Mutual Wasserstein Discrepancy Minimization for Sequential Recommendation

Ziwei Fan
University of Illinois at Chicago
USA
zfan20@uic.edu

Zhiwei Liu
Salesforce AI Research
USA
zhiweiliu@salesforce.com

Hao Peng*
Beihang University
China
penghao@act.buaa.edu.cn

Philip S. Yu
University of Illinois at Chicago
USA
psyu@uic.edu

ABSTRACT

Self-supervised sequential recommendation significantly improves recommendation performance by maximizing mutual information with well-designed data augmentations. However, the mutual information estimation is based on the calculation of Kullback–Leibler divergence with several limitations, including asymmetrical estimation, the exponential need of the sample size, and training instability. Also, existing data augmentations are mostly stochastic and can potentially break sequential correlations with random modifications. These two issues motivate us to investigate an alternative robust mutual information measurement capable of modeling uncertainty and alleviating KL divergence's limitations.

To this end, we propose a novel self-supervised learning framework based on the **M**utual Wasser**Stein** discrepancy minimization (MStein) for the sequential recommendation. We propose the Wasserstein Discrepancy Measurement to measure the mutual information between augmented sequences. Wasserstein Discrepancy Measurement builds upon the 2-Wasserstein distance, which is more robust, more efficient in small batch sizes, and able to model the uncertainty of stochastic augmentation processes. We also propose a novel contrastive learning loss based on Wasserstein Discrepancy Measurement. Extensive experiments on four benchmark datasets demonstrate the effectiveness of MStein over baselines. More quantitative analyses show the robustness against perturbations and training efficiency in batch size. Finally, improvements analysis indicates better representations of popular users/items with significant uncertainty. The source code is in https://github.com/zfan20/MStein.

CCS CONCEPTS

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

*Corresponding author

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583529

KEYWORDS

Sequential Recommendation, Wasserstein Distance, Mutual Information, Self-supervised Learning

ACM Reference Format:
Ziwei Fan, Zhiwei Liu, Hao Peng, and Philip S. Yu. 2023. Mutual Wasserstein Discrepancy Minimization for Sequential Recommendation. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3543507.3583529

1 INTRODUCTION

Recommender systems have been a prevalent and crucial component in several application scenarios [24, 25, 40]. Among existing personalized recommendations, sequential recommendation (SR) attracts increasing interest for its scalability and performance. SR predicts the next preferred item for each user by modeling the sequential behaviors of the user and capturing item-item transition correlations.

Existing works of SR include Markovian approaches [16, 40], convolution-based approaches [43, 51], RNN-based approaches [18, 54], and Transformer-based methods [20, 23, 42]. The recent success of Self-Supervised Learning (SSL) further improves SR [26, 50, 56] by alleviating the data sparsity issue and improving robustness with novel data augmentations and contrastive loss, *i.e.,* InfoNCE. As the widely used SSL framework, contrastive learning (CL) constructs positive and negative pairs via data augmentation strategies. The commonly adopted CL loss InfoNCE maximizes the mutual information of positive pairs among all pairs. With data augmentations and mutual information maximization, SSL-based SR methods capture more robust user preferences.

Despite the effectiveness of SSL for SR, we argue that existing SSL for SR methods still suffer critical issues in both data augmentations and the mutual information maximization CL loss due to the following reasons:

- **Stochasticity of Data Augmentations:** Most data augmentation techniques are random augmentations, such as the random sequence crop from CL4Rec and dropout augmentation from DuoRec. Different random augmentations can be viewed as augmentation distributions, and the perturbed sequences are realized samples from augmentation distributions. However, existing CL

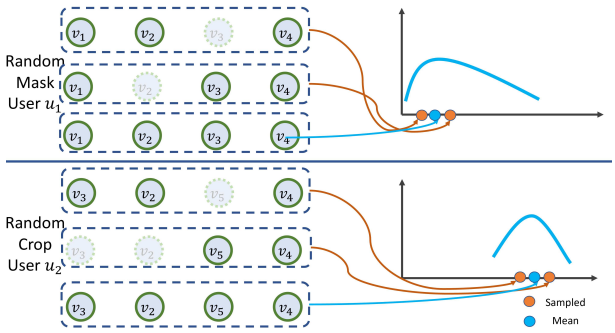


Figure 1: Motivation Examples. We only show the random mask and crop as examples but other data augmentations can also be viewed as sampling from augmentation distributions.

methods only measure the similarities between realized samples without considering the uncertainty of augmentation distributions. Ignoring uncertain information affects the stability of training and robustness of item embeddings learning against noises.

- **Limitations of KL-divergence-based Mutual Information Measurement:** Existing methods mainly adopt the InfoNCE as CL loss, which is building upon mutual information maximization. Although DuoRec [37] adopts the alignment and uniformity losses, it still follows [47] to interpret InfoNCE as the combination of alignment and uniformity. We argue that mutual information measurement has several limitations, which originate from KL-divergence, including asymmetrical estimation, the exponential need for sample size, and instability against small perturbations. All these issues might significantly affect the modeling in CL, especially when augmentations are stochastic and potentially destroy the sequential correlations.

As shown in Fig. (1), each data augmentation strategy has user-specific augmentation distributions. In Fig. (1), user u_1 has a smoother distribution than the user u_2. Different augmentation strategies introduce different levels of uncertainty and the possibility of breaking the sequential correlations. In this example, the random crop is more likely to break the user u_2's sequential correlations. Therefore, depending on users, different data augmentations follow augmentation distributions with varying uncertainties, and modeling this stochasticity becomes crucial for learning more robust embeddings. Moreover, distributions of users u_1 and u_2 have limited overlapping. In this case, the calculation of KL-divergence becomes unstable.

These issues motivate us to model the stochasticity of data augmentations and address the limitations of KL-divergence-based mutual information measurement. However, developing a framework that achieves these two goals simultaneously is nontrivial because it demands the framework to: (1) consider uncertainty information in modeling stochasticity data augmentations; (2) measure mutual information with uncertainty signals while still bypassing the limitations of KL-divergence.

In this research, we first theoretically analyze the limitations of KL-divergence in mutual information measurement and demonstrate the necessity of proposing the alternative. Then, based on the theoretical analysis, we propose the Wasserstein Discrepancy Measurement, which measures mutual information with a 2-Wasserstein distance between distributions. Furthermore, we introduce how to adapt Wasserstein Discrepancy Measurement to the contrastive learning framework by proposing the **M**utual Wasser**Stein** discrepancy minimization (MStein). Finally, we build the proposed MStein with a novel contrastive learning loss based on Wasserstein discrepancy measurement, which greatly advances both STOSA [10] for stochastic embeddings and CoSeRec [26] for data augmentation in sequential recommendation.

We summarize our four major contributions as follows:

- We propose the Wasserstein Discrepancy Measurement as the alternative to existing mutual information measurement based on KL-divergence, which has several theoretically proven limitations as the core component of InfoNCE contrastive loss.
- On top of the proposed Wasserstein Discrepancy Measurement, we propose the mutual Wasserstein discrepancy minimization as a novel contrastive learning loss and demonstrate its superiority in modeling augmentation stochasticity and robust representation learning against perturbations.
- We show that alignment and uniformity properties are exactly optimized in the proposed Wasserstein Discrepancy Measurement, and the original version of alignment and uniformity optimizations is a version of Wasserstein Discrepancy Measurement.
- Extensive experiments first demonstrate the effectiveness of MStein in generating recommendations over state-of-the-art baselines DuoRec, CL4Rec, and CoSeRec with improvements on all standard metrics from 2.24% to 20.10%. Further analysis in robustness and batch sizes show the advantages of adopting Wasserstein Discrepancy Measurement in mutual information measurement.

2 RELATED WORK

Several topics are related to this paper, including sequential and self-supervised recommendations. We first introduce relevant work in the sequential recommendation, which is the problem setting in our paper. Then we discuss some related works in the self-supervised recommendation. Lastly, we discuss existing works on uncertainty modeling and distinguish our proposed framework from them.

2.1 Sequential Recommendation

Sequential Recommendation (SR) [10, 11, 27, 35] formats each user's temporal interactions as a sequence and encodes the sequential behaviors as the user preference. The fundamental idea of SR is modeling item-item transitions within sequences and inferring user preferences from transitions. The earliest work is Markovian approaches, including most notable works FPMC [40] and Fossil [16]. Based on the Markov Chain's ability to learn orders of transitions, FPMC learns first-order item-item transitions, assuming that the next item depends on only the previous item. Fossil further combines FPMC and matrix factorization by additional item similarities information. With the perspective of viewing the sequence as an image, another line of works based on Convolution Neural Network (CNN) emerges, such as Caser [43], CosRec [51], and NextItNet [53]. Caser proposes vertical/horizontal convolution operations on the sequence embedding. CosRec interprets the

sequence as a tensor and adopts the 2d convolution. NextItNet utilizes 1D dilated convolution in the sequence embedding. Recurrent Neural Network (RNN) has shown remarkable performance in the recommendation to utilize sequential information further. The relevant works based on RNN include GRU4Rec [18], HGN [28, 36], HRNN [38], and RRN [48]. These methods adopt different RNN frameworks for SR. HGN and HRNN both propose hierarchical RNNs for SR. With the inspiration of the recent success of the self-attention mechanism, Transformer-based approaches increasingly attract interest in SR. SASRec [20] firstly adopts Transformer in SR. BERT4Rec [42] extends SASRec with bi-directional self-attention module. Multiple following works [8, 9, 23, 27, 49] build upon SASRec and further enhance SR. The advantage of self-attention for SR is its ability to model long-term dependency within the sequence.

2.2 Self-Supervised Contrastive Learning

Self-Supervised Learning (SSL) [33] advances the representation learning in multiple areas, including computer vision [4, 14, 15], NLP [13, 32], graph learning [52, 58] and recommender systems [12, 26, 37, 50, 56]. Contrastive learning (CL), as the most widely adopted approach in SSL, improves the model encoder with data augmentations and contrastive loss. Data augmentation strategies generate perturbed views of original input data. Then, contrastive learning pulls embeddings of views from the same input closer while pushing away embeddings of views from different inputs. Both data augmentations and contrastive loss are crucial components in CL. The most used contrastive loss is InfoNCE [33], which adopts the categorical cross-entropy loss and aims to maximize mutual information of two variables (in CL, two perturbed views). [34] proposes a contrastive learning loss based on Euclidean distance to generalize contrastive predictive coding.

Depending on application scenarios, several data augmentation strategies are proposed, and most are stochastic processes for perturbing the original input. In computer vision, SimCLR [4] proposes several simple stochastic data augmentations for images in the contrastive learning framework, including distortion, crop, and blurring. For recommendation, most existing works in CL for SR propose reorder, mask, and crop as augmentations [50]. S3-Rec [56] utilizes items' attributes and develops a self-supervised pre-training framework for SR. ICLRec [6] applies intent learning into the CL framework. EC4SRec [46] extends CoSeRec [26] with contrastive signals selected by explanation methods. DuoRec [37] proposes supervised contrastive signals and dropout as unsupervised signals in CL for SR. Existing works commonly propose augmentation methods specific to the application, and most augmentations are stochastic.

2.3 Uncertainty Modeling

Modeling uncertainty information has been attracting interest from the research community [2, 17, 41, 44]. The uncertainty information is categorized into aleatoric and epistemic uncertainty [19]. Aleatoric uncertainty describes the stochastic uncertainty that is unavailable to be known, while epistemic uncertainty describes the systematic uncertainty that is known but hard to measure. The most common approach to consider uncertainty information modeling is adopting the Gaussian embedding. For example, DVNE [57] represents nodes as distributions. DDN [55] and PMLAM [29] interpret users/items as distributions for recommendation. DT4SR [9] and STOSA [10] both represent items as Gaussian and develop self-attention adaptive to Gaussian embeddings for SR. Moreover, [45] proposes the AUR framework for modeling interactions' aleatoric uncertainty.

3 PRELIMINARIES

3.1 Problem Definition

In SR, we have a set of users $\mathcal{U}$ and items $\mathcal{V}$ and the associated interactions. We can sort interacted items of each user $u \in \mathcal{U}$ based on timestamps in a sequence as $\mathcal{S}^u = [v_1^u, v_2^u, \ldots, v_{|\mathcal{S}^u|}^u]$, where $v_i^u \in \mathcal{V}$ is the i-th interacted item in the sequence. For each user, SR generates a top-N recommendation list of items as the most likely preferred items in the next action. Formally, we predict the score $p\left(v_{|\mathcal{S}^u|+1}^{(u)} = v \,|\mathcal{S}^u\right)$ and rank scores of all items to generate top-N list. The core challenge in SR lies in how to model the user's action sequence $\mathcal{S}^u$.

3.2 Stochastic Transformer for SR

Transformer has been the successful backbone model for SR [20, 42] because of its self-attention module for modeling weights from all items in the sequence. However, the original Transformer architecture fails to model uncertainty information in sequences. Among existing Transformer variants, STOSA [10] extends the Transformer to introduce stochastic embeddings and a Wasserstein self-attention module for modeling uncertainty information in SR. In this research, we build on STOSA to model uncertainty information. Specifically, STOSA represents items as Gaussian distributions with mean and covariance embeddings, which together are defined as stochastic embeddings as follows:

$$\mathbf{E}^{\mu} = \mathrm{Emb}_{\mu}(\mathcal{S}^u), \mathbf{E}^{\Sigma} = \mathrm{Emb}_{\Sigma}(\mathcal{S}^u). \tag{1}$$

For the item v_i in $\mathcal{S}^u$, its stochastic embeddings proposed by STOSA includes $\mathbf{E}_{v_i}^{\mu}$ and $\mathbf{E}_{v_i}^{\Sigma}$ and parameterizes the Gaussian distribution $\mathcal{N}(\mathbf{E}_{v_i}^{\mu}, \mathrm{diag}(\mathbf{E}_{v_i}^{\Sigma}))$. To calculate the self-attention values on a specific pair of items (v_i, v_j), the Wasserstein self-attention adopts the negative 2-Wasserstence distance as follows:

$$\mathbf{A}_{ij} = -(W_2(v_i, v_j)), = -\left(||\mu_{v_i} - \mu_{v_j}||_2^2 + ||\Sigma_{v_i}^{1/2} - \Sigma_{v_j}^{1/2}||_{\mathrm{F}}^2\right), \tag{2}$$

where $W_2(\cdot,\cdot)$ denotes the 2-Wasserstence distance, $\mu_{v_i} = \mathbf{E}_{v_i}^{\mu} W_K^{\mu}$, $\Sigma_{v_i} = \mathrm{ELU}\left(\mathrm{diag}(\hat{\mathbf{E}}_{v_i}^{\Sigma} W_K^{\Sigma})\right) + 1$, $\Sigma_{v_j} = \mathrm{ELU}\left(\mathrm{diag}(\hat{\mathbf{E}}_{v_j}^{\Sigma} W_Q^{\Sigma})\right) + 1$, $\mu_{v_j} = \mathbf{E}_{v_j}^{\mu} W_Q^{\mu}$, and ELU is the Exponential Linear Unit activation function, W_K^{μ}, W_K^{Σ}, W_Q^{μ}, and W_Q^{Σ} are linear mappings for stochastic embeddings. STOSA also has *Feed-forward Neural Networks*, *Residual Connection*, and *Layer Normalization* modules, similar to original Transformer. Hence, we formulate the STOSA sequence encoder as:

$$\mathbf{h}_u = (\mathbf{h}_u^{\mu}, \mathbf{h}_u^{\Sigma}) = \mathrm{StosaEnc}(\mathcal{S}^u), \tag{3}$$

where $\mathbf{h}_u^{\mu}$ and $\mathbf{h}_u^{\Sigma}$ are stochastic sequence embeddings of $\mathcal{S}^u$, and for each timestep t, $\mathbf{h}_{u,t} = (\mathbf{h}_{u,t}^{\mu}\mathbf{h}_{u,t}^{\Sigma})$ encodes the next-item representation. The overall optimization loss is defined as follows:

$$\mathcal{L}_{\text{rec}} = \sum_{\mathcal{S}^u \in \mathcal{S}} \sum_{t=1}^{|\mathcal{S}^u|} -\log(\sigma(W_2(\mathbf{h}_{u,t}, v_t^-) - W_2(\mathbf{h}_{u,t}, v_t^+))) + \lambda \ell_{pvn}, \tag{4}$$

where v_t^+ is the ground truth next item stochastic embedding, v_t^- denotes the negative sampled item embedding, $\sigma(\cdot)$ denotes the sigmoid activation function, the stochastic embedding tables (μ, Σ) are optimized simultaneously, ℓ_{pvn} is the positive-vs-negative loss proposed by STOSA.

3.3 InfoNCE for Contrastive Learning

Contrastive loss is the core component of Contrastive Learning (CL). InfoNCE [33] is the most widely used contrastive loss. Minimizing the InfoNCE is equivalent to maximizing the lower bound of mutual information. Specifically, given a batch of N user sequences, random data augmentations generate two perturbed views of each sequence, concluding that there are $2N$ sequences, N positive pairs of sequences, and $4N^2 - 2N$ negative pairs in the InfoNCE calculation. We introduce contrastive loss based on the original Transformer encoder. For the batch of N user sequences $\mathcal{B}$, the augmentated pairs $S_{\mathcal{B}}$ are:

$$S_{\mathcal{B}} = \{\mathcal{S}_a^{u_1}, \mathcal{S}_b^{u_1}, \mathcal{S}_a^{u_2}, \mathcal{S}_b^{u_2}, \cdots, \mathcal{S}_a^{u_N}, \mathcal{S}_b^{u_N}\}, \tag{5}$$

where subscripts a and b denote two perturbed versions of $\mathcal{S}^u$. The InfoNCE for a pair of augmented sequences $(\mathcal{S}_a^{u_i}, \mathcal{S}_b^{u_i})$ is calculated as follows:

$$\mathcal{L}_{cl}(\mathbf{h}_a^{u_i}, \mathbf{h}_b^{u_i}) = -\log \frac{\exp(\text{sim}(\mathbf{h}_a^{u_i}, \mathbf{h}_b^{u_i}))}{\exp(\text{sim}(\mathbf{h}_a^{u_i}, \mathbf{h}_b^{u_i})) + \sum_{j \in S_{\mathcal{B}}^-} \exp(\text{sim}(\mathbf{h}_a^{u_i}, \mathbf{h}^j))}, \tag{6}$$

where $\mathbf{h}_a^{u_i}$ and $\mathbf{h}_b^{u_i}$ are sequence embeddings of two perturbed sequences versions learned from the encoder, $S_{\mathcal{B}}^- = S_{\mathcal{B}} - \{\mathcal{S}_a^{u_i}, \mathcal{S}_b^{u_i}\}$ denotes the negative augmented sequence pairs, and the $\text{sim}(\cdot)$ denotes the cosine similarity.

4 WASSERSTEIN DISCREPANCY MEASUREMENT

In this section, we first recall the definition of mutual information and its connection with InfoNCE in the setting of contrastive learning. Then we discuss the disadvantages of KL-divergence-based mutual information measurement. Finally, we introduce the proposed Wasserstein Discrepancy Measurement in mutual information measurement to alleviate these disadvantages.

4.1 InfoNCE and Mutual Information

InfoNCE in contrastive learning is first adopted in Contrastive Predictive Coding (CPC) [33]. The mutual information is maximized when InfoNCE is optimized. Formally, in contrastive learning, we denote the randomly augmented sequences of the user u_i as $(x_a^{u_i} = \mathcal{S}_a^{u_i}, x_b^{u_i} = \mathcal{S}_b^{u_i})$ and $(x_a^{u_i}, x_b^{u_i})$ are random variables following random augmentation distributions. The connection between InfoNCE and mutual information of $(x_a^{u_i}, x_b^{u_i})$ is given as:

$$\begin{aligned}
\mathcal{L}_{cl} &= -\mathbb{E}_{\mathcal{S}^{u_i} \in \mathcal{B}} \log \left[\frac{\frac{p(x_a^{u_i}|x_b^{u_i})}{p(x_b^{u_i})}}{\frac{p(x_a^{u_i}|x_b^{u_i})}{p(x_b^{u_i})} + \sum_{x^- \in S_{\mathcal{B}}^-} \frac{p(x_a^{u_i}|x^-)}{p(x^-)}} \right] \\
&\approx \mathbb{E}_{\mathcal{S}^{u_i} \in \mathcal{B}} \log \left[1 + \frac{p(x_b^{u_i})}{p(x_a^{u_i}|x_b^{u_i})} (2N-1) \mathbb{E}_{x^- \in S_{\mathcal{B}}^-} \frac{p(x_a^{u_i}|x^-)}{p(x^-)} \right] \\
&= \mathbb{E}_{\mathcal{S}^{u_i} \in \mathcal{B}} \log \left[1 + \frac{p(x_b^{u_i})}{p(x_a^{u_i}|x_b^{u_i})} (2N-1) \right] \\
&\geq \mathbb{E}_{\mathcal{S}^{u_i} \in \mathcal{B}} \log \left[\frac{p(x_b^{u_i})}{p(x_a^{u_i}|x_b^{u_i})} (2N-1) \right] \\
&= -\mathbb{E}_{\mathcal{S}^{u_i} \in \mathcal{B}} \log \left[\frac{p(x_a^{u_i}|x_b^{u_i})}{p(x_b^{u_i})} \right] + \log(2N-1) \\
&= \mathbb{E}_{\mathcal{S}^{u_i} \in \mathcal{B}} - D_{\text{KL}} \left(p(x_a^{u_i}), p(x_b^{u_i}) \right) + \log(2N-1) \\
&= \mathbb{E}_{\mathcal{S}^{u_i} \in \mathcal{B}} - I \left(x_a^{u_i}, x_b^{u_i} \right) + \log(2N-1),
\end{aligned} \tag{7}$$

where $D_{(KL)}$ denotes the KL-divergence, $I(x, y)$ is the mutual information between random variables x and y.

Eq. (7) proves that optimizing $\mathcal{L}_{cl}$ simultaneously maximizes mutual information as $I\left(x_a^{u_i}, x_b^{u_i}\right) \geq \log(2N-1) - \mathcal{L}_{cl}$. It also shows that when the batch size N grows larger, we can better approximate the mutual information, which has been demonstrated in related works [26, 50]. We argue that the existing InfoNCE relies on KL-divergence to measure the mutual information between variables of data augmentations. Several deficiencies from KL-divergence limit the representation learning by InfoNCE.

4.2 Limitations of KL Divergence

As mutual information estimation utilizes KL-divergence to measure the similarity of distribution, mutual information estimation shares the limitations of KL-divergence. We argue that there are three limitations to KL-divergence, including asymmetrical estimation, the exponential need for sample size, and training instability.

4.2.1 Assymetrical Estimation. As given in Eq. (7), the KL-divergence between $(x_a^{u_i}, x_b^{u_i})$ is calculated as:

$$\begin{aligned}
D_{\text{KL}} \left(p(x_a^{u_i}), p(x_b^{u_i}) \right) &= \mathbb{E}_{\mathcal{B}} \log \left[\frac{p(x_a^{u_i}|x_b^{u_i})}{p(x_b^{u_i})} \right] \\
&\neq D_{\text{KL}} \left(p(x_b^{u_i}), p(x_a^{u_i}) \right).
\end{aligned} \tag{8}$$

We can conclude that the estimation from KL-divergence is asymmetric. However, the goal of contrastive learning in InfoNCE is to maximize the similarity between augmented pairs from the same user, *i.e.,* $(x_a^{u_i}, x_b^{u_i})$, and minimize the similarity between pairs from other users, *i.e.,* $(x_a^{u_j}, x_b^{u_j}, \cdots)$. The KL-divergence requires both $D_{\text{KL}} \left(p(x_a^{u_i}), p(x_b^{u_i}) \right)$ and $D_{\text{KL}} \left(p(x_b^{u_i}), p(x_a^{u_i}) \right)$ to be small. We need to calculate more when we consider KL-divergence in negative pairs. In such cases, it requires more data and considers more pairs in InfoNCE to accurately estimate mutual information.

4.2.2 Exponential Need of Sample Size. As derived by [31, 34], the mutual information estimation based on KL-divergence has the high-confidence lower bound on N samples that cannot be larger than $O(\ln N)$. With the application in contrastive learning InfoNCE, we have a similar theorem for mutual information estimation.

Theorem 1. *Let $p(x_a)$ and $p(x_b)$ be two user sequence augmented distributions, and A denotes the set of augmented sequences with sample size N from $p(x_a)$, and B denotes the set of augmented sequences with sample size N from $p(x_b)$, respectively. Let δ be the confidence bound and let $F(A, B, \delta)$ be a real-valued function with augmented sets A, B, and the confidence parameter δ. With probability $1-\delta$, we have*

$$D_{KL}(p(x_a), p(x_b)) \geq F(A, B, \delta), \tag{9}$$

then with at least $1-4\delta$ probability that

$$\ln N \geq F(A, B, \delta). \tag{10}$$

As mutual information is measured by KL-divergence, from [31, 34], we can conclude that the mutual information bound is $N = \exp(I(x_a, x_b))$. The N in contrastive learning denotes the batch size. Therefore, the high-confidence mutual information lower bound estimation requires exponential sample sizes, which also matches with the derivation from Eq. (7).

4.2.3 Training Instability. As demonstrated in analysis in WGAN [1], KL-divergence and Jensen-Shannon divergence both encounter unstable vanishing gradients when distributions are non-overlapping. KL-divergence can be infinite when sampled data have small probabilities close to 0. As defined in Eq. (7), when $p(x_b^{u_i})$ has sampled points with probabilities $p(x_b^{u_i}) \approx 0$, the infinite KL-divergence happens. This case happens when the randomness of augmentations is large or user sequences are easily broken. Thus, it might cause training instability in mutual information estimation.

4.3 Wasserstein Discrepancy Measurement

With these three limitations of KL-divergence, it is desirable to propose an alternative to KL-divergence in mutual information estimation. In this research, we propose Wasserstein Discrepancy Measurement in mutual information estimation. Formally, we define the Wasserstein Discrepancy Measurement with the negative 2-Wasserstein distance as follows:

$$I_{W_2}\left(x_a^{u_i}, x_b^{u_i}\right) \stackrel{\text{def}}{=} -W_2(x_a^{u_i}, x_b^{u_i}) \propto \frac{p(x_a^{u_i}|x_b^{u_i})}{p(x_b^{u_i})}, \tag{11}$$

where $-W_2(x_a^{u_i}, x_b^{u_i})$ measures the negative 2-Wasserstein distribution distance between $\mathcal{N}(\mathbf{E}^{\mu}_{x_a^{u_i}}, \text{diag}(\mathbf{E}^{\Sigma}_{x_a^{u_i}}))$ and $\mathcal{N}(\mathbf{E}^{\mu}_{x_b^{u_i}}, \text{diag}(\mathbf{E}^{\Sigma}_{x_b^{u_i}}))$. 2-Wasserstein distribution distance measures information gain from the metric learning perspective. Wasserstein Discrepancy Measurement measures the negative optimal transport cost [3] between augmentation distributions, which helps stabilize the gradient calculation and alleviates the training instability limitation. Moreover, 2-Wasserstein distance is symmetric, which indicates $W_2(x_a^{u_i}, x_b^{u_i}) = W_2(x_b^{u_i}, x_a^{u_i})$, and further demonstrates less need of batch size in estimating mutual information, compared with KL-divergence.

5 MUTUAL WASSERSTEIN DISCREPANCY MINIMIZATION

Considering the stochasticity of data augmentations with stochastic modeling, we propose Wasserstein Discrepancy Measurement in the InfoNCE framework. We minimize Wasserstein discrepancy measurement $\mathcal{L}_{\text{MStein}}$ (equivalent to maximizing the mutual information $I_{W_2}\left(x_a^{u_i}, x_b^{u_i}\right)$) as follows:

$$\begin{aligned}\mathbb{E}_{\mathcal{S}^{u_i}\in\mathcal{B}} I_{W_2}\left(x_a^{u_i}, x_b^{u_i}\right) &\geq \mathbb{E}_{\mathcal{S}^{u_i}\in\mathcal{B}} - \mathcal{L}_{\text{MStein}}\left(\mathbf{h}_a^{u_i}, \mathbf{h}_b^{u_i}\right) \\ &= \mathbb{E}_{\mathcal{S}^{u_i}\in\mathcal{B}} \log \frac{\exp\left(-W_2(\mathbf{h}_a^{u_i}, \mathbf{h}_b^{u_i})\right)}{\exp\left(-W_2(\mathbf{h}_a^{u_i}, \mathbf{h}_b^{u_i})\right) + \sum_{j\in S_{\mathcal{B}}^-} \exp\left(-W_2(\mathbf{h}_a^{u_i}, \mathbf{h}^j)\right)},\end{aligned} \tag{12}$$

where $\left(\mathbf{h}_{\mu}^{u_i}, \mathbf{h}_{\Sigma}^{u_i}\right) = \text{StosaEnc}(\mathcal{S}^{u_i})$ are encoded stochastic output representations, and the 2-Wasserstein distance on encoded distribution is $-W_2(\mathbf{h}_a^{u_i}, \mathbf{h}_b^{u_i}) = -\left(||\mu_{x_a^{u_i}} - \mu_{x_b^{u_i}}||_2^2 + ||\Sigma_{x_a^{u_i}}^{1/2} - \Sigma_{x_b^{u_i}}^{1/2}||_{\text{F}}^2\right)$, which is the sum of two $L2$-errors on both mean embeddings and the square root of covariance embeddings. We measure the Wasserstein discrepancy of all augmented sequence pairs. The discrepancy is minimized for positive pairs, while the discrepancy is maximized for negative pairs. With $\mathcal{L}_{\text{MStein}}$, both stochasticities of augmentations and sequential behaviors are modeled. Moreover, adopting the 2-Wasserstein distance to measure the mutual information requires less batch size with symmetric estimation and more stable training.

5.1 Approixmating Lipschitz Continuity for Robustness

The stable training stability originates from the approximating Lipschitz continuity of STOSA and MStein. Intuitively, a model is Lipschitz continuous when a certain amount of inputs bounds its embedding output with no more than Lipschitz constant times that amount [21]. Lipschitz continuity is closely related to the robustness of the model against perturbations, which is also a necessary component in contrastive learning robustness. We utilize the demonstration from [21] that the dot-product self-attention module is not Lipschitz, but the self-attention based on the $L2$ norm is Lipschitz instead. The 2-Wasserstein distance with the diagonal covariance is the sum of two $L2$ errors on both mean embeddings and the square root of covariance embeddings. The Wasserstein self-attention proposed by STOSA approximates Lipschitz. Moreover, the approximated Lipschitz continuity of the encoder further derives the Lipschitz approximation of the proposed $\mathcal{L}_{\text{MStein}}$, which improves the robustness. We empirically demonstrate the robustness of MStein in experiment Section 6.3. In the actual implementation, we relax the requirement that $W_Q = W_K$ in the Wasserstein self-attention module to approximate Lipschitz continuity for better flexibility and better performances.

5.2 Exact Optimization of Alignment and Uniformity

We further show that mutual Wasserstein discrepancy minimization exactly optimizes two important properties, alignment and uniformity [47]. Specifically, by decomposing the $\mathcal{L}_{\text{MStein}}(\mathbf{h}_a^{u_i}, \mathbf{h}_b^{u_i})$, we

obtain the alignment component from the nominator of Eq. (12) as:

$$||\mu_{x_a^{u_i}} - \mu_{x_b^{u_i}}||_2^2 + ||\Sigma_{x_a^{u_i}}^{1/2} - \Sigma_{x_b^{u_i}}^{1/2}||_{\mathrm{F}}^2, \tag{13}$$

and the uniformity component from the denominator of Eq. (12) as:

$$\begin{aligned} &-\log \sum \exp\left(-W_2(\mathbf{h}_a^{u_i}, \mathbf{h}_b^{u_j})\right) \\ &= \log \sum \exp\left(||\mu_{x_a^{u_i}} - \mu_{x_b^{u_j}}||_2^2\right) \cdot \exp\left(||\Sigma_{x_a^{u_i}}^{1/2} - \Sigma_{x_b^{u_j}}^{1/2}||_{\mathrm{F}}^2\right) \\ &= \log \sum \exp\left(||\mu_{x_a^{u_i}} - \mu_{x_b^{u_j}}||_2^2\right) + \log \sum \exp\left(||\Sigma_{x_a^{u_i}}^{1/2} - \Sigma_{x_b^{u_j}}^{1/2}||_{\mathrm{F}}^2\right) \end{aligned} \tag{14}$$

Commonality: Both Eq. (13) and Eq. (14) have similar forms as the original alignment and uniformity as defined in [37, 47]. The Eq. (13) also adopts the Euclidean distance between embeddings (alignment on representations) and the Eq. (14) also adopts the exponential Euclidean distance on all pairs of augmented sequences (uniformity on representations).

Differences and Novelty: The differences between our proposed $\mathcal{L}_{\mathrm{MStein}}(\mathbf{h}_a^{u_i}, \mathbf{h}_b^{u_i})$ and [37, 47] from two perspectives: (1). We introduce the alignment and uniformity optimizations also on the covariance embeddings, with the advantage of pulling similar users' augmentation distributions together (*i.e.,* **distribution alignment**) and enforcing the distributions to be as distinguishable as possible (*i.e.,* **distribution uniformity**); (2). the alignment and uniformity terms proposed in [47] are **asymptotically** optimized by the contrastive loss and are not induced from the original formulation of the contrastive loss. However, our proposed $\mathcal{L}_{\mathrm{MStein}}(\mathbf{h}_a^{u_i}, \mathbf{h}_b^{u_i})$ induces and optimizes **exactly** the alignment and uniformity terms. In other words, the alignment and uniformity optimizations proposed in [47] and the Euclidean metric used in CL by [34] can be viewed as a special case of our $\mathcal{L}_{\mathrm{MStein}}(\mathbf{h}_a^{u_i}, \mathbf{h}_b^{u_i})$, which adopts Euclidean distance instead of Wasserstein distance.

5.3 Optimization and Prediction

With mutual Wasserstein discrepancy minimization $\mathcal{L}_{\mathrm{MStein}}$, we finalize the optimization loss with the recommendation loss from Eq. (4) as follows:

$$\mathcal{L} = \mathcal{L}_{rec} + \beta \mathcal{L}_{\mathrm{MStein}}, \tag{15}$$

where β is the hyper-parameter for adjusting the contribution of contrastive loss with mutual Wasserstein discrepancy minimization. The final recommendation list is generated by calculating the Wasserstein distance of the sequence encoded distribution embeddings $(\mathbf{h}_u^{\mu}, \mathbf{h}_u^{\Sigma})$ and all items' stochastic embeddings. The distances on all items are sorted in the ascending order to produce the top-N.

6 EXPERIMENTS

In this section, we demonstrate the effectiveness of the proposed MStein in multiple aspects, including performances over baselines, robustness against perturbations, and analysis of performances over different batch sizes. We answer the following research questions (RQs) in experiments:

- **RQ1:** Is MStein generating better recommendations than state-of-the-art baselines?
- **RQ2:** Is MStein more robust to noisy and limited data?
- **RQ3:** Does MStein need smaller batch sizes?
- **RQ4:** Where are improvements of MStein from?

6.1 Baselines

We compare the proposed MStein with three groups of recommendation methods. The first group includes static recommendation methods. We present BPRMF [39] due to the page limitation. The second group of methods include state-of-the-art sequential recommendation methods without self-supervised module, including Caser [43], SASRec [20], BERT4Rec [42], and STOSA [10]. The third group contains most recent sequential recommendation methods with self-supervised learning, including CL4Rec [50], DuoRec [37], and CoSeRec [26]. We also introduce a variant that builds upon CL methods with SASRec as base backbone but uses WDM as CL loss, which is CoSeRec(WDM) by converting the sequence output embeddings as $[mean\_emb; ELU(cov\_emb) + 1]$. Note that we use only **one training negative sample** for models with the Cross-Entropy loss (*e.g.,* DuoRec) because we observe that the number of negative samples significantly affects the recommendation performance [5, 7, 30].

6.2 Overall Comparisons (RQ1)

As demonstrated in the overall comparison results Table 1, we can conclude the superiority of MStein over all baselines in all metrics. We have the following observations:

- Among all models, the proposed MStein achieves the consistently best performance in all metrics over all evaluated datasets. The improvements range from 0.9% to 20.10% in all metrics, proving the effectiveness of MStein in SR. In the most challenging task top-1 recommendation, MStein obtains the most significant improvements. In the entire list ranking metric MRR, MStein achieves 2.53% to 9.90% improvements over the best baseline. We attribute these improvements to several characteristics of MStein: (1). a novel mutual information estimation based on the 2-Wasserstein distance; (2). the uncertainty modeling for stochastic data augmentation processes in self-supervised learning; (3). the robust modeling from WDM.
- Comparing the self-supervised learning SR methods (CL4Rec, DuoRec, and CoSeRec), MStein still achieves significant improvements among them. Although MStein adopts the same data augmentations as CoSeRec, the performance improvements stem from the stochastic modeling of data augmentations and more accurate and robust mutual information estimation. Furthermore, CL4Rec and CoSeRec generate better performances among these baselines as both provide manually designed data augmentations. These observations demonstrate the benefits of modeling the uncertainty of data augmentation processes and the proposed Wasserstein Discrepancy Measurement.
- In static models and SR methods, SR methods outperform the static models. This observation demonstrates the necessity of sequential information in recommendations. STOSA achieves the best performance in all SR methods, and the SASRec is the second best, showing that the self-attention module benefits SR. STOSA first introduces stochastic embeddings for modeling sequential uncertainty and demonstrates its effectiveness over other SR methods.

Table 1: Overall Performance Comparison Table. The best results are bold and the best baseline results are underlined, respectively. 'Improve.' indicates the relative improvement against the best baseline performance.

Dataset	Metric	BPRMF	Caser	SASRec	BERT4Rec	STOSA	CL4Rec	DuoRec	CoSeRec	CoSeRec(WDM)	MStein	Improv.
Beauty	Recall@1	0.0082	0.0112	0.0129	0.0119	0.0193	0.0156	0.0158	0.0188	0.0189	**0.0220**	+14.39%
	Recall@5	0.0300	0.0309	0.0416	0.0396	0.0504	0.0538	0.0505	0.0508	0.0524	**0.0551**	+2.24%
	NDCG@5	0.0189	0.0214	0.0274	0.0257	0.0351	0.0349	0.0310	0.0351	0.0359	**0.0392**	+11.69%
	Recall@10	0.0471	0.0407	0.0633	0.0595	0.0707	0.0726	0.0685	0.0738	0.0760	**0.0774**	+4.78%
	NDCG@10	0.0245	0.0246	0.0343	0.0321	0.0416	0.0412	0.0375	0.0425	0.0435	**0.0463**	+9.00%
	MRR	0.0216	0.0231	0.0291	0.0294	0.0360	0.0356	0.0325	0.0365	0.0368	**0.0398**	+9.11%
Tools	Recall@1	0.0062	0.0056	0.0103	0.0059	0.0120	0.0112	0.0108	0.0112	0.0114	**0.0144**	+20.10%
	Recall@5	0.0216	0.0129	0.0284	0.0189	0.0312	0.0314	0.0304	0.0318	**0.0344**	0.0334	+8.17%
	NDCG@5	0.0139	0.0091	0.0194	0.0123	0.0217	0.0208	0.0201	0.0216	0.0230	**0.0242**	+11.11%
	Recall@10	0.0334	0.0193	0.0427	0.0319	0.0468	0.0404	0.0401	0.0453	**0.0487**	0.0472	+4.06%
	NDCG@10	0.0177	0.0112	0.0240	0.0165	0.0267	0.0226	0.0234	0.0260	0.0276	**0.0286**	+6.90%
	MRR	0.0154	0.0106	0.0207	0.0160	0.0226	0.0212	0.0202	0.0223	0.0234	**0.0248**	+9.90%
Toys	Recall@1	0.0084	0.0089	0.0193	0.0110	0.0240	0.0220	0.0215	0.0222	0.0228	**0.0266**	+10.73%
	Recall@5	0.0301	0.0240	0.0551	0.0300	0.0577	0.0617	0.0580	0.0584	0.0616	**0.0637**	+3.17%
	NDCG@5	0.0194	0.0210	0.0377	0.0206	0.0412	0.0424	0.0401	0.0408	0.0426	**0.0457**	+7.78%
	Recall@10	0.0460	0.0262	0.0797	0.0466	0.0800	0.0764	0.0784	0.0791	**0.0852**	0.0845	+6.50%
	NDCG@10	0.0245	0.0231	0.0456	0.0260	0.0481	0.0454	0.0461	0.0474	0.0502	**0.0524**	+8.91%
	MRR	0.0216	0.0221	0.0385	0.0244	0.0415	0.0417	0.0400	0.0405	0.0425	**0.0453**	+8.67%
Office	Recall@1	0.0073	0.0069	0.0198	0.0137	0.0234	0.0230	0.0221	0.0245	0.0267	**0.0277**	+13.33%
	Recall@5	0.0214	0.0302	0.0656	0.0485	0.0677	0.0709	0.0665	0.0718	0.0703	**0.0740**	+3.13%
	NDCG@5	0.0144	0.0186	0.0428	0.0309	0.0461	0.0471	0.0456	0.0483	0.0485	**0.0512**	+5.93%
	Recall@10	0.0306	0.0550	0.0989	0.0848	0.1021	0.1091	0.1005	0.1024	0.1052	**0.1155**	+5.96%
	NDCG@10	0.0173	0.0266	0.0534	0.0426	0.0572	0.0594	0.0556	0.0598	0.0597	**0.0627**	+4.90%
	MRR	0.0162	0.0268	0.0457	0.0408	0.0502	0.0511	0.0482	0.0516	0.0519	**0.0529**	+2.53%

6.3 Robustness Analysis (RQ2)

We argue that MStein is more robust with the newly proposed mutual Wasserstein discrepancy minimization process. We validate the robustness from two perspectives, including the robustness against noisy interactions and data sizes. The comparison is conducted in MStein and CoSeRec because both adopt the same data augmentation techniques.

6.3.1 Sensitivity to Noisy Interactions. We show the sensitivity analysis of MStein against noisy interactions in Fig. (2) in all datasets. Fig. (2) shows the MRR performance over different noise ratios for the CoSeRec and the proposed MStein. We can observe that MStein is more robust to noisy interactions than CoSeRec. Specifically, for example, in the Beauty dataset analysis in Fig. (2a), when the noise ratio is 0.4 for MStein and 0.3 for CoSeRec, the performances are similar. This observation shows the robustness of MStein against noisy interaction with Wasserstein discrepancy measurement as MStein and CoSeRec adopt the same data augmentation strategies. We can also see that the performance of CoSeRec drops significantly in the Toys dataset when the noise ratio is large (0.9), while MStein still achieves satisfactory performance.

6.3.2 Sensitivity to Data Size. The sensitivity of MStein against the data size is shown in Fig. (3). In Fig. (3), we present the performance comparison between MStein and CoSeRec in varying data sizes. We can observe that MStein consistently outperforms CoSeRec in all varying data size ratios, demonstrating the superiority of MStein in SR. Moreover, MStein is more stable than CoSeRec, especially in Tools and Office datasets, as shown in Fig. (3c) and Fig. (3d) respectively. It demonstrates that MStein is more robust than CoSeRec against the dataset size, potentially due to the collaborative transitivity from stochastic embedding modeling and the newly proposed Wasserstein discrepancy measurement.

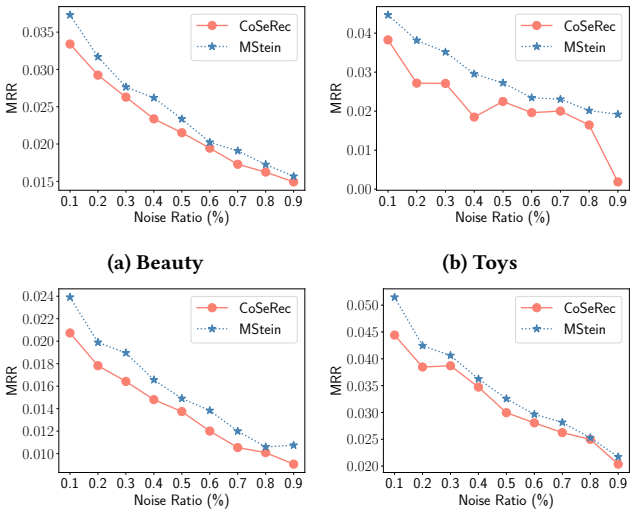
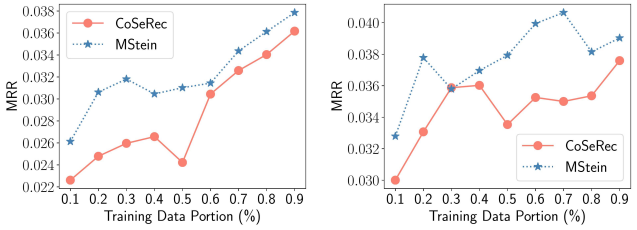
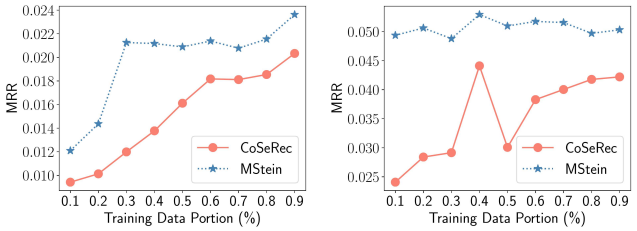


Figure 2: MRR over Different Noise Ratios.

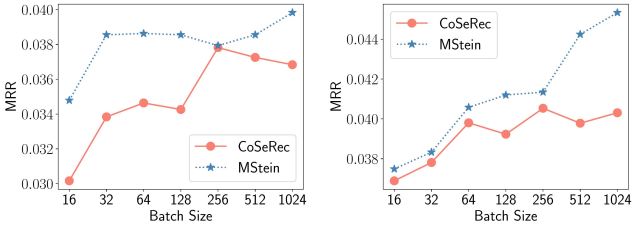


(a) Beauty (b) Toys

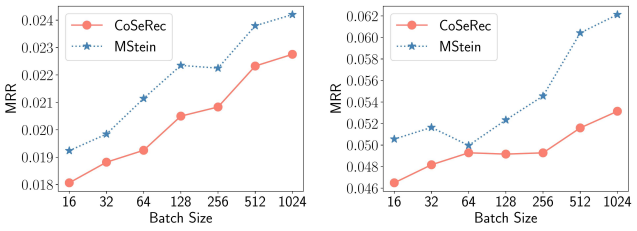


(c) Tools (d) Office

Figure 3: MRR over Different Training Data Portions.



(a) Beauty (b) Toys



(c) Tools (d) Office

Figure 4: MRR over Different Batch Sizes.

6.4 Sensitivity to Batch Size (RQ3)

As we argue that the proposed Wasserstein discrepancy measurement alleviates the exponential need for the sample size of KL divergence in mutual information estimation, we conduct the sensitivity analysis to batch sizes in Fig. (4). We compare the proposed MStein with CoSeRec in this analysis as the same set of data augmentations is applied. In contrastive learning, larger batch sizes improve the model performance significantly [33]. Fig. (4) demonstrates the beneficial effect of using larger batch sizes. Moreover, in all four datasets, MStein achieves comparative performances with much smaller batch sizes. For example, in Fig. (4a), MStein obtains MRR as 0.035 when batch size is 16 (2^4) while CoSeRec needs the batch size 128 (2^7). This observation validates the superiority of MStein over CoSeRec in need of sample size, where CoSeRec needs exponential sample sizes in InfoNCE measurement due to the inherent KL divergence limitation.

6.5 Improvements Analysis (RQ4)

We visualize the improvements of users items in Appendix Fig. (5) and items in Fig. (6) in on all datasets. We separate users and items in groups based on the number of interactions. For each group, we average NDCG@5 over the group users/items. We observe that the distributions of users/items based on the number of interactions follow the long-tail distributions shown in the bar chart. In most datasets, the performance increases as the number of interactions grow. The proposed MStein achieves better performance than SASRec, STOSA, and CoSeRec. The improvements come from the groups with the longest sequences and the second longest sequences. It verifies the strength of MStein for modeling stochastic augmentations because data augmentations for long sequences provide richer perspectives of sequences. For short sequences, data augmentations can easily break the sequential correlations. Long sequences and popular items have larger uncertainty, and stochastic augmentations provide more informative signals in contrastive learning. This observation also happens to the item perspective. MStein also achieves better performance in popular items.

7 CONCLUSIONS

We study the connection between mutual information and InfoNCE and discuss the limitations of mutual information estimation based on KL-divergence, including asymmetrical estimation, the exponential need for sample size, and the training instability. We propose an alternative choice of mutual information estimation based on Wasserstein distance, which is Wasserstein Discrepancy Measurement. With the proposed Wasserstein Discrepancy Measurement, we formulate the mutual Wasserstein discrepancy minimization in the InfoNCE framework as MStein. Extensive experiments on four benchmark datasets demonstrate the superiority of MStein using Wasserstein Discrepancy Measurement in mutual information estimation. Additional robustness analysis proves that MStein is more robust against noisy interactions and variants of data sizes.

ACKNOWLEDGMENTS

This paper was supported by the National Key R&D Program of China through grant 2022YFB3104703, NSFC through grant 62002007, Natural Science Foundation of Beijing Municipality through grant 4222030, S&T Program of Hebei through grant 21340301D, the Fundamental Research Funds for the Central Universities, and Xiaomi Young Scholar Funds for Beihang University. Philip S. Yu was supported by NSF under grants III-1763325, III-1909323, III-2106758, and SaTC-1930941. For any correspondence, please refer to Hao Peng.

REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*. PMLR, 214–223.

[2] Aleksandar Bojchevski and Stephan Günnemann. 2018. Deep Gaussian Embedding of Graphs: Unsupervised Inductive Learning via Ranking. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1ZdKJ-0W

[3] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. 2020. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*. PMLR, 1542–1553.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[5] Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. 2017. On sampling strategies for neural network-based collaborative filtering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 767–776.

[6] Yongjun Chen, Zhiwei Liu, Jia Li, Julian McAuley, and Caiming Xiong. 2022. Intent Contrastive Learning for Sequential Recommendation. In *Proceedings of the ACM Web Conference 2022*. 2172–2182.

[7] Jingtao Ding, Yuhan Quan, Xiangnan He, Yong Li, and Depeng Jin. 2019. Reinforced Negative Sampling for Recommendation with Exposure Data.. In *IJCAI*. 2230–2236.

[8] Ziwei Fan, Zhiwei Liu, Chen Wang, Peijie Huang, Hao Peng, and S Yu Philip. 2022. Sequential Recommendation with Auxiliary Item Relationships via Multi-Relational Transformer. In *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 525–534.

[9] Ziwei Fan, Zhiwei Liu, Shen Wang, Lei Zheng, and Philip S Yu. 2021. Modeling sequences as distributions with uncertainty for sequential recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3019–3023.

[10] Ziwei Fan, Zhiwei Liu, Yu Wang, Alice Wang, Zahra Nazari, Lei Zheng, Hao Peng, and Philip S Yu. 2022. Sequential recommendation via stochastic self-attention. In *Proceedings of the ACM Web Conference 2022*. 2036–2047.

[11] Ziwei Fan, Zhiwei Liu, Jiawei Zhang, Yun Xiong, Lei Zheng, and Philip S Yu. 2021. Continuous-time sequential recommendation with temporal graph collaborative transformer. In *Proceedings of the 30th ACM international conference on information & knowledge management*. 433–442.

[12] Ziwei Fan, Alice Wang, and Zahra Nazari. 2023. Episodes Discovery Recommendation with Multi-Source Augmentations. *arXiv preprint arXiv:2301.01737* (2023).

[13] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821* (2021).

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems* 33 (2020), 21271–21284.

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[16] Ruining He and Julian McAuley. 2016. Fusing similarity models with markov chains for sparse sequential recommendation. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 191–200.

[17] Shizhu He, Kang Liu, Guoliang Ji, and Jun Zhao. 2015. Learning to represent knowledge graphs with gaussian embedding. In *Proceedings of the 24th ACM international on conference on information and knowledge management*. 623–632.

[18] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).

[19] Eyke Hüllermeier and Willem Waegeman. 2019. Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction. (2019).

[20] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 197–206.

[21] Hyunjik Kim, George Papamakarios, and Andriy Mnih. 2021. The lipschitz constant of self-attention. In *International Conference on Machine Learning*. PMLR, 5562–5571.

[22] Walid Krichene and Steffen Rendle. 2020. On sampled metrics for item recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1748–1757.

[23] Jiacheng Li, Yujie Wang, and Julian McAuley. 2020. Time interval aware self-attention for sequential recommendation. In *Proceedings of the 13th international conference on web search and data mining*. 322–330.

[24] Xu Lin, Panagiotis Ilia, and Jason Polakis. 2020. Fill in the blanks: Empirical analysis of the privacy threats of browser form autofill. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*. 507–519.

[25] Xu Lin, Panagiotis Ilia, Saumya Solanki, and Jason Polakis. 2022. Phish in Sheep's Clothing: Exploring the Authentication Pitfalls of Browser Fingerprinting. In *31st USENIX Security Symposium (USENIX Security 22)*. 1651–1668.

[26] Zhiwei Liu, Yongjun Chen, Jia Li, Philip S Yu, Julian McAuley, and Caiming Xiong. 2021. Contrastive self-supervised sequential recommendation with robust augmentation. *arXiv preprint arXiv:2108.06479* (2021).

[27] Zhiwei Liu, Ziwei Fan, Yu Wang, and Philip S. Yu. 2021. *Augmenting Sequential Recommendation with Pseudo-Prior Items via Reversely Pre-Training Transformer*. Association for Computing Machinery, New York, NY, USA, 1608–1612. https://doi.org/10.1145/3404835.3463036

[28] Chen Ma, Peng Kang, and Xue Liu. 2019. Hierarchical gating networks for sequential recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 825–833.

[29] Chen Ma, Liheng Ma, Yingxue Zhang, Ruiming Tang, Xue Liu, and Mark Coates. 2020. Probabilistic metric learning with adaptive margin for top-K Recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1036–1044.

[30] Kelong Mao, Jieming Zhu, Jinpeng Wang, Quanyu Dai, Zhenhua Dong, Xi Xiao, and Xiuqiang He. 2021. SimpleX: A Simple and Strong Baseline for Collaborative Filtering. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1243–1252.

[31] David McAllester and Karl Stratos. 2020. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 875–884.

[32] Yu Meng, Chenyan Xiong, Payal Bajaj, Paul Bennett, Jiawei Han, Xia Song, et al. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems* 34 (2021), 23102–23114.

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[34] Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron Van den Oord, Sergey Levine, and Pierre Sermanet. 2019. Wasserstein dependency measure for representation learning. *Advances in Neural Information Processing Systems* 32 (2019).

[35] Bo Peng, Zhiyun Ren, Srinivasan Parthasarathy, and Xia Ning. 2020. M2: Mixed models with preferences, popularities and transitions for next-basket recommendation. *arXiv preprint arXiv:2004.01646* (2020).

[36] Bo Peng, Zhiyun Ren, Srinivasan Parthasarathy, and Xia Ning. 2021. HAM: hybrid associations models for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[37] Ruihong Qiu, Zi Huang, Hongzhi Yin, and Zijian Wang. 2022. Contrastive learning for representation degeneration problem in sequential recommendation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 813–823.

[38] Massimo Quadrana, Alexandros Karatzoglou, Balázs Hidasi, and Paolo Cremonesi. 2017. Personalizing session-based recommendations with hierarchical recurrent neural networks. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 130–137.

[39] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[40] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*. 811–820.

[41] Chi Sun, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2018. Gaussian word embedding with a wasserstein distance loss. *arXiv preprint arXiv:1808.07016* (2018).

[42] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1441–1450.

[43] Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 565–573.

[44] Luke Vilnis and Andrew McCallum. 2014. Word representations via gaussian embedding. *arXiv preprint arXiv:1412.6623* (2014).

[45] Chenxu Wang, Fuli Feng, Yang Zhang, Qifan Wang, Xunhan Hu, and Xiangnan He. 2022. Rethinking Missing Data: Aleatoric Uncertainty-Aware Recommendation. *arXiv preprint arXiv:2209.11679* (2022).

[46] Lei Wang, Ee-Peng Lim, Zhiwei Liu, and Tianxiang Zhao. 2022. Explanation guided contrastive learning for sequential recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2017–2027.

[47] Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*. PMLR, 9929–9939.

[48] Chao-Yuan Wu, Amr Ahmed, Alex Beutel, Alexander J Smola, and How Jing. 2017. Recurrent recommender networks. In *Proceedings of the tenth ACM international conference on web search and data mining*. 495–503.

[49] Liwei Wu, Shuqing Li, Cho-Jui Hsieh, and James Sharpnack. 2020. SSE-PT: Sequential recommendation via personalized transformer. In *Fourteenth ACM Conference on Recommender Systems*. 328–337.
[50] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive learning for sequential recommendation. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE, 1259–1273.
[51] An Yan, Shuo Cheng, Wang-Cheng Kang, Mengting Wan, and Julian McAuley. 2019. CosRec: 2D convolutional neural networks for sequential recommendation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2173–2176.
[52] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems* 33 (2020), 5812–5823.
[53] Fajie Yuan, Alexandros Karatzoglou, Ioannis Arapakis, Joemon M Jose, and Xiangnan He. 2019. A simple convolutional generative network for next item recommendation. In *Proceedings of the twelfth ACM international conference on web search and data mining*. 582–590.
[54] Lei Zheng, Ziwei Fan, Chun-Ta Lu, Jiawei Zhang, and Philip S Yu. 2019. Gated Spectral Units: Modeling Co-evolving Patterns for Sequential Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1077–1080.
[55] Lei Zheng, Chaozhuo Li, Chun-Ta Lu, Jiawei Zhang, and Philip S Yu. 2019. Deep Distribution Network: Addressing the Data Sparsity Issue for Top-N Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1081–1084.
[56] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1893–1902.
[57] Dingyuan Zhu, Peng Cui, Daixin Wang, and Wenwu Zhu. 2018. Deep variational network embedding in wasserstein space. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2827–2836.
[58] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. 2021. Graph contrastive learning with adaptive augmentation. In *Proceedings of the Web Conference 2021*. 2069–2080.

A DATA STATISTICS

We present the detailed datasets statistics in Table 2. We evaluate all models in four benchmark datasets from the public Amazon review dataset[1]. In the Amazon reviews dataset, there are multiple categories of product interactions with timestamps from users. We choose *Beauty*, *Toys and Games* (Toys), *Tools and Home* (Tools), and *Office Products* (Office) categories in our experiments as these four categories are widely used benchmark datasets [9, 10, 20, 23, 42, 43]. We treat the presence of user-item reviews as user-item interactions. For each user, we sort the interacted items based on the timestamp to form the interaction sequence. In each user sequence, we use the last interaction for testing and the second to last one for validation. We adopt the standard 5-core pre-processing step on users [9, 10, 20, 23, 42, 43] to filter out users with less than five interactions. We present detailed datasets statistics in Table 2.

Table 2: Datasets Statistics.

Dataset	#users	#items	#interactions	density	avg. interactions per user
Beauty	22,363	12,101	198,502	0.05%	8.3
Toys	19,412	11,924	167,597	0.07%	8.6
Tools	16,638	10,217	134,476	0.08%	8.1
Office	4,905	2,420	53,258	0.44%	10.8

[1]http://deepyeti.ucsd.edu/jianmo/amazon/index.html

B EVALUATION

We generate the top-N recommendation list for each user based on the sequence-item Wasserstein distance in ascending order. We **rank all items** for all models so that no sampling bias is introduced in evaluation [22]. The evaluation includes standard top-N ranking metrics, Recall@N, NDCG@N, and MRR. We report the average results over all test users. The test results are reported based on the best validation results. We report metrics in multiple Ns, including $N = \{1, 5, 10\}$, which are widely adopted in [10, 20, 42].

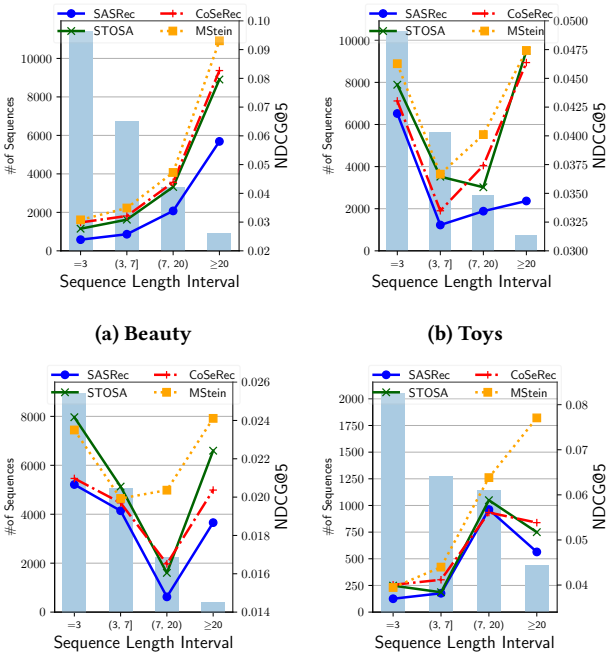
C HYPER-PARAMETERS GRID SEARCH

We implement MStein with Pytorch. We grid search all parameters and report the test performance based on the best validation results. For all baselines, we search the embedding dimension in $\{64, 128\}$. As the proposed model has both mean and covariance embeddings, we only search for $\{32, 64\}$ for MStein for the fair comparison. We also search max sequence length from $\{50, 100\}$. We tune the learning rate in $\{10^{-3}, 10^{-4}\}$, search the $L2$ regularization weight from $\{10^{-1}, 10^{-2}, 10^{-3}\}$, dropout rate from $\{0.3, 0.5, 0.7\}$. For sequential methods, we search number of layers from $\{1, 2, 3\}$, and number of heads in $\{1, 2, 4\}$. We adopt the early stopping strategy that model optimization stops when the validation MRR does not increase for 50 epochs. The followings are the model specific hyper-parameters search ranges of baselines: The third group consists of sequential recommendation methods:

- **BPR**[2]: BPR is the most classical collaborative filtering method for personalized ranking with implicit feedbacks. We search the learning rate in $\{10^{-3}, 10^{-4}\}$ and $L2$ regularization weight from $\{10^{-1}, 10^{-2}, 10^{-3}\}$.
- **Caser**[3]: A CNN-based sequential recommendation method that views the sequence embedding matrix as an image and applies convolution operators to it. We search the length L from $\{5, 10\}$, and T from $\{1, 3, 5\}$.
- **SASRec**[4]: The state-of-the-art sequential method that depends on the Transformer architecture. We search the dropout rate from $\{0.3, 0.5, 0.7\}$.
- **BERT4Rec**[5]: This method extends SASRec to model bidirectional item transitions with standard Cloze objective. We search the mask probability from the range of $\{0.1, 0.2, 0.3, 0.5, 0.7\}$.
- **STOSA**[6]: A metric learning-base sequential method that models items as distributions and proposes a Wasserstein self-attention module. We search the dropout rate from $\{0.3, 0.5, 0.7\}$.
- **CL4Rec:**[7] A sequential recommendation method that introduces masking, reorder, and cropping data augmentations in the contrastive learning framework. We search the masking rate from $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ and cropping ratio from $\{0.1, 0.2, 0.3, 0.4, 0.5\}$.
- **DuoRec:**[8] This method introduces unsupervised Dropout and supervised semantic augmentations in self-supervised learning for sequential recommendation.

[2]https://github.com/xiangwang1223/neural_graph_collaborative_filtering
[3]https://github.com/graytowne/caser_pytorch
[4]https://github.com/RUCAIBox/CIKM2020-S3Rec
[5]https://github.com/FeiSun/BERT4Rec
[6]https://github.com/zfan20/STOSA
[7]https://github.com/YChen1993/CoSeRec
[8]https://github.com/RuihongQiu/DuoRec

- **CoSeRec:**[9] This method extends CL4Rec with additional data augmentation techniques.

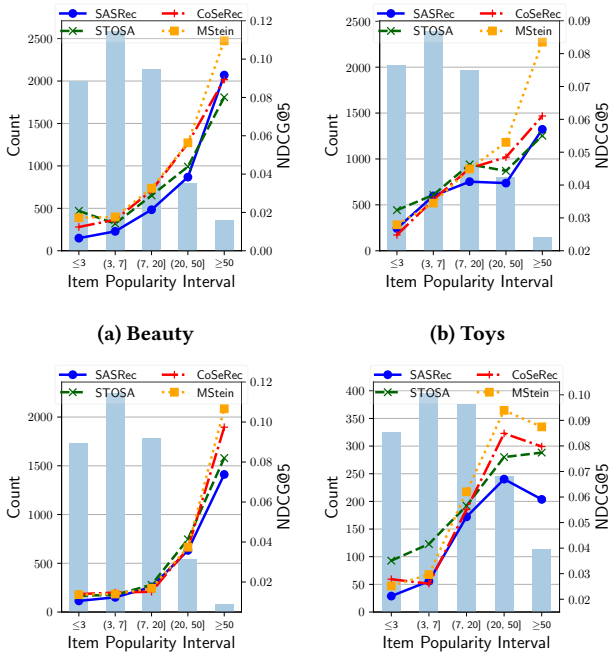


(a) Beauty (b) Toys

(c) Tools (d) Office

Figure 5: NDCG@5 on different sequences based on length.

[9]https://github.com/YChen1993/CoSeRec



(a) Beauty (b) Toys

(c) Tools (d) Office

Figure 6: NDCG@5 on different items based on popularity.

D USERS AND ITEMS IMPROVEMENTS ANALYSIS ON ALL DATASETS

Detailed analysis and observations can be found in Section 6.5.