Accurate and Diverse Recommendations via Propensity-Weighted Linear Autoencoders

Kazuma Onishi
onishi.kazuma.l5@elms.hokudai.ac.jp
Hokkaido University
Hokkaido, Japan

Katsuhiko Hayashi
katsuhiko-hayashi@g.ecc.u-tokyo.ac.jp
The University of Tokyo
Tokyo, Japan

Hidetaka Kamigaito
kamigaito.h@is.naist.jp
Nara Institute of Science and Technology
Nara, Japan

Abstract

In real-world recommender systems, user-item interactions are Missing Not At Random (MNAR), as interactions with popular items are more frequently observed than those with less popular ones. Missing observations shift recommendations toward frequently interacted items, which reduces the diversity of the recommendation list. To alleviate this problem, Inverse Propensity Scoring (IPS) is widely used and commonly models propensities based on a power-law function of item interaction frequency. However, we found that such power-law-based correction overly penalizes popular items and harms their recommendation performance. We address this issue by redefining the propensity score to allow broader item recommendation without excessively penalizing popular items. The proposed score is formulated by applying a sigmoid function to the logarithm of the item observation frequency, maintaining the simplicity of power-law scoring while allowing for more flexible adjustment. Furthermore, we incorporate the redefined propensity score into a linear autoencoder model, which tends to favor popular items, and evaluate its effectiveness. Experimental results revealed that our method substantially improves the diversity of items in the recommendation list without sacrificing recommendation accuracy. The source code of our experiments is available on GitHub at https://github.com/cars1015/IPS-LAE.

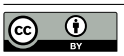
CCS Concepts

• **Information systems** → **Recommender systems**.

Keywords

Recommender Systems; Collaborative Filtering; Inverse Propensity Scoring; Diversity

ACM Reference Format:
Kazuma Onishi, Katsuhiko Hayashi, and Hidetaka Kamigaito. 2025. Accurate and Diverse Recommendations via Propensity-Weighted Linear Autoencoders. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP 2025), December 7–10, 2025, Xi'an, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3767695.3769512



SIGIR-AP 2025, Xi'an, China

ACM ISBN 979-8-4007-2218-9/2025/12
https://doi.org/10.1145/3767695.3769512

1 Introduction

In today's world, users face an overwhelming number of products such as movies and music. Recommender systems have become essential for helping users discover preferred items and supporting various services. In recent years, implicit feedback recommendation [6], which relies on feedback such as clicks, views, and purchases, has been widely adopted in practice. The data used in these systems is biased toward popular items, and user feedback also tends to concentrate on a small subset of items [1, 31]. As a result, item frequency follows a long-tail distribution.

While it is important to recommend popular items that are favored by many users, they are often already recognized by users and are less likely to offer new experiences. In contrast, recently added or less popular items that align with individual preferences can provide novelty and serendipity [5, 27]. Therefore, recommending both popular and less popular items is important, and such diversity has been shown to improve user satisfaction and service profit [11, 29]. However, achieving such diversity is challenging, as real-world recommendation data contains missing values that amplify popularity bias. In recommendation data, a zero value indicating no interactions indicates not only disinterest but also missingness due to unobserved items. The missingness mechanism is Missing Not At Random (MNAR) [19, 28], and it tends to occur more often in less popular items that are shown to fewer users than in widely displayed popular items. If recommendation models do not address the bias caused by missing data, recommendations become increasingly skewed toward popular items, concentrating user feedback on those items and promoting the lack of diversity [12, 32]. Therefore, reducing the bias caused by missing data is an important challenge in recommender systems. This issue can be addressed by collecting complete, unbiased data or by utilizing auxiliary information such as user profiles and contextual information [3, 8]. However, in real-world settings, collecting such complete data is costly, and auxiliary information is often unavailable or insufficient.

Inverse Propensity Scoring (IPS) [9, 17–19, 22, 28] is often employed as a simple yet effective approach to reduce this bias, even under such constraints. It estimates the probability that feedback on each item is correctly observed and uses its inverse as a weight to reduce the bias in the observed data. In implicit feedback recommendation, propensity scores based on a power-law function of item frequency have been widely used under the simple assumption that observation probability is proportional to item popularity. However, since the inverse of the power-law function decreases monotonically on a logarithmic scale, applying it as weights tends to excessively suppress popular items.

Promoting diversity is important, but popular items are generally of high quality or reflect current trends, and degrading their recommendation accuracy may lead to the loss of valuable recommendation opportunities [8, 30, 31]. To address this issue, we redefine the propensity scores to appropriately reduce popularity bias while avoiding excessive penalization of popular items.

Furthermore, we apply the redefined propensity scores to linear autoencoder models [13, 24, 26]. Models such as matrix factorization [6, 16] and VAE-based models [10, 20] typically represent users and items in low-dimensional spaces, which limits their expressive power and reduces item coverage in recommendations [14]. In contrast, high-dimensional linear autoencoder models [24] are capable of representing a wider range of items, yet their recommendations remain biased toward popular items, failing to fully leverage their expressive potential. Weighting based on the redefined propensity scores addresses this problem and achieves a substantial improvement in item coverage while maintaining or improving recommendation accuracy.

1.1 Notation and Preliminaries

Vectors are represented by boldface lowercase letters, e.g., $\mathbf{a}$. The i-th element of a vector $\mathbf{a}$ is represented by $\mathbf{a}_i$. $\mathbf{0}_D$ and $\mathbf{1}_D$ represent D-dimensional all-zero and all-one vectors, respectively. Matrices are represented by boldface capital letters, e.g., $\mathbf{A}$. The i-th row of a matrix $\mathbf{A}$ is represented by $\mathbf{A}_{i*}$, and the j-th column of $\mathbf{A}$ is represented by $\mathbf{A}_{*j}$. The element (i, j) of a matrix $\mathbf{A}$ is denoted by $\mathbf{A}_{ij}$. $\mathbf{A}^\top$ and $\mathbf{A}^{-1}$ denote the transpose and inverse of a matrix $\mathbf{A}$, respectively. $\mathbf{I}_D$ denotes the D-dimensional identity matrix. $\mathrm{diag}(\mathbf{A})$ is the diagonal of a square matrix $\mathbf{A}$. $\mathrm{diagMat}(\mathbf{a})$ denotes the diagonal matrix whose diagonal is the vector $\mathbf{a}$.

Let U be the set of users and I the set of items. Recommendation based on implicit feedback utilizes a binary user-item interaction matrix $\mathbf{X} \in \{0, 1\}^{|U|\times|I|}$. Here, $\mathbf{X}_{ui} = 1$ indicates that user u has interacted with item i, and $\mathbf{X}_{ui} = 0$ otherwise. Item-based collaborative filtering learns a similarity matrix $\mathbf{B} \in \mathbb{R}^{|I|\times|I|}$ that captures the similarity between items, where $\mathbf{B}_{ij}$ denotes the similarity between item i and item j. The relevance score $\mathbf{S}_{ui}$ of item i for user u is given by: $\mathbf{S}_{ui} = \mathbf{X}_{u*}\mathbf{B}_{*i}$. In Top-N recommendation, items are ranked in descending order of the relevance score, and the recommendation list is generated by selecting the top-N items.

2 Related Work

Inverse Propensity Scoring (IPS) is an effective method for mitigating biases caused by Missing Not At Random (MNAR) data, and has been applied not only in recommender systems but also in other domains [7, 15]. In implicit feedback recommendation [6], propensity scores based on a simple power-law assumption of item frequency have been commonly used [9, 17, 18, 22, 28]. Yang et al. [28] proposed unbiased evaluation metrics for implicit feedback recommendation by decomposing the observation probability of user-item interactions into the exposure and interaction probabilities, assuming that the exposure probability follows a power-law distribution. IPS has also been employed in loss functions for learning unbiased user-item relevance from biased data [9, 17, 18].

Although IPS offers a theoretically sound and practically simple approach to bias reduction, IPS-based models face challenges as their performance heavily depends on the accuracy of the propensity score and inverse propensity weighting introduces high variance. Several methods, such as joint learning [33] and clipping [18], have been proposed to address these issues, but setting appropriate propensity scores remains a major challenge [3, 8].

3 Shallow Linear Autoencoders

In this section, we introduce shallow linear autoencoder models that learn a full-rank item-item similarity matrix $\mathbf{B}$, which can be regarded as representing items in the $|I|$-dimensional vector.

3.1 EASE: Linear AutoEncoder with Diagonal Constraints

Steck [24] introduced a linear model named EASE, which provides a closed-form solution and often achieves higher performance than deep learning methods. It minimizes the L2-regularized squared error under a zero-diagonal constraint on the similarity matrix, which avoids trivial identity solutions such as $\mathbf{B}_{ii} = 1$.

$$\arg\min_{\mathbf{B}} \left\{ ||\mathbf{X} - \mathbf{X}\mathbf{B}||_F^2 + \lambda ||\mathbf{B}||_F^2 \right\} \quad \text{s.t.} \quad \mathrm{diag}\,(\mathbf{B}) = \mathbf{0}. \tag{1}$$

Here, λ is a hyperparameter that controls the degree of L2 regularization. The optimization problem can be solved using Lagrange multipliers:

$$L = ||\mathbf{X} - \mathbf{X}\mathbf{B}||_F^2 + \lambda ||\mathbf{B}||_F^2 + 2\boldsymbol{\alpha}^\top \mathrm{diag}\,(\mathbf{B}) \tag{2}$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{|I|}]$ denotes the vector of Lagrange multipliers. By setting the derivative of Eq. (2) to zero, the similarity matrix $\mathbf{B}$ is estimated as:

$$\widehat{\mathbf{B}}_{\mathrm{EASE}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_{|I|})^{-1}(\mathbf{X}^\top\mathbf{X} - \mathrm{diagMat}\,(\boldsymbol{\alpha})). \tag{3}$$

We define $\widehat{\mathbf{P}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_{|I|})^{-1}$ and imposing the constraint $\mathrm{diag}(\mathbf{B}) = 0$, the vector $\boldsymbol{\alpha}$ is determined as follows.

$$\boldsymbol{\alpha} = \mathbf{1}_{|I|} \oslash \mathrm{diag}(\widehat{\mathbf{P}}) - \lambda\mathbf{1}_{|I|}. \tag{4}$$

Here, $\oslash$ denotes element-wise division. By substituting Eq. (4) into Eq. (3), a closed-form solution can be obtained:

$$\widehat{\mathbf{B}}_{\mathrm{EASE}} = \mathbf{I}_{|I|} - \widehat{\mathbf{P}}\,\mathrm{diagMat}\,(\mathbf{1}_{|I|} \oslash \mathrm{diag}(\widehat{\mathbf{P}})). \tag{5}$$

Several variants of EASE have been proposed. EDLAE [26] introduces random dropout as regularization. MRF [25] and SANSA [21] improve scalability by using approximate factorization methods to obtain a sparse $\widehat{\mathbf{B}}$ without explicitly computing $\widehat{\mathbf{P}}$.

3.2 Limitation of EASE

Although EASE achieves high performance, it tends to overfit to popular items [13]. Applying singular value decomposition (SVD) to the user-item interaction matrix $\mathbf{X}$ yields $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and Σ is a diagonal matrix containing the top r singular values ($\sigma_1 > \cdots > \sigma_r$). Under this decomposition, the similarity matrix learned by EASE can be expressed as follows.

$$\begin{aligned}\widehat{\mathbf{B}}_{\mathrm{EASE}} = \mathbf{V}\,\mathrm{diag}\Big(\frac{\sigma_1^2}{\sigma_1^2+\lambda}, \ldots, \frac{\sigma_r^2}{\sigma_r^2+\lambda}\Big)\mathbf{V}^\top \\ - \mathbf{V}\,\mathrm{diag}\Big(\frac{1}{\sigma_1^2+\lambda}, \ldots, \frac{1}{\sigma_r^2+\lambda}\Big)\mathbf{V}^\top\,\mathrm{diagMat}(\boldsymbol{\alpha}).\end{aligned} \tag{6}$$

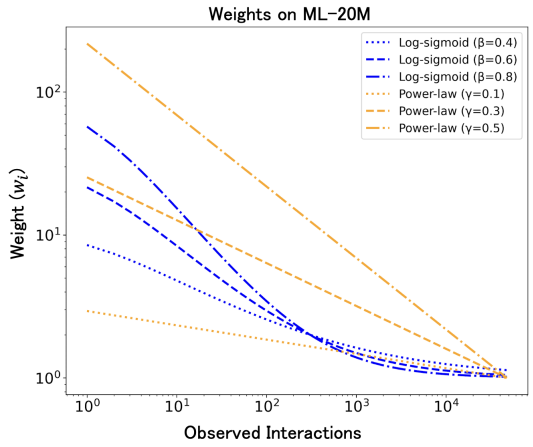


Figure 1: Comparison of the behavior of power-law-based and log-sigmoid-based weights on the ML-20M dataset. The power-law-based weights normalize the propensity score before taking its inverse.

The first term reduces the influence of principal components with small singular values through L2 regularization, and the second term derived from the zero-diagonal constraint also reduces their contribution. Principal components with large singular values tend to correspond to popular items, and those with small singular values correspond to less popular items [13]. As a result, despite its high dimensionality, EASE tends to be biased toward popular items. This tendency is also observed in EDLAE, MRF, and SANSA.

4 Proposed Method

4.1 Popularity Bias

User interactions in implicit feedback recommendation [6], such as clicks, purchases, and views, are Missing Not At Random (MNAR) [19, 28]. Popular items are more frequently exposed and interacted with, while less popular ones are rarely displayed and observed. As a result, observed interactions are biased toward popular items rather than reflecting users' true preferences [1, 31], and models trained on such data have a popularity bias. To address this, Inverse Propensity Scoring (IPS) has been proposed, assigning each item a weight equal to the inverse of its propensity score. The propensity score p_{ui} is defined as the probability that an interaction with item i is observed, given that user u is interested in it.

4.2 Estimating Propensity Score

Implicit feedback recommendation lacks auxiliary user information, and we assume that the propensity score p_{ui} is user-independent, i.e., $p_{ui} = p_i$ [22, 28]. Under this assumption, we define p_i as follows:

$$p_i = \frac{N_i^*}{N_i} \tag{7}$$

where N_i^* denotes the number of observed interactions with item i, and N_i represents the number of interactions in the ground truth data that fully reflects user preferences. However, since the ground truth data is unavailable, Eq. (7) cannot be directly computed.

4.2.1 Power-Law-Based Propensity Score. The frequency distribution of observed interactions empirically follows a power-law. Based on this observation, Steck [22] modeled p_i as a power-law function of N_i, deriving an approximation $p_i \propto (N_i^*)^\gamma$. This power-law-based propensity score has been widely used [9, 17, 18, 22, 28]. Although simple and convenient, the propensity score monotonically increases on a logarithmic scale with the number of observed interactions. The marginal utility of increasing the number of observed interactions is defined as the incremental contribution to the propensity score as follows.

$$\frac{dp_i}{dN_i^*} \propto \frac{\gamma}{(N_i^*)^{1-\gamma}}. \tag{8}$$

Eq. (8) indicates that when $\gamma < 1$, the marginal utility diminishes as N_i^* increases, while it does not diminish when $\gamma \geq 1$. Moreover, regardless of the value of γ, doubling or tripling N_i^* produces the same proportional increase in the propensity score for both popular and less popular items. In contrast, in practice the contribution of additional observations to an item's recognition depends on item popularity, and it is considered smaller for items that are already widely recognized. Therefore, the utility obtained from observed interactions should follow the law of diminishing marginal utility [2], but the power-law modeling contradicts this law.

As a result, the difference in inverse weights between popular and less popular items becomes excessively large and unfairly suppresses popular items. Furthermore, power-law-based weighting introduces large discrepancies even among popular items. Since popular items are already well recognized, differences in observed interactions likely reflect the size of the user groups who prefer them rather than differences in observation probability. Thus, excessive correction among popular items degrades recommendation accuracy. In addition, although defined as an observation probability, power-law-based propensity score may exceed 1. To address these issues, we propose a novel propensity score.

4.2.2 Proposed Propensity Score. The proposed propensity score is defined as a sigmoid function of the logarithm of observed interactions:

$$p_i = \frac{1}{1 + \exp\left(-\alpha - \beta \cdot \log(N_i^* + 1)\right)} \tag{9}$$

where β is a hyperparameter that controls the slope of the sigmoid function. A larger value makes the propensity score more sensitive to differences in observed interactions. α is an intercept that adjusts the center of the sigmoid function. Although it can be treated as a hyperparameter, for simplicity, we set it so that the propensity score is 0.5 when the log-observed interaction equals the midpoint of the minimum and maximum values:

$$\alpha = -\beta \cdot \frac{\log(\min_i N_i^* + 1) + \log(\max_i N_i^* + 1)}{2}. \tag{10}$$

We use the inverse of the defined propensity score $w_i = 1/p_i$ as the weight for each item.

The proposed propensity score takes the logarithm of the observed interactions as input, so that the utility varies with item popularity.

$$\frac{d}{dN_i^*} \log(N_i^* + 1) = \frac{1}{N_i^* + 1}. \tag{11}$$

For less popular items with small N_i^* that are not sufficiently recognized, the utility gained from observed interactions is large, whereas for popular items with large N_i^* it decreases. In this way, the law of diminishing marginal utility [2] is explicitly incorporated into the propensity score.

Moreover, by applying a sigmoid function, the propensity score is bounded within $[0, 1]$, and can be further adjusted according to item popularity as follows:

- Less popular items: These items are not recognized by many users. Since p_i is close to 0, the weight w_i becomes large, thereby promoting their recommendation.
- Moderately popular items: These items are partially recognized, but it is difficult to determine whether fewer observations than highly popular items are due to the size of the preferring user group or to insufficient exposure. By tuning the intercept α and slope β, the strength of the correction can be flexibly adjusted to fit the data.
- Highly popular items: These items are widely recognized and sufficiently exposed. Since p_i asymptotically approaches 1, w_i is prevented from becoming excessively small, and correction among highly popular items is also suppressed.

Fig. 1 compares the behavior of the proposed log-sigmoid-based weighting with power-law-based weighting. While both assign large weights to less popular items, the proposed log-sigmoid-based weighting reduces excessive weight differences between popular and less popular items and among popular items.

4.3 Incorporating Inverse Propensity Weights

The inverse propensity weights are applied to the target values in the objective function as weights, following [23].

$$\arg\min_{\mathbf{B}} \left\{ ||\mathbf{X}\,\mathrm{diag}(\mathbf{w}) - \mathbf{X}\mathbf{B}||_F^2 + \lambda ||\mathbf{B}||_F^2 \right\} \quad \text{s.t.} \quad \mathrm{diag}\,(\mathbf{B}) = \mathbf{0}. \tag{12}$$

Here, $\mathbf{w} = [w_1, \ldots, w_{|I|}]$ denotes a vector of item-wise weights. The solution to Eq. (12) has a closed-form solution [23].

$$\widehat{\mathbf{B}}_{weighted} = \widehat{\mathbf{B}}\,\mathrm{diag}(\mathbf{w}). \tag{13}$$

This solution can be applied after the similarity matrix has been learned. This enables flexible adaptation to popularity changes by simply multiplying with updated weights $\mathbf{w}$, without retraining.

5 Experiments

5.1 Experimental Setup

5.1.1 Datasets and Baseline Models. To verify the effectiveness of the proposed weighting, we conducted experiments using three datasets:

- MovieLens 20 Million (ML-20M): 136,677 users and 20,108 movies with about 10 million interactions.
- Netflix Prize (Netflix): 463,435 users and 17,769 movies with about 56.9 million interactions.
- Million Song Data (MSD): 571,355 users and 41,140 songs with about 34 million interactions.

We employed the linear autoencoder models EASE [24], EDLAE [26], RDLAE [13], and SANSA [21]. RDLAE is a variant of EDLAE that mitigates popularity bias by relaxing its diagonal constraint. For SANSA, we used the Cholesky factorization variant implemented with CHOLMOD [4] and imposed a target density $d\%$ on the Cholesky factor to control sparsity.

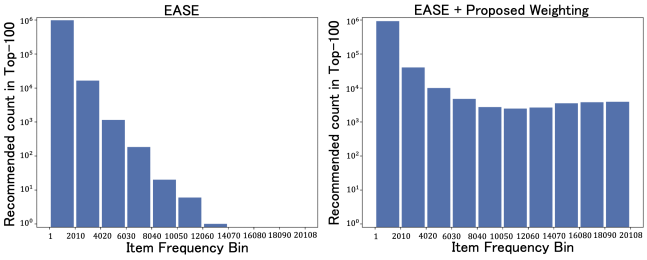


Figure 2: Distribution of recommended items before and after weighting. Items are sorted into 10 bins in descending order of their counts in the training data.

5.1.2 Metrics and Evaluation protocols. We evaluate recommendation accuracy using Recall@K and Normalized Discounted Cumulative Gain at K (NDCG@K). Recall@K measures the proportion of relevant items in the top-K recommendations, and NDCG@K takes into account the ranking positions of those relevant items. In addition, to measure the diversity of recommendations, we use Coverage@K, which indicates the proportion of unique items recommended across all users.

To ensure reproducibility, we follow the preprocessing procedure used in [10]. The evaluation was performed under a strong generalization, where the training, validation, and test sets were constructed by splitting users into disjoint sets. The hyperparameters were determined by grid search. In particular, for the inverse propensity score weighting, the coefficient β was selected from the range $[0.1, 0.2, \ldots, 0.9]$ based on its ability to improve Coverage@100 without significantly degrading NDCG@100. Our implementation and hyperparameter settings are available at https://github.com/cars1015/IPS-LAE.

5.2 Results

Tab. 1 shows the experimental results. Despite its simple implementation, the proposed weighting improves catalog coverage while maintaining or even enhancing recommendation accuracy in linear autoencoder models. In particular, it improved Coverage@100 by over 280% on ML-20M and achieved an approximately 150% improvement on Netflix Prize, without sacrificing accuracy. Moreover, even on MSD, where Coverage@100 was already high, adjusting the correction strength enabled us to further improve coverage without sacrificing recommendation accuracy. Regarding SANSA, Coverage@K showed limited improvement under high sparsity, but this limitation was alleviated by allowing a sufficient level of density.

5.3 Analysis

5.3.1 Effect of Proposed Weighting. Fig. 2 shows that the proposed log-sigmoid-based weighting introduced in Section 4.2 significantly contributes to improving the diversity of the recommendation list. While unweighted EASE concentrates recommendations on popular items, the proposed weighting promotes a broader range, covering both popular and less popular items.

Table 1: Comparison of linear autoencoder-based models with and without weighting by the proposed inverse propensity score.

Method	Unweighted				Weighted			
	Recall@20	Recall@50	NDCG@100	Coverage@100	Recall@20	Recall@50	NDCG@100	Coverage@100
ML-20M								
EASE	0.3913	0.5210	0.4203	0.2134	▲ **0.3924**	▲ **0.5239**	▲ **0.4217**	▲ **0.7133**
EDLAE	0.3925	0.5242	0.4240	0.2152	▲ **0.3954**	▲ **0.5264**	▲ **0.4253**	▲ **0.7585**
RDLAE	0.3933	0.5265	0.4252	0.2685	▲ **0.3943**	▲ **0.5270**	▲ **0.4257**	▲ **0.7423**
SANSA (d%=0.5)	0.3860	0.5144	**0.4169**	0.2131	▲ **0.3868**	▲ **0.5153**	▼ 0.4164	▲ **0.2746**
SANSA (d%=1)	0.3885	**0.5183**	0.4183	0.2194	▲ **0.3889**	▼ 0.5175	▲ **0.4189**	▲ **0.5205**
SANSA (d%=5)	0.3906	**0.5205**	0.4200	0.2140	▲ **0.3924**	▼ 0.5203	▲ **0.4212**	▲ **0.7347**
Netflix								
EASE	0.3617	0.4451	0.3934	0.4933	▲ **0.3623**	▲ **0.4454**	▲ **0.3938**	▲ **0.8491**
EDLAE	**0.3655**	0.4494	**0.3979**	0.5179	▼ 0.3650	▲ **0.4495**	▼ 0.3978	▲ **0.8391**
RDLAE	**0.3658**	**0.4496**	**0.3982**	0.5707	▼ 0.3649	▼ 0.4494	▼ 0.3977	▲ **0.8437**
SANSA (d%=0.5)	0.3545	0.4370	0.3863	0.4989	▲ **0.3553**	▲ **0.4383**	▲ **0.3865**	▲ **0.6404**
SANSA (d%=1)	0.3557	0.4389	0.3864	0.5043	▲ **0.3579**	▲ **0.4410**	▲ **0.3893**	▲ **0.7086**
SANSA (d%=5)	0.3577	0.4419	0.3899	0.5066	▲ **0.3587**	▲ **0.4422**	▲ **0.3906**	▲ **0.8270**
MSD								
EASE	0.3331	**0.4281**	0.3893	0.9773	▲ **0.3332**	**0.4281**	▲ **0.3904**	▲ **0.9832**
EDLAE	0.3335	**0.4294**	**0.3912**	0.9811	▲ **0.3337**	▼ 0.4285	▼ 0.3908	▲ **0.9859**
RDLAE	**0.3341**	**0.4292**	**0.3914**	0.9803	▼ 0.3337	▼ 0.4285	▼ 0.3908	▲ **0.9859**
SANSA (d%=0.5)	**0.3323**	**0.4269**	0.3887	0.9819	▼ 0.3322	▼ 0.4261	▲ **0.3892**	▲ **0.9859**
SANSA (d%=1)	**0.3329**	**0.4274**	0.3892	0.9828	▼ 0.3323	▼ 0.4269	▲ **0.3896**	▲ **0.9867**
SANSA (d%=5)	**0.3330**	**0.4279**	0.3892	0.9818	▼ 0.3325	▼ 0.4272	▲ **0.3898**	▲ **0.9858**

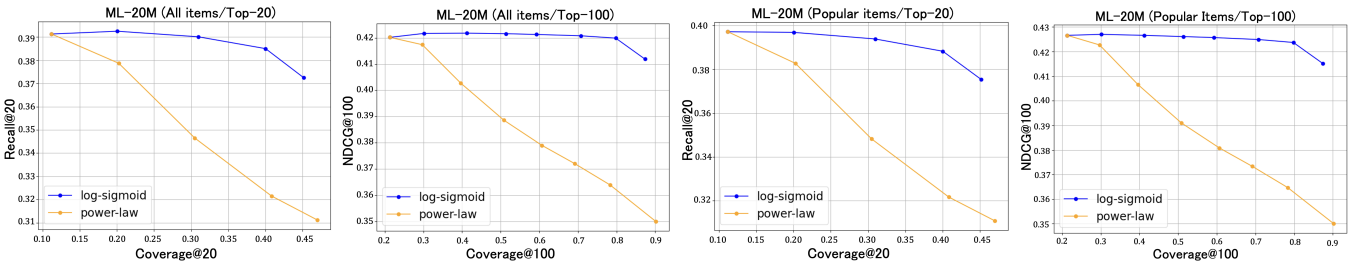


Figure 3: Comparison of performance as Coverage@K increases between power-law-based and log-sigmoid-based weighting, evaluated on all items and popular items.

5.3.2 Comparison of Power-law-based and Proposed Weighting. Fig. 3 compares how recommendation accuracy changes with Coverage@K for power-law-based and log-sigmoid-based weighting, applied to the similarity matrix learned by unweighted EASE. The comparison is conducted under two settings, where one evaluates all items and the other evaluates only the top 20% most popular items. Coverage@K is measured across all items in both settings. As shown in Fig. 3, unlike power-law-based weighting that reduces accuracy as Coverage@K improves, the proposed weighting maintains high accuracy while improving Coverage@K. However, the maximum Coverage@K achievable by the proposed weighting is limited to the range shown in Fig. 3, whereas power-law-based weighting can achieve even higher Coverage@K at the cost of significant accuracy loss.

We also evaluated Clipped IPS [18], which sets a lower bound on power-law-based propensity scores using a positive constant C as $\bar{p}_i = \max\{p_i, C\}$, to prevent the inverse propensity weights for less popular items from becoming excessively large compared to those for popular items. Tab. 2 shows the changes in recommendation performance when clipping is applied to power-law-based weights, where the exponent parameter γ is set to 0.5. As the clipping threshold C increases, Coverage@K drops significantly, reducing the diversity of the recommendation list, yet the improvement in recommendation accuracy remains limited. This suggests that while Clipped IPS reduces the excessive penalization of popular items similarly to the log-sigmoid-based weighting, it fails to suppress overcorrection among popular items, which contribute to the degradation of recommendation accuracy.

Table 2: Effect of Clipped IPS on performance. C denotes the clipping threshold.

ML-20M	Recall@20	Recall@50	NDCG@100	Coverage@100
EASE (C=0)	0.2930	0.4188	0.3137	0.9886
EASE (C=0.01)	0.2955	0.4261	0.3181	0.8102
EASE (C=0.03)	0.2978	0.4324	0.3223	0.5057
EASE (C=0.05)	0.3008	0.4367	0.3264	0.3149
EASE (C=0.1)	0.3110	0.4508	0.3384	0.3150

5.3.3 Impact of Dimensionality Reduction. Tab. 3 shows the performance after reducing the item-side dimensionality of the user-item

Table 3: Performance after Dimensionality Reduction. 'SVD' denotes Singular Value Decomposition, and $\phi_{\text{log-sigmoid}}$ indicates log-sigmoid-based weighting (Section 4.2).

ML-20M	Recall@20	Recall@50	NDCG@100	Coverage@100
SVD	0.3905	0.5195	0.4189	0.1851
SVD+$\phi_{log-sigmoid}$	0.3900	0.5199	0.4187	0.4685

interaction matrix X to 2048 dimensions via singular value decomposition. Applying the proposed weighting to the reduced EASE decreases Coverage@100 compared to the weighted EASE without dimensionality reduction but improves it over the unweighted reduced EASE. This suggests that while the proposed weighting helps improve diversity even in low-dimensional models, its effect tends to be limited compared to high-dimensional models.

6 Conclusion

We proposed a novel propensity scoring function for Inverse Propensity Scoring (IPS) in implicit recommendation, defined as a sigmoid function of the logarithm of the observed item interactions. The inverse weighting based on the proposed score retains the simplicity of traditional power-law-based weighting, while mitigating excessive correction for popular items and promoting the recommendation of less popular items. We further applied the proposed weighting to linear autoencoder models. These are state-of-the-art collaborative filtering models that are high-dimensional and expressive, but tend to overfit to popular items. Experiments showed that applying the proposed weighting to linear autoencoder models significantly improved recommendation diversity without sacrificing accuracy.

For future work, important challenges include enhancing diversity in low-dimensional recommendation methods while maintaining accuracy, and extending the proposed IPS to other domains with long-tail distributions.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP24K 02993.

References

[1] Himan Abdollahpouri and Masoud Mansoury. 2020. Multi-sided Exposure Bias in Recommendation. arXiv:2006.15772 [cs.IR] https://arxiv.org/abs/2006.15772

[2] Daniel Bernoulli. 1954. Exposition of a New Theory on the Measurement of Risk. *Econometrica* 22, 1 (1954), 23–36. http://www.jstor.org/stable/1909829

[3] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and Debias in Recommender System: A Survey and Future Directions. *ACM Trans. Inf. Syst.* 41, 3, Article 67 (Feb. 2023), 39 pages. doi:10.1145/3564284

[4] Yanqing Chen, Timothy A. Davis, William W. Hager, and Sivasankaran Rajamanickam. 2008. Algorithm 887: CHOLMOD, Supernodal Sparse Cholesky Factorization and Update/Downdate. *ACM Trans. Math. Softw.* 35, 3, Article 22 (Oct. 2008), 14 pages. doi:10.1145/1391989.1391995

[5] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the Fourth ACM Conference on Recommender Systems* (Barcelona, Spain) *(RecSys '10)*. Association for Computing Machinery, New York, NY, USA, 257–260. doi:10.1145/1864708.1864761

[6] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *2008 Eighth IEEE International Conference on Data Mining*. 263–272. doi:10.1109/ICDM.2008.22

[7] Himanshu Jain, Yashoteja Prabhu, and Manik Varma. 2016. Extreme Multi-label Loss Functions for Recommendation, Tagging, Ranking & Other Missing Label Applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 935–944. doi:10.1145/2939672.2939756

[8] Anastasiia Klimashevskaia, Dietmar Jannach, Mehdi Elahi, and Christoph Trattner. 2024. A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction* 34, 5 (July 2024), 1777–1834. doi:10.1007/s11257-024-09406-0

[9] Jae-woong Lee, Seongmin Park, and Jongwuk Lee. 2021. Dual Unbiased Recommender Learning for Implicit Feedback. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1647–1651. doi:10.1145/3404835.3463118

[10] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 689–698. doi:10.1145/3178876.3186150

[11] Siyi Liu and Yujia Zheng. 2020. Long-tail Session-based Recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 509–514. doi:10.1145/3383313.3412222

[12] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback Loop and Bias Amplification in Recommender Systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 2145–2148. doi:10.1145/3340531.3412152

[13] Jaewan Moon, Hye-young Kim, and Jongwuk Lee. 2023. It's Enough: Relaxing Diagonal Constraints in Linear Autoencoders for Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 1639–1648. doi:10.1145/3539618.3591704

[14] Naoto Ohsaka and Riku Togashi. 2023. Curse of "Low" Dimensionality in Recommender Systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 537–547. doi:10.1145/3539618.3591659

[15] Kazuma Onishi and Katsuhiko Hayashi. 2025. A Simple but Effective Closed-form Solution for Extreme Multi-label Learning. arXiv:2501.10179 [cs.IR] https://arxiv.org/abs/2501.10179

[16] Steffen Rendle, Walid Krichene, Li Zhang, and Yehuda Koren. 2022. Revisiting the Performance of iALS on Item Recommendation Benchmarks. In *Proceedings of the 16th ACM Conference on Recommender Systems* (Seattle, WA, USA) *(RecSys '22)*. Association for Computing Machinery, New York, NY, USA, 427–435. doi:10.1145/3523227.3548486

[17] Yuta Saito. 2020. Unbiased Pairwise Learning from Biased Implicit Feedback. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval* (Virtual Event, Norway) *(ICTIR '20)*. Association for Computing Machinery, New York, NY, USA, 5–12. doi:10.1145/3409256.3409812

[18] Yuta Saito, Suguru Yaginuma, Yuta Nishino, Hayato Sakata, and Kazuhide Nakata. 2020. Unbiased Recommender Learning from Missing-Not-At-Random Implicit Feedback. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (Houston, TX, USA) *(WSDM '20)*. Association for Computing Machinery, New York, NY, USA, 501–509. doi:10.1145/3336191.3371783

[19] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48* (New York, NY, USA) *(ICML'16)*. JMLR.org, 1670–1679.

[20] Ilya Shenbin, Anton Alekseev, Elena Tutubalina, Valentin Malykh, and Sergey I. Nikolenko. 2020. RecVAE: A New Variational Autoencoder for Top-N Recommendations with Implicit Feedback. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (Houston, TX, USA) *(WSDM '20)*. Association for Computing Machinery, New York, NY, USA, 528–536. doi:10.1145/3336191.3371831

[21] Martin Spišák, Radek Bartyzal, Antonín Hoskovec, Ladislav Peska, and Miroslav Tůma. 2023. Scalable Approximate NonSymmetric Autoencoder for Collaborative Filtering. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore, Singapore) *(RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 763–770. doi:10.1145/3604915.3608827

[22] Harald Steck. 2011. Item popularity and recommendation accuracy. In *Proceedings of the Fifth ACM Conference on Recommender Systems* (Chicago, Illinois, USA) *(RecSys '11)*. Association for Computing Machinery, New York, NY, USA, 125–132. doi:10.1145/2043932.2043957

[23] Harald Steck. 2019. Collaborative Filtering via High-Dimensional Regression. arXiv:1904.13033 [cs.IR] https://arxiv.org/abs/1904.13033

[24] Harald Steck. 2019. Embarrassingly Shallow Autoencoders for Sparse Data. In *The World Wide Web Conference (WWW '19)*. ACM. doi:10.1145/3308558.3313710

[25] Harald Steck. 2019. *Markov random fields for collaborative filtering.* Curran Associates Inc., Red Hook, NY, USA.

[26] Harald Steck. 2020. Autoencoders that don't overfit towards the identity. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 1644, 11 pages.

[27] Saúl Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems* (Chicago, Illinois, USA) *(RecSys '11)*. Association for Computing Machinery, New York, NY, USA, 109–116. doi:10.1145/2043932.2043955

[28] Longqi Yang, Yin Cui, Yuan Xuan, Chenyang Wang, Serge Belongie, and Deborah Estrin. 2018. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 279–287. doi:10.1145/3240323.3240355

[29] Hongzhi Yin, Bin Cui, Jing Li, Junjie Yao, and Chen Chen. 2012. Challenging the long tail recommendation. *Proc. VLDB Endow.* 5, 9 (May 2012), 896–907. doi:10.14778/2311906.2311916

[30] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal Intervention for Leveraging Popularity Bias in Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 11–20. doi:10.1145/3404835.3462875

[31] Zihao Zhao, Jiawei Chen, Sheng Zhou, Xiangnan He, Xuezhi Cao, Fuzheng Zhang, and Wei Wu. 2023. Popularity Bias is not Always Evil: Disentangling Benign and Harmful Bias for Recommendation. *IEEE Trans. on Knowl. and Data Eng.* 35, 10 (Oct. 2023), 9920–9931. doi:10.1109/TKDE.2022.3218994

[32] Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. 2021. Disentangling User Interest and Conformity for Recommendation with Causal Embedding. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 2980–2991. doi:10.1145/3442381.3449788

[33] Ziwei Zhu, Yun He, Yin Zhang, and James Caverlee. 2020. Unbiased Implicit Recommendation and Propensity Estimation via Combinational Joint Learning. In *Proceedings of the 14th ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 551–556. doi:10.1145/3383313.3412210