Graph Convolution Network based Recommender Systems: Learning Guarantee and Item Mixture Powered Strategy

Anonymous Author(s)
Affiliation
Address
email

Abstract

Inspired by their powerful representation ability on graph-structured data, Graph Convolution Networks (GCNs) have been widely applied to recommender systems, and have shown superior performance. Despite their empirical success, there is a lack of theoretical explorations such as generalization properties. In this paper, we take a first step towards establishing a generalization guarantee for GCN-based recommendation models under inductive and transductive learning. We mainly investigate the roles of graph normalization and non-linear activation, providing some theoretical understanding, and construct extensive experiments to further verify these findings empirically. Furthermore, based on the proven generalization bound and the challenge of existing models in discrete data learning, we propose Item Mixture (IMix) to enhance recommendation. It models discrete spaces in a continuous manner by mixing the embeddings of positive-negative item pairs, and its effectiveness can be strictly guaranteed from empirical and theoretical aspects.

1 Introduction

With the explosive growth of online information, Internet users rely on recommender systems to alleviate such information overload [1, 2, 3]. In order to take full advantage of the rich graph structure with high-hop neighbors in recommender systems, recent studies have begun to apply Graph Convolution Networks (GCNs) [4] to recommender systems. As expected, GCN-based models have emerged as one of the most attractive approaches to recommender systems due to their powerful graph representation abilities. However, these works only demonstrate the effectiveness of GCNs empirically and do not provide theoretical guarantees.

GCN-based Recommendation. [5] was the first to introduce GCN into recommender systems. Specifically, GCN in their approach was applied to the user-user and item-item graphs to learn auxiliary information, respectively. GC-MC [6] proposed to apply GCN to the user-item interaction graph to learn the relations between users and items. SpectralCF [7] directly discovered all possible connectivity between users and items from the spectral domain of the user-item graph. However, SpectralCF fails to capture the information in high order. NGCF [8] further proposed employing the element-wise product between users and embeddings to augment the item features that users care about and the user features that items are selected by. Some works [8, 9, 10, 11, 12] observed that removing the non-linear activation not only simplifies the update operation, but also retains or enhances the recommendation performance. LightGCN [9] further removed the feature transformation to simplify the design of GCN. However, these changes only depend on the empirical results, without theoretical guarantees. In addition to the user-item interaction matrix, there exist studies combining the social network [13] and knowledge graph [14] to enhance the representations of users and items. In this paper, we mainly consider the user-item interaction matrix, i.e. bipartite graph.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Generalization Analysis for GNNs. Due to the numerical success [15, 5, 16], several theoretical studies have been devoted to establishing generalization guarantees for GNNs. In inductive learning, [17] studied the generalization capability of GNNs by the VC-dimension, which is polynomial w.r.t. the number of parameters and the number of nodes in the graph. [18] provided the first data-dependent bounds via Rademacher Complexity for message passing GNNs and did not allow the analysis of graph information. Subsequently, [19] derived a tighter bound for GNNs based on the PAC-Bayes framework. In transductive learning, [20] analyzed the uniform stability of (single layer) GCNs and thereby deriving their generalization guarantees. They proved that the graph filter, which makes the largest absolute eigenvalue independent of the graph size, satisfies this uniform stability criterion. [21] derived a Rademacher complexity based bound for two layers GCN model. [22] provided a more informative bound according to the relations between normalized graph and input feature at the first layer. However, these works mentioned above only focused on node or graph classification tasks, different from recommendations. The recent works [23, 24] have studied the link prediction and graph embedding tasks; however, they did not concern with the specific models.

Contributions and paper structure. The main contributions are shown as follows:

1) We quantify the model complexity for GCN-based recommender systems on the bipartite graph via Covering number. And then, in Section 3, we relate the Covering number of function space with a specific radius to the generalization error under the inductive and transductive settings, leading to the generalization guarantees.

2) Motivated by the proven bound that generalization needs to compromise empirical risk and model complexity and the fact that the recommendation model overfits the sampled discrete data, we propose a simple but effective augmentation strategy, Item mixture (IMix) in Section 4. It makes discrete sample space continuous by mixing positive-negative item pairs about the same user. Notably, we theoretically demonstrate that it is beneficial for recommendation generalization.

3) In Section 5, extensive evaluations are conducted on two datasets, Gowalla and Yelp2018. First, to confirm the previous theoretical foundation, we present the corresponding numerical results, mainly including graph normalization and non-linear activation. Second, we validate the effectiveness of the proposed IMix on different models, and the experimental results consistently show that the recommendation performance is significantly improved after configuring IMix.

We conclude in Section 6. All proofs and more experimental results are provided in the appendix.

2 Preliminary

In this section, we first formulate the problem of collaborative filtering (CF) based recommender system. We then describe the GCN-based CF models in detail.

2.1 Problem Formulation

We use $u \in \mathcal{U}$ and $i \in \mathcal{I}$ to denote the user and item, and use $\mathcal{R} = \{(u,i)|u \in \mathcal{U}, i \in \mathcal{I}\}$ to denote the collected user-item interactions. Let $\mathcal{Z} = \{(u,i)|u \in \mathcal{U}, i \in \mathcal{I}\}$, $\mathcal{D}$ to be a fixed but unknown distribution over $\mathcal{Z} \times \mathcal{Z} \times \{-1, +1\}$. For any triplet $(z, z', y) \sim \mathcal{D}$, $y = +1$ if the preference between user-item pair z is higher than z', otherwise $y = -1$. Let $f \in \mathcal{F} : \mathcal{U} \times \mathcal{I} \to \mathbb{R}$ be a recommender system. For any user-item pair $z \in \mathcal{Z}$, the model predicts a preference score $f(z)$ by the learned user representation $\boldsymbol{e}_u$ and item representation $\boldsymbol{e}_i$. In this paper, we consider the inner-product as the preference function, i.e., $f(z) = \boldsymbol{e}_u^T \boldsymbol{e}_i$. To optimize the recommendation ability of f, we adopt a pairwise ranking loss as follows, for any $(z, z', y) \sim \mathcal{D}$,

$$loss = l(y(f(z) - f(z'))).$$

2.2 GCN-based Recommendation

Let $\boldsymbol{x}_u, \boldsymbol{x}_i \in \mathbb{R}^d$ be the user and item feature, where d is the input dimension. Without loss of generality, we fix the collected user-item interactions as matrix $\boldsymbol{R} \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{I}|}$, where each entry $\boldsymbol{R}_{ui} = 1$ if user u has interacted with item i. Let $\boldsymbol{A}$ and $\tilde{\boldsymbol{A}}$ to be the adjacency matrix the user-item graph and the corresponding laplacian matrix, which are both block anti-diagonal matrices since the

graph is bipartite. We formalize these two matrices as follows.

$$\boldsymbol{A} = \begin{bmatrix} \mathbf{0} & \boldsymbol{R} \\ \boldsymbol{R}^T & \mathbf{0} \end{bmatrix} \text{ and } \tilde{\boldsymbol{A}} = \begin{bmatrix} \mathbf{0} & \tilde{\boldsymbol{R}}_u \\ \tilde{\boldsymbol{R}}_i & \mathbf{0} \end{bmatrix}.$$

Following [25], the existing GCN-based collaborative filtering models can be generalized into two steps. Note that we omit the update formula of items for the sake of brevity.

(1) Update operation. The most straightforward layer-wise updations are inspired by the GNN techniques, and they can be summarized as follows,

$$\mathbf{h}_u^{l+1} = \phi\left(\mathbf{W}_1^l \mathbf{h}_u^l + \sum_{i \in \mathcal{N}_u} \tilde{a}_{ui} \mathbf{W}_1^l \mathbf{h}_i^l\right), \tag{1}$$

where $\tilde{a}_{ui} = [\widetilde{\boldsymbol{R}}_u]_{ui}$ represents the assigned weights of user u to item i. Some recent works [8] further adopt the element-wise product between nodes with their neighbors to leverage the consistency between user preferences and item attributes. This operation is formulated as,

$$\mathbf{h}_u^{l+1} = \phi\left(\mathbf{W}_1^l \mathbf{h}_u^l + \sum_{i \in \mathcal{N}_u} \tilde{a}_{ui} \left(\mathbf{W}_1^l \mathbf{h}_i^l + \mathbf{W}_2^l \left(\mathbf{h}_i^l \odot \mathbf{h}_u^l\right)\right)\right). \tag{2}$$

In this paper, we focus on the more complex operation defined in Eq. 2 since the Eq. 1 can be seen as a special case of it when $\boldsymbol{W}_2^l = \boldsymbol{O}$.

(2) Final node representation. The CF-based recommender systems require the representation of users and items for the final prediction task. The mainstream approaches include linear combination and concatenation, respectively defined as,

$$\text{Linear combination:} \quad \boldsymbol{e}_u = \sum\nolimits_{l=0}^{L} \alpha^l \mathbf{h}_u^l,$$
$$\text{Concatenation:} \quad \boldsymbol{e}_u = \mathbf{h}_u^0 \oplus \mathbf{h}_u^1 \oplus \cdots \oplus \mathbf{h}_u^L,$$

where α^l is a learnable parameter or hyper-parameter, L is the network depth. In this paper, for convenience, we only consider the embeddings of the last layer as the final representation, i.e., $\boldsymbol{e}_u = \mathbf{h}_u^L$. We will prove that the generalization analysis can be easily extended to other integration operations in appendix.

3 Generalization for GCN-based Recommender Systems

We first investigate the model complexity of GCN-based recommender system via Covering number with specific radius in Section 3.1. Then we link the Covering number and generalization errors for inductive and transductive learning in Section 3.2, which finally yield the meaningful bounds.

Assumptions. To establish the generalization guarantee, let us immediately make some mild assumptions that are easy to implement. For the loss function, assuming $l : \mathbb{R} \to [-B, B]$ is any bounded C_l-$lipschitz$ continuous function. For the non-linear activation function, we assume ϕ is C_ϕ-$lipschitz$ continuous with some $C_\phi > 0$, $\phi(0) = 0$. For the weight matrices, we assume that $\boldsymbol{W}_1^l, \boldsymbol{W}_2^l \in \mathbb{R}^{d \times d}$ are shared across layers, and have bounded norms $\|\boldsymbol{W}_1^l\|_2 \leq B_1$ and $\|\boldsymbol{W}_2^l\|_2 \leq B_2$. For the input feature, we assume $\max_{u \in \mathcal{U}} \|\boldsymbol{x}_u\|_2 \leq B_u$ and $\max_{i \in \mathcal{I}} \|\boldsymbol{x}_i\|_2 \leq B_i$.

3.1 Covering Number for GCN-based Recommender Systems

In classical learning theory, the complexity of function class is closely related to the generalization ability [26], and the typical complexity measurements are VC-Dimension [27] and Rademacher complexity [28]. In this paper, we adopt the Covering number [29] to investigate the model complexity of GCN-based recommender systems. Based on the properties of bipartite graphs, we study the complexity by combining the recommendations of GCN-based models for any user-item pairs.

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

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

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

$$\begin{aligned}
&\log \mathcal{N}\left(\mathcal{F}|S, \epsilon, \|\cdot\|\right) \leq 2d^2 \log \frac{8M(\gamma+1)\sqrt{ndB_1B_2b}}{\epsilon}, \ where \\
&M = C_\phi \mathcal{C}^{2L-1}(B_u + B_i)^2 \frac{(2\mathcal{C})^L - 1}{2\mathcal{C} - 1}, \\
&\mathcal{C} = C_\phi[B_1 + \gamma(B_1 + B_2 b)].
\end{aligned}$$

Remark 1. The key idea of the proof is to exploit the properties of the bipartite graph in the recommender system. To investigate the complexity of the recommendation for any user-item pair generated by GCN-based models, we establish the connection between the complexities of function space and weight space. Here $\mathcal{C}$ describes the layer-wise update complexity, which is the important term for the following generalization analysis. According to the norm bound of hidden states, the assumption can be naturally satisfied when the activation function is bounded. And we will discuss the role of unbounded activation in Section 3. Note that the layer-wise complexity reduces to $\mathcal{C} = C_\phi(\gamma+1)B_1$ when ignoring element-wise between users and items, which means that there is no need to discuss the value of b.

3.2 Generalization Bound for Inductive & Transductive Learning

Transductive and inductive learning are both commonly used in GNNs and recommender systems. The primary difference between these two settings is whether the testing set depends on the training set. The testing samples in inductive learning are drawn from some unknown distribution, while in transductive learning, they are sampled without replacement from a fixed data set. In the following, we first relate the generalization error for two settings to the Covering number of GCN-based recommendations, then derive the final generalization guarantees using Lemma 1.

Inductive Learning. Given a training set $S_m = \{(z_i, z_i', y_i)\}_{i=1}^m$, we assume that all samples in S_m are i.i.d. according to $\mathcal{D}$, denoted as $S_m \sim \mathcal{D}^m$. Under this inductive setting, the empirical error over a sample S_m and the corresponding generalization error is defined as follows,

$$\begin{aligned}
&\text{Empirical error:} \qquad \hat{\mathcal{L}}_m(f) = \frac{1}{m}\sum\nolimits_{(z_i, z_i', y_i)\in S_m} l(y_i(f(z_i) - f(z_i')), \\
&\text{Generalization error:} \qquad \mathcal{L}(f) = \mathbb{E}_{(z,z',y)\sim\mathcal{D}}[l(y(f(z) - f(z'))].
\end{aligned} \tag{3}$$

Under inductive learning, we can derive the following generalization error bound by the complexity term according to the Covering number.

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*

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

Transductive Learning. For the transductive learning setting, the full samples are observed prior to learning. Given a fixed set $S = \{(z_i, z_i', y_i)\}_{i=1}^{m+u}$, a labeled training set S_m with size m is selected from S uniformly at random. The goal of transductive learning is to predict the labels of the samples in testing set $S_u = S \backslash S_m$. Under this transductive setting, the formulation of empirical error $\hat{\mathcal{L}}_m(f)$ is similar to Eq. 3, and the generalization error is defined as follows,

$$\hat{\mathcal{L}}_u(f) = \frac{1}{u}\sum\nolimits_{(z_i, z_i', y_i)\in S_u} l(y_i(f(z_i) - f(z_i')).$$

Following the same strategy adopted in proving Lemma 2, we again need to perform a generalization bound under transductive setting via Covering number of function class. As shown above, the

generalization and empirical error under transductive setting do not depend on any underlying distributions. The important difference induces two versions of generalization analysis. For the transductive learning, we consider the transductive Rademacher complexity [30] defined on both the training and testing data. Then the Covering number bounded generalization error can be derived by the interdependence between model complexity measurements [31, 32].

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*

$$\mathcal{L}_u(f) \leq \hat{\mathcal{L}}_m(f) + 2C_l \inf_{\alpha>0} \left(\frac{4\alpha}{\sqrt{m+u}} + \frac{12}{m+u} \int_{\alpha}^{2e\sqrt{m+u}} \sqrt{\log \mathcal{N}(\mathcal{F}|S, \epsilon, \|\cdot\|)} d\epsilon \right) + Bc_0 Q_1 \sqrt{\min(m,u)} + 2B\sqrt{\frac{Q_2}{2} Q_1 \ln \frac{1}{\delta}}.$$

Armed with Lemma 2, Lemma 3 and Lemma 1, we now present the generalization error bound for inductive and transductive learning.

Theorem 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,*

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

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

$$\mathcal{L}_u(f) \leq \hat{\mathcal{L}}_m(f) + \frac{24C_l}{\sqrt{m+u}} \mathcal{C}^{2L}(B_u + B_i)^2 d \sqrt{2\log(8(m+u)M(\gamma+1)\sqrt{dB_1B_2b})} + \frac{4C_l\sqrt{2mu}}{(m+u)^2} + Bc_0 Q_1 \sqrt{\min(m,u)} + 2B\sqrt{\frac{Q_2}{2} Q_1 \ln \frac{1}{\delta}}.$$

Remark 2. Theorem 1 states that the gap between generalization error and empirical error converges uniformly to 0 with the rate of $\sqrt{(\log m)/m}$. We first discuss several obvious findings about the dependency of the bound on some terms. For the hidden dimension d, the generalization bound scales as $\mathcal{O}(d\sqrt{\log d})$; for the input feature norms B_u and B_i, the bound scales as $\mathcal{O}((B_u + B_i)^2\sqrt{\log(B_u + B_i)})$. The most important term is the layer-wise update complexity $\mathcal{C}$, which depends on the value of γ according to the graph structure, the spectral norms of weight matrices B_1 and B_2, and the value of C_ϕ and b according to the non-linear activation function. A natural idea is to constrain the weight matrices during training such that the generalization bound would not grow exponentially with the network depth L, which is a regularization techniques commonly used in machine learning. The subsequent discussions focus on the γ, C_ϕ and b to investigate the influence of graph normalization and non-linear activation functions.

Role of graph normalization. It is easy to see that γ is equal to the maximum degree $D_{\max}$ of the user-item interaction graph without graph normalization. Therefore, fixed other terms in $\mathcal{C}$, the generalization bound grows as $O(D_{\max}^L)$. For the random-walk graph, i.e., $\widetilde{\boldsymbol{A}} = \boldsymbol{D}^{-1}\boldsymbol{A}$, $\gamma = 1$, and for symmetric normalized graph, i.e., $\widetilde{\boldsymbol{A}} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$, $\gamma \leq \sqrt{D_{\max}/D_{\min}}$, and hence, the growth is much smaller than the unnormalized graph.

Role of non-linear activation. As pointed above, the activation functions affect the generalization performance through two quantities: the Lipschitz constant C_ϕ and the norm bound of hidden states b. Due to the most activation functions satisfy the condition $C_\phi \leq 1$, and particularly LeakyReLU, ReLU, and Tanh are 1-$Lipschitz$, C_ϕ presents negligible influence. For the norm bound b, while the bounded activation naturally provides explicit value, the unbounded function can also satisfy the assumption by constraints on the weight matrices. As discussed in Remark 2, weight regularization is widely used in machine learning and can help improve model generalization performance to prevent overfitting. In addition, one recent work [33] gives a new theoretical finding that non-linearity does not bring considerable expressive power to improve representation learning in GNNs. Therefore, this generalization bound provides theoretical support for removing the activation function to simplify the model without incurring a performance penalty.

4 Item Mixture Powered Recommendation

The upper bound in Theorem 1 inspires us that the generalization needs to balance empirical risk and model complexity. Most models [34, 8] inevitably lead to excessive complexity in pursuit of high expressiveness, but simple models may lead to higher empirical errors, so it is valuable to reduce the complexity while ensuring model expressiveness. In addition, the recommendation algorithm is trained based on the limited discrete data obtained by sampling, making the model perform well in the discrete space of sampling and prone to overfitting. Considering the above and inspired by Mixup [35], we propose Item Mixture (IMix) to power recommendation, which forces the model to deal with regions between discrete samples continuously by mixing any two positive-negative item pairs. In addition, IMix does not make changes to the model, while we can theoretically prove that it reduces the model complexity, and then enjoys generalization. Below we give specific details.

Since the original representations of items are discrete and unprocessable (e.g., orange or apple), we suggest mixing their embeddings. Specifically, for a triplet $(\boldsymbol{e}_u, \boldsymbol{e}_i, \boldsymbol{e}_{i'})$ computed according to Section 2.2, where user u prefers item i to i', we arbitrarily sample another embedding triplet $(\boldsymbol{e}_u, \boldsymbol{e}_j, \boldsymbol{e}_{j'})$ of the same user. We can obtain mixed embedding triplet $(\boldsymbol{e}_u, \widetilde{\boldsymbol{e}}_i, \widetilde{\boldsymbol{e}}_{i'})$, where $\widetilde{\boldsymbol{e}}_i = \lambda \boldsymbol{e}_i + (1-\lambda)\boldsymbol{e}_j$, and $\widetilde{\boldsymbol{e}}_{i'} = \lambda \boldsymbol{e}_{i'} + (1-\lambda)\boldsymbol{e}_{j'}$. $\lambda \sim Beta(\alpha, \beta)$ for $\alpha, \beta > 0$, and we set $D_\lambda = Beta(\alpha, \beta)$. So in IMix, we will train the recommendation model on the new mixed embeddings. The standard loss and IMix loss are defined as

$$\begin{aligned} &\text{Standard loss:} \qquad \mathcal{L}_m^{std} = \frac{1}{m}\sum\nolimits_{k=1}^{m} l(\boldsymbol{e}_{u_k}^T(\boldsymbol{e}_{i_k} - \boldsymbol{e}_{i'_k})), \\ &\text{IMix loss:} \qquad \mathcal{L}_m^{IMix} = \frac{1}{m}\sum\nolimits_{k=1}^{m} l(\boldsymbol{e}_{u_k}^T(\widetilde{\boldsymbol{e}}_{i_k} - \widetilde{\boldsymbol{e}}_{i'_k})). \end{aligned} \tag{5}$$

Considering the motivation above, IMix preserves the model expressiveness without modification of the model structure. Besides, it transforms the learning of discrete sample pairs into the learning of continuous regions between sample pairs, which may helps to alleviate the overfitting.

Notably, although IMix is inspired by the generalization bound in Theorem 1, it is decoupled from the model structure. Therefore, it is theoretically applicable to any embedding-based recommendation model, and we will show its performance in non-GCN-based models in Appendix A.3.

Next, we theoretically guarantee the effectiveness of the proposed IMix. Following a similar approach in [36], we adopt second-order Taylor expansion to state the IMix training loss is approximately equivalent to standard empirical error with a regularization term.

Lemma 4. *Consider a training set S_m, and denote $\hat{\boldsymbol{E}}_{u_k} = \frac{1}{m}\sum_{k=1}^{m} y_k(\boldsymbol{e}_{i_k} - \boldsymbol{e}_{i'_k})$, $\hat{\Sigma}_X = \frac{1}{m}\sum_{k=1}^{m}(\boldsymbol{e}_{i_k} - \boldsymbol{e}_{i'_k})(\boldsymbol{e}_{i_k} - \boldsymbol{e}_{i'_k})^T$ and $\omega(\lambda) = \frac{1-\lambda}{\lambda}$. If the loss function l is twice differentiable, then the IMix loss defined in Eq. 5 has the following approximation,*

$$\mathcal{L}_m^{mix} \approx \mathcal{L}_m^{std} + \frac{1}{m}\sum\nolimits_{k=1}^{m}\left[\mathbb{E}_\lambda[\omega(\lambda)]l'(\eta_k)\boldsymbol{e}_{u_k}^T\hat{\boldsymbol{E}}_{u_k} + \frac{1}{2}\mathbb{E}_\lambda[\omega(\lambda)^2]l''(\eta_k)\boldsymbol{e}_{u_k}^T\hat{\Sigma}_X\boldsymbol{e}_{u_k}\right].$$

Then we are ready to investigate the generalization bound considering the function class $\mathcal{F}_\tau^{mix}$ related to the dual problem of the above regularized term. The details for this function class are provided in Appendix F.2.

Remark 3. Assuming that the loss function l satisfies $l'(x)\mathbb{E}_\lambda[\omega(\lambda)] \le l''(x)\mathbb{E}_\lambda[\omega(\lambda)^2] \le \rho \le \tau, \forall x \in \mathbb{R}$. Then with probability of at least $1-\delta$, for any function $f \in \mathcal{F}_\tau^{mix}$, we have,

$$\mathcal{L}(f) \le \hat{\mathcal{L}}(f) + 4C_l\sqrt{\frac{rank(\Sigma_X)\tau}{\rho m}} + 2B\sqrt{\frac{2\log 4/\delta}{m}}.$$

If we consider the function class with a general regularization technique on the weight matrices, as we discussed above, the constraint on weights will ultimately control the node representation. Therefore, for the general condition, we focus on the function class $\mathcal{F}_\tau^{std} = \{\mathcal{F} | \mathbb{E}_{(u,i,i')}\left[\|\boldsymbol{e}_u\|_2^2 + \|\boldsymbol{e}_i\|_2^2\right] \le \tau\}$, then the similar proof strategy would yield the generalization error bound $\mathcal{L}(f) \le \hat{\mathcal{L}}(f) + 4C_l\sqrt{\frac{\tau^2}{m}} + 2B\sqrt{\frac{2\log 4/\delta}{m}}$. we can conclude that when $rank(\Sigma_X) < \tau$, IMix has a better generalization. It is easily met because τ tends to be large to pursue high performance. For example, in the experiment, the embedding size d is 64 (noted that $rank(\Sigma_X) < d$), while the minimum value of τ is 130.8.

5 Experiments

In this section, we conduct extensive experiments to validate our theoretical findings and the effectiveness of the proposed IMix.

Dataset. We follow existing work [8, 9], using two datasets: (1) Gowalla, contains 1,027,370 check-in information from 29,858 users to 40,981 locations; (2) Yelp2018, comprises 31,668 users who have reviewed 38,048 items (e.g., restaurants, bars) with 1,561,406 times. The training set, validation set, and test set are randomly sampled in a ratio of $7:1:2$.

Model Structure. Two competitive GCN-based algorithms, NGCF [8] and LightGCN [9], are used to construct experiments. Among them, NGCF explicitly models the high-order connectivity between users and items through message construction and message aggregation, thereby improving the expressiveness of embedding. LightGCN, on the other hand, is keenly aware that the nonlinear activation function and feature transformation of NGCF have no specific semantics in the recommendation, so it is simplified to only a domain message aggregation module. Correspondingly, these models are configured with the proposed IMix denoted as NGCF-IMix and LightGCN-IMix, respectively.

We use the public implementation version of LightGCN [1] and follow the default parameters of their algorithms unless otherwise specified. More details are provided in Appendix A.1.

Metrics. The recommendation performance is measured by the positive items' rankings in the test set. We use two ranking metrics widely used in recommender systems [37, 38, 39]: Recall and NDCG. Recall at a cutoff K, denoted as Recall@K, denotes the proportion of positive items predicted to be correct. NDCG at a cutoff K, denoted as NDCG@K, measures the reward for the positive items' position in the top-K recommendation list. For these two metrics, we truncate the ranked list K to 20.

5.1 Numerical Discussion

In this subsection, we discuss the implication of the generalization bound defined in Theorem. 1, and the results are built on NGCF because it is more similar to traditional GCNs.

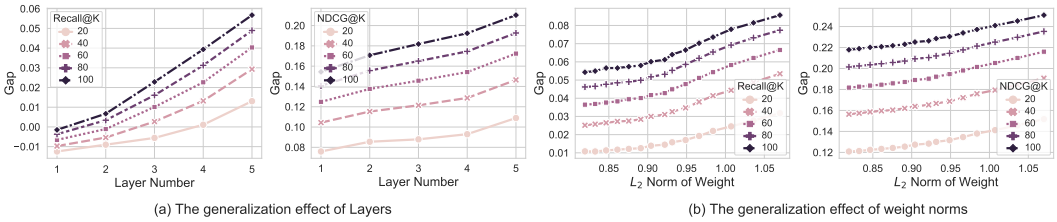
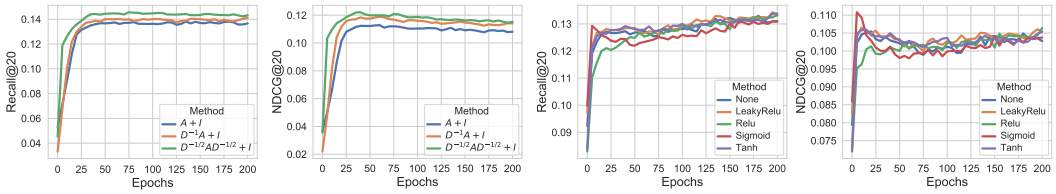


Figure 1: (a): the effect of number of layers L on Generalization Performance on Gowalla. (b): the effect of weight norms B_1, B_2 on Generalization Performance on Gowalla.

Quantitative Analysis. We investigate the effects of two major values on the generalization gap, which represents the differences between recommendation performances in training and testing sets. We consider the number of layers L, as shown in the left part of Fig. 1. As the number of layers increases, the generalization gaps of ndcg@K and recall@K significantly increase. The empirical results imply that the generalization error is proportional to the number of layers, consistent with the theoretical analysis. We also consider the spectral norms of weight matrices B_1 and B_2, as shown in the right part of Fig. 1. We compare the generalization abilities with different weight norms. As the norms increase, the generalization gaps increase steadily. Compared with the number of layers, we can observe that the weight norms increase more slowly. The theoretical reason is that the generalization performance is L-th order with respect to the weight norms B_1 and B_2, and exponential with respect to the number of layers L.

The Role of Normalized Graph. We compare the generalization performances with different normalized graphs, as shown in the left part of Fig. 2. It is clear that the two normalized graphs have higher generalizations than the unnormalized graph. This result is consistent with the theoretical findings. The comparison between two normalized graphs depends on the specific user-item interaction graph.

[1] `https://github.com/kuandeng/LightGCN`



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

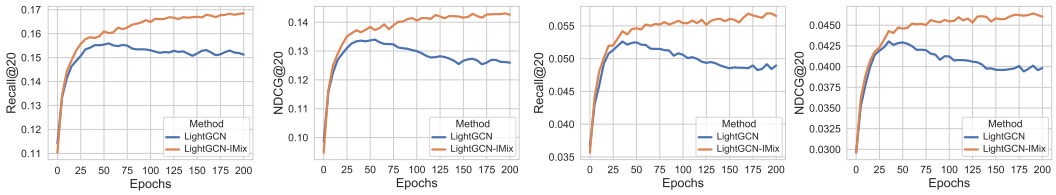
Figure 2: (a): the effect of normalized graph on Generalization Performance on Gowalla. (b): the effect of activation function on Generalization Performance on Gowalla.

The Role of Non-linear Activation Function. We study the role of non-linear activation function in GCN-based recommender systems. Since most activation functions are all 1-Lipschitz continuous, we compare the generalization performances with the popular activation functions, including LeakyReLU, ReLU, Sigmoid, and Tanh. Furthermore, we compare these results with the model without activation function. As shown in the right part of Fig. 2, these models present similar performances, which provide theoretical support for removing activation functions. Although the Sigmoid presents great recommendations at first, it overfitted as the training continued, while the performance of other models is still improved, and their optimal performance tends to be unified.

5.2 Performance Analysis of IMix

Table 1: Overall evaluation. All models' performance is improved after configured with IMix.

Dataset		Gowalla		Yelp2018	
Layer #	**Method**	**Recall@20**	**NDCG@20**	**Recall@20**	**NDCG@20**
1 layer	NGCF	0.1308	0.1032	0.0419	0.0336
	NGCF-IMix	0.1451(+10.93%)	0.1146(+11.05%)	0.0492(+17.42%)	0.0360(+7.14%)
2 layer	NGCF	0.1274	0.1050	0.0431	0.0348
	NGCF-IMix	0.1454(+14.13%)	0.1164(+10.86%)	0.0491(+13.92%)	0.0391(+12.36%)
3 layer	NGCF	0.1303	0.1082	0.0435	0.0353
	NGCF-IMix	0.1496(+14.81%)	0.1191(+10.07%)	0.0504(+15.86%)	0.0405(+14.73%)
1 layer	LightGCN	0.1556	0.1340	0.0526	0.0427
	LightGCN-IMix	0.1699(+9.19%)	0.1436(+7.16%)	0.0562(+6.84%)	0.0458(+7.26%)
2 layer	LightGCN	0.1672	0.1425	0.0564	0.0462
	LightGCN-IMix	0.1765(+5.56%)	0.1507(+5.75%)	0.0621(+10.10%)	0.0507(+9.74%)
3 layer	LightGCN	0.1759	0.1500	0.0602	0.0492
	LightGCN-IMix	0.1793(+1.93%)	0.1524(+1.6%)	0.0644(+6.98%)	0.0527(+7.11%)



(a) Gowalla (b) Yelp2018

Figure 3: Test curves (Recall@20 and NDCG@20) for 200 epochs of training.

Effectiveness Validation. We apply the proposed IMix to NGCF and LightGCN and perform a detailed comparison, as shown in Table 1. The percentage of relative improvement is shown in parentheses. Obviously, the proposed augmentation strategy (corresponding to NGCF-IMix and LightGCN-IMix) significantly improves the performance of the corresponding models (NGCF and LightGCN). For example, in the Gowalla dataset, the best performance of NGCF on Recall@20 is 0.1308, while with IMix configured, the performance improves to an encouraging 0.1496. In addition,

the improvements of IMix on NGCF are larger than that on LightGCN. We reasonably believe that LightGCN is more expressive (better performance), and the benefit of the boosting strategy decreases accordingly. Nonetheless, the performance on LightGCN is also a pleasant surprise.

Besides, Fig. 3 presents the performance curves regarding Recall and NDCG for the test set during the training of LightGCN-based models (LightGCN and LightGCN-IMix). Throughout the training process, LightGCN-IMix consistently outperforms LightGCN. After 50 epochs, LightGCN shows different degrees of overfitting. On the contrary, the performance of LightGCN-IMix is still further improved, and their gap is gradually enlarged. We reasonably believe that the proposed strategy adds a data-dependent regularization term to LightGCN (Lemma 4), which helps to alleviate overfitting.

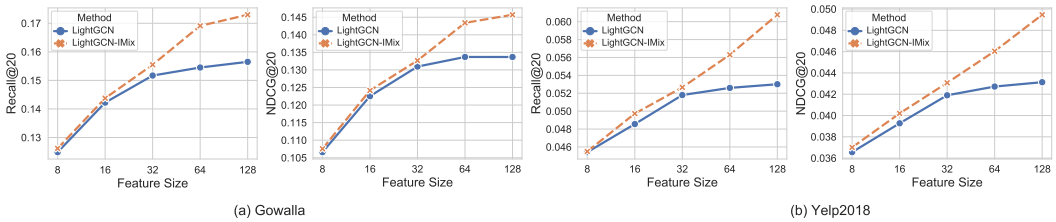


Figure 4: Performance comparison of different embedding sizes.

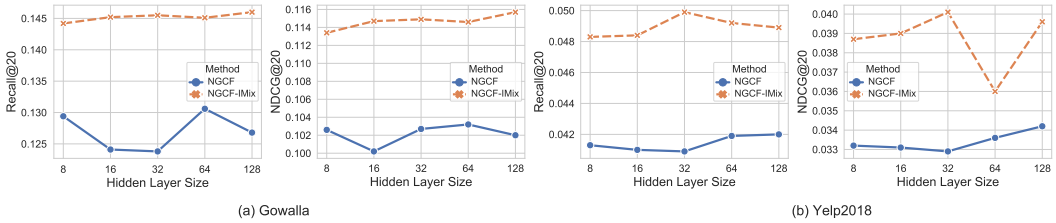


Figure 5: Performance comparison of different hidden layer size.

Sensitivity w.r.t. Model Hyperparameters. We also explore the impact of the model's hyperparameters: different feature sizes (dimension of x_u and x_i) and hidden layer sizes (dimension of W_1^l and W_2^l). Here, the number of layers is set to 1, and the results are shown in Fig. 4 and Fig. 5.

In Fig. 4, we find that the larger the feature size, the more beneficial it is to the recommendation. Moreover, LightGCN-IMix is better than LightGCN, and even more, the better the performance of the original model, the better the improvement after configuring IMix, which highlights the superiority of the proposed IMix.

Fig. 5 reveals that the hidden layer size does not cause a significant change in model performance, while when the hidden layer size is only 8, the performance is sufficiently satisfactory, which provides support for the practicality of GCN-based algorithms. In addition, with similar conclusions to Fig. 4 and Table 1, the IMix-configured models outperform the original models in all cases. In conclusion, we validate the effectiveness of IMix.

6 Conclusion and Future Work

GCNs achieve superior empirical success in recommender systems but lack theoretical understanding. In this paper, we provide a model complexity analysis for GCN-based recommendations via Covering number, and then induce the generalization error bounds in inductive and transductive learning. We report several theoretical findings and again verify them via experimental results. Inspired by the generalization analysis, we introduce an augmentation strategy, Item Mixture, to improve the recommendation generalization by mixing item embeddings.

There are several interesting future directions. First, it would be interesting to further analyze the user-item bipartite graph with social relationship and knowledge graph. The second is to extend our analysis to investigate the plausibility of the other GNN-based variants. The third is to study the role of graph spectral properties in generalization ability.

References

[1] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198, 2016.

[2] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 974–983, 2018.

[3] Fajie Yuan, Xiangnan He, Alexandros Karatzoglou, and Liguang Zhang. Parameter-efficient transfer from sequential behaviors for user modeling and recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1469–1478, 2020.

[4] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[5] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5115–5124, 2017.

[6] Rianne van den Berg, Thomas N Kipf, and Max Welling. Graph convolutional matrix completion. *arXiv preprint arXiv:1706.02263*, 2017.

[7] Lei Zheng, Chun-Ta Lu, Fei Jiang, Jiawei Zhang, and Philip S Yu. Spectral collaborative filtering. In *Proceedings of the 12th ACM conference on recommender systems*, pages 311–319, 2018.

[8] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*, pages 165–174, 2019.

[9] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648, 2020.

[10] Lei Chen, Le Wu, Richang Hong, Kun Zhang, and Meng Wang. Revisiting graph based collaborative filtering: A linear residual graph convolutional network approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:27–34, 04 2020.

[11] Le Wu, Yonghui Yang, Kun Zhang, Richang Hong, Yanjie Fu, and Meng Wang. Joint item recommendation and attribute inference: An adaptive graph convolutional network approach. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 679–688, 2020.

[12] Zhiwei Liu, Lin Meng, Fei Jiang, Jiawei Zhang, and Philip S Yu. Deoscillated graph collaborative filtering. *arXiv preprint arXiv:2011.02100*, 2020.

[13] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. Graph neural networks for social recommendation. In *The World Wide Web Conference*, pages 417–426, 2019.

[14] Hongwei Wang, Fuzheng Zhang, Mengdi Zhang, Jure Leskovec, Miao Zhao, Wenjie Li, and Zhongyuan Wang. Knowledge-aware graph neural networks with label smoothness regularization for recommender systems. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 968–977, 2019.

[15] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pages 2323–2332. PMLR, 2018.

[16] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.

[17] Franco Scarselli, Ah Chung Tsoi, and Markus Hagenbuchner. The vapnik–chervonenkis dimension of graph and recursive neural networks. *Neural Networks*, 108:248–259, 2018.

[18] Vikas Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks. In *International Conference on Machine Learning*, pages 3419–3430. PMLR, 2020.

[19] Renjie Liao, Raquel Urtasun, and Richard Zemel. A pac-bayesian approach to generalization bounds for graph neural networks. *arXiv preprint arXiv:2012.07690*, 2020.

[20] Saurabh Verma and Zhi-Li Zhang. Stability and generalization of graph convolutional neural networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1539–1548, 2019.

[21] Shaogao Lv. Generalization bounds for graph convolutional neural networks via rademacher complexity. *arXiv preprint arXiv:2102.10234*, 2021.

[22] Pascal Esser, Leena Chennuru Vankadara, and Debarghya Ghoshdastidar. Learning theory can (sometimes) explain generalisation in graph neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[23] Yihan Gao, Chao Zhang, Jian Peng, and Aditya Parameswaran. The importance of norm regularization in linear graph embedding: Theoretical analysis and empirical demonstration. *arXiv preprint arXiv:1802.03560*, 2018.

[24] Atsushi Suzuki, Atsushi Nitanda, Linchuan Xu, Kenji Yamanishi, Marc Cavazza, et al. Generalization bounds for graph embedding using negative sampling: Linear vs hyperbolic. *Advances in Neural Information Processing Systems*, 34, 2021.

[25] Shiwen Wu, Fei Sun, Wentao Zhang, and Bin Cui. Graph neural networks in recommender systems: a survey. *arXiv preprint arXiv:2011.02260*, 2020.

[26] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[27] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

[28] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[29] Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

[30] Ran El-Yaniv and Dmitry Pechyony. Transductive rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 35:193–234, 2009.

[31] Minshuo Chen, Xingguo Li, and Tuo Zhao. On generalization bounds of a family of recurrent neural networks. *arXiv preprint arXiv:1910.12947*, 2019.

[32] Peter Bartlett, Dylan J Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1706.08498*, 2017.

[33] Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. *arXiv preprint arXiv:1905.10947*, 2019.

[34] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. Self-supervised graph learning for recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 726–735, 2021.

[35] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[36] Linjun Zhang, Zhun Deng, Kenji Kawaguchi, Amirata Ghorbani, and James Zou. How does mixup help with robustness and generalization? *arXiv preprint arXiv:2010.04819*, 2020.

[37] Haoyu Wang, Defu Lian, and Yong Ge. Binarized collaborative filtering with distilling graph convolutional networks. *arXiv preprint arXiv:1906.01829*, 2019.

[38] Shuyi Ji, Yifan Feng, Rongrong Ji, Xibin Zhao, Wanwan Tang, and Yue Gao. Dual channel hypergraph collaborative filtering. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2020–2029, 2020.

[39] Xiang Wang, Hongye Jin, An Zhang, Xiangnan He, Tong Xu, and Tat-Seng Chua. Disentangled graph collaborative filtering. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 1001–1010, 2020.

[40] Yehuda Koren, Steffen Rendle, and Robert Bell. Advances in collaborative filtering. *Recommender systems handbook*, pages 91–142, 2022.

[41] Xiaotian Han, Zhimeng Jiang, Ninghao Liu, and Xia Hu. G-mixup: Graph data augmentation for graph classification. *arXiv preprint arXiv:2202.07179*, 2022.

[42] Yiwei Wang, Wei Wang, Yuxuan Liang, Yujun Cai, and Bryan Hooi. Mixup for node and graph classification. In *Proceedings of the Web Conference 2021*, pages 3663–3674, 2021.

[43] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[44] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

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? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (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]

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]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (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)? [No]
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]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (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]

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.

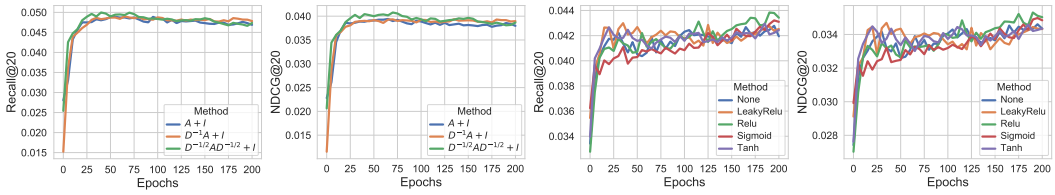
Notation	Size	Meaning
u, i, z		The user, the item and the user-item pair.
$\mathcal{U}, \mathcal{I}, \mathcal{R}, \mathcal{Z}$		The sets of users, items, user-item interactions and user-item pairs.
$\boldsymbol{R}$	$\lvert\mathcal{U}\rvert \times \lvert\mathcal{I}\rvert$	The user-item interaction matrix.
$\boldsymbol{A}, \tilde{\boldsymbol{A}}$	$(\lvert\mathcal{U}\rvert + \lvert\mathcal{I}\rvert) \times (\lvert\mathcal{U}\rvert + \lvert\mathcal{I}\rvert)$	The adjacency matrix and the corresponding normalized graph filter.
$\mathcal{D}$		The unknown data distribution.
S_u, S_m		The testing set in transductive learning and the training set.
$\boldsymbol{W}_1^l, \boldsymbol{W}_2^l$	$d \times d$	The two weight matrices.
$\boldsymbol{e}_u, \boldsymbol{e}_i$	d	The learned embeddings of the user u and the item i.
$\mathbf{h}_u^l, \mathbf{h}_i^l$	d	The hidden states of the user u and the item i.
$D_{\max}, D_{\min}$	Constant	The maximum and minimum degrees of node.

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



(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.

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.

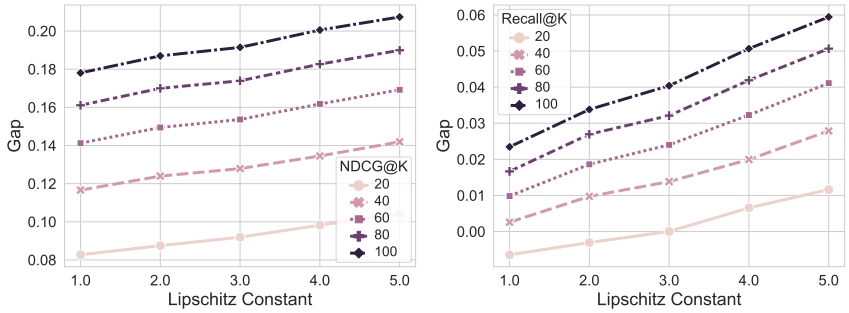


Figure 7: The effect of Lipschitz constant on Generalization performance on Gowalla.

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

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.

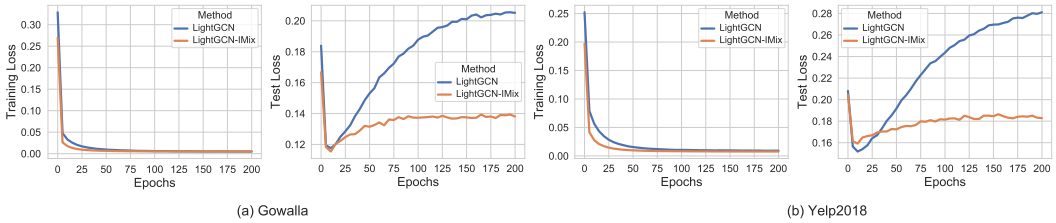


Figure 8: Training curves (training loss) and test curves (test loss) for 200 epochs of training.

A.5 Comparison of Transductive Learning and Inductive Learning

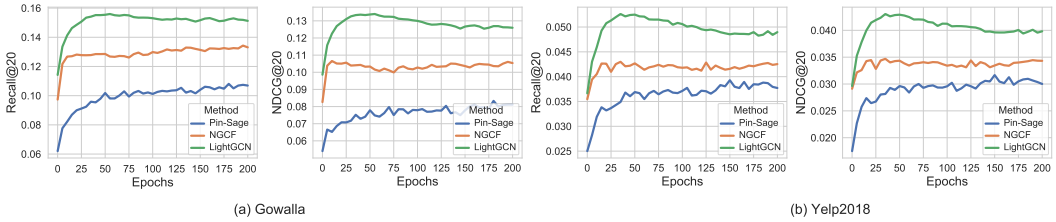


Figure 9: Comparison of test curve (Recall@20 and NDCG@20) of transductive learning (LightGCN, NGCF) and inductive learning (Pin-Sage).

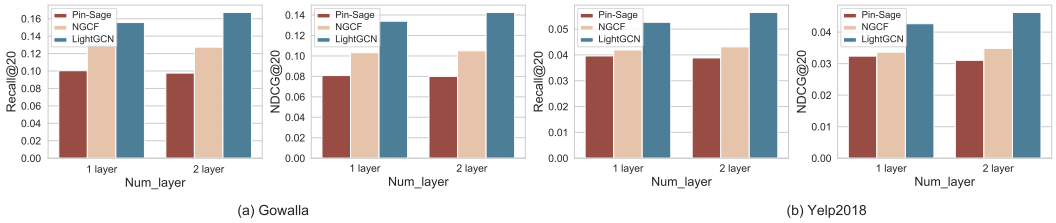


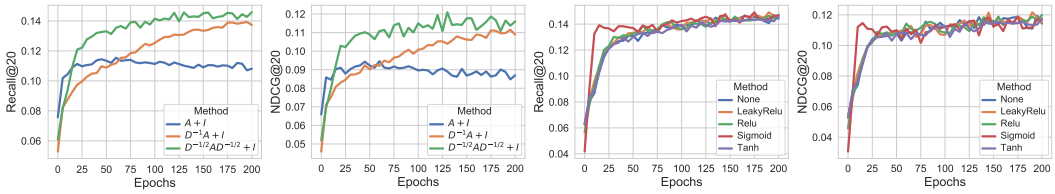
Figure 10: Performance comparison between transductive learning (LightGCN, NGCF) and inductive learning (Pin-Sage) after training.

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 [9] (and NGCF [8]) 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 [40], 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.

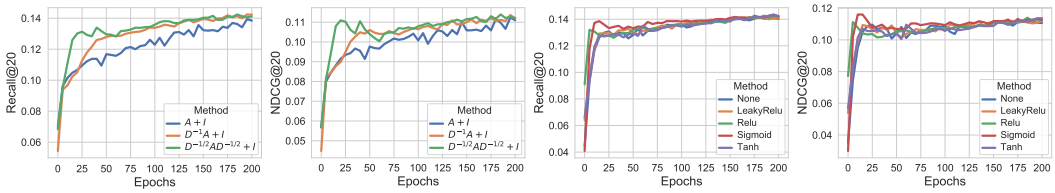
A.6 Additional Experiments on GCN-based Recommendations

In addition to LightGCN and NGCF, we also use two other GCN-based models, SpectralCF [7] and GCMC [6], 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.

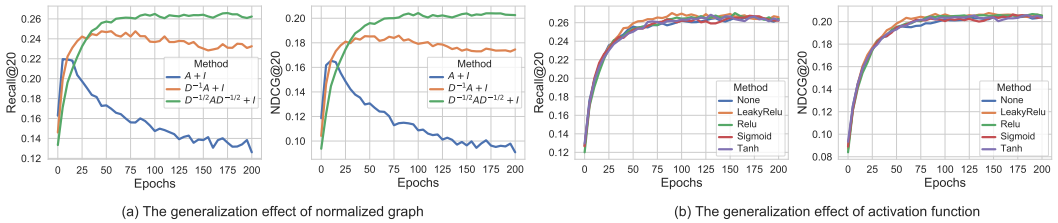
Table 4: Overall evaluation. SpectralCF and GCMC's performance is improved after configured with IMix.

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%)

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.



(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.

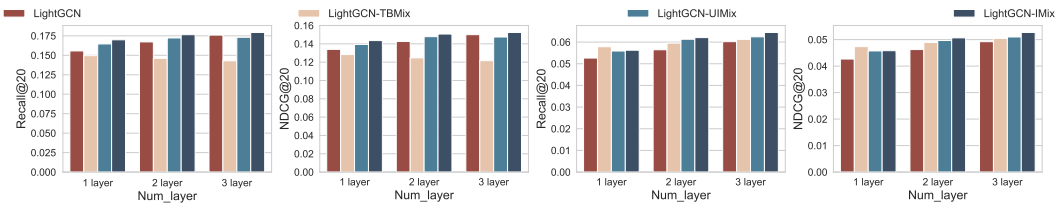
Numerical Discussion. We construct experiments on the effect of normalized graph and activation function on recommendation performance, as shown in Fig. 13 (a) and Fig. 13 (b). We can get similar but more obvious findings as in Gowalla (in Fig. 2) and Yelp2018 (in Fig. 6): (1) The performance on the 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 [9].

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



(a) Gowalla (b) Yelp2018

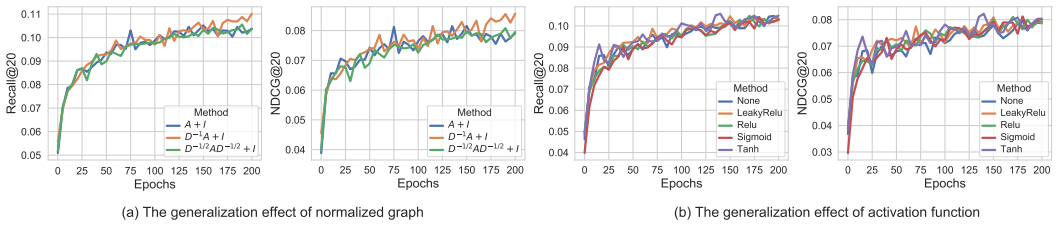
Figure 14: The performance comparison of LightGCN configured various Mixup strategies in Gowalla and Yelp2018.

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 [41] 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) [42] 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 [42].

Performance Comparison with Mixup-based Methods. We compare IMix with the strategy in [42] (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, 7, 9, 8]. 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).



(a) The generalization effect of normalized graph

(b) The generalization effect of activation function

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 raw features (Fig. 2, Fig. 6, Fig. 11, Fig. 12, Fig. 13). This may be that the node feature weakens the over-reliance on the graph structure. But it is undeniable that our conclusion 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.

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.

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%)

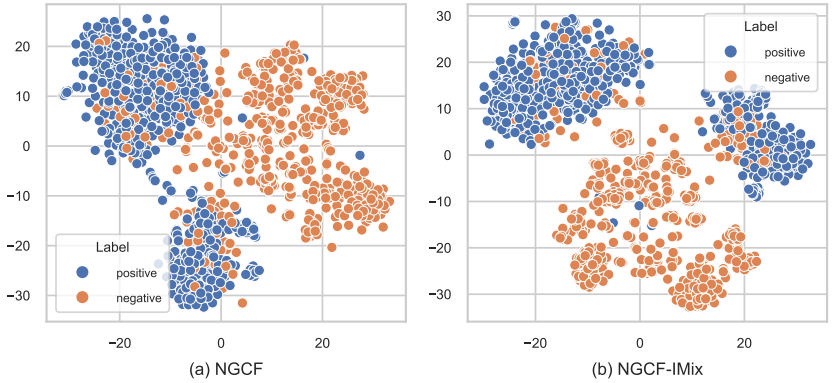


Figure 16: The t-SNE visualization of the embeddings of positive and negative items for a single user with or without IMix.

A.10 Visualization

To intuitively measure the quality of model learning, we use the t-SNE tool [43] 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 norm, and then we derive the Covering number of function class $\mathcal{F}$ using this lemma.

Lemma B.1 ([31]). *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.

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*

$$\log \mathcal{N}\left(\mathcal{F}|S, \epsilon, \|\cdot\|_2\right) \leq d^2 \log\left(1 + \frac{4(\gamma+1)MB_1\sqrt{nd}}{\epsilon}\right)\left(1 + \frac{4\gamma bMB_2\sqrt{nd}}{\epsilon}\right).$$

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

$$\begin{aligned}
&\log \mathcal{N}\left(\mathcal{F}|S, \epsilon, \|\cdot\|_2\right) \leq 2d^2 \log \frac{8M(\gamma+1)\sqrt{ndB_1B_2b}}{\epsilon}, \ where \\
&M = C_\phi \mathcal{C}^{2L-1}(B_u + B_i)^2 \frac{(2\mathcal{C})^L - 1}{2\mathcal{C} - 1}, \\
&\mathcal{C} = C_\phi[B_1 + \gamma(B_1 + B_2 b)].
\end{aligned}$$

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

Step 1: Max norm of user representation and item representation. We first bound the maximum spectral norm of user embedding 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

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

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

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

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 $\left\|\mathbf{u}_{l-1}\right\|_\infty \leq b$, we can get

$$\max_{u \in \mathcal{U}} \|\mathbf{u}_l\|_2 \leq C_\phi B_1 \left\|\mathbf{u}_{l-1}\right\|_2 + C_\phi \gamma \left(B_1 + B_2 b\right) \left\|\mathbf{w}_{l-1}\right\|_2. \tag{6}$$

Similarly,

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

Combining Eq. 6 and Eq. 7, we have

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

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

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

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

$$T_l = \max_{u\in\mathcal{U},w\in\mathcal{I}} (\|\mathbf{u}_l\|_2 + \|\mathbf{w}_l\|_2) \leq \mathcal{C}^l T_0 \leq \mathcal{C}^l (B_u + B_i). \tag{8}$$

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_w and $f_{w'}$.

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

$$\begin{aligned}
&\max_{u\in\mathcal{U}} \|\mathbf{u}_l - \mathbf{u}'_l\|_2 \\
&= \left\|\phi\left(\boldsymbol{W}_1\mathbf{u}_{l-1} + \mathbf{n}_l\right) - \phi\left(\boldsymbol{W}'_1\mathbf{u}'_{l-1} + \mathbf{n}'_l\right)\right\|_2 \\
&\leq C_\phi \left\|\left(\boldsymbol{W}_1\mathbf{u}_{l-1} + \mathbf{n}_l\right) - \left(\boldsymbol{W}'_1\mathbf{u}'_{l-1} + \mathbf{n}'_l\right)\right\|_2 \\
&\leq C_\phi \left(\left\|\boldsymbol{W}_1\mathbf{u}_{l-1} - \boldsymbol{W}'_1\mathbf{u}'_{l-1}\right\|_2 + \|\mathbf{n}_l - \mathbf{n}'_l\|_2\right).
\end{aligned} \tag{9}$$

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. 9.

$$\begin{aligned}
\left\|\boldsymbol{W}_1\mathbf{u}_{l-1} - \boldsymbol{W}'_1\mathbf{u}'_{l-1}\right\|_2 &\leq \|\boldsymbol{W}_1\mathbf{u}_{l-1} - \boldsymbol{W}'_1\mathbf{u}_{l-1}\|_2 + \left\|\boldsymbol{W}'_1\mathbf{u}_{l-1} - \boldsymbol{W}'_1\mathbf{u}'_{l-1}\right\|_2 \\
&\leq \|\boldsymbol{W}_1 - \boldsymbol{W}'_1\|_2 \|\mathbf{u}_{l-1}\|_2 + \|\boldsymbol{W}'_1\|_2 \left\|\mathbf{u}_{l-1} - \mathbf{u}'_{l-1}\right\|_2 \\
&\leq \|\boldsymbol{W}_1 - \boldsymbol{W}'_1\|_2 \|\mathbf{u}_{l-1}\|_2 + B_1 \left\|\mathbf{u}_{l-1} - \mathbf{u}'_{l-1}\right\|_2.
\end{aligned} \tag{10}$$

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

$$\begin{aligned}
&\|\mathbf{n}_l - \mathbf{n}'_l\|_2 \\
&\leq \sum_{w\in\mathcal{N}_u} |\tilde{a}_{uw}| \left\|\left(\boldsymbol{W}_1\mathbf{w}'_{l-1} - \boldsymbol{W}'_1\mathbf{w}'_{l-1}\right) + \left(\boldsymbol{W}_2\left(\mathbf{u}_{l-1}\odot\mathbf{w}_{l-1}\right) - \boldsymbol{W}'_2\left(\mathbf{u}'_{l-1}\odot\mathbf{w}'_{l-1}\right)\right)\right\|_2 \\
&\leq \sum_{w\in\mathcal{N}_u} |\tilde{a}_{uw}| \left(\left\|\boldsymbol{W}_1\mathbf{w}'_{l-1} - \boldsymbol{W}'_1\mathbf{w}'_{l-1}\right\|_2 + \left\|\boldsymbol{W}_2\left(\mathbf{u}_{l-1}\odot\mathbf{w}_{l-1}\right) - \boldsymbol{W}'_2\left(\mathbf{u}'_{l-1}\odot\mathbf{w}'_{l-1}\right)\right\|_2\right).
\end{aligned} \tag{11}$$

Similar to Eq. 10,

$$\left\|\boldsymbol{W}_1\mathbf{w}_{l-1} - \boldsymbol{W}'_1\mathbf{w}'_{l-1}\right\|_2 \leq \|\boldsymbol{W}_1 - \boldsymbol{W}'_1\|_2 \|\mathbf{w}_{l-1}\|_2 + B_1 \left\|\mathbf{w}_{l-1} - \mathbf{w}'_{l-1}\right\|_2. \tag{12}$$

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

$$\begin{aligned}
&\left\|\boldsymbol{W}_2\left(\mathbf{u}_{l-1}\odot\mathbf{w}_{l-1}\right) - \boldsymbol{W}'_2\left(\mathbf{u}'_{l-1}\odot\mathbf{w}'_{l-1}\right)\right\|_2 \\
&\leq \|\boldsymbol{W}_2 - \boldsymbol{W}'_2\|_2 \|\mathbf{u}_{l-1}\odot\mathbf{w}_{l-1}\|_2 + B_2 \left\|\mathbf{u}_{l-1}\odot\mathbf{w}_{l-1} - \mathbf{u}'_{l-1}\odot\mathbf{w}'_{l-1}\right\|_2 \\
&\leq \|\boldsymbol{W}_2 - \boldsymbol{W}'_2\|_2 \|\mathbf{u}_{l-1}\odot\mathbf{w}_{l-1}\|_2 + B_2 \left(\left\|\mathbf{u}_{l-1}\odot\left(\mathbf{w}_{l-1} - \mathbf{w}'_{l-1}\right)\right\|_2 + \left\|\left(\mathbf{u}_{l-1} - \mathbf{u}'_{l-1}\right)\odot\mathbf{w}'_{l-1}\right\|_2\right) \\
&\leq \|\boldsymbol{W}_2 - \boldsymbol{W}'_2\|_2 \|\mathbf{u}_{l-1}\|_2 \|\mathbf{w}_{l-1}\|_\infty + B_2 \left(\|\mathbf{u}_{l-1}\|_\infty \left\|\mathbf{w}_{l-1} - \mathbf{w}'_{l-1}\right\|_2 + \left\|\mathbf{u}_{l-1} - \mathbf{u}'_{l-1}\right\|_2 \left\|\mathbf{w}'_{l-1}\right\|_\infty\right).
\end{aligned} \tag{13}$$

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. 12 and Eq.13 into Eq. 11, 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

$$
\begin{aligned}
&\|\mathbf{n}_l - \mathbf{n}'_l\|_2 \\
&\leq \gamma \left(\|\boldsymbol{W}_1 - \boldsymbol{W}'_1\|_2 \|\mathbf{w}_{l-1}\|_2 + B_1 \left\|\mathbf{w}_{l-1} - \mathbf{w}'_{l-1}\right\|_2 \right) \\
&+ \gamma b \left(\|\boldsymbol{W}_2 - \boldsymbol{W}'_2\|_2 \|\mathbf{u}_{l-1}\|_2 + B_2 \left\|\mathbf{w}_{l-1} - \mathbf{w}'_{l-1}\right\|_2 + B_2 \left\|\mathbf{u}_{l-1} - \mathbf{u}'_{l-1}\right\|_2 \right).
\end{aligned} \tag{14}
$$

Combining Eq. 10 and Eq. 14, we have

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

Similarly, we have

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

Combining Eq. 15 and Eq. 16, we have

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

The last inequality is due to $C_\phi \left[(\gamma + 1) B_1 + 2\gamma b B_2 \right] \leq 2 C_\phi \left[(\gamma + 1) B_1 + \gamma b B_2 \right] = 2\mathcal{C}$. According to the norm bound for node representations completed in Eq. 8 and expanding the recursion, we can get the following result,

$$
\begin{aligned}
\Delta_l &= \max_{u \in \mathcal{U}} \|\mathbf{u}_l - \mathbf{u}'_l\|_2 + \max_{w \in \mathcal{I}} \|\mathbf{w}_l - \mathbf{w}'_l\|_2 \\
&\leq C_\phi C^{l-1} (B_u + B_i) \left[(\gamma + 1) \|\boldsymbol{W}_1 - \boldsymbol{W}'_1\|_2 + \gamma b \|\boldsymbol{W}_2 - \boldsymbol{W}'_2\|_2 \right] + 2\mathcal{C} \Delta_{l-1} \\
&\leq C_\phi C^{l-1} (B_u + B_i) \left[(\gamma + 1) \|\boldsymbol{W}_1 - \boldsymbol{W}'_1\|_2 + \gamma b \|\boldsymbol{W}_2 - \boldsymbol{W}'_2\|_2 \right] \frac{(2\mathcal{C})^l - 1}{2\mathcal{C} - 1}.
\end{aligned} \tag{17}
$$

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

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

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

(Case 1) Final layer:

$$
\begin{aligned}
\Lambda_L &\leq \max_{u \in \mathcal{U}, w \in \mathcal{I}} \left| \mathbf{u}_L^T \mathbf{w}_L - \mathbf{u}'_L{}^T \mathbf{w}'_L \right| \\
&\leq \left| \mathbf{u}_L^T \mathbf{w}_L - \mathbf{u}_L{}^T \mathbf{w}'_L \right| + \left| \mathbf{u}_L^T \mathbf{w}'_L - \mathbf{u}'_L{}^T \mathbf{w}'_L \right| \\
&\leq \|\mathbf{u}_L\|_2 \|\mathbf{w}_L - \mathbf{w}'_L\|_2 + \|\mathbf{u}_L - \mathbf{u}'_L\|_2 \|\mathbf{w}'_L\|_2.
\end{aligned} \tag{18}
$$

Similarly,

$$
\Lambda_L \leq \|\mathbf{u}'_L\|_2 \|\mathbf{w}_L - \mathbf{w}'_L\|_2 + \|\mathbf{u}_L - \mathbf{u}'_L\|_2 \|\mathbf{w}_L\|_2. \tag{19}
$$

Combining Eq. 18 and Eq. 19, we have

$$
\begin{aligned}
\Lambda_L &\leq \left(\|\mathbf{u}_L\|_2 + \|\mathbf{u}'_L\|_2 \right) \|\mathbf{w}_L - \mathbf{w}'_L\|_2 + \|\mathbf{u}_L - \mathbf{u}'_L\|_2 \left(\|\mathbf{w}_L\|_2 + \|\mathbf{w}'_L\|_2 \right) \\
&\leq \frac{1}{2} \left(\|\mathbf{u}_L\|_2 + \|\mathbf{u}'_L\|_2 + \|\mathbf{w}_L\|_2 + \|\mathbf{w}'_L\|_2 \right) \left(\|\mathbf{u}_L - \mathbf{u}'_L\|_2 + \|\mathbf{w}_L - \mathbf{w}'_L\|_2 \right) \\
&\leq T_L \Delta_L.
\end{aligned} \tag{20}
$$

(Case 2) Linear combination:

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

(Case 3) Concatenation:

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

For convenience, we only consider the outputs of the last layer as the final representations. Substituting Eq. 8 and Eq. 17 in to Eq. 20, we will derive the following change bound of recommendations predicted by two GCN-based models with different parameters.

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

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. Since $\|\boldsymbol{a}\|_2 \leq \sqrt{n}\|\boldsymbol{a}\|_\infty$ for any vector $\boldsymbol{a} \in \mathbb{R}^n$, let $M = C_\phi C^{2L-1}(B_u + B_i)^2 \frac{(2\mathcal{C})^L - 1}{2\mathcal{C} - 1}$, we have

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

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

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

According to Lemma B.1, we have

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

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

$$\log \mathcal{N}\left(\mathcal{F}, \epsilon, \|\cdot\|_2\right) \leq d^2 \log\left(1+\frac{4(\gamma+1)MB_1\sqrt{nd}}{\epsilon}\right)\left(1+\frac{4\gamma bMB_2\sqrt{nd}}{\epsilon}\right).$$

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

$$\begin{aligned}\log \mathcal{N}\left(\mathcal{F}, \epsilon, \|\cdot\|_2\right) &\leq d^2 \log \frac{64M^2nd\gamma(\gamma+1)B_1B_2b}{\epsilon^2} \\ &\leq 2d^2 \log \frac{8M\sqrt{ndB_1B_2b}(\gamma+1)}{\epsilon}.\end{aligned}$$

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 $S = \{x_i\}_{i=1}^n$ are sampled independently from $\mathcal{X}$ according to $\mathcal{D}$. Let $\mathcal{F}$ be a family of functions $\mathcal{X}$ to $\mathbb{R}$. The empirical Rademacher complexity of $\mathcal{F}$ with respect to the samples S is defined as*

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

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 ([44])**.** *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 $\mathcal{G}$,*

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

Lemma C.2 (Contraction Lemma [26])**.** *Let $\Phi : \mathbb{R} \to \mathbb{R}$ be an $L-lipschitz$ function, Then for a real-valued function set $\mathcal{G}$,*

$$\Re_m\left(\Phi\circ\mathcal{G}\right) \leq L\Re_m\left(\mathcal{G}\right).$$

Lemma C.3 (Extension of [32])**.** *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*

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

Then we provide the proof of Lemma 2.

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*

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

Proof. Let $l\circ\mathcal{H} = \{(z, z', y) \to l(y(f(z) - f(z'))) : f \in \mathcal{F}\}$ 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

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

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

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

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

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

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*

$$\sigma_i \triangleq \begin{cases} 1 & \textit{with probability} \quad p; \\ -1 & \textit{with probability} \quad p; \\ 0 & \textit{with probability} \quad 1-2p. \end{cases} \tag{21}$$

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 ([30])**.** *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 $\boldsymbol{v} \in \mathcal{V}$,*

$$\mathcal{L}_u(f) \leq \hat{\mathcal{L}}_m(f) + \Re_{m+u}(\mathcal{F}, \frac{mu}{(m+u)^2}) + B_{\max}c_0Q_1\sqrt{\min(m,u)} + B\sqrt{\frac{Q_2}{2}Q_1\ln\frac{1}{\delta}}.$$

Lemma D.2 (Contraction lemma for TRC [30])**.** *Let $\Phi : \mathbb{R} \to \mathbb{R}$ be an $L-lipschitz$ function, Then for a real-valued function set $\mathcal{G}$,*

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

Lemma D.3 (Massart's lemma for TRC)**.** *Let $\mathcal{A} \subseteq \mathbb{R}^m$ be a finite set, with $r = \max_{\mathbf{x}\in\mathcal{A}} \|\mathbf{x}\|_2$, 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. 21 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*

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

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*

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

Proof. Let $l \circ \mathcal{H} = \{(z, z', y) \to l(y(f(z) - f(z'))) : f \in \mathcal{F}\}$ 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

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

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

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

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

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

The proof is complete. □

D.1 Proof of Lemma D.3

Proof. The proof follows similar to the one for inductive Rademacher random variables [26].

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 [32].

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

$$\begin{aligned}
\hat{\mathfrak{R}}_n(\mathcal{F}) &= \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{f}\in\mathcal{F}} \sum_{j=1}^{n} \sigma_j f(x_j) \\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{f}\in\mathcal{F}} \left[\sum_{j=1}^{n} \sigma_j \left(f(x_j) - v_j^N[f]\right) + \sum_{i=1}^{N-1}\sum_{j=1}^{n} \sigma_j \left(v_j^i[f] - v_j^{i+1}[f]\right) - \sum_{j=1}^{n} \sigma_j v_j^1[f]\right] \\
&\leq \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{f}\in\mathcal{F}} \left[\sum_{j=1}^{n} \sigma_j \left(f(x_j) - v_j^N[f]\right)\right] + \frac{1}{n}\sum_{i=1}^{N-1} \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{f}\in\mathcal{F}} \left[\sum_{j=1}^{n} \sigma_j \left(v_j^i[f] - v_j^{i+1}[f]\right)\right] \\
&+ \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{f}\in\mathcal{F}} \left[\sum_{j=1}^{n} \sigma_j v_j^1[f]\right]. \qquad (22)
\end{aligned}$$

The last inequality is due to $\sigma_i \in \{-1, +1, 0\}$. For the first term in Eq. 22, 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. 22, let $W_i = \{v_j^i[f] - v_j^{i+1}[f] | f \in \mathcal{F}\}$, then we have

$$\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{f}\in\mathcal{F}} \left[\sum_{j=1}^{n} \sigma_j \left(v_j^i[f] - v_j^{i+1}[f]\right)\right] \leq \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{w}\in W_i} \left[\sum_{j=1}^{n} \sigma_j w_t\right]. \qquad (23)$$

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,

$$\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{w}\in W_i} \left[\sum_{j=1}^{n} \sigma_j w_t\right] \leq \sup_{\boldsymbol{w}\in W_i} \|\boldsymbol{w}\|_2 \sqrt{2\log|W_i|} \leq 2 \sup_{\boldsymbol{w}\in W_i} \|\boldsymbol{w}\|_2 \sqrt{\log|V_{i+1}|}. \qquad (24)$$

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. 23 and Eq. 24, 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. 22, 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.

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

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

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

The proof is complete. □

E Proof of Theorem 1

Theorem 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,*

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

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

$$\begin{aligned}\mathcal{L}_u(f) \leq& \hat{\mathcal{L}}_m(f) + \frac{24C_l}{\sqrt{m+u}}\mathcal{C}^{2L}(B_u+B_i)^2 d\sqrt{2\log(8(m+u)M(\gamma+1)\sqrt{dB_1B_2b})} \\ &+ \frac{4C_l\sqrt{2mu}}{(m+u)^2} + Bc_0Q_1\sqrt{\min(m,u)} + 2B\sqrt{\frac{Q_2}{2}Q_1\ln\frac{1}{\delta}}.\end{aligned}$$

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

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

Using Eq. 8, 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.

$$\mathcal{L}(f) \leq \hat{\mathcal{L}}_m(f) + 4C_l\inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{m}} + \frac{12}{m}\int_{\alpha}^{2e\sqrt{m}}\sqrt{\log\mathcal{N}\left(\mathcal{F}|S_1,\epsilon,\|\cdot\|_2\right)}d\epsilon\right) + 4B\sqrt{\frac{2\log 4/\delta}{m}} \tag{25}$$

where

$$\begin{aligned}&4C_l\inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{m}} + \frac{12}{m}\int_{\alpha}^{2e\sqrt{m}}\sqrt{\log\mathcal{N}\left(\mathcal{F}|S_1,\epsilon,\|\cdot\|_2\right)}d\epsilon\right) \\ &\leq 4C_l\inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{m}} + \frac{12d}{m}\int_{\alpha}^{2e\sqrt{m}}\sqrt{2\log\frac{8M(\gamma+1)\sqrt{mdB_1B_2b}}{\epsilon}}d\epsilon\right) \\ &\leq 4C_l\inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{m}} + \frac{24ed}{\sqrt{m}}\sqrt{2\log\frac{8M(\gamma+1)\sqrt{mdB_1B_2b}}{\alpha}}\right) \\ &= 4C_l\inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{m}} + \frac{6dC^{2L}(B_u+B_i)^2}{\sqrt{m}}\sqrt{2\log\frac{8M(\gamma+1)\sqrt{mdB_1B_2b}}{\alpha}}\right).\end{aligned} \tag{26}$$

Set $\alpha = \sqrt{\frac{1}{m}}$ and substitute Eq. 26 into Eq. 25, we have

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

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

$$\begin{aligned}\mathcal{L}_u(f) \leq & \hat{\mathcal{L}}_m(f) + \frac{24C_l}{\sqrt{m+u}}\mathcal{C}^{2L}(B_u+B_i)^2 d\sqrt{2\log(8(m+u)M(\gamma+1)\sqrt{dB_1B_2b})} \\ & + \frac{4C_l\sqrt{2mu}}{(m+u)^2} + Bc_0Q_1\sqrt{\min(m,u)} + 2B\sqrt{\frac{Q_2}{2}Q_1\ln\frac{1}{\delta}}.\end{aligned}$$

The proof is complete. □

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 $y = -1$. 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 $-y$. 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. If $y_i = y_j$, we compute the mixed embedding triplet $(\boldsymbol{e}_u, \widetilde{\boldsymbol{e}}_i, \widetilde{\boldsymbol{e}}_{i'}) = (\boldsymbol{e}_u, \boldsymbol{e}_i + \frac{1-\lambda}{\lambda}\boldsymbol{e}_j, \boldsymbol{e}_{i'} + \frac{1-\lambda}{\lambda}\boldsymbol{e}_{j'})$. Otherwise, we compute the mixed embedding triplet $(\boldsymbol{e}_u, \widetilde{\boldsymbol{e}}_i, \widetilde{\boldsymbol{e}}_{i'}) = (\boldsymbol{e}_u, \boldsymbol{e}_i + \frac{1-\lambda}{\lambda}\boldsymbol{e}_{j'}, \boldsymbol{e}_{i'} + \frac{1-\lambda}{\lambda}\boldsymbol{e}_j)$. For convenience, we only consider the embedding triplet with label $y = +1$ in main text.

F.1 Proof of Lemma 4

Lemma 4. *Suppose that* $\hat{\Sigma}_X = \frac{1}{m}\sum_{k=1}^{m}(\boldsymbol{e}_{i_k} - \boldsymbol{e}_{i'_k})(\boldsymbol{e}_{i_k} - \boldsymbol{e}_{i'_k})^T$ *and* $\mathcal{D}_x$ *is the empirical distribution of the items in training dataset* S_m*. If the loss function* l *is twice differentiable, then the IMix loss defined in Eq. 5 has the following approximation,*

$$\mathcal{L}_m^{mix}(S_m) \approx \mathcal{L}_m^{std}(S_m) + \frac{1}{m}\mathbb{E}_{\lambda\sim\mathcal{D}_\lambda}\left(\frac{(1-\lambda)^2}{\lambda^2}\right)\sum\nolimits_{k=1}^{m} l''(y_k\boldsymbol{e}_{u_k}^T(\boldsymbol{e}_{i_k} - \boldsymbol{e}_{i'_i}))\boldsymbol{e}_{u_k}^T\hat{\Sigma}_X\boldsymbol{e}_{u_k}.$$

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

$$\begin{aligned}\mathcal{L}_m^{IMix} &= \frac{1}{m}\sum\nolimits_{k=1}^{m} l(y_k\boldsymbol{e}_{u_k}^T(\widetilde{\boldsymbol{e}}_{i_k} - \widetilde{\boldsymbol{e}}_{i'_k})) \\ &= \frac{1}{m}\mathbb{E}_{\lambda\sim Beta(\alpha,\beta)}\mathbb{E}_{(j,j')\sim D_i} l(y_k\boldsymbol{e}_{u_k}^T(\widetilde{\boldsymbol{e}}_{i_k} - \widetilde{\boldsymbol{e}}_{i'_k})).\end{aligned}$$

Denote the randomness (of λ and (j, j')) by ξ, using the second-order Taylor expansion, we can approximate the IMix loss as follows,

$$l(y_k\boldsymbol{e}_{u_k}^T(\widetilde{\boldsymbol{e}}_{i_k} - \widetilde{\boldsymbol{e}}_{i'_k})) \approx l(y_k\boldsymbol{e}_{u_k}^T(\boldsymbol{e}_{i_k} - \boldsymbol{e}_{i'_k})) + \mathcal{R}_1 + \mathcal{R}_2.$$

For convenience, let $\omega(\lambda) = \frac{1-\lambda}{\lambda}$, $\eta_k = y_k\boldsymbol{e}_{u_k}^T(\boldsymbol{e}_{i_k} - \boldsymbol{e}_{i'_k})$. Then we have

$$\begin{aligned}\mathcal{R}_1 &= \mathbb{E}_\xi l'(\eta_k)\boldsymbol{e}_{u_k}^T(\widetilde{\boldsymbol{e}}_{i_k} - \boldsymbol{e}_{i_k})y_k - l'(\eta_k)\boldsymbol{e}_{u_k}^T(\widetilde{\boldsymbol{e}}_{i'_k} - \boldsymbol{e}_{i'_k})y_k \\ &= \mathbb{E}_\xi l'(\eta_k)\boldsymbol{e}_{u_k}^T[(\widetilde{\boldsymbol{e}}_{i_k} - \boldsymbol{e}_{i_k}) - (\widetilde{\boldsymbol{e}}_{i'_k} - \boldsymbol{e}_{i'_k})]y_k \\ &= \mathbb{E}_\lambda[\omega(\lambda)]\cdot l'(\eta_k)\boldsymbol{e}_{u_k}^T\hat{\boldsymbol{E}}_{u_k}.\end{aligned}$$

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

where $\hat{\boldsymbol{E}}_{u_k}=\frac{1}{m}\sum_{k=1}^{m}y_k(\boldsymbol{e}_{i_k}-\boldsymbol{e}_{i'_k})$, $\hat{\Sigma}_X=\frac{1}{m}\sum_{k=1}^{m}(\boldsymbol{e}_{i_k}-\boldsymbol{e}_{i'_k})(\boldsymbol{e}_{i_k}-\boldsymbol{e}_{i'_k})^T$. The last equality is due to the training set is symmetric and $(u,i,i')\in S_m$ for any $u\in D_u$ and $(i,i')\in D_i$. Then we have the following second-order Taylor approximation of the IMix loss.

$$
\mathcal{L}_m^{mix}\approx\mathcal{L}_m^{std}+\frac{1}{m}\sum\nolimits_{k=1}^{m}\left[\mathbb{E}_\lambda[\omega(\lambda)]l'(\eta_k)\boldsymbol{e}_{u_k}^T\hat{\boldsymbol{E}}_{u_k}+\frac{1}{2}\mathbb{E}_\lambda[\omega(\lambda)^2]l''(\eta_k)\boldsymbol{e}_{u_k}^T\hat{\Sigma}_X\boldsymbol{e}_{u_k}\right].
$$

The proof is complete. □

F.2 Proof of Remark 3

Lemma F.1 ([44])**.** *We first present a fundamental uniform convergence theory, then we establish the generalization bounds for via Rademacher complexity. 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 $\mathcal{G}$,*

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

Remark 3. Let $\tau>0$, $\boldsymbol{E}_u=\mathbb{E}y(\boldsymbol{e}_i-\boldsymbol{e}_{i'})$ and $\Sigma_X=\mathbb{E}\left[(\boldsymbol{e}_i-\boldsymbol{e}_{i'})(\boldsymbol{e}_i-\boldsymbol{e}_{i'})^T\right]$, we shed light upon the generalization bound by investigating the following function class:

$$
\mathcal{F}_\tau^{mix}=\{\mathcal{F}|\mathbb{E}_{(u,i,i')}\left[\mathbb{E}_\lambda[\omega(\lambda)]l'(\eta_k)\boldsymbol{e}_u^T\hat{\boldsymbol{E}}_u+\frac{1}{2}\mathbb{E}_\lambda[\omega(\lambda)^2]l''(\eta_k)\boldsymbol{e}_u^T\hat{\Sigma}_X\boldsymbol{e}_u\right]\le\tau\}.,
$$

Assuming that the loss function l satisfies $l'(x)\mathbb{E}_\lambda[\omega(\lambda)]\le l''(x)\mathbb{E}_\lambda[\omega(\lambda)^2]\le\rho,\forall x\in\mathbb{R}$ on the basis of the previous assumptions. Then with probability of at least $1-\delta$, for any function $f\in\mathcal{F}_\tau^{mix}$, we have,

$$
\mathcal{L}(f)\le\hat{\mathcal{L}}(f)+4C_l\sqrt{\frac{rank(\Sigma_X)\tau}{\rho}}+2B\sqrt{\frac{2\log 2/\delta}{m}}.
$$

For the general condition, we focus on the function class $\mathcal{F}_\tau^{std}=\{\mathcal{F}|\mathbb{E}_{(u,i,i')}\|\boldsymbol{e}_u\|_2^2+\|\boldsymbol{e}_i\|_2^2\le\tau\}$, then the generalization error bound is

$$
\mathcal{L}(f)\le\hat{\mathcal{L}}(f)+4C_l\sqrt{4\tau^2}+2B\sqrt{\frac{2\log 2/\delta}{m}}.
$$

Proof. Let $\widetilde{\boldsymbol{x}}_k=\Sigma_X^{\dagger/2}(\boldsymbol{e}_{i_k}-\boldsymbol{e}_{i'_k})$ and $\boldsymbol{v}_k=\Sigma_X^{1/2}\boldsymbol{e}_{u_k}$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{E}_\lambda[\omega(\lambda)]l'(\eta_k)\boldsymbol{e}_u^T\hat{\boldsymbol{E}}_u+\frac{1}{2}\mathbb{E}_\lambda[\omega(\lambda)^2]l''(\eta_k)\boldsymbol{e}_u^T\hat{\Sigma}_X\boldsymbol{e}_u\right]\le\tau \\
&\Rightarrow\mathbb{E}\left[\frac{1}{2}l''(\eta_k)\mathbb{E}_\lambda[\omega(\lambda)^2]\boldsymbol{e}_u^T\hat{\Sigma}_X\boldsymbol{e}_u+l'(\eta_k)\mathbb{E}_\lambda[\omega(\lambda)]\sqrt{\boldsymbol{e}_u^T\hat{\Sigma}_X\boldsymbol{e}_u}\right]\le\tau \\
&\Rightarrow\mathbb{E}\left[\boldsymbol{e}_u^T\hat{\Sigma}_X\boldsymbol{e}_u-2\sqrt{\boldsymbol{e}_u^T\hat{\Sigma}_X\boldsymbol{e}_u}\right]\le\frac{\tau}{\rho} \\
&\Rightarrow\mathbb{E}\left[\left(\sqrt{\boldsymbol{e}_u^T\hat{\Sigma}_X\boldsymbol{e}_u}-1\right)^2\right]\le\frac{\tau}{\rho}+1 \\
&\Rightarrow\mathbb{E}\left[\sqrt{\boldsymbol{e}_u^T\hat{\Sigma}_X\boldsymbol{e}_u}\right]\le 1+\sqrt{\frac{\tau}{\rho}+1} \qquad \text{Jensen's inequality.}
\end{aligned}
$$

Here $\rho \geq 2l''(x)\mathbb{E}_\lambda[\omega(\lambda)^2], \forall x \in \mathbb{R}$. We further assume that $\tau > \rho$, then we can bound the empirical Rademacher complexity as follows,

$$\begin{aligned}\hat{\Re}_{S_m}(\mathcal{F}) &\leq \mathbb{E}_{\boldsymbol{\sigma}}\left\{\sup_{\mathbb{E}\left[\sqrt{\boldsymbol{e}_{u_k}^T \Sigma_X \boldsymbol{e}_{u_k}}\right] \leq 2\sqrt{\frac{\tau}{\rho}}} \frac{1}{m}\sum_{k=1}^{m} \sigma_i l(y_k \boldsymbol{e}_{u_k}^T(\boldsymbol{e}_{i_k} - \boldsymbol{e}_{i'_k}))\right\} \\ &\leq C_l \mathbb{E}_{\boldsymbol{\sigma}}\left\{\sup_{\mathbb{E}\left[\sqrt{\boldsymbol{e}_{u_k}^T \Sigma_X \boldsymbol{e}_{u_k}}\right] \leq 2\sqrt{\frac{\tau}{\rho}}} \frac{1}{m}\sum_{k=1}^{m} \sigma_i \boldsymbol{e}_{u_k}^T(\boldsymbol{e}_{i_k} - \boldsymbol{e}_{i'_k})\right\}.\end{aligned}$$

The last inequality uses Lemma C.2. Then we have

$$\begin{aligned}\hat{\Re}_{S_m}(\mathcal{F}) &\leq C_l \mathbb{E}_{\boldsymbol{\sigma}}\left\{\sup_{\mathbb{E}\|\boldsymbol{v}_k\|_2 \leq 2\sqrt{\frac{\tau}{\rho}}} \frac{1}{m}\sum_{k=1}^{m} \sigma_i \boldsymbol{v}_k^T \widetilde{\boldsymbol{x}}_k\right\} \\ &\leq \frac{1}{m} \cdot C_l \cdot 2\sqrt{\frac{\tau}{\rho}} \cdot \mathbb{E}_{\boldsymbol{\sigma}}\left\|\sum_{k=1}^{m} \sigma_i \widetilde{\boldsymbol{x}}_k\right\|_2 \\ &\leq \frac{1}{m} \cdot C_l \cdot 2\sqrt{\frac{\tau}{\rho}} \cdot \sqrt{\mathbb{E}_{\boldsymbol{\sigma}}\left\|\sum_{k=1}^{m} \sigma_i \widetilde{\boldsymbol{x}}_k\right\|_2^2} \\ &\leq \frac{1}{m} \cdot C_l \cdot 2\sqrt{\frac{\tau}{\rho}} \cdot \sqrt{\sum_{k=1}^{m} \widetilde{\boldsymbol{x}}_k^T \widetilde{\boldsymbol{x}}_k}.\end{aligned}$$

Therefore, we have the following bounded Rademacher complexity.

$$\Re_m(\mathcal{F}) \leq \frac{1}{m}\mathbb{E}_{S_m} \cdot C_l \cdot 2\sqrt{\frac{\tau}{\rho}} \cdot \sqrt{\sum_{k=1}^{m} \widetilde{\boldsymbol{x}}_k^T \widetilde{\boldsymbol{x}}_k} \leq 2C_l\sqrt{\frac{rank(\Sigma_X)\tau}{\rho m}}.$$

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

$$\Re_m(\mathcal{F}) \leq 2C_l\sqrt{\tau^2}.$$

Finally, based on Lemma F.1, we can derive the corresponding generalization error bound using these two Rademcaher complexities.

The proof is complete. □