Examining the Bias of In-Batch Sampling in Similarity Learning with Two-Tower Models

Yaxu Liu
National Taiwan University
d08944045@ntu.edu.tw
Mohamed bin Zayed University
of Artificial Intelligence
yaxu.liu@mbzuai.ac.ae

Li-Chung Lin
National Taiwan University
r08922141@ntu.edu.tw

Chih-Jen Lin
National Taiwan University
cjlin@csie.ntu.edu.tw
Mohamed bin Zayed University
of Artificial Intelligence
chihjen.lin@mbzuai.ac.ae

Abstract

Two-tower models are widely used for applications involving learning similarities between pairs of entities, such as user-item pairs in recommender systems. These models are commonly trained using stochastic gradient methods. However, uniformly sampling data often leads to problematic batches that lack positive pairs, especially when positives are a minority of the dataset. Instead, a strategy known as in-batch sampling is widely adopted to ensure the presence of positive pairs and the training efficiency. Nevertheless, in-batch sampling introduces its own issues, such as mistaking positives for negatives and oversampling popular pairs, resulting in significant performance degradation. In this work, we provide the first systematic analysis of these issues, showing that they all arise from the inconsistency between the expected objective under in-batch sampling and the full-data objective. We refer to this inconsistency as the bias of in-batch sampling. To validate our analysis, we design an unbiased batch loss and conduct rigorous experiments to compare unbiased and biased losses. The results provide strong empirical confirmation of our theoretical findings.

1 Introduction

Numerous applications involve the problem of learning similarities between pairs of entities, labeled as the left and right entities in this work. For instance, in recommender systems, the similarity between a user-item pair might indicate the user's preference for the item. Similarly, in multi-label classification, an instance is assigned to a label if the instance-label pair exhibits high similarity. A prevalent method for learning such similarities involves training two-tower models (Yuan et al., 2021) to represent each entity. In this approach, we hope to map similar left and right entities (i.e., positive pairs) to vectors close to each other in the embedding space. Conversely, we map dissimilar entities (i.e., negative pairs) to vectors that are far apart.

Stochastic gradient methods are commonly adopted to train two-tower models, especially those with deep neural networks. These methods randomly choose a subset (i.e., a batch) of the whole dataset (including all positive and negative pairs) to estimate the gradient for model updates at each training step. It is typically assumed that uniformly sampled batches can ensure that, in expectation, the stochastic gradients align with the one computed over the entire dataset.

However, because positive pairs are rare in typical datasets for similarity learning, selecting positive pairs by uniform sampling is challenging. The resulting batches often contain no positives, making the stochastic gradients noisy and slowing down the training process. To mitigate this issue, the sampling process in each training step must ensure that positive pairs are included, thereby improving the quality of the stochastic gradients and speeding up training.

A widely adopted strategy to ensure some positive pairs in a batch is *in-batch sampling* by Gillick et al. (2019) and Karpukhin et al. (2020). This sampling strategy first selects a subset of positive pairs. Then, a batch is defined as the set of all pairwise combinations between the left and right entities of the selected positive pairs. In this batch, the selected positives remain labeled

Proceedings of the 29$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

as positives, while all other pairwise combinations are treated as negatives, regardless of their actual ground truth labels. With in-batch sampling, we can guarantee that every batch contains positives.

While in-batch sampling addresses the shortcomings of uniform sampling, it introduces challenges. The first issue is misclassifying positive pairs as negative pairs (Gupta et al., 2024), while the second is over-sampling specific pairs (Yi et al., 2019). Notably, Gupta et al. (2024) reported that smaller batch sizes significantly degrade model performance. These previous works, however, only partially explained the above issues.

In this work, we show that all the aforementioned issues on model performance share a common root cause. In-batch sampling is inherently non-uniform, leading to a learning objective different from that of uniform sampling. We refer to this difference as the bias of in-batch sampling and provide a theoretical analysis of how the bias causes all the issues.

To validate our theoretical findings, we introduce an unbiased batch loss and compare it with the biased batch losses in our experiments. However, the experiments are particularly challenging, since we must ensure that performance differences are solely due to the losses but not other factors such as model convergence. Unfortunately, the learning problem with two-tower models is non-convex, and stochastic gradient methods may diverge or reach different points. To ensure stable convergence, we must adopt a prohibitively time-consuming setting by training models with sufficiently small learning rates. Such a careful and rigorous design helps to ensure the validity of our experiments.

We list the main notations in Appendix A and summarize our main contributions as follows:

- We explain the unique aspects of data sampling in training two-tower models and summarize three sampling strategies, including the in-batch sampling.
- We provide the first systematic analysis of in-batch sampling by examining its expected objective, thereby revealing the bias as the root cause of several issues reported in prior works.
- By introducing an unbiased batch loss, we conduct rigorous experiments with carefully controlled settings to validate our analysis.

The experimental code is available at `https://www.csie.ntu.edu.tw/~cjlin/papers/inbatch_bias/`.

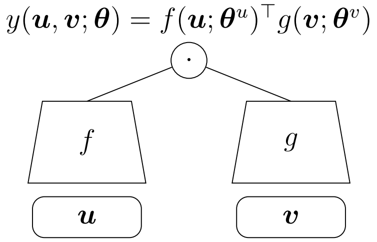

$y(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\theta}) = f(\boldsymbol{u}; \boldsymbol{\theta}^u)^\top g(\boldsymbol{v}; \boldsymbol{\theta}^v)$
$\cdot$
f
g
$\boldsymbol{u}$
$\boldsymbol{v}$

Figure 1: An illustration of the two-tower model, adopted from Yuan et al. (2021).

2 Learning Two-Tower Models

A similarity learning problem corresponds to a binary label matrix $\boldsymbol{Y} \in \{0,1\}^{m \times n}$ representing the similarity of m left entities and n right entities, where 1 means two entities are similar (i.e., they form a positive pair) while 0 does not. The set of pairs labelled with 1s is denoted as $\mathbb{O} \subseteq [m] \times [n]$, where $[m]$ and $[n]$ are integer sets $\{1, \cdots, m\}$ and $\{1, \cdots, n\}$. Typically, $|\mathbb{O}| \ll mn$, and $\boldsymbol{Y}$ is very sparse.

Given feature vectors of a left entity $\boldsymbol{u} \in \mathbb{R}^{D_u}$ and a right entity $\boldsymbol{v} \in \mathbb{R}^{D_v}$, our goal is to find a similarity function y with parameters $\boldsymbol{\theta}$, $y(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\theta}) : \mathbb{R}^{D_u} \times \mathbb{R}^{D_v} \to \mathbb{R}$, such that

$$y(\boldsymbol{u}_i, \boldsymbol{v}_j; \boldsymbol{\theta}) \approx Y_{ij}, \forall (i,j) \in [m] \times [n]. \tag{1}$$

A common structure of y is the two-tower model,

$$y(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\theta}) = y(\boldsymbol{u}, \boldsymbol{v}; [\boldsymbol{\theta}^u; \boldsymbol{\theta}^v]) = f(\boldsymbol{u}; \boldsymbol{\theta}^u)^\top g(\boldsymbol{v}; \boldsymbol{\theta}^v), \tag{2}$$

where $f : \mathbb{R}^{D_u} \to \mathbb{R}^k$ and $g : \mathbb{R}^{D_v} \to \mathbb{R}^k$ with parameters $\boldsymbol{\theta}^u$ and $\boldsymbol{\theta}^v$ encode the left and right entities into embedding vectors, as illustrated in Figure 1. While (2) may seem restrictive on the structure of y, a recent comparative study in recommender systems (Rendle et al., 2020) supports using the dot product to compute the similarity between the embeddings of two entities.

By considering all mn pairs, we learn $\boldsymbol{\theta}$ through solving

$$\min_{\boldsymbol{\theta}} \quad L(\boldsymbol{\theta}), \tag{3}$$

where

$$L(\boldsymbol{\theta}) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \ell(Y_{ij}, \hat{Y}_{ij}) \tag{4}$$

is the objective function and $\hat{Y}_{ij} = y(\boldsymbol{u}_i, \boldsymbol{v}_j; \boldsymbol{\theta})$. For $\ell(\cdot, \cdot)$, we use a differentiable point-wise loss function convex in $\hat{Y}_{ij}$,[1] like the square loss $\ell(a,b) = \frac{1}{2}(a-b)^2$.

[1] A point-wise loss function evaluates the error for each data point (i,j) independently (Rendle, 2022). Another popular loss is the Softmax loss, but we do not explore it in this work since its sampled estimators are proved to be definitely biased (Lin et al., 2025). Please check Appendix F.2 for more details.

Through defining $\ell_{ij}^{+} := \ell(1, \hat{Y}_{ij})$ and $\ell_{ij}^{-} := \ell(0, \hat{Y}_{ij})$ for subsequent discussion, we reformulate (4) as follows:

$$\begin{aligned}(4) =& \frac{1}{mn}\left(\sum_{(i,j)\in\mathbb{O}} \ell_{ij}^{+} + \sum_{(i,j)\notin\mathbb{O}} \ell_{ij}^{-}\right)\\ =& \frac{1}{mn}\left(\sum_{(i,j)\in\mathbb{O}} \ell_{ij}^{+} - \sum_{(i,j)\in\mathbb{O}} \ell_{ij}^{-} + \sum_{i=1}^{m}\sum_{j=1}^{n} \ell_{ij}^{-}\right). \quad (5)\end{aligned}$$

2.1 Stochastic Gradient Method for Training Two-Tower Models

The stochastic gradient (SG) method is widely used in machine learning, particularly in the case with limited computational resources, as illustrated by Bottou et al. (2018). In empirical risk minimization (ERM), the objective is typically defined as the average of training losses over the entire dataset. For example, in multi-class classification, the objective is

$$\frac{1}{N}\sum_{i=1}^{N} \text{Loss}(\text{Prediction\_of\_Instance}_i, \text{Label}_i), \quad (6)$$

where N is the total number of instances. The SG method approximates the objective (6) by selecting a subset of instances, known as a batch $\mathbb{B}$. Then, the stochastic gradient is computed as the gradient of the following batch loss.

$$\frac{1}{|\mathbb{B}|}\sum_{i\in\mathbb{B}} \text{Loss}(\text{Prediction\_of\_Instance}_i, \text{Label}_i). \quad (7)$$

This method is efficient as the stochastic gradient involves only a subset of the data.

However, in our case, the objective (4) involves a double summation, $\sum_{i=1}^{m}\sum_{j=1}^{n}$, which deviates from the standard form of the ERM in (6). For example, in extending (6) for multi-class classification to multi-label scenarios, we have m instances and n labels. Each label $j \in [n]$ has a feature vector $\boldsymbol{v}_j$, and instance $\boldsymbol{u}_i$ can be associated with multiple labels. From the similarity learning problem in (5), if we select a batch $\mathbb{B}$ consisting of some (i, j) pairs from the full set $[m] \times [n]$, the batch loss $\hat{L}(\boldsymbol{\theta})$ is:

$$\hat{L}(\boldsymbol{\theta}) := \frac{1}{|\mathbb{B}|}\sum_{(i,j)\in\mathbb{B}} \ell(Y_{ij}, \hat{Y}_{ij}). \quad (8)$$

However, the double summation structure in (4) and the properties of similarity learning problems cause a uniformly sampled $\mathbb{B}$ to be no longer suitable. For example, many datasets for two-tower models contain very few positives, so the selected $\mathbb{B}$ may not contain any positive pairs. Thus, the sampling strategy for two-tower models requires additional considerations.

3 Data Sampling for Two-Tower Models

In this section, we first show that data sampling on only one of the double summations mentioned above can be impractical due to computational and memory constraints. We then review different sampling strategies applied to both summations, including the in-batch sampling. In the end, we discuss issues of in-batch sampling identified in recent studies.

3.1 Sampling on Only One Summation of $\sum_{i=1}^{m}\sum_{j=1}^{n}$

We discussed in Section 2.1 that the learning problem of two-tower models is an extension of the traditional ERM in (6), as the objective involves $\sum_{i=1}^{m}\sum_{j=1}^{n}$ instead of $\sum_{i=1}^{N}$. Thus, a natural sampling scheme similar to the batch loss in (7) is to sample along the first summation on instances while retaining all n labels. Then, a batch $\mathbb{B}$ must be at least of size n, and it is often computationally infeasible when n is very large. Some works, like Gupta et al. (2024), have provided a memory-efficient implementation to perform distributed training. Nevertheless, retaining all labels in a batch remains challenging. Alternatively, we may sample labels while maintaining all m instances. This setting may be even less feasible, as the number of instances is usually greater than the number of labels.

3.2 Sampling on Both Summations of $\sum_{i=1}^{m}\sum_{j=1}^{n}$

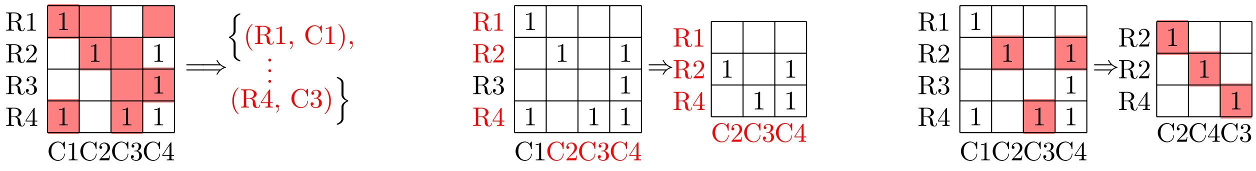
Instead of sampling from only one summation, the double summations in (4) allow us to sample from both (i.e., the row and column dimensions of the label matrix $\boldsymbol{Y}$), offering the flexibility to manage computational and memory costs. This section summarizes three strategies that sample data from both summations of $\sum_{i=1}^{m}\sum_{j=1}^{n}$.[2]

3.2.1 Naive Sampling

Naive sampling is to uniformly sample pairs from $[m] \times [n]$ to form a batch $\mathbb{B}$, as shown in Figure 2a. While conceptually simple, this strategy is ineffective in practice for two reasons.

1. **High computation cost:** For large values of m and n, in a small batch the sampled pairs generally have distinct left (and right) entities. That is, a left (or right) entity seldom occurs in two or more sampled pairs. Thus, to compute (8), we must encode $|\mathbb{B}|$ left

[2] Details of corresponding batch losses are in Appendix B.



(a) **Naive Sampling** (b) **Cross Sampling** (c) **In-Batch Sampling**

Figure 2: Three strategies for $m = n = 4$ and $|\mathbb{B}| = 9$. The left side of each sub-figure is the full label matrix $\boldsymbol{Y}$, while the right side is a batch. Squares marked with 1s are the positive pairs, while the others are negative pairs. Red indicates what is being sampled. **Naive Sampling**: The batch consists of uniformly selected samples from all pairs; see squares in red. **Cross Sampling**: The batch consists of a grid formed from sampled rows and columns; see rows and columns in red. **In-Batch Sampling**: The batch consists of samples of positive pairs and pairs sharing the same rows and columns as the positive pairs. Only the sampled positive pairs in the red squares are treated as positive. The other sampled pairs are always treated as negative, even if their original labels are positive.

and $|\mathbb{B}|$ right entities into embedding vectors with f and g, respectively. The cost on a batch is

$$|\mathbb{B}|F(f) + |\mathbb{B}|F(g), \tag{9}$$

where $F(f)$ and $F(g)$ are computation costs of f and g respectively. This sampling incurs a higher computational cost than the cross and in-batch sampling discussed later.

2. **Batches without positives:** Due to the high sparsity of $\boldsymbol{Y}$, sampled batches often lack positives, especially when the batch size is small. Gradients calculated on batches without positives are poor approximations of the gradient calculated on the full data, thereby slowing the model training.

3.2.2 Cross Sampling

To address the high computation cost in naive sampling, a sampling method used by Yuan et al. (2021), which we call cross sampling, aims to reduce the number of computations of $f(\cdot)$ and $g(\cdot)$. We first sample $\sqrt{|\mathbb{B}|}$ left and $\sqrt{|\mathbb{B}|}$ right entities. Then, a batch includes all $|\mathbb{B}|$ combinations of left and right entities, as shown in the example in Figure 2b.

With this strategy, we reduce the computation cost from (9) to

$$\sqrt{|\mathbb{B}|}F(f) + \sqrt{|\mathbb{B}|}F(g). \tag{10}$$

However, this strategy still suffers from the problem that a batch may contain no positive pairs.

3.2.3 In-Batch Sampling

To have a batch with positives, previous works, like Karpukhin et al. (2020), Yang et al. (2020), and Gupta et al. (2024), proposed in-batch sampling. As illustrated in Figure 2c, this method proceeds in two steps. First, we uniformly sample a subset $\hat{\mathbb{O}}$ from the set of positive pairs $\mathbb{O}$. Then, we construct a batch using all $|\hat{\mathbb{O}}|^2$ combinations of left and right entities appearing in $\hat{\mathbb{O}}$, keeping duplicated entities. The batch of all $|\hat{\mathbb{O}}|^2$ combinations is organized into a square matrix with $\hat{\mathbb{O}}$ being the diagonal elements. We consider the diagonal elements positive while the off-diagonal elements negative.

Taking $|\hat{\mathbb{O}}| = \sqrt{|\mathbb{B}|}$, we obtain a batch with size $|\mathbb{B}|$. The computation cost is the same as in (10). However, contrary to cross sampling, since we choose $\hat{\mathbb{O}}$ from $\mathbb{O}$, in-batch sampling ensures that each batch contains $|\hat{\mathbb{O}}|$ positives. Because of addressing issues of the other two strategies, in-batch sampling has seen widespread adoption. However, in the next section, we show that in-batch sampling brings its own issues.

3.3 Issues of In-Batch Sampling

We highlight three issues of in-batch sampling.

3.3.1 Issue 1: Batches Generated by In-Batch Sampling Mistake Positives for Negatives

In Figure 2c, the pairs (R2, C2), (R2, C4), and (R4, C4) are mistaken as negatives in the batch, even though they are actually positives in $\boldsymbol{Y}$. Works like Gupta et al. (2024) and Chuang et al. (2020) have reported a performance decline caused by this issue.

3.3.2 Issue 2: In-Batch Sampling Over-Samples Popular Samples

As sampling happens on the positive set $\mathbb{O}$ but not $[m] \times [n]$, the rows and columns with more positives have higher chances to be sampled, and so do the pairs in these rows and columns. An illustration in Figure 2c shows that pairs from R2 and R4 are sampled, as R1 and R3 each contain only one positive. Many previous works, like Yi et al. (2019), Zhou et al. (2021), and Chen et al. (2022), have reported that model performance can be improved if the oversampled pairs are reweighted to reduce their contribution. To quantify the over-sampling, we define

$$\text{popularity of pair } (i,j) := |\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|,$$

where $|\mathbb{O}_{i,:}|$ and $|\mathbb{O}_{:,j}|$ are the numbers of positives of the ith row and the jth column of $\boldsymbol{Y}$. A pair with a larger $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$ is sampled more frequently.

3.3.3 Issue 3: Tuning Batch Sizes is Crucial for Test Performance

As shown on the right side of Figure 2c, among the $|\hat{\mathbb{O}}| \times |\hat{\mathbb{O}}|$ sampled pairs, each row contains only one positive. Thus, the positive-negative ratio of a batch is

$$|\hat{\mathbb{O}}| : (|\hat{\mathbb{O}}|^2 - |\hat{\mathbb{O}}|) = 1 : (|\hat{\mathbb{O}}| - 1).$$

Clearly, this ratio depends on the size $|\hat{\mathbb{O}}|$ chosen to construct each batch. Many works using two-tower models with in-batch sampling have highlighted the impact of batch size on test performance. Examples include Karpukhin et al. (2020) and Qu et al. (2021) in question answering as well as Dahiya et al. (2021) and Gupta et al. (2024) in multi-label classification.

4 In-Batch Sampling from the Perspective of the Expected Objective

The issues discussed in Section 3.3, particularly the last two, suggest that in-batch sampling is inherently non-uniform. Thus, we suspect that the expected objective of in-batch sampling may differ from (5), the objective of the similarity learning problem on the whole label matrix $\boldsymbol{Y}$.

In this section, we examine the expected objective of in-batch sampling and connect it to the issues discussed in Section 3.3. We also clarify how our analysis differs from the closely related work of Rawat et al. (2021), which studies non-uniform data sampling from the perspective of the expected objective.

4.1 Bias of In-Batch Sampling

For in-batch sampling, we derive in Appendix B that the loss for a selected batch is

$$\frac{1}{|\hat{\mathbb{O}}|^2}\left(\sum_{(i,j)\in\hat{\mathbb{O}}}\ell^+_{ij} + \left(\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\ell^-_{ij} - \sum_{(i,j)\in\hat{\mathbb{O}}}\ell^-_{ij}\right)\right). \tag{11}$$

From Figure 2c, the first term of (11) corresponds to the diagonal elements while the remaining terms correspond to the off-diagonal elements. While $\hat{\mathbb{O}}$ is a set of random elements, its size $|\hat{\mathbb{O}}|$ is a fixed hyper-parameter. Therefore, in Appendix E.2, we derive that

$$\begin{aligned}\mathbb{E}_{\hat{\mathbb{O}}}\left[(11)\right] =& \frac{1}{|\mathbb{O}||\hat{\mathbb{O}}|}\left(\sum_{(i,j)\in\mathbb{O}}\ell^+_{ij} - \frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}\right.\\ &\left.+\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\sum_{i=1}^{m}\sum_{j=1}^{n}|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|\ell^-_{ij}\right).\end{aligned} \tag{12}$$

To compare with (5), we multiply (12) by a constant,

$$\begin{aligned}\frac{|\mathbb{O}||\hat{\mathbb{O}}|}{mn}(12) =& \frac{1}{mn}\left(\sum_{(i,j)\in\mathbb{O}}\ell^+_{ij} - \frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}\right.\\ &\left.+\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\sum_{i=1}^{m}\sum_{j=1}^{n}|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|\ell^-_{ij}\right).\end{aligned} \tag{13}$$

Clearly, (13) differs from (5). Thus,

$$\min\nolimits_{\boldsymbol{\theta}}\ \mathbb{E}_{\hat{\mathbb{O}}}\left[(11)\right] \not\equiv \min\nolimits_{\boldsymbol{\theta}}\ (5).$$

We refer to this disparity as the bias. In the next section, we analyze how this bias leads to the issues in Section 3.3.

4.2 The Issues of In-Batch Sampling Stem from Two Terms: $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$ and $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$

By comparing (13) to (5), we see that their difference comes from two extra terms $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$ and $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$ in (13). Through investigation we connect these two terms to the issues in Section 3.3. We start with rewriting (13) to isolate positive and negative pairs as follows:

$$\begin{aligned}&\frac{1}{mn}\left(\sum_{(i,j)\in\mathbb{O}}\left(\ell^+_{ij} + \frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\left(|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|-1\right)\ell^-_{ij}\right)\right.\\ &\left.+\sum_{(i,j)\notin\mathbb{O}}\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|\ell^-_{ij}\right).\end{aligned} \tag{14}$$

4.2.1 Term $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$ Causes Mistaking Positives as Negatives and Over-Sampling Popular Samples

Let us consider a simplified case $|\hat{\mathbb{O}}| = |\mathbb{O}|$, meaning that all positive pairs in $\boldsymbol{Y}$ are used to form the batch. We also assume $\boldsymbol{Y}$ has at least one positive pair in each row and column.[3] The design of in-batch sampling

[3] Appendix E.1 justifies this assumption.

then implies that all (i,j) pairs in $\boldsymbol{Y}$ are included in the batch. In this case, (14) becomes

$$\frac{1}{mn}\left(\sum_{(i,j)\in\mathbb{O}}\underbrace{\left(\ell^+_{ij}+\left(|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|-1\right)\ell^-_{ij}\right)}_{\text{(a)}} + \sum_{(i,j)\notin\mathbb{O}}\underbrace{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|\ell^-_{ij}}_{\text{(b)}}\right). \tag{15}$$

We have the following observations.

- For a positive pair $(i,j)\in\mathbb{O}$, Term (a) in (15) contains two loss terms ℓ^+_{ij} and ℓ^-_{ij}. The former tries to fit the pair (i,j) as a positive, while the latter does the opposite. Since $(|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|-1)$ acts as a weight of ℓ^-_{ij}, the more popular a positive pair (i,j) is, the more inclined we are to misclassify (i,j) as a negative pair. This misclassification may harm model performance, a situation corresponding to the issue in Section 3.3.1.
- For a negative pair $(i,j)\notin\mathbb{O}$, Term (b) in (15) penalizes ℓ^-_{ij} according to its popularity $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$. Thus, the more popular a negative pair (i,j) is, the more inclined we are to classify (i,j) as a negative pair. This tendency corresponds to Section 3.3.2 that over-sampling popular (i,j) pairs negatively impacts the model performance.

4.2.2 Term $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$ Relates to Why Batch Size may Affect the Performance

To illustrate the effect of $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$, let us consider another simplified case $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|=1$ for every (i,j) pair. In this situation, $|\mathbb{O}_{i,:}|=|\mathbb{O}_{:,j}|=1, \forall i,j$, so $\boldsymbol{Y}$ contains exactly one positive element in each row and each column. Then, (14) reduces to

$$\frac{1}{mn}\left(\sum_{(i,j)\in\mathbb{O}}\ell^+_{ij}+\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\sum_{(i,j)\notin\mathbb{O}}\ell^-_{ij}\right). \tag{16}$$

Obviously, the second term is proportional to $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$, and the smaller $|\hat{\mathbb{O}}|$ is, the less important the second term becomes. Tuning $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$ may affect the model performance by balancing between ℓ^+_{ij} and ℓ^-_{ij}. This situation corresponds to Section 3.3.3, which indicates that batch size greatly affects the model performance.

4.3 Differences between Our Analysis and Rawat et al. (2021)

We are not the only one to recognize the biases in recent data sampling strategies. A related work (Rawat et al., 2021) also investigated the disparity on the expected objective.[4] We describe the differences between our analysis and theirs.

While we discuss data sampling over both instances and labels, Rawat et al. (2021) only considered the sampling on labels. They consider a multi-class setting so that every instance has only one positive label and multiple negative labels.[5] As a result, their analysis cannot directly cover the multi-label setting in this work. For example, Section 4.2.1 analyzes the popularity term $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$, but Rawat et al. (2021) does not account for $|\mathbb{O}_{i,:}|$ because $|\mathbb{O}_{i,:}|=1$ in their multi-class setting.

Besides, the discussion in Rawat et al. (2021) covered generic negative (label) sampling strategies instead of focusing on the in-batch sampling. Therefore, their analysis does not explicitly derive the two interpretable bias terms, $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$ and $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$, both of which are the keys in this work to explain the known issues of in-batch sampling.

5 Rigorous Validation with an Unbiased Batch Loss

To validate our analysis, we must introduce an unbiased batch loss for comparing with different biased batch losses. Through a careful experimental design, we can examine the influence of each factor (term) and check whether the empirical observations align with our analysis.

To derive the unbiased loss, recall that the two terms, $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$ and $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$, contribute to the issues of in-batch sampling by acting as weight factors for ℓ^-_{ij} in (13). An intuitive idea to eliminate these terms is to reweight ℓ^-_{ij} by multiplying it with the reciprocals of $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$ and $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$. Given this idea, we introduce the following unbiased batch loss for in-batch sampling.

$$\frac{1}{mn}\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\left(\sum_{(i,j)\in\hat{\mathbb{O}}}\ell^+_{ij}+\frac{|\mathbb{O}|-1}{|\hat{\mathbb{O}}|-1}\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\frac{\ell^-_{ij}}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|} - \sum_{(i,j)\in\hat{\mathbb{O}}}\left(\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\hat{\mathbb{O}}|-1}+1\right)\ell^-_{ij}\right). \tag{17}$$

Appendix E.5 shows that

$$\mathbb{E}_{\hat{\mathbb{O}}}\left[(17)\right] = \text{ the original objective in (5).}$$

[4] The work (Rawat et al., 2021) refers to the expected objective as "the implicit loss" in their context.

[5] For multi-label problems, they argue that these problems can be reduced to a multi-class setting by replicating each instance into as many copies as its number of positive labels, where each copy retains exactly one positive label and treats all remaining labels as negative.

Another unbiased loss was proposed earlier by Yuan et al. (2021) and Krichene et al. (2019). In Appendix F.3, we compare their unbiased loss with ours. The results show that models trained with both unbiased losses converge at the same rate, but our method incurs lower computational cost per training step, making it preferable for both subsequent experiments and practical use.

We also would like to note that the work by Rawat et al. (2021) discussed in Section 4.3 did not derive an unbiased loss like (17). As a result, their experimental evaluation uses the loss without sampling as the unbiased reference. Because this reference involves no sampling, their setting is not entirely suitable for comparing biased and unbiased losses. Specifically, other factors (e.g., the variance of sampling) may also affect the comparison results.

Obtaining the unbiased loss is only the first step in our validation process. A serious obstacle for our experiments is that optimization convergence may affect the model performance and then the comparison among different batch losses. It is known that the learning problem in (3) with two-tower models is non-convex, and SG methods may diverge or converge to any stationary point. While it is impossible to ensure the convergence to the global minimum, to reduce the negative impact of poor convergence, we adopt a conservative and time-consuming experimental setting: we carefully select a sufficiently small learning rate to ensure stable convergence (i.e., a steady decrease in the objective function value).[6]

6 Experiments

The main goal of our experiments is to study how the bias affects model performance. By comparing batch losses that differ in some terms, we examine the effect of each individual term as well as the overall impact. Table 1 lists all batch losses considered for experiments and their corresponding expected objectives.

As discussed in Section 5, mitigating the SG convergence issue requires a sufficiently small learning rate. To select such a learning rate, we design an automatic scheme detailed in Appendix D. The whole process is extremely time-consuming since each learning rate must be carefully verified.

To keep the overall experimental setup tractable, we adopt a simple two-tower model, where both towers consist of a single linear layer with $k = 64$ hidden units. We select two medium-sized datasets, `ml1m` and `EUR-Lex`.[7] These choices allow us to perform rigorous comparisons across different losses while keeping all experiments feasible.

Since batch size is critical to Issue 3 in Section 3.3.3, we study its effect using the batch ratio, defined as $|\hat{\mathbb{O}}|^2/|\mathbb{O}|^2$. We consider this normalized form of batch size for the independence on data size. We experiment with the following batch ratios: $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$.

We consider two widely-used losses, the logistic regression loss, $\ell(a, b) = \log(1 + \exp(-ab))$, and the square loss, $\ell(a, b) = \frac{1}{2}(a - b)^2$.

For model evaluation, we apply the precision and recall at K (i.e., P@K and R@K), where $K \in \{1, 5, 25\}$. Due to limited space, we only present P@5, while other results are in Appendix G.1.

6.1 Ablation Study on $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$ and $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$ to Investigate Their Individual Effects

This section compares different batch losses to validate our analysis in Section 4.2. In Table 1, we give each batch loss a name and list the corresponding expected objective.

6.1.1 The Effect of $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$

We refer to the batch loss in (26) as $\hat{L}_{\text{Popularity}}(\boldsymbol{\theta})$. Figure 3 shows that $\hat{L}_{\text{Popularity}}(\boldsymbol{\theta})$ consistently performs worse than $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$, which corresponds to the unbiased batch loss in (17). We explain this performance gap by examining properties of their expected objectives summarized in Table 1. The expected objective of $\hat{L}_{\text{Popularity}}(\boldsymbol{\theta})$ is in (15). It is affected by the term $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$, which, as described in Section 4.2.1, can significantly degrade model performance by mistaking positives for negatives and over-penalizing true negatives. Similarly, $\hat{L}_{\text{In-Batch}}(\boldsymbol{\theta})$ in (28) is consistently worse than $\hat{L}_{\text{Pos-Neg}}(\boldsymbol{\theta})$ in (27) because the former is additionally affected by $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$.

We note that the P@5 scores of the models trained on `ml1m` with $\hat{L}_{\text{Popularity}}(\boldsymbol{\theta})$ are poor, though the situation is not as serious for the dataset `EUR-Lex`. We can attribute this result to the larger average value of $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$ in `ml1m` (i.e., 12640.8) than `EUR-Lex` (i.e., 115.2). All the aforementioned findings confirm the harmful influence of $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$ on model performance.

[6] Furthermore, we train models with full gradient as an extra reference for our experiments with the SG method. Detailed settings and results of the experiments using full gradient are in Appendix G.2.

[7] The details of the two datasets are provided in Appendix C.

Table 1: A list of the batch losses used in our experiments. This table also indicates whether the expected objectives of these losses are affected by $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$ and $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$. Detailed definitions of these losses are provided in Appendix D.1.

Batch Loss	$\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ (17)	$\hat{L}_{\text{Unbiased-}\omega}(\boldsymbol{\theta})$ (25)	$\hat{L}_{\text{Popularity}}(\boldsymbol{\theta})$ (26)	$\hat{L}_{\text{Pos-Neg}}(\boldsymbol{\theta})$ (27)	$\hat{L}_{\text{In-Batch}}(\boldsymbol{\theta})$ (28)
Expected Objective	(5)	(19)	(15)	(16)	(13)
Affected by $\|\mathbb{O}_{i,:}\|\|\mathbb{O}_{:,j}\|$	No		Yes	No	Yes
Affected by $\frac{\|\hat{\mathbb{O}}\|-1}{\|\mathbb{O}\|-1}$	No		No	Yes	Yes

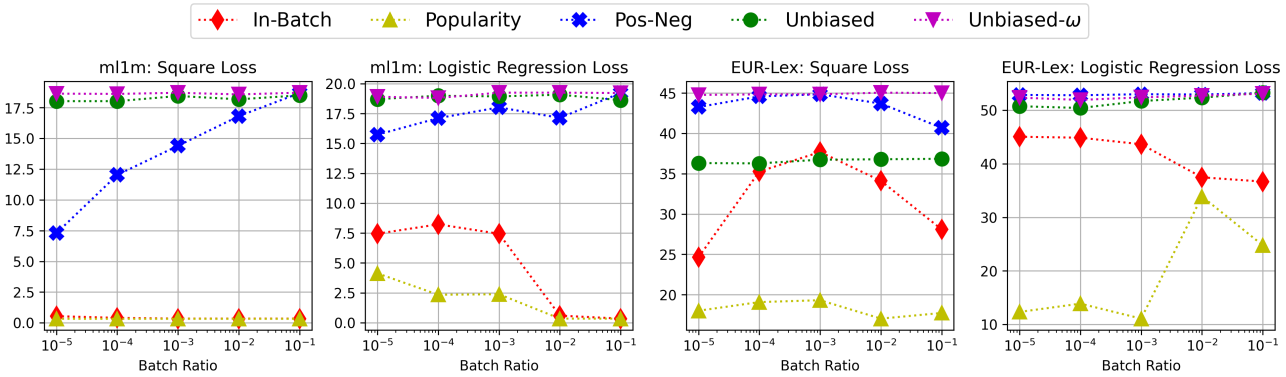


Figure 3: Precision at 5 (P@5) on the testing data for Datasets `ml1m` and `EUR-Lex`. The two batch losses we propose, $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ and $\hat{L}_{\text{Unbiased-}\omega}(\boldsymbol{\theta})$, demonstrate consistent stability across all batch ratios. In particular, $\hat{L}_{\text{Unbiased-}\omega}(\boldsymbol{\theta})$ achieves the best performance when an appropriate ω is used.

6.1.2 The Effect of $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$

We refer to the batch loss in (27) as $\hat{L}_{\text{Pos-Neg}}(\boldsymbol{\theta})$. Figure 3 shows that $\hat{L}_{\text{Pos-Neg}}(\boldsymbol{\theta})$ exhibits varying performance across different batch ratios, whereas $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ remains consistently stable. According to Table 1, we can attribute this difference to the fact that the expected objective of $\hat{L}_{\text{Pos-Neg}}(\boldsymbol{\theta})$ in (16) includes an additional term $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$ when compared to (5), the expected objective of $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$. The following equation connects this term to the batch ratio:

$$\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1} = \frac{\sqrt{\text{batch ratio} \times |\mathbb{O}|^2} - 1}{|\mathbb{O}|-1}. \tag{18}$$

Section 4.2.2 shows that $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$ implicitly adjusts the balance between ℓ^+ and ℓ^-, so (18) implies that the performance of $\hat{L}_{\text{Pos-Neg}}(\boldsymbol{\theta})$ may fluctuate with the batch ratio. Thus, we can interpret $\hat{L}_{\text{Pos-Neg}}(\boldsymbol{\theta})$ as a form of cost-sensitive learning, like Hsieh et al. (2015), where positives and negatives get different weights.

However, using $|\hat{\mathbb{O}}|$ as a hyper-parameter in $\hat{L}_{\text{Pos-Neg}}(\boldsymbol{\theta})$ for cost-sensitive learning may be not ideal because the range of $|\hat{\mathbb{O}}|$ is bounded by, for example, the amount of memory available. A better way is to introduce a separate hyper-parameter ω into $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$, leading to a cost-sensitive variant $\hat{L}_{\text{Unbiased-}\omega}(\boldsymbol{\theta})$. In Appendix E.6, we discuss this variant and prove that

$$\mathbb{E}_{\hat{\mathbb{O}}}\left[\hat{L}_{\text{Unbiased-}\omega}(\boldsymbol{\theta})\right] = \frac{1}{mn}\left(\sum_{(i,j)\in\mathbb{O}} \ell_{ij}^+ + \omega \sum_{(i,j)\notin\mathbb{O}} \ell_{ij}^-\right). \tag{19}$$

Compared to the expected objective of $\hat{L}_{\text{Pos-Neg}}(\boldsymbol{\theta})$ in (16), (19) balances between ℓ^+ and ℓ^- by ω instead of $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$. Figure 3 shows that if ω is properly set, $\hat{L}_{\text{Unbiased-}\omega}(\boldsymbol{\theta})$ can consistently achieve the best performance across all batch ratios. In this experiment, we set ω to $\frac{|\hat{\mathbb{O}}|^*-1}{|\mathbb{O}|-1}$, where $|\hat{\mathbb{O}}|^*$ denotes the batch size yielding the best result for $\hat{L}_{\text{Pos-Neg}}(\boldsymbol{\theta})$.

6.2 Combined Effect of $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$ and $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$

In Figure 3, the performance of $\hat{L}_{\text{In-Batch}}(\boldsymbol{\theta})$ significantly varies, for which we see two interesting points. First, $\hat{L}_{\text{In-Batch}}(\boldsymbol{\theta})$ never performs worse than $\hat{L}_{\text{Popularity}}(\boldsymbol{\theta})$. Second, as the batch ratio approaches one (i.e., $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1} \to 1$), the performance of the two batch losses tends to be close. We can understand these observations from the expected objective of $\hat{L}_{\text{In-Batch}}(\boldsymbol{\theta})$ in (14), which involves $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}| \times \frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$, the product of the two terms that we are interested in. First, $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1} \in (0,1]$ serves as a ratio to mitigate the negative effects introduced by the term $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$, thereby improving the model performance. Second, as $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1} \to 1$, $\hat{L}_{\text{In-Batch}}(\boldsymbol{\theta})$ gradually reduces to $\hat{L}_{\text{Popularity}}(\boldsymbol{\theta})$, leading to similar performance.

6.3 The Potential of the Unbiased Loss

Our in-depth analysis of in-batch sampling leads us to derive an unbiased batch loss $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ in (17). Even though we currently regard $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ as a tool to validate our theoretical findings, $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ and $\hat{L}_{\text{Unbiased-}\omega}(\boldsymbol{\theta})$ demonstrate remarkable stability across different batch ratios in Figure 3. This observation suggests that $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ is a unified solution to address the issues of in-batch sampling effectively. Furthermore, we show in Appendix F.1 that, compared to $\hat{L}_{\text{In-Batch}}(\boldsymbol{\theta})$ of in-batch sampling, $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ incurs very low additional computational cost. These advantages indicate the potential of the derived unbiased loss for two-tower models.

7 Conclusions

This work presents the first systematic analysis of in-batch sampling for training two-tower models. We novelly point out that all the known issues stem from the inherent bias in the expected objective. To validate our findings, we derive an unbiased batch loss by a simple reweighting scheme. Through rigorous settings, we experimentally compare biased batch losses with the unbiased one. The results fully confirm our theoretical analysis and demonstrate the advantages of using an unbiased batch loss in training two-tower models.

A potential direction for future work is to extend our analysis beyond point-wise losses. For example, ranking-based losses (e.g., pairwise losses) are also common for two-tower models. For experiments, an extension to use more complicated models may give further insight.

Acknowledgements

We would like to thank Bowen Yuan, Jui-Nan Yen, and Sheng-Wei Chen for helpful discussions. This work was supported in part by National Science and Technology Council of Taiwan grants NSTC-113-2222-E-002-005-MY3 and NSTC-114-2634-F-002-007.

References

Awasthi, P., Dikkala, N., and Kamath, P. (2022). Do more negative samples necessarily hurt in contrastive learning? In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1101–1116.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311.

Chen, J., Lian, D., Li, Y., Wang, B., Zheng, K., and Chen, E. (2022). Cache-augmented inbatch importance resampling for training recommender retriever. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chuang, C.-Y., Robinson, J., Lin, Y.-C., Torralba, A., and Jegelka, S. (2020). Debiased contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 8765–8775.

Dahiya, K., Saini, D., Mittal, A., Shaw, A., Dave, K., Soni, A., Jain, H., Agarwal, S., and Varma, M. (2021). DeepXML: A deep extreme multi-label learning framework applied to short text documents. In *Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 31–39.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159.

Gillick, D., Kulkarni, S., Lansing, L., Presta, A., Baldridge, J., Ie, E., and Garcia-Olano, D. (2019). Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 528–537.

Gupta, N., Devvrit, Rawat, A. S., Bhojanapalli, S., Jain, P., and Dhillon, I. S. (2024). Dual-encoders for extreme multi-label classification. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Hsieh, C.-J., Natarajan, N., and Dhillon, I. (2015). PU learning for matrix completion. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 2445–2453.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Krichene, W., Mayoraz, N., Rendle, S., Zhang, L., Yi, X., Hong, L., Chi, E., and Anderson, J. (2019). Efficient training on very large corpora via gramian estimation. In *International Conference on Learning Representations*.

Lin, L.-C., Liu, Y., and Lin, C.-J. (2025). Sampled estimators for softmax must be biased. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38.

Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528.

Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W. X., Dong, D., Wu, H., and Wang, H. (2021). RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 5835–5847.

Rawat, A. S., Menon, A. K., Jitkrittum, W., Jayasumana, S., Yu, F. X., Reddi, S. J., and Kumar, S. (2021). Disentangling sampling and labeling bias for learning in large-output spaces. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8890–8901.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3980–3990.

Rendle, S. (2022). Item recommendation from implicit feedback. In *Recommender Systems Handbook*, pages 143–171. Springer US.

Rendle, S., Krichene, W., Zhang, L., and Anderson, J. (2020). Neural collaborative filtering vs. matrix factorization revisited. In *Fourteenth ACM Conference on Recommender Systems*, page 240–248.

Yang, J., Yi, X., Cheng, D. Z., Hong, L., Li, Y., Wang, S. X., Xu, T., and Chi, E. H. (2020). Mixed negative sampling for learning two-tower neural networks in recommendations. In *Companion Proceedings of the Web Conference (WWW)*, pages 441–447.

Yi, X., Yang, J., Hong, L., Cheng, D. Z., Heldt, L., Kumthekar, A., Zhao, Z., Wei, L., and Chi, E. (2019). Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, page 269–277.

Yu, H.-F., Bilenko, M., and Lin, C.-J. (2017). Selection of negative samples for one-class matrix factorization. In *Proceedings of SIAM International Conference on Data Mining (SDM)*.

Yuan, B., Li, Y.-S., Quan, P., and Lin, C.-J. (2021). Efficient optimization methods for extreme similarity learning with nonlinear embeddings. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Zhou, C., Ma, J., Zhang, J., Zhou, J., and Yang, H. (2021). Contrastive learning for debiased candidate generation in large-scale recommender systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 3985–3995.

Examining the Bias of In-Batch Sampling in Similarity Learning with Two-Tower Models: Supplementary Materials

A Main Notations

In Table 2, we list the main notations used in this work.

Table 2: Main notations

Notation	Description
(i, j)	(left entity i, right entity j) pair
m, n	numbers of left entities and right entities
$\boldsymbol{Y}, Y_{ij}$	the ground-truth label matrix and its element for the (i, j) pair
$\hat{\boldsymbol{Y}}, \hat{Y}_{ij}$	the similarity matrix predicted by $y(\cdot)$ and its element for the (i, j) pair
$\mathbb{O}, \hat{\mathbb{O}}$	the full set and a subset of positive pairs in $\boldsymbol{Y}$
$\lvert\mathbb{O}_{i,:}\rvert, \lvert\mathbb{O}_{:,j}\rvert$	numbers of positives of the ith row and the jth column of $\boldsymbol{Y}$
ℓ, ℓ^+, ℓ^-	a point-wise loss, and its variants for positive and negative pairs
ω	a hyper-parameter used to balance between ℓ^+ and ℓ^-
$y(\cdot), \boldsymbol{\theta}$	a similarity function and its parameters
$L(\boldsymbol{\theta}), \hat{L}(\boldsymbol{\theta})$	the expected objective function and a batch loss
$f(\cdot), g(\cdot)$	models of the two towers for left and right entities
$F(f)$, $F(g)$	cost of operations related to $f(\cdot)$, and $g(\cdot)$

B Batch Losses for Different Sampling Strategies

Generally, the batch loss is computed as the average of losses of all data points in a batch. Here, we present the batch losses corresponding to different sampling strategies.

- **Batch loss for Naive Sampling:** Given $\mathbb{B}$ the batch sampled from $[m] \times [n]$, we have

$$\frac{1}{|\mathbb{B}|} \sum_{(i,j)\in\mathbb{B}} \ell(Y_{ij}, \hat{Y}_{ij}), \tag{20}$$

- **Batch loss for Cross Sampling:** Given $R \subseteq [m]$ the $\hat{m}$ sampled rows and $C \subseteq [n]$ the $\hat{n}$ sampled columns, we have

$$\frac{1}{\hat{m}\hat{n}} \sum_{i\in R} \sum_{j\in C} \ell(Y_{ij}, \hat{Y}_{ij}). \tag{21}$$

- **Batch loss for In-Batch Sampling:** Given $\hat{\mathbb{O}}$ the sampled batch from $\mathbb{O}$, we build a new label matrix $\bar{\boldsymbol{Y}}$ of size $|\hat{\mathbb{O}}| \times |\hat{\mathbb{O}}|$ (see an example in Figure 2c). In $\bar{\boldsymbol{Y}}$, we only take diagonal elements as positives and the others as negatives. We define two mapping functions that map row and column indices of $\bar{\boldsymbol{Y}}$ back to the indices in $\hat{\mathbb{O}}$,

$$\begin{cases} \phi : \{1, \cdots, |\hat{\mathbb{O}}|\} \to \{i : (i, j) \in \hat{\mathbb{O}}\}, \\ \psi : \{1, \cdots, |\hat{\mathbb{O}}|\} \to \{j : (i, j) \in \hat{\mathbb{O}}\}. \end{cases}$$

Table 3: Data statistics.

Dataset	m	n	$\lvert\mathbb{O}\rvert$	Average $\lvert\mathbb{O}_{i,:}\rvert\lvert\mathbb{O}_{:,j}\rvert$
`ml1m`	6,037	3,513	517,770	12,640.8
`EUR-Lex`	15,449	3,801	82,265	115.2

Note that samples of $\hat{\mathbb{O}}$ corresponds to different (ϕ, ψ) pairs. The sequence $\left(\phi(1), \cdots, \phi(|\hat{\mathbb{O}}|)\right)$ contains indices of taking i from iterating $(i, j') \in \hat{\mathbb{O}}$. Similarly, we take j from iterating $(i', j) \in \hat{\mathbb{O}}$ to have ψ. Therefore,

$$\sum_{a=1}^{|\hat{\mathbb{O}}|}\sum_{b=1}^{|\hat{\mathbb{O}}|}(\cdots)_{\phi(a)\psi(b)} = \sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}(\cdots)_{ij}. \tag{22}$$

For instance, in the example of Figure 2c, $\hat{\mathbb{O}} = \{(2,2), (2,4), (4,1)\}$. We have

$$\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}(\cdots)_{ij} = \sum_{i=2,2,4}\sum_{j=2,4,1}(\cdots)_{ij},$$

indicating the sum over all elements in the square generated by $\hat{\mathbb{O}}$. Then, the batch loss is

$$\begin{align}
&\frac{1}{|\hat{\mathbb{O}}|^2}\sum_{a=1}^{|\hat{\mathbb{O}}|}\sum_{b=1}^{|\hat{\mathbb{O}}|}\ell\left(\bar{Y}_{ab}, \hat{Y}_{\phi(a)\psi(b)}\right) \tag{23}\\
=&\frac{1}{|\hat{\mathbb{O}}|^2}\sum_{a=1}^{|\hat{\mathbb{O}}|}\ell^{+}_{\phi(a)\psi(a)} + \frac{1}{|\hat{\mathbb{O}}|^2}\sum_{a=1}^{|\hat{\mathbb{O}}|}\sum_{b=1}^{|\hat{\mathbb{O}}|}\ell^{-}_{\phi(a)\psi(b)} - \frac{1}{|\hat{\mathbb{O}}|^2}\sum_{a=1}^{|\hat{\mathbb{O}}|}\ell^{-}_{\phi(a)\psi(a)} \notag\\
=&\frac{1}{|\hat{\mathbb{O}}|^2}\sum_{(i,j)\in\hat{\mathbb{O}}}\ell^{+}_{ij} + \frac{1}{|\hat{\mathbb{O}}|^2}\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\ell^{-}_{ij} - \frac{1}{|\hat{\mathbb{O}}|^2}\sum_{(i,j)\in\hat{\mathbb{O}}}\ell^{-}_{ij} \tag{24}\\
=&\frac{1}{|\hat{\mathbb{O}}|^2}\sum_{(i,j)\in\hat{\mathbb{O}}}\left(\ell^{+}_{ij} - \ell^{-}_{ij}\right) + \frac{1}{|\hat{\mathbb{O}}|^2}\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\ell^{-}_{ij}, \notag
\end{align}$$

where (24) is from (22).

C Statistics of Datasets

We show the statistics of the two datasets, `ml1m` and `EUR-Lex`, in Table 3. The first dataset, `ml1m`, corresponding to positive–unlabeled learning in recommender systems, is from the MovieLens dataset: `https://grouplens.org/datasets/movielens/1m/`. Since `ml1m` is a rating-based dataset, we follow Yu et al. (2017) to binarize the ratings in the original set. We consider the pairs with rating ≥ 4 as positive while the rest including the unrated items as negative. The second dataset, `EUR-Lex`, is a standard benchmark for multi-label text classification and from the LibSVM datasets: `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.

D More Details of Experimental Settings

The stochastic gradient method used in our experiments is AdaGrad (Duchi et al., 2011). To address the convergence issues discussed in Section 5, we adopt a prohibitively time-consuming but rigorous procedure to select a sufficiently small learning rate. The procedure relies on repeatedly restarting training with progressively smaller learning rates to achieve stable convergence.

The complete experimental procedure, including the learning rate selection, is summarized below for each dataset and batch loss:

1. Begin the learning rate selection with an intentionally large value, 2^{18}, which inevitably leads to divergence.
2. Train the model. Within each epoch, evaluate the training objective function and all test metrics 100 times.
3. If, at any checkpoint, the objective function value diverges (i.e., results in NaN or is greater than 100 times the initial value), we stop training, halve the learning rate, and restart from Step 2.
4. Otherwise, continue training and keep track of the best value achieved so far for every test metric. If none of the metrics improves over its best recorded value for 10 consecutive epochs, we stop training and mark this learning rate as "usable."
5. Report the best test metrics under the usable learning rate.

D.1 Batch Losses Used in Our Experiments

Here we provide the definitions of all batch losses used in our experiments.

- **Unbiased**: The batch loss proposed in Section 5.

$$\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta}) := (17),$$

which is unbiased with respect to (5). We also define $L_{\text{Unbiased}}(\boldsymbol{\theta}) := (5)$.

- **Unbiased-ω**: A variant of $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ introduced in Section 6.1.2.

$$\begin{aligned}\hat{L}_{\text{Unbiased-}\omega}(\boldsymbol{\theta}) :=& \frac{1}{mn}\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\left(\sum_{(i,j)\in\hat{\mathbb{O}}}\ell^{+}_{ij} + \omega\frac{|\mathbb{O}|-1}{|\hat{\mathbb{O}}|-1}\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\frac{\ell^{-}_{ij}}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\right.\\ &\left.-\omega\sum_{(i,j)\in\hat{\mathbb{O}}}\left(\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\hat{\mathbb{O}}|-1}+1\right)\ell^{-}_{ij}\right),\end{aligned} \tag{25}$$

where ω is a hyper-parameter used to balance between ℓ^{+} and ℓ^{-}. In Appendix E.6, we prove that $\mathbb{E}_{\hat{\mathbb{O}}}\left[\hat{L}_{\text{Unbiased-}\omega}(\boldsymbol{\theta})\right] = (19)$. We further define $L_{\text{Unbiased-}\omega}(\boldsymbol{\theta}) := (19)$.

- **Popularity**: To study the individual effect of $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$, we need a batch loss corresponding to an objective only affected by $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$. That is, the objective should not be related to the term $\frac{|\mathbb{O}|-1}{|\hat{\mathbb{O}}|-1}$. Recall that $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ debiases in-batch sampling by reweighting. We discard the scaling factor $\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}$ used in $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ for alleviating the effect of $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$ and obtain

$$\hat{L}_{\text{Popularity}}(\boldsymbol{\theta}) := \frac{1}{mn}\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\left(\sum_{(i,j)\in\hat{\mathbb{O}}}\ell^{+}_{ij} + \frac{|\mathbb{O}|-1}{|\hat{\mathbb{O}}|-1}\left(\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\ell^{-}_{ij} - \sum_{(i,j)\in\hat{\mathbb{O}}}\ell^{-}_{ij}\right)\right). \tag{26}$$

By this way, the objective may be affected by $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$ but not $\frac{|\mathbb{O}|-1}{|\hat{\mathbb{O}}|-1}$. In Appendix E.3, we prove that $\mathbb{E}_{\hat{\mathbb{O}}}\left[\hat{L}_{\text{Popularity}}(\boldsymbol{\theta})\right] = (15)$, which satisfies the properties that we hope to have. We further define $L_{\text{Popularity}}(\boldsymbol{\theta}) := (15)$.

- **Pos-Neg**: The goal is to study the effect of the term $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$. Similar to the design of $L_{\text{Popularity}}(\boldsymbol{\theta})$, we only preserve the bias caused by this term. We modify $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ by multiplying every ℓ^{-}_{ij} with $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$, resulting in

$$\hat{L}_{\text{Pos-Neg}}(\boldsymbol{\theta}) := \frac{1}{mn}\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\left(\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\frac{\ell^{-}_{ij}}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|} + \sum_{(i,j)\in\hat{\mathbb{O}}}\left(\ell^{+}_{ij} - \left(\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\mathbb{O}|-1} + \frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\right)\ell^{-}_{ij}\right)\right). \tag{27}$$

In Appendix E.4, we prove that $\mathbb{E}_{\hat{\mathbb{O}}}\left[\hat{L}_{\text{Pos-Neg}}(\boldsymbol{\theta})\right] = (16)$, where (16) is only affected by $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$. We further define $L_{\text{Pos-Neg}}(\boldsymbol{\theta}) := (16)$.

- **In-Batch**: To study in-batch sampling, we already have a batch loss in (11). To facilitate the comparison with other batch losses, we also multiply (11) by $\frac{|\mathbb{O}||\hat{\mathbb{O}}|}{mn}$ and define

$$\hat{L}_{\text{In-Batch}}(\boldsymbol{\theta}) := \frac{|\mathbb{O}||\hat{\mathbb{O}}|}{mn}(11). \tag{28}$$

Since $\frac{|\mathbb{O}||\hat{\mathbb{O}}|}{mn}$ is a constant,

$$\mathbb{E}_{\hat{\mathbb{O}}}\left[\hat{L}_{\text{In-Batch}}(\boldsymbol{\theta})\right] = \frac{|\mathbb{O}||\hat{\mathbb{O}}|}{mn}\mathbb{E}_{\hat{\mathbb{O}}}\left[(11)\right] = (13).$$

where the last equality is from

$$\mathbb{E}_{\hat{\mathbb{O}}}\left[(11)\right] = (12), \text{ and } \frac{|\mathbb{O}||\hat{\mathbb{O}}|}{mn}(12) = (13).$$

We further define $L_{\text{In-Batch}}(\boldsymbol{\theta}) := (13)$.

E Proofs

E.1 Preliminaries

Following common practice, we have the following assumptions.

Assumption E.1. Any column or row of $\boldsymbol{Y}$, the label matrix of the training data, has at least one positive entry. That is,

$$\forall i \in [m], j \in [n], |\mathbb{O}_{i,:}| \geq 1 \text{ and } |\mathbb{O}_{:,j}| \geq 1,$$

where $\mathbb{O}_{i,:} = \{j : (i,j) \in \mathbb{O}\}$, $\mathbb{O}_{:,j} = \{i : (i,j) \in \mathbb{O}\}$, and $|\mathbb{O}_{i,:}|$ and $|\mathbb{O}_{:,j}|$ are their sizes.

Take multi-label classification for example. This assumption means that in the training data, each label has at least one positive instance, and each instance has at least one positive label. Such a situation generally holds. With this assumption, $\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}$ is well-defined in our discussion.

Assumption E.2. At each training step, each batch is sampled without replacement.

Throughout all the proofs in this work, we will be using Assumptions E.1 and E.2, and the indicator function $\mathbb{1}$.

Definition E.3.

$$\mathbb{1}_{\phi} = \begin{cases} 1 & \text{if } \phi \text{ is true,} \\ 0 & \text{otherwise.} \end{cases}$$

We will be using the following identity,

$$\mathbb{E}\left[\mathbb{1}_{\phi}\right] = \Pr(\phi), \tag{29}$$

as well as the following lemmas.

Lemma E.4. *(Change of Variables) Given a function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$,*

$$\sum_{(i,j')\in\mathbb{O}} \sum_{(i',j)\in\mathbb{O}} f(i,j) = \sum_{i=1}^{m}\sum_{j=1}^{n} |\mathbb{O}_{i,:}||\mathbb{O}_{:,j}| f(i,j).$$

Proof.

$$\begin{aligned}\sum_{(i,j')\in\mathbb{O}} \sum_{(i',j)\in\mathbb{O}} f(i,j) &= \sum_{i=1}^{m}\sum_{j'\in\mathbb{O}_{i,:}}\sum_{j=1}^{n}\sum_{i'\in\mathbb{O}_{:,j}} f(i,j) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} |\mathbb{O}_{i,:}||\mathbb{O}_{:,j}| f(i,j).\end{aligned}$$

□

Lemma E.5. *Let $\hat{\mathbb{O}}$ be uniformly sampled without replacement from $\mathbb{O}$. Given a function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$,*

$$\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}} f(i,j)\right] = \frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}} f(i,j).$$

Proof.

$$\begin{aligned}\mathbb{E}_{\hat{\mathbb{O}}}\left[\sum_{(i,j)\in\hat{\mathbb{O}}} f(i,j)\right] &= \mathbb{E}_{\hat{\mathbb{O}}}\left[\sum_{(i,j)\in\mathbb{O}} \mathbb{1}_{(i,j)\in\hat{\mathbb{O}}}\cdot f(i,j)\right] \\ &= \sum_{(i,j)\in\mathbb{O}} \mathbb{E}_{\hat{\mathbb{O}}}\left[\mathbb{1}_{(i,j)\in\hat{\mathbb{O}}}\right] f(i,j) \\ &= \frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}} f(i,j), \qquad (30)\end{aligned}$$

where the last equality follows from (29) and

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{(i,j)\in\hat{\mathbb{O}}}\right] &= \Pr\left((i,j)\in\hat{\mathbb{O}}\right) \\ &= \frac{\#\text{ possible }\hat{\mathbb{O}}\text{'s with }(i,j)\in\hat{\mathbb{O}}}{\#\text{ possible }\hat{\mathbb{O}}\text{'s}} \\ &= \frac{\binom{|\mathbb{O}|-1}{|\hat{\mathbb{O}}|-1}}{\binom{|\mathbb{O}|}{|\hat{\mathbb{O}}|}} \\ &= \frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}.\end{aligned}$$

□

Lemma E.6. *Let $\hat{\mathbb{O}}$ be uniformly sampled without replacement from $\mathbb{O}$. Given a function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$,*

$$\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}} f(i,j)\right] = \frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}} f(i,j) + \frac{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}{|\mathbb{O}|(|\mathbb{O}|-1)}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|f(i,j) - \sum_{(i,j)\in\mathbb{O}} f(i,j)\right).$$

Proof.

$$\begin{aligned}
\mathbb{E}_{\hat{\mathbb{O}}}\left[\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}} f(i,j)\right] &= \mathbb{E}_{\hat{\mathbb{O}}}\left[\sum_{(i,j')\in\mathbb{O}}\sum_{(i',j)\in\mathbb{O}} \mathbb{1}_{(i,j')\in\hat{\mathbb{O}}}\cdot\mathbb{1}_{(i',j)\in\hat{\mathbb{O}}}\cdot f(i,j)\right]\\
&= \sum_{(i,j')\in\mathbb{O}}\sum_{(i',j)\in\mathbb{O}} \mathbb{E}\left[\mathbb{1}_{(i,j')\in\hat{\mathbb{O}}}\cdot\mathbb{1}_{(i',j)\in\hat{\mathbb{O}}}\right] f(i,j)\\
&= \sum_{(i,j)\in\mathbb{O}} \mathbb{E}\left[\mathbb{1}_{(i,j)\in\hat{\mathbb{O}}}\cdot\mathbb{1}_{(i,j)\in\hat{\mathbb{O}}}\right] f(i,j)\\
&\quad+\sum_{(i,j')\in\mathbb{O}}\sum_{\substack{(i',j)\in\mathbb{O}\\(i',j)\neq(i,j')}} \mathbb{E}\left[\mathbb{1}_{(i,j')\in\hat{\mathbb{O}}}\cdot\mathbb{1}_{(i',j)\in\hat{\mathbb{O}}}\right] f(i,j)\\
&= \frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}} f(i,j) + \frac{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}{|\mathbb{O}|(|\mathbb{O}|-1)}\sum_{(i,j')\in\mathbb{O}}\sum_{\substack{(i',j)\in\mathbb{O}\\(i',j)\neq(i,j')}} f(i,j) \qquad (31)\\
&= \frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}} f(i,j) + \frac{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}{|\mathbb{O}|(|\mathbb{O}|-1)}\left(\sum_{(i,j')\in\mathbb{O}}\sum_{(i',j)\in\mathbb{O}} f(i,j) - \sum_{(i,j)\in\mathbb{O}} f(i,j)\right)\\
&= \frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}} f(i,j) + \frac{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}{|\mathbb{O}|(|\mathbb{O}|-1)}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|f(i,j) - \sum_{(i,j)\in\mathbb{O}} f(i,j)\right), \qquad (32)
\end{aligned}$$

where (31) follows from (30) and the derivation in (33), and (32) follows from Lemma E.4. We show the details of (33) below.

$$\begin{aligned}
\mathbb{E}_{\hat{\mathbb{O}}}\left[\mathbb{1}_{(i,j')\in\hat{\mathbb{O}}}\cdot\mathbb{1}_{(i',j)\in\hat{\mathbb{O}}}\right] &= \frac{\#\text{ possible }\hat{\mathbb{O}}\text{'s with }(i,j')\in\hat{\mathbb{O}}\text{ and }(i',j)\in\hat{\mathbb{O}}}{\#\text{ possible }\hat{\mathbb{O}}\text{'s}}\\
&\overset{(a)}{=} \frac{\binom{|\mathbb{O}|-2}{|\hat{\mathbb{O}}|-2}}{\binom{|\mathbb{O}|}{|\hat{\mathbb{O}}|}}\\
&= \frac{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}{|\mathbb{O}|(|\mathbb{O}|-1)},. \qquad (33)
\end{aligned}$$

where the equality (a) holds because $(i,j')\neq(i',j)$. □

E.2 Proof for the Expectation of (11)

Taking expectation on (11), we have

$$\mathbb{E}\left[\frac{1}{|\hat{\mathbb{O}}|^2}\left(\sum_{(i,j)\in\hat{\mathbb{O}}}(\ell_{ij}^{+}-\ell_{ij}^{-}) + \sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\ell_{ij}^{-}\right)\right] = \frac{1}{|\hat{\mathbb{O}}|^2}\left(\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}(\ell_{ij}^{+}-\ell_{ij}^{-})\right] + \mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\ell_{ij}^{-}\right]\right).$$

Applying Lemma E.5, the first expectation is

$$\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}(\ell_{ij}^{+}-\ell_{ij}^{-})\right] = \frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}}(\ell_{ij}^{+}-\ell_{ij}^{-}).$$

Applying Lemma E.6, the second expectation is

$$\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\ell_{ij}^{-}\right] = \frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}}\ell_{ij}^{-} + \frac{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}{|\mathbb{O}|(|\mathbb{O}|-1)}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|\ell_{ij}^{-} - \sum_{(i,j)\in\mathbb{O}}\ell_{ij}^{-}\right).$$

Thus, the expectation of (11) is

$$
\begin{aligned}
&\frac{1}{|\hat{\mathbb{O}}|^2}\left(\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\left(\ell^+_{ij}-\ell^-_{ij}\right)\right]+\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\ell^-_{ij}\right]\right)\\
=&\frac{1}{|\hat{\mathbb{O}}|^2}\left(\frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}}\left(\ell^+_{ij}-\ell^-_{ij}\right)+\frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}+\frac{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}{|\mathbb{O}|(|\mathbb{O}|-1)}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|\ell^-_{ij}-\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}\right)\right)\\
=&\frac{1}{|\mathbb{O}||\hat{\mathbb{O}}|}\left(\sum_{(i,j)\in\mathbb{O}}\left(\ell^+_{ij}-\ell^-_{ij}\right)+\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}+\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|\ell^-_{ij}-\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}\right)\right)\\
=&\frac{1}{|\mathbb{O}||\hat{\mathbb{O}}|}\left(\sum_{(i,j)\in\mathbb{O}}\ell^+_{ij}+\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|\ell^-_{ij}-\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}\right)\right).
\end{aligned}
$$

E.3 Proof of $\mathbb{E}\left[(26)\right]=(15)$

Taking expectation on $\hat{L}_{\text{Popularity}}(\boldsymbol{\theta})$ in (26), we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{mn}\left(\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\sum_{(i,j)\in\hat{\mathbb{O}}}\ell^+_{ij}+\frac{|\mathbb{O}|(|\mathbb{O}|-1)}{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}\left(\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\ell^-_{ij}-\sum_{(i,j)\in\hat{\mathbb{O}}}\ell^-_{ij}\right)\right)\right]\\
=&\frac{1}{mn}\left(\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\ell^+_{ij}\right]+\frac{|\mathbb{O}|(|\mathbb{O}|-1)}{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}\left(\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\ell^-_{ij}\right]-\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\ell^-_{ij}\right]\right)\right).
\end{aligned}
$$

Applying Lemma E.5, the first expectation is

$$
\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\ell^+_{ij}\right]=\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}}\ell^+_{ij}=\sum_{(i,j)\in\mathbb{O}}\ell^+_{ij},
$$

and the third expectation is

$$
\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\ell^-_{ij}\right]=\frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}.
$$

Applying Lemma E.6, the second expectation is

$$
\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\ell^-_{ij}\right]=\frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}+\frac{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}{|\mathbb{O}|(|\mathbb{O}|-1)}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|\ell^-_{ij}-\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}\right).
$$

Then, we have

$$
\begin{aligned}
&\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\ell^+_{ij}\right]+\frac{|\mathbb{O}|(|\mathbb{O}|-1)}{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}\left(\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\ell^-_{ij}\right]-\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\ell^-_{ij}\right]\right)\\
=&\sum_{(i,j)\in\mathbb{O}}\ell^+_{ij}+\frac{|\mathbb{O}|(|\mathbb{O}|-1)}{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}\left(\frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}+\frac{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}{|\mathbb{O}|(|\mathbb{O}|-1)}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|\ell^-_{ij}-\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}\right)-\frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}\right)\\
=&\sum_{(i,j)\in\mathbb{O}}\ell^+_{ij}+\frac{|\mathbb{O}|(|\mathbb{O}|-1)}{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}\left(\frac{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}{|\mathbb{O}|(|\mathbb{O}|-1)}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|\ell^-_{ij}-\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}\right)\right)\\
=&\sum_{(i,j)\in\mathbb{O}}\ell^+_{ij}+\sum_{i=1}^{m}\sum_{j=1}^{n}|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|\ell^-_{ij}-\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}.
\end{aligned}
$$

We complete the proof by multiplying $\frac{1}{mn}$ with the last equation above.

E.4 Proof of $\mathbb{E}\left[(27)\right] = (16)$

Taking the expectation of $\hat{L}_{\text{Pos-Neg}}(\boldsymbol{\theta})$ in (27), we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{mn}\left(\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\sum_{(i,j)\in\hat{\mathbb{O}}}\ell_{ij}^{+} + \frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell_{ij}^{-}\right.\right.\\
&\left.\left.-\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\sum_{(i,j)\in\hat{\mathbb{O}}}\left(\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\mathbb{O}|-1}+\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\right)\ell_{ij}^{-}\right)\right]\\
=&\frac{1}{mn}\left(\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\ell_{ij}^{+}\right] + \frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell_{ij}^{-}\right]\right.\\
&\left.-\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\left(\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\mathbb{O}|-1}+\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\right)\ell_{ij}^{-}\right]\right).
\end{aligned}$$

Applying Lemma E.5, the first expectation is

$$\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\ell_{ij}^{+}\right] = \frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}}\ell_{ij}^{+} = \sum_{(i,j)\in\mathbb{O}}\ell_{ij}^{+},$$

and the third expectation is

$$\begin{aligned}
\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\left(\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\mathbb{O}|-1}+\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\right)\ell_{ij}^{-}\right] =&\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}}\left(\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\mathbb{O}|-1}+\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\right)\ell_{ij}^{-}\\
=&\sum_{(i,j)\in\mathbb{O}}\left(\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\mathbb{O}|-1}+\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\right)\ell_{ij}^{-}\\
=&\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\mathbb{O}|-1}\sum_{(i,j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\ell_{ij}^{-}+\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\sum_{(i,j)\in\mathbb{O}}\ell_{ij}^{-}.
\end{aligned}$$

Applying Lemma E.6, the second expectation is

$$\begin{aligned}
&\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell_{ij}^{-}\right]\\
=&\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell_{ij}^{-}+\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\frac{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}{|\mathbb{O}|(|\mathbb{O}|-1)}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell_{ij}^{-}-\sum_{(i,j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell_{ij}^{-}\right)\\
=&\sum_{(i,j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell_{ij}^{-}+\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\ell_{ij}^{-}-\sum_{(i,j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell_{ij}^{-}\right)\\
=&(1-\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1})\sum_{(i,j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell_{ij}^{-}+\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\ell_{ij}^{-}\\
=&\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\mathbb{O}|-1}\sum_{(i,j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell_{ij}^{-}+\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\ell_{ij}^{-}.
\end{aligned}$$

Then,

$$
\begin{aligned}
&\frac{1}{mn}\left(\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\ell^+_{ij}\right]+\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}\right]\right.\\
&\left.-\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\left(\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\mathbb{O}|-1}+\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\right)\ell^-_{ij}\right]\right)\\
=&\frac{1}{mn}\left(\sum_{(i,j)\in\mathbb{O}}\ell^+_{ij}+\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\mathbb{O}|-1}\sum_{(i,j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}+\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\ell^-_{ij}\right.\\
&\left.-\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\mathbb{O}|-1}\sum_{(i,j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}-\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}\right)\\
=&\frac{1}{mn}\left(\sum_{(i,j)\in\mathbb{O}}\ell^+_{ij}+\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\ell^-_{ij}-\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}\right)\\
=&\frac{1}{mn}\left(\sum_{(i,j)\in\mathbb{O}}\ell^+_{ij}+\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\ell^-_{ij}-\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}\right)\right).
\end{aligned}
$$

E.5 Proof of $\mathbb{E}\left[(17)\right]=(5)$

Taking expectation on $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ in (17), we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{mn}\left(\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\sum_{(i,j)\in\hat{\mathbb{O}}}\ell^+_{ij}+\frac{|\mathbb{O}|(|\mathbb{O}|-1)}{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}\right.\right.\\
&\left.\left.-\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\sum_{(i,j)\in\hat{\mathbb{O}}}\left(\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\hat{\mathbb{O}}|-1}+1\right)\ell^-_{ij}\right)\right]\\
=&\frac{1}{mn}\left(\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\ell^+_{ij}\right]+\frac{|\mathbb{O}|(|\mathbb{O}|-1)}{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}\right]\right.\\
&\left.-\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\left(\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\hat{\mathbb{O}}|-1}+1\right)\ell^-_{ij}\right]\right).
\end{aligned}
$$

Applying Lemma E.5, the first expectation is

$$
\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\ell^+_{ij}\right]=\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}}\ell^+_{ij}=\sum_{(i,j)\in\mathbb{O}}\ell^+_{ij},
$$

and the third expectation is

$$
\begin{aligned}
\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\left(\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\hat{\mathbb{O}}|-1}+1\right)\ell^-_{ij}\right]=&\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}}\left(\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\hat{\mathbb{O}}|-1}+1\right)\ell^-_{ij}\\
=&\sum_{(i,j)\in\mathbb{O}}\left(\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\hat{\mathbb{O}}|-1}+1\right)\ell^-_{ij}\\
=&\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\hat{\mathbb{O}}|-1}\sum_{(i,j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\ell^-_{ij}+\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}.
\end{aligned}
$$

Applying Lemma E.6, the second expectation is

$$\begin{aligned}
&\frac{|\mathbb{O}|(|\mathbb{O}|-1)}{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}\right]\\
=&\frac{|\mathbb{O}|(|\mathbb{O}|-1)}{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}\frac{|\hat{\mathbb{O}}|}{|\mathbb{O}|}\sum_{(i,j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}\\
&+\frac{|\mathbb{O}|(|\mathbb{O}|-1)}{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}\frac{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}{|\mathbb{O}|(|\mathbb{O}|-1)}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}-\sum_{(i,j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}\right)\\
=&\frac{|\mathbb{O}|-1}{|\hat{\mathbb{O}}|-1}\sum_{(i,j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}+\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\ell^-_{ij}-\sum_{(i,j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}\right)\\
=&(\frac{|\mathbb{O}|-1}{|\hat{\mathbb{O}}|-1}-1)\sum_{(i,j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}+\sum_{i=1}^{m}\sum_{j=1}^{n}\ell^-_{ij}\\
=&\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\hat{\mathbb{O}}|-1}\sum_{(i,j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}+\sum_{i=1}^{m}\sum_{j=1}^{n}\ell^-_{ij}.
\end{aligned}$$

Then,

$$\begin{aligned}
&\frac{1}{mn}\left(\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\ell^+_{ij}\right]+\frac{|\mathbb{O}|(|\mathbb{O}|-1)}{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}\right]\right.\\
&\left.-\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\left(\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\hat{\mathbb{O}}|-1}+1\right)\ell^-_{ij}\right]\right)\\
=&\frac{1}{mn}\left(\sum_{(i,j)\in\mathbb{O}}\ell^+_{ij}+\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\hat{\mathbb{O}}|-1}\sum_{(i,j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}+\sum_{i=1}^{m}\sum_{j=1}^{n}\ell^-_{ij}-\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\hat{\mathbb{O}}|-1}\sum_{(i,j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\ell^-_{ij}-\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}\right)\\
=&\frac{1}{mn}\left(\sum_{(i,j)\in\mathbb{O}}\ell^+_{ij}+\sum_{i=1}^{m}\sum_{j=1}^{n}\ell^-_{ij}-\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}\right).
\end{aligned}$$

E.6 The Cost-Sensitive Batch Loss (25) and Its Expectation

Let ω be the weight hyper-parameter. We define a cost-sensitive batch loss as

$$\begin{aligned}
\hat{L}_{\text{Unbiased-}\omega}(\boldsymbol{\theta})=&\frac{1}{mn}\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\left(\sum_{(i,j)\in\hat{\mathbb{O}}}\ell^+_{ij}+\omega\frac{|\mathbb{O}|-1}{|\hat{\mathbb{O}}|-1}\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\frac{\ell^-_{ij}}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\right.\\
&\left.-\omega\sum_{(i,j)\in\hat{\mathbb{O}}}\left(\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\hat{\mathbb{O}}|-1}+1\right)\ell^-_{ij}\right).
\end{aligned}$$

Taking expectation on $\hat{L}_{\text{Unbiased-}\omega}(\boldsymbol{\theta})$, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{mn}\left(\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\sum_{(i,j)\in\hat{\mathbb{O}}}\ell_{ij}^{+}+\omega\frac{|\mathbb{O}|(|\mathbb{O}|-1)}{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell_{ij}^{-}\right.\right.\\
&\left.\left.-\omega\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\sum_{(i,j)\in\hat{\mathbb{O}}}\left(\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\hat{\mathbb{O}}|-1}+1\right)\ell_{ij}^{-}\right)\right]\\
=&\frac{1}{mn}\left(\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\ell_{ij}^{+}\right]+\omega\frac{|\mathbb{O}|(|\mathbb{O}|-1)}{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell_{ij}^{-}\right]\right.\\
&\left.-\omega\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\left(\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\hat{\mathbb{O}}|-1}+1\right)\ell_{ij}^{-}\right]\right).
\end{aligned}$$

The expectations are the same as those in Appendix E.5. So we have

$$\begin{aligned}
&\frac{1}{mn}\left(\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\ell_{ij}^{+}\right]+\omega\frac{|\mathbb{O}|(|\mathbb{O}|-1)}{|\hat{\mathbb{O}}|(|\hat{\mathbb{O}}|-1)}\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}}\sum_{(i',j)\in\hat{\mathbb{O}}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell_{ij}^{-}\right]\right.\\
&\left.-\omega\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}|}\mathbb{E}\left[\sum_{(i,j)\in\hat{\mathbb{O}}}\left(\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\hat{\mathbb{O}}|-1}+1\right)\ell_{ij}^{-}\right]\right)\\
=&\frac{1}{mn}\left(\sum_{(i,j)\in\mathbb{O}}\ell_{ij}^{+}+\omega\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\hat{\mathbb{O}}|-1}\sum_{(i,j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell_{ij}^{-}+\omega\sum_{i=1}^{m}\sum_{j=1}^{n}\ell_{ij}^{-}\right.\\
&\left.-\omega\frac{|\mathbb{O}|-|\hat{\mathbb{O}}|}{|\hat{\mathbb{O}}|-1}\sum_{(i,j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|}\ell_{ij}^{-}-\omega\sum_{(i,j)\in\mathbb{O}}\ell_{ij}^{-}\right)\\
=&\frac{1}{mn}\left(\sum_{(i,j)\in\mathbb{O}}\ell_{ij}^{+}+\omega\sum_{i=1}^{m}\sum_{j=1}^{n}\ell_{ij}^{-}-\omega\sum_{(i,j)\in\mathbb{O}}\ell_{ij}^{-}\right)\\
=&\frac{1}{mn}\left(\sum_{(i,j)\in\mathbb{O}}\ell_{ij}^{+}+\omega\sum_{(i,j)\notin\mathbb{O}}\ell_{ij}^{-}\right).
\end{aligned}$$

F More on Objective Functions

F.1 Time and Space Complexity of (17)

In this section, we analyze the time and space complexity of $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ and show that it has the same time complexity as $\hat{L}_{\text{In-Batch}}(\boldsymbol{\theta})$.

Compared to $\hat{L}_{\text{In-Batch}}(\boldsymbol{\theta})$, $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ in (17) introduces two additional types of terms: those involving $|\hat{\mathbb{O}}|$ and $|\mathbb{O}|$, and those involving $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$. Since all the terms involving $|\hat{\mathbb{O}}|$ and $|\mathbb{O}|$ are scalars, they do not affect the computational complexity. We therefore focus on the terms involving $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$.

A straightforward implementation is to compute $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$ on the fly from $\boldsymbol{Y}$ each time $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ is evaluated. However, this implementation is inefficient. For a fixed label matrix Y, the quantities $|\mathbb{O}_{i,:}|$ and $|\mathbb{O}_{:,j}|$ remain constant for each given i and j. Recomputing recomputation is therefore redundant. A more efficient strategy is to precompute $|\mathbb{O}_{i,:}|$ for all $i\in[m]$ and $|\mathbb{O}_{:,j}|$ for all $j\in[n]$, and store them for reuse during training. The strategy requires only $O(m+n)$ additional space.

With the caches of $|\mathbb{O}_{i,:}|$ for all $i\in[m]$ and $|\mathbb{O}_{:,j}|$ for all $j\in[n]$, the terms involving $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$ can be accessed in

Table 4: Best P@5 on the testing data for Datasets `EUR-Lex` and `ml1m`. “SQ” and “LR” mean the square and logistic losses.

Batch Ratio	Batch Loss	`EUR-Lex`		`ml1m`	
		SQ	LR	SQ	LR
10^{-5}	$\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$	36.31	50.75	18.03	18.71
	$\hat{L}_{\text{Sogram}}(\boldsymbol{\theta})$	36.17	50.51	18.07	18.16
10^{-4}	$\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$	36.27	50.42	18.03	18.99
	$\hat{L}_{\text{Sogram}}(\boldsymbol{\theta})$	36.28	49.79	18.01	18.77
10^{-3}	$\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$	36.73	51.72	18.45	18.94
	$\hat{L}_{\text{Sogram}}(\boldsymbol{\theta})$	36.64	51.32	18.26	18.65
10^{-2}	$\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$	36.78	52.35	18.18	19.08
	$\hat{L}_{\text{Sogram}}(\boldsymbol{\theta})$	36.75	52.49	18.27	18.97
10^{-1}	$\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$	36.84	53.28	18.50	18.62
	$\hat{L}_{\text{Sogram}}(\boldsymbol{\theta})$	36.59	53.45	18.49	18.75

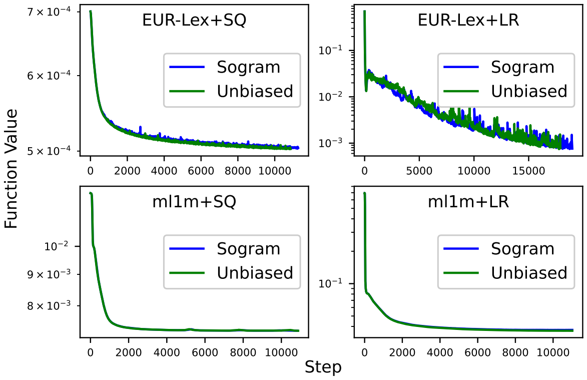


Figure 4: The function value versus training step on $\hat{L}_{\text{Sogram}}(\boldsymbol{\theta})$ and $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$. We consider two datasets, `EUR-Lex` and `ml1m`, and two losses, SQ: square loss and LR: logistic regression loss.

constant time for any $(i,j) \in [m] \times [n]$. Consequently, computing the term

$$\sum_{(i,j')\in\hat{\mathbb{O}}} \sum_{(i',j)\in\hat{\mathbb{O}}} \frac{1}{|\mathbb{O}_{i,:}|} \frac{1}{|\mathbb{O}_{:,j}|} \ell_{ij}^{-}$$

in $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ requires $O(|\hat{\mathbb{O}}|^2)$ time. Therefore, $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ and $\hat{L}_{\text{In-Batch}}(\boldsymbol{\theta})$ still have the same time complexity.

F.2 An Extended Discussion on the Softmax Loss

This work considers point-wise losses, which are still widely used in recommendation systems and multi-label classification. We have already proven that an unbiased batch loss exists for in-batch subsampling. We pointed out in Section 7 that studying other losses is a future direction. However, some losses for two-tower models are inherently biased. For example, Softmax, a list-wise loss, has been considered with in-batch sampling to train two-tower models in contrastive learning (Awasthi et al., 2022) and semantic textual similarity (Reimers and Gurevych, 2019). Lin et al. (2025) have proven that for the Softmax loss, it is impossible to find an unbiased batch loss as what we did in this work.

F.3 Compare (17) with Another Unbiased Batch Loss

Besides $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ in (17), Yuan et al. (2021) and Krichene et al. (2019) derived another unbiased batch loss for two-tower models under point-wise losses. Different from $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ or $\hat{L}_{\text{In-Batch}}(\boldsymbol{\theta})$, their approach requires two sets $\hat{\mathbb{O}}_1$ and $\hat{\mathbb{O}}_2$ from $\mathbb{O}$. We refer to their batch loss as $\hat{L}_{\text{Sogram}}(\boldsymbol{\theta})$, which takes the following form:

$$\hat{L}_{\text{Sogram}}(\boldsymbol{\theta}) := \frac{1}{mn}\left(\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}_1|}\sum_{(i,j')\in\hat{\mathbb{O}}_1}\left(\ell^+_{ij'}-\ell^-_{ij'}\right)+\frac{|\mathbb{O}|^2}{|\hat{\mathbb{O}}_1||\hat{\mathbb{O}}_2|}\sum_{(i,j')\in\hat{\mathbb{O}}_1}\sum_{(i',j)\in\hat{\mathbb{O}}_2}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}\right), \tag{34}$$

The proof in Appendix F.4 shows that $\mathbb{E}\left[(34)\right] = (5)$, the objective that we intend to minimize.

However, (34) is significantly more expensive than (17). For the first term in (34), we compute all $|\hat{\mathbb{O}}_1|$ left and right entities appearing in $\hat{\mathbb{O}}_1$ while for the second term in (34), we need to additionally compute all $|\hat{\mathbb{O}}_2|$ right entities appearing in $\hat{\mathbb{O}}_2$. Because $|\hat{\mathbb{O}}_1|^2 = |\hat{\mathbb{O}}_2|^2 = |\mathbb{B}|$, the cost of (34) is

$$|\hat{\mathbb{O}}_1|F(f)+|\hat{\mathbb{O}}_1|F(g)+|\hat{\mathbb{O}}_2|F(g)=\sqrt{|\mathbb{B}|}F(f)+2\sqrt{|\mathbb{B}|}F(g), \tag{35}$$

which is $\sqrt{|\mathbb{B}|}F(g)$ more than the value in (10) that indicates the cost of (17).

Merely comparing computational costs per step between (34) and (17) is insufficient. We must also evaluate the convergence speed and the model performance. By using the same experimental setting in Section 6, we show the comparison between $\hat{L}_{\text{Sogram}}(\boldsymbol{\theta})$ and $\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$ in Table 4. Clearly, their model performance is similar. To check the convergence speed, Figure 4 shows the function value versus the training step. For the batch size, we consider “batch ratio$=10^{-3}$” in (18). Both losses yield similar convergence speed. In conclusion, given the similar performance and convergence speed, the proposed unbiased loss (17) in this work is preferable due to its lower computational cost at each training step.

F.4 Proof of $\mathbb{E}\left[(34)\right] = (5)$

Taking expectation on (34), we have

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{mn}\left(\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}_1|}\sum_{(i,j')\in\hat{\mathbb{O}}_1}\left(\ell^+_{ij'}-\ell^-_{ij'}\right)+\frac{|\mathbb{O}|^2}{|\hat{\mathbb{O}}_1||\hat{\mathbb{O}}_2|}\sum_{(i,j')\in\hat{\mathbb{O}}_1}\sum_{(i',j)\in\hat{\mathbb{O}}_2}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}\right)\right]\\ =&\frac{1}{mn}\left(\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}_1|}\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}_1}\left(\ell^+_{ij'}-\ell^-_{ij'}\right)\right]+\frac{|\mathbb{O}|^2}{|\hat{\mathbb{O}}_1||\hat{\mathbb{O}}_2|}\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}_1}\sum_{(i',j)\in\hat{\mathbb{O}}_2}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}\right]\right).\end{aligned}$$

Applying Lemma E.5, the first expectation is

$$\begin{aligned}\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}_1|}\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}_1}\left(\ell^+_{ij'}-\ell^-_{ij'}\right)\right]&=\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}_1|}\frac{|\hat{\mathbb{O}}_1|}{|\mathbb{O}|}\sum_{(i,j')\in\mathbb{O}}\left(\ell^+_{ij'}-\ell^-_{ij'}\right)\\&=\sum_{(i,j')\in\mathbb{O}}\left(\ell^+_{ij'}-\ell^-_{ij'}\right).\end{aligned}$$

For the second expectation,

$$\begin{aligned}\frac{|\mathbb{O}|^2}{|\hat{\mathbb{O}}_1||\hat{\mathbb{O}}_2|}\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}_1}\sum_{(i',j)\in\hat{\mathbb{O}}_2}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}\right]&=\frac{|\mathbb{O}|^2}{|\hat{\mathbb{O}}_1||\hat{\mathbb{O}}_2|}\mathbb{E}\left[\sum_{(i,j')\in\mathbb{O}}\sum_{(i',j)\in\mathbb{O}}\mathbb{1}_{(i,j')\in\hat{\mathbb{O}}_1}\cdot\mathbb{1}_{(i',j)\in\hat{\mathbb{O}}_2}\cdot\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}\right]\\&=\frac{|\mathbb{O}|^2}{|\hat{\mathbb{O}}_1||\hat{\mathbb{O}}_2|}\sum_{(i,j')\in\mathbb{O}}\sum_{(i',j)\in\mathbb{O}}\mathbb{E}\left[\mathbb{1}_{(i,j')\in\hat{\mathbb{O}}_1}\cdot\mathbb{1}_{(i',j)\in\hat{\mathbb{O}}_2}\right]\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}.\end{aligned}$$

Since $\hat{\mathbb{O}}_1$ and $\hat{\mathbb{O}}_2$ are independently sampled from $\mathbb{O}$, we have

$$\begin{aligned}\mathbb{E}_{\hat{\mathbb{O}}_1,\hat{\mathbb{O}}_2}\left[\mathbb{1}_{(i,j')\in\hat{\mathbb{O}}_1}\cdot\mathbb{1}_{(i',j)\in\hat{\mathbb{O}}_2}\right]&=\mathbb{E}\left[\mathbb{1}_{(i,j')\in\hat{\mathbb{O}}_1\wedge(i',j)\in\hat{\mathbb{O}}_2}\right]\\&=P((i,j')\in\hat{\mathbb{O}}_1\wedge(i',j)\in\hat{\mathbb{O}}_2)\\&=P((i,j')\in\hat{\mathbb{O}}_1)P((i',j)\in\hat{\mathbb{O}}_2)\\&=\frac{|\hat{\mathbb{O}}_1|}{|\mathbb{O}|}\cdot\frac{|\hat{\mathbb{O}}_2|}{|\mathbb{O}|},\end{aligned}$$

where $P(\cdot)$ means the probability. Then,

$$\begin{aligned}\frac{|\mathbb{O}|^2}{|\hat{\mathbb{O}}_1||\hat{\mathbb{O}}_2|}\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}_1}\sum_{(i',j)\in\hat{\mathbb{O}}_2}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}\right] &= \frac{|\mathbb{O}|^2}{|\hat{\mathbb{O}}_1||\hat{\mathbb{O}}_2|}\sum_{(i,j')\in\mathbb{O}}\sum_{(i',j)\in\mathbb{O}}\frac{|\hat{\mathbb{O}}_1|}{|\mathbb{O}|}\cdot\frac{|\hat{\mathbb{O}}_2|}{|\mathbb{O}|}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}\\ &= \sum_{(i,j')\in\mathbb{O}}\sum_{(i',j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}\\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}\ell^-_{ij},\end{aligned}$$

where the last equality follows from

$$\sum_{(i,j')\in\mathbb{O}}\sum_{(i',j)\in\mathbb{O}}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}f(i,j) = \sum_{i=1}^{m}\sum_{j=1}^{n}f(i,j).$$

The above equation holds if we let $g(i,j) := \frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}f(i,j)$ and apply Lemma E.4 to the function g with the assumption that $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}| > 0$ for all (i,j). Consequently,

$$\begin{aligned}&\frac{1}{mn}\left(\frac{|\mathbb{O}|}{|\hat{\mathbb{O}}_1|}\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}_1}\left(\ell^+_{ij'}-\ell^-_{ij'}\right)\right]+\frac{|\mathbb{O}|^2}{|\hat{\mathbb{O}}_1||\hat{\mathbb{O}}_2|}\mathbb{E}\left[\sum_{(i,j')\in\hat{\mathbb{O}}_1}\sum_{(i',j)\in\hat{\mathbb{O}}_2}\frac{1}{|\mathbb{O}_{i,:}|}\frac{1}{|\mathbb{O}_{:,j}|}\ell^-_{ij}\right]\right)\\ =&\frac{1}{mn}\left(\sum_{(i,j')\in\mathbb{O}}\left(\ell^+_{ij'}-\ell^-_{ij'}\right)+\sum_{i=1}^{m}\sum_{j=1}^{n}\ell^-_{ij}\right)\\ =&\frac{1}{mn}\left(\sum_{(i,j)\in\mathbb{O}}\ell^+_{ij}+\sum_{i=1}^{m}\sum_{j=1}^{n}\ell^-_{ij}-\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}\right).\end{aligned}$$

G Full Experimental Results

G.1 Full Results of Using Stochastic Gradient

Details of experimental settings are in Appendix D. All results are shown in Table 5.

In this table, most observations align with the findings from Sections 6.1 and 6.2. The only exception occurs for the dataset `EUR-Lex` under the square loss, where $\hat{L}_{\text{In-Batch}}(\boldsymbol{\theta})$ gives the best P@1 and R@1 under batch ratio $= 10^{-3}$. However, we argue that this superior performance is not generalizable due to the following reasons.

- The superior performance does not hold for the logistic regression loss on the same dataset. Moreover, in the `ml1m` dataset, which has a larger (and therefore, more harmful) $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$, $\hat{L}_{\text{In-Batch}}(\boldsymbol{\theta})$'s performance remains very poor.
- For evaluation criteria other than P@1 and R@1, $\hat{L}_{\text{In-Batch}}(\boldsymbol{\theta})$'s results do not stand out.

G.2 Full Results of Using Full Gradient

We additionally provide the results of using full gradient for solving the optimization problem. This setting avoids the sampling variance in stochastic gradient methods. Because we no longer consider a batch, $|\hat{\mathbb{O}}| = |\mathbb{O}|$. In this case, $L_{\text{In-Batch}}(\boldsymbol{\theta})$ reduces to $L_{\text{Popularity}}(\boldsymbol{\theta})$ while $L_{\text{Pos-Neg}}(\boldsymbol{\theta})$ reduces to $L_{\text{Ubiased}}(\boldsymbol{\theta})$. However, these reductions are not what we want. Our actual goal is to analyze $L_{\text{In-Batch}}(\boldsymbol{\theta})$ and $L_{\text{Pos-Neg}}(\boldsymbol{\theta})$ corresponding to a specific batch size $|\hat{\mathbb{O}}|$ (or equivalently, a specific batch ratio $|\hat{\mathbb{O}}|^2/|\mathbb{O}|^2$). To emulate the behavior of the objectives corresponding to different batch ratios while still using full gradient, we replace the factor $\frac{|\hat{\mathbb{O}}|-1}{|\mathbb{O}|-1}$ in $L_{\text{In-Batch}}(\boldsymbol{\theta})$ and $L_{\text{Pos-Neg}}(\boldsymbol{\theta})$ with a tunable factor ω, leading to

$$L_{\text{In-Batch-}\omega}(\boldsymbol{\theta}) = \frac{1}{mn}\left(\sum_{(i,j)\in\mathbb{O}}\ell^+_{ij}-\omega\sum_{(i,j)\in\mathbb{O}}\ell^-_{ij}+\omega\sum_{i=1}^{m}\sum_{j=1}^{n}|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|\ell^-_{ij}\right),$$

and

$$L_{\text{Pos-Neg-}\omega}(\boldsymbol{\theta}) = \frac{1}{mn}\left(\sum_{(i,j)\in\mathbb{O}} \ell_{ij}^{+} + \omega \sum_{(i,j)\notin\mathbb{O}} \ell_{ij}^{-}\right).$$

Then, to simulate the situation of using a batch ratio $|\hat{\mathbb{O}}|^2/|\mathbb{O}|^2$ via stochastic gradient, we follow (16) and (18) to set

$$\omega = \frac{|\hat{\mathbb{O}}| - 1}{|\mathbb{O}| - 1}. \tag{36}$$

Note that $L_{\text{Pos-Neg-}\omega}(\boldsymbol{\theta})$ is the same as $L_{\text{Ubiased-}\omega}(\boldsymbol{\theta})$ in (19).

For optimization, we used the limited-memory BFGS method with line search (Liu and Nocedal, 1989). The stopping condition is

$$\|\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t)\| \leq 10^{-7}.$$

where $\boldsymbol{\theta}_t$ is the results at the t-th iteration.

Results in Table 6 roughly align with our earlier observations of running stochastic gradient methods.

- $L_{\text{Pos-Neg-}\omega}(\boldsymbol{\theta})$ outperforms $L_{\text{In-Batch-}\omega}(\boldsymbol{\theta})$, and $L_{\text{Unbiased}}(\boldsymbol{\theta})$ outperforms $L_{\text{Popularity}}(\boldsymbol{\theta})$. Notably, the better-performing methods are those without the term $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$. This results reinforces the earlier conclusion about the negative impact of the term $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$.
- $L_{\text{In-Batch-}\omega}(\boldsymbol{\theta})$'s superiority over $L_{\text{Popularity}}(\boldsymbol{\theta})$ suggests that $\omega < 1$ mitigates the negative influence of the term $|\mathbb{O}_{i,:}||\mathbb{O}_{:,j}|$.
- The good performance under suitable ω of $L_{\text{Pos-Neg-}\omega}(\boldsymbol{\theta})$, which is also $L_{\text{Ubiased-}\omega}(\boldsymbol{\theta})$, indicates that we can select proper ω (for example, via validation) to conduct cost-sensitive learning on positive/negative pairs.

Table 5: Results by using stochastic gradient. We show P@{1, 5, 25} and R@{1, 5, 25} on the test sets. "SQ" and "LR" mean the square and logistic regression losses. Results marked with "*" indicate models that have optimization issues. The underlined values are the best ones within the same column.

(a) P@1 and R@1

Batch Loss	Batch Ratio	P@1				R@1			
		`EUR-Lex`		`ml1m`		`EUR-Lex`		`ml1m`	
		SQ	LR	SQ	LR	SQ	LR	SQ	LR
$\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$	10^{-5}	59.25	72.83*	25.54	26.49	11.69	14.78*	3.50	3.81
	10^{-4}	59.38	74.85*	25.57	26.51	11.74	15.17*	3.51	3.72
	10^{-3}	59.92	76.95*	26.72	23.86	11.85	15.66*	3.71	3.24
	10^{-2}	60.31	77.98*	25.94	26.30	11.95	15.84*	3.55	3.65
	10^{-1}	60.23	78.53	26.39	26.49	11.92	15.95	3.56	3.81
$\hat{L}_{\text{In-Batch}}(\boldsymbol{\theta})$	10^{-5}	31.10	63.23	0.58	10.62	6.27	12.72	0.11	1.88
	10^{-4}	60.93	62.85	0.40	10.74	12.29	12.63	0.08	1.89
	10^{-3}	67.68	60.26	0.28	10.27	13.75	12.10	0.03	1.85
	10^{-2}	63.65	58.40	0.28	0.85	12.99	11.74	0.03	0.27
	10^{-1}	50.82	56.51	0.28	0.59	10.38	11.40	0.03	0.10
$\hat{L}_{\text{Popularity}}(\boldsymbol{\theta})$	10^{-5}	29.91	23.00	0.28	5.42	6.13	4.51	0.03	0.75
	10^{-4}	32.03	25.30	0.28	3.15	6.56	4.93	0.03	0.36
	10^{-3}	32.70	22.28	0.28	3.49	6.67	4.33	0.03	0.49
	10^{-2}	31.28	0.62	0.28	0.66	6.40	0.10	0.03	0.11
	10^{-1}	31.85	40.91	0.28	0.28	6.51	8.18	0.03	0.03
$\hat{L}_{\text{Pos-Neg}}(\boldsymbol{\theta})$	10^{-5}	54.83	75.11	8.79	21.58	11.09	15.22	1.72	3.16
	10^{-4}	58.89	75.55	15.98	23.18	11.95	15.31	2.62	3.49
	10^{-3}	62.64	76.51	20.08	22.19	12.69	15.50	3.14	3.31
	10^{-2}	64.48	76.79	23.29	23.90	13.02	15.60	3.80	3.46
	10^{-1}	63.21	77.72	26.44	27.44	12.68	15.75	4.09	4.10

(b) P@5 and R@5

Batch Loss	Batch Ratio	P@5				R@5			
		`EUR-Lex`		`ml1m`		`EUR-Lex`		`ml1m`	
		SQ	LR	SQ	LR	SQ	LR	SQ	LR
$\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$	10^{-5}	36.31	50.75*	18.03	18.71	34.92	43.88*	11.85	12.53
	10^{-4}	36.27	50.42*	18.03	18.99	34.87	49.36*	11.77	12.55
	10^{-3}	36.73	51.72*	18.45	18.94	35.31	50.58*	12.19	10.39
	10^{-2}	36.78	52.35*	18.18	19.08	35.33	51.28*	11.92	11.80
	10^{-1}	36.84	53.28	18.50	18.62	35.44	52.20	12.01	12.61
$\hat{L}_{\text{In-Batch}}(\boldsymbol{\theta})$	10^{-5}	24.63	45.05	0.54	7.45	24.43	44.07	0.44	6.58
	10^{-4}	35.26	44.87	0.40	8.23	34.50	43.84	0.32	6.89
	10^{-3}	37.68	43.65	0.34	7.44	37.11	42.72	0.17	6.11
	10^{-2}	34.11	37.47	0.34	0.59	33.65	36.79	0.17	0.93
	10^{-1}	28.08	36.68	0.34	0.34	27.75	35.98	0.17	0.26
$\hat{L}_{\text{Popularity}}(\boldsymbol{\theta})$	10^{-5}	18.02	12.34	0.34	4.13	17.82	11.77	0.17	2.63
	10^{-4}	19.09	13.89	0.34	2.34	18.83	13.32	0.17	1.44
	10^{-3}	19.32	11.09	0.34	2.38	19.02	10.75	0.17	1.48
	10^{-2}	17.04	33.90	0.34	0.34	16.87	0.51	0.17	0.29
	10^{-1}	17.74	24.78	0.34	0.34	17.51	24.26	0.17	0.17
$\hat{L}_{\text{Pos-Neg}}(\boldsymbol{\theta})$	10^{-5}	43.25	52.86	7.29	15.74	42.61	51.92	6.67	11.15
	10^{-4}	44.57	52.80	12.03	17.12	43.78	51.73	9.44	12.05
	10^{-3}	44.83	52.99	14.40	18.02	43.88	51.94	10.87	11.12
	10^{-2}	43.68	52.94	16.81	17.14	42.56	51.92	12.50	11.70
	10^{-1}	40.67	53.20	18.69	19.22	39.38	52.19	13.38	13.29

(c) P@25 and P@25

Batch Loss	Batch Ratio	P@25				R@25			
		EUR-Lex		ml1m		EUR-Lex		ml1m	
		SQ	LR	SQ	LR	SQ	LR	SQ	LR
$\hat{L}_{\text{Unbiased}}(\boldsymbol{\theta})$	10^{-5}	12.36	13.67*	10.08	10.44	58.74	65.35*	30.45	32.25
	10^{-4}	12.41	15.55*	10.11	10.41	58.96	74.43*	30.64	32.20
	10^{-3}	12.42	15.81*	10.23	9.35	59.06	75.74*	31.07	27.99
	10^{-2}	12.51	16.15*	10.21	10.15	59.49	77.33*	30.76	30.77
	10^{-1}	12.52	16.35	10.27	10.41	59.57	78.33	31.12	32.22
$\hat{L}_{\text{In-Batch}}(\boldsymbol{\theta})$	10^{-5}	10.39	14.26	0.59	4.71	50.41	68.50	2.34	18.66
	10^{-4}	12.26	14.20	0.47	5.19	59.16	68.14	1.93	19.65
	10^{-3}	12.17	14.04	0.33	4.70	58.77	67.39	0.84	17.52
	10^{-2}	10.67	13.15	0.33	0.36	51.82	63.27	0.84	2.67
	10^{-1}	8.68	13.02	0.33	0.33	42.09	62.59	0.84	0.84
$\hat{L}_{\text{Popularity}}(\boldsymbol{\theta})$	10^{-5}	5.94	4.97	0.33	3.14	28.92	23.57	0.84	8.60
	10^{-4}	6.26	5.67	0.33	1.70	30.50	26.99	0.84	4.97
	10^{-3}	6.21	4.38	0.33	1.68	30.19	21.09	0.84	4.37
	10^{-2}	5.16	0.34	0.33	0.33	25.17	1.44	0.84	0.86
	10^{-1}	5.42	9.27	0.33	0.33	26.40	44.53	0.84	0.84
$\hat{L}_{\text{Pos-Neg}}(\boldsymbol{\theta})$	10^{-5}	$\underline{16.42}$	$\underline{16.83}$	5.29	9.30	$\underline{78.75}$	$\underline{80.62}$	20.63	30.09
	10^{-4}	16.40	16.73	7.63	10.01	78.52	80.14	27.14	31.99
	10^{-3}	15.99	16.68	8.86	9.40	76.55	79.96	30.15	30.07
	10^{-2}	15.16	16.58	9.94	9.96	72.45	79.51	33.45	31.31
	10^{-1}	13.90	16.38	$\underline{10.68}$	$\underline{10.79}$	66.20	78.55	$\underline{34.04}$	$\underline{33.83}$

Table 6: Results by using full gradient. We show P@{1, 5, 25} and R@{1, 5, 25} on the test sets. The batch ratios here are emulated by ω via (18) and (36), rather than by explicit sampling. "SQ" and "LR" mean the square and logistic regression losses. The underlined values are the best ones within the same column.

(a) P@1 and R@1

Objective	Batch Ratio Emulated by ω	P@1				R@1			
		EUR-Lex		ml1m		EUR-Lex		ml1m	
		SQ	LR	SQ	LR	SQ	LR	SQ	LR
$L_{\text{Unbiased}}(\boldsymbol{\theta})$	-	59.35	77.59	24.23	22.78	11.74	15.74	3.30	3.08
$L_{\text{In-Batch-}\omega}(\boldsymbol{\theta})$	10^{-5}	55.73	60.41	0.82	10.16	11.30	12.13	0.11	1.72
	10^{-4}	61.58	60.26	0.35	9.52	12.50	12.13	0.07	1.75
	10^{-3}	67.63	60.16	0.28	8.82	13.74	12.07	0.03	1.55
	10^{-2}	61.94	60.16	0.28	8.92	12.70	12.04	0.03	1.66
	10^{-1}	47.79	60.05	0.28	8.23	9.80	12.05	0.03	1.49
$L_{\text{Popularity}}(\boldsymbol{\theta})$	-	30.87	52.45	0.28	7.69	6.32	10.51	0.03	1.47
$L_{\text{Pos-Neg-}\omega}(\boldsymbol{\theta})$	10^{-5}	53.69	73.89	5.93	16.05	10.81	15.02	1.26	2.41
	10^{-4}	57.77	74.20	10.28	17.40	11.66	15.04	1.66	2.33
	10^{-3}	62.17	74.49	14.61	18.79	12.56	15.12	2.56	2.62
	10^{-2}	62.82	76.02	21.63	18.70	12.60	15.41	3.67	2.58
	10^{-1}	62.51	77.08	26.18	21.79	12.53	15.67	3.98	2.86

(b) P@5 and R@5

Objective	Batch Ratio Emulated by ω	P@5				R@5			
		EUR-Lex		ml1m		EUR-Lex		ml1m	
		SQ	LR	SQ	LR	SQ	LR	SQ	LR
$L_{\text{Unbiased}}(\boldsymbol{\theta})$	-	37.01	51.12	17.07	16.35	35.66	50.01	11.47	10.55
$L_{\text{In-Batch-}\omega}(\boldsymbol{\theta})$	10^{-5}	33.04	43.93	0.85	7.39	32.45	42.96	0.60	6.03
	10^{-4}	34.89	44.09	0.34	7.23	34.22	43.10	0.25	5.91
	10^{-3}	36.05	44.46	0.34	6.70	35.43	43.42	0.17	5.82
	10^{-2}	32.88	44.53	0.34	6.77	32.50	43.50	0.17	5.82
	10^{-1}	26.28	44.34	0.34	6.16	25.99	43.26	0.17	5.92
$L_{\text{Popularity}}(\boldsymbol{\theta})$	-	18.22	38.35	0.34	5.89	17.97	37.55	0.17	5.62
$L_{\text{Pos-Neg-}\omega}(\boldsymbol{\theta})$	10^{-5}	42.63	49.82	5.77	11.93	41.99	48.77	5.67	8.40
	10^{-4}	44.13	49.65	8.34	13.23	43.36	48.66	6.97	9.00
	10^{-3}	44.53	49.88	11.70	13.86	43.53	48.90	9.87	8.77
	10^{-2}	43.34	50.60	16.21	13.98	42.19	49.59	12.40	9.69
	10^{-1}	40.65	50.79	18.68	15.45	39.40	49.80	13.40	9.70

(c) P@25 and R@25

Objective	Batch Ratio Emulated by ω	P@25				R@25			
		EUR-Lex		ml1m		EUR-Lex		ml1m	
		SQ	LR	SQ	LR	SQ	LR	SQ	LR
$L_{\text{Unbiased}}(\boldsymbol{\theta})$	-	12.30	15.74	9.98	9.30	58.40	75.38	30.76	28.06
$L_{\text{In-Batch-}\omega}(\boldsymbol{\theta})$	10^{-5}	12.41	14.64	0.82	4.74	59.67	70.16	3.31	19.00
	10^{-4}	12.19	14.82	0.34	4.78	58.83	70.96	1.32	19.15
	10^{-3}	11.77	14.92	0.33	4.66	57.02	71.52	0.84	18.77
	10^{-2}	10.35	15.01	0.33	4.58	50.25	71.94	0.84	18.58
	10^{-1}	8.41	15.00	0.33	4.50	40.88	71.82	0.84	18.38
$L_{\text{Popularity}}(\boldsymbol{\theta})$	-	5.98	13.31	0.33	4.35	29.14	63.84	0.84	18.59
$L_{\text{Pos-Neg-}\omega}(\boldsymbol{\theta})$	10^{-5}	16.29	16.29	4.87	7.81	78.14	78.00	20.38	24.62
	10^{-4}	16.30	16.13	5.95	8.01	78.09	77.30	24.63	25.33
	10^{-3}	15.94	16.09	8.15	8.37	76.32	77.12	30.12	25.01
	10^{-2}	15.09	16.11	9.93	8.45	72.09	77.21	33.45	27.31
	10^{-1}	13.85	15.94	10.62	8.80	65.95	76.40	33.93	26.46