Notations

Let $[L] = \{1, \ldots, L\}$. For a vector $\boldsymbol{v} \in \mathbb{R}^n$, we denote its Euclidean norm by $\|\boldsymbol{v}\|_2 = \sqrt{\boldsymbol{v}^T \boldsymbol{v}}$, and the infinity norm by $\|\boldsymbol{v}\|_\infty = \max_i |\boldsymbol{v}_i|$. For a matrix $\boldsymbol{W} \in \mathbb{R}^{m \times n}$, we denote its Frobenius norm by $\|\boldsymbol{W}\|_F = \sqrt{trace(\boldsymbol{W}^T \boldsymbol{W})}$, the infinity norm by $\|\boldsymbol{W}\|_\infty = \max\limits_{1 \le i \le m} \sum_{j=1}^{n} |\boldsymbol{W}|_{ij}$, and the spectral norm by $\|\boldsymbol{W}\|_2 = \sigma_{\max}(\boldsymbol{W})$. For ease of exposition, we summarize the notations used throughout the paper in Table 2.

Table 2: Summary of notations.

Notation	Meaning
u, i, z	The user, the item and the user-item pair.
y	The label.
$\mathcal{U}, \mathcal{I}, \mathcal{R}, \mathcal{Z}$	The sets of users, items, user-item interactions and user-item pairs.
$\boldsymbol{R}$	The user-item interaction matrix.
$\boldsymbol{A}, \tilde{\boldsymbol{A}}$	The adjacency matrix and the corresponding normalized graph filter.
$\mathcal{D}$	The unknown data distribution.
d	The dimension size.
S_m	The training set.
S, S_u	The fixed set and testing set in transductive learning.
$\boldsymbol{W}_1^l, \boldsymbol{W}_2^l$	The two weight matrices at layer l.
B_1, B_2	The l_2-norm bounds of $\boldsymbol{W}_1^l$ and $\boldsymbol{W}_2^l$.
$\boldsymbol{x}_u, \boldsymbol{x}_i$	The input features of user u and item i.
B_u, B_i	The l_2-norm bounds of any $\boldsymbol{x}_u$ and any $\boldsymbol{x}_i$.
$\boldsymbol{e}_u, \boldsymbol{e}_i$	The learned embeddings of user u and item i.
$\mathbf{h}_u^l, \mathbf{h}_i^l$	The hidden states of user u and item i at layer l.
l, ϕ	The loss function and the non-linear activation.
B	The bound of loss function l.
C_l, C_ϕ	The lipschitz constant of l and ϕ.
$D_{\max}$, $D_{\min}$	The maximum and minimum degrees of node in graph.

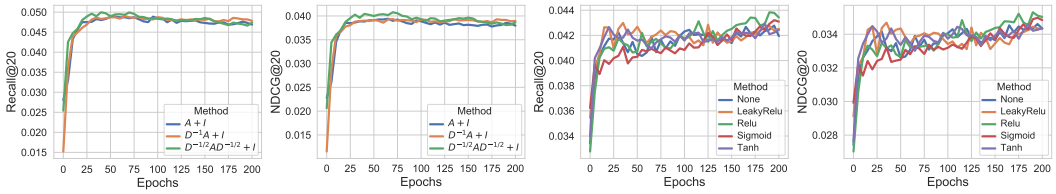
A Experiments

A.1 Hyper-parameter Settings

Following LightGCN, we use its default parameters in most cases. Specifically, unless otherwise specified, the features are initialized using the Xavier method, and their sizes are fixed to 64. For the number of layers, we learn the range from 1 to 3. The regularization factor is fixed to $1e-4$, and the node dropout ratio is set to 0. For NGCF with hidden layer parameters, we set the hidden layer size to 64 and use LeakyRelu as its default activation function. For the proposed strategy, only two parameters, α and β about the beta distribution, both are set to 0.5. The learning rate of all models was set to 0.002 and trained with Adam optimizer for 200 epochs. Additionally, an early stop strategy is used during training.

A.2 Additional Experiments for Numerical Discussion

The Role of Normalized Graph. Fig. 6 shows the performance effect of different Normalized graphs (left) and non-linear activation functions (right) in Yelp2018. Obviously, we can get a similar conclusion to Section 5.1 of the main text.



(a) The generalization effect of normalized graph (b) The generalization effect of activation function

Figure 6: (a): the effect of normalized graph on Generalization Performance on Yelp2018. (b): the effect of activation function on Generalization Performance on Yelp2018.

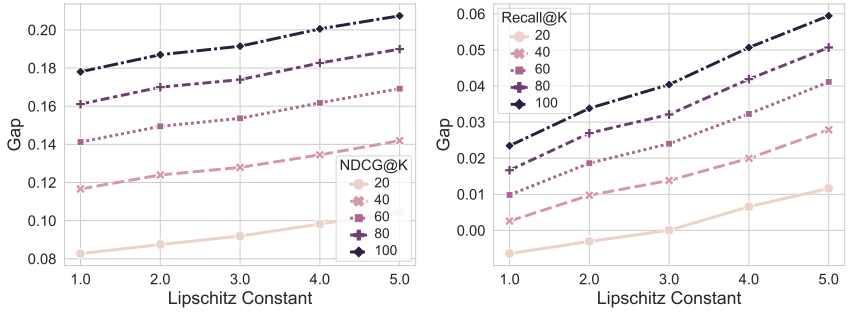


Figure 7: The effect of Lipschitz constant on Generalization performance on Gowalla.

The Role of Non-linear Activation Function. In Section 5.1 of the main text, we explore the impact of different non-linear activation functions on performance. To take a deeper look at how the activation function affects the generalization performance, we consider non-linear activation functions with different Lipschitz constants. Specifically, we multiply the LeakyReLU by different weights, then compare their effects on generalization abilities. From Fig. 7, it can be observed that the generalization gap significantly increases with the Lipschitz constants. These results once again confirm our theoretical findings on the generalization error.

A.3 Performance Analysis on Non-GCN-based Models

Table 3: Overall evaluation on Non-GCN-based models. All models' performance is imporved after using Item Mixtures.

Dataset	**Gowalla**		**Yelp2018**	
Method	**Recall@20**	**NDCG@20**	**Recall@20**	**NDCG@20**
BPRMF	0.1338	0.1144	0.0402	0.0325
BPRMF-IMix	0.1402(+4.81%)	0.1190(+3.99%)	0.0430(+6.91%)	0.0345(+6.09%)
GRMF	0.1354	0.1151	0.0418	0.0335
GRMF-IMix	0.1393(+2.83%)	0.1191(+3.48%)	0.0504(+20.64%)	0.0402(+20.26%)
NGRMF	0.1317	0.1097	0.0408	0.0331
NGRMF-IMix	0.1403(+6.52)	0.1214(+10.65)	0.0432(+5.90%)	0.0349(+5.25%)

In theory, the proposed strategy is suitable for non-GCN recommendation models. We experiment on three factorization-based recommendation models: (1) BPRMF is a traditional matrix decomposition algorithm that decomposes the user-commodity interaction matrix into user embedding and commodity embedding and uses BPR as the loss function. (2) GRMF is a smoothed version of BPRMF that adds graph Laplacian regularization to the loss function. (3) NGRMF is a variant of GRMF that performs normalization on graph Laplacian. Correspondingly, the models configured with the proposed IMix are denoted as BPRMF-IMix, GRMF-IMix, and NGRMF-IMix, respectively. The comparison results are shown in Table 3. As expected, IMix still works well on non-GCN-based models, with an average improvement of 7.94% regarding Recall@20 and 8.29% regarding NDCG@20, which reveals the potential of the proposed enhancement strategy.

A.4 Loss Curve during Training

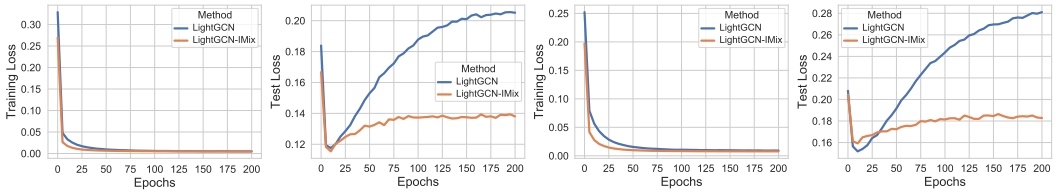


Figure 8: Training curves (training loss) and test curves (test loss) for 200 epochs of training.

Fig. 8 presents the model training and testing curves with and without IMix. During the whole training process, the training loss of LightGCN-IMix converges faster, which indicates that the linear and continuous processing of IMix is more suitable for the training data. Secondly, consistent with the find observed in Fig. 3 of the main text, LightGCN has serious overfitting, while LightGCN-IMix is much lighter, which confirms the positive role of IMix in alleviating overfitting. It is also the key motivation to propose IMix. Furthermore, this observation also validates the theoretical analysis in Section 4 of the main text; that is, IMix essentially adds a regularization term to ensure model generalization.

A.5 Comparison of Transductive Learning and Inductive Learning

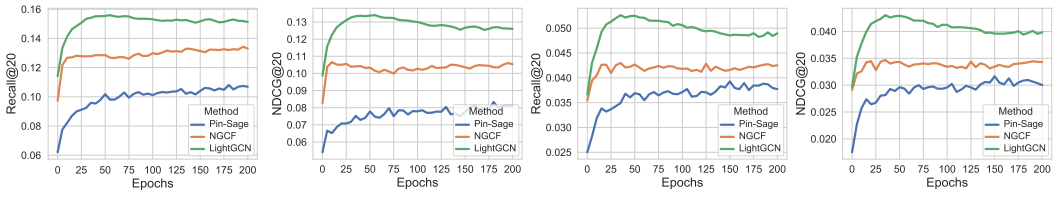


Figure 9: Comparison of test curve (Recall@20 and NDCG@20) of transductive learning (LightGCN, NGCF) and inductive learning (Pin-Sage).

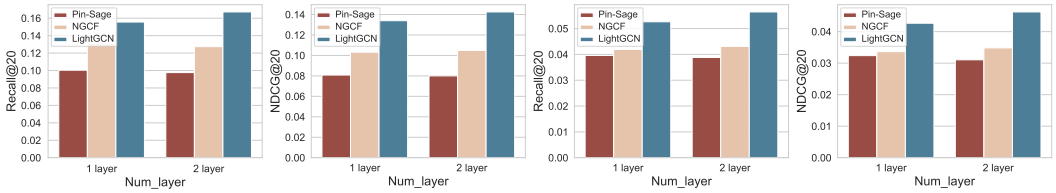


Figure 10: Performance comparison between transductive learning (LightGCN, NGCF) and inductive learning (Pin-Sage) after training.

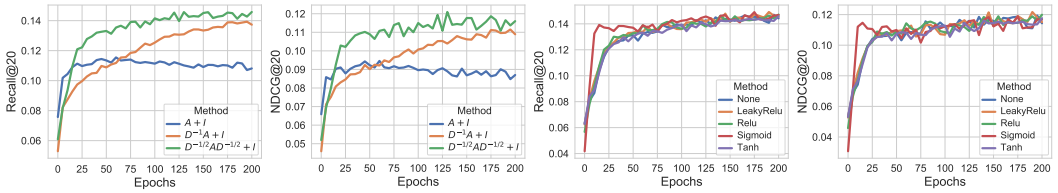
This section compares the GCN-based recommender systems under transductive and inductive learning. In the recommendation field, typical models for these two settings are LightGCN [12] (and NGCF [11]) and Pin-Sage [2], respectively. The difference in model structure makes them unable to carry out a fair numerical analysis. therefore, we only compare their performance, and the results are shown in Fig. 9 and Fig. 10.

In Fig. 9, we find that the performance of Pin-Sage with inductive learning is inferior to LightGCN and NGCF throughout the training process, and this gap persists until the end of training, as shown in Fig. 10. We reasonably suspect that in collaborative filtering, the recommendation results mainly depend on historical interactions rather than node features [44], which reflects that the performance of GCN-based models is more dependent on its neighbors. However, Pin-Sage uses neighbor sampling

to ensure model's flexibility and generalization, which loses a large part of valuable information and damages the recommendation performance.

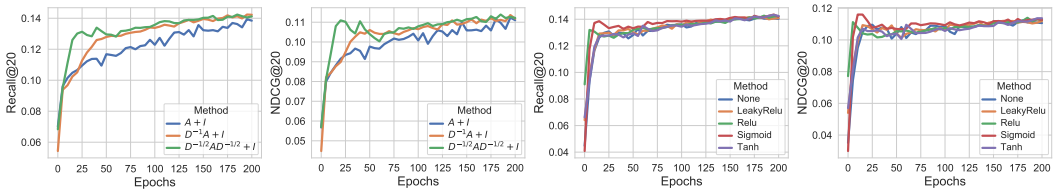
A.6 Additional Experiments on GCN-based Recommendations

In addition to LightGCN and NGCF, we also use two other GCN-based models, SpectralCF [10] and GCMC [9], to verify the effectiveness of our theoretical findings and the proposed IMix. Among them, SpectralCF directly finds all possible connectivity between users and items from the spectral domain of the user-item graph, while GCMC uses a graph auto-encoder to transmit messages on the interaction graph to find the potential interests of users.



(a) The generalization effect of normalized graph (b) The generalization effect of activation function

Figure 11: Numerical analysis based on SpectralCF. (a): the generalization effect of normalized graph on Gowalla. (b): the generalization effect of activation function on Gowalla.



(a) The generalization effect of normalized graph (b) The generalization effect of activation function

Figure 12: Numerical analysis based on GCMC. (a): the generalization effect of normalized graph on Gowalla. (b): the generalization effect of activation function on Gowalla.

Numerical Discussion. We evaluate the role of normalized graph and non-linear activation function on SpectralCF as well as GCMC, and the results are illustrated in Fig. 11 and Fig. 12. We can get similar findings to Section 5.1 and A.2: (1) Graph normalization helps to improve recommendation performance (Fig. 11 (a) and Fig. 12 (a)). (2) The activation function has little effect on the model performance (Fig. 11 (b) and Fig. 12 (b)). This further supports the theoretical findings in Section 3.2.

Effectiveness of IMix. In Section 5.2, we have verified the effectiveness of IMIx on LightGCN and NGCF. In this section, we further apply IMix to SpectralCF and GCMC (denoted as SpectralCF-IMix and GCMC-IMix), and the results are shown in Table 4. It can be clearly found that the models' recommendation performance equipped IMix can significantly outperform the original model under different structures. Specifically, the average improvement on Recall@20 is 11.86%, while the average improvement is 12.70% concerning NDCG@20. These encouraging results emphasize the potential of IMix on GCN-based recommender systems.

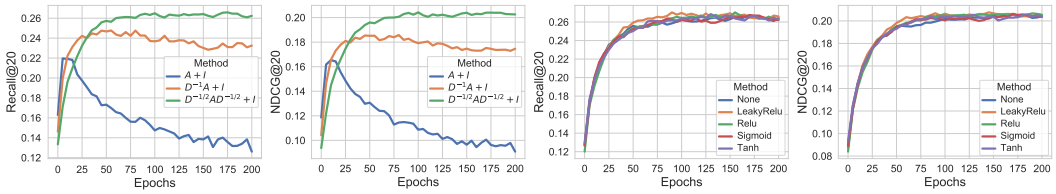
A.7 Experiments on LastFM

In this section, we introduce LastFM to enhance our work. LastFM is a music recommendation dataset consisting of 1,027,370 ratings for 4449 artists from 1892 user ratings collected from the music website of Last.fm. The item is defined as an artist. The training, validation, and test sets are randomly divided in a ratio of 7:1:2.

Numerical Discussion. We construct experiments on the effect of normalized graph and activation function on recommendation performance, as shown in Fig. 13 (a) and Fig. 13 (b). We can get similar

Table 4: Overall evaluation. SpectralCF and GCMC's performance is improved after configured with IMix.

Dataset		Gowalla		Yelp2018	
Layer#	**Method**	**Recall@20**	**NDCG@20**	**Recall@20**	**NDCG@20**
1 layer	SpectralCF	0.1424	0.1171	0.0496	0.0400
	SpectralCF-IMix	0.1588(+11.52%)	0.1314(+12.21%)	0.0546(+10.08%)	0.0443(+10.75%)
2 layer	SpectralCF	0.1435	0.1162	0.0501	0.0407
	SpectralCF-IMix	0.1531(+6.69%)	0.1265(+8.86%)	0.0543(+8.38%)	0.0438(+7.62%)
3 layer	SpectralCF	0.1395	0.1161	0.0493	0.0397
	SpectralCF-IMix	0.1540(+10.39%)	0.1281(+10.34%)	0.0560(+13.59%)	0.0452(+13.85%)
1 layer	GCMC	0.1328	0.1142	0.0484	0.0387
	GCMC-IMix	0.1491(+12.27%)	0.1201(+5.16%)	0.0546(+12.81%)	0.0441(+13.95%)
2 layer	GCMC	0.1324	0.1099	0.0472	0.0377
	GCMC-IMix	0.1533(+15.79%)	0.1354(+23.20%)	0.0554(+17.37%)	0.0440(+16.71%)
3 layer	GCMC	0.1443	0.1142	0.0480	0.0382
	GCMC-IMix	0.1570(+8.80%)	0.1288(+12.78%)	0.0550(+14.58%)	0.0447(+17.02%)



(a) The generalization effect of normalized graph

(b) The generalization effect of activation function

Figure 13: (a): the effect of normalized graph on Generalization Performance on LastFM. (b): the effect of activation function on Generalization Performance on LastFM. Here the model is LightGCN.

but more obvious findings as in Gowalla (in Fig. 2) and Yelp2018 (in Fig. 6): (1) The performance on the unnormalized graph is significantly worse than the other two normalized graphs, and it starts to decline after 5 rounds. This underscores our point in Section 3.2 that the normalized graph can yield generalization gains. (2) The effect of different activation functions on recommendation performance is negligible. Therefore, it is proved that removing the activation function can simplify the model without incurring a performance penalty, which is consistent with the findings in [12].

Table 5: Performance comparison with and without IMix in LastFM.

Last.fm		LightGCN		NGCF	
Layer#	**Method**	**Recall@20**	**NDCG@20**	**Recall@20**	**NDCG@20**
1 layer	Without IMix	0.2694	0.2068	0.2341	0.1746
	With IMix	0.2731(+1.37%)	0.2108(+1.95%)	0.2635(+12.50%)	0.1986(+13.76%)
2 layer	Without IMix	0.2731	0.2101	0.2299	0.1714
	With IMix	0.2732(+0.04%)	0.2129(+1.28%)	0.2630(+14.41%)	0.1991(+16.15%)
3 layer	Without IMix	0.2700	0.2090	0.2347	0.1738
	With IMix	0.2585(-4.25%)	0.1986(-5.17%)	0.2549(+8.58%)	0.1915(+10.12%)

Effectiveness of IMix. We evaluate the performance of LightGCN and NGCF using IMix in LastFM. Although on 3-layer-LightGCN, our method has a drop of about 5%, in other cases, the IMix models' performance has improved, and the improvement is particularly obvious in NGCF, with an average improvement is 11.83% concerning Recall@20.

A.8 Comparison with Mixup-based Methods

Differences from Recent Mixup-based Methods. More recently, some Mixup-based strategies have been introduced in GNNs, but our work is quite different from them: (1) G-Mixup proposed in

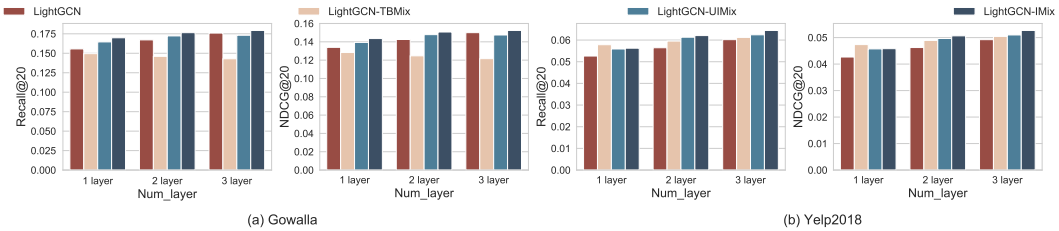


Figure 14: The performance comparison of LightGCN configured various Mixup strategies in Gowalla and Yelp2018.

[45] studies graph classification tasks, while our recommendation task is essentially link prediction. Besides, G-Mixup enhances the model by interpolating the graphon of different classes of graphs, but we only mix the item embeddings of the same users; (2) [46] proposed to mix two-branch features and used Mixup at each layer, while our IMix does not introduce additional convolution branches and only performs item mixture on the last layer. In addition, we give the theoretical guarantee of IMix, which is lacking in [46].

Performance Comparison with Mixup-based Methods. We compare IMix with the strategy in [46] (two branch Mixup, denoted as TBMix) and UIMix (a variant of IMix that mixes both user and item embeddings). Note that we do not consider the version that only mixes user embeddings, because it must ensure that the positive and negative items are the same, i.e., $e_i = e_j$ and $e_{i'} = e_{j'}$, otherwise, a quintuple $(\tilde{e}_u, e_i, e_{i'}, e_j, e_{j'})$ instead of triplet $(\tilde{e}_u, e_i, e_{i'})$. We apply these strategies on LightGCN, denoted LightGCN-TBMix, and LightGCN-UIMix, respectively, and the results are shown in Fig. 14. First, although TBMix shows good performance on Yelp2018, it performs poorly on Gowalla, even inferior to LightGCN. Second, the proposed IMix is the best among the compared methods. Third, the performance of UIMix, which mixes both user and item embeddings, is not as good as IMix, which only mixes item embeddings. One possible reason is that the recommender system has two types of feature spaces: user space and item space. IMix mixes in the item space of the same user (same user space), while UIMix mixes across spaces (different user space and different item space), which may lead to confusion in learning.

A.9 Experiments on node-features-included Models

In collaborative filtering, due to the lack of rich features, random embeddings are generally used as initial features and learned during training, which is especially common in GCN-based recommender systems [2, 10, 12, 11]. In this section, however, we explore whether recommendation models using raw features can achieve similar findings. To solve the misalignment of user and item features, we use a Multilayer Perceptron to map them into the space with the same dimensionality and apply LightGCN and NGCF on it (denoted as F-LightGCN and F-NGCF).

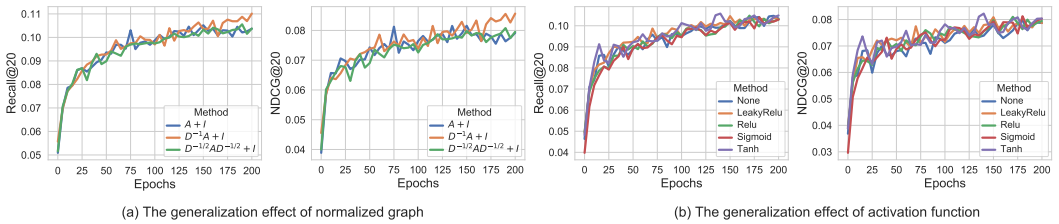


Figure 15: Numerical analysis based on F-LightGCN. (a): the generalization effect of normalized graph on Gowalla. (b): the generalization effect of activation function on Gowalla.

Numerical Discussion. We construct experiments on the effect of the normalized graph and activation function based on F-LightGCN, as shown in Fig. 15. In Fig. 15 (a), we find that the performance gap of whether the graph is normalized or not is much smaller than the model that does not use the node features (Fig. 2). This may be that the node features weaken the high dependence on the graph structure. But it is undeniable that our finding still holds that using graph normalization helps model

generalization. Second, the performance of different activation functions is still indistinguishable. This provides support for simplifying the removal of nonlinear layers in GCN-based recommender systems.

Table 6: Performance comparison of node-features-included model configured IMix.

Dataset		Gowalla		Yelp2018	
Layer#	**Methods**	**Recall@20**	**NDCG@20**	**Recall@20**	**NDCG@20**
1 layer	F-NGCF	0.1060	0.0820	0.0338	0.0268
	F-NGCF-IMix	0.1217(+14.81%)	0.0965(+17.69%)	0.0422(+24.85%)	0.0337(+25.75%)
2 layer	F-NGCF	0.1056	0.0822	0.0337	0.0269
	F-NGCF-IMix	0.1254(+18.75%)	0.0968(+17.76%)	0.0431(+27.89%)	0.0346(+28.62%)
3 layer	F-NGCF	0.1046	0.0830	0.0376	0.0301
	F-NGCF-IMix	0.1277(+22.08%)	0.1008(+21.45%)	0.0442(+17.55%)	0.0350(+16.28%)
1 layer	F-LightGCN	0.1334	0.1124	0.0445	0.0361
	F-LightGCN-IMix	0.1552(+16.34%)	0.1334(+18.68%)	0.0530(+19.10%)	0.0435(+20.50%)
2 layer	F-LightGCN	0.1498	0.1273	0.0498	0.0403
	F-LightGCN-IMix	0.1588(+6.01%)	0.1359(+6.76%)	0.0572(+14.86%)	0.0468(+16.13%)
3 layer	F-LightGCN	0.1578	0.1348	0.0548	0.0450
	F-LightGCN-IMix	0.1606(+1.77%)	0.1361(+0.96%)	0.0576(+5.11%)	0.0468(+4.00%)

Numerical Discussion. We use the proposed augmentation strategy IMix (denoted as F-NGCF-IMix and F-LightGCN-IMix) on F-NGCF as well as F-LightGCN, and the results are shown in Table 6. The performance of the models using IMix has been gratifyingly improved. Considering Recall@20, the average improvement is 15.76%, and the maximum improvement is an astonishing 27.89%. These results again validate the ability of IMix to enhance recommendations.

A.10 Visualization

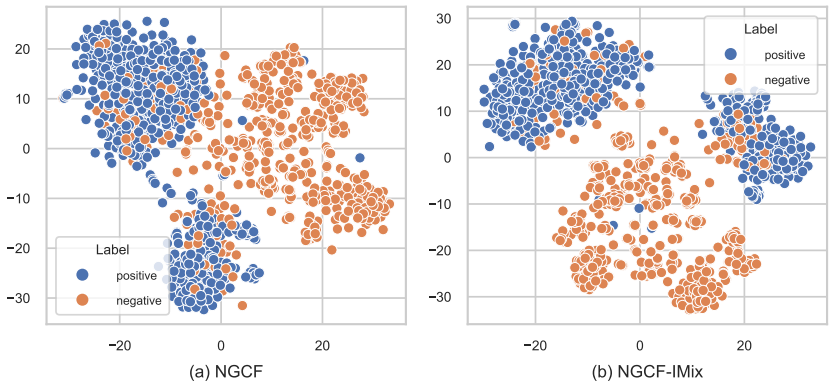


Figure 16: The t-SNE visualization of the embeddings of positive and negative items for a single user with or without IMix.

To intuitively measure the quality of model learning, we use the t-SNE tool [47] to visualize the embeddings of positive and negative items of the same user learned by NGCF and NGCF-IMix. We experiment on Gowalla and randomly sample the same number of negative items as positive items. The visualization results are shown in Fig. 16. Blue nodes denote the embeddings of positive items, and orange items represent the embeddings of negative ones. We can observe that compared to NGCF, NGCF-IMix can better distinguish positive and negative items, which indicates that using the proposed IMix has more potential to learn user preferences.

B Proof of Lemma 1

We first present a lemma according to the covering number of matrices with bounded spectral norm.

Lemma B.1 ([34]). *Let $\mathcal{G} = \{\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2} : \|\boldsymbol{A}\|_2 \leq \lambda\}$ be the set of matrices with bounded spectral norm and $\epsilon \geq 0$ be given. The covering number $\mathcal{N}(\mathcal{G}, \epsilon, \|\cdot\|_F)$ is upper bounded by*

$$\mathcal{N}(\mathcal{G}, \epsilon, \|\cdot\|_F) \leq \left(1 + 2\frac{\min\{\sqrt{d_1}, \sqrt{d_2}\}\lambda}{\epsilon}\right)^{d_1 d_2}$$

Then we provide the proof of Lemma 1. For ease of understanding, we summarize the proof strategy used in this section as follows.

- Investigate the change in recommendations predicted by models due to the change in weights. (**Step 1**&**Step 2**&**Step 3**)
- Derive the covering numbers of weight matrices with bounded norms using Lemma B.1 and yield final result. (**Step 4**)

Lemma 1 (Covering number bound)**.** *Under Assumptions, we further assume that $\|\mathbf{h}_u^l\|_\infty, \|\mathbf{h}_i^l\|_\infty \leq b$ for any $u \in \mathcal{U}$, $i \in \mathcal{I}$, and $l = [L]$, let $\gamma = \|\widetilde{\boldsymbol{A}}\|_\infty$. Given a sample set S with size n, the covering number of $\mathcal{F}$ over S with specific ϵ is bounded as*

$$\log \mathcal{N}\left(\mathcal{F}_{|S}, \epsilon, \|\cdot\|_\infty\right) \leq d^2 \log\left(1 + \frac{4(\gamma+1)MB_1\sqrt{d}}{\epsilon}\right)\left(1 + \frac{4\gamma b M B_2 \sqrt{d}}{\epsilon}\right).$$

Moreover, when $\epsilon \leq 4M\sqrt{d}\max\{(\gamma+1)B_1, \gamma b B_2\}$,

$$\begin{aligned}
&\log \mathcal{N}\left(\mathcal{F}_{|S}, \epsilon, \|\cdot\|_\infty\right) \leq 2d^2 \log \frac{8M(\gamma+1)\sqrt{dB_1B_2b}}{\epsilon}, \ where \\
&M = C_\phi \mathcal{C}^{2L-1}(B_u + B_i)^2 \frac{(2\mathcal{C})^L - 1}{2\mathcal{C} - 1}, \\
&\mathcal{C} = C_\phi[B_1 + \gamma(B_1 + B_2 b)].
\end{aligned}$$

Proof. Let f to be a GCN-based recommendation model defined in Section 2.2 with a set of weight matrices $\{\boldsymbol{W}_1, \boldsymbol{W}_2\}$, and f' with $\{\boldsymbol{W}_1', \boldsymbol{W}_2'\}$. For any user-item (u, i), we first perform that the preference score predicted by f and f' can be bounded by parameters. Then we derive the $l - \infty-$covering number bound for the function class $\mathcal{F}$. For convenience, we denote the user representation for user u and the item representation for item w at layer l generated by model f as $\boldsymbol{u}_l$ and $\boldsymbol{w}_l$, and those generated by model f' as $\boldsymbol{u}_l'$ and $\boldsymbol{w}_l'$.

Step 1: Max norm of user representation and item representation. We first bound the maximum spectral norm of user embeddings at layer l.

Let $\mathbf{n}_l = \sum_{w \in \mathcal{N}_u} \tilde{a}_{uw} \left(\boldsymbol{W}_1 \mathbf{w}_{l-1} + \boldsymbol{W}_2 \left(\mathbf{u}_{l-1} \odot \mathbf{w}_{l-1}\right)\right)$, since $\phi(0) = 0$, we have

$$\begin{aligned}
\max_{u \in \mathcal{U}, w \in \mathcal{I}} \|\mathbf{u}_l\|_2 &= \max_{u \in \mathcal{U}, w \in \mathcal{I}} \|\phi\left(\boldsymbol{W}_1 \mathbf{u}_{l-1} + \mathbf{n}_l\right)\|_2 \\
&= \max_{u \in \mathcal{U}, w \in \mathcal{I}} \|\phi\left(\boldsymbol{W}_1 \mathbf{u}_{l-1} + \mathbf{n}_l\right) - \phi(\mathbf{0})\|_2 \\
&\leq C_\phi \cdot \max_{u \in \mathcal{U}, w \in \mathcal{I}} \left\| \boldsymbol{W}_1 \mathbf{u}_{l-1} + \sum_{w \in \mathcal{N}_u} \tilde{a}_{uw} \left(\boldsymbol{W}_1 \mathbf{w}_{l-1} + \boldsymbol{W}_2 \left(\mathbf{u}_{l-1} \odot \mathbf{w}_{l-1}\right)\right) \right\|_2 \\
&\leq C_\phi \cdot \max_{u \in \mathcal{U}, w \in \mathcal{I}} \left[\|\boldsymbol{W}_1 \mathbf{u}_{l-1}\|_2 + \sum_{w \in \mathcal{N}_u} |\tilde{a}_{uw}| \left(\|\boldsymbol{W}_1 \mathbf{w}_{l-1}\|_2 + \|\boldsymbol{W}_2 \left(\mathbf{u}_{l-1} \odot \mathbf{w}_{l-1}\right)\|_2\right) \right] \\
&\leq C_\phi \cdot \max_{u \in \mathcal{U}, w \in \mathcal{I}} \|\boldsymbol{W}_1\|_2 \|\mathbf{u}_{l-1}\|_2 \\
&+ C_\phi \cdot \left(\max_{u \in \mathcal{U}} \sum_{w \in \mathcal{N}_u} |\tilde{a}_{uw}|\right) \cdot \max_{u \in \mathcal{U}, w \in \mathcal{I}} \left(\|\boldsymbol{W}_1\|_2 \|\mathbf{w}_{l-1}\|_2 + \|\boldsymbol{W}_2\|_2 \|\mathbf{u}_{l-1} \odot \mathbf{w}_{l-1}\|_2\right).
\end{aligned}$$

The first inequality is due to the activation function $\phi(\cdot)$ is $C_\phi - lipschitz$. Since $\|\boldsymbol{W}_1\|_2 \leq B_1, \|\boldsymbol{W}_2\|_2 \leq B_2$ and

$$\max_{u\in\mathcal{U}} \sum_{w\in\mathcal{N}_u} |\tilde{a}_{uw}| = \|\widetilde{\boldsymbol{R}}_u\|_\infty \leq \|\widetilde{\boldsymbol{A}}\|_\infty = \gamma,$$

we have

$$\begin{aligned}\max_{u\in\mathcal{U},w\in\mathcal{I}} \|\mathbf{u}_l\|_2 &\leq C_\phi \cdot \max_{u\in\mathcal{U},w\in\mathcal{I}} \left[B_1 \|\mathbf{u}_{l-1}\|_2 + \gamma B_1 \|\mathbf{w}_{l-1}\|_2 + \gamma B_2 \|\mathbf{u}_{l-1} \odot \mathbf{w}_{l-1}\|_2\right] \\ &\leq C_\phi \cdot \max_{u\in\mathcal{U},w\in\mathcal{I}} \left[B_1 \|\mathbf{u}_{l-1}\|_2 + \gamma \left(B_1 + B_2 \|\mathbf{u}_{l-1}\|_\infty\right) \|\mathbf{w}_{l-1}\|_2\right].\end{aligned}$$

The last inequality is due to $\|\boldsymbol{a} \odot \boldsymbol{b}\|_2 \leq \|\boldsymbol{a}\|_2 \|\boldsymbol{b}\|_\infty$ for any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$. Together with the assumption about the infinity norm bound of node representation at any layer, which means $\|\mathbf{u}_{l-1}\|_\infty \leq b$, we can get

$$\max_{u\in\mathcal{U},w\in\mathcal{I}} \|\mathbf{u}_l\|_2 \leq C_\phi \cdot \max_{u\in\mathcal{U},w\in\mathcal{I}} \left[B_1 \|\mathbf{u}_{l-1}\|_2 + \gamma \left(B_1 + B_2 b\right) \|\mathbf{w}_{l-1}\|_2\right]. \tag{7}$$

Similarly,

$$\max_{u\in\mathcal{U},w\in\mathcal{I}} \|\mathbf{w}_l\|_2 \leq C_\phi \cdot \max_{u\in\mathcal{U},w\in\mathcal{I}} \left[B_1 \|\mathbf{w}_{l-1}\|_2 + \gamma \left(B_1 + B_2 b\right) \|\mathbf{u}_{l-1}\|_2\right]. \tag{8}$$

Combining Eq. (7) and Eq. (8), we have

$$T_l = \max_{u\in\mathcal{U},w\in\mathcal{I}} \left(\|\mathbf{u}_l\|_2 + \|\mathbf{w}_l\|_2\right) \leq C_\phi T_{l-1} \left[(\gamma+1)B_1 + \gamma B_2 b\right].$$

Let $T_l \leq \mathcal{C} T_{l-1}$, where $\mathcal{C} = C_\phi \left[(\gamma+1)B_1 + \gamma B_2 b\right]$. Since we assume that $\max_{u\in\mathcal{U}} \|\boldsymbol{x}_u\|_2 \leq B_u$ and $\max_{i\in\mathcal{I}} \|\boldsymbol{x}_i\|_2 \leq B_i$, then we have

$$T_0 = \max_{u\in\mathcal{U},w\in\mathcal{I}} \left(\|\mathbf{u}_0\|_2 + \|\mathbf{w}_0\|_2\right) = \max_{u\in\mathcal{U},i\in\mathcal{I}} \left(\|\boldsymbol{x}_u\|_2 + \|\boldsymbol{x}_i\|_2\right) \leq B_u + B_i,$$

and expanding the recursion $T_l \leq \mathcal{C} T_{l-1}$, we get

$$T_l = \max_{u\in\mathcal{U},w\in\mathcal{I}} \left(\|\mathbf{u}_l\|_2 + \|\mathbf{w}_l\|_2\right) \leq \mathcal{C}^l T_0 \leq \mathcal{C}^l (B_u + B_i). \tag{9}$$

Step 2: Max change of user representation and item representation. We first bound the maximum difference between user representations at layer l generated by f and f'.

Let $\mathbf{n}_l = \sum_{w\in\mathcal{N}_u} \tilde{a}_{uw} \left(\boldsymbol{W}_1 \mathbf{w}_{l-1} + \boldsymbol{W}_2 \left(\mathbf{u}_{l-1} \odot \mathbf{w}_{l-1}\right)\right)$, then we have

$$\begin{aligned}\max_{u\in\mathcal{U},w\in\mathcal{I}} \|\mathbf{u}_l - \mathbf{u}'_l\|_2 &= \max_{u\in\mathcal{U},w\in\mathcal{I}} \left\|\phi\left(\boldsymbol{W}_1 \mathbf{u}_{l-1} + \mathbf{n}_l\right) - \phi\left(\boldsymbol{W}'_1 \mathbf{u}'_{l-1} + \mathbf{n}'_l\right)\right\|_2 \\ &\leq C_\phi \cdot \max_{u\in\mathcal{U},w\in\mathcal{I}} \left\|\left(\boldsymbol{W}_1 \mathbf{u}_{l-1} + \mathbf{n}_l\right) - \left(\boldsymbol{W}'_1 \mathbf{u}'_{l-1} + \mathbf{n}'_l\right)\right\|_2 \\ &\leq C_\phi \cdot \max_{u\in\mathcal{U},w\in\mathcal{I}} \left(\left\|\boldsymbol{W}_1 \mathbf{u}_{l-1} - \boldsymbol{W}'_1 \mathbf{u}'_{l-1}\right\|_2 + \|\mathbf{n}_l - \mathbf{n}'_l\|_2\right).\end{aligned} \tag{10}$$

Here the first inequality is due to the activation function $\phi(\cdot)$ is $C_\phi - lipschitz$. Then we proceed to bound the two terms in Eq. (10).

$$\begin{aligned}\left\|\boldsymbol{W}_1 \mathbf{u}_{l-1} - \boldsymbol{W}'_1 \mathbf{u}'_{l-1}\right\|_2 &\leq \|\boldsymbol{W}_1 \mathbf{u}_{l-1} - \boldsymbol{W}'_1 \mathbf{u}_{l-1}\|_2 + \left\|\boldsymbol{W}'_1 \mathbf{u}_{l-1} - \boldsymbol{W}'_1 \mathbf{u}'_{l-1}\right\|_2 \\ &\leq \|\boldsymbol{W}_1 - \boldsymbol{W}'_1\|_2 \|\mathbf{u}_{l-1}\|_2 + \|\boldsymbol{W}'_1\|_2 \left\|\mathbf{u}_{l-1} - \mathbf{u}'_{l-1}\right\|_2 \\ &\leq \|\boldsymbol{W}_1 - \boldsymbol{W}'_1\|_2 \|\mathbf{u}_{l-1}\|_2 + B_1 \left\|\mathbf{u}_{l-1} - \mathbf{u}'_{l-1}\right\|_2.\end{aligned} \tag{11}$$

The last inequality is due to $\|\boldsymbol{W}_1\|_2 \leq B_1$.

$$\begin{aligned}&\max_{u\in\mathcal{U},w\in\mathcal{I}} \|\mathbf{n}_l - \mathbf{n}'_l\|_2 \\ &\leq \max_{u\in\mathcal{U},w\in\mathcal{I}} \sum_{w\in\mathcal{N}_u} |\tilde{a}_{uw}| \left\|\left(\boldsymbol{W}_1 \mathbf{w}'_{l-1} - \boldsymbol{W}'_1 \mathbf{w}'_{l-1}\right) + \left(\boldsymbol{W}_2 \left(\mathbf{u}_{l-1} \odot \mathbf{w}_{l-1}\right) - \boldsymbol{W}'_2 \left(\mathbf{u}'_{l-1} \odot \mathbf{w}'_{l-1}\right)\right)\right\|_2 \\ &\leq \max_{u\in\mathcal{U},w\in\mathcal{I}} \left(\left\|\boldsymbol{W}_1 \mathbf{w}'_{l-1} - \boldsymbol{W}'_1 \mathbf{w}'_{l-1}\right\|_2 + \left\|\boldsymbol{W}_2 \left(\mathbf{u}_{l-1} \odot \mathbf{w}_{l-1}\right) - \boldsymbol{W}'_2 \left(\mathbf{u}'_{l-1} \odot \mathbf{w}'_{l-1}\right)\right\|_2\right) \\ &\times \left(\max_{u\in\mathcal{U}} \sum_{w\in\mathcal{N}_u} |\tilde{a}_{uw}|\right).\end{aligned} \tag{12}$$

Similar to Eq. (11),

$$\left\|\boldsymbol{W}_1\mathbf{w}_{l-1}-\boldsymbol{W}_1'\mathbf{w}_{l-1}'\right\|_2 \leq \|\boldsymbol{W}_1-\boldsymbol{W}_1'\|_2\,\|\mathbf{w}_{l-1}\|_2 + B_1\left\|\mathbf{w}_{l-1}-\mathbf{w}_{l-1}'\right\|_2. \tag{13}$$

And since $\|\boldsymbol{W}_2\|_2 \leq B_2$, we have

$$\begin{aligned}
&\left\|\boldsymbol{W}_2\left(\mathbf{u}_{l-1}\odot\mathbf{w}_{l-1}\right)-\boldsymbol{W}_2'\left(\mathbf{u}_{l-1}'\odot\mathbf{w}_{l-1}'\right)\right\|_2\\
&\leq \|\boldsymbol{W}_2-\boldsymbol{W}_2'\|_2\,\|\mathbf{u}_{l-1}\odot\mathbf{w}_{l-1}\|_2 + B_2\left\|\mathbf{u}_{l-1}\odot\mathbf{w}_{l-1}-\mathbf{u}_{l-1}'\odot\mathbf{w}_{l-1}'\right\|_2\\
&\leq \|\boldsymbol{W}_2-\boldsymbol{W}_2'\|_2\,\|\mathbf{u}_{l-1}\odot\mathbf{w}_{l-1}\|_2 + B_2\left(\left\|\mathbf{u}_{l-1}\odot\left(\mathbf{w}_{l-1}-\mathbf{w}_{l-1}'\right)\right\|_2 + \left\|\left(\mathbf{u}_{l-1}-\mathbf{u}_{l-1}'\right)\odot\mathbf{w}_{l-1}'\right\|_2\right)\\
&\leq \|\boldsymbol{W}_2-\boldsymbol{W}_2'\|_2\,\|\mathbf{u}_{l-1}\|_2\,\|\mathbf{w}_{l-1}\|_\infty + B_2\left(\|\mathbf{u}_{l-1}\|_\infty\left\|\mathbf{w}_{l-1}-\mathbf{w}_{l-1}'\right\|_2 + \left\|\mathbf{u}_{l-1}-\mathbf{u}_{l-1}'\right\|_2\left\|\mathbf{w}_{l-1}'\right\|_\infty\right).
\end{aligned} \tag{14}$$

The last inequality is due to $\|\boldsymbol{a}\odot\boldsymbol{b}\|_2 \leq \|\boldsymbol{a}\|_2\|\boldsymbol{b}\|_\infty$ for any $\boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^d$. Together with the assumption about the infinity norm bound of node representation at any layer, which means $\|\mathbf{u}_{l-1}\|_\infty \leq b$ and $\left\|\mathbf{w}_{l-1}'\right\|_\infty \leq b$. Substituting Eq. (13) and Eq.(14) into Eq. (12), and since

$$\max_{u\in\mathcal{U}}\sum_{w\in\mathcal{N}_u}|\tilde{a}_{uw}| = \|\widetilde{\boldsymbol{R}}_u\|_\infty \leq \|\widetilde{\boldsymbol{A}}\|_\infty = \gamma,$$

we can get

$$\begin{aligned}
&\max_{u\in\mathcal{U},w\in\mathcal{I}}\|\mathbf{n}_l-\mathbf{n}_l'\|_2\\
&\leq \gamma\cdot\max_{u\in\mathcal{U},w\in\mathcal{I}}\left(\|\boldsymbol{W}_1-\boldsymbol{W}_1'\|_2\,\|\mathbf{w}_{l-1}\|_2 + B_1\left\|\mathbf{w}_{l-1}-\mathbf{w}_{l-1}'\right\|_2\right)\\
&+\gamma b\cdot\max_{u\in\mathcal{U},w\in\mathcal{I}}\left(\|\boldsymbol{W}_2-\boldsymbol{W}_2'\|_2\,\|\mathbf{u}_{l-1}\|_2 + B_2\left\|\mathbf{w}_{l-1}-\mathbf{w}_{l-1}'\right\|_2 + B_2\left\|\mathbf{u}_{l-1}-\mathbf{u}_{l-1}'\right\|_2\right).
\end{aligned} \tag{15}$$

Combining Eq. (11) and Eq. (15), we have

$$\begin{aligned}
&\max_{u\in\mathcal{U},w\in\mathcal{I}}\|\mathbf{u}_l-\mathbf{u}_l'\|_2\\
&\leq C_\phi\cdot\max_{u\in\mathcal{U},w\in\mathcal{I}}\left[\|\mathbf{u}_{l-1}\|_2\left(\|\boldsymbol{W}_1-\boldsymbol{W}_1'\|_2+\gamma b\,\|\boldsymbol{W}_2-\boldsymbol{W}_2'\|_2\right)+\gamma\,\|\mathbf{w}_{l-1}\|_2\,\|\boldsymbol{W}_1-\boldsymbol{W}_1'\|_2\right]\\
&+C_\phi\cdot\max_{u\in\mathcal{U},w\in\mathcal{I}}\left[\left\|\mathbf{u}_{l-1}-\mathbf{u}_{l-1}'\right\|_2\left(B_1+\gamma B_2 b\right)+\left\|\mathbf{w}_{l-1}-\mathbf{w}_{l-1}'\right\|_2\gamma\left(B_1+B_2 b\right)\right].
\end{aligned} \tag{16}$$

Similarly, we have

$$\begin{aligned}
&\max_{u\in\mathcal{U},w\in\mathcal{I}}\|\mathbf{w}_l-\mathbf{w}_l'\|_2\\
&\leq C_\phi\cdot\max_{u\in\mathcal{U},w\in\mathcal{I}}\left[\|\mathbf{w}_{l-1}\|_2\left(\|\boldsymbol{W}_1-\boldsymbol{W}_1'\|_2+\gamma b\,\|\boldsymbol{W}_2-\boldsymbol{W}_2'\|_2\right)+\gamma\,\|\mathbf{u}_{l-1}\|_2\,\|\boldsymbol{W}_1-\boldsymbol{W}_1'\|_2\right]\\
&+C_\phi\cdot\max_{u\in\mathcal{U},w\in\mathcal{I}}\left[\left\|\mathbf{w}_{l-1}-\mathbf{w}_{l-1}'\right\|_2\left(B_1+\gamma B_2 b\right)+\left\|\mathbf{u}_{l-1}-\mathbf{u}_{l-1}'\right\|_2\gamma\left(B_1+B_2 b\right)\right].
\end{aligned} \tag{17}$$

Combining Eq. (16) and Eq. (17), we have

$$\begin{aligned}
\Delta_l &= \max_{u\in\mathcal{U},w\in\mathcal{I}}\left(\|\mathbf{u}_l-\mathbf{u}_l'\|_2+\|\mathbf{w}_l-\mathbf{w}_l'\|_2\right)\\
&= C_\phi T_{l-1}\left[(\gamma+1)\,\|\boldsymbol{W}_1-\boldsymbol{W}_1'\|_2+\gamma b\,\|\boldsymbol{W}_2-\boldsymbol{W}_2'\|_2\right]+C_\phi\Delta_{l-1}\left[(\gamma+1)B_1+2\gamma b B_2\right]\\
&\leq C_\phi T_{l-1}\left[(\gamma+1)\,\|\boldsymbol{W}_1-\boldsymbol{W}_1'\|_2+\gamma b\,\|\boldsymbol{W}_2-\boldsymbol{W}_2'\|_2\right]+2\mathcal{C}\Delta_{l-1}.
\end{aligned}$$

The last inequality is due to $C_\phi\left[(\gamma+1)B_1+2\gamma b B_2\right] \leq 2C_\phi\left[(\gamma+1)B_1+\gamma b B_2\right] = 2\mathcal{C}$. According to the norm bound for node representations completed in Eq. (9) and expanding the recursion, we can get the following result,

$$\begin{aligned}
\Delta_l &= \max_{u\in\mathcal{U},w\in\mathcal{I}}\left(\|\mathbf{u}_l-\mathbf{u}_l'\|_2+\|\mathbf{w}_l-\mathbf{w}_l'\|_2\right)\\
&\leq C_\phi C^{l-1}(B_u+B_i)\left[(\gamma+1)\,\|\boldsymbol{W}_1-\boldsymbol{W}_1'\|_2+\gamma b\,\|\boldsymbol{W}_2-\boldsymbol{W}_2'\|_2\right]+2\mathcal{C}\Delta_{l-1}\\
&\leq C_\phi C^{l-1}(B_u+B_i)\left[(\gamma+1)\,\|\boldsymbol{W}_1-\boldsymbol{W}_1'\|_2+\gamma b\,\|\boldsymbol{W}_2-\boldsymbol{W}_2'\|_2\right]\frac{(2\mathcal{C})^l-1}{2\mathcal{C}-1}.
\end{aligned} \tag{18}$$

Step 3: Max change of recommendation. For any user-item (u, i), we first perform that the difference between preference scores predicted by f and f' can be bounded by w and w'. Then we derive the covering number bound for the function class $\mathcal{F}$ with specific radius.

$$\Lambda_L = \max_{(u,i)\in S} |f(u,i) - f'(u,i)| = \max_{(u,i)\in S} \left| \boldsymbol{e}_u^T \boldsymbol{e}_i - \boldsymbol{e}_u'^T \boldsymbol{e}_i' \right|.$$

There exists three types of integration operations to obtain the final node representations.

(Case 1) Final layer:

$$\begin{aligned}
\Lambda_L &\leq \max_{u\in\mathcal{U}, w\in\mathcal{I}} \left| \mathbf{u}_L^T \mathbf{w}_L - \mathbf{u}_L'^T \mathbf{w}_L' \right| \\
&\leq \max_{u\in\mathcal{U}, w\in\mathcal{I}} \left(\left| \mathbf{u}_L^T \mathbf{w}_L - \mathbf{u}_L^T \mathbf{w}_L' \right| + \left| \mathbf{u}_L^T \mathbf{w}_L' - \mathbf{u}_L'^T \mathbf{w}_L' \right| \right) \\
&\leq \max_{u\in\mathcal{U}, w\in\mathcal{I}} \left(\|\mathbf{u}_L\|_2 \|\mathbf{w}_L - \mathbf{w}_L'\|_2 + \|\mathbf{u}_L - \mathbf{u}_L'\|_2 \|\mathbf{w}_L'\|_2 \right). \qquad (19)
\end{aligned}$$

Similarly,

$$\Lambda_L \leq \max_{u\in\mathcal{U}, w\in\mathcal{I}} \left(\|\mathbf{u}_L'\|_2 \|\mathbf{w}_L - \mathbf{w}_L'\|_2 + \|\mathbf{u}_L - \mathbf{u}_L'\|_2 \|\mathbf{w}_L\|_2 \right). \qquad (20)$$

Combining Eq. (19) and Eq. (20), we have

$$\begin{aligned}
\Lambda_L &\leq \max_{u\in\mathcal{U}, w\in\mathcal{I}} \left[\left(\|\mathbf{u}_L\|_2 + \|\mathbf{u}_L'\|_2 \right) \|\mathbf{w}_L - \mathbf{w}_L'\|_2 + \|\mathbf{u}_L - \mathbf{u}_L'\|_2 \left(\|\mathbf{w}_L\|_2 + \|\mathbf{w}_L'\|_2 \right) \right] \\
&\leq \frac{1}{2} \max_{u\in\mathcal{U}, w\in\mathcal{I}} \left(\|\mathbf{u}_L\|_2 + \|\mathbf{u}_L'\|_2 + \|\mathbf{w}_L\|_2 + \|\mathbf{w}_L'\|_2 \right) \left(\|\mathbf{u}_L - \mathbf{u}_L'\|_2 + \|\mathbf{w}_L - \mathbf{w}_L'\|_2 \right) \\
&\leq T_L \Delta_L. \qquad (21)
\end{aligned}$$

(Case 2) Linear combination:

$$\begin{aligned}
\Lambda_L &\leq \max_{u\in\mathcal{U}, w\in\mathcal{I}} \left| \left(\sum_{l=0}^{L} \alpha^l \mathbf{u}_l \right)^T \left(\sum_{l=0}^{L} \alpha^l \mathbf{w}_l \right) - \left(\sum_{l=0}^{L} \alpha^l \mathbf{u}_l' \right)^T \left(\sum_{l=0}^{L} \alpha^l \mathbf{w}_l' \right) \right| \\
&\leq \frac{1}{2} \max_{u\in\mathcal{U}, w\in\mathcal{I}} \left(\left\| \sum_{l=0}^{L} \alpha^l \mathbf{u}_l \right\|_2 + \left\| \sum_{l=0}^{L} \alpha^l \mathbf{u}_l' \right\|_2 + \left\| \sum_{l=0}^{L} \alpha^l \mathbf{w}_l \right\|_2 + \left\| \sum_{l=0}^{L} \alpha^l \mathbf{w}_l' \right\|_2 \right) \\
&\times \left(\left\| \sum_{l=0}^{L} \alpha^l \left(\mathbf{u}_l - \mathbf{u}_l' \right) \right\|_2 + \left\| \sum_{l=0}^{L} \alpha^l \left(\mathbf{w}_l - \mathbf{w}_l' \right) \right\|_2 \right) \\
&\leq \frac{1}{2} \max_{u\in\mathcal{U}, w\in\mathcal{I}} \left(\sum_{l=0}^{L} |\alpha^l| \left(\|\mathbf{u}_l\|_2 + \|\mathbf{u}_l'\|_2 + \|\mathbf{w}_l\|_2 + \|\mathbf{w}_l'\|_2 \right) \right) \left(\sum_{l=0}^{L} |\alpha^l| \left(\|\mathbf{u}_l - \mathbf{u}_l'\|_2 + \|\mathbf{w}_l - \mathbf{w}_l'\|_2 \right) \right) \\
&\leq \left(\sum_{l=0}^{L} |\alpha^l| T_l \right) \left(\sum_{l=0}^{L} |\alpha^l| \Delta_l \right).
\end{aligned}$$

(Case 3) Concatenation:

$$\begin{aligned}
\Lambda_L &\leq \max_{u\in\mathcal{U}, w\in\mathcal{I}} \left| \left(\mathbf{u}_0 \oplus \cdots \oplus \mathbf{u}_L \right)^T \left(\mathbf{w}_0 \oplus \cdots \oplus \mathbf{w}_L \right) - \left(\mathbf{u}_0' \oplus \cdots \oplus \mathbf{u}_L' \right)^T \left(\mathbf{w}_0' \oplus \cdots \oplus \mathbf{w}_L' \right) \right| \\
&= \max_{u\in\mathcal{U}, w\in\mathcal{I}} \left| \sum_{l=0}^{L} \mathbf{u}_l^T \mathbf{w}_l - \sum_{l=0}^{L} \mathbf{u}_l'^T \mathbf{w}_l' \right| \\
&\leq \max_{u\in\mathcal{U}, w\in\mathcal{I}} \sum_{l=0}^{L} \left| \mathbf{u}_l^T \mathbf{w}_l - \mathbf{u}_l'^T \mathbf{w}_l' \right| \\
&\leq \sum_{l=0}^{L} T_l \Delta_l.
\end{aligned}$$

For convenience, we only consider the outputs of the last layer as the final representations. Substituting Eq. (9) and Eq. (18) in to Eq. (21), we will derive the following change bound of recommendations predicted by two GCN-based models with different parameters.

$$\begin{aligned}\Lambda_L &= \max_{u\in\mathcal{U},w\in\mathcal{I}} |f(u,i) - f'(u,i)| \\ &\leq C_\phi C^{2L-1}(B_u + B_i)^2 \left[(\gamma+1)\left\|\boldsymbol{W}_1 - \boldsymbol{W}_1'\right\|_2 + \gamma b \left\|\boldsymbol{W}_2 - \boldsymbol{W}_2'\right\|_2\right] \frac{(2\mathcal{C})^L - 1}{2\mathcal{C}-1}.\end{aligned}$$

Step 4: Covering number bound of the function class $\mathcal{F}$ over a sample set S with size n. Let $S = \{z_i\}_{i=1}^n$, we denote a vector $\boldsymbol{f}_S = (f(z_1), \ldots, f(z_n))$ as the recommendations predicted by the model $f \in \mathcal{F}$ for all sample in sample set S. Let $M = C_\phi \mathcal{C}^{2L-1}(B_u + B_i)^2 \frac{(2\mathcal{C})^L - 1}{2\mathcal{C}-1}$, we have

$$\|\boldsymbol{f}_S\|_\infty = \Lambda_L \leq M\left[(\gamma+1)\left\|\boldsymbol{W}_1 - \boldsymbol{W}_1'\right\|_2 + \gamma b \left\|\boldsymbol{W}_2 - \boldsymbol{W}_2'\right\|_2\right].$$

Since for any matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_2 \leq \|\boldsymbol{A}\|_F$, we can find a $\frac{\epsilon}{2(\gamma+1)M}-$cover for $\boldsymbol{W}_1$ and a $\frac{\epsilon}{2\gamma bM}-$cover for $\boldsymbol{W}_2$ to bound

$$\begin{aligned}\|\boldsymbol{W}_1 - \boldsymbol{W}_1^{'}\|_2 \leq \|\boldsymbol{W}_1 - \boldsymbol{W}_1^{'}\|_F \leq \frac{\epsilon}{(\gamma+1)M}, \\ \|\boldsymbol{W}_2 - \boldsymbol{W}_2^{'}\|_2 \leq \|\boldsymbol{W}_2 - \boldsymbol{W}_2^{'}\|_F \leq \frac{\epsilon}{\gamma bM}.\end{aligned}$$

According to Lemma B.1, we have

$$\begin{aligned}\mathcal{N}(\boldsymbol{W}_1, \frac{\epsilon}{2(\gamma+1)M}, \|\cdot\|_F) \leq \left(1 + \frac{4(\gamma+1)MB_1\sqrt{d}}{\epsilon}\right)^{d^2}, \\ \mathcal{N}(\boldsymbol{W}_2, \frac{\epsilon}{2\gamma bM}, \|\cdot\|_F) \leq \left(1 + \frac{4\gamma bMB_2\sqrt{d}}{\epsilon}\right)^{d^2}.\end{aligned}$$

Thus, we have a way of obtaining the covering number of the function class $\mathcal{F}$ to bound $\Lambda_L \leq 2\epsilon$,

$$\log\mathcal{N}\left(\mathcal{F}_{|S}, \epsilon, \|\cdot\|_\infty\right) \leq d^2 \log\left(1 + \frac{4(\gamma+1)MB_1\sqrt{d}}{\epsilon}\right)\left(1 + \frac{4\gamma bMB_2\sqrt{d}}{\epsilon}\right).$$

Moreover, when $\epsilon \leq 4M\sqrt{d}\max\{(\gamma+1)B_1, \gamma bB_2\}$,

$$\begin{aligned}\log\mathcal{N}\left(\mathcal{F}_{|S}, \epsilon, \|\cdot\|_\infty\right) &\leq d^2 \log\frac{64M^2 d\gamma(\gamma+1)B_1B_2b}{\epsilon^2} \\ &\leq 2d^2 \log\frac{8M\sqrt{dB_1B_2b}(\gamma+1)}{\epsilon}.\end{aligned}$$

The proof is complete. □

C Proof of Lemma 2

We first present the definition of the inductive empirical Rademacher complexity and some lemmas that we will use in the later proofs.

Definition C.1 (Inductive empirical Rademacher complexity)**.** *Let $\mathcal{D}$ be a probability distribution over $\mathcal{X}$. Suppose that the examples $S = \{x_i\}_{i=1}^m$ are sampled independently from $\mathcal{X}$ according to $\mathcal{D}$. Let $\mathcal{F}$ be a family of functions mapping $\mathcal{X}$ to $\mathbb{R}$. The empirical Rademacher complexity of $\mathcal{F}$ with respect to the samples S is defined as*

$$\hat{\mathfrak{R}}_S(\mathcal{F}) \triangleq \mathbb{E}_{\boldsymbol{\sigma}}\left\{\sup_{\boldsymbol{f}\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i)\right\},$$

where $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_m)^T$, with σ_i is independent uniform random variables taking values in $\{-1, +1\}$.

Lemma C.1 ([48]). *For any bounded loss function $l : \mathbb{R} \to [-B, B]$, given a training set $D \sim \mathcal{D}^m$, with probability of at least $1 - \delta$, for any function g in a class $\mathcal{G}$,*

$$\mathcal{L}_G(g) \leq \hat{\mathcal{L}}(g) + 2\hat{\mathfrak{R}}_S(l \circ \mathcal{G}) + 4B\sqrt{\frac{2\log 4/\delta}{m}}.$$

Lemma C.2 (Contraction Lemma [29]). *Let $\Phi : \mathbb{R} \to \mathbb{R}$ be an $L - lipschitz$ function, then for a real-valued function set $\mathcal{G}$,*

$$\mathfrak{R}_m(\Phi \circ \mathcal{G}) \leq L\mathfrak{R}_m(\mathcal{G}).$$

Lemma C.3 (Extension of [35]). *Let $\mathcal{F}$ be a real-valued function class taking values in $[-e, e]$, and assume that $\mathbf{0} \in \mathcal{F}$. Then the empirical Rademacher complexity of $\mathcal{F}$ can be bounded as*

$$\hat{\mathfrak{R}}_S(\mathcal{F}) \leq \inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{n}} + \frac{12}{n}\int_{\alpha}^{2e\sqrt{n}}\sqrt{\log \mathcal{N}\left(\mathcal{F}_{|S}, \epsilon, \|\cdot\|_2\right)}d\epsilon\right).$$

Then we provide the proof of Lemma 2. The following summarizes the proof strategy and the roles of the above lemmas.

- Derive a Rademacher complexity-based generalization error bound using Lemma C.1 and Lemma C.2.
- Yield the final result with respect to discrete covering number via chaining using Lemma C.3.

Lemma 2. *Let $\mathcal{F}$ be a real-valued function class taking values in $[-e, e]$, and assume that $\mathbf{0} \in \mathcal{F}$. Under assumptions, for any function f in a class $\mathcal{F}$, with probability of at least $1 - \delta$ over an i.i.d. size-m training set, we have*

$$\mathcal{L}(f) \leq \hat{\mathcal{L}}_m(f) + 4C_l \inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{m}} + \frac{12}{m}\int_{\alpha}^{2e\sqrt{m}}\sqrt{\log \mathcal{N}\left(\mathcal{F}_{|S_m}, \epsilon, \|\cdot\|_2\right)}d\epsilon\right) + 4B\sqrt{\frac{2\log 4/\delta}{m}}.$$

Proof. Let $l \circ \mathcal{H} = \{(z, z', y) \to l(h) | h \in \mathcal{H}\}$ and $\mathcal{H} = \{(z, z', y) \to y(f(z) - f(z')) | f \in \mathcal{F}\}$. We denote $S_1 = \{(z_i, y_i)\}_{i=1}^m$ as the sample derived from S_m by keeping only the first element of each pair and $S_2 = \{(z_i', y_i)\}_{i=1}^m$ the one obtained by keeping only the second element. Using Lemma C.1, Lemma C.2 and the assumption about loss function, we have

$$\begin{aligned}\mathcal{L}(f) &\leq \hat{\mathcal{L}}_m(f) + 2\hat{\mathfrak{R}}_{S_m}(l \circ \mathcal{H}) + 4B\sqrt{\frac{2\log 4/\delta}{m}} \\ &\leq \hat{\mathcal{L}}_m(f) + 2C_l\hat{\mathfrak{R}}_{S_m}(\mathcal{H}) + 4B\sqrt{\frac{2\log 4/\delta}{m}}.\end{aligned}$$

Here, since $\sigma_i y_i$ and σ_i have same distribution, $\hat{\mathfrak{R}}_m(\mathcal{H})$ can be bounded as follows,

$$\begin{aligned}\hat{\mathfrak{R}}_{S_m}(\mathcal{H}) &= \mathbb{E}_{\boldsymbol{\sigma}}\left\{\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m \sigma_i y_i(f(z_i) - f(z_i'))\right\} \\ &= \mathbb{E}_{\boldsymbol{\sigma}}\left\{\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m \sigma_i(f(z_i) - f(z_i'))\right\} \\ &\leq \mathbb{E}_{\boldsymbol{\sigma}}\left\{\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m \sigma_i f(z_i) + \sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m \sigma_i f(z_i')\right\} \\ &= \hat{\mathfrak{R}}_{S_1}(\mathcal{F}) + \hat{\mathfrak{R}}_{S_2}(\mathcal{F}).\end{aligned}$$

Considering the distribution $\mathcal{D}$ as symmetric, we have $\hat{\mathfrak{R}}_{S_1}(\mathcal{F}) = \hat{\mathfrak{R}}_{S_2}(\mathcal{F})$, and then $\hat{\mathfrak{R}}_{S_m}(\mathcal{H}) \leq 2\hat{\mathfrak{R}}_{S_1}(\mathcal{F})$. Together with Lemma C.3, we can derive the following generalization error bounded by covering number.

$$\begin{aligned}\mathcal{L}(f) &\leq \hat{\mathcal{L}}_m(f) + 4C_l\hat{\mathfrak{R}}_{S_1} + 4B\sqrt{\frac{2\log 4/\delta}{m}} \\ &\leq \hat{\mathcal{L}}_m(f) + 4C_l \inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{m}} + \frac{12}{m}\int_{\alpha}^{2e\sqrt{m}}\sqrt{\log \mathcal{N}\left(\mathcal{F}_{|S_1}, \epsilon, \|\cdot\|_2\right)}d\epsilon\right) + 4B\sqrt{\frac{2\log 4/\delta}{m}}.\end{aligned}$$

The proof is complete. □

D Proof of Lemma 3

Similarly, we first present the definition of the transductive Rademacher complexity (TRC) and some TRC version of lemmas that we will use in the later proofs.

Definition D.1 (Transductive Rademacher complexity)**.** *Let $\mathcal{F} \subseteq \mathbb{R}^{m+u}$ and $p \in [0, 1/2]$. Let $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_{m+u})^T$ be a vector of i.i.d. random variables such that*

$$\sigma_i \triangleq \begin{cases} 1 & \textit{with probability} \quad p; \\ -1 & \textit{with probability} \quad p; \\ 0 & \textit{with probability} \quad 1-2p. \end{cases} \tag{22}$$

The transductive Rademacher complexity with parameter p is

$$\Re_{m+u}(\mathcal{F}, p) \triangleq \left(\frac{1}{m} + \frac{1}{u}\right) \cdot \mathbb{E}_{\boldsymbol{\sigma}} \left\{ \sup_{\boldsymbol{f} \in \mathcal{F}} \boldsymbol{\sigma}^T \boldsymbol{f} \right\}.$$

Lemma D.1 ([33])**.** *Let $B_1 \leq 0$, $B_2 \geq 0$ and $\mathcal{V}$ be a (possibly infinite) set of real-valued vectors in $[B_1, B_2]^{m+u}$. Let $B = B_2 - B_1$ and $B_{\max} = \max(|B_1|, |B_2|)$. Let $Q_1 = \frac{1}{u} + \frac{1}{m}$, $Q_2 = \frac{m+u}{(m+u-1/2)(1-1/2(\max(m,u)))}$ and $c_0 = \sqrt{\frac{32 \log(4e)}{3}} < 5.05$. Then with probability of at least $1-\delta$ over random partitions of S, for all $\boldsymbol{v} \in \mathcal{V}$,*

$$\mathcal{L}_u(f) \leq \hat{\mathcal{L}}_m(f) + \Re_{m+u}(\mathcal{F}, \frac{mu}{(m+u)^2}) + B_{\max} c_0 Q_1 \sqrt{\min(m,u)} + B\sqrt{\frac{Q_1 Q_2}{2} \ln \frac{1}{\delta}}.$$

Lemma D.2 (Contraction lemma for TRC [33])**.** *Let $\Phi : \mathbb{R} \to \mathbb{R}$ be an $L-lipschitz$ function, Then for a real-valued function set $\mathcal{G}$,*

$$\Re_m(\Phi \circ \mathcal{G}, p) \leq L \Re_m(\mathcal{G}, p).$$

Lemma D.3 (Massart's lemma for TRC)**.** *Let $\mathcal{A} \subseteq \mathbb{R}^m$ be a finite set of vectors, $r = \max_{\mathbf{x} \in \mathcal{A}} \|\mathbf{x}\|_2$, then*

$$\frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{x} \in \mathcal{A}} \sum_{i=1}^{m} \sigma_i x_i \right] \leq \frac{r\sqrt{2 \log |\mathcal{A}|}}{m},$$

where σ_is are transductive Rademacher random variables defined in Eq. (22) and $\boldsymbol{x} = \{x_1, \ldots, x_m\}$.

Lemma D.4. *Let $\mathcal{F}$ be a real-valued function class taking values in $[-e, e]$, and assume that $\mathbf{0} \in \mathcal{F}$. Then the transductive Rademacher complexity of $\mathcal{F}$ can be bounded as*

$$\Re_S(\mathcal{F}, p) \leq \inf_{\alpha > 0} \left(\frac{4\alpha}{\sqrt{n}} + \frac{12}{n} \int_{\alpha}^{2e\sqrt{n}} \sqrt{\log \mathcal{N}\left(\mathcal{F}_{|S}, \epsilon, \|\cdot\|_2\right)} d\epsilon \right).$$

The proofs of Lemma D.3 and Lemma D.4 follow the proofs for inductive Rademacher complexity, and will be provided in Section D.1 and Section D.2.

Lemma 3. *Let $\mathcal{F}$ be a real-valued function class taking values in $[-e, e]$, and assume that $\mathbf{0} \in \mathcal{F}$. Let $Q_1 = \frac{1}{u} + \frac{1}{m}, Q_2 = \frac{m+u}{(m+u-1/2)(1-1/2(\max(m,u)))}$ and $c_0 < 5.05$. Under assumptions, for any function f in a class $\mathcal{F}$, with probability of at least $1-\delta$ over random partitions of S, we have*

$$\begin{aligned} \mathcal{L}_u(f) \leq & \hat{\mathcal{L}}_m(f) + 2C_l \inf_{\alpha > 0} \left(\frac{4\alpha}{\sqrt{m+u}} + \frac{12}{m+u} \int_{\alpha}^{2e\sqrt{m+u}} \sqrt{\log \mathcal{N}\left(\mathcal{F}_{|S}, \epsilon, \|\cdot\|_2\right)} d\epsilon \right) \\ & + B c_0 Q_1 \sqrt{\min(m,u)} + 2B\sqrt{\frac{Q_2}{2} Q_1 \ln \frac{1}{\delta}}. \end{aligned}$$

Proof. Let $l \circ \mathcal{H} = \{(z, z', y) \to l(h) | h \in \mathcal{H}\}$ and $\mathcal{H} = \{(z, z', y) \to y(f(z) - f(z')) | f \in \mathcal{F}\}$. We denote $S_1 = \{(z_i, y_i)\}_{i=1}^{m+u}$ as the sample derived from S by keeping only the first element of

each pair and $S_2 = \{(z_i', y_i)\}_{i=1}^{m+u}$ the one obtained by keeping only the second element. Using Lemma D.1, Lemma C.2 and the assumption about loss function, we have

$$\begin{aligned}\mathcal{L}_u(f) &\leq \hat{\mathcal{L}}_m(f) + 2\Re_S(l \circ \mathcal{H}) + Bc_0Q_1\sqrt{\min(m,u)} + 2B\sqrt{\frac{Q_2}{2}Q_1\ln\frac{1}{\delta}} \\ &\leq \hat{\mathcal{L}}_m(f) + 2C_l\Re_S(\mathcal{H}) + Bc_0Q_1\sqrt{\min(m,u)} + 2B\sqrt{\frac{Q_2}{2}Q_1\ln\frac{1}{\delta}}.\end{aligned}$$

Here, since $\sigma_i y_i$ and σ_i have same distribution, $\Re_S(\mathcal{H})$ can be bounded as follows,

$$\begin{aligned}\Re_S(\mathcal{H}) &= \mathbb{E}_{\boldsymbol{\sigma}}\left\{\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i y_i(f(z_i) - f(z_i'))\right\} \\ &= \mathbb{E}_{\boldsymbol{\sigma}}\left\{\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i (f(z_i) - f(z_i'))\right\} \\ &\leq \mathbb{E}_{\boldsymbol{\sigma}}\left\{\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i) + \sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i')\right\} \\ &= \Re_{S_1}(\mathcal{F}) + \Re_{S_2}(\mathcal{F}).\end{aligned}$$

Considering the distribution $\mathcal{D}$ as symmetric, we have $\Re_{S_1}(\mathcal{F}) = \Re_{S_2}(\mathcal{F})$, and then $\Re_S(\mathcal{H}) \leq 2\Re_{S_1}(\mathcal{F})$. Together with Lemma C.3, we can derive the following generalization error bounded by Covering number.

$$\begin{aligned}\mathcal{L}(f) &\leq \hat{\mathcal{L}}_m(f) + 4C_l\Re_{S_1}(\mathcal{F}) + Bc_0Q_1\sqrt{\min(m,u)} + 2B\sqrt{\frac{Q_2}{2}Q_1\ln\frac{1}{\delta}} \\ &\leq \hat{\mathcal{L}}_m(f) + 4C_l\inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{m}} + \frac{12}{m}\int_{\alpha}^{2e\sqrt{m}}\sqrt{\log\mathcal{N}\left(\mathcal{F}_{|S_1}, \epsilon, \|\cdot\|_2\right)}d\epsilon\right) \\ &+ Bc_0Q_1\sqrt{\min(m,u)} + 2B\sqrt{\frac{Q_2}{2}Q_1\ln\frac{1}{\delta}}.\end{aligned}$$

The proof is complete. □

D.1 Proof of Lemma D.3

Proof. The proof follows similar to the one for inductive Rademacher random variables [29].

For $\forall t > 0$, using Jensen's inequality, we have

$$\begin{aligned}\exp\left(t\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{x}\in\mathcal{A}}\sum_{i=1}^{m}\sigma_i x_i\right]\right) &\leq \mathbb{E}_{\boldsymbol{\sigma}}\left[\exp\left(t\sup_{\boldsymbol{x}\in\mathcal{A}}\sum_{i=1}^{m}\sigma_i x_i\right)\right] \\ &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{x}\in\mathcal{A}}\exp\left(t\sum_{i=1}^{m}\sigma_i x_i\right)\right] \\ &\leq \sum_{\boldsymbol{x}\in\mathcal{A}}\mathbb{E}_{\boldsymbol{\sigma}}\left[\exp\left(t\sum_{i=1}^{m}\sigma_i x_i\right)\right] \\ &= \sum_{\boldsymbol{x}\in\mathcal{A}}\mathbb{E}_{\boldsymbol{\sigma}}\left[\prod_{i=1}^{m}\exp\left(t\sigma_i x_i\right)\right] \\ &= \sum_{\boldsymbol{x}\in\mathcal{A}}\prod_{i=1}^{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\exp\left(t\sigma_i x_i\right)\right].\end{aligned}$$

Since $\mathbb{E}_{\boldsymbol{\sigma}}[\sigma_i x_i] = 0$ and $-|x_i| \leq \sigma_i x_i \leq |x_i|$, using Hoeffding's lemma, we have

$$\begin{aligned}
\exp\left(t\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{x}\in\mathcal{A}}\sum_{i=1}^{m}\sigma_i x_i\right]\right) &\leq \sum_{\boldsymbol{x}\in\mathcal{A}}\prod_{i=1}^{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\exp\left(t\sigma_i x_i\right)\right] \\
&\leq \sum_{\boldsymbol{x}\in\mathcal{A}}\prod_{i=1}^{m}\exp\left(\frac{t^2(2x_i)^2}{8}\right) \\
&= \sum_{\boldsymbol{x}\in\mathcal{A}}\exp\left(\frac{t^2}{2}\sum_{i=1}^{m}x_i^2\right) \\
&\leq \sum_{\boldsymbol{x}\in\mathcal{A}}\exp\left(\frac{t^2 r^2}{2}\right) \\
&\leq |\mathcal{A}|\exp\left(\frac{t^2 r^2}{2}\right).
\end{aligned}$$

Therefore,

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{x}\in\mathcal{A}}\sum_{i=1}^{m}\sigma_i x_i\right] \leq \frac{\log|\mathcal{A}|}{t} + \frac{tr^2}{2}.$$

Set $t = \frac{\sqrt{2\log|\mathcal{A}|}}{t}$, we can complete the proof.

$$\frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{x}\in\mathcal{A}}\sum_{i=1}^{m}\sigma_i x_i\right] \leq \frac{r\sqrt{2\log|\mathcal{A}|}}{m}.$$

The proof is complete. □

D.2 Proof of Lemma D.4

Proof. The proof follows similar to the one for inductive empirical Rademacher complexity [35].

Let N be arbitrary and let $\varepsilon_i = 2e\sqrt{n}2^{-(i-1)}$ for each $i \in [N]$. We denote the outputs of function $f \in \mathcal{F}$ over training set $S = \{x_1, \ldots, x_n\}$ as a vector $\boldsymbol{f}_S = (f(x_1), \ldots, f(x_n))$. For each i let V_i denote the cover achieving $\mathcal{N}(\mathcal{F}_{|S}, \varepsilon_i, \|\cdot\|_2)$, so that

$$\forall f \in \mathcal{F}, \exists \boldsymbol{v} \in V_i, \|\boldsymbol{f}_S - \boldsymbol{v}\|_2 \leq \varepsilon_i,$$

and $|V_i| = \mathcal{N}(\mathcal{F}_{|S}, \varepsilon_i, \|\cdot\|_2)$. For a fixed $f \in \mathcal{F}$, let $\boldsymbol{v}^i[f]$ denote the nearest element in V_i. Then we have

$$\begin{aligned}
\Re_n\left(\mathcal{F}, p\right) &= \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{\boldsymbol{f}\in\mathcal{F}}\sum_{j=1}^{n}\sigma_j f(x_j) \\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{\boldsymbol{f}\in\mathcal{F}}\left[\sum_{j=1}^{n}\sigma_j\left(f(x_j) - v_j^N[f]\right) + \sum_{i=1}^{N-1}\sum_{j=1}^{n}\sigma_j\left(v_j^i[f] - v_j^{i+1}[f]\right) - \sum_{j=1}^{n}\sigma_j v_j^1[f]\right] \\
&\leq \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{\boldsymbol{f}\in\mathcal{F}}\left[\sum_{j=1}^{n}\sigma_j\left(f(x_j) - v_j^N[f]\right)\right] + \frac{1}{n}\sum_{i=1}^{N-1}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{\boldsymbol{f}\in\mathcal{F}}\left[\sum_{j=1}^{n}\sigma_j\left(v_j^i[f] - v_j^{i+1}[f]\right)\right] \\
&\quad + \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{\boldsymbol{f}\in\mathcal{F}}\left[\sum_{j=1}^{n}\sigma_j v_j^1[f]\right]. \qquad (23)
\end{aligned}$$

The last inequality is due to $\sigma_i \in \{-1, +1, 0\}$. For the first term in Eq. (23), using Cauchy-Schwarz inequality, since $(\sigma_j)^2 \leq 1$, we have

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{\boldsymbol{f}\in\mathcal{F}}\left[\sum_{j=1}^{n}\sigma_j\left(f(x_j) - v_j^N[f]\right)\right] \leq \frac{1}{n}\sqrt{\mathbb{E}_{\boldsymbol{\sigma}}\sum_{j=1}^{n}(\sigma_j)^2}\sqrt{\sup_{\boldsymbol{f}\in\mathcal{F}}\left(f(x_j) - v_j^N[f]\right)^2} \leq \frac{\varepsilon_N}{\sqrt{n}}.$$

For the second term in Eq. (23), let $W_i = \{v_j^i[f] - v_j^{i+1}[f] | f \in \mathcal{F}\}$, then we have

$$\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{f} \in \mathcal{F}} \left[\sum_{j=1}^{n} \sigma_j \left(v_j^i[f] - v_j^{i+1}[f] \right) \right] \leq \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{w} \in W_i} \left[\sum_{j=1}^{n} \sigma_j w_t \right]. \tag{24}$$

Since $\varepsilon_{i+1} \leq \varepsilon_i$, we can get $|W_i| \leq |V_i||V_{i+1}| \leq |V_{i+1}|^2$, which implies W_i is a finite set. Using Lemma D.3, the following inequality holds,

$$\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{w} \in W_i} \left[\sum_{j=1}^{n} \sigma_j w_t \right] \leq \sup_{\boldsymbol{w} \in W_i} \|\boldsymbol{w}\|_2 \sqrt{2 \log |W_i|} \leq 2 \sup_{\boldsymbol{w} \in W_i} \|\boldsymbol{w}\|_2 \sqrt{\log |V_{i+1}|}. \tag{25}$$

And furthermore

$$\begin{aligned}
\sup_{\boldsymbol{w} \in W_i} \|\boldsymbol{w}\|_2 &= \sup_{f \in \mathcal{F}} \left\| \boldsymbol{v}^i[f] - \boldsymbol{v}^{i+1}[f] \right\|_2 \\
&\leq \sup_{f \in \mathcal{F}} \left\| \boldsymbol{v}^i[f] - \boldsymbol{f}_S \right\|_2 + \sup_{f \in \mathcal{F}} \left\| \boldsymbol{f}_S - \boldsymbol{v}^{i+1}[f] \right\|_2 \\
&\leq \varepsilon_i + \varepsilon_{i+1} \\
&= 3\varepsilon_{i+1}.
\end{aligned}$$

Combining with Eq. (24) and Eq. (25), we have

$$\frac{1}{n} \sum_{i=1}^{N-1} \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{f} \in \mathcal{F}} \left[\sum_{j=1}^{n} \sigma_j \left(v_j^i[f] - v_j^{i+1}[f] \right) \right] \leq \frac{6}{n} \sum_{i=1}^{N-1} \varepsilon_{i+1} \sqrt{\log |V_{i+1}|}.$$

For the third term in Eq. (23), if we set $V_1 = \{\mathbf{0}\}$, $\|\boldsymbol{f}_S - \mathbf{0}\|_2 \leq 2e\sqrt{n} = \varepsilon_1$, which is a cover achieving $\mathcal{N}(\mathcal{F}_{|S}, \varepsilon_1, \|\cdot\|_2)$. Then the following equation holds,

$$\frac{1}{n} \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{f} \in \mathcal{F}} \left[\sum_{j=1}^{n} \sigma_j v_j^1[f] \right] = 0.$$

Collecting all terms, we can get the final result.

$$\begin{aligned}
\mathfrak{R}_n(\mathcal{F}, p) &\leq \frac{\varepsilon_N}{\sqrt{n}} + \frac{6}{n} \sum_{i=1}^{N-1} \varepsilon_{i+1} \sqrt{\mathcal{N}(\mathcal{F}_{|S}, \varepsilon_{i+1}, \|\cdot\|_2)} \\
&\leq \frac{\varepsilon_N}{\sqrt{n}} + \frac{12}{n} \sum_{i=1}^{N-1} (\varepsilon_i - \varepsilon_{i+1}) \sqrt{\mathcal{N}(\mathcal{F}_{|S}, \varepsilon_i, \|\cdot\|_2)} \\
&\leq \frac{\varepsilon_N}{\sqrt{n}} + \frac{12}{n} \int_{\varepsilon_{N+1}}^{2e\sqrt{n}} \sqrt{\mathcal{N}(\mathcal{F}_{|S}, \varepsilon, \|\cdot\|_2)}.
\end{aligned}$$

Finally, select any $\alpha > 0$ and take N be the largest integer with $\varepsilon_{N+1} > \alpha$. Then $\varepsilon_N < 4\varepsilon_{N+2} < 4\alpha$, we have

$$\mathfrak{R}_n(\mathcal{F}, p) \leq \frac{4\alpha}{\sqrt{n}} + \frac{12}{n} \int_{\varepsilon_{N+1}}^{2e\sqrt{n}} \sqrt{\mathcal{N}(\mathcal{F}_{|S}, \varepsilon, \|\cdot\|_2)}.$$

The proof is complete. □

E Proof of Proposition 1

The main the strategy in this proof is to derive the integral over discrete covering number.

Proposition 1 (Generalization Bound)**.** *Under assumptions, for any function f in a class $\mathcal{F}$, in inductive learning, with probability of at least $1 - \delta$, we have,*

$$\begin{aligned}
\mathcal{L}(f) \leq & \hat{\mathcal{L}}_m(f) + \frac{24 C_l}{\sqrt{m}} \mathcal{C}^{2L} (B_u + B_i)^2 d \sqrt{2 \log \left(8mM(\gamma+1) \sqrt{d B_1 B_2 b} \right)} \\
& + \frac{16 C_l}{m} + 4B \sqrt{\frac{2 \log 4/\delta}{m}}.
\end{aligned}$$

Accordingly, in transductive learning, with probability of at least $1-\delta$*, we have,*

$$\begin{aligned}\mathcal{L}_u(f) \leq & \hat{\mathcal{L}}_m(f) + \frac{24C_l}{\sqrt{m+u}}\mathcal{C}^{2L}(B_u+B_i)^2 d\sqrt{2\log\left(8(m+u)M(\gamma+1)\sqrt{dB_1B_2b}\right)} \\ & + \frac{4C_l\sqrt{2mu}}{(m+u)^2} + Bc_0Q_1\sqrt{\min(m,u)} + 2B\sqrt{\frac{Q_1Q_2}{2}\ln\frac{1}{\delta}}.\end{aligned}$$

Proof. First, we need to derive the bound $[-e, e]$ for any $f \in \mathcal{F}$.

$$\max_{u\in\mathcal{U},w\in\mathcal{I}} \left|\mathbf{u}_L^T\mathbf{w}_L\right| \leq \max_{u\in\mathcal{U},w\in\mathcal{I}} \|\mathbf{u}_L\|_2\|\mathbf{w}_L\|_2 \leq \frac{1}{4}\max_{u\in\mathcal{U},w\in\mathcal{I}} \left(\|\mathbf{u}_L\|_2 + \|\mathbf{w}_L\|_2\right)^2 = \frac{1}{4}T_L^2.$$

Using Eq. (9), we have $e = \frac{1}{4}\mathcal{C}^{2L}(B_u+B_i)^2$. Together with Lemma 1 and Lemma 2, the generalization error can be bounded as follows.

$$\mathcal{L}(f) \leq \hat{\mathcal{L}}_m(f) + 4C_l \inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{m}} + \frac{12}{m}\int_{\alpha}^{2e\sqrt{m}} \sqrt{\log\mathcal{N}\left(\mathcal{F}_{|S_1}, \epsilon, \|\cdot\|_2\right)}d\epsilon\right) + 4B\sqrt{\frac{2\log 4/\delta}{m}}. \tag{26}$$

Since for any vector $\boldsymbol{a} \in \mathbb{R}^n$, $\|\boldsymbol{a}\|_2 \leq \sqrt{n}\|\boldsymbol{a}\|_\infty$, we have

$$\begin{aligned}& 4C_l \inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{m}} + \frac{12}{m}\int_{\alpha}^{2e\sqrt{m}} \sqrt{\log\mathcal{N}\left(\mathcal{F}_{|S_1}, \epsilon, \|\cdot\|_2\right)}d\epsilon\right) \\ & \leq 4C_l \inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{m}} + \frac{12d}{m}\int_{\alpha}^{2e\sqrt{m}} \sqrt{2\log\frac{8M(\gamma+1)\sqrt{mdB_1B_2b}}{\epsilon}}d\epsilon\right) \\ & \leq 4C_l \inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{m}} + \frac{24ed}{\sqrt{m}}\sqrt{2\log\frac{8M(\gamma+1)\sqrt{mdB_1B_2b}}{\alpha}}\right) \\ & = 4C_l \inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{m}} + \frac{6dC^{2L}(B_u+B_i)^2}{\sqrt{m}}\sqrt{2\log\frac{8M(\gamma+1)\sqrt{mdB_1B_2b}}{\alpha}}\right).\end{aligned} \tag{27}$$

Set $\alpha = \sqrt{\frac{1}{m}}$ and substitute Eq. (27) into Eq. (26), we have

$$\begin{aligned}\mathcal{L}(f) \leq & \hat{\mathcal{L}}_m(f) + \frac{24C_l}{\sqrt{m}}\mathcal{C}^{2L}(B_u+B_i)^2 d\sqrt{2\log\left(8mM(\gamma+1)\sqrt{dB_1B_2b}\right)} \\ & + \frac{16C_l}{m} + 4B\sqrt{\frac{2\log 4/\delta}{m}}.\end{aligned}$$

Similarly, we can derive the generalization error bound for transductive learning.

$$\begin{aligned}\mathcal{L}_u(f) \leq & \hat{\mathcal{L}}_m(f) + \frac{24C_l}{\sqrt{m+u}}\mathcal{C}^{2L}(B_u+B_i)^2 d\sqrt{2\log\left(8(m+u)M(\gamma+1)\sqrt{dB_1B_2b}\right)} \\ & + \frac{4C_l\sqrt{2mu}}{(m+u)^2} + Bc_0Q_1\sqrt{\min(m,u)} + 2B\sqrt{\frac{Q_1Q_2}{2}\ln\frac{1}{\delta}}.\end{aligned}$$

The proof is complete. □

F Proofs of Section 4

We consider a training set $S_m = \{(u_k, i_k, i'_k)\}_{k=1}^m$ with labels $\{y_k\}_{k=1}^m$, where $y_k = 1$ if user u_k prefers item i_k to i'_k, otherwise $y_k = 0$. Denote by D_u all users in training set and D_i all item pairs in training set. We assume that $(u, i, i') \in S_m$ for any $u \in D_u$ and $(i, i') \in D_i$. We further assume that

the training set S_m is symmetric, which implies if $(u, i, i') \in S_m$ with label y, then $(u, i', i) \in S_m$ with label $1 - y$. We also assume that the samplings of user u and item-pair (i, i') are independent.

Recall the Item Mixture strategy proposed in Section 4. Since the recommendation is invariant to the scaling of the embedding, so it suffices to consider the following definition. For a triplet (u, i, i') with label y_i, we arbitrarily sample another triplet (u, j, j') with label y_j.

F.1 Proof of Lemma 4

In this section, we prove by second-order Taylor expansion that the IMix loss approximates the standard loss plus a regularization term.

Lemma 4. *Consider the symmetric dataset S_m and denote $\hat{\Sigma} = \frac{1}{m}\sum_{k=1}^{m}(\boldsymbol{e}_{i_k} - \boldsymbol{e}_{i'_k})(\boldsymbol{e}_{i_k} - \boldsymbol{e}_{i'_k})^T$, the second-order approximation of IMix loss defined in Eq. (5) is given by*

$$\mathcal{L}_m^{mix} \approx \mathcal{L}_m^{std} + \mathbb{E}_\lambda\left[\frac{(1-\lambda)^2}{\lambda^2}\right] \cdot \frac{1}{2m}\sum\nolimits_{k=1}^{m}\left[\frac{e^{\eta_k}}{(1+e^{\eta_k})^2}\boldsymbol{e}_{u_k}^T\hat{\Sigma}\boldsymbol{e}_{u_k}\right].$$

where $\eta_k = \boldsymbol{e}_{u_k}^T(\boldsymbol{e}_{i_k} - \boldsymbol{e}_{i'_k})$.

Proof. The IMix loss function over training set S_m is defined as

$$\begin{aligned}\mathcal{L}_m^{IMix} &= \frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_{\lambda\sim D_\lambda}E_{(j,j')\sim D_i}\ell\left(\boldsymbol{e}_{u_k}^T(\widetilde{\boldsymbol{e}}_{i_k} - \widetilde{\boldsymbol{e}}_{i'_k}), y_k\right)\\ &= \frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_{\lambda\sim D_\lambda}E_{(j,j')\sim D_i}\left[\log\left(1+\exp\left(\boldsymbol{e}_{u_k}^T(\widetilde{\boldsymbol{e}}_{i_k} - \widetilde{\boldsymbol{e}}_{i'_k})\right)\right) - y_k\boldsymbol{e}_{u_k}^T(\widetilde{\boldsymbol{e}}_{i_k} - \widetilde{\boldsymbol{e}}_{i'_k})\right].\end{aligned} \tag{28}$$

Denote the randomness (of λ and (j, j')) by ξ, since the training set S_m is symmetric, we have

$$\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_\xi - y_k\boldsymbol{e}_{u_k}^T(\widetilde{\boldsymbol{e}}_{i_k} - \widetilde{\boldsymbol{e}}_{i'_k}) = \frac{1}{m}\sum_{k=1}^{m} -y_k\boldsymbol{e}_{u_k}^T(\boldsymbol{e}_{i_k} - \boldsymbol{e}_{i'_k}).$$

Further using the second-order Taylor expansion, we can approximate the first term in Eq. (28) as follows,

$$\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_\xi\left[\log\left(1+\exp\left(\boldsymbol{e}_{u_k}^T(\widetilde{\boldsymbol{e}}_{i_k} - \widetilde{\boldsymbol{e}}_{i'_k})\right)\right)\right] \overset{2nd-order\ approx.}{=} \frac{1}{m}\sum_{k=1}^{m}\log\left(1+e^{\eta_k}\right) + \mathcal{R}_1 + \mathcal{R}_2,$$

where $\eta_k = \boldsymbol{e}_{u_k}^T(\boldsymbol{e}_{i_k} - \boldsymbol{e}_{i'_k})$. Since the training set is symmetric, we have

$$\mathcal{R}_1 = \frac{1}{m}\sum_{k=1}^{m}\frac{1}{1+e^{-\eta_k}} \cdot \mathbb{E}_\xi\boldsymbol{e}_{u_k}^T\left[(\widetilde{\boldsymbol{e}}_{i_k} - \widetilde{\boldsymbol{e}}_{i'_k}) - (\boldsymbol{e}_{i_k} - \boldsymbol{e}_{i'_k})\right] = 0.$$

Based on the assumption that $(u, i, i') \in S_m$ for any $u \in D_u$ and $(i, i') \in D_i$, we can get

$$\begin{aligned}\mathcal{R}_2 &= \frac{1}{2m}\sum_{k=1}^{m}\frac{e^{\eta_k}}{(1+e^{\eta_k})^2} \cdot \mathbb{E}_\xi\boldsymbol{e}_{u_k}^T\left[(\widetilde{\boldsymbol{e}}_{i_k} - \widetilde{\boldsymbol{e}}_{i'_k}) - (\boldsymbol{e}_{i_k} - \boldsymbol{e}_{i'_k})\right]\left[(\widetilde{\boldsymbol{e}}_{i_k} - \widetilde{\boldsymbol{e}}_{i'_k}) - (\boldsymbol{e}_{i_k} - \boldsymbol{e}_{i'_k})\right]^T\boldsymbol{e}_{u_k}\\ &= \mathbb{E}_\lambda\left[(1-\lambda)^2\right] \cdot \frac{1}{2m}\sum_{k=1}^{m}\frac{e^{\eta_k}}{(1+e^{\eta_k})^2}\boldsymbol{e}_{u_k}^T\hat{\Sigma}\boldsymbol{e}_{u_k},\end{aligned}$$

Then we have the following second-order Taylor approximation of the IMix loss.

$$\mathcal{L}_m^{mix} \approx \mathcal{L}_m^{std} + \mathbb{E}_\lambda\left[(1-\lambda)^2\right] \cdot \frac{1}{2m}\sum\nolimits_{k=1}^{m}\left[\frac{e^{\eta_k}}{(1+e^{\eta_k})^2}\boldsymbol{e}_{u_k}^T\hat{\Sigma}\boldsymbol{e}_{u_k}\right].$$

The proof is complete. □

F.2 Proof of Remark 3

We first present a fundamental uniform convergence theory, then we establish the generalization bounds for via Rademacher complexity.

Lemma F.1 ([48]). *For any bounded loss function $l : \mathbb{R} \to [-B, B]$, with probability of at least $1 - \delta$, for any function g in a class $\mathcal{G}$,*

$$\mathcal{L}_G(g) \leq \hat{\mathcal{L}}(g) + 2\mathbb{E}_{S_m \sim \mathcal{D}^m} \hat{\mathfrak{R}}_m(l \circ \mathcal{G}) + B\sqrt{\frac{2\log 2/\delta}{m}}.$$

Remark 3. Let $\psi(u) = \frac{e^u}{(1+e^u)^2}$, we shed light upon the generalization bound by investigating the following function class:

$$\mathcal{F}_\tau^{mix} = \{\mathcal{F}, \text{such that } \mathbb{E}_{u_k, i_k, i'_k}\left[\frac{e^{\eta_k}}{(1+e^{\eta_k})^2} \boldsymbol{e}_{u_k}^T \Sigma \boldsymbol{e}_{u_k}\right] \leq \tau\}.$$

Assuming that the distribution of $(\boldsymbol{e}_i - \boldsymbol{e}_{i'})$ is $\rho-$retentive for some $\rho \in (0, 1/2]$, that is, if for any non-zero vector $\boldsymbol{e}_u$, $\left[\mathbb{E}_{(i,i')}\psi(\boldsymbol{e}_u^T(\boldsymbol{e}_i - \boldsymbol{e}_{i'}))\right]^2 \geq \rho \min\{1, \mathbb{E}_{(i,i')}\left(e_u^T(\boldsymbol{e}_i - \boldsymbol{e}_{i'})\right)^2\}$. For any $f \in \mathcal{F}_\tau^{mix}$, with probability of at least $1 - \delta$, we have,

$$\mathcal{L}(f) \leq \hat{\mathcal{L}}(f) + 2C_l\left(\max\{(\frac{\tau}{\rho})^{1/4}, (\frac{\tau}{\rho})^{1/2}\}\sqrt{\frac{rank(\Sigma)}{|D_i|}}\right) + B\sqrt{\frac{2\log 2/\delta}{m}}.$$

For the general condition, we focus on the function class as follows,

$$\mathcal{F}_\tau^{std} = \{\mathcal{F} | \mathbb{E}_{(u,i)}\left[\|\boldsymbol{e}_u\|_2^2 + \|\boldsymbol{e}_i\|_2^2\right] \leq \tau\}.$$

Then the generalization error bound is

$$\mathcal{L}(f) \leq \hat{\mathcal{L}}(f) + 2C_l\sqrt{\frac{\tau^2}{|D_i|}} + B\sqrt{\frac{2\log 2/\delta}{m}}.$$

Proof. Let $\widetilde{\boldsymbol{x}}_i = \Sigma^{\dagger/2}(\boldsymbol{e}_i - \boldsymbol{e}_{i'})$ and $\boldsymbol{v}_u = \Sigma^{1/2}\boldsymbol{e}_u$, we can bound the empirical Rademacher complexity as follows,

$$\begin{aligned}
\hat{\mathfrak{R}}_{S_m}(\mathcal{F}_\tau^{mix}) &= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathbb{E}\left[\frac{e^{\eta_k}}{(1+e^{\eta_k})^2}\boldsymbol{e}_{u_k}^T \Sigma \boldsymbol{e}_{u_k}\right] \leq \tau} \frac{1}{m}\sum_{k=1}^m \sigma_k \boldsymbol{e}_{u_k}^T(\boldsymbol{e}_{i_k} - \boldsymbol{e}_{i'_k}) \\
&= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathbb{E}\left[\frac{e^{\eta_k}}{(1+e^{\eta_k})^2}\boldsymbol{e}_{u_k}^T \Sigma \boldsymbol{e}_{u_k}\right] \leq \tau} \frac{1}{m}\sum_{k=1}^m \sigma_k \boldsymbol{v}_{u_k}^T \widetilde{\boldsymbol{x}}_{i_k} \\
&\leq \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathbb{E}\|\boldsymbol{v}_{u_k}^T\|_2^2 \leq (\frac{\tau}{\rho})^{1/2} \vee \frac{\tau}{\rho}} \frac{1}{m}\sum_{k=1}^m \sigma_k \boldsymbol{v}_{u_k}^T \widetilde{\boldsymbol{x}}_{i_k}.
\end{aligned}$$

The last inequality is due to the samplings of user and item-pair are independent.

According to the assumptions on training set and Jensen's inequality, we have

$$\begin{aligned}
\hat{\Re}_{S_m}(\mathcal{F}_\tau^{mix}) &\leq \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathbb{E}\|\boldsymbol{v}_{u_k}^T\|_2^2 \leq (\frac{\tau}{\rho})^{1/2} \vee \frac{\tau}{\rho}} \frac{1}{m} \sum_{k=1}^{m} \sigma_k \boldsymbol{v}_{u_k}^T \widetilde{\boldsymbol{x}}_{i_k} \\
&\leq \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathbb{E}\|\boldsymbol{v}_{u_k}^T\|_2 \leq (\frac{\tau}{\rho})^{1/4} \vee (\frac{\tau}{\rho})^{1/2}} \frac{1}{m} \sum_{k=1}^{m} \sigma_k \boldsymbol{v}_{u_k}^T \widetilde{\boldsymbol{x}}_{i_k} \\
&\leq \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \sum_{u \in D_u} \sup_{\mathbb{E}\|\boldsymbol{v}_u^T\|_2 \leq (\frac{\tau}{\rho})^{1/4} \vee (\frac{\tau}{\rho})^{1/2}} \sum_{(i,i') \in D_i} \sigma_{ui} \boldsymbol{v}_u^T \widetilde{\boldsymbol{x}}_i \\
&\leq \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \sum_{u \in D_u} \sup_{\mathbb{E}\|\boldsymbol{v}_u^T\|_2 \leq (\frac{\tau}{\rho})^{1/4} \vee (\frac{\tau}{\rho})^{1/2}} \|\boldsymbol{v}_u^T\|_2 \left\| \sum_{(i,i') \in D_i} \sigma_i \widetilde{\boldsymbol{x}}_i \right\|_2 \\
&\leq \frac{1}{m} \sum_{u \in D_u} \sup_{\mathbb{E}\|\boldsymbol{v}_u^T\|_2 \leq (\frac{\tau}{\rho})^{1/4} \vee (\frac{\tau}{\rho})^{1/2}} \|\boldsymbol{v}_u^T\|_2 \sqrt{\mathbb{E}_{\boldsymbol{\sigma}} \left\| \sum_{(i,i') \in D_i} \sigma_i \widetilde{\boldsymbol{x}}_i \right\|_2^2} \\
&\leq \frac{1}{m} \sum_{u \in D_u} \sup_{\mathbb{E}\|\boldsymbol{v}_u^T\|_2 \leq (\frac{\tau}{\rho})^{1/4} \vee (\frac{\tau}{\rho})^{1/2}} \|\boldsymbol{v}_u^T\|_2 \sqrt{\sum_{(i,i') \in D_i} \widetilde{\boldsymbol{x}}_i^T \widetilde{\boldsymbol{x}}_i}.
\end{aligned}$$

Therefore, we have the following bounded Rademacher complexity.

$$\begin{aligned}
\Re_m(\mathcal{F}_\tau^{mix}) &\leq \frac{\sqrt{|D_i|}}{m} \cdot \sqrt{rank(\Sigma)} \cdot \sum_{u \in D_u} \sup_{\mathbb{E}\|\boldsymbol{v}_u^T\|_2 \leq (\frac{\tau}{\rho})^{1/4} \vee (\frac{\tau}{\rho})^{1/2}} \|\boldsymbol{v}_u^T\|_2 \\
&\leq \frac{1}{\sqrt{|D_i|}} \cdot \sqrt{rank(\Sigma)} \cdot (\frac{\tau}{\rho})^{1/4} \vee (\frac{\tau}{\rho})^{1/2} \\
&= \max\{(\frac{\tau}{\rho})^{1/4}, (\frac{\tau}{\rho})^{1/2}\} \sqrt{\frac{rank(\Sigma)}{|D_i|}}.
\end{aligned}$$

Similarly, for the general condition, we can bound the empirical Rademacher complexity as follows,

$$\Re_m(\mathcal{F}) \leq \sqrt{\frac{\tau^2}{|D_i|}}.$$

Suppose there exists a loss function, which is $C_l - lipschitz$ continuous and bounded by $[-B, B]$, based on Lemma F.1, we can derive the corresponding generalization error bound using these two Rademacher complexities.

The proof is complete. □