Towards Reliable Item Sampling for Recommendation Evaluation

Dong Li[1], **Ruoming Jin**[1], **Zhenming Liu**[2], **Bin Ren**[2], **Jing Gao**[3], **Zhi Liu**[3]

[1] Kent State University,
[2] College of William & Mary,
[3] iLambda,
{dli12, rjin1}@kent.edu, {zliu, bren}@cs.wm.edu, {jgao, zliu}@ilambda.com

Abstract

Since Rendle and Krichene argued that commonly used sampling-based evaluation metrics are "inconsistent" with respect to the global metrics (even in expectation), there have been a few studies on the sampling-based recommender system evaluation. Existing methods try either mapping the sampling-based metrics to their global counterparts or more generally, learning the empirical rank distribution to estimate the top-K metrics. However, despite existing efforts, there is still a lack of rigorous theoretical understanding of the proposed metric estimators, and the basic item sampling also suffers from the "blind spot" issue, i.e., estimation accuracy to recover the top-K metrics when K is small can still be rather substantial. In this paper, we provide an in-depth investigation into these problems and make two innovative contributions. First, we propose a new item-sampling estimator that explicitly optimizes the error with respect to the ground truth, and theoretically highlights its subtle difference against prior work. Second, we propose a new adaptive sampling method that aims to deal with the "blind spot" problem and also demonstrate the expectation-maximization (EM) algorithm can be generalized for such a setting. Our experimental results confirm our statistical analysis and the superiority of the proposed works. This study helps lay the theoretical foundation for adopting item sampling metrics for recommendation evaluation and provides strong evidence for making item sampling a powerful and reliable tool for recommendation evaluation.

Introduction

As personalization and recommendation continue to play an integral role in the emerging AI-driven economy (Fayyaz et al. 2020; Zhao et al. 2021; Peng, Sugiyama, and Mine 2022; Jin et al. 2021a; Chen et al. 2022), proper and rigorous evaluation of recommendation models become increasingly important in recent years for both academic researchers and industry practitioners (Gruson et al. 2019; Cremonesi et al. 2011; Dacrema et al. 2019; Rendle 2019). Particularly, ever since Krichene and Rendle (2020, 2022) pointed out the "inconsistent" issue of item-sampling based evaluation of commonly used (top-K) evaluation metrics, such as Recall (Hit-Ratio)/Precision, Average Precision (AP) and Normalized Discounted Cumulative Gain (NDCG), (other than AUC), it has emerged as a major controversy being hotly debated among recommendation community.

Specifically, the item-sampling strategy calculates the top-K evaluation metrics using only a small set of item samples (Koren 2008; Cremonesi, Koren, and Turrin 2010; He et al. 2017; Ebesu, Shen, and Fang 2018; Hu et al. 2018; Krichene et al. 2019; Wang et al. 2019; Yang et al. 2018a,b). Krichene and Rendle (2020, 2022) show that the top-K metrics based on the samples differ from the global metrics using all the items. They suggested a cautionary use (avoiding if possible) of the sampled metrics for recommendation evaluation. Due to the ubiquity of sampling methodology, it is not only of theoretical importance but also of practical interest to understand item-sampling evaluation. Indeed, since the number of items in any real-world recommendation system is typically quite large (easily in the order of tens of millions), efficient model evaluation based on item sampling can be very useful for recommendation researchers and practitioners.

To address the discrepancy between item sampling results and the exact top-K recommendation evaluation, Krichene and Rendle (2020) have proposed a few estimators in recovering global top-K metrics from sampling. Concurrently, Li et al. (2020) showed that for the top-K Hit-Ratios (Recalls) metric, there is an (approximately) linear relationship between the item-sampling top-k and the global top-K metrics ($K = f(k)$, where f is approximately linear). In another recent study, Jin et al. (2021b) developed solutions based on MLE (Maximal Likelihood Estimation) and ME (Maximal Entropy) to learn the *empirical rank distribution*, which is then used to estimate global top-K metrics.

Despite these latest works on item-sampling estimation (Li et al. 2020; Krichene and Rendle 2020; Jin et al. 2021b), there remain some major gaps in making item-sampling reliable and accurate for top-K metrics. Specifically, the following important problems are remaining open:

(i) What is the optimal estimator given the basic item sampling? All the earlier estimation methods do not establish any optimality results with respect to the estimation errors (Krichene and Rendle 2020; Jin et al. 2021b).

(ii) What can we do for the problem of the basic item sampling, which appears to have a fundamental limitation that prevents us from recovering the global rank distributions accurately? For the offline recommendation evaluation, we

typically are interested in the top-ranked items and top-K metrics, when K is relatively small, say less than 50. However, the current item sampling seems to have a "blind spot" for the top-rank distribution. For example, when there are $n = 100$ samples and $N = 10k$, the estimation granularity is only at around 1% $(1/n)$ level (Krichene and Rendle 2020; Li et al. 2020). We can only infer that the top items in the samples are top 1% (top 100) in the global rank, while we can not further tell whether the top items in the sample set are in, for example, top-50, without increasing the sampling size. Given this, even with the best estimator for the item sampling, we may still not be able to provide accurate results for the top-K metrics. A remedy is increasing the sampling size, but it can significantly increase the estimation cost too, limiting the benefits of item sampling. Can we sample the items in a more intelligent manner to circumvent the "blind spot" while keeping estimation cost low (and the sample size small)? To address the above open questions, we make the following contributions in this paper:

- We derive an optimal item-sampling estimator and highlight subtle differences from the BV estimators derived by Krichene and Rendle (2020), and point out the potential issues of BV estimator because it fails to link the user population size with the estimation variance. To the best of our knowledge, this is the first estimator that directly optimizes the estimation errors.
- We address the limitation of the current item sampling approaches by proposing a new adaptive sampling method. This provides a simple and effective remedy that helps avoid the sampling "blind spot", and significantly improves the accuracy of the estimated metrics with low sample complexity.
- We perform a thorough experimental evaluation of the proposed item-sampling estimator and the new adaptive sampling method. The experimental results further confirm the statistical analysis and the superiority of newly proposed estimators.

Our results help lay the theoretical foundation for adopting item sampling metrics for recommendation evaluation and offer a simple yet effective new adaptive sampling approach to help recommendation practitioners and researchers to apply item sampling-based approaches to speed up offline evaluation. The following is organized: Section Overview introduces the item sampling based top-K evaluation framework and reviews the related work (in Appendix); Section New Estimator introduces the new optimal estimator minimizing its mean squared error with respect to the ground-truth; Section Adaptive presents the new adaptive sampling and estimation method; Section Experiments discusses the experimental results; and finally, Section Conclusion concludes the paper. The full report is at https://arxiv.org/abs/2211.15743.

Overview

The appendix highlights the key notations for evaluating recommendation algorithms used throughout the paper. Given M users and N items. To evaluate the quality of recommender models, each testing user u hides an already clicked (or so-called target) item i_u, and compares it with the rest of the items, derives a rank R_u. The recommendation model is considered to be effective if it ranks i_u at the top (small R_u). Formally, given a recommendation model, a metric function (denoted as a metric $\mathcal{M}$) maps each rank R_u to a real-valued score, and then averages over all the users in the test set:

$$T = \frac{1}{M}\sum_{u=1}^{M}\mathcal{M}(R_u) \tag{1}$$

And the corresponding top-K evaluation metric:

$$T = \frac{1}{M}\sum_{u=1}^{M}\mathbf{1}_{R_u \le K}\cdot\mathcal{M}(R_u) \tag{2}$$

where $\mathbf{1}_x = 1$ if x is True, 0 otherwise.

The commonly used function $\mathcal{M}$ of evaluation metrics (Krichene and Rendle 2020) are Recall, Precision, AUC, NDCG, and AP. For example:

$$Recall@K = \frac{1}{M}\sum_{u=1}^{M}\mathbf{1}_{R_u \le K} \tag{3}$$

Item-Sampling Top-K Evaluation

In the item-sampling-based top-K evaluation scenario, for a given user u and his/her relevant item i_u, another $n-1$ items from the entire item set I are sampled. The union of sampled items and i_u is I_u ($i_u \in I_u$, $|I_u| = n$). The recommendation model then returns the rank of i_u among I_u, denoted as r_u (again, R_u is the rank against the entire set of items I).

Given this, a list of studies (Koren 2008; He et al. 2017) simply replaces R_u with r_u for (top-K) evaluation. Sampled evaluation metric/performance denoted as :

$$T_S \triangleq \frac{1}{M}\sum_{u=1}^{M}\mathbf{1}_{r_u \le K}\cdot\mathcal{M}(r_u) \tag{4}$$

It's obvious that r_u and R_u differ substantially, for example, $r_u \in [1, n]$ whereas $R_u \in [1, N]$. Therefore, for the same K, the item-sampling top-K metrics and the global top-K metrics correspond to distinct measures (no direct relationship): $T \neq T_S$ ($Recall@K \neq Recall_S@K$). This problem is highlighted in (Krichene and Rendle 2020; Rendle 2019), referring to these two metrics being *inconsistent*. From the perspective of statistical inference, the basic sampling-based top-K metric $T_S@K$ is not a reasonable or good *estimator* (Lehmann and Casella 2006) of $T@K$.

Li et al. (2020) showed that for some of the most commonly used metrics, the top-K Recall/HitRatio, there is a mapping function f (approximately linear), such that $Recall@f(k) \approx Recall_S@k$. Thus, they give an intuitive explanation on how to look at a sampling top-K metrics (on Recall) linking it the same global metric but at a different rank/location.

There are two recent works (see related works in Appendix) studying the general metric estimation problem based on the item sampling metrics. Specifically, given the sampling ranked results in the test set, $\{r_u\}_{u=1}^{M}$, how to infer/approximate the T from Eq. (1) or more commonly Eq. (2), without the knowledge $\{R_u\}_{u=1}^{M}$?

Two Problems: We consider two fundamental (open) questions for item-sampling estimation: 1) What is the optimal estimator following for (Krichene and Rendle 2020)? (Their methods do not directly target minimizing the estimation errors). 2) By solving the first problem and the MLE method from (Jin et al. 2021b), we observe the best effort using the basic item sampling still fails to recover accurately on the global top-K metrics when K is small, as well as the rank distribution for the top spots. Those "blind spots" seem to stem from the inherent (*information*) limitation of item sampling methods, not from the estimators. How can we effectively address such item-sampling limitations? Next, we will introduce methods to address these two problems.

New Estimator for Item-Sampling

In this section, we introduce a new estimator which aims to directly minimize the expected errors between the item-sampling-based top-K metrics and the global top-K metrics. Here, we consider a similar strategy as (Krichene and Rendle 2020) though our objective function is different and aims to explicitly minimize the expected error. We aim to search for a *sampled metric* $\widehat{\mathcal{M}}(r)$ to approach $\widehat{T} \approx T$:

$$\begin{aligned}\widehat{T} &= \sum_{r=1}^{n} \tilde{P}(r) \cdot \widehat{\mathcal{M}}(r) = \frac{1}{M}\sum_{u=1}^{M} \widehat{\mathcal{M}}(r_u) \\ &\approx \frac{1}{M}\sum_{u=1}^{M} \mathcal{M}(R_u) = \sum_{R=1}^{N} \tilde{P}(R) \cdot \mathcal{M}(R) = T\end{aligned}$$

where $\tilde{P}(r) = \frac{1}{M}\sum\limits_{r=1}^{M} \mathbf{1}_{r_u = r}$ is the empirical sampled rank distribution and $\widehat{\mathcal{M}}(r)$ is the adjusted discrete metric function. An immediate observation is:

$$\mathbb{E}\,\widehat{T} = \sum_{r=1}^{n} \mathbb{E}[\tilde{P}(r)] \cdot \widehat{\mathcal{M}}(r) = \sum_{r=1}^{n} P(r) \cdot \widehat{\mathcal{M}}(r) \tag{5}$$

Following the classical statistical inference (Casella and Berger 2002), the optimality of an estimator is measured by Mean Squared Error (more derivation in Appendix):

$$\mathbb{E}[\widehat{T} - \mathbb{E}\,T]^2 = \mathbb{E}[\widehat{T} - \sum_{R=1}^{N} P(R)\mathcal{M}(R)]^2 \tag{6}$$

$$\begin{aligned}=&\Big(\sum_{r=1}^{n}\sum_{R=1}^{N} P(r|R)P(R)\widehat{\mathcal{M}}(r) - \sum_{R=1}^{N} P(R)\mathcal{M}(R)\Big)^2 \\ +&\,\mathbb{E}[\sum_{r=1}^{n}\sum_{R=1}^{N} \tilde{P}(r|R)P(R)\widehat{\mathcal{M}}(r) - \sum_{r=1}^{n}\sum_{R=1}^{N} P(r|R)P(R)\widehat{\mathcal{M}}(r)]^2\end{aligned}$$

Remark that $\tilde{P}(r|R)$ is the empirical conditional sampling rank distribution given a global rank R. We next use Jensen's inequality to bound the first term in (Eq. (6)). Specifically, we may treat $\sum_{r=1}^{n} P(r|R)\widehat{\mathcal{M}}(r) - \mathcal{M}(R)$ as a random variable and use $(\mathbb{E}\,X)^2 \le \mathbb{E}\,X^2$ to obtain

$$\begin{aligned}&\Big(\sum_{r=1}^{n}\sum_{R=1}^{N} P(r|R)P(R)\widehat{\mathcal{M}}(r) - \sum_{R=1}^{N} P(R)\mathcal{M}(R)\Big)^2 \\ &\le \sum_{R=1}^{N} P(R)\Big(\sum_{r=1}^{n} P(r|R)\widehat{\mathcal{M}}(r) - \mathcal{M}(R)\Big)^2\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\mathbb{E}[\widehat{T} - \sum_{R=1}^{N} P(R)\mathcal{M}(R)]^2 \\ &\le \underbrace{\sum_{R=1}^{N} P(R)\Big\{\Big(\sum_{r=1}^{n} P(r|R)\widehat{\mathcal{M}}(r) - \mathcal{M}(R)\Big)^2}_{\mathcal{L}_1} \\ &\underbrace{+\,\mathbb{E}[\sum_{r=1}^{n} \tilde{P}(r|R)\widehat{\mathcal{M}}(r) - \sum_{r=1}^{n} P(r|R)\widehat{\mathcal{M}}(r)]^2\Big\}}_{\mathcal{L}_2}.\end{aligned}$$

Let $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$, which gives an upper bound on the expected MSE. Therefore, our goal is to find $\widehat{\mathcal{M}}(r)$ to minimize $\mathcal{L}$. We remark that a seemingly innocent application of Jensen's inequality results in an optimization objective that possesses a range of interesting properties:

1. Statistical structure. The objective has a variance-bias trade-off interpretation, i.e.,

$$\mathcal{L}_1 = \sum_{R=1}^{N} P(R)\Big(\mathbb{E}(\widehat{\mathcal{M}}(r)|R) - \mathcal{M}(R)\Big)^2 \tag{7}$$

$$\mathcal{L}_2 = \sum_{R=1}^{N} \frac{1}{M} Var(\widehat{\mathcal{M}}(r)|R) \tag{8}$$

where $\mathcal{L}_1$ can be interpreted as a bias term and $\mathcal{L}_2$ can be interpreted as a variance term. Note that while Krichene and Rendle (2020) also introduce a variance-bias tradeoff objective, their objective is constructed from heuristics and contains a hyper-parameter (that determines the relative weight between bias and variance) that needs to be tuned in an adhoc manner. Here, because our objective is constructed from direct optimization of the MSE, it is more principled and also removes dependencies on hyperparameters. See next Subsection for proving Eq. (8) (Eq. (7) is trivial) and Subsection Closed Form for more comparison against estimators proposed in (Krichene and Rendle 2020).

2. Algorithmic structure. while the objective is not convex, we show that the objective can be expressed in a compact manner using matrices and we can find the optimal solution in a fairly straightforward manner. In other words, Jensen's inequality substantially simplifies the computation at the cost of having a looser upper bound.

3. Practical performance. Our experiments also confirm that the new estimator is effective, which suggests that Jensen's inequality makes only inconsequential and moderate performance impact on the estimator's quality.

Analysis of $\mathcal{L}_2$

To analyze $\mathcal{L}_2$, let us take a close look at $\tilde{P}(r|R)$. Formally, let X_r be the random variable representing the number of items at rank r in the item-sampling data whose original rank in the entire item set is R. Then, we rewrite $\tilde{P}(r|R) = \frac{X_r}{M \cdot P(R)}$. Furthermore, it is easy to observe $(X_1, \cdots X_n)$ follows the multinomial distribution

$Multi(P(1|R), \cdots, P(n|R))$ (See $P(r|R)$ defined in Appendix. We also have:

$$\begin{aligned} \mathbb{E}[X_r] &= M \cdot P(R) \cdot P(r|R) \\ Var[X_r] &= M \cdot P(R) \cdot P(r|R)(1 - P(r|R)) \end{aligned} \tag{9}$$

Next, let us define a new random variable $\mathcal{B} \triangleq \sum_r^n \widehat{\mathcal{M}}(r) X_r$, which is the weighted sum of random variables under a multinomial distribution. According to Appendix, its variance is give by:

$$Var[\mathcal{B}] = \mathbb{E}[\sum_{r=1}^{n} X_r \widehat{\mathcal{M}}(r) - \sum_{r=1}^{n} \mathbb{E}[X_r]\widehat{\mathcal{M}}(r)]^2$$

$$= M \cdot P(R)\Big(\sum_r^n \widehat{\mathcal{M}}^2(r) P(r|R) - \big(\sum_r^n \widehat{\mathcal{M}}(r) P(r|R)\big)^2\Big)$$

$\mathcal{L}_2$ can be re-written (see Appendix) as:

$$\mathcal{L}_2 = \sum_{R=1}^{N} \frac{1}{M} Var(\widehat{\mathcal{M}}(r)|R)$$

Closed Form Solution and its Relationship to Bias-Variance Estimator

We can rewrite $\mathcal{L}$ as a matrix format (see Appendix) and optimize it corresponding to a constraint least square optimization and its solution:

$$\mathcal{L} = ||\sqrt{D}A\mathbf{x} - \sqrt{D}\mathbf{b}||_F^2 + \frac{1}{M}||\sqrt{\Lambda_1}\mathbf{x}||_F^2 - \frac{1}{M}||A\mathbf{x}||_F^2 \tag{10}$$

$$\mathbf{x} = \Big(A^T D A - \frac{1}{M}A^T A + \frac{1}{M}\Lambda_1\Big)^{-1} A^T D \mathbf{b} \tag{11}$$

where M is the number of users and $\mathcal{M}$ is metric function, $diagM(\cdot)$ is a diagonal matrix:

$$\mathbf{x} = \begin{bmatrix} \widehat{\mathcal{M}}(r=1) \\ \vdots \\ \widehat{\mathcal{M}}(r=n) \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} \mathcal{M}(R=1) \\ \vdots \\ \mathcal{M}(R=N) \end{bmatrix} \in \mathbb{R}^N$$

$$A_{R,r} = P(r|R) \in \mathbb{R}^{N \times n} \quad D = diagM\big(P(R)\big) \in \mathbb{R}^{N \times N}$$

$$\Lambda_1 = diagM\big(\sum_{R=1}^{N} P(r|R)\big) \in \mathbb{R}^{n \times n}$$

Relationship to the BV Estimator: The bias-variance trade-off is given by (Krichene and Rendle 2020):

$$\mathcal{L}_{BV} = \underbrace{\sum_{R=1}^{N} P(R)(\mathbb{E}[\widehat{\mathcal{M}}(r)|R] - \mathcal{M}(R))^2}_{\mathcal{L}_1}$$

$$+ \underbrace{\sum_{R=1}^{N} P(R) \cdot \gamma \cdot Var[\widehat{\mathcal{M}}(r)|R]}_{\mathcal{L}_2}$$

We observe the main difference between the BV and our new estimator is on the $\mathcal{L}_2$ components (variance components): for our estimator, each $Var[\widehat{\mathcal{M}}(r)|R]$ is regularized by $1/M$ (M is the number of testing users), where in BV, this term is regularized by $P(R)\gamma$. Our estimator reveals that as the number of users increases, the variance in the $\mathcal{L}_2$ components will continue to decrease, whereas the BV estimator does not consider this factor. Thus, as the user size increases, BV estimator still needs to deal with $\mathcal{L}_2$ or has to manually adjust γ.

Finally, both BV and the new estimator rely on prior distribution $P(R)$, which is unknown. In (Krichene and Rendle 2020), the uniform distribution is used for the estimation purpose. In this paper, we propose to leverage the latest approaches in (Jin et al. 2021b) which provide a more accurate estimation of $P(R)$ for this purpose. The experimental results in Section Experiments will confirm the validity of using such distribution estimations.

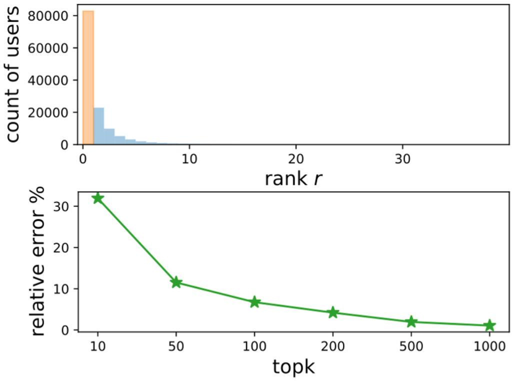

Figure 1: Top is distribution of r_u with sample set size $n = 100$. Rank $r = 1$ is highlighted. Bottom is the relative error of MLE estimator for different top-K. The result is obtained by EASE model (Steck 2019) over ml-20m dataset.

Adaptive Item-Sampling Estimator

Blind Spot and Adaptive Sampling

In recommendation, top-ranked items are vital, thus it's more crucial to obtain an accurate estimation for these top items. However current sampling approaches treat all items equally and particularly have difficulty in recovering the global top-K metrics when K is small. At top of Fig. 1, we plot the distribution of target items' rank in the sample set and observe that most target items rank top 1 (highlighted in red). This could lead to the "blind spot" problem - when K gets smaller, the estimation of basic estimators is more inaccurate (see bottom of Fig. 1). Intuitively, when $r_u = 1$, it does not mean its global rank R_u is 1, instead, its expected global rank may be around 100 (assuming $N = 10K$ and sample set size $n = 100$). And the estimation granularity is only at around 1% ($1/n$) level. This blind spot effect brings a big drawback for current estimators.

Based the on the above discussion, we propose an adaptive sampling strategy, which increases acceptable test sample size for users whose target item ranks top (say $r_u = 1$) in the sampled data. When $r_u = 1$, we continue doubling the sample size until $r_u \neq 1$ or until the sample size reaches

Algorithm 1: Adaptive Sampling Process

INPUT: Recommender Model RS, test user set $\mathcal{U}$, initial size n_0, terminal size n_{max}
OUTPUT: $\{(u, r_u, n_u)\}$

```
for all $u \in \mathcal{U}$ do
    sampling $n_0 - 1$ items, form the sample set $I_u^s$
    $n_u = n_0, r_u = RS(i_u, I_u^s)$
    while $r_u = 1$ and $n_u \neq n_{max}$ do
        sampling extra $n_u$ items, form the new set $I_u^s$
        $n_u = 2n_u, r_u = RS(i_u, I_u^s)$
    end while
    record $n_u, r_u$ for user $u$
end for
```

a predetermined ceiling. See Algorithm 1 (and detailed explanation in Appendix). The benefits of this adaptive strategy are two folds: *high granularity*, with more items sampled, the counts of $r_u = 1$ shall reduce, which could further improve the estimating accuracy; *efficiency*, we iteratively sample more items for users whose $r_u = 1$ and the empirical experiments (Table 2) confirm that small average adaptive sample size (compared to uniform sample size) is able to achieve significantly better performance.

Maximum Likelihood Estimation by EM

To utilize the adaptive item sampling for estimating the global top-K metrics, we review two routes: 1) approaches from (Krichene and Rendle 2020) and our aforementioned new estimators in this paper; 2) methods based on MLE and EM (Jin et al. 2021b). Since every user has a different number of item samples, we found the first route is hard to extend (which requires an equal sample size); but luckily the second route is much more flexible and can be easily generalized to this situation.

To begin with, we note that for any user u (his/her test item ranks r_u in the sample set (with size n_u) and ranks R_u (unknown)), its rank r_u follows a binomial distribution:

$$P(r = r_u | R = R_u; n_u) = Bin(r_u - 1; n_u - 1, \theta_u) \quad (12)$$

Given this, let $\mathbf{\Pi} = (\pi_1, \dots, \pi_R, \dots, \pi_N)^T$ be the parameters of the mixture of binomial distributions, π_R is the probability for user population ranks at position R globally. And then we have $p(r_u|\mathbf{\Pi}) = \sum_{R=1}^{N} \pi_R \cdot p(r_u|\theta_R; n_u)$, where $p(r_u|\theta_R; n_u) = Bin(r_u - 1; n_u - 1, \theta_R)$. We can apply the maximal likelihood estimation (MLE) to learn the parameters of the mixture of binomial distributions (MB), which naturally generalizes the EM procedure (details see Appendix) used in (Jin et al. 2021b), where each user has the same n samples:

$$\phi(R_{uk}) = P(R_u = k | r_u; \boldsymbol{\pi}^{old})$$
$$\pi_k^{new} = \frac{1}{M} \sum_{u=1}^{M} \phi(R_{uk})$$

When the process converges, we obtain $\mathbf{\Pi}^*$ and use it to estimate $\boldsymbol{P}$, i.e., $\widehat{P}(R) = \pi_R^*$. Then, we can use $\widehat{P}(R)$ to estimate the desired metric $metric@K$. The overall time complexity is linearly with respect to the sample size $O(t \sum n_u)$ where t is the iteration number.

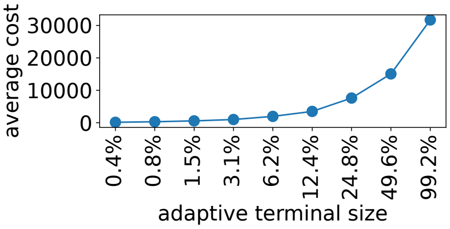

Figure 2: Sample efficiency w.r.t terminal size (ratio of total items). The illustration result is obtained by EASE model (Steck 2019) over $yelp$ dataset.

Sampling Size UpperBound

Now, we consider how to determine the terminal size n_{max}. We take the post-analysis over the different terminal sizes and investigate the average sampling cost, which introduces the concept *sampling efficiency*, see Fig. 2. Formally, we first select a large number $n_{max} \approx N$ and repeat the aforementioned adaptive sampling process. For each user, his/her sampling set size could be one of $\{n_0, n_1 = 2n_0, n_2 = 4n_0, \dots, n_t = n_{max}\}$. And there are m_j users whose sample set size is n_j $(j = 0, 1, \dots, t)$. The average sampling cost for each size n_j can be defined heuristically:

$$C_j = \frac{(M - \sum_{p=0}^{j-1} m_p) \times (n_j - n_{j-1})}{m_j} \qquad j \neq 0, t \quad (13)$$
$$C_0 = \frac{M \times n_0}{m_0}$$

The intuition behind Eq. (13) is: at j-th iteration, we independently sample $n_j - n_{j-1}$ items for total $M - \sum_{p=0}^{j-1} m_p$ users, and there are m_j users whose rank $r_u > 1$. C_j is the average items to be sampled to get a user whose $r_u > 1$, which reflects sampling efficiency. In Fig. 2, we can see that when the sample reaches 12.4% (of total items, around 3200 for $yelp$ dataset) the sampling efficiency will reduce quickly (the average cost C_j increases fast). Such post-analysis provides insights on how to balance the sample size and sampling efficiency. In this case, we observe 12.4% can be a reasonable choice. Even though different datasets can pick up different thresholds, we found in practice $10\% \sim 15\%$ can serve as a default choice to start and achieve pretty good performance for the estimation accuracy.

Experiments

In this section, we evaluate the performance of the new proposed estimator in Eq. (11) compared to the baselines from (Krichene and Rendle 2020; Jin et al. 2021b) and validate the effectiveness and efficiency of adaptive sampling-based MLE (**adaptive MLE**). Specifically, we aim to answer three questions:

dataset	Models	sample set size 100						
		baseline			this paper			
		MES	MLE	BV	BV_MES	BV_MLE	MN_MES	MN_MLE
pinterest-20	EASE	5.86±2.26	5.54±1.85	8.11±2.00	**5.05±1.46**	5.14±1.46	**5.00±1.39**	5.10±1.34
	MultiVAE	4.17±2.91	3.34±2.07	**2.75±1.61**	2.89±1.74	**2.88±1.74**	**2.75±1.66**	**2.75±1.68**
	NeuMF	5.17±2.74	4.28±1.95	4.23±1.79	3.83±1.59	3.84±1.72	**3.60±1.50**	**3.76±1.44**
	itemKNN	5.90±2.20	5.80±1.60	8.93±1.70	**5.11±1.22**	5.31±1.25	**5.09±1.15**	5.26±1.14
	ALS	4.19±2.37	3.44±1.68	3.17±1.34	3.05±1.39	3.07±1.42	**2.86±1.27**	**2.90±1.28**
yelp	EASE	8.08±4.94	7.89±4.70	18.60±2.78	6.10±3.74	6.56±3.90	**4.84±2.17**	**5.61±2.30**
	MultiVAE	9.33±6.61	7.67±4.94	9.70±3.22	6.84±4.10	6.80±4.04	**4.30±1.27**	**4.35±1.31**
	NeuMF	15.09±6.24	15.47±5.55	22.40±3.17	**13.14±4.55**	13.92±4.70	**13.46±2.43**	14.50±2.45
	itemKNN	9.25±4.87	9.62±4.88	23.24±2.16	**7.69±4.09**	8.15±4.17	**7.74±2.08**	8.75±2.08
	ALS	14.31±3.96	13.68±3.51	15.14±1.86	13.43±3.16	13.26±3.08	**11.68±0.88**	**11.57±0.83**

Table 1: The average relative errors between estimated $Recall@K$ (K from 1 to 50) and the true ones. Unit is $\%$. In each row, the smallest two results are highlighted in bold, indicating the most accurate results. Sample set size $n = 100$.

Question 1. How do the new estimators in Section New Estimator perform compared to estimators based on learning empirical distributions (i.e., MLE, MES in (Jin et al. 2021b)), and to the BV approach in (Krichene and Rendle 2020)?
Question 2. How effective and efficient the adaptive item-sampling evaluation method **adaptive MLE** is, compared with the best estimators for the basic (non-adaptive) item sampling methods in Section New Estimator?
Question 3. How accurately can these estimators find the best model (in terms of the global top-K metric) among a list of recommendation models?

Experimental Setup We take three widely-used datasets for recommendation system research, $pinterest-20$, $yelp$, $ml-20m$. See also Appendix for dataset statistics. We follow the work (Jin et al. 2021b) to adopt five popular recommendation algorithms, including three non-deep-learning methods (itemKNN (Deshpande and Karypis 2004), ALS (Hu, Koren, and Volinsky 2008), and EASE (Steck 2019)) and two deep learning ones (NeuMF (He et al. 2017) and MultiVAE (Liang et al. 2018)). Three most popular top-K metrics: $Recall$, $NDCG$, and AP are utilized for evaluating the recommendation models.

Estimating Procedure There are M users and N items. Each user u is associated with a target item i_u. The learned recommendation algorithm/model A would compute the ranks $\{R_u\}_{u=1}^M$ among all items called global ranks and the ranks $\{r_u\}_{u=1}^M$ among sampled items called sampled ranks. Without the knowledge of $\{R_u\}_{u=1}^M$, the estimator tries to estimate the global metric defined in Eq. (2) based on $\{r_u\}_{u=1}^M$. We repeat experiments 100 times, deriving 100 distinct $\{r_u\}_{u=1}^M$ set. Below reports the experimental results that are averaged on these $\mathbf{100}$ repeat.

Due to the space limitation, we report representative experimental results in the following and leave baselines, dataset statistics, and more results in the Appendix.

Q1. Accuracy of the Estimators for Basic Item-Sampling

Here, we aim to answer Question 1: comparing the new estimators in perform against the state-of-the-art methods from (Krichene and Rendle 2020; Jin et al. 2021b) under the basic item sampling scheme. Different from (Jin et al. 2021b), where authors list all the estimating as well as the true results for $metric@10$ and conclude, here we would quantify the accuracy of each estimator in terms of relative error, leading to a more rigorous and reliable comparison. Specifically, we compute the true global $metric@k$ (k from 1 to 50), then we average the absolute relative error between the estimated $metric@k$ from each estimator and the true one.

The estimators include BV (with the tradeoff parameter $\gamma = 0.01$) from (Krichene and Rendle 2020), MLE (Maximal Likelihood Estimation), MES (Maximal Entropy with Squared distribution distance, where $\eta = 0.001$) from (Jin et al. 2021b) and newly proposed estimators Eq. (11). Table 1 (see Appendix for a complete result version.) presents the average relative error of the estimators in terms of $Recall@K$ (k from 1 to 50). The results of $NDCG@K$ and $AP@K$ are in Appendix. We highlight the most and the second-most accurate estimator. For instance, for model $EASE$ in dataset $pinterest-20$ (line 1 of Table 1), the estimator MN_MES is the most accurate one with 5.00% average relative error compared to its global $Recall@K$ (K from 1 to 50).

Overall, we observe from Table 1 that MN_MES and MN_MLE are among the most or the second-most accurate estimators. And in most cases, they outperform the others significantly. Meantime, they have a smaller deviation compared to their prior estimators MES and MLE. In addition, we also notice that the estimators with the knowledge of some reasonable prior distribution (BV_MES, MN_MES, BV_MLE, MN_MLE) could achieve more accurate results than the others. This indicates that these estimators could better help the distribution to converge.

Q2. Performance of Adaptive Sampling

Here, we aim to answer Question 2: comparing the adaptive item sampling method **adaptive MLE**, comparing with the best estimators for the basic (non-adaptive) item sampling methods, such as BV_MES, MN_MES, BV_MLE, MN_MLE.

Table 2 (complete version in appendix) presents the average relative error of the estimators in terms of $NDCG@K$

Dataset	Models	Fix Sample				Adaptive Sample	
		BV_MES	BV_MLE	MN_MES	MN_MLE	average size	adaptive MLE
		sample set size n = 500					
pinterest-20	EASE	3.87±2.13	4.33±2.23	4.17±2.45	4.33±2.50	307.74±1.41	**1.46±0.63**
	MultiVAE	2.66±1.75	2.58±1.75	3.26±2.14	3.07±2.09	286.46±1.48	**1.67±0.70**
	NeuMF	2.82±1.98	2.79±1.96	3.24±2.30	3.22±2.27	259.77±1.28	**1.73±0.83**
	itemKNN	3.99±2.22	4.18±2.24	4.23±2.47	4.06±2.44	309.56±1.31	**1.42±0.67**
	ALS	3.35±1.89	3.24±1.87	3.90±2.24	3.80±2.23	270.75±1.22	**1.84±1.07**
		sample set size n = 500					
yelp	EASE	5.63±3.73	5.48±3.66	4.03±2.53	4.04±2.53	340.79±2.03	**3.55±2.00**
	MultiVAE	7.68±5.86	7.48±5.74	5.77±3.87	5.64±3.82	288.70±2.24	**5.09±2.60**
	NeuMF	9.34±5.50	9.55±5.50	8.43±4.07	8.91±4.13	290.62±2.11	**4.43±2.55**
	itemKNN	5.01±2.99	4.99±2.96	**3.65±2.28**	3.71±2.29	369.16±2.51	3.67±2.73
	ALS	13.39±7.34	13.94±7.61	12.57±5.46	13.67±5.80	297.07±2.29	**5.48±3.34**

Table 2: The average relative errors between estimated $NDCG@K$ (K from 1 to 50) and the true ones. Unit is $\%$. In each row, the smallest relative error is highlighted, indicating the most accurate result.

Dataset	Top-K	Metrics	Fix Sample				AdaptiveSample
			BV_MES	BV_MLE	MN_MES	MN_MLE	adaptive MLE
			sample set size n = 500				sample set size 260 $\sim$ 310
pinterest-20	10	RECALL	69	73	67	69	78
		NDCG	58	59	58	60	84
		AP	54	57	52	52	68
	20	RECALL	69	73	70	74	81
		NDCG	69	73	68	73	79
		AP	57	60	54	56	69

Table 3: Accuracy of predicting the winner models for different datasets. Values in the table are the number of correct predictions over 100 repeats. The larger number, the better estimator

(k from 1 to 50). We highlight the most accurate estimator. For the basic item sampling, we choose 500 sample size for datasets $pinterest-20$ and $yelp$, and 1000 sample size for dataset $ml-20m$. The upperbound threshold n_{max} is set to 3200.

We observe that adaptive sampling uses much less sample size (typically $200-300$ vs 500 on $pinterest-20$, $yelp$ datasets and $700-800$ vs 1000 on $ml-20m$ dataset). Particularly, the relative error of the adaptive sampling is significantly smaller than that of the basic sampling methods. On the first ($pinterest-20$) and third ($ml-20m$) datasets, the relative errors have reduced to less than 2%. In other words, the adaptive method has been much more effective (in terms of accuracy) and efficient (in terms of sample size). This also confirms the benefits in addressing the "blind spot" issue, which provides higher resolution to recover global K metrics for small K ($K \leq 50$ here).

Q3. Estimating the Winner

Table 3 (complete version in appendix) indicates the results of among the 100 repeats, how many times an estimator could match the best recommendation algorithms for a given $metric@K$. More concretely, for a global $metric@K$ (compute from Eq. (2)), there exists a recommendation algorithm/model which performs best called the winner. The estimator could also estimate $metric@K$ for each algorithm based on its $\{r_u\}_{u=1}^{M}$. Among this estimated $metric@K$, one can also find the best recommendation model. Intuitively, if the estimator is accurate enough, it could find the same winner as the truth. Thus, we count the success time that an estimator can find as another measure of estimating accuracy. From Table 3, We observe that newly proposed adaptive estimators could achieve competitive or even better results to the baselines with much less average sample costs.

Conclusion

In this paper, we propose first item-sampling estimators which explicitly optimize its mean square error with respect to the ground truth. Then we highlight the subtle difference between the estimators from (Krichene and Rendle 2020) and ours, and point out the potential issue of the former - failing to link the user size with the variance. Furthermore, we address the limitation of the current item sampling approach, which typically does not have sufficient granularity to recover the top K global metrics especially when K is small. We then propose an effective adaptive item-sampling method. The experimental evaluation demonstrates the new adaptive item sampling significantly improves both the sampling efficiency and estimation accuracy. Our results provide a solid step toward making item sampling available for recommendation research and practice. In the future, we would like to further investigate how to combine item-sampling working with user sampling to speed up the offline evaluation.

References

Casella, G.; and Berger, R. 2002. *Statistical Inference*. Thomson Learning.

Chen, Z.; Silvestri, F.; Wang, J.; Zhang, Y.; Huang, Z.; Ahn, H.; and Tolomei, G. 2022. GREASE: Generate Factual and Counterfactual Explanations for GNN-based Recommendations. *arXiv preprint arXiv:2208.04222*.

Cremonesi, P.; Garzotto, F.; Negro, S.; Papadopoulos, A.; and Turrin, R. 2011. Comparative evaluation of recommender system quality. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, 1927–1932.

Cremonesi, P.; Koren, Y.; and Turrin, R. 2010. Performance of Recommender Algorithms on Top-n Recommendation Tasks. In *RecSys'10*.

Dacrema, M. F.; Boglio, S.; Cremonesi, P.; and Jannach, D. 2019. A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research. arXiv:1911.07698.

Deshpande, M.; and Karypis, G. 2004. Item-Based Top-N Recommendation Algorithms. *ACM Trans. Inf. Syst.*

Ebesu, T.; Shen, B.; and Fang, Y. 2018. Collaborative Memory Network for Recommendation Systems. In *SIGIR'18*.

Fayyaz, Z.; Ebrahimian, M.; Nawara, D.; Ibrahim, A.; and Kashef, R. 2020. Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities. *applied sciences*, 10(21): 7748.

Gruson, A.; Chandar, P.; Charbuillet, C.; McInerney, J.; Hansen, S.; Tardieu, D.; and Carterette, B. 2019. Offline Evaluation to Make Decisions About PlaylistRecommendation Algorithms. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, 420–428. New York, NY, USA: Association for Computing Machinery. ISBN 9781450359405.

He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; and Chua, T.-S. 2017. Neural Collaborative Filtering. WWW '17.

Hu, B.; Shi, C.; Zhao, W. X.; and Yu, P. S. 2018. Leveraging Meta-Path Based Context for Top- N Recommendation with A Neural Co-Attention Model. In *KDD'18*.

Hu, Y.; Koren, Y.; and Volinsky, C. 2008. Collaborative filtering for implicit feedback datasets. In *ICDM'08*.

Jin, R.; Li, D.; Gao, J.; Liu, Z.; Chen, L.; and Zhou, Y. 2021a. Towards a better understanding of linear models for recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 776–785.

Jin, R.; Li, D.; Mudrak, B.; Gao, J.; and Liu, Z. 2021b. On Estimating Recommendation Evaluation Metrics under Sampling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 4147–4154.

Koren, Y. 2008. Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model. In *KDD'08*.

Krichene, W.; Mayoraz, N.; Rendle, S.; Zhang, L.; Yi, X.; Hong, L.; Chi, E. H.; and Anderson, J. R. 2019. Efficient Training on Very Large Corpora via Gramian Estimation. In *ICLR'2019*.

Krichene, W.; and Rendle, S. 2020. On Sampled Metrics for Item Recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20.

Krichene, W.; and Rendle, S. 2022. On sampled metrics for item recommendation. *Commun. ACM*, 65(7): 75–83.

Lehmann, E. L.; and Casella, G. 2006. *Theory of point estimation*. Springer Science & Business Media.

Li, D.; Jin, R.; Gao, J.; and Liu, Z. 2020. On Sampling Top-K Recommendation Evaluation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20.

Liang, D.; Krishnan, R. G.; Hoffman, M. D.; and Jebara, T. 2018. Variational Autoencoders for Collaborative Filtering. In *WWW'18*.

Peng, S.; Sugiyama, K.; and Mine, T. 2022. Less is More: Reweighting Important Spectral Graph Features for Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, 1273–1282. New York, NY, USA: Association for Computing Machinery. ISBN 9781450387323.

Rendle, S. 2019. Evaluation metrics for item recommendation under sampling. *arXiv preprint arXiv:1912.02263*.

Steck, H. 2019. Embarrassingly Shallow Autoencoders for Sparse Data. *WWW'19*.

Wang, X.; Wang, D.; Xu, C.; He, X.; Cao, Y.; and Chua, T. 2019. Explainable Reasoning over Knowledge Graphs for Recommendation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI2019*.

Yang, L.; Bagdasaryan, E.; Gruenstein, J.; Hsieh, C.-K.; and Estrin, D. 2018a. OpenRec: A Modular Framework for Extensible and Adaptable Recommendation Algorithms. In *WSDM'18*.

Yang, L.; Cui, Y.; Xuan, Y.; Wang, C.; Belongie, S. J.; and Estrin, D. 2018b. Unbiased Offline Recommender Evaluation for Missing-Not-at-Random Implicit Feedback. In *RecSys'18*.

Zhao, W. X.; Mu, S.; Hou, Y.; Lin, Z.; Chen, Y.; Pan, X.; Li, K.; Lu, Y.; Wang, H.; Tian, C.; Min, Y.; Feng, Z.; Fan, X.; Chen, X.; Wang, P.; Ji, W.; Li, Y.; Wang, X.; and Wen, J.-R. 2021. RecBole: Towards a Unified, Comprehensive and Efficient Framework for Recommendation Algorithms. In *In Proceedings of Conference on Information and Knowledge Management (CIKM '21)*.