Matrix Factorization (MF) based algorithms, where the preference of a specific user to an item is conveyed by the inner product between their embeddings [60, 6]. Recently, a milestone study known as *Collaborative Metric Learning* (CML) [18] pointed out that the inner product involved in MF violates the triangle inequality, resulting in a sub-optimal topological embedding space. To fix this, CML proposes a novel framework to overcome such a problem by borrowing the strength from metric learning. Practically, CML has achieved promising performance over a series of RS benchmark datasets. Hereafter, many efforts have been made along the research direction to improve CML [44, 41, 36, 2, 54, 47, 65, 62, 43]. More discussions of the related work are presented in Appendix.A.

However, through the lens of a critical example in the practical scenarios (shown in Sec.2.2), we notice that users usually have multiple categories of preferences in real-world RS. Moreover, such interest groups are often not equally distributed, where the amount of some groups dominates the others. Unfortunately, as shown in Fig.3, in this case, the existing studies might induce preference bias since they tend to meet the majority interest while missing the other potential preference. Therefore, in this paper, we ask:

> *How to develop an effective CML-based algorithm to accommodate the diversity of user preferences?*

Contributions. In search of an answer, we propose a novel algorithm called *Diversity-Promoting Collaborative Metric Learning (DPCML)*. The key idea is to explore the diversity of user interest which spans multiple groups of items. To this end, we propose a multi-vector user representation strategy, where each user has a set of C embeddings. To find out the score of a given item embedding $\boldsymbol{g}_v$, we aggregate the results from the user embeddings $\boldsymbol{g}_u^1, \boldsymbol{g}_u^2, \cdots, \boldsymbol{g}_u^C$ by taking the minimum distance $s(u,v) = \min_c \|\boldsymbol{g}_u^c - \boldsymbol{g}_v\|^2$. Then we will recommend the item with the smallest s value.

In this way, we can focus on all potential items that fit one of the users' interests well, both for the majority and the minority interests. Meanwhile, we observe that the diversity of the embeddings among the same user representation set also plays an important role in better achieving our goal. Therefore, we further present a novel diversity control regularization scheme.

Taking a step further, we continue to ask the following question:

> *Could CML generalize well under the multi-vector representation strategy?*

To the best of our knowledge, such a problem remains barely explored in the existing literature. To solve the problem, we then proceed to explore the generalization bound for DPCML algorithm. Here the major challenges fall into two aspects: 1) The risk of DPCML could not be expressed as a sum of independently identically distributed (i.i.d.) loss terms, making the standard Rademacher Complexity-based [3, 33] theoretical arguments unavailable; 2) The annoying minimum operation are not continuous, which cannot be analyzed easily in the Rademacher complexity framework. Facing these challenges, we employ the covering number and ϵ-net arguments to derive the generalization bound. On top of this, we show that DPCML could induce a small generalization error with high probability. This supports the effectiveness of DPCML from a theoretical perspective.

Finally, we conduct empirical studies over a range of RS benchmark datasets that demonstrate the superiority of DPCML.

2 Methodology

2.1 Preliminary

In this paper, we focus on how to develop an effective CML-based recommendation system on top of the implicit feedback signals (say clicks, browses, and bookmarks). Assume there are a pool of users and items in the system, denoted by $\mathcal{U} = \{u_1, u_2, \ldots, u_{|\mathcal{U}|}\}$ and $\mathcal{I} = \{v_1, v_2 \ldots, v_{|\mathcal{I}|}\}$,
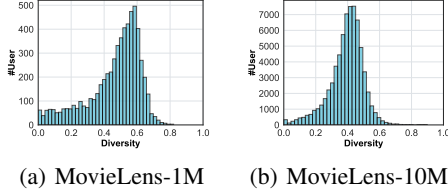
(a) MovieLens-1M   (b) MovieLens-10M

Figure 1: Statistics of preference diversity.
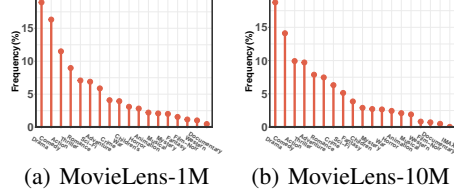

(a) MovieLens-1M   (b) MovieLens-10M

Figure 2: The item category distribution.

respectively. For each user $u_i \in \mathcal{U}, i = 1, 2, \ldots, |\mathcal{U}|$, let $\mathcal{D}_{u_i}^+ = \{v_1^+, v_2^+, \ldots, v_{n_i^+}^+\}$ denote the set of items that user u_i has interacted with (i.e., observed user-item interactions) and the rest of the items (i.e., unobserved interactions) are denoted by $\mathcal{D}_{u_i}^- = \{v_1^-, v_2^-, \ldots, v_{n_i^-}^-\}$, where n_i^+, n_i^- are the number of observed/unobserved interactions of user u_i. We have $\mathcal{I} = \mathcal{D}_{u_i} = \mathcal{D}_{u_i}^+ \cup \mathcal{D}_{u_i}^-$ and $|\mathcal{I}| = n_i^+ + n_i^-$. In the standard settings of OCCF, one usually assumes that users tend to have a higher preference for the items contained in $\mathcal{D}_{u_i}^+$ than the items in $\mathcal{D}_{u_i}^-$. Therefore, given a target user $u_i \in \mathcal{U}$ and his/her historical interaction records, the goal of RS is to discover the most interested N items by recommending the items with the top-N (bottom-N) score. The top-N item list is denoted as $\mathcal{I}_N^{u_i}$.

2.2 Motivating Example

We start by a definition of the preference diversity of users.

Definition 1 (Preference Diversity). Assume that there exists an attribute set $\mathcal{T} = \{\mathcal{T}(v_1), \mathcal{T}(v_2), \ldots, \mathcal{T}(v_{|\mathcal{I}|})\}$ in a typical RS, where $\mathcal{T}(v_j) = \{t_1, t_2, \ldots, t_{T_j}\}$ contains the attribute information of item v_j (e.g., the genres of a movie) and T_j is the number of attributes. Given a user u_i and interaction records $\mathcal{D}_{u_i}^+$, the preference diversity is defined as follows:

$$\mathsf{Div}(u_i) = \frac{\sum_{v_j, v_k \in \mathcal{D}_{u_i}^+, v_j \neq v_k} \mathbb{I}\left[\mathcal{T}(v_j) \cap \mathcal{T}(v_k) = \varnothing\right]}{|\mathcal{D}_{u_i}^+|(|\mathcal{D}_{u_i}^+| - 1)},$$

where $\mathbb{I}(x)$ is an indicator function, i.e., returns 1 if the condition x holds, otherwise 0 is returned.

Remark 1. Intuitively, the range of $\mathsf{Div}(u_i)$ is among $[0, 1]$, and its value measures the diversity of u_i's preference to a certain extent. That is to say, if items among the historical interaction records of users are irrelevant, there should induce a large value (e.g., $\mathsf{Div}(u_i) = 1$), implying the diversity of their preferences. If the opposite is the case, the value is small. This means users may have narrow interests where only some unique attributes appeal to them.

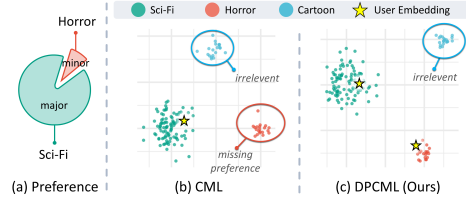(a) Preference   (b) CML   (c) DPCML (Ours)

Figure 3: An illustration shows the benefit of our proposed algorithm when a user has multiple categories of preferences. Taking movies as an example, we assume that Sci-Fi/Horror is the majority/minority interest of the user while Cartoon is an irrelevant movie type. It is easy to see that if the item embeddings are distributed as shown in the figure, we can hardly find a single user embedding that simultaneously captures both interests.

Based on Def.1, we visualize the user preferences on two real-world benchmark datasets, including **MovieLens-1m** and **MovieLens-10m**. The detailed information of datasets is listed in Tab.1 in Appendix.C. Here we adopt the movie genres as the attribute set $\mathcal{T}$ because such information is easy to obtain. The results are shown in Fig.1. From the results, we can make the following observations. First, only a few users have limited interest. Moreover, most of the users have a diversity value spaning $(0, 0.8]$, suggesting that they have multiple categories of interests. Finally, one can notice that there are very few users with high preference diversity (at the lower-right corner) in both figures. This is a convincing case in the real-world recommendation since most users usually have interests in a couple of movie genres but not all.

Motivation and Discussion. Through the above example, the key information is that users usually have multiple categories of preference in real-world recommendations. This poses a critical challenge to the current CML framework. Specifically, following the convention of RS, the existing CML-based methods leverage unique representations of users to model their preferences. Facing the multiplicity of user intentions, such a paradigm may induce preference bias due to the limited expressiveness, especially when the item category distribution is imbalanced. Fig.2 visualizes the item distribution on MovieLens-1m and MovieLens-10m datasets. We see that both of them are imbalanced. In this case, as shown in Fig.3-(b), CML would pay more attention to the **majority** interest of users making the unique user embedding close to the items with the science fiction (Sci-Fi) category. In this way, the **minority** interest of the user (i.e., Horror movies) would be ignored by the method, inducing performance degradation. This motivates us to explore diversity-promoting strategies on top of CML.

2.3 Diversity-Promoting Collaborative Metric Learning

Recall that the critical recipe behind CML-based algorithms is to seek a metric space (usually adopting the Euclidean space) such that user preferences could be naturally specified by their distance toward different items. To do this, the traditional CML-based methods usually represent each user and each item as a vector, respectively. Different from them, taking the preference diversity of users into account, we propose to adopt C ($C > 1$) different embeddings for each user and represent each item as one single vector in a joint Euclidean space.

Concretely, each user u_i is projected into the metric space via the following lookup transformations [49, 50, 55]:

$$\boldsymbol{g}_{u_i}^c = \boldsymbol{P}_c^\top \boldsymbol{u}_i, \;\; \forall c, u_i, \; c \in [C], \;\; u_i \in \mathcal{U}, \tag{1}$$

where $\boldsymbol{g}_{u_i}^c \in \mathbb{R}^d$ is a representation vector of user u_i; $[C]$ is the set $\{1, 2, \ldots, C\}$; $\boldsymbol{P}_c \in \mathbb{R}^{|\mathcal{U}| \times d}$ is a learned transformation weight; d is the dimension of space and $\boldsymbol{u}_i \in \mathbb{R}^{|\mathcal{U}|}$ is a one-hot encoding that the nonzero elements correspond to its index of a particular user u_i.

Similarly, we apply the following transformation to each item v_j:

$$\boldsymbol{g}_{v_j} = \boldsymbol{Q}^\top \boldsymbol{v}_j, \;\; \forall v_j \in \mathcal{I}, \tag{2}$$

where $\boldsymbol{g}_{v_j} \in \mathbb{R}^d$ is the embedding of item v_j; $\boldsymbol{Q} \in \mathbb{R}^{|\mathcal{I}| \times d}$ is the learned transformation weight and $\boldsymbol{v}_j \in \mathbb{R}^{|\mathcal{I}|}$ is a one-hot embedding of item v_j.

In what follows, given a target user u_i, we need to find out a score function to express the user preference toward an item in the context of multiple representations of users. Here we define the score function by taking the minimum item-user Euclidean distance among the user embedding set:

$$s(u_i, v_j) = \min_{c \in [C]} \|\boldsymbol{g}_{u_i}^c - \boldsymbol{g}_{v_j}\|^2, \forall v_j \in \mathcal{I}. \tag{3}$$

Equipped with this formulation, we focus on the potential items that fit one of the user preferences. If user u_i has interacted with item v_j, there should be a small value with respect to $s(u_i, v_j)$. If the opposite is the case, we then expect to see a large $s(u_i, v_j)$. Mathematically, the following inequality should be satisfied to reflect the relative preference of u_i in the learned Euclidean space:

$$s(u_i, v_j^+) < s(u_i, v_k^-), \forall v_j^+ \in \mathcal{D}_{u_i}^+, \; \forall v_k^- \in \mathcal{D}_{u_i}^-. \tag{4}$$

Therefore, given the whole sample set $\mathcal{D} = \underset{u_i \in \mathcal{U}}{\cup} \mathcal{D}_{u_i}$, we adopt the following pairwise learning problems [18, 44, 24, 57] to achieve such goal:

$$\min_{\boldsymbol{g}} \;\; \hat{\mathcal{R}}_{\mathcal{D}, \boldsymbol{g}}, \tag{5}$$

where, $\forall v_j^+ \in \mathcal{D}_{u_i}^+$, $\forall v_k^- \in \mathcal{D}_{u_i}^-$, we have

$$\hat{\mathcal{R}}_{\mathcal{D}, \boldsymbol{g}} = \frac{1}{|\mathcal{U}|} \sum_{u_i \in \mathcal{U}} \frac{1}{n_i^+ n_i^-} \sum_{j=1}^{n_i^+} \sum_{k=1}^{n_i^-} \ell_g^{(i)}(v_j^+, v_k^-),$$

4

$$\ell_g^{(i)}(v_j^+, v_k^-) = \max(0, \lambda + s(u_i, v_j^+) - s(u_i, v_k^-)). \tag{6}$$

and $\lambda > 0$ is a safe margin.

According to (5), we have the following explanations. At first, optimizing the above problem could pull the observed items close to the users and push the unobserved items apart from the observed items. This achieves our goal of preserving user preferences in the Euclidean space. Then, as shown in Fig.3-(c), equipped with a multiple set of representations for each user, DPCML would exploit different user vectors to focus on different interest groups. In this sense, the minority interest groups can also be modeled well. Last but not least, one appealing property is that, DPCML also preserves the triangle inequality for the items falling into the same interest group.

2.4 Diversity Control Regularization Scheme

In practice, we note that a proper regularization scheme is crucial to accommodate the multi-vector representation strategy. Here we focus on the diversity within the embedding sets of a given user. Such diversity is defined as the average pairwise distance among the C user embeddings for user u_i, i.e.,

$$\delta_{\boldsymbol{g}, u_i} = \frac{1}{2C(C-1)} \sum_{c_1, c_2 \in [C]} \|\boldsymbol{g}_{u_i}^{c_1} - \boldsymbol{g}_{u_i}^{c_2}\|^2.$$

Based on the definition, we argue that one should attain a proper $\delta_{\boldsymbol{g}, u_i}$ to get a good performance since extremely large/small values of $\delta_{\boldsymbol{g}, u_i}$ might be harmful for the generalization error. It is easy to see that if $\delta_{\boldsymbol{g}, u_i}$ is extremely small, the embeddings for a given user are very close to each other such that the multi-vector representation strategy degenerates to the original single-vector representation. This increases the model complexity with few performance gains and obviously will induce overfitting. On the other hand, a too large diversity might also induce overfitting. It might be a bit confusing at first glance. But, imagine that when some noise observations or extremely rare interests far away from the normal patterns exist in the data, having a large diversity will make it easier to overfit such data. Moreover, it is also a natural assumption that a user's interests should not be too different, as validated in Fig.1. In this sense, the distance across different user embeddings should remain at a moderate magnitude.

Therefore, controlling a proper diversity is essential for the multi-vector representation. To do this, we put forward the following diversity control regularization scheme (DCRS):

$$\hat{\Omega}_{\mathcal{D}, \boldsymbol{g}} = \frac{1}{|\mathcal{U}|} \sum_{u_i \in \mathcal{U}} \psi_{\boldsymbol{g}}(u_i), \tag{7}$$

where, we have

$$\psi_{\boldsymbol{g}}(u_i) = \max\left(0, \delta_1 - \delta_{\boldsymbol{g}, u_i}\right) + \max\left(0, \delta_{\boldsymbol{g}, u_i} - \delta_2\right),$$

and δ_1, δ_2 are two threshold parameters with $\delta_1 \leq \delta_2$. Intuitively, optimizing (7) ensures that the diversity of user's vectors lies between δ_1 and δ_2.

2.5 Optimization

Finally, we arrive at the following optimization problem for our proposed DPCML:

$$\min_{\boldsymbol{g}} \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) := \hat{\mathcal{R}}_{\mathcal{D}, \boldsymbol{g}} + \eta \cdot \hat{\Omega}_{\mathcal{D}, \boldsymbol{g}}, \tag{8}$$

where η is a trade-off hyper-parameter.

When the training is completed, one can easily carry out recommendations by choosing the items with the smallest $s(u_i, v_j), \forall v_j, v_j \in \mathcal{I}$.

2.6 General Framework of Joint Accessibility

Now, we expect to provide another intriguing perspective of our proposed method. As we discussed in Sec.A.2, equipped with a multiple set of representations for each user, our proposed algorithm could be treated as a generalized framework against the joint accessibility issue. To see this, if we

restrict the user and item embeddings within a unit sphere, then the score function (3) degenerates to :

$$s(u_i, v_j) = \min_{c \in [C]} \left(1 - \hat{\boldsymbol{g}}_{u_i}^c \boldsymbol{g}_{v_j} \right),$$
$$s.t. \ \|\boldsymbol{g}_{u_i}^c\| = 1, \forall u_i \in \mathcal{U},$$
$$\|\boldsymbol{g}_{v_j}\| = 1, \forall v_j \in \mathcal{I},$$

(9)

where $\hat{\boldsymbol{g}}_{u_i}^c \in \mathbb{R}^{1 \times d}$ represents the transpose vector of $\boldsymbol{g}_{u_i}^c \in \mathbb{R}^d$. Therefore, to minimize (9), one only needs to maximize the following equivalent problem:

$$\hat{s}(u_i, v_j) = \max_{c \in [C]} \hat{\boldsymbol{g}}_{u_i}^c \boldsymbol{g}_{v_j},$$
$$s.t. \ \|\hat{\boldsymbol{g}}_{u_i}^c\| = 1, \forall u_i \in \mathcal{U},$$
$$\|\boldsymbol{g}_{v_j}\| = 1, \forall v_j \in \mathcal{I},$$

(10)

which is exactly the original form of the joint accessibility model.

3 Generalization Analysis

In this section, we present a systematic theoretical analysis of the generalization ability of our proposed algorithm. Following the standard learning theory, deriving a uniform upper bound of the generalization error relies on the proper measure of its complexity over the given hypothesis space $\mathcal{H}$. The most common complexity to achieve this is the Rademacher complexity [3, 33, 23], which is derived from the symmetrization technique as an upper bound for the largest deviation over a given hypothesis space $\mathcal{H}$:

$$\mathbb{E}_{\mathcal{D}} \left[\sup_{f \in \mathcal{H}} \mathbb{E}_{\mathcal{D}}(\hat{\mathcal{R}}_{\mathcal{D}}) - \hat{\mathcal{R}}_{\mathcal{D}} \right].$$

However, the standard symmetrization technique requires the empirical risk $\hat{\mathcal{R}}_{\mathcal{D}}$ to be a sum of independent terms, which is not applicable for the CML-based methods since they usually involve a sum of pairwise terms in (5). For instance, with respect to (5), the terms $\ell_g^{(i)}(v_j^+, v_k^-)$ and $\ell_g^{(i)}(\tilde{v}_j^+, \tilde{v}_k^-)$ are interdependent as long as one of them is the same (i.e., $v_j^+ = \tilde{v}_j^+$ or $v_k^- = \tilde{v}_k^-$).

Therefore, we turn to leverage another complexity measure, i.e., covering number, to overcome this difficulty. The necessary notations are summarized as follows.

Definition 2 (ϵ-Covering). [21] Let $(\mathcal{F}, \rho)$ be a (pseudo) metric space, and $\mathcal{G} \subseteq \mathcal{F}$. $\{f_1, \ldots, f_K\}$ is said to be an ϵ-covering of $\mathcal{G}$ if $\mathcal{G} \subseteq \bigcup_{i=1}^{K} \mathcal{B}(f_i, \epsilon)$, i.e., $\forall g \in \mathcal{G}, \exists i$ such that $\rho(g, f_i) \leq \epsilon$.

Definition 3 (Covering Number). [21] According to the notations in Def.2, the covering number of $\mathcal{G}$ with radius ϵ is defined as:

$$\mathcal{N}(\epsilon; \mathcal{G}, \rho) = \min\{n : \exists \epsilon - covering \ over \ \mathcal{G} \ with \ size \ n\}$$

With the above definitions, we further have the following assumption and lemma to help us derive the generalization bound.

Assumption 1 (Basic Assumptions). We assume that all the embeddings of users and items are chosen from the following embedding hypothesis space:

$$\mathcal{H}_R = \left\{ \boldsymbol{g} : \boldsymbol{g} \in \mathbb{R}^d, \|\boldsymbol{g}\| \leq r \right\},$$

(11)

where $\boldsymbol{g}_{u_i}^c \in \mathcal{H}_R, u_i \in \mathcal{U}, c \in [C]$ and $\boldsymbol{g}_{v_j} \in \mathcal{H}_R, v_j \in \mathcal{I}$.

Lemma 1. [28, 25, 64] The covering number of the hypothesis class $\mathcal{H}_R$ has the following upper bound:

$$\log \mathcal{N}(\epsilon; \mathcal{H}_R, \rho) \leq d \log \left(\frac{3r}{\epsilon} \right),$$

(12)

where d is the dimension of embedding space.

Based on the above introductions, we have the following results. ***Due to space limitations, please refer to Appendix.B for all proofs in detail.***

Theorem 1 (Generalization Upper Bound of DPCML). *Let $\mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})]$ be the population risk of $\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})$. Then, $\forall \, \boldsymbol{g} \in \mathcal{H}_R$, with high probability, the following inequation holds:*

$$\left| \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) - \mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})] \right| \leq \sqrt{\frac{2d \log\left(3r\tilde{N}\right)}{\tilde{N}}}, \tag{13}$$

where we have

$$\tilde{N} = \left(4r^2 \sqrt{\left(\frac{(4+\eta)^2}{|\mathcal{U}|} + \frac{2}{|\mathcal{U}|^2} \sum_{u_i \in \mathcal{U}} \left(\frac{1}{n_i^+} + \frac{1}{n_i^-} \right) \right)} \right)^{-2}$$

Intriguingly, we see that our derived bound does not depend on C. This is consistent with the over-parameterization phenomenon [8, 34]. On top of Thm.1, we have the following corollary.

Corollary 1. *DPCML could enjoy a smaller generalization error than CML.*

Therefore, we can conclude that DPCML generalizes to unseen data better than single-vector CML and thus improves the recommendation performance. This supports the superiority of our proposed DPCML from a theoretical perspective. In addition, we also empirically demonstrate this in the experiment Sec.4.3 and Appendix.C.6.4.

Table 1: Basic Information of the Datasets. %Density is defined as $\frac{\#Ratings}{\#Users \times \#Items} \times 100\%$.

Datasets	MovieLens-1M	Steam-200k	CiteULike-T	MovieLens-10M
Domain	Movie	Game	Paper	Movie
#Users	6,034	3,757	5,219	69,167
#Items	3,953	5,113	25,975	10,019
#Ratings	575,271	115,139	125,580	5,003,437
%Density	2.4118%	0.5994%	0.0926%	0.7220%

4 Experiments

In this section, our proposed method is applied to a wide range of real-world recommendation datasets to show its superiority. ***Please refer to Appendix.C for more results about experiments***.

4.1 Experimental Setups

To begin with, we perform empirical experiments on several common recommendation benchmarks: *MovieLens-1m*, *Steam-200k*, *CiteULike* and *MovieLens-10m*. The detailed statistics in terms of these datasets are summarized in Tab.1. For the datasets with explicit feedbacks, we follow the previous works [14, 44] and transfer them into implicit feedback. Secondly, we evaluate the performance with five metrics, including **Precision** (P@N), **Recall** (R@N), **Normalized Discounted Cumulative Gain** (NDCG@N), **Mean Average Precision** (MAP), and **Mean Reciprocal Rank** (MRR). Moreover, we compared our proposed method with 14 competitive competitors: a) Item-based CF method, **itemKNN** [27]; b) MF-based algorithms, including the combination of MF and deep learning models and multi-vector MF-based approaches: **GMF**, **MLP**, **NeuMF** [14], **M2F** [12] and **MGMF** [12]; c) CML-based methods, including **UniS** [35], **PopS** [53], **2stS** [44], **HarS** [18, 11], **TransCF** [36], **LRML** [41], **AdaCML** [62] and **HLR** [43].

4.2 Overall Performance

The experimental results of all the involved competitors are shown in Tab.2 and Tab.5 (in Appendix.C.5). Consequently, we can draw the following conclusions: 1) In most cases, the best

Table 2: Performance comparisons on MovieLens-1m and Steam-200k datasets.

|	Type	Method	P@3	R@3	NDCG@3	P@5	R@5	NDCG@5	MAP	MRR
MovieLens-1m	Item-based	itemKNN	12.24	2.90	12.41	12.43	4.29	12.79	8.34	26.16
	MF-based	GMF	14.10	2.81	14.33	14.28	4.08	14.73	8.29	29.51
		MLP	13.95	2.78	14.22	14.06	3.98	14.56	8.30	29.39
		NeuMF	16.43	3.20	16.87	16.73	4.68	17.40	9.69	33.23
		M2F	8.61	1.84	9.36	7.60	2.30	8.67	2.95	20.40
		MGMF	17.38	3.51	18.08	17.63	5.05	18.52	10.12	35.15
	CML-based	UniS	17.56	3.71	17.89	18.34	5.60	18.79	12.40	35.77
		PopS	12.96	3.11	13.30	12.82	4.41	13.40	7.59	28.61
		2stS	21.07	4.84	21.35	21.81	7.07	22.29	14.42	40.36
		HarS	24.88	5.86	25.38	24.89	8.25	25.77	15.74	45.15
		TransCF	10.03	1.84	10.31	10.90	3.09	11.20	7.07	23.66
		LRML	17.15	3.52	17.56	17.45	5.12	18.08	10.42	34.36
		AdaCML	19.06	4.12	19.31	19.74	6.23	20.20	13.30	37.36
		HLR	21.10	4.80	21.53	21.61	7.06	22.28	13.95	40.71
	Ours	DPCML1	19.12	4.14	19.34	19.90	6.27	20.29	13.24	37.55
		DPCML2	25.18	6.06	25.64	25.35	8.51	26.16	16.09	45.32
Steam-200k	Item-based	itemKNN	12.58	9.47	13.23	6.47	3.90	7.23	11.74	23.33
	MF-based	GMF	12.57	6.17	13.29	14.22	6.86	15.39	9.72	28.38
		MLP	17.07	9.63	17.49	16.89	8.49	17.67	15.15	34.54
		NeuMF	17.36	9.65	17.95	17.41	8.79	18.45	15.11	35.55
		M2F	11.33	5.69	11.95	11.44	5.73	12.98	6.43	25.05
		MGMF	12.51	6.14	13.25	14.45	6.88	15.55	9.63	28.40
	CML-based	UniS	20.71	11.97	21.42	20.92	10.36	21.61	18.88	40.10
		PopS	18.05	11.58	18.76	14.94	7.98	15.78	15.13	34.04
		2stS	25.20	14.62	26.20	23.97	11.91	25.35	21.48	46.17
		HarS	26.66	15.74	27.93	24.94	12.78	26.63	23.25	48.84
		TransCF	13.30	6.61	13.58	15.26	7.09	15.89	11.08	26.29
		LRML	14.91	7.48	15.43	16.49	8.06	17.51	12.24	31.89
		AdaCML	23.02	13.19	23.38	22.35	11.31	23.23	19.88	42.03
		HLR	20.30	11.65	20.96	19.79	9.88	20.94	17.06	39.26
	Ours	DPCML1	25.39	14.84	26.56	23.88	12.11	25.25	22.26	46.79
		DPCML2	29.88	17.13	31.22	28.70	14.51	30.56	24.10	51.95
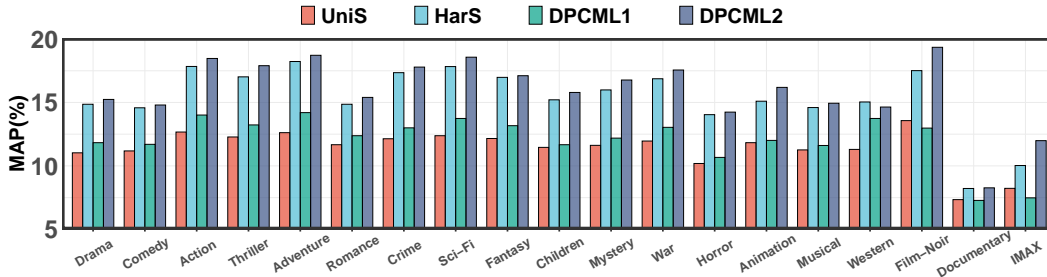
Figure 4: Fine-grained performance over each interest group on MovieLens-10m dataset.

performance of CML-based methods consistently surpasses the best MF-based competitors. This suggests that it is necessary to develop CML-based RS algorithms. 2) Our proposed method consistently surpasses all the competitors significantly on all datasets, except the results for MAP and MRR on CiteULike. Even for the failure results, the performance is fairly competitive compared with the competitors. This shows the effectiveness of our proposed algorithm. 3) Compared with studies targeting joint accessibility (i.e., M2F and MGMF), our proposed method significantly outperforms M2F and MGMF on all benchmark datasets. This shows the advantage of the CML-based paradigm that deserves more attention along this direction in future work.

4.3 Quantitative Analysis

Fine-grained Performance Comparison. Fig.4 presents the MAP metric over each interest group (movie genre) on MovieLens-10m. We can observe that our proposed framework could not only significantly outperform their single-vector counterparts in the majority interests but also improve the performance of minority groups in most cases. Especially, compared with HarS, the performance

Table 3: The diversity performance comparison of CML-based algorithms on Steam-200k and MovieLens-1m datasets. Here a higher value implies more diverse recommendation results.

Steam-200k				
Method	MaxDiv@3	MaxDiv@5	MaxDiv@10	MaxDiv@20
UniS	1.354	4.750	23.520	117.927
HarS	1.752	6.809	40.378	236.794
DPCML1 w/o DCRS	1.643	5.857	30.425	155.193
DPCML1	1.822	6.713	34.727	179.065
DPCML2 w/o DCRS	2.958	11.398	65.398	365.458
DPCML2	2.977	11.472	65.952	369.876
MovieLens-1m				
UniS	1.739	6.142	30.127	140.095
HarS	2.443	8.826	46.390	244.078
DPCML1 w/o DCRS	1.623	5.857	29.500	140.057
DPCML1	1.744	6.195	30.755	145.615
DPCML2 w/o DCRS	2.827	10.423	55.612	292.089
DPCML2	3.144	11.498	60.696	313.086

improvement of DPCML2 on minority interests is sharp. This shows that DPCML could reasonably focus on potentially interesting items even with the imbalanced item distribution.

Recommendation Diversity Evaluation. We test the performance of DPCML against CML-based competitors with *max-sum diversification (MaxDiv)* [4]. The diversity results are shown in Tab.3. We observe that: a) For methods within the same negative sampling strategy (i.e., UniS and DPCML1, HarS and DPCML2), our proposed DPCML could achieve relatively higher max-sum values. This suggests the improvement of DPCML in terms of promoting recommendation diversity. b) In most cases (except for DPCML1 w/o DCRS on the MovieLens-1m dataset), DPCML outperforms other competitors even without regularization. c) Most importantly, equipped with the regularization term DCRS, DPCML could achieve better diversification results against w/o DCRS.

Effect of the Diversity Control Regularization. Fig.5 illustrates a 3D-barplot based on the results of grid search on Steam-200k. From the results, we can observe that the proposed regularization scheme could significantly boost performance on all metrics. Moreover, there would induce different performances with different diversity values. This suggests that controlling a proper diversity of the embeddings for the same user is essential to accommodate their preferences better.

Empirical Justification of Corol.1. Fig.7 shows the empirical results on Steam-200k dataset. Based on these results, we can see that, with the increase of C, the empirical risk (i.e., training loss) of DPCML ($C > 1$) is significantly smaller than CML ($C = 1$). In addition, DPCML could substantially improve the performance of the validation/test set. Thus, we can conclude that DPCML could induce a smaller generalization error than traditional CML. This is consistent with Corol.1.

Sensitive Analysis of C. Fig.8 demonstrates the performance of DPCML methods with different C on Steam-200k dataset. We observe that a proper C could significantly improve the performance. Besides, leveraging C too aggressively for DPCML2 may adversely hurt the performance since models optimized with hard samples are more likely to lead to the over-fitting problem with the increasing parameters.

Sensitivity analysis of η. We investigate the sensitivity of $\eta \in \{0, 1, 3, 5, 10, 20, 30\}$ for recommendation results on the Steam-200k dataset. The experimental results are listed in Tab.6 and Tab.7. We can conclude that a proper η (roughly 10) could significantly improve the performance, suggesting the essential role of the proposed diversity control regularization scheme.
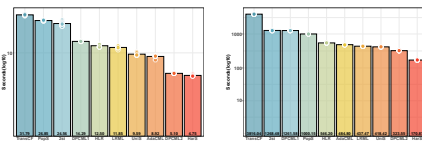
Training Efficiency. Since DPCML includes multiple user representations, it will inevitably introduce extra complexity to the overall optimization. We further investigate the efficiency of our proposed algorithm, presented in Fig.6. This trend suggests that our proposed algorithm could achieve competitive performance with acceptable efficiency.

Ablation Studies of DCRS. In order to show the effectiveness of our proposed DCRS, we compare its performance with the its three variants: **a) w/o DCRS**. **b) DCRS** $- \delta_1$ and **c) DCRS** $- \delta_2$. Please refer to Appendix.C.6.8 to the details of each variant. The empirical results on Steam-200k dataset are provided in Tab.8 and Tab.9. From the above results, we can see that: In most cases, only
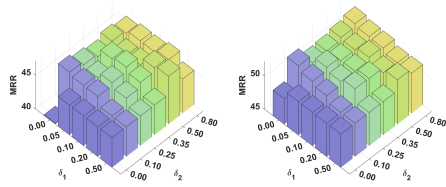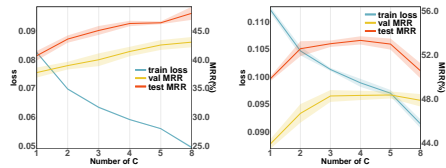
(a) DPCML1 (MRR)　　(b) DPCML2 (MRR)　　　(a) CiteULike　　(b) MovieLens-10m

Figure 5: Sensitivity analysis about δ_1 and δ_2 on Steam-200k datasets.

Figure 6: Training efficiency comparison among CML-based competitors.

(a) DPCML1　　(b) DPCML2　　　(a) P@3　　(b) P@5

Figure 7: Empirical justification of Corol.1　Figure 8: Sensitive Analysis of different C.
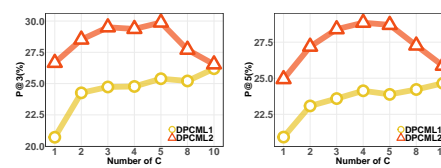
employing one of the two terms of DCRS could still improve the recommendation performance. However, none of them could outperform our proposed method. This strengthens the effectiveness of our proposed regularization scheme.

DCRS for MF-based Systems. We attempt to apply the proposed diversity control regularization scheme (DCRS) for M2F [52, 12]. In addition, we further explore the effectiveness of DCRS for the general framework of joint accessibility (GFJA, Eq.(9) in the main paper). The experimental results are summarized in Tab.10. The overall performance suggests the superiority of our proposed method against the current multi-vector-based competitors.

5　Conclusion & Future Remarks

This paper pays attention to developing an effective CML-based algorithm when users have multiple categories of interests. First, we point out that the current CML framework might induce preference bias, especially when the item category distribution is imbalanced. To this end, we propose a novel algorithm called DPCML. The key idea is to include multiple representations for each user in the model design. Meanwhile, a novel *diversity control regularization* scheme is specifically tailored to serve our purpose better. To see the generalization ability of DPCML on unseen test data, we also provide high probability upper bounds for the generalization error. Finally, the experiments over a range of benchmark datasets speak to the efficacy of DPCML. However, following the paradigm of CML, a possible limitation of DPCML is that it generally applies to implicit feedback but not explicit feedback since CML only cares about the relative preference ranking instead of concrete magnitude. In the future, we will explore how to improve the recommendation diversity based on explicit feedback.

6　Acknowledgements

References

[1] Askari, B., Szlichta, J., and Salehi-Abari, A. Variational autoencoders for top-k recommendation with implicit feedback. In *SIGIR*, pp. 2061–2065, 2021.

[2] Bao, S., Xu, Q., Ma, K., Yang, Z., Cao, X., and Huang, Q. Collaborative preference embedding against sparse labels. In *ACM MM*, pp. 2079–2087, 2019.

[3] Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. In *COLT*, volume 2111, pp. 224–240, 2001.

[4] Borodin, A., Lee, H. C., and Ye, Y. Max-sum diversification, monotone submodular functions and dynamic updates. In *SIGMOD/PODS*, pp. 155–166, 2012.

[5] Canévet, O. and Fleuret, F. Efficient sample mining for object detection. In *ACML*, pp. 48–63, 2014.

[6] Chen, J., Lian, D., and Zheng, K. Improving one-class collaborative filtering via ranking-based implicit regularizer. In *AAAI*, pp. 37–44, 2019.

[7] Curmei, M., Dean, S., and Recht, B. Quantifying availability and discovery in recommender systems via stochastic reachability. In *ICML*, pp. 2265–2275, 2021.

[8] Dar, Y., Muthukumar, V., and Baraniuk, R. G. A farewell to the bias-variance tradeoff? an overview of the theory of overparameterized machine learning. 2021.

[9] Dean, S., Rich, S., and Recht, B. Recommendations and user agency: the reachability of collaboratively-filtered information. In *FAT*, pp. 436–445, 2020.

[10] Ding, J., Quan, Y., He, X., Li, Y., and Jin, D. In *IJCAI*, pp. 2230–2236, 2019.

[11] Gajic, B., Amato, A., and Gatta, C. Fast hard negative mining for deep metric learning. *Pattern Recognition*, 112:107795, 2021.

[12] Guo, W., Krauth, K., Jordan, M. I., and Garg, N. The stereotyping problem in collaboratively filtered recommender systems. In *EAAMO*, pp. 6:1–6:10, 2021.

[13] He, X., Zhang, H., Kan, M., and Chua, T. Fast matrix factorization for online recommendation with implicit feedback. In *SIGIR*, pp. 549–558, 2016.

[14] He, X., Liao, L., Zhang, H., Nie, L., Hu, X., and Chua, T. Neural collaborative filtering. In *WWW*, pp. 173–182, 2017.

[15] He, X., Du, X., Wang, X., Tian, F., Tang, J., and Chua, T. Outer product-based neural collaborative filtering. In *IJCAI*, pp. 2227–2233, 2018.

[16] Heckel, R. and Ramchandran, K. The sample complexity of online one-class collaborative filtering. In *ICML*, pp. 1452–1460, 2017.

[17] Henriques, J. F., Carreira, J., Caseiro, R., and Batista, J. Beyond hard negative mining: Efficient detector learning via block-circulant decomposition. In *ICCV*, pp. 2760–2767, 2013.

[18] Hsieh, C.-K., Yang, L., Cui, Y., Lin, T.-Y., Belongie, S., and Estrin, D. Collaborative metric learning. In *WWW*, pp. 193–201, 2017.

[19] Jiang, M., Cui, P., Chen, X., Wang, F., Zhu, W., and Yang, S. Social recommendation with cross-domain transferable knowledge. *IEEE TKDE*, 27(11):3084–3097, 2015.

[20] Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[21] Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: isoperimetry and processes*. 1991.

[22] Lee, D., Kang, S., Ju, H., Park, C., and Yu, H. Bootstrapping user and item representations for one-class collaborative filtering. In *SIGIR*, pp. 1513–1522, 2021.

[23] Lei, Y., Ding, L., and Bi, Y. Local rademacher complexity bounds based on covering numbers. *Neurocomputing*, 218:320–330, 2016.

[24] Lei, Y., Ledent, A., and Kloft, M. Sharper generalization bounds for pairwise learning. In *NeurIPs*, 2020.

[25] Li, S. and Liu, Y. Sharper generalization bounds for clustering. In *ICML*, pp. 6392–6402, 2021.

[26] Li, Z., Xu, Q., Jiang, Y., Ma, K., Cao, X., and Huang, Q. Neural collaborative preference learning with pairwise comparisons. *IEEE Trans. Multim.*, 23:1977–1989, 2021.

[27] Linden, G., Smith, B., and York, J. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, (1):76–80, 2003.

[28] Long, P. M. and Sedghi, H. Generalization bounds for deep convolutional neural networks. In *ICLR 2020*, 2020.

[29] Lv, Y., Zheng, Y., Wei, F., Wang, C., and Wang, C. AICF: attention-based item collaborative filtering. *Adv. Eng. Informatics*, 44:101090:1–11, 2020.

[30] Ma, J., Zhou, C., Cui, P., Yang, H., and Zhu, W. Learning disentangled representations for recommendation. In *NeurIPS*, pp. 5712–5723, 2019.

[31] Ma, J., Zhou, C., Yang, H., Cui, P., Wang, X., and Zhu, W. Disentangled self-supervision in sequential recommenders. In *KDD*, pp. 483–491, 2020.

[32] McDiarmid, C. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pp. 195–248. 1998.

[33] Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. MIT Press, 2012.

[34] Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. In *ICLR*, 2020.

[35] Pan, R., Zhou, Y., Cao, B., Liu, N. N., Lukose, R. M., Scholz, M., and Yang, Q. One-class collaborative filtering. In *ICDM*, pp. 502–511, 2008.

[36] Park, C., Kim, D., Xie, X., and Yu, H. Collaborative translational metric learning. In *ICDM*, pp. 367–376, 2018.

[37] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

[38] Rendle, S. and Freudenthaler, C. Improving pairwise learning for item recommendation from implicit feedback. In *WSDM*, pp. 273–282, 2014.

[39] Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. BPR: bayesian personalized ranking from implicit feedback. In *UAI*, pp. 452–461, 2009.

[40] Takács, G. and Tikk, D. Alternating least squares for personalized ranking. In *RecSys*, pp. 83–90, 2012.

[41] Tay, Y., Tuan, L. A., and Hui, S. C. Latent relational metric learning via memory-based attention for collaborative ranking. In *WWW*, pp. 729–739, 2018.

[42] Togashi, R., Kato, M., Otani, M., and Satoh, S. Density-ratio based personalised ranking from implicit feedback. In *WWW*, pp. 3221–3233, 2021.

[43] Tran, V., Salha-Galvan, G., Hennequin, R., and Moussallam, M. Hierarchical latent relation modeling for collaborative metric learning. In *RecSys*, pp. 302–309, 2021.

[44] Tran, V.-A., Hennequin, R., Royo-Letelier, J., and Moussallam, M. Improving collaborative metric learning with efficient negative sampling. In *SIGIR*, pp. 1201–1204, 2019.

[45] Wang, C., Zhou, T., Chen, C., Hu, T., and Chen, G. Off-policy recommendation system without exploration. In *PAKDD*, volume 12084, pp. 16–27. Springer, 2020.

[46] Wang, H., Chen, B., and Li, W. Collaborative topic regression with social regularization for tag recommendation. In *IJCAI*, pp. 2719–2725, 2013.

[47] Wang, H., Li, Y., and Frimpong, F. Group recommendation via self-attention and collaborative metric learning model. *IEEE Access*, 7:164844–164855, 2019.

[48] Wang, M., Gong, M., Zheng, X., and Zhang, K. Modeling dynamic missingness of implicit feedback for recommendation. In *NeurIPS*, pp. 6670–6679, 2018.

[49] Wang, W., Feng, F., He, X., Nie, L., and Chua, T. Denoising implicit feedback for recommendation. In *WSDM*, pp. 373–381, 2021.

[50] Wang, X., He, X., Wang, M., Feng, F., and Chua, T. Neural graph collaborative filtering. In *ACM SIGIR*, pp. 165–174, 2019.

[51] Wang, X., Wang, R., Shi, C., Song, G., and Li, Q. Multi-component graph convolutional collaborative filtering. In *AAAI*, pp. 6267–6274, 2020.

[52] Weston, J., Weiss, R. J., and Yee, H. Nonlinear latent factorization by embedding multiple user interests. In *RecSys*, pp. 65–68, 2013.

[53] Wu, G., Volkovs, M., Soon, C. L., Sanner, S., and Rai, H. Noise contrastive estimation for one-class collaborative filtering. In *SIGIR*, pp. 135–144, 2019.

[54] Wu, H., Zhou, Q., Nie, R., and Cao, J. Effective metric learning with co-occurrence embedding for collaborative recommendations. *Neural Networks*, 124:308–318, 2020.

[55] Wu, Q., Zhang, H., Gao, X., Yan, J., and Zha, H. Towards open-world recommendation: An inductive model-based collaborative filtering approach. In *ICML*, pp. 11329–11339, 2021.

[56] Xu, D., Ruan, C., Körpeoglu, E., Kumar, S., and Achan, K. Rethinking neural vs. matrix-factorization collaborative filtering: the theoretical perspectives. In *ICML*, pp. 11514–11524, 2021.

[57] Yang, T. and Ying, Y. AUC maximization in the era of big data and AI: A survey. *ACM Computing Surveys*, 2022.

[58] Yang, Z., Ding, M., Zhou, C., Yang, H., Zhou, J., and Tang, J. Understanding negative sampling in graph representation learning. In *KDD*, pp. 1666–1676, 2020.

[59] Yao, Y., Tong, H., Yan, G., Xu, F., Zhang, X., Szymanski, B. K., and Lu, J. Dual-regularized one-class collaborative filtering with implicit feedback. *WWW*, 22(3):1099–1129, 2019.

[60] Zhang, Q. and Ren, F. Prior-based bayesian pairwise ranking for one-class collaborative filtering. *Neurocomputing*, 440:365–374, 2021.

[61] Zhang, Q. and Ren, F. Double bayesian pairwise learning for one-class collaborative filtering. *Knowl. Based Syst.*, 229:107339, 2021.

[62] Zhang, T., Zhao, P., Liu, Y., Xu, J., Fang, J., Zhao, L., Sheng, V. S., and Cui, Z. Adacml: Adaptive collaborative metric learning for recommendation. In *DASFAA*, volume 11447, pp. 301–316, 2019.

[63] Zheng, Y., Tang, B., Ding, W., and Zhou, H. A neural autoregressive approach to collaborative filtering. In *ICML*, pp. 764–773, 2016.

[64] Zhou, D. The covering number in learning theory. *J. Complex.*, 18(3):739–767, 2002.

[65] Zhou, X., Liu, D., Lian, J., and Xie, X. Collaborative metric learning with memory network for multi-relational recommender systems. In *IJCAI*, pp. 4454–4460, 2019.

Checklist

1. For all authors...

    (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

    (b) Did you describe the limitations of your work? [N/A]

    (c) Did you discuss any potential negative societal impacts of your work? [No] There are no immediate societal impacts associated with our research.

    (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

    (a) Did you state the full set of assumptions of all theoretical results? [Yes]

    (b) Did you include complete proofs of all theoretical results? [Yes] See the supplementary materials.

3. If you ran experiments...

    (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]

    (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Sec.C.4 in the Appendix.

    (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

    (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

    (a) If your work uses existing assets, did you cite the creators? [Yes]

    (b) Did you mention the license of the assets? [Yes] See Sec.C

    (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

    (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] We use open-source software and datasets.

    (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

    (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

    (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

    (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Contents

A Prior Arts

In this section, we briefly review the closely related studies along with our main topic.

A.1 One-Class Collaborative Filtering

In many real-world applications, the vast majority of interactions are implicitly expressed by users' behaviors, e.g., downloads of movies, clicks of products and browses of news. In order to develop RS from such implicit feedback, researchers usually formulate the recommendation task as the *One-Class Collaborative Filtering* (OCCF) problem [35, 59, 16, 61, 22, 60].

Matrix Factorization (MF) based Algorithm. Over the past decades, the Matrix Factorization (MF)-based algorithms are one of the most classical OCCF solutions [15, 63, 39, 6]. The key idea of MF is to factorize the user-item interaction into an inner product between user and item embeddings, respectively. For example, [13] proposes an item-oriented MF method with implicit feedback. Neural Collaborative Filtering (NCF) [14] develops a general framework that unifies the MF and the neural networks together, and then regards the recommendation task as a regression problem.

Collaborative Metric Learning based Algorithm. As mentioned in the Sec.1, *Collaborative Metric Learning* (CML) [18] presents a very early trial to replace the inner product scoring function with Euclidean distance. Ever since CML [18] has shown great power in RS, there raises a new direction to develop recommendation algorithms. Typically, the existing methods fall into two camps. The first camp pursues model-based improvement over the original CML [36, 41], with a goal to improve the formulation of the latent embeddings. Representative work includes [36, 41, 62, 43]. Moreover, since CML employs a negative sampling strategy [58, 38] to mitigate the high computational burden of pairwise learning, another camp attempts to improve the effectiveness of the negative sampling process [5], such as uniform sampling [18], popularity-based sampling [53], two-stage negative sampling [44] and Hard negative sampling [11, 10].

This paper falls into the first camp. We focus on a challenging scenario where a user has multiple categories of interests. Unfortunately, in this case, the current literature equipped with unique user representation might induce preference bias, especially when the item category distribution is imbalanced.

A.2 Recommendation against Joint Accessibility

Recently, some studies [12, 7, 9] have pointed out a *joint accessibility* problem in the recommendation, which determines the opportunities of users to discover interesting content. More precisely, joint accessibility measures whether an item candidate with size K could be jointly accessed by a user in a Top-K recommendation [12]. In other words, joint accessibility also somewhat captures a fundamental requirement of content diversity. If there are sufficient preference records of a target user, he/she should be able to be recommended any combination of K items that he/she may be interested in. In this direction, noteworthy is the work present in [12], which provides the theoretically necessary and sufficient conditions to meet the joint accessibility. Subsequently, [12] proposes an alternative MF-based model (M2F) to improve joint accessibility. Formally, with respect to each user, it assigns m feature vectors to users, and thus the predicted score of each item is defined as $s(j) = \max_{i \in [m]} \boldsymbol{u}_i^\top \boldsymbol{v}_j$, where $\boldsymbol{u}_i, i \in [m]$ is the i-th user latent vector; $\boldsymbol{v}_j$ is the item feature and $[m] = \{1, \ldots, m\}$. Finally, M2F adopts the least square [40] loss to recover the missing values in the user-item matrix.

The existing line of such work merely focuses on the MF-based algorithms, while we take a further step to explore the problem under the context of CML. It is also interesting to note that, under mild conditions, we could see that M2F is a particular case of our method (shown in Sec.2.6). In this sense, we generalize the original idea of joint accessibility.

B Generalization Bounds and its proofs

B.1 Preliminary Lemmas

In this section, we first briefly review some preparatory knowledge for the proof.

Definition 4 (Bounded Difference Property). Given a group of independent random variables $X_1, X_2, \cdots, X_n$ where $X_t \in \mathbb{X}, \forall t$, $f(X_1, X_2, \cdots, X_n)$ is satisfied with the bounded difference property, if there exists some non-negative constants $c_1, c_2, \cdots, c_n$, such that:

$$\sup_{x_1, x_2, \cdots, x_n, x'_t} |f(x_1, \cdots, x_n) - f(x_1, \cdots, x_{t-1}, x'_t, \cdots, x_n)| \leq c_t, \ \forall t, 1 \leq t \leq n. \tag{14}$$

Hereafter, if any function f holds the Bounded Difference Property, the following Mcdiarmid's inequality is always satisfied.

Lemma 2 (Mcdiarmid's Inequality [32]). *Assume we have n independent random variables $X_1, X_2, \ldots, X_n$ that all of them are chosen from the set $\mathcal{X}$. For a function $f : \mathcal{X} \rightarrow \mathbb{R}$, $\forall t, 1 \leq t \leq n$, if the following inequality holds:*

$$\sup_{x_1, x_2, \cdots, x_n, x'_t} |f(x_1, \cdots, x_n) - f(x_1, \cdots, x_{t-1}, x'_t, \cdots, x_n)| \leq c_t, \ \forall t, 1 \leq t \leq n.$$

with $\boldsymbol{x} \neq \boldsymbol{x}'$, then for all $\epsilon > 0$, we have

$$\mathbb{P}[\mathbb{E}(f) - f \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{t=1}^{n} c_t^2}\right),$$

$$\mathbb{P}[f - \mathbb{E}(f) \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{t=1}^{n} c_t^2}\right).$$

Lemma 3 (Union bound/Boole's inequality). *Given the countable or finite set of events E_i, the probability that at least one event happens is less than or equal to the sum of all probabilities of the events happened individually, i.e.,*

$$\mathbb{P}\left[\underset{i}{\cup} E_i\right] \leq \sum_i \mathbb{P}[E_i] \tag{15}$$

Lemma 4 (ϕ-Lipschitz Continuous). *Given a set $\mathcal{X}$ and a function $f : \mathcal{X} \rightarrow \mathbb{R}$, if f is continuously differentiable on $\mathcal{X}$ such that, $\forall x, y \in \mathcal{X}$, the following condition holds with a real constant ϕ:*

$$\|f(x) - f(y)\| \leq \phi \|x - y\|.$$

Thereafter, f is said to be a ϕ-Lipschitz continuous function.

B.2 Key Lemmas

Restate of Definition 2 (ϵ-Covering). [21] Let $(\mathcal{F}, \rho)$ be a (pesudo) metric space, and $\mathcal{G} \subseteq \mathcal{F}$. $\{f_1, \ldots, f_n\}$ is said to be an ϵ-covering of $\mathcal{G}$ if $\mathcal{G} \subseteq \overset{n}{\underset{i=1}{\cup}} \mathcal{B}(f_i, \epsilon)$, i.e., $\forall g \in \mathcal{G}, \exists i$ such that $\rho(g, f_i) \leq \epsilon$.

Restate of Definition 3 (Covering Number). [21] According to the notations in Def.B.2, the covering number of $\mathcal{G}$ with radius ϵ is defined as:

$$\mathcal{N}(\epsilon; \mathcal{G}, \rho) = \min\{n : \exists \epsilon - covering \ over \ \mathcal{G} \ with \ size \ n\}$$

Restate of Assumption 1 (Basic Assumptions). We assume that all the embeddings of users and items are chosen from the following embedding hypothesis space:

$$\mathcal{H}_R = \left\{\boldsymbol{g} : \boldsymbol{g} \in \mathbb{R}^d, \|\boldsymbol{g}\| \leq r\right\}, \tag{16}$$

where $\boldsymbol{g}_{u_i}^c \in \mathcal{H}_R, u_i \in \mathcal{U}, c \in [C]$ and $\boldsymbol{g}_{v_j} \in \mathcal{H}_R, v_j \in \mathcal{I}$.

Restate of Lemma 1. [28, 25, 64] The covering number of the hypothesis class $\mathcal{H}_R$ has the following upper bound:

$$\log \mathcal{N}(\epsilon; \mathcal{H}_R, \rho) \leq d \log\left(\frac{3r}{\epsilon}\right), \tag{17}$$

where d is the dimension of embedding space.

In what follows, we will present the key lemmas to derive the upper bounds.

Lemma 5. *Let ε be the generalization error between $\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})$ and $\mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})]$. Then by constructing an σ-covering $\{\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_n\}$ of $\mathcal{H}_R$ with $\sigma = \frac{\varepsilon}{16r(4+\eta)}$, the following inequality holds*

$$\mathbb{P}\left[\sup_{\boldsymbol{g}\in\mathcal{B}(\boldsymbol{g}_l,\sigma)}\left|\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) - \mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})]\right| \leq \varepsilon\right] \geq \mathbb{P}\left[\left|\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}_l) - \mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}_l)]\right| \leq \frac{\varepsilon}{2}\right], \;\; \forall l \in [n], \qquad (18)$$

Proof. Assume there exists an σ-covering $\{\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_n\}$ of $\mathcal{H}_R$. To prove (18), we turn to prove the following inequality:

$$\left|\,|\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) - \mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})]| - |\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}_l) - \mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}_l)]|\,\right| \leq \frac{\varepsilon}{2}, \; \forall l \in [n]. \qquad (19)$$

Note that, we have

$$\left|\,|\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) - \mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})]| - |\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}_l) - \mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}_l)]|\,\right| \overset{(**)}{\leq} \left|\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) - \mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})] - \left(\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}_l) - \mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}_l)]\right)\right|$$

$$\overset{(*)}{\leq} \left|\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) - \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}_l)\right| + \left|\mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}_l)] - \mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})]\right|, \qquad (20)$$

where $(*)$ and $(**)$ follows the facts $|x+y| \leq |x| + |y|$ and $||x| - |y|| \leq |x-y|$, respectively.

Then, to achieve (19), we only need to show that the following inequation holds:

$$\left|\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) - \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}_l)\right| \leq \frac{\varepsilon}{4}, \;\; \forall l \in [n]. \qquad (21)$$

Recall that

$$\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) = \hat{\mathcal{R}}_{\mathcal{D},\boldsymbol{g}} + \eta \cdot \hat{\Omega}_{\mathcal{D},\boldsymbol{g}}, \qquad (22)$$

where

$$\hat{\mathcal{R}}_{\mathcal{D},\boldsymbol{g}} = \frac{1}{|\mathcal{U}|}\sum_{u_i\in\mathcal{U}}\frac{1}{n_i^+ n_i^-}\sum_{j=1}^{n_i^+}\sum_{k=1}^{n_i^-}\ell_g^{(i)}(v_j^+, v_k^-),$$

$$\ell_g^{(i)}(v_j^+, v_k^-) = \max(0, \lambda + s(u_i, v_j^+) - s(u_i, v_k^-)), \qquad (23)$$

$$s(u_i, v_j) = \min_{c\in[C]}\|\boldsymbol{g}_{u_i}^c - \boldsymbol{g}_{v_j}\|^2, \forall\, v_j, v_j \in \mathcal{I}$$

and

$$\hat{\Omega}_{\mathcal{D},\boldsymbol{g}} = \frac{1}{|\mathcal{U}|}\sum_{u_i\in\mathcal{U}}\left(\max\left(0, \delta_1 - \delta_{\boldsymbol{g},u_i}\right) + \max(0, \delta_{\boldsymbol{g},u_i} - \delta_2)\right),$$

$$\delta_{\boldsymbol{g},u_i} = \frac{1}{2C(C-1)}\sum_{c_1,c_2\in C}\|\boldsymbol{g}_{u_i}^{c_1} - \boldsymbol{g}_{u_i}^{c_2}\|^2. \qquad (24)$$

Let us define some intermediate variables:

$$\hat{\mathcal{R}}_{\mathcal{D}_{u_i},\boldsymbol{g}} = \frac{1}{n_i^+ n_i^-}\sum_{j=1}^{n_i^+}\sum_{k=1}^{n_i^-}\ell_g^{(i)}(v_j^+, v_k^-),$$

$$\hat{\Omega}_{\mathcal{D}_{u_i},\boldsymbol{g}} = \max\left(0, \delta_1 - \delta_{\boldsymbol{g},u_i}\right) + \max\left(0, \delta_{\boldsymbol{g},u_i} - \delta_2\right), \qquad (25)$$

$$\Delta_{\boldsymbol{g},\boldsymbol{g}_l}(c_1, c_2) = \left(\|\boldsymbol{g}_{u_i}^{c_1} - \boldsymbol{g}_{v_j^+}\|^2 - \|\tilde{\boldsymbol{g}}_{u_i}^{c_2} - \tilde{\boldsymbol{g}}_{v_j^+}\|^2\right).$$

In this sense, we have

$$\left|\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) - \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}_l)\right| = \left|\hat{\mathcal{R}}_{\mathcal{D},\boldsymbol{g}} + \eta \cdot \hat{\Omega}_{\mathcal{D},\boldsymbol{g}} - \hat{\mathcal{R}}_{\mathcal{D},\boldsymbol{g}_l} - \eta \cdot \hat{\Omega}_{\mathcal{D},\boldsymbol{g}_l}\right|$$

$$\leq \underbrace{\left|\hat{\mathcal{R}}_{\mathcal{D},\boldsymbol{g}} - \hat{\mathcal{R}}_{\mathcal{D},\boldsymbol{g}_l}\right|}_{(1)} + \underbrace{\eta\left|\Omega_{\mathcal{D},\boldsymbol{g}} - \cdot\Omega_{\mathcal{D},\boldsymbol{g}_l}\right|}_{(2)} \qquad (26)$$

18

Subsequently, in terms of (1), we first consider a specific user u_i with her/his corresponding interaction records $\mathcal{D}_{u_i}$. We have

$$
\begin{aligned}
\left|\hat{\mathcal{R}}_{\mathcal{D}_{u_i},\boldsymbol{g}} - \hat{\mathcal{R}}_{\mathcal{D}_{u_i},\boldsymbol{g}_l}\right| &= \left| \frac{1}{n_i^+ n_i^-} \sum_{j=1}^{n_i^+} \sum_{k=1}^{n_i^-} \ell_{\boldsymbol{g}}^{(i)}(v_j^+, v_k^-) - \frac{1}{n_i^+ n_i^-} \sum_{j=1}^{n_i^+} \sum_{k=1}^{n_i^-} \ell_{\boldsymbol{g}_l}^{(i)}(v_j^+, v_k^-) \right| \\
&\leq \frac{1}{n_i^+ n_i^-} \sum_{j=1}^{n_i^+} \sum_{k=1}^{n_i^-} \left| \ell_{\boldsymbol{g}}^{(i)}(v_j^+, v_k^-) - \ell_{\boldsymbol{g}_l}^{(i)}(v_j^+, v_k^-) \right| \\
&\overset{(a)}{\leq} \frac{1}{n_i^+ n_i^-} \sum_{j=1}^{n_i^+} \sum_{k=1}^{n_i^-} \left| s(u_i, v_j^+) - s(u_i, v_k^-) - \tilde{s}_l(u_i, v_j^+) + \tilde{s}_l(u_i, v_k^-) \right| \\
&\overset{(*)}{\leq} \frac{1}{n_i^+ n_i^-} \sum_{j=1}^{n_i^+} \sum_{k=1}^{n_i^-} \left(\left| s(u_i, v_j^+) - \tilde{s}_l(u_i, v_j^+) \right| + \left| \tilde{s}_l(u_i, v_k^-) - s(u_i, v_k^-) \right| \right)
\end{aligned}
\tag{27}
$$

where (a) follows the Lem.4 and $\ell_{\boldsymbol{g}}^{(i)}$ is apparently a 1-Lipschitz continuous function.

In terms of $\left| s(u_i, v_j^+) - \tilde{s}_l(u_i, v_j^+) \right|$, the following equation holds:

$$
\begin{aligned}
\left| s(u_i, v_j^+) - \tilde{s}_l(u_i, v_j^+) \right| &= \left| \min_{c_1 \in [C]} \|\boldsymbol{g}_{u_i}^{c_1} - \boldsymbol{g}_{v_j^+}\|^2 - \min_{c_2 \in [C]} \|\tilde{\boldsymbol{g}}_{u_i}^{c_2} - \tilde{\boldsymbol{g}}_{v_j^+}\|^2 \right| \\
&= \left| \min_{c_1 \in [C]} \max_{c_2 \in [C]} \Delta_{\boldsymbol{g},\boldsymbol{g}_l}(c_1, c_2) \right| \\
&= \max \left\{ \min_{c_1 \in [C]} \max_{c_2 \in [C]} \Delta_{\boldsymbol{g},\boldsymbol{g}_l}(c_1, c_2), \max_{c_1 \in [C]} \min_{c_2 \in [C]} \Delta_{\boldsymbol{g}_l,\boldsymbol{g}}(c_2, c_1) \right\}
\end{aligned}
\tag{28}
$$

Moreover, we have

$$
\begin{aligned}
&\min_{c_1 \in [C]} \max_{c_2 \in [C]} \Delta_{\boldsymbol{g},\boldsymbol{g}_l}(c_1, c_2) \\
&\leq \max_{c_1 = c_2, c_1, c_2 \in [C]} \Delta_{\boldsymbol{g},\boldsymbol{g}_l}(c_1, c_2) \\
&\leq \max_{c_1 = c_2, c_1, c_2 \in [C]} \left| \Delta_{\boldsymbol{g},\boldsymbol{g}_l}(c_1, c_2) \right| \\
&= \max_{c \in [C]} \left| \left(\|\boldsymbol{g}_{u_i}^c - \boldsymbol{g}_{v_j^+}\| + \|\tilde{\boldsymbol{g}}_{u_i}^c - \tilde{\boldsymbol{g}}_{v_j^+}\| \right) \left(\|\boldsymbol{g}_{u_i}^c - \boldsymbol{g}_{v_j^+}\| - \|\tilde{\boldsymbol{g}}_{u_i}^c - \tilde{\boldsymbol{g}}_{v_j^+}\| \right) \right| \\
&\overset{(**)}{\leq} \max_{c \in [C]} \left(\|\boldsymbol{g}_{u_i}^c - \boldsymbol{g}_{v_j^+}\| + \|\tilde{\boldsymbol{g}}_{u_i}^c - \tilde{\boldsymbol{g}}_{v_j^+}\| \right) \left(\|\boldsymbol{g}_{u_i}^c - \boldsymbol{g}_{v_j^+} - \tilde{\boldsymbol{g}}_{u_i}^c + \tilde{\boldsymbol{g}}_{v_j^+}\| \right) \\
&\leq 4r \left(\max_{c \in [C]} \|\boldsymbol{g}_{u_i}^c - \tilde{\boldsymbol{g}}_{u_i}^c\| + \|\tilde{\boldsymbol{g}}_{v_j^+} - \boldsymbol{g}_{v_j^+}\| \right) \\
&\leq 8r\sigma
\end{aligned}
\tag{29}
$$

where $(**)$ follows the fact $\|x\| - \|y\| \leq |x - y|$.

Similarly, we have

$$
\begin{aligned}
\left| \tilde{s}_l(u_i, v_k^-) - s(u_i, v_k^-) \right| &\leq 4r \left(\max_{c \in [C]} \|\boldsymbol{g}_{u_i}^c - \tilde{\boldsymbol{g}}_{u_i}^c\| + \|\tilde{\boldsymbol{g}}_{v_k^-} - \boldsymbol{g}_{v_k^-}\| \right) \\
&\leq 8r\sigma
\end{aligned}
\tag{30}
$$

Thus, we have

$$
\left| \hat{\mathcal{R}}_{\mathcal{D},g} - \hat{\mathcal{R}}_{\mathcal{D},\tilde{g}_l} \right| = \left| \hat{\mathcal{R}}_{\mathcal{D}_{u_i},\boldsymbol{g}} - \hat{\mathcal{R}}_{\mathcal{D}_{u_i},\boldsymbol{g}_l} \right| \leq 16r\sigma
\tag{31}
$$

Therefore, for all users, we also have

$$\left| \hat{\mathcal{R}}_{\mathcal{D},\boldsymbol{g}} - \hat{\mathcal{R}}_{\mathcal{D},\boldsymbol{g}_l} \right| \leq 16r\sigma. \tag{32}$$

With respect to (2), we also first consider a specific user u_i, i.e.,

$$
\begin{aligned}
\eta \left| \hat{\Omega}_{\mathcal{D}_{u_i},\boldsymbol{g}} - \hat{\Omega}_{\mathcal{D}_{u_i},\boldsymbol{g}_l} \right| &\overset{(a)}{\leq} 2\eta \left| \delta_{\boldsymbol{g},u_i} - \delta_{\boldsymbol{g}_l,u_i} \right| \\
&= \frac{\eta}{C(C-1)} \left| \sum_{c_1,c_2 \in C} \left\| \boldsymbol{g}_{u_i}^{c_1} - \boldsymbol{g}_{u_i}^{c_2} \right\|^2 - \sum_{c_1,c_2 \in C} \left\| \tilde{\boldsymbol{g}}_{u_i}^{c_1} - \tilde{\boldsymbol{g}}_{u_i}^{c_2} \right\|^2 \right| \\
&\leq \frac{\eta}{C(C-1)} \left| \sum_{c_1,c_2 \in C} \left(\left\| \boldsymbol{g}_{u_i}^{c_1} - \boldsymbol{g}_{u_i}^{c_2} \right\| + \left\| \tilde{\boldsymbol{g}}_{u_i}^{c_1} - \tilde{\boldsymbol{g}}_{u_i}^{c_2} \right\| \right) \left(\left\| \boldsymbol{g}_{u_i}^{c_1} - \boldsymbol{g}_{u_i}^{c_2} \right\| - \left\| \tilde{\boldsymbol{g}}_{u_i}^{c_1} - \tilde{\boldsymbol{g}}_{u_i}^{c_2} \right\| \right) \right| \\
&\leq \frac{\eta}{C(C-1)} \sum_{c_1,c_2 \in C} \left| \left(\left\| \boldsymbol{g}_{u_i}^{c_1} - \boldsymbol{g}_{u_i}^{c_2} \right\| + \left\| \tilde{\boldsymbol{g}}_{u_i}^{c_1} - \tilde{\boldsymbol{g}}_{u_i}^{c_2} \right\| \right) \left(\left\| \boldsymbol{g}_{u_i}^{c_1} - \boldsymbol{g}_{u_i}^{c_2} \right\| - \left\| \tilde{\boldsymbol{g}}_{u_i}^{c_1} - \tilde{\boldsymbol{g}}_{u_i}^{c_2} \right\| \right) \right| \\
&\overset{(**)}{\leq} \frac{4\eta r}{C(C-1)} \sum_{c_1,c_2 \in C} \left(\left\| \boldsymbol{g}_{u_i}^{c_1} - \tilde{\boldsymbol{g}}_{u_i}^{c_1} \right\| + \left\| \tilde{\boldsymbol{g}}_{u_i}^{c_2} - \boldsymbol{g}_{u_i}^{c_2} \right\| \right) \\
&\leq 4\eta r \left(\max_{c \in [C]} \left\| \boldsymbol{g}_{u_i}^{c} - \tilde{\boldsymbol{g}}_{u_i}^{c} \right\| \right) \\
&\leq 4\eta r\sigma
\end{aligned}
\tag{33}
$$

where (a) follows the Lem.4 and $(**)$ follows $|\|x\| - \|y\|| \leq |x - y|$.

In like wise, we have

$$\eta \left| \hat{\Omega}_{\mathcal{D},\boldsymbol{g}} - \hat{\Omega}_{\mathcal{D},\boldsymbol{g}_l} \right| \leq 4\eta r\sigma. \tag{34}$$

Finally, based on (32) and (34), we have

$$\left| \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) - \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}_l) \right| \leq 4r\sigma(4 + \eta). \tag{35}$$

Based on this, by further choosing $\sigma = \frac{\varepsilon}{16r(4+\eta)}$, we could construct the covering number $\mathcal{N}_1$ and $\mathcal{N}_2$ with respect to users and items, respectively, i.e.,

$$
\begin{aligned}
&\mathcal{N}_1 \left(\frac{\varepsilon}{16r(4+\eta)}, \mathcal{H}_R, \rho_1 \right), \quad \rho_1 = \max_{c \in [C]} \left\| \boldsymbol{g}_{u_i}^{c} - \tilde{\boldsymbol{g}}_{u_i}^{c} \right\|, \quad \forall u_i \in \mathcal{U}, \\
&\mathcal{N}_2 \left(\frac{\varepsilon}{16r(4+\eta)}, \mathcal{H}_R, \rho_2 \right), \quad \rho_2 = \left\| \tilde{\boldsymbol{g}}_{v_j} - \boldsymbol{g}_{v_j} \right\|, \quad \forall v_j \in \mathcal{I},
\end{aligned}
\tag{36}
$$

such that the following inequality holds:

$$\left| \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) - \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}_l) \right| \leq \frac{\varepsilon}{4}.$$

This completed the proof. □

Lemma 6 (Bounded Difference Property of DPCML). *Let $\mathcal{D}$ and $\mathcal{D}'$ be two independent datasets where exactly one instance is different instead of a term. We conclude that $\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})$ satisfies the bounded difference property (Lem.4).*

Proof. We need to seek the upper bound of

$$\sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| \hat{\mathcal{L}}_{\mathcal{D}'}(\boldsymbol{g}) - \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) \right|.$$

To achieve this, notice that, such difference between $\mathcal{D}$ and $\mathcal{D}'$ could be caused by either the user side or the item side. Therefore, we have the following three possible cases:

- **Case 1:** Only one user is different, i.e.,

$$\mathcal{D} = \bigcup_{u_i \in \mathcal{U}} \mathcal{D}_{u_i}, \quad \mathcal{D}' = (\mathcal{D} \setminus \mathcal{D}_{u_t}) \cup \mathcal{D}_{u'_t}, \ \ \forall t, t = 1, 2, \ldots, |\mathcal{U}|. \tag{37}$$

Under this circumstance, we have

$$\sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| \hat{\mathcal{L}}_{\mathcal{D}'}(\boldsymbol{g}) - \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) \right| \overset{(b)}{\leq} \underbrace{\sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| \hat{\mathcal{R}}_{\mathcal{D},\boldsymbol{g}} - \hat{\mathcal{R}}_{\mathcal{D}',\boldsymbol{g}} \right|}_{(3)} + \underbrace{\sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| \eta \hat{\Omega}_{\mathcal{D},\boldsymbol{g}} - \eta \hat{\Omega}_{\mathcal{D}',\boldsymbol{g}} \right|}_{(4)} \tag{38}$$

where (b) is achieved by the inequality: $\sup(x + y) \leq \sup(x) + \sup(y)$.

Based on (38), in what follows, we will show the upper bound of term (3) and (4), respectively.

At first, we define some intermediate variables:

$$\hat{\mathcal{R}}_{\mathcal{D}_{u_i},\boldsymbol{g}} = \frac{1}{n_i^+ n_i^-} \sum_{j=1}^{n_i^+} \sum_{k=1}^{n_i^-} \ell_{\boldsymbol{g}}^{(i)}(v_j^+, v_k^-),$$

$$\phi_{\boldsymbol{g}}(c_1, c_2) = \|\boldsymbol{g}_{u_t}^{c_1} - \boldsymbol{g}_{v_j^+}\|^2 - \|\boldsymbol{g}_{u'_t}^{c_2} - \boldsymbol{g}_{v_j^+}\|^2, \forall c_1, c_2, c_1, c_2 \in [C]$$

Then, with respect to term (3), we have

$$\sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| \hat{\mathcal{R}}_{\mathcal{D},\boldsymbol{g}} - \hat{\mathcal{R}}_{\mathcal{D}',\boldsymbol{g}} \right| = \frac{1}{|\mathcal{U}|} \sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| \hat{\mathcal{R}}_{\mathcal{D}_{u_t},\boldsymbol{g}} - \hat{\mathcal{R}}_{\mathcal{D}_{u'_t},\boldsymbol{g}} \right|$$

$$= \frac{1}{|\mathcal{U}|} \sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| \frac{1}{n_i^+ n_i^-} \sum_{j=1}^{n_i^+} \sum_{k=1}^{n_i^-} \ell_{\boldsymbol{g}}^{(t)}(v_j^+, v_k^-) - \frac{1}{n_i^+ n_i^-} \sum_{j=1}^{n_i^+} \sum_{k=1}^{n_i^-} \ell_{\boldsymbol{g}}^{(t')}(v_j^+, v_k^-) \right|$$

$$\leq \frac{1}{|\mathcal{U}|} \frac{1}{n_i^+ n_i^-} \sup_{\boldsymbol{g} \in \mathcal{H}_R} \sum_{j=1}^{n_i^+} \sum_{k=1}^{n_i^-} \left| \ell_{\boldsymbol{g}}^{(t)}(v_j^+, v_k^-) - \ell_{\boldsymbol{g}}^{(t')}(v_j^+, v_k^-) \right|$$

$$\overset{(b)}{\leq} \frac{1}{|\mathcal{U}|} \frac{1}{n_i^+ n_i^-} \sum_{j=1}^{n_i^+} \sum_{k=1}^{n_i^-} \sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| \ell_{\boldsymbol{g}}^{(t)}(v_j^+, v_k^-) - \ell_{\boldsymbol{g}}^{(t')}(v_j^+, v_k^-) \right|$$

$$\overset{(a)}{\leq} \frac{1}{|\mathcal{U}|} \frac{1}{n_i^+ n_i^-} \sum_{j=1}^{n_i^+} \sum_{k=1}^{n_i^-} \sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| s(u_t, v_j^+) - s(u_t, v_k^-) - \left(s(u'_t, v_j^+) - s(u'_t, v_k^-) \right) \right|$$

$$\overset{(*)}{\leq} \frac{1}{|\mathcal{U}|} \frac{1}{n_i^+ n_i^-} \sum_{j=1}^{n_i^+} \sum_{k=1}^{n_i^-} \sup_{\boldsymbol{g} \in \mathcal{H}_R} \left(\left| s(u_t, v_j^+) - s(u'_t, v_j^+) \right| + \left| s(u'_t, v_k^-) - s(u_t, v_k^-) \right| \right)$$

$$\overset{(b)}{\leq} \frac{1}{|\mathcal{U}|} \frac{1}{n_i^+ n_i^-} \sum_{j=1}^{n_i^+} \sum_{k=1}^{n_i^-} \left(\sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| s(u_t, v_j^+) - s(u'_t, v_j^+) \right| + \sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| s(u'_t, v_k^-) - s(u_t, v_k^-) \right| \right) \tag{39}$$

For $\sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| s(u_t, v_j^+) - s(u'_t, v_j^+) \right|$, the following results hold:

$$\sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| s(u_t, v_j^+) - s(u'_t, v_j^+) \right| = \sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| \min_{c_1 \in [C]} \|\boldsymbol{g}_{u_t}^{c_1} - \boldsymbol{g}_{v_j^+}\|^2 - \min_{c_2 \in [C]} \|\boldsymbol{g}_{u'_t}^{c_2} - \boldsymbol{g}_{v_j^+}\|^2 \right|$$

$$\leq \sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| \min_{c_1 \in [C]} \max_{c_2 \in [C]} \left(\|\boldsymbol{g}_{u_t}^{c_1} - \boldsymbol{g}_{v_j^+}\|^2 - \|\boldsymbol{g}_{u'_t}^{c_2} - \boldsymbol{g}_{v_j^+}\|^2 \right) \right|$$

$$\leq \max \left\{ \min_{c_1 \in [C]} \max_{c_2 \in [C]} \phi_{\boldsymbol{g}}(c_1, c_2), \max_{c_1 \in [C]} \min_{c_2 \in [C]} \phi_{\boldsymbol{g}}(c_2, c_1) \right\} \tag{40}$$

According to (40), we can go a step further:

$$
\begin{aligned}
\min_{c_1\in[C]} \max_{c_2\in[C]} \phi_{\boldsymbol{g}}(c_1,c_2) &\leq \max_{c_1=c_2,c1,c2\in[C]} \phi_{\boldsymbol{g}}(c_1,c_1) \\
&\leq \max_{c\in[C]} |\phi_{\boldsymbol{g}}(c,c)| \\
&= \max_{c\in[C]} \left| \|\boldsymbol{g}_{u_t}^c - \boldsymbol{g}_{v_j^+}\|^2 - \|\boldsymbol{g}_{u_t'}^c - \boldsymbol{g}_{v_j^+}\|^2 \right| \\
&= \max_{c\in[C]} \left| \left(\|\boldsymbol{g}_{u_t}^c - \boldsymbol{g}_{v_j^+}\| + \|\boldsymbol{g}_{u_t'}^c - \boldsymbol{g}_{v_j^+}\| \right) \left(\|\boldsymbol{g}_{u_t}^c - \boldsymbol{g}_{v_j^+}\| - \|\boldsymbol{g}_{u_t'}^c - \boldsymbol{g}_{v_j^+}\| \right) \right| \\
&\overset{(**)}{\leq} 4r \max_{c\in[C]} \|\boldsymbol{g}_{u_t}^c - \boldsymbol{g}_{u_t'}^c\| \\
&\leq 8r^2
\end{aligned}
\tag{41}
$$

Based on the result of (41), we have the following result for (39)

$$
\frac{1}{|\mathcal{U}|} \sup_{\boldsymbol{g}\in\mathcal{H}_R} \left| \hat{\mathcal{R}}_{\mathcal{D}_{u_t},\boldsymbol{g}} - \hat{\mathcal{R}}_{\mathcal{D}_{u_t'},\boldsymbol{g}} \right| \leq \frac{16r^2}{|\mathcal{U}|}
\tag{42}
$$

With respect to (4), recall that, we have

$$
\hat{\Omega}_{\mathcal{D},\boldsymbol{g}} = \frac{1}{|\mathcal{U}|} \sum_{u_i\in\mathcal{U}} \psi_{\boldsymbol{g}}(u_i),
$$

where

$$
\psi_{\boldsymbol{g}}(u_i) = \max\left(0, \delta_1 - \delta_{\boldsymbol{g},u_i}\right) + \max\left(0, \delta_{\boldsymbol{g},u_i} - \delta_2\right),
$$
$$
\delta_{\boldsymbol{g},u_i} = \frac{1}{2C(C-1)} \sum_{c_1,c_2\in C} \|\boldsymbol{g}_{u_i}^{c_1} - \boldsymbol{g}_{u_i}^{c_2}\|^2,
$$

Moreover, let us define some intermediate variables:

$$
\psi_{\boldsymbol{g},\delta_1}(u_i,u_j) = \max\left(0, \delta_1 - \delta_{\boldsymbol{g},u_i}\right) - \max\left(0, \delta_1 - \delta_{\boldsymbol{g},u_j}\right),
$$
$$
\psi_{\boldsymbol{g},\delta_2}(u_i,u_j) = \max\left(0, \delta_{\boldsymbol{g},u_i} - \delta_2\right) - \max\left(0, \delta_{\boldsymbol{g},u_j} - \delta_2\right).
$$

In this sense, in terms of (38), the following result holds:

$$
\begin{aligned}
\sup_{\boldsymbol{g}\in\mathcal{H}_R} \left| \hat{\mathcal{L}}_{\mathcal{D}'}(\boldsymbol{g}) - \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) \right| &= \eta \cdot \left| \frac{1}{|\mathcal{U}|} \psi_{\boldsymbol{g}}(u_t) - \frac{1}{|\mathcal{U}|} \psi_{\boldsymbol{g}}(u_t') \right| \\
&= \frac{\eta}{|\mathcal{U}|} \cdot |\psi_{\boldsymbol{g}}(u_t) - \psi_{\boldsymbol{g}}(u_t')| \\
&\overset{(*)}{\leq} \frac{\eta}{|\mathcal{U}|} \cdot \left(|\psi_{\boldsymbol{g},\delta_1}(u_t,u_t')| + |\psi_{\boldsymbol{g},\delta_2}(u_t,u_t')| \right) \\
&\overset{(a)}{\leq} \frac{2\eta}{|\mathcal{U}|} \cdot \left| \delta_{\boldsymbol{g},u_t} - \delta_{\boldsymbol{g},u_t'} \right| \\
&= \frac{\eta}{C(C-1)|\mathcal{U}|} \left| \sum_{c_1,c_2\in C} \left(\|\boldsymbol{g}_{u_t}^{c_1} - \boldsymbol{g}_{u_t}^{c_2}\|^2 - \|\boldsymbol{g}_{u_t'}^{c_1} - \boldsymbol{g}_{u_t'}^{c_2}\|^2 \right) \right| \\
&\overset{(*)}{\leq} \frac{\eta}{C(C-1)|\mathcal{U}|} \sum_{c_1,c_2\in C} \left| \|\boldsymbol{g}_{u_t}^{c_1} - \boldsymbol{g}_{u_t}^{c_2}\|^2 - \|\boldsymbol{g}_{u_t'}^{c_1} - \boldsymbol{g}_{u_t'}^{c_2}\|^2 \right| \\
&\leq \frac{4r^2\eta}{|\mathcal{U}|}
\end{aligned}
\tag{43}
$$

where $(*)$ achieves via the inequality $|x+y| \leq |x| + |y|$ and (a) follows the Lem.4.

Finally, in this case, we have

$$
\sup_{\boldsymbol{g}\in\mathcal{H}_R} \left| \hat{\mathcal{L}}_{\mathcal{D}'}(\boldsymbol{g}) - \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) \right| \leq \frac{16r^2 + 4r^2\eta}{|\mathcal{U}|}.
\tag{44}
$$

- **Case 2:** Only one positive item is different. In this case, we consider such difference occurs in the positive item $v_{t_1}^+$ with respect to a specific user u_i and there are $|\mathcal{U}|$ cases for all users. Mathematically, we have

$$\mathcal{D}_{u_i} = \{v_j^+\}_{j=1}^{n_i^+} \cup \{v_k^-\}_{k=1}^{n_i^-}, \quad \mathcal{D}'_{u_i} = (\mathcal{D}_{u_i}\backslash\{v_{t_1}^+\}) \cup \{\tilde{v}_{t_1}^+\}, \tag{45}$$

where $\forall t_1, t_1 = 1, 2, \ldots, n_i^+$ and $n_i^+ + n_i^- = |\mathcal{I}|$. Then, it is obvious that in this case only the first term in (22) contributes to the upper bound. According to this observation, the upper bound could be simplified as follows:

$$\sup_{\boldsymbol{g}\in\mathcal{H}_R} \left|\hat{\mathcal{L}}_{\mathcal{D}'}(\boldsymbol{g}) - \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})\right| = \sup_{\boldsymbol{g}\in\mathcal{H}_R} \left|\hat{\mathcal{R}}_{\mathcal{D}}(\boldsymbol{g}) - \hat{\mathcal{R}}_{\mathcal{D}'}(\boldsymbol{g})\right|$$

$$= \sup_{\boldsymbol{g}\in\mathcal{H}_R} \left|\hat{\mathcal{R}}_{\mathcal{D}_{u_i}}(\boldsymbol{g}) - \hat{\mathcal{R}}_{\mathcal{D}'_{u_i}}(\boldsymbol{g})\right|, \tag{46}$$

where again we denote

$$\hat{\mathcal{R}}_{\mathcal{D}_{u_i}}(\boldsymbol{g}) = \frac{1}{|\mathcal{U}|} \cdot \sum_{j=1}^{n_i^+}\sum_{k=1}^{n_i^-} \ell_{\boldsymbol{g}}^{(i)}(v_j^+, v_k^-),$$

and

$$\ell_{\boldsymbol{g}}^{(i)}(v_j^+, v_k^-) = \max(0, \lambda + s(u_i, v_j^+) - s(u_i, v_k^-)).$$

Let

$$\Delta_{\boldsymbol{g}}(c_1, c_2) = \|\boldsymbol{g}_{u_i}^{c_1} - \boldsymbol{g}_{v_{t_1}^+}\|^2 - \|\boldsymbol{g}_{u_i}^{c_2} - \boldsymbol{g}_{\tilde{v}_{t_1}^+}\|^2. \tag{47}$$

Then, since v_j^+ and $\tilde{v}_j^+$ are different in this case, we have

$$\sup_{\boldsymbol{g}\in\mathcal{H}_R}\left|\hat{\mathcal{L}}_{\mathcal{D}'}(\boldsymbol{g}) - \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})\right| = \frac{1}{|\mathcal{U}|}\sup_{\boldsymbol{g}\in\mathcal{H}_R}\left|\frac{1}{n_i^+ n_i^-}\sum_{k=1}^{n_i^-}\ell_{\boldsymbol{g}}^{(i)}(v_{t_1}^+, v_k^-) - \frac{1}{n_i^+ n_i^-}\sum_{k=1}^{n_i^-}\ell_{\boldsymbol{g}}^{(i)}(\tilde{v}_{t_1}^+, v_k^-)\right|$$

$$\overset{(*)}{\leq} \frac{1}{|\mathcal{U}|n_i^+ n_i^-}\sup_{\boldsymbol{g}\in\mathcal{H}_R}\sum_{k=1}^{n_i^-}\left|\ell_{\boldsymbol{g}}^{(i)}(v_{t_1}^+, v_k^-) - \ell_{\boldsymbol{g}}^{(i)}(\tilde{v}_{t_1}^+, v_k^-)\right|$$

$$\overset{(b)}{\leq} \frac{1}{|\mathcal{U}|n_i^+ n_i^-}\sum_{k=1}^{n_i^-}\left(\sup_{\boldsymbol{g}\in\mathcal{H}_R}\left|\ell_{\boldsymbol{g}}^{(i)}(v_{t_1}^+, v_k^-) - \ell_{\boldsymbol{g}}^{(i)}(\tilde{v}_{t_1}^+, v_k^-)\right|\right)$$

$$\overset{(a)}{\leq} \frac{1}{|\mathcal{U}|n_i^+ n_i^-}\sum_{k=1}^{n_i^-}\sup_{\boldsymbol{g}\in\mathcal{H}_R}\left|s(u_i, v_{t_1}^+) - s(u_i, \tilde{v}_{t_1}^+)\right|$$

$$= \frac{1}{|\mathcal{U}|n_i^+ n_i^-}\sum_{k=1}^{n_i^-}\sup_{\boldsymbol{g}\in\mathcal{H}_R}\left|\min_{c_1\in[C]}\|\boldsymbol{g}_{u_i}^{c_1} - \boldsymbol{g}_{v_{t_1}^+}\|^2 - \min_{c_2\in[C]}\|\boldsymbol{g}_{u_i}^{c_2} - \boldsymbol{g}_{\tilde{v}_{t_1}^+}\|^2\right|$$

$$= \frac{1}{|\mathcal{U}|n_i^+ n_i^-}\sum_{k=1}^{n_i^-}\sup_{\boldsymbol{g}\in\mathcal{H}_R}\left|\min_{c_1\in[C]}\max_{c_2\in[C]}\left(\|\boldsymbol{g}_{u_i}^{c_1} - \boldsymbol{g}_{v_{t_1}^+}\|^2 - \|\boldsymbol{g}_{u_i}^{c_2} - \boldsymbol{g}_{\tilde{v}_{t_1}^+}\|^2\right)\right|$$

$$\tag{48}$$

According to (47), we have

$$\sup_{\boldsymbol{g}\in\mathcal{H}_R}\left|\hat{\mathcal{L}}_{\mathcal{D}'}(\boldsymbol{g}) - \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})\right| \leq \frac{1}{|\mathcal{U}|n_i^+ n_i^-}\sum_{k=1}^{n_i^-}\sup_{\boldsymbol{g}\in\mathcal{H}_R}\left|\min_{c_1\in[C]}\max_{c_2\in[C]}\Delta_{\boldsymbol{g}}(c_1, c_2)\right|$$

$$= \frac{1}{|\mathcal{U}|n_i^+ n_i^-}\sum_{k=1}^{n_i^-}\left|\max\left\{\min_{c_1\in[C]}\max_{c_2\in[C]}\Delta_{\boldsymbol{g}}(c_1, c_2), \max_{c_1\in[C]}\min_{c_2\in[C]}\Delta_{\boldsymbol{g}}(c_2, c_1)\right\}\right|$$

$$\tag{49}$$

It is easy to show that,

$$
\begin{aligned}
\min_{c_1 \in [C]} \max_{c_2 \in [C]} \Delta_{\boldsymbol{g}}(c_1, c_2) &\leq \max_{c_1 = c_2, c_1, c_2 \in [C]} \Delta_{\boldsymbol{g}}(c_1, c_2) \\
&\leq \max_{c_1 = c_2, c_1, c_2 \in [C]} |\Delta_{\boldsymbol{g}}(c_1, c_2)| \\
&= \max_{c \in [C]} \left| \|\boldsymbol{g}_{u_i}^c - \boldsymbol{g}_{v_{t_1}^+}\|^2 - \|\boldsymbol{g}_{u_i}^c - \boldsymbol{g}_{\tilde{v}_{t_1}^+}\|^2 \right| \\
&= \max_{c \in [C]} \left| \left(\|\boldsymbol{g}_{u_i}^c - \boldsymbol{g}_{v_{t_1}^+}\| + \|\boldsymbol{g}_{u_i}^c - \boldsymbol{g}_{\tilde{v}_{t_1}^+}\| \right) \left(\|\boldsymbol{g}_{u_i}^c - \boldsymbol{g}_{v_{t_1}^+}\| - \|\boldsymbol{g}_{u_i}^c - \boldsymbol{g}_{\tilde{v}_{t_1}^+}\| \right) \right| \\
&\overset{(**)}{\leq} \left(\|\boldsymbol{g}_{u_i}^c - \boldsymbol{g}_{v_{t_1}^+}\| + \|\boldsymbol{g}_{u_i}^c - \boldsymbol{g}_{\tilde{v}_{t_1}^+}\| \right) \left(\|\boldsymbol{g}_{u_i}^c - \boldsymbol{g}_{v_{t_1}^+} - \boldsymbol{g}_{u_i}^c + \boldsymbol{g}_{\tilde{v}_{t_1}^+}\| \right) \\
&\leq 8r^2
\end{aligned}
\tag{50}
$$

In the same way, we also have

$$
\max_{c_1 \in [C]} \min_{c_2 \in [C]} \Delta_{\boldsymbol{g}}(c_2, c_1) \leq 8r^2
\tag{51}
$$

Therefore, applying (50) and (51) to (49), we have

$$
\sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| \hat{\mathcal{L}}_{\mathcal{D}'}(\boldsymbol{g}) - \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) \right| \leq \frac{8r^2}{|\mathcal{U}| n_i^+}
\tag{52}
$$

- **Case 3:** Only one negative item is different. In this case, we assume such difference occurs in the negative item $v_{t_2}^-$ with respect to a specific user u_i, and there are also $|\mathcal{U}|$ cases for all users. Mathematically, we have

$$
\mathcal{D}_i = \{v_j^+\}_{j=1}^{n_i^+} \cup \{v_k^-\}_{k=1}^{n_i^-}, \quad \mathcal{D}_i' = (\mathcal{D}_i \backslash \{v_{t_2}^-\}) \cup \{\tilde{v}_{t_2}^-\}.
\tag{53}
$$

where $\forall t_2, t_2 = 1, 2, \dots, n_i^-$.

Similarly, if v_k^- and $\tilde{v}_k^-$ are different, we can also hold

$$
\sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| \hat{\mathcal{L}}_{\mathcal{D}'}(\boldsymbol{g}) - \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) \right| \leq \frac{8r^2}{|\mathcal{U}| n_i^-}
\tag{54}
$$

Finally, taking all above three cases into account, one can conclude that $\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})$ is satisfied with the bounded difference property (Lem.4).

This completed the proof. $\qquad \square$

Lemma 7. *Equipped with Lem.5 and Lem.6, the following inequality holds:*

$$
\mathbb{P}\left[\left| \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}_l) - \mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}_l)] \right| \geq \frac{\varepsilon}{2} \right] \leq 2\exp\left(\frac{-\varepsilon^2 \tilde{N}}{2} \right),
$$

where

$$
\tilde{N} = \left(4r^2 \sqrt{ \left(\frac{(4+\eta)^2}{|\mathcal{U}|} + \frac{2}{|\mathcal{U}|^2} \sum_{u_i \in \mathcal{U}} \left(\frac{1}{n_i^+} + \frac{1}{n_i^-} \right) \right) } \right)^{-2}
$$

Proof. The proof could be easily achieved by applying Lem.2 on top of Lem.5 and Lem.6. $\qquad \square$

B.3   Proof of the Main Result

B.3.1   Proof of Thm.1

Restate of Theorem 1 (Generalization Upper Bound of DPCML). Let $\mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})]$ be the population risk of $\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})$. Then, $\forall \, \boldsymbol{g}, \boldsymbol{g} \in \mathcal{H}_R$, with high probability, the following inequation holds:

$$\left| \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) - \mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})] \right| \leq \sqrt{\frac{2d \log\left(3r\tilde{N}\right)}{\tilde{N}}}, \tag{55}$$

where we have

$$\tilde{N} = \left(4r^2 \sqrt{\left(\frac{(4+\eta)^2}{|\mathcal{U}|} + \frac{2}{|\mathcal{U}|^2} \sum_{u_i \in \mathcal{U}} \left(\frac{1}{n_i^+} + \frac{1}{n_i^-} \right) \right)} \right)^{-2}.$$

Proof. Step 1. In order to obtain the generalization bound, we need to first figure out the following probability:

$$\mathbb{P}\left[\sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) - \mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})] \right| \geq \varepsilon \right],$$

where ε is the generalization error and usually a very small value.

Denote the covering number of σ-covering in Lem.5 as $\mathcal{N}_3(\sigma; \mathcal{H}_R, \rho_3)$. Then, according to Def.B.2, Def.B.2, Lem.3 and Lem.5, we have

$$\mathbb{P}\left[\sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) - \mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})] \right| \geq \varepsilon \right] \leq \mathbb{P}\left[\sup_{\boldsymbol{g} \in \bigcup_{l=1}^{\mathcal{N}_3} \mathcal{B}(\boldsymbol{g}_l, \sigma)} \left| \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) - \mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})] \right| \geq \varepsilon \right]$$

$$\overset{Lem.3}{\leq} \sum_{l=1}^{\mathcal{N}_3} \mathbb{P}\left[\sup_{\boldsymbol{g} \in \mathcal{B}(\boldsymbol{g}_l, \sigma)} \left| \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) - \mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})] \right| \geq \varepsilon \right] \tag{56}$$

$$\overset{Lem.5}{\leq} \sum_{l=1}^{\mathcal{N}_3} \mathbb{P}\left[\left| \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}_l) - \mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}_l)] \right| \geq \frac{\varepsilon}{2} \right]$$

where, without the loss of generality, we denote the covering number as $\mathcal{N}_3$ for short.

Note that, from Lem.5 we have $\sigma = \frac{\varepsilon}{16r(4+\eta)}$, and

$$\mathcal{N}_3(\sigma; \mathcal{H}_R, \rho_3) \leq \mathcal{N}_1\left(\frac{\varepsilon}{16r(4+\eta)}, \mathcal{H}_R, \rho_1 \right) \cdot \mathcal{N}_2\left(\frac{\varepsilon}{16r(4+\eta)}, \mathcal{H}_R, \rho_2 \right).$$

Therefore, we further have

$$\mathbb{P}\left[\sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) - \mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})] \right| \geq \varepsilon \right] \leq \mathcal{N}_1 \cdot \mathcal{N}_2 \cdot \mathbb{P}\left[\left| \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}_l) - \mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}_l)] \right| \geq \frac{\varepsilon}{2} \right] \tag{57}$$

Step 2. Now, according to Lem.7, we have

$$\mathbb{P}\left[\sup_{\boldsymbol{g} \in \mathcal{H}_R} \left| \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g}) - \mathbb{E}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{g})] \right| \geq \varepsilon \right] \leq 2\mathcal{N}_1 \cdot \mathcal{N}_2 \cdot \exp\left(\frac{-\varepsilon^2 \tilde{N}}{2} \right). \tag{58}$$

Then with Lem.1 and by further choosing

$$\varepsilon = \sqrt{\frac{2d}{\tilde{N}} \log\left(3r\tilde{N}\right)},$$

25

we have:

$$\mathbb{P}\left[\sup_{\boldsymbol{g}\in\mathcal{H}_R}\left|\hat{\mathcal{L}}_\mathcal{D}(\boldsymbol{g}) - \mathbb{E}[\hat{\mathcal{L}}_\mathcal{D}(\boldsymbol{g})]\right| \geq \sqrt{\frac{2d\log\left(3r\tilde{N}\right)}{\tilde{N}}}\right] \leq 2\left(\frac{3rB^2}{2d\log\left(3r\tilde{N}\right)}\right)^d,\tag{59}$$

where again

$$\tilde{N} = \left(4r^2\sqrt{\left(\frac{(4+\eta)^2}{|\mathcal{U}|} + \frac{2}{|\mathcal{U}|^2}\sum_{u_i\in\mathcal{U}}\left(\frac{1}{n_i^+} + \frac{1}{n_i^-}\right)\right)}\right)^{-2},$$

and

$$B = 16r(4+\eta)\tag{60}$$

Therefore, we can conclude that, with high probability,

$$\left|\hat{\mathcal{L}}_\mathcal{D}(\boldsymbol{g}) - \mathbb{E}[\hat{\mathcal{L}}_\mathcal{D}(\boldsymbol{g})]\right| \leq \sqrt{\frac{2d\log\left(3r\tilde{N}\right)}{\tilde{N}}}, \ \ \forall\, \boldsymbol{g}, \boldsymbol{g} \in \mathcal{H}_R.\tag{61}$$

This completed the proof. □

B.3.2 Proof of Corol.1

Restate of Corollary 1. On the top of Thm.1, DPCML could enjoy a smaller generalization error than CML.

Proof. For simplication of notations, let $\mathcal{X}_{=1}$ and $\mathcal{X}_{>1}$ be the feasible regions of CML ($C = 1$) and DPCML ($C > 1$), and $\hat{\mathcal{L}}_{=1}(\boldsymbol{g})$ and $\hat{\mathcal{L}}_{>1}(\boldsymbol{g})$ be the empirical risks of CML ($C = 1$) and DPCML ($C > 1$), respectively. Then, since DPCML leverages $\min_{c\in C}\|\boldsymbol{g}_{u_i}^c - \boldsymbol{g}_{v_j}\|^2$ as the distance, which can be regarded as a minimum of multiple single version CML, it is easy to know that the feasible solution of CML is also included in DPCML, i.e., $\mathcal{X}_{=1} \subseteq \mathcal{X}_{>1}$. Therefore, we can conclude that $\hat{\mathcal{L}}_{>1}(\boldsymbol{g}) \leq \hat{\mathcal{L}}_{=1}(\boldsymbol{g})$. Denote $\Delta = \sqrt{\frac{2d\log\left(3r\tilde{N}\right)}{\tilde{N}}}$ as the residuals between $\mathbb{E}[\hat{\mathcal{L}}(\boldsymbol{g})]$ and $\hat{\mathcal{L}}(\boldsymbol{g})$. Moreover, we have $\Delta_{\text{DPCML}} = \Theta(\Delta_{\text{CML}})$ since Δ in our bound does not depend on C. This is consistent with the over-parameterization phenomenon [8, 34]. According to Thm.1, we see that $\mathbb{E}[\hat{\mathcal{L}}_*(\boldsymbol{g})] \leq \hat{\mathcal{L}}_*(\boldsymbol{g}) + \Delta$, where $*$ represents $= 1$ or > 1. Therefore, we can conclude that DPCML could enjoy a smaller generalization error than the traditional CML. We also empirically demonstrate this in the experiment Sec.4.3. □

C Experiments

C.1 Dataset

We perform empirical experiments on several public and real-world benchmark datasets, including:

- **MovieLens**[2] - One of the most popular benchmark datasets with many versions. Specifically, it includes explicit user-item ratings ranging from 1 to 5 and movie types in terms of various movies. We adopt **MovieLens-1m**[3] and **MovieLens-10m**[4] here to evaluate the performance. To obtain the implicit preference feedback, if the score of item v_j rated by user u_i is no less than 4, we regard item v_j as a positive item for user u_i following the previous and successful research [14].

- **CiteULike**[5] [46] - An implicit feedback dataset that includes the preferences of users toward different articles. There are two configurations of CiteULike collected from CiteULike and Google Scholar. Following [18], we adopt **CiteULike-T** here to evaluate the performance.

---

[2]https://grouplens.org/datasets/movielens/
[3]https://grouplens.org/datasets/movielens/1m/
[4]https://grouplens.org/datasets/movielens/10m/
[5]http://www.citeulike.org/faq/data.adp

- **Steam-200k**[6] - This dataset is collected from the Steam which is the world's most popular PC gaming hub. The observed behaviors of users include 'purchase' and 'play' signals. In order to obtain the implicit feedback, if a user has purchased a game as well as the playing hours $play > 0$, we treat this game as a positive item.

The detailed statistics in terms of these datasets are summarized in Tab.4.

Table 4: Basic Information of the Datasets. %Density is defined as $\frac{\#Ratings}{\#Users \times \#Items} \times 100\%$.

Datasets	MovieLens-1M	Steam-200k	CiteULike-T	MovieLens-10M
Domain	Movie	Game	Paper	Movie
#Users	6,034	3,757	5,219	69,167
#Items	3,953	5,113	25,975	10,019
#Ratings	575,271	115,139	125,580	5,003,437
%Density	2.4118%	0.5994%	0.0926%	0.7220%

C.2 Evaluation Metrics

In some typical recommendation systems, users often care about the top-N items in recommendation lists, so the most relevant items should be ranked first as much as possible. Motivated by this, we evaluate the performance of competitors and our algorithm with the following extensively adopted metrics, including:

- **Precision** (P@N) counts the proportion that the ground-truth items are among the Top-N recommended list.

$$\text{P@}N = \frac{1}{|\mathcal{U}|} \sum_{u_i \in \mathcal{U}} \frac{|\mathcal{D}_{u_i}^+ \cap I_N^{u_i}|}{N}$$

  where again $\mathcal{D}_{u_i}^+$ is the set of ground-truth items of user u_i; $I_N^{u_i}$ is the top-N recommendation list for user u_i; and $|\cdot|$ means the size of set.

- **Recall** (R@N) is defined as the number of the ground-truth items in top-N recommendation list divided by the amount of totally ground-truth items. This reflects the ability of model to find the relevant items.

$$\text{R@}N = \frac{1}{|\mathcal{U}|} \sum_{u_i \in \mathcal{U}} \frac{|\mathcal{D}_{u_i}^+ \cap I_N^{u_i}|}{|\mathcal{D}_{u_i}^+|}$$

- **Normalized Discounted Cumulative Gain** (NDCG@N) counts the ground-truth items in the top-N recommendation list with a position weighting strategy, i.e., assigning a larger value on top items than bottom ones.

$$\text{NDCG@}N = \frac{1}{|\mathcal{U}|} \sum_{u_i \in \mathcal{U}} \frac{\text{DCG}_{u_i}\text{@}N}{\text{IDCG}_{u_i}\text{@}N}$$

  Specifically, the $\text{DCG}_{u_i}\text{@}N$ and $\text{IDCG}_{u_i}\text{@}N$ are defined as:

$$\text{DCG}_{u_i}\text{@}N = \sum_{j=1}^{N} \frac{1 \cdot \mathbb{I}(I_{N,j}^{u_i} \in \mathcal{D}_{u_i}^+)}{\log_2(j+1)},$$

$$\text{IDCG}_{u_i}\text{@}N = \sum_{k=1}^{\min(N,|\mathcal{D}_{u_i}^+|)} \frac{1}{\log_2(k+1)},$$

  where $I_{N,j}^{u_i}$ respresents the j-th item in the top-N recommendation list; $\mathbb{I}(\cdot)$ is an indicator function that returns 1 if the statement is true and returns 0, otherwise.

---

[6]https://www.kaggle.com/tamber/steam-video-games

- **Mean Average Precision** (MAP) is an extension of Average Precision(AP). AP is the average of precision values at all positions where ground-truth items are found.

$$\text{AP}_{u_i} = \frac{1}{|\mathcal{D}_{u_i}^+|} \sum_{j=1}^{|\hat{I}_{u_i}|} \frac{|\mathcal{D}_{u_i}^+ \cap \hat{I}_{u_i,1:j}| \cdot \mathbb{I}(j \in \mathcal{D}_{u_i}^+)}{rank_j^{u_i}}$$

$$\text{MAP} = \frac{1}{|\mathcal{U}|} \sum_{u_i \in \mathcal{U}} \text{AP}_{u_i}$$

  where different from $I_N^{u_i}$, $\hat{I}_{u_i}$ is the recommendation rankings in terms of all items for user u_i; $\hat{I}_{u_i,1:j}$ represents the top-j recommendation list for user u_i; and $rank_j^{u_i}$ means the ranking of item j in $\hat{I}_{u_i}$.

- **Mean Reciprocal Rank** (MRR) takes the rank of each recommended item into account. It is the average of reciprocal ranks of the desired item:

$$\text{MRR} = \frac{1}{|\mathcal{U}|} \sum_{u_i \in \mathcal{U}} \sum_{j=1}^{|\hat{I}_{u_i}|} \frac{1}{rank_j^{u_i}} \cdot \mathbb{I}(\hat{I}_{u_i,j} \in \mathcal{D}_{u_i}^+)$$

Note that, for all the above metrics, the higher the metric is, the better the algorithm achieves.

C.3 Competitors

The involved competitors roughly fall into three groups here, including:

1) Item-based collaborative filtering algorithm.

- **itemKNN** [27] is designed on the criterion of the k-nearest neighborhood (KNN), which directly considers the similarity (such as cosine similarity) between the candidate and the previously interacted items to make the recommendations.

2) MF-based algorithms including the combination of MF and deep learning network and multi-vector MF-based methods.

- **Generalized Matrix Factorization** (GMF) adopts a linear kernel to capture the preference of users such that it is more expressive than the traditional MF algorithms.
- **Multi-Layer Perceptron** (MLP) leverages a multi-layer perceptron endowed with reasonable flexibility and non-linearity to model the users' preference toward items.
- **Neural network-based Collaborative Filtering** (NeuMF) [7] [14] is a seminal and competitive deep learning based recommendation framework. Specifically, NCF integrates the GMF and MLP algorithms and makes recommendation via regarding the recommendation task as a regression problem.
- **Multi-vector MF** (M2F) [12] is a state-of-the-art MF-based recommendation algorithm, which models the diversity preference of users by assigning them multiple embeddings in the dot-product space. This could be regarded as a competitive baseline to figure out the superiority of our proposed algorithm.
- **Multi-vector GMF** (MGMF). Consider that the original algorithm [12] might be specifically tailored for the explicit feedback rather than the implicit signals, we further apply a multiple set of users' representations to GMF [14].

3) CML-based recommendation competitors.

- **Uniform Negative Sampling** (UniS) [35] in terms of each user, uniformly sample S items from unobserved interactions as negative instances to optimize the pairwise ranking loss.
- **Popularity-based Negative Sampling** (PopS) [53] samples S negative candidates from unobserved interactions based on their popularity/frequencies.

---

[7] https://github.com/guoyang9/NCF

- **Two-Stage Negative Sampling** (2stS) [8][44] adopts a two-stage sampling strategy. 1) A candidate set of items are sampled based on their popularity; 2) according to their inner product values with anchors (positive items), the most informative samples are selected from this candidate.

- **Hard Negative Sampling** (HarS) [9] is similar to the negative sample mining process broadly used in metric learning [11, 10]. It can be divided into two steps: a) uniformly sample S candidates from unobserved items; b) select the hard (the most distant) item from the candidates as negative in the light of the distance between the aimed user and item candidates.

- **Collaborative Translational Metric Learning (TransCF)** [36] is a translation-based method. Specifically, such translation-based algorithms employ $\boldsymbol{d}(i,j) = ||\boldsymbol{g}_{u_i} + \boldsymbol{g}_{r_{ij}} - \boldsymbol{g}_{v_j}||^2$ as the distance/score between user u_i and item v_j instead of $||\boldsymbol{g}_{u_i} - \boldsymbol{g}_{v_j}||^2$, where $\boldsymbol{g}_{r_{ij}}$ is a specific translation vector for u_i and v_j. In light of this, TransCF discovers such user–item translation vectors via the users' relationships with their neighbor items.

- **Latent Relational Metric Learning (LRML)** [41] is also a translation-based CML method. As a whole, the key idea of LRML is similar to TransCF. The main difference is how to access the translation vectors effectively. Concretely, TransCF leverages the neighborhood information of users and items to acquire the translation vectors while LRML introduces an attention-based memory-augmented neural architecture to learn the exclusive and optimal translation vectors.

- **AdaCML** [62] learns an adaptive user representation via a memory component and an attention mechanism to accurately model the implicit relationships of user-item pairs and users' interests.

- **HLR** [43] is a state-of-the-art CML-based approach that employs the memory-based attention networks to hierarchically capture users' preferences from both latent user-item and item-item relations.

C.4 Implementation Details

We implement our model with PyTorch [10] [37] and employ *Adam* [20] as the optimizer. In terms of all benchmark datasets, each user interactions are divided into training/validation/test sets with a $0.6 : 0.2 : 0.2$ split ratio. According to this, to ensure that each user has at least one positive interaction in training/validation/test, users who have less than five interactions are filtered out from these datasets. We adopt grid search for all methods to select the best parameters based on the validation set and report the corresponding performance on the test set. Specifically, for all competitors, the batch size is set to 256 and the learning rate is searched within $\{0.001, 0.003, 0.01\}$. The number of epochs is set as 100. The dimension of embedding d is fixed as 100, and the margin λ is searched within $\{1.0, 1.5, 2.0\}$. The number of user representations C is tuned among $\{2, 3, 4, 5\}$. For the regularization term, η is searched within $\{10, 20, 30\}$, $\delta_1 \in \{0., 0.05, 0.1, 0.2, 0.5\}$ and $\delta_2 \in \{0.1, 0.25, 0.35, 0.5, 0.8\}$. For our proposed algorithm, following [14, 26], we also adopt the negative sampling strategy to avoid the heavy burden of pairwise learning. Here we employ two negative sampling strategies i.e., uniform [35] and hard negative sampling [17], denoted as **DPCML1** and **DPCML2** respectively. To make sure a reasonable comparison, we set the sampling constant $S = 10$ for all methods. For the other parameters of baseline models, we follow their tuning strategies in the original papers. Finally, in terms of the top-N recommendation, we evaluate the performance at $N \in \{3, 5\}$, respectively.

C.5 Overall Performance

The experimental results of all the involved competitors are shown in Tab.2 and Tab.5 . Consequently, we can draw the following conclusions: 1) In most cases, the best performance of CML-based methods consistently surpasses the best MF-based competitors. This suggests that it is necessary to develop CML-based RS algorithms. 2) Our proposed method consistently surpasses all the competitors significantly on all datasets, except the results for MAP and MRR on CiteULike. Even for the failure results, the performance is fairly competitive compared with the competitors. This shows the effectiveness of our proposed algorithm. 3) Compared with studies targeting joint accessibility (i.e., M2F and MGMF), our proposed method significantly outperforms M2F and MGMF on all benchmark datasets. This shows the advantage of the CML-based paradigm that deserves more attention along this direction in future work.

---

[9]https://github.com/changun/CollMetric
[10]https://pytorch.org/

Table 5: Performance comparisons on CiteULike and MovieLens-10m datasets. The best and second-best are highlighted in bold and underlined, respectively.

	Type	Method	P@3	R@3	NDCG@3	P@5	R@5	NDCG@5	MAP	MRR
CiteULike	Item-based	itemKNN	1.20	0.83	1.23	1.15	0.77	1.16	1.44	3.78
	MF-based	GMF	1.86	0.96	2.05	2.15	0.97	2.40	1.34	5.53
		MLP	2.06	1.08	2.22	2.40	1.16	2.61	1.52	12.37
		NeuMF	2.06	1.08	2.21	2.36	1.16	2.57	1.54	12.22
		M2F	1.76	0.90	1.97	1.87	0.93	2.18	0.93	4.53
		MGMF	2.31	1.23	2.48	2.42	1.12	2.71	1.51	6.18
	CML-based	UniS	7.34	3.71	7.48	9.54	5.13	10.02	5.59	17.27
		PopS	5.41	2.94	5.77	6.75	3.62	7.23	4.61	14.39
		2stS	6.40	3.35	6.77	8.27	4.29	8.81	4.99	15.87
		HarS	8.44	4.41	8.82	10.43	5.60	11.25	**6.67**	**20.08**
		TransCF	5.79	3.03	6.09	7.45	3.93	7.84	4.54	14.50
		LRML	2.52	1.33	2.58	3.06	1.64	3.19	1.91	6.45
		AdaCML	7.04	3.75	7.31	8.70	4.52	9.18	5.57	17.31
		HLR	2.03	1.08	2.20	2.25	1.13	2.52	1.45	5.86
	Ours	DPCML1	7.78	4.04	8.14	10.03	5.33	10.64	6.08	18.75
		DPCML2	**8.70**	**4.59**	**9.06**	**10.96**	**5.85**	**11.47**	6.44	19.96
MovieLens-10m	Item-based	itemKNN	11.44	3.70	11.78	12.27	4.93	12.63	8.25	25.85
	MF-based	GMF	13.55	3.87	13.91	14.67	5.41	15.13	9.14	28.91
		MLP	15.27	4.93	15.46	16.08	6.53	16.38	12.77	32.21
		NeuMF	15.19	5.02	15.27	16.09	6.65	16.24	12.76	31.87
		M2F	7.03	1.41	7.21	7.55	2.23	7.98	2.50	15.17
		MGMF	14.62	4.26	15.15	15.53	5.96	16.26	10.30	31.07
	CML-based	UniS	10.15	2.84	10.33	11.19	4.08	11.38	8.92	24.24
		PopS	8.61	3.06	8.96	8.34	3.76	8.84	6.08	20.97
		2stS	16.47	4.89	16.72	17.62	6.87	18.06	12.89	33.75
		HarS	17.00	4.97	17.16	18.34	6.96	18.70	13.14	34.20
		TransCF	11.00	3.70	10.91	11.62	4.94	11.61	7.99	23.67
		LRML	13.72	3.96	13.98	14.53	5.58	15.08	8.99	28.77
		AdaCML	13.65	4.00	13.82	14.64	5.52	14.98	11.13	29.58
		HLR	15.13	5.12	14.94	16.40	7.00	16.23	13.40	31.66
	Ours	DPCML1	12.73	3.82	13.05	13.12	5.07	13.72	10.32	28.65
		DPCML2	**18.00**	**5.46**	**18.37**	**18.97**	**7.37**	**19.57**	**14.01**	**36.44**

C.6 Quantitative Analysis

C.6.1 Fine-grained Performance Comparison

At first, we expect to examine the performance improvements of our proposed method with respect to minority preferences. To do this, we show the fine-grained MAP metric over each interest group (movie genre) on MovieLens-10m. The empirical results are reported in Fig.4. Here the x-axis is organized with a descending order according to the item category distribution presented in Fig.2(b). We can observe that our proposed framework could not only significantly outperform their single-vector counterparts in the majority interests but also improve the performance of minority groups in most cases. Especially, compared with HarS, the performance improvement of DPCML2 on minority interests is sharp. This shows that DPCML could reasonably focus on potentially interesting items even with the imbalanced item distribution to alleviate the preference bias induced by the conventional CML-based studies.

C.6.2 Recommendation Diversity Evaluation

Furthermore, to show the improvement of DPCML in promoting diversity, we evaluate the performance of DPCML against CML-based competitors with a metric called *max-sum diversification (MaxDiv)* [4], i.e.,

$$MaxDiv@N = \frac{1}{|\mathcal{U}|} \sum_{u_i \in \mathcal{U}} \sum_{\substack{v_i, v_j \in \mathcal{I}_{u_i}^N, \\ v_i \neq v_j}} s(v_i, v_j),$$

where $s(v_i, v_j) = \|\boldsymbol{g}_{v_i} - \boldsymbol{g}_{v_j}\|^2$ is the square of Euclidean distance between item v_i and v_j; $\mathcal{I}_{u_i}^N$ is the top-N recommendation items for user u_i.
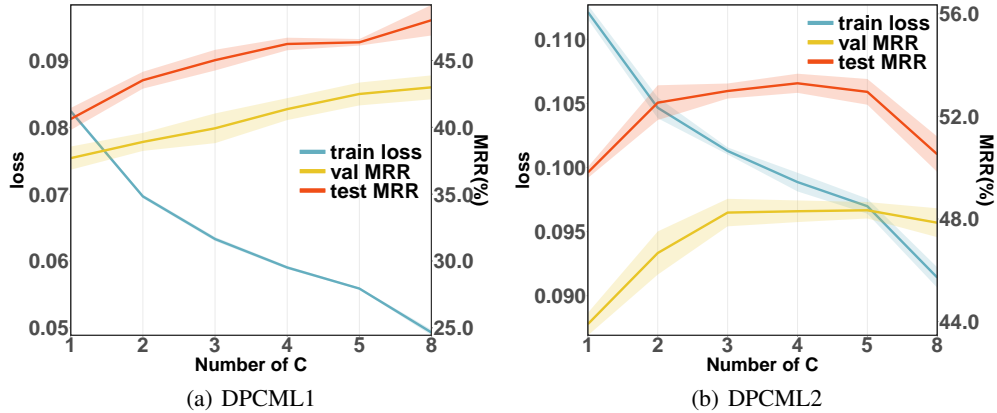
Figure 9: Empirical justification of Thm.1.

Generally speaking, MaxDiv@N measures the recommendation diversification by considering item-side similarity, where a high value implies that the recommendation results are relatively diverse. Then, we compare DPCML with the following competitors for a fair evaluation: a) **UniS** b) **HarS** c) **DPCML1 without (w/o) DCRS** and d) **DPCML2 without (w/o) DCRS**. The experiments are conducted on the Steam-200k and MovieLens-1m datasets with $N \in \{3, 5, 10, 20\}$.

The diversity results are shown in Tab.3. We observe that: a) For methods within the same negative sampling strategy (i.e., UniS and DPCML1, HarS and DPCML2), our proposed DPCML could achieve relatively higher max-sum values. This suggests the improvement of DPCML in terms of promoting recommendation diversity. b) In most cases (except for DPCML1 w/o DCRS on the MovieLens-1m dataset), DPCML outperforms other competitors even without regularization. c) Most importantly, equipped with the regularization term DCRS, DPCML could achieve better diversification results against w/o DCRS. This once again shows the rationality/importance of DCRS.

C.6.3 Effect of the Diversity Control Regularization

Next, we study the effectiveness of our proposed diversity control regularization scheme. To do this, we analyze the influence of two main hyper-parameters, δ_1 and δ_2. We illustrate a 3D-barplot based on the results of grid search on Steam-200k. The results are presented in Fig.5, Fig.11 and Fig.12. For a clear comparison, $\delta_1 = \delta_2 = 0$ represents the standard single-vector counterparts performances and $\delta_1 > \delta_2$ indicates the results of DPCML removing the diversity control regularization scheme. Moreover, we set the trade-off coefficient $\eta = 10$ and the representation number $C = 5$ here. From these results, we can observe that the proposed regularization scheme could significantly boost performance on all metrics, which demonstrates the effectiveness of the diversity control regularization term. In addition, one can see that there would induce different performances with different diversity values. This suggests that controlling a proper diversity of the embeddings for the same user is essential to accommodate their preferences better.

C.6.4 Empirical Justification of Corol.1

We conduct the empirical studies of Corol.1 on the Steam-200k dataset. Expressly, we set $C \in \{1, 2, 3, 4, 5, 8\}$ and record the results of training loss and validation/test MRR metric. Moreover, the experiments are repeated 5 times. The empirical results are shown in Fig.9, where the shades represent the variance among 5 experiments. Based on these results, we can see that, with the increase of C, the empirical risk (i.e., training loss) of DPCML ($C > 1$) is significantly smaller than CML ($C = 1$). In addition, DPCML could substantially improve the performance of the validation/test set. Thus, we can conclude that DPCML could induce a smaller generalization error than traditional CML. Overall, this experiment empirically suggests the correctness of Corol.1.
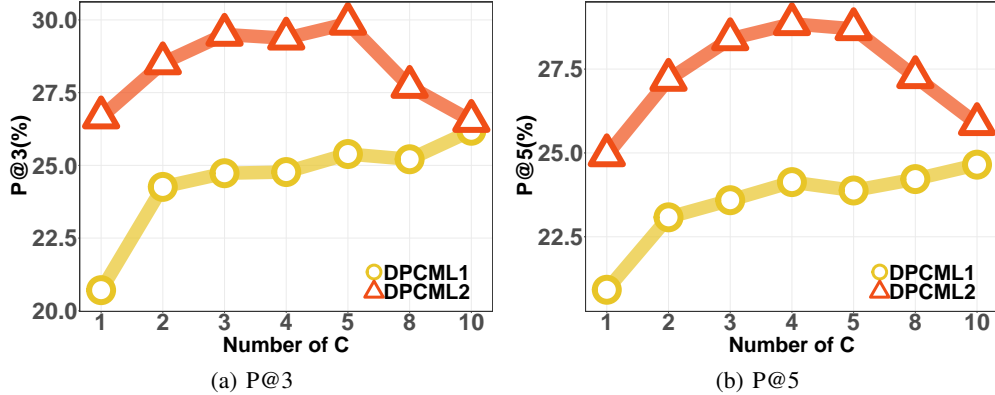
(a) P@3   (b) P@5

Figure 10: Sensitive Analysis of different C.

C.6.5 Sensitive Analysis of C

Fig.10 demonstrates the P@3 and P@5 performance of DPCML methods with different C on Steam-200k dataset. We observe that a proper C could significantly improve the performance. Besides, leveraging C too aggressively for DPCML2 may adversely hurt the performance since models optimized with hard samples are more likely to lead to the over-fitting problem with the increasing parameters.

Table 6: Sensitivity analysis for DPCML1 ($C = 5$) on the Steam-200k dataset.

η	P@3	R@3	NDCG@3	P@5	R@5	NDCG@5	MAP	MRR
1	25.04	14.65	26.01	24.60	12.55	25.81	21.65	45.55
3	24.67	14.43	25.50	23.88	12.25	24.96	21.56	44.73
5	25.24	14.91	26.65	23.80	12.17	25.34	22.17	47.23
10	25.39	14.84	26.56	23.88	12.11	25.25	22.26	46.79
20	24.60	14.34	25.79	24.03	12.05	25.17	21.87	46.20
30	25.23	14.69	26.19	24.25	12.08	25.58	21.94	46.00

Table 7: Sensitivity analysis for DPCML2 ($C = 5$) on the Steam-200k dataset.

η	P@3	R@3	NDCG@3	P@5	R@5	NDCG@5	MAP	MRR
1	28.55	16.35	29.92	27.82	13.94	29.65	22.90	50.57
3	28.68	16.32	29.96	27.71	13.90	29.59	23.13	50.19
5	29.34	16.82	30.45	27.98	13.95	29.75	23.42	50.62
10	29.88	17.13	31.22	28.70	14.51	30.56	24.10	51.95
20	29.81	17.12	31.08	29.11	14.65	30.77	24.35	51.90
30	29.43	16.99	30.67	28.96	14.53	30.56	24.50	51.36

C.6.6 Sensitivity analysis of η

We investigate the sensitivity of $\eta \in \{0, 1, 3, 5, 10, 20, 30\}$ for recommendation results on the Steam-200k dataset. The experimental results are listed in Tab.6 and Tab.7 for DPCML1 and DPCML2, respectively. We can conclude that a proper η (roughly 10) could significantly improve the performance, suggesting the essential role of the proposed diversity control regularization scheme.

C.6.7 Training Efficiency

Since DPCML includes multiple user representations, it will inevitably introduce extra complexity to the overall optimization. We further investigate the training overheads of our proposed algorithm.
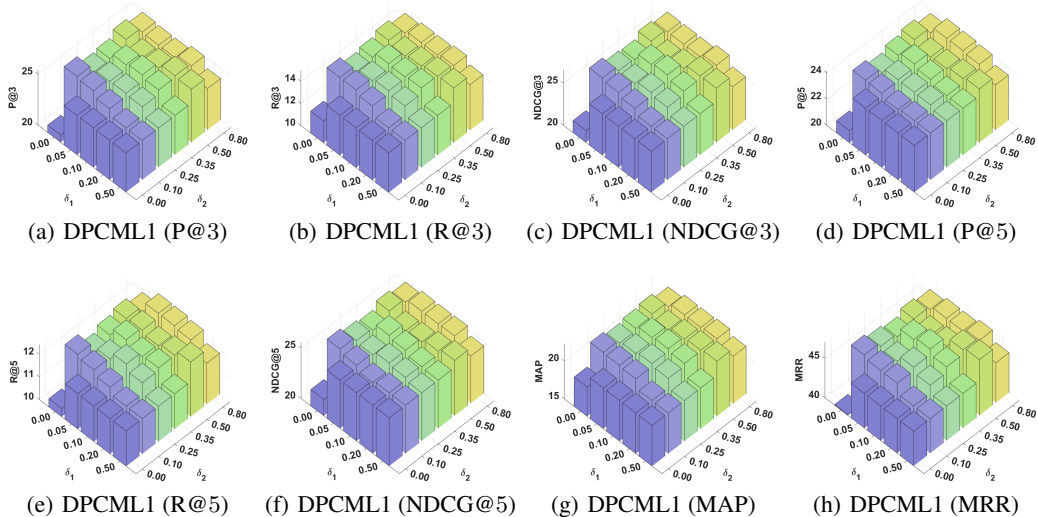
Figure 11: Sensitivity against δ_1 and δ_2 for DPCML1 on Steam-200k datasets. The x- and y-axis stand for the value of δ_1 and δ_2 respectively, and the z-axis shows the performance.

Fig.13 shows the efficiency performance. Specifically, every method is conducted 10 epochs, and the average running time across 10 epochs is reported at the bottom of the boxplot. This trend suggests that our proposed algorithm could achieve competitive performance with acceptable efficiency.

C.6.8 Ablation Studies of Diversity Control Regularization Scheme (DCRS)

In order to show the effectiveness of our proposed DCRS, we compare its performance with the following three variants of DCRS:

- **w/o DCRS**: This is a variant of our method where no regularization is adopted at all.
- **DCRS$-\delta_1$**: This is a variant of our method where the punishment on a **large** diversity is **removed**. In other words, we will use the following regularization term:

$$\psi_{\boldsymbol{g}}(u_i) = \max(0, \delta_1 - \delta_{\boldsymbol{g},u_i}).$$

- **DCRS$-\delta_2$**: This is a variant of our method where the punishment on a **small** diversity is **removed**. In other words, we will use the following regularization term:

$$\psi_{\boldsymbol{g}}(u_i) = \max(0, \delta_{\boldsymbol{g},u_i} - \delta_2).$$

The empircal results on Steam-200k dataset are provided in Tab.8 and Tab.9. From the above results, we can see that: In most cases, only employing one of the two terms of DCRS could still improve the recommendation performance. However, none of them could outperform our proposed method. This strengthens the effectiveness of our proposed regularization scheme.

Table 8: Ablation studies of DPCML1 on Steam-200k dataset.

Method	P@3	R@3	NDCG@3	P@5	R@5	NDCG@5	MAP	MRR
w/o DCRS	23.86	13.06	24.90	23.57	11.56	24.77	20.37	44.38
DCRS-δ_1	24.28	14.38	25.61	22.48	11.35	24.13	21.02	45.76
DCRS-δ_2	24.56	14.36	25.41	23.82	11.97	24.74	21.67	45.14
DCRS	**25.39**	**14.84**	**26.56**	**23.88**	**12.11**	**25.25**	**22.26**	**46.79**

C.6.9 The Effectiveness of DCRS for MF-based Systems

To see this, we attempt to apply the proposed diversity control regularization scheme (DCRS) for M2F [52, 12]. In addition, we further explore the effectiveness of DCRS for the general framework
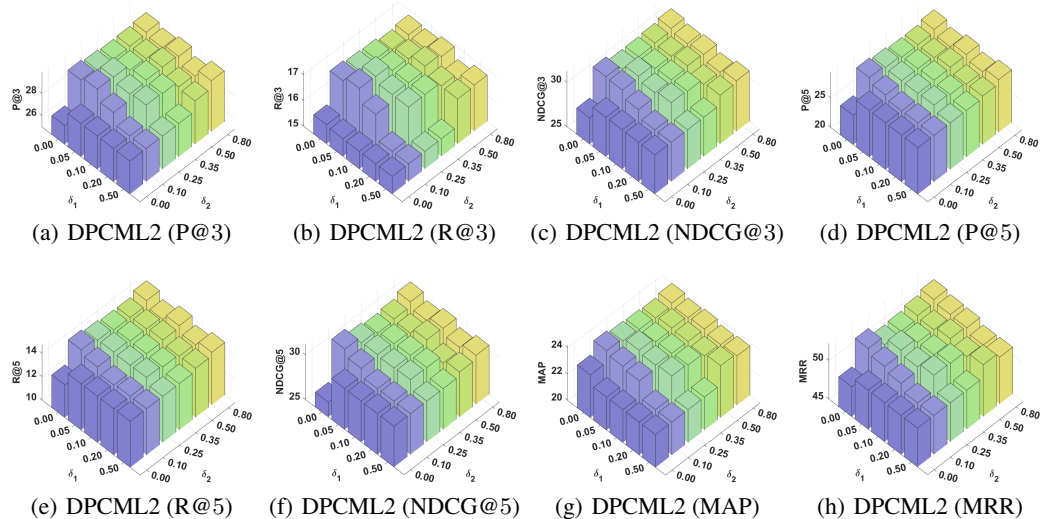
33

(a) DPCML2 (P@3) (b) DPCML2 (R@3) (c) DPCML2 (NDCG@3) (d) DPCML2 (P@5)

(e) DPCML2 (R@5) (f) DPCML2 (NDCG@5) (g) DPCML2 (MAP) (h) DPCML2 (MRR)

Figure 12: Sensitivity against δ_1 and δ_2 for DPCML2 on Steam-200k datasets. The x- and y-axis stand for the value of δ_1 and δ_2 respectively, and the z-axis shows the performance.

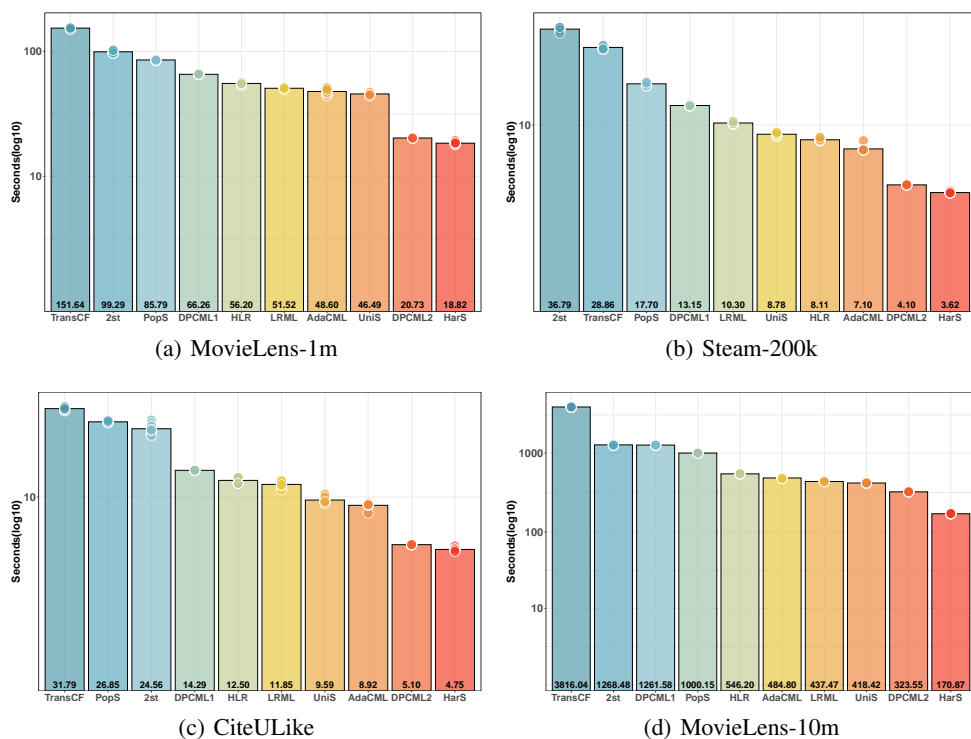(a) MovieLens-1m (b) Steam-200k

(c) CiteULike (d) MovieLens-10m

Figure 13: Training efficiency comparison among CML-based competitors.

of joint accessibility (GFJA, Eq.(9) in the main paper). Here we also conduct a grid search to choose the best performance of M2F with DCRS on the Steam-200k and MovieLens-1m datasets, where the parameters space stays the same as DPCML. The experimental results are summarized in Tab.10. From the above results, we can draw the following observations: 1) The proposed DCRS does not work well for MF-based models. A possible reason here is that the metric space of MF-based and CML-based methods are intrinsically different. MF adopts the inner-product space while CML adopts the Euclidean space. In this paper, we merely consider the DCRS for Euclidean space.

Table 9: Ablation studies of DPCML2 on Steam-200k dataset.

Method	P@3	R@3	NDCG@3	P@5	R@5	NDCG@5	MAP	MRR
w/o DCRS	27.96	15.68	29.42	27.85	13.94	29.56	22.50	50.03
DCRS-δ_1	28.28	16.05	29.60	27.25	13.75	29.17	22.63	50.00
DCRS-δ_2	29.26	16.83	30.61	28.47	14.28	30.16	23.86	51.14
DCRS	**29.88**	**17.13**	**31.22**	**28.70**	**14.51**	**30.56**	**24.10**	**51.95**

The corresponding strategy for the inner-product space is left as future work. 2) In most metrics, GFJA+DCRS could outperform GFJA significantly, which supports the advantages of our proposed DCRS. 3) Compared with M2F, the performance gain of GFJA is sharp on both datasets. This suggests the superiority of our proposed method against the current multi-vector-based competitors.

Table 10: Performance comparison of joint accessibility model equipped with DCRS on the Steam-200k and MovieLens-1m datasets.

Steam-200k								
Method	P@3	R@3	NDCG@3	P@5	R@5	NDCG@5	MAP	MRR
M2F	11.33	5.69	11.95	11.44	5.73	12.98	6.43	25.05
M2F+DCRS	10.92	5.58	11.49	10.89	5.48	12.37	6.25	24.26
GFJA	21.53	**12.60**	22.52	20.37	**10.16**	21.49	19.32	40.69
GFJA+DCRS	**21.63**	12.40	**22.72**	**20.38**	9.98	**21.74**	**19.53**	**40.92**
MovieLens-1m								
M2F	8.61	1.84	9.36	7.60	2.30	8.67	2.95	20.40
M2F+DCRS	7.59	1.49	8.16	7.10	2.02	7.92	2.53	18.51
GFJA	15.79	3.19	16.11	16.02	4.77	16.66	11.04	32.54
GFJA+DCRS	**16.71**	**3.54**	**16.94**	**17.24**	**5.27**	**17.71**	**11.75**	**33.87**