On Sampling Strategies for Neural Network-based Collaborative Filtering

Ting Chen
University of California, Los Angeles
Los Angeles, CA 90095
tingchen@cs.ucla.edu

Yizhou Sun
University of California, Los Angeles
Los Angeles, CA 90095
yzsun@cs.ucla.edu

Yue Shi∗
Yahoo! Research
Sunnyvale, CA 94089
yueshi@acm.org

Liangjie Hong
Etsy Inc.
Brooklyn, NY 11201
lhong@etsy.com

1 INTRODUCTION

Collaborative Filtering (CF) has been one of the most effective methods in recommender systems, and methods like matrix factorization [17, 18, 27] are widely adopted. However, one of its limitation is the dealing of “cold-start” problem, where there are few or no observed interactions for new users or items, such as in news recommendation. To overcome this problem, hybrid methods are proposed to incorporate side information [7, 25, 28], or item content information [11, 31] into the recommendation algorithm. Although these methods can deal with side information to some extent, they are not effective for extracting features in complicated data, such as image, audio and text. On the contrary, deep neural networks have been shown very powerful at extracting complicated features from those data automatically [15, 19]. Hence, it is natural to combine deep learning with traditional collaborative filtering for recommendation tasks, as seen in recent studies [1, 4, 32, 37].

In this work, we generalize several state-of-the-art neural network-based recommendation algorithms [1, 4, 30], and propose a more general framework that combines both collaborative filtering and deep neural networks in a unified fashion. The framework inherits the best of two worlds: (1) the power of collaborative filtering at capturing user preference via their interaction with items, and (2) that of deep neural networks at automatically extracting high-level features from content data. However, it also comes with a price. Traditional CF methods, such as sparse matrix factorization [17, 27], are usually fast to train, while the deep neural networks in general are much more computationally expensive [19]. Combining these two models in a new recommendation framework can easily increase computational cost by hundreds of times, thus require a new design of the training algorithm to make it more efficient.

We tackle the computational challenges by first establishing a connection between the loss functions and the user-item interaction bipartite graph. We realize the key issue when combining the CF and deep neural networks is in: the loss function terms are defined on links while major computation burdens are located at nodes. We call this type of loss functions “graph-based” loss functions, for which varied mini-batch sampling strategies can have different computational costs. Based on the insight, three novel sampling strategies are proposed, which can significantly improve the training efficiency of the proposed framework (up to ×30 times speedup in our experiments), as well as improving the recommendation performance. Theoretical analysis is also provided for both the computational cost and the convergence. We believe the study of sampling strategies have further implications on general graph-based loss functions, and would also enable more research under the neural network-based recommendation framework.

ACM Reference format:
https://doi.org/10.1145/3097983.3098202

∗Now at Facebook.

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

KDD ’17, August 13-17, 2017, Halifax, NS, Canada
© 2017 Copyright held by the owner/author(s). Publication rights licensed to Association for Computing Machinery.
ACM ISBN 978-1-4503-4887-4/17/08...$15.00
https://doi.org/10.1145/3097983.3098202
inefficient, as they do not take into account the node computations that can be potentially shared across links/data points.

Inspired by the connection established, we propose three novel sampling strategies for the general framework that can take coupled computation costs across user-item interactions into consideration. The first strategy is Stratified Sampling, which try to amortize costly node computation by partitioning the links into different groups based on nodes (called stratum), and sample links based on these groups. The second strategy is Negative Sharing, which is based on the observation that interaction/link computation is fast, so once a mini-batch of user-item tuples are sampled, we share the nodes for more links by creating additional negative links between nodes in the same batch. Both strategies have their pros and cons, and to keep their advantages while avoid their weakness, we form the third strategy by combining the above two strategies. Theoretical analysis of computational cost and convergence is also provided.

Our contributions can be summarized as follows.

- We propose a general hybrid recommendation framework (Neural Network-based Collaborative Filtering) combining CF and content-based methods with deep neural networks, which generalize several state-of-the-art approaches.
- We establish a connection between the loss functions and the user-item interaction graph, based on which, we propose sampling strategies that can significantly improve training efficiency (up to \( \times 30 \) times faster in our experiments) as well as the recommendation performance of the proposed framework.
- We provide both theoretical analysis and empirical experiments to demonstrate the superiority of the proposed methods.

2 A GENERAL FRAMEWORK FOR NEURAL NETWORK-BASED COLLABORATIVE FILTERING

In this section, we propose a general framework for neural network-based Collaborative Filtering that incorporates both interaction and content information.

2.1 Text Recommendation Problem

In this work, we use the text recommendation task \([1, 4, 31, 32]\) as an illustrative application for the proposed framework. However, the proposed framework can be applied to more scenarios such as music and video recommendations.

We use \( x_u \) and \( x_v \) to denote features of user \( u \) and item \( v \), respectively. In text recommendation setting, we set \( x_u \) to one-hot vector indicating \( u \)'s user id (i.e. a binary vector with only one at the \( u \)-th position)\(^1\), and \( x_v \) as the text sequence, i.e. \( x_v = (w_1, w_2, \ldots, w_L) \). A response matrix \( \hat{R} \) is used to denote the historical interactions between users and articles, where \( \hat{R}_{uv} \) indicates interaction between a user \( u \) and an article \( v \), such as “click-or-not” and “like-or-not”. Furthermore, we consider \( \hat{R} \) as implicit feedback in this work, which means only positive interactions are provided, and non-interactions are treated as negative feedback implicitly.

---

\(^1\)Other user profile features can be included, if available.
In this work, we consider two types of loss functions that can handle recommendation tasks with implicit feedback, namely, pointwise loss functions and pairwise loss functions. Pointwise loss functions have been applied to such problems in many existing work. In [1, 30, 32], mean square loss (MSE) has been applied where “negative terms” are weighted less. And skip-gram (SG) loss has been successfully utilized to learn robust word embedding [22].

These two loss functions are summarized in Table 1. Note that we use a weighted expectation term over all negative samples, which can be approximated with small number of samples. We can also abstract the pointwise loss functions into the following form:

\[
L_{\text{pointwise}} = \mathbb{E}_{u \sim P_d(u)} \left[ \mathbb{E}_{v \sim P_n(v|u)} c_{u,v} L^+(u,v|\theta) + \mathbb{E}_{v' \sim P_n(v'|u)} c_{u,v'} L^-(u,v'|\theta) \right]
\]

(1)

where \(P_d\) is (empirical) data distribution, \(P_n\) is user-defined negative data distribution, \(c\) is user-defined weights for the different user-item pairs, \(\theta\) denotes the set of all parameters, \(L^+ (u,v|\theta)\) denotes the loss function on a single positive pair \((u,v)\), and \(L^- (u,v|\theta)\) denotes the loss on a single negative pair. Generally speaking, given a user \(u\), pointwise loss function encourages her score with positive items \(v\), and discourage her score with negative items \(v'\).

When it comes to ranking problem as commonly seen in implicit feedback setting, some have argued that the pairwise loss would be advantageous [26, 33], as pairwise loss encourages ranking of positive items above negative items for the given user. Different from pointwise counterparts, pairwise loss functions are defined on a triplet of \((u,v,v')\), where \(v\) is a positive item and \(v'\) is a negative item to the user \(u\). Table 1 also gives two instances of such loss functions used in existing papers [26, 33] (with \(\gamma\) being the predefined “margin” parameter). We can also abstract pairwise loss functions by the following form:

\[
L_{\text{pairwise}} = \mathbb{E}_{u \sim P_d(u)} \mathbb{E}_{v \sim P_n(v|u)} c_{u,v} L^+(u,v,v'|\theta) - \mathbb{E}_{v' \sim P_n(v'|u)} c_{u,v'} L^-(u,v,v'|\theta)
\]

(2)

where the notations are similarly defined as in Eq. 1 and \(L^+ (u,v,v'|\theta)\) denotes the loss function on the triplet \((u,v,v')\).

### 2.4 Stochastic Gradient Descent Training and Computational Challenges

To train the model, we use stochastic gradient descent based algorithms [3, 16], which are widely used for training matrix factorization and neural networks. The main flow of the training algorithm is summarized in Algorithm 1. By adopting the functional embedding with (deep) neural networks, we can increase the power of the model, but it also comes with a cost. Figure 2 shows the training time (for CiteULike data) with different item functions \(g(\cdot)\), namely linear embedding taking item id as feature (equivalent to conventional MF), CNN-based content embedding, and RNN/LSTM-based content embedding. We see orders of magnitude increase of training time for the latter two embedding functions, which may create barriers to adopt models under this framework.

![Figure 2: Model training time per epoch with different types of item functions (in log-scale).](image)

**Algorithm 1 Standard model training procedure**

```text
while not converged do
    // mini-batch sampling
do
    draw a mini-batch of user-item tuples \((u, v)^2\)
    // forward pass
    compute \(f(x_u), g(x_v)\) and their interaction \(\tilde{f}_u^T g_v\)
    compute the loss function \(L\)
    // backward pass
    compute gradients and apply SGD updates
end while
```

\(2^\text{draw a mini-batch of user-item triplets \((u, v, v')\) if a pairwise loss function is adopted.}\)
3 MINI-BATCH SAMPLING STRATEGIES FOR EFFICIENT MODEL TRAINING

In this section, we propose and discuss different sampling strategies that can improve the efficiency of the model training.

3.1 Computational Cost in a Graph View

Before the discussion of different sampling strategies, we motivate our readers by first making a connection between the loss functions and the bipartite graph of user-item interactions. In the loss functions laid out before, we observed that each loss function term in Eq. 1, namely, \( L(u, v) \), involves a pair of user and item, which corresponds to a link in their interaction graph. And two types of links corresponding to two types of loss terms in the loss functions, i.e., positive links/terms and negative links/terms. Similar analysis holds for pairwise loss in Eq. 2, though there are slight differences as each single loss function corresponds to a pair of links with opposite signs on the graph. We can also establish a correspondence between user/item functions and nodes in the graph, i.e., \( f(u) \) to user node \( u \) and \( g(v) \) to item node \( v \). The connection is illustrated in Figure 3. Since the loss functions are defined over the links, we name them “graph-based” loss functions to emphasize the connection.

The key observation for graph-based loss functions is that: the loss functions are defined over links, but the major computational burden are located at nodes (due to the use of costly \( g(\cdot) \) function). Since each node is associated with multiple links, which are corresponding to multiple loss function terms, the computational costs of loss functions over links are coupled (as they may share the same nodes) when using mini-batch based SGD. Hence, varied sampling strategies yield different computational costs. For example, when we put links connected to the same node together in a mini-batch, the computational cost can be lowered as there are fewer \( g(\cdot) \) to compute\(^{3}\). This is in great contrast to conventional optimization problems, where each loss function term dose not couple with others in terms of computation cost.

\(^{3}\)This holds for both forward and backward computation. For the latter, the gradient from different links can be aggregated before back-propagating to \( g(\cdot) \).

3.2 Existing Mini-Batch Sampling Strategies

In standard SGD sampler, (positive) data samples are drawn uniformly at random for gradient computation. Due to the appearance of negative samples, we draw negative samples from some predefined probability distribution, i.e. \( (u', v') \sim P_n(u', v') \). We call this approach “IID Sampling”, since each positive link is dependently and identical distributed, and the same holds for negative links (with a different distribution).

Many existing algorithms with graph-based loss functions [1, 22, 29] adopt the “Negative Sampling” strategy, in which \( k \) negative samples are drawn whenever a positive example is drawn. The negative samples are sampled based on the positive ones by replacing the items in the positive samples. This is illustrated in Algorithm 2 and Figure 4(a).

**Algorithm 2 Negative Sampling** [1, 21, 29]

```
Require: number of positive links in a mini-batch \( b \), number of negative links per positive one: \( k \)
for each of \( b \) positive links do
    draw \( k \) negative links by replacing true item \( v \) with \( v' \sim P_n(v') \)
end for
```

The IID Sampling strategy dose not take into account the property of graph-based loss functions, since samples are completely independent of each other. Hence, the computational cost in a single mini-batch cannot be amortized across different samples, leading to very extensive computations with (deep) neural networks. The Negative Sampling does not really help, since the item function computation cost \( t_g \) is the dominant one. To be more specific, consider a mini-batch with \( b(1 + k) \) links sampled by IID Sampling or Negative Sampling, we have to conduct item based \( g(\cdot) \) computation \( b(1 + k) \) times, since items in a mini-batch are likely to be non-overlapping with sufficient large item sets.

3.3 The Proposed Sampling Strategies

3.3.1 Stratified Sampling (by Items)

Motivated by the connection between the loss functions and the bipartite interaction graph as shown in Figure 3, we propose to sample links that share nodes, in particular those with high computational cost (i.e. \( t_g \) for item function \( g(\cdot) \) in our case). By doing so, the computational cost within a mini-batch can be amortized, since fewer costly functions are computed (in both forward and backward propagations).

In order to achieve this, we (conceptually) partition the links, which correspond to loss function terms, into strata. A stratum in the strata is a set of links on the bipartite graph sharing the same source or destination node. Instead of drawing links directly for training, we will first draw stratum and then draw both positive and negative links. Since we want each stratum to share the same item, we can directly draw an item and then sample its links. The details are given in Algorithm 3 and illustrated in Figure 4(b).

Compared to Negative Sampling in Algorithm 2, there are several differences: (1) Stratified Sampling can be based on either item or user, but in the negative sampling only negative items are drawn; and (2) each node in stratified sampling can be associated with
This can be achieved by creating a complete bipartite graph for a mini-batch by adding negative links between all non-interaction pairs between users and items. Using this strategy, we can draw additional negative samples. Since the main computational cost of training is on the node computation and the node set is fixed given the batch of positive links, we can share the nodes for negative links without increasing much of computational burdens. Based on this idea, Algorithm 4 summarizes an extremely simple sampling procedure, and it is illustrated in Figure 4(c).

Since Negative Sharing avoids sampling \( k \) negative links, it only contains \( b \) items while in Negative Sampling contains \( b(1 + k) \) items. So it can provide \( (1 + k) \) times speedup compared to Negative Sampling (assuming \( t_g \gg t_f, t_i \)), and total interaction cost is still insignificant. Given the batch size \( b \) is usually larger than \( k \) (e.g., \( b = 512, k = 20 \) in our experiments), much more negative links (e.g. \( 512 \times 511 \)) will also be considered, this is helpful for both faster convergence and better performance, which is shown in our experiments. However, as the number of negative samples increases, the performance and the convergence will not be improved linearly, diminishing return is expected.

### 3.3.3 Stratified Sampling with Negative Sharing

The two strategies above can both reduce the computational cost by smarter sampling of the mini-batch. However, they both have weakness: Stratified Sampling cannot deal with pairwise loss and it is still dependent on the number of negative examples \( k \), and Negative Sharing introduces a lot of negative samples which may be unnecessary due to diminishing return.

The good news is, the two sampling strategies are proposed from different perspectives, and combining them together can preserve their advantages while avoid their weakness. This leads to the Stratified Sampling with Negative Sharing, which can be applied to both pointwise and pairwise loss functions, and it can have flexible ratio between positive and negative samples (i.e. more positive links given the same negative links compared to Negative Sharing). To do so, basically we sample positive links according to Stratified Sampling, and then sample/create negative links by treating non-interactions as negative links. The details are given in Algorithm 5 and illustrated in Figure 4(d).

---

**Algorithm 3** Stratified Sampling (by Items)

Require: number of positive links in a mini-batch: \( b \), number of positive links per stratum: \( s \), number of negative links per positive one: \( k \)

repeat
  draw an item \( u \sim P_d(v) \)
  draw \( s \) positive users \( \{u\} \) of \( v \) uniformly at random
  draw \( k \times s \) negative users \( \{u'\} \sim P_d(u') \)
until a mini-batch of \( b \) positive links are sampled

**Algorithm 4** Negative Sharing

Require: number of positive links in a mini-batch: \( b \)
draw \( b \) positive user-item pairs \( \{(u, v)\} \) uniformly at random
construct negative pairs by connecting non-linked users and items in the batch
Computationally, Stratified Sampling with Negative Sharing only involve $b/s$ item nodes in a mini-batch, so it can provide the same $s(1 + k)$ times speedup over Negative Sampling as Stratified Sampling (by Items) does, but it will utilize much more negative links compared to Negative Sampling. For example, in our experiments with $b = 512, s = 4$, we have 127 negative links per positive one, much larger than $k = 10$ in Negative Sampling, and only requires 1/4 times of $g(\cdot)$ computations compared to Negative Sampling.

3.3.4 Implementation Details. When the negative/noise distribution $P_n$ is not unigram\(^4\), we need to adjust the loss function in order to make sure the stochastic gradient is unbiased. For pointwise loss, each of the negative term is adjusted by multiplying a weight of $P_n(v)/P_n(v')$, for pairwise loss, each term based on a triplet of $(u, v, v')$ is adjusted by multiplying a weight of $P_n(v')/P_n(v)$ where $v'$ is the sampled negative item.

Instead of sampling, we prefer to use shuffling as much as we can, which produces unbiased samples while yielding zero variance. This can be a useful trick for achieving better performance when the number of drawn samples are not large enough for each loss terms. For IID and Negative Sampling, this can be easily done for positive links by simply shuffling them. As for the Stratified Sampling (w/wo. Negative Sharing), instead of shuffling the positive links directly, we shuffle the randomly formed strata (where each stratum contains roughly a single item)\(^5\). All other necessary sampling operations required are sampling from discrete distributions, which can be done in $O(1)$ with Alias method.

In Negative Sharing (w/wo. Stratified Sampling), We can compute the user-item interactions with more efficient operator, i.e. replacing the vector dot product between each pair of $(f, g)$ with matrix multiplication between $(F, G)$, where $F = [f_{u1}, \ldots, f_{un}]$, $G = [g_{v1}, \ldots, g_{vnm}]$. Since matrix multiplication is higher in BLAS level than vector multiplication [14], even we increase the number of interactions, with medium matrix size (e.g. 1000 x 1000) it does not affect the computational cost much in practice.

3.4 Computational Cost and Convergence Analysis

Here we provide a summary for the computational cost for different sampling strategies discussed above, and also analyze their convergence. Two aspects that can lead to speedup are analyzed: (1) the computational cost for a mini-batch, i.e. per iteration, and (2) the number of iterations required to reach some referenced loss.

3.4.1 Computational Cost. To fairly compare different sampling strategies, we fix the same number of positive links in each of the mini-batch, which correspond to the positive terms in the loss function. Table 2 shows the computational cost of different sampling strategies for a given mini-batch. Since $t_g \gg t_f, t_i$ in practice, we approximate the theoretical speedup per iteration by comparing the number of $t_g$ computation. We can see that the proposed sampling strategies can provide $(1 + k)$, by Negative Sharing, or $s(1 + k)$, by Stratified Sampling (w/wo. Negative Sharing), times speedup for each iteration compared to IID Sampling or Negative Sampling. As for the number of iterations required to reach a reference loss, it is related to number of negative samples utilized, which is analyzed below.

3.4.2 Convergence Analysis. We want to make sure the SGD training under the proposed sampling strategies can converge correctly. The necessary condition for this to hold is the stochastic gradient estimator has to be unbiased, which leads us to the following lemma.

**Lemma 1.** (unbiased stochastic gradient) Under sampling Algorithm 2, 3, 4, and 5, we have $E_t[\nabla L_t(\theta^t)] = \nabla L(\theta^*)$. In other words, the stochastic mini-batch gradient equals to true gradient in expectation.

This holds for both pointwise loss and pairwise loss. It is guaranteed since we draw samples stochastically and re-weight certain samples accordingly. The detailed proof can be found in the supplementary material.

Given this lemma, we can further analyze the convergence behavior of the proposed sampling behaviors. Due to the highly non-linear and non-convex functions composed by (deep) neural networks, the convergence rate is usually difficult to analyze. So we show the SGD with the proposed sampling strategies follow a local convergence bound (similar to [10, 24]).

**Proposition 1.** (local convergence) Suppose $L$ has $\sigma$-bounded gradient; let $\eta_t = \eta = c/\sqrt{T}$ where $c = \sqrt{\frac{2L(0^0) - L(0^*)}{L(0^*)}}$, and $0^*$ is the minimizer to $L$. Then, the following holds for the proposed sampling strategies given in Algorithm 2, 3, 4, 5

\[
\min_{0 \leq t \leq T-1} \mathbb{E}[||\nabla L(\theta^t)||^2] \leq \frac{2(L(0^0) - L(0^*))}{T} \sigma
\]

The detailed proof is also given in the supplementary material. Furthermore, utilizing more negative links in each mini-batch can lower the expected stochastic gradient variance. As shown in [35, 36], the reduction of variance can lead to faster convergence. This suggests that Negative Sharing (w/wo. Stratified Sampling) has better convergence than the Stratified Sampling (by Items).

4 EXPERIMENTS

4.1 Data Sets

Two real-world text recommendation data sets are used for the experiments. The first data set CiteULike, collected from CiteULike.org, is provided in [31]. The CiteULike data set contains users bookmarking papers, where each paper is associated with a title
and an abstract. The second data set is a random subset of Yahoo! News data set, which contains users clicking on news presented at Yahoo! There are 5,551 users and 16,980 items, and total of 204,986 positive interactions in CiteULike data. As for Yahoo! News data, there are 10,000 users, 58,579 items and 515,503 interactions.

Following [4], we select a portion (20%) of items to form the pool of test items. All user interactions with those test items are held-out during training, only the remaining user-item interactions are used as training data, which simulates the scenarios for recommending newly-emerged text articles.

### 4.2 Experimental Settings

The main purpose of experiments is to compare the efficiency and effectiveness of our proposed sampling strategies against existing ones. So we mainly compare Stratified Sampling, Negative Sharing, and Stratified Sampling with Negative Sharing, against IID sampling and Negative Sampling. It is worth noting that several existing state-of-the-art models [1, 4, 30] are special cases of our framework (e.g. using MSE-loss/Log-loss with CNN or RNN), so they are compared to other loss functions under our framework.

#### Evaluation Metrics

For recommendation performance, we follow [1, 32] and use recall@M. As pointed out in [32], the precision is not a suitable performance measure since non interactions may be due to (1) the user is not interested in the item, or (2) the user does not pay attention to its existence. More specifically, for each user, we rank candidate test items based on the predicted scores, and then compute recall@M based on the list. Finally the recall@M is averaged over all users.

As for the computational cost, we mainly measure it in three dimensions: the training time for each iteration (or epoch equivalently, since batch size is fixed for all methods), the number of iterations needed to reach a referenced loss, and the total amount of computation time needed to reach the same loss. In our experiments, we use the smallest loss obtained by IID sampling in the maximum 30 epochs as referenced loss. Noted that all time measure mentioned here is in Wall Time.

#### Parameter Settings

The key parameters are tuned with validation set, while others are simply set to reasonable values. We adopt Adam [16] as the stochastic optimizer. We use the same batch size $b = 512$ for all sampling strategies, we use the number of positive link per sampled stratum $s = 4$, learning rate is set to 0.001 for MSE-loss, and 0.01 for others. $y$ is set to 0.1 for Hinge-loss, and 10 for others. $\lambda$ is set to 8 for MSE-loss, and 128 for others. We set number of negative examples $k = 10$ for convolutional neural networks, and $k = 5$ for RNN/LSTM due to the GPU memory limit. All experiments are run with Titan X GPUs. We use unigram noise/negative distribution.

For CNN, we adopt the structure similar in [15], and use 50 filters with filter size of 3. Regularization is added using both weight decay on user embedding and dropout on item embedding. For RNN, we use LSTM [12] with 50 hidden units. For both models, the dimensions of user and word embedding are set to 50. Early stop is utilized, and the experiments are run to maximum 30 epochs.

### 4.3 Speedup Under Different Sampling Strategies

Table 3 breaks down the speedup into (1) speedup for training on a given mini-batch, (2) number of iterations (to reach referenced cost) speedup, and (3) the total speedup, which is product of the first two. Different strategies are compared against IID Sampling. It is shown that Negative Sampling has similar computational cost as IID Sampling, which fits our projection. All three proposed sampling strategies can significantly reduce the computation cost within a mini-batch. Moreover, the Negative Sharing and Stratified Sampling with Negative Sharing can further improve the convergence w.r.t. Table 3: Comparisons of speedup for different sampling strategies against IID Sampling: per iteration, # of iteration, and total speedup.

<table>
<thead>
<tr>
<th>Model Sampling</th>
<th>CiteULike</th>
<th>News</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Per it.</td>
<td># of it.</td>
</tr>
<tr>
<td>Negative CNN</td>
<td>1.02</td>
<td>1.00</td>
</tr>
<tr>
<td>Stratified CNN</td>
<td>8.83</td>
<td>0.97</td>
</tr>
<tr>
<td>N.S.</td>
<td>8.42</td>
<td>2.31</td>
</tr>
<tr>
<td>Strat. w. N.S.</td>
<td>15.53</td>
<td>1.87</td>
</tr>
<tr>
<td>Negative LSTM</td>
<td>0.99</td>
<td>0.96</td>
</tr>
<tr>
<td>Stratified</td>
<td>3.1</td>
<td>0.77</td>
</tr>
<tr>
<td>N.S.</td>
<td>2.87</td>
<td>2.45</td>
</tr>
<tr>
<td>Strat. w. N.S.</td>
<td>3.4</td>
<td>2.22</td>
</tr>
</tbody>
</table>
the number of iterations, which demonstrates the benefit of using larger number of negative examples.

Figure 5 and 6 shows the convergence curves of both loss and test performance for different sampling strategies (with CNN + SG-loss). In both figures, we measure progress every epoch, which is equivalent to a fixed number of iterations since all methods have the same batch size $b$. In both figures, we can observe mainly two types of convergences behavior. Firstly, in terms of number of iterations, Negative Sharing (w./wo. Stratified Sampling) converge fastest, which attributes to the number of negative samples used. Secondly, in terms of wall time, Negative Sharing (w./wo. Stratified Sampling) and Stratified Sampling (by Items) are all significantly faster than baseline sampling strategies, i.e. IID Sampling and Negative Sampling. It is also interesting to see that that overfitting occurs earlier as convergence speeds up, which does no harm as early stopping can be used.

For Stratified Sampling (w./wo. negative sharing), the number of positive links per stratum $s$ can also play a role to improve speedup as we analyzed before. As shown in Figure 7, the convergence time as well as recommendation performance can both be improved with a reasonable $s$, such as 4 or 8 in our case.

4.4 Recommendation Performance Under Different Sampling Strategies

It is shown in above experiments that the proposed sampling strategies are significantly faster than the baselines. But we would also like to further access the recommendation performance by adopting the proposed strategies.

Table 4 compares the proposed sampling strategies with CNN/RNN models and four loss functions (both pointwise and pairwise). We can see that IID Sampling, Negative Sampling and Stratified Sampling (by Items) have similar recommendation performances, which is expected since they all utilize same amount of negative links. For Negative Sharing and Stratified Sampling with Negative Sharing, since there are much more negative samples utilized, their performances are significantly better. We also observe that the current recommendation models based on MSE-loss [1, 30] can be improved by others such as SG-loss and pairwise loss functions [4].

To further investigate the superior performance brought by Negative Sharing. We study the number of negative examples $k$ and
Recall@50 for different sampling strategies under different models and losses.

<table>
<thead>
<tr>
<th>Model</th>
<th>Sampling</th>
<th>CiteULike</th>
<th>News</th>
</tr>
</thead>
<tbody>
<tr>
<td>IID</td>
<td></td>
<td>0.4746</td>
<td>0.4437</td>
</tr>
<tr>
<td>Negative</td>
<td></td>
<td>0.4725</td>
<td>0.4408</td>
</tr>
<tr>
<td>CNN</td>
<td>Stratified</td>
<td>0.4761</td>
<td>0.4394</td>
</tr>
<tr>
<td>Negative Sharing</td>
<td></td>
<td>0.4866</td>
<td>0.4423</td>
</tr>
<tr>
<td>Stratified with N.S.</td>
<td></td>
<td><strong>0.4890</strong></td>
<td><strong>0.4535</strong></td>
</tr>
<tr>
<td>IID</td>
<td></td>
<td>0.4479</td>
<td>0.4718</td>
</tr>
<tr>
<td>Negative</td>
<td></td>
<td>0.4371</td>
<td>0.4668</td>
</tr>
<tr>
<td>LSTM</td>
<td>Stratified</td>
<td>0.4344</td>
<td>0.4685</td>
</tr>
<tr>
<td>Negative Sharing</td>
<td></td>
<td>0.4629</td>
<td>0.4839</td>
</tr>
<tr>
<td>Stratified with N.S.</td>
<td></td>
<td><strong>0.4742</strong></td>
<td><strong>0.4877</strong></td>
</tr>
</tbody>
</table>

Figure 8: The number of negatives VS performances.

The convergence performance. Figure 8 shows the test performance against various k. As shown in the figure, we observe a clear diminishing return in the improvement of performance. However, the performance seems still increasing even we use 20 negative examples, which explains why our proposed method with negative sharing can result in better performance.

5 RELATED WORK

Collaborative filtering [18] has been one of the most effective methods in recommender systems, and methods like matrix factorization [17, 27] are widely adopted. While many papers focus on the explicit feedback setting such as rating prediction, implicit feedback is found in many real-world scenarios and studied by many papers as well [13, 23, 26]. Although collaborative filtering techniques are powerful, they suffer from the so-called "cold-start" problem since side/content information is not well leveraged. To address the issue and improve performance, hybrid methods are proposed to incorporate side information [5, 7, 25, 28, 32], as well as content information [4, 11, 31, 32].

Deep Neural Networks (DNNs) have been showing extraordinary abilities to extract high-level features from raw data, such as video, audio, and text [8, 15, 34]. Compared to traditional feature detectors, such as SIFT and n-grams, DNNs and other embedding methods [5, 6, 29] can automatically extract better features that produce higher performance in various tasks. To leverage the extraordinary feature extraction or content understanding abilities of DNNs for recommender systems, recent efforts are made in combining collaborative filtering and neural networks [1, 4, 30, 32]. [32] adopts autoencoder for extracting item-side text information for article recommendation, [1] adopts RNN/GRU to better understand the text content. [4] proposes to use CNN and pairwise loss functions, and also incorporate unsupervised text embedding. The general functional embedding framework in this work subsumes existing models [1, 4, 30].

Stochastic Gradient Descent [3] and its variants [16] have been widely adopted in training machine learning models, including neural networks. Samples are drawn uniformly at random (IID) so that the stochastic gradient vector equals to the true gradient in expectation. In the setting where negative examples are overwhelming, such as in word embedding (e.g., Word2Vec [22]) and network embedding (e.g., LINE [29]) tasks, negative sampling is utilized. Recent efforts have been made to improve SGD convergence by (1) reducing the variance of stochastic gradient estimator, or (2) distributing the training over multiple workers. Several sampling techniques, such as stratified sampling [35] and importance sampling [36] are proposed to achieve the variance reduction. Different from their work, we improve sampling strategies in SGD by reducing the computational cost of a mini-batch while preserving, or even increasing, the number of data points in the mini-batch. Sampling techniques are also studied in [9, 39] to distribute the computation of matrix factorization, their objectives in sampling strategy design are reducing the parameter overlapping and cache miss. We also find that the idea of sharing negative examples is exploited to speed up word embedding training in [14].

6 DISCUSSIONS

While it is discussed under content-based collaborative filtering problem in this work, the study of sampling strategies for "graph-based" loss functions have further implications. The IID sampling strategy is simple and popular for SGD-based training, since the
loss function terms usually do not share the common computations. So no matter how a mini-batch is formed, it almost bears the same amount of computation. This assumption is shattered by models that are defined under graph structure, with applications in social and knowledge graph mining [2], image caption ranking [20], and so on. For those scenarios, we believe better sampling strategies can result in much faster training than that with IID sampling.

We would also like to point out limitations of our work. The first one is the setting of implicit feedback. When the problem is posed under explicit feedback, Negative Sharing can be less effective since the constructed negative samples may not overlap with the explicit negative ones. The second one is the assumption of efficient computation for interaction functions. When we use neural networks as interaction functions, we may need to consider constructing negative samples more wisely for Negative Sharing as it will also come with a noticeable cost.

7 CONCLUSIONS AND FUTURE WORK

In this work, we propose a hybrid recommendation framework, combining conventional collaborative filtering with (deep) neural networks. The framework generalizes several existing state-of-the-art recommendation models, and embody potentially more powerful ones. To overcome the high computational cost brought by combining “cheap” CF with “expensive” NN, we first establish the connection between the loss functions and the user-item interaction bipartite graph, and then point out the computational costs can vary with different sampling strategies. Based on this insight, we propose three novel sampling strategies that can significantly improve the training efficiency of the proposed framework, as well as its performance.

In the future, there are some promising directions. Firstly, based on the efficient sampling techniques of this paper, we can more efficiently study different neural networks and auxiliary information for building hybrid recommendation models. Secondly, we can also study the effects of negative sampling distributions and its affect on the design of more efficient sampling strategies. Lastly but not least, it would also be interesting to apply our sampling strategies in a distributed training environments where multi-GPUs and multi-machines are considered.

ACKNOWLEDGEMENTS

The authors would like to thank anonymous reviewers for helpful suggestions. The authors would also like to thank NVIDIA for the donation of one Titan X GPU. This work is partially supported by NSF CAREER #1741634.

REFERENCES

SUPPLEMENTARY MATERIAL

A PROOFS

Here we give the proofs for both the lemma and the proposition introduced in the main paper. For brevity, throughout we assume by default the loss function $L$ is the pointwise loss of Eq. (1) in the main paper. Proofs are only given for the pointwise loss, but it can be similarly derived for the pairwise loss. We start by first introducing some definitions.

**Definition 1.** A function $f$ is $L$-smooth if there is a constant $L$ such that
\[
\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|
\]

Such an assumption is very common in the analysis of first-order methods. In the following proof, we assume any loss functions $L$ is $L$-smooth.

**Property 1.** (Quadratic Upper Bound) A $L$-smooth function $f$ has the following property
\[
f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|y - x\|^2
\]

**Definition 2.** We say a function $f$ has a $\sigma$-bounded gradient if $\|\nabla f_i(\theta)\|^2 \leq \sigma$ for all $i \in [n]$ and any $\theta \in \mathbb{R}^d$.

For each training iteration, we first sample a mini-batch of links (denoted by $B$) of both positive links ($B^+$) and negative links ($B^-$), according to the sampling algorithm (one of the Algorithm 2, 3, 4, 5), and then the stochastic gradient is computed and applied to the parameters as follows:

\[
\theta^{t+1} = \theta^t - \frac{\eta_t}{m} \sum_{(u, v) \in B} c_{uv}^+ \nabla L^+(\theta | u, v) + \frac{\eta_t}{n} \sum_{(u, v) \in B} c_{uv}^- \nabla L^-(\theta | u, v)
\]

Here we use $L^+(\theta | u, v)$ to denote the gradient of loss function $L^+\theta$ given a pair of $(u, v)$. And $m, n$ are the number of positive and negative links in the batch $B$, respectively.

**Lemma 1.** (unbiased stochastic gradient) Under sampling Algorithm 2, 3, 4, 5, we have $E_B[\nabla L_B(\theta^t)] = \nabla L(\theta^t)$. In other words, the stochastic mini-batch gradient equals to true gradient in expectation.

**Proof.** Below we prove this lemma for each for the sampling Algorithm. For completeness, we also show the proof for Uniform Sampling as follows. The main idea is show the expectation of stochastic gradient computed in a randomly formed mini-batch equal to the true gradient of objective in Eq. 1.

**IID Sampling.** The positive links in the batch $B$ are i.i.d. samples from $P_d(u, v)$ (i.e. drawn uniformly at random from all positive links), and the negative links in $B$ are i.i.d. samples from $P_d(u)P_d(v)$, thus we have
\[
E_B[\nabla L_B(\theta^t)] = \frac{1}{m} \sum_{i=1}^m E_{(u, v) \sim P_d(u, v)}[c_{uv}^+ \nabla L^+(\theta | u, v)] + \frac{1}{n} \sum_{i=1}^n E_{(u, v) \sim P_d(u)P_d(v)}[c_{uv}^- \nabla L^-(\theta | u, v)]
\]
\[
E_{u \sim P_d(u)}\left[ E_{v \sim P_d(v | u)}[c_{uv}^+ \nabla L^+(\theta | u, v)] + E_{v' \sim P_d(v')}[c_{uv}^- \nabla L^-(\theta | u, v')]ight]
\]
\[
= \nabla L(\theta^t)
\]

The first equality is due to the definition of sampling procedure, the second equality is due to the definition of expectation, and the final equality is due to the definition of pointwise loss function in Eq. 1.

**Negative Sampling.** In Negative Sampling, we have batch $B$ consists of i.i.d. samples of $m$ positive links, and conditioning on each positive link, $k$ negative links are sampled by replacing items in the same i.i.d. manner. Positive links are sampled from $P_d(u, v)$, and negative items are sampled from $P_d(v')$, thus we have
\[
E_B[\nabla L_B(\theta^t)] = \frac{1}{m} \sum_{i=1}^m E_{(u, v) \sim P_d(u, v)} \frac{1}{k} \sum_{j=1}^k E_{v' \sim P_d(v')}[c_{uv}^+ \nabla L^+(\theta | u, v) + c_{uv}^- \nabla L^-(\theta | u, v')]
\]
\[
= E_{u \sim P_d(u)}\left[ E_{v \sim P_d(v | u)}[c_{uv}^+ \nabla L^+(\theta | u, v)] + E_{v' \sim P_d(v')}[c_{uv}^- \nabla L^-(\theta | u, v')]ight]
\]
\[
= \nabla L(\theta^t)
\]

The first equality is due to the definition of sampling procedure, and the second equality is due to the properties of joint probability distribution and expectation.
Stratified Sampling (by Items). In Stratified Sampling (by Items), a batch $B$ consists of links samples drawn in two steps: (1) draw an item $v \sim P_d(v)$, and (2) draw positive users $u \sim P_d(u|v)$ and negative users $u' \sim P_d(u)$ respectively. Additionally, negative terms are also re-weighted, thus we have

$$E_B[\nabla L_B(\theta^t)] = E_{u \sim P_d(u)} \left[ \frac{1}{m} \sum_{i=1}^{m} E_{u \sim P_d(u)} [c^+_{uv} \nabla L_+^+(\theta|u,v)] + \frac{1}{n} \sum_{j=1}^{n} E_{u \sim P_d(u)} [c^-_{uv} N_{P_d(v)} \nabla L_-^-(\theta|u,v)] \right]$$

$$= E_{(u,v) \sim P_d(u,v)} [c^+_{uv} \nabla L_+^+(\theta|u,v)] + E_{(u,v) \sim P_d(u,v)} [c^-_{uv} N_{P_d(v)} \nabla L_-^-(\theta|u,v)]$$

$$= E_{(u,v) \sim P_d(u,v)} [c^+_{uv} \nabla L_+^+(\theta|u,v)] + E_{(u,v) \sim P_d(u,v)} [c^-_{uv} \nabla L_-^-(\theta|u,v)]$$

$$= \nabla L(\theta^t)$$

The first equality is due to the definition of sampling procedure, and the second, third and the forth equality is due to the properties of joint probability distribution and expectation.

Negative Sharing. In Negative Sharing, we only draw positive links uniformly at random (i.e. $(u,v) \sim P_d(u,v)$), while constructing negative links from sharing the items in the batch. So the batch $B$ we use for computing gradient consists of both $m$ positive links and $m(m-1)$ negative links.

Although we do not draw negative links directly, we can still calculate their probability according to the probability distribution from which we draw the positive links. So a pair of constructed negative link in the batch is drawn from $(u,v) \sim P_d(u,v) = P_d(v) P_d(u|v)$. Additionally, negative terms are also re-weighted, we have

$$E_B[\nabla L_B(\theta^t)] = \frac{1}{m} \sum_{i=1}^{m} E_{u \sim P_d(u)} [c^+_{uv} \nabla L_+^+(\theta|u,v)] + \frac{1}{m(m-1)} \sum_{j=1}^{m(m-1)} E_{(u,v) \sim P_d(u,v)} [c^-_{uv} N_{P_d(v)} \nabla L_-^-(\theta|u,v)]$$

$$= E_{u \sim P_d(u)} \left[ \sum_{i=1}^{m} E_{u \sim P_d(u)} [c^+_{uv} \nabla L_+^+(\theta|u,v)] + E_{v \sim P_d(v)} [c^-_{uv} \nabla L_-^-(\theta|u,v)] \right]$$

$$= \nabla L(\theta^t)$$

The first equality is due to the definition of sampling procedure, and the second equality is due to the properties of joint probability distribution and expectation.

Stratified Sampling with Negative Sharing. Under this setting, we follow a two-step sampling procedure: (1) draw an item $v \sim P_d(v)$, and (2) draw positive users $u \sim P_d(u|v)$ and negative links are constructed from independently drawn items in the same batch. So the batch $B$ consists of $m$ positive links and $n$ negative links.

We can use the same method as in Negative Sharing to calculate the probability of sampled negative links, which is also $(u,v) \sim P_d(u,v)$. Again, negative terms are re-weighted, thus we have

$$E_B[\nabla L_B(\theta^t)] = \frac{1}{m} \sum_{i=1}^{m} E_{u \sim P_d(u)} [c^+_{uv} \nabla L_+^+(\theta|u,v)] + \frac{1}{n} \sum_{j=1}^{n} E_{u \sim P_d(u)} [c^-_{uv} N_{P_d(v)} \nabla L_-^-(\theta|u,v)]$$

$$= E_{(u,v) \sim P_d(u,v)} [c^+_{uv} \nabla L_+^+(\theta|u,v)] + E_{(u,v) \sim P_d(u,v)} [c^-_{uv} \nabla L_-^-(\theta|u,v)]$$

$$= \nabla L(\theta^t)$$

The first equality is due to the definition of sampling procedure, and the second, third and fourth equality is due to the properties of joint probability distribution and expectation.

Proposition 1. Suppose $L$ has $\sigma$-bounded gradient; let $\eta_t = \eta = c/\sqrt{T}$ where $c = \sqrt{2(L(0^0) - L(0^t))}$, and $\theta^t$ is the minimizer to $L$. Then, the following holds for the sampling strategies given in Algorithm 2, 3, 4, 5

$$\min_{0 \leq t \leq T-1} \mathbb{E} \left[ \left\| \nabla L(\theta^t) \right\|^2 \right] \leq \frac{2(L(0^0) - L(0^t))}{T} \Delta$$
Proof. With the property of $L$-smooth function $\mathcal{L}$, we have

$$\mathbb{E}[\mathcal{L}(\theta^{t+1})] \leq \mathbb{E}[\mathcal{L}(\theta^t)] + \langle \nabla \mathcal{L}(\theta^t), \theta^{t+1} - \theta^t \rangle + \frac{L}{2} \|\theta^{t+1} - \theta^t\|^2$$

(9)

By applying the stochastic update equation, lemma 1, i.e. $\mathbb{E}[\nabla \mathcal{L}(\theta^t)] = \nabla \mathcal{L}(\theta^t)$, we have

$$\mathbb{E}[\langle \nabla \mathcal{L}(\theta^t), \theta^{t+1} - \theta^t \rangle] + \frac{L}{2} \|\theta^{t+1} - \theta^t\|^2 \leq \eta_t \mathbb{E}[\|\nabla \mathcal{L}(\theta^t)\|^2] + \frac{L\eta_t^2}{2} \mathbb{E}[\|\nabla \mathcal{L}(\theta^t)\|^2]$$

(10)

Combining results in Eq. ?? and ??, with assumption that the function $\mathcal{L}$ is $\sigma$-bounded, we have

$$\mathbb{E}[\mathcal{L}(\theta^{t+1})] \leq \mathbb{E}[\mathcal{L}(\theta^t)] + \eta_t \mathbb{E}[\|\nabla \mathcal{L}(\theta^t)\|^2] + \frac{L\eta_t^2}{2b} \sigma^2$$

Rearranging the above equation we obtain

$$\mathbb{E}[\|\nabla \mathcal{L}(\theta^t)\|^2] \leq \frac{1}{\eta_t} \left( \mathbb{E}[\mathcal{L}(\theta^t)] - \mathbb{E}[\mathcal{L}(\theta^{t+1})] + \frac{L\eta_t}{2b} \sigma^2 \right)$$

(11)

By summing Eq. ?? from $t = 0$ to $T - 1$ and setting $\eta = c/\sqrt{T}$, we have

$$\min_t \mathbb{E}[\|\nabla \mathcal{L}(\theta^t)\|^2] \leq \frac{1}{c\sqrt{T}} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla \mathcal{L}(\theta^t)\|^2] \leq \frac{1}{c\sqrt{T}} \left( \mathbb{E}[\mathcal{L}(\theta^0)] - \mathbb{E}[\mathcal{L}(\theta^*)] + \frac{Lc}{2\sqrt{T}} \sigma^2 \right)$$

(12)

By setting

$$\epsilon = \sqrt{\frac{2(\mathbb{E}[\mathcal{L}(\theta^0)] - \mathbb{E}[\mathcal{L}(\theta^*)])}{L\sigma^2}}$$

We obtain the desired result.

\[\square\]

B  VECTOR DOT PRODUCT VERSUS MATRIX MULTIPLICATION

Here we provide some empirical evidence for the computation time difference of replacing vector dot product with matrix multiplication. Since vector dot product can be batched by element-wise matrix multiplication followed by summing over each row. We compare two operations between two square matrices of size $n$: (1) element-wise matrix multiplication, and (2) matrix multiplication. A straightforward implementation of the former has algorithmic complexity of $O(n^2)$, while the latter has $O(n^3)$. However, modern computation devices such as GPUs are better optimized for the latter, so when the matrix size is relatively small, their computation time can be quite similar. This is demonstrated in Figure ???. In our choice of batch size and embedding dimension, $n \ll 1000$, so the computation time is comparable. Furthermore, $t_i \ll t_g$, so even several times increase would also be ignorable.
C FUNCTIONAL EMBEDDING VERSUS FUNCTIONAL REGULARIZATION

In this work we propose a functional embedding framework, in which the embedding of a user/item is obtained by some function such as neural networks. We notice another approach is to penalize the distance between user/item embedding and the function output (instead of equate them directly as in functional embedding), which we refer as functional regularization, and it is used in [32]. More specifically, functional regularization emits following form of loss function:

\[ L(h_u, h_v) + \lambda \| h_u - f(x_u) \|^2 \]

Here we point out its main issue, which does not appear in Functional Embedding. In order to equate the two embedding vectors, we need to increase \( \lambda \). However, setting large \( \lambda \) will slow down the training progress under coordinate descent. The gradient w.r.t. \( h_u \) is

\[ \nabla_{h_u} L(h_u, h_v) + \frac{\lambda}{2} (h_u - f(x_u)) \]

so when \( \lambda \) is large, \( h_u^{t+1} \rightarrow f(x_u) \), which means \( h_u \) cannot be effectively updated by interaction information.