Mutual Wasserstein Discrepancy Minimization for Sequential Recommendation

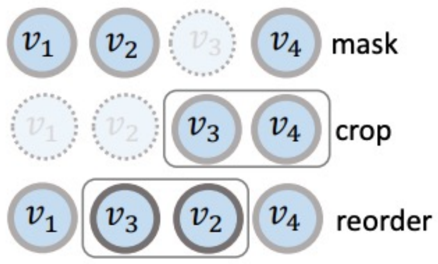
Ziwei Fan∗, Zhiwei Liu†, Hao Peng§, Philip S. Yu∗

∗Department of Computer Science, University of Illinois Chicago, Chicago, USA
†Salesforce AI Research, Palo Alto, USA
§School of Cyber Science and Technology, Beihang University, Beijing, China

https://github.com/zfan20/MStein

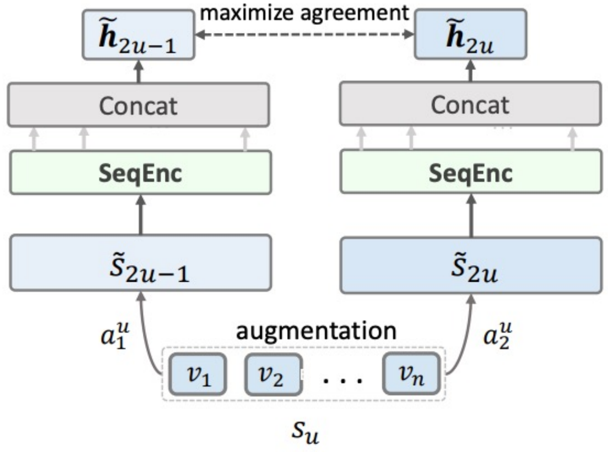
Background

- Sequential Recommendation (SR) models the dynamic user behaviors.
- Self-supervised Sequential Recommendation.



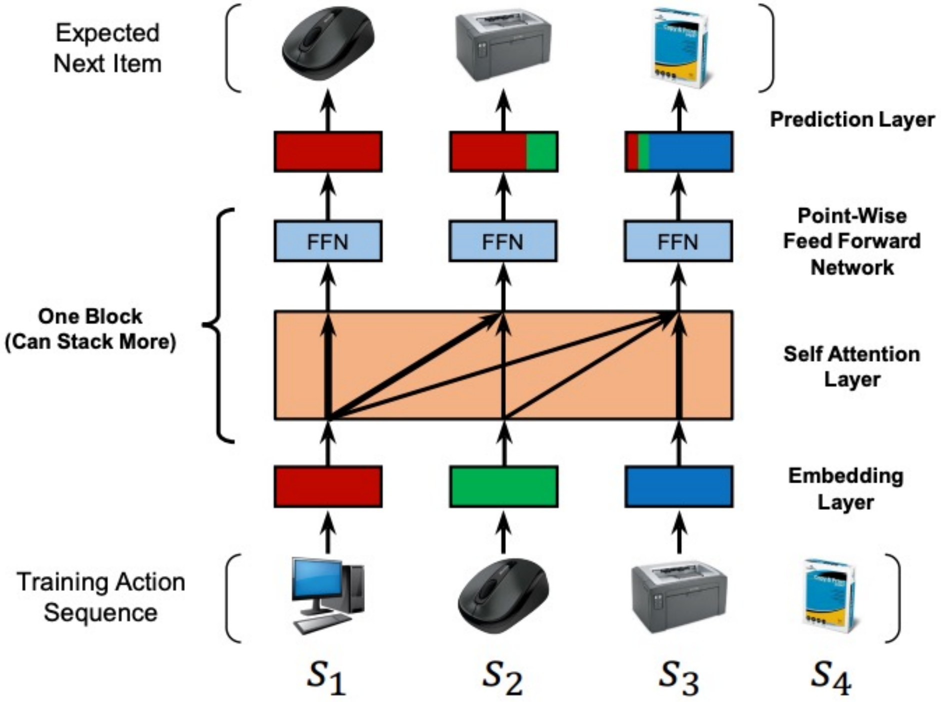
(a) Random

Sequence Augmentation



(c) Contrastive Self-supervised Learning

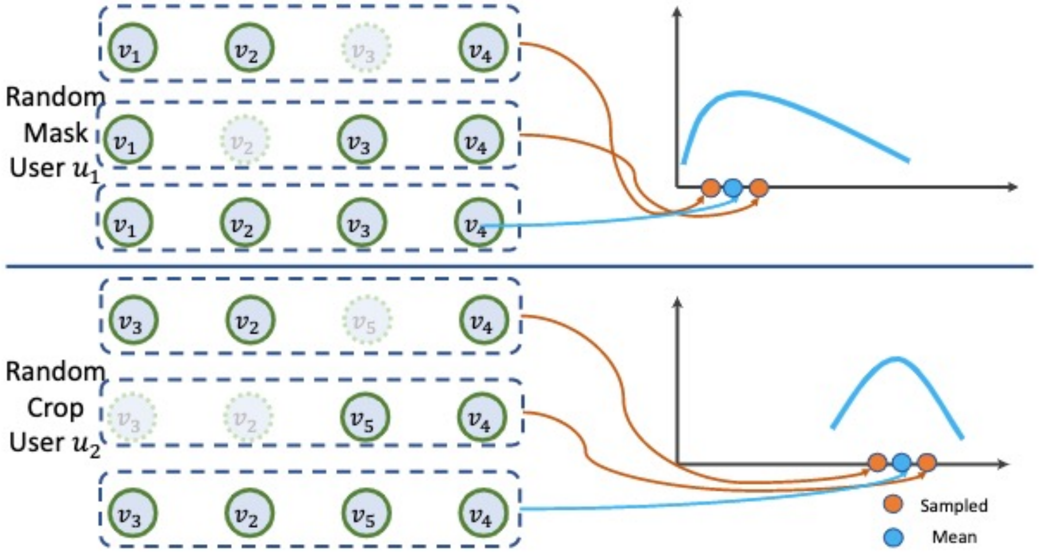
CoSeRec and CL4Rec, 2021

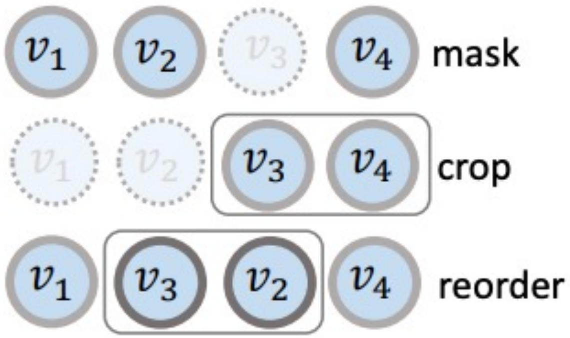


SASRec, ICDM 2018

Motivation 1

- Data Augmentations are stochastic
  - Sampling from augmentation distributions with uncertainty.





(a) Random

Sequence Augmentation

Motivation 2

- Existing methods mainly adopt the InfoNCE as contrastive learning loss
  - Based on KL-divergence to measure information gain.
- Several limitations of KL-divergence
  - Asymmetrical estimation.
  - Exponential need of samples.
  - Training instability.

InfoNCE is based on KL-divergence

- Augmented samples. $(x_a^{u_i}, x_b^{u_i})$
- When the batch size N grows larger, we can better approximate
  the mutual information.

$$\mathcal{L}_{cl} \geq \mathbb{E}_{\mathcal{S}^{u_i} \in \mathcal{B}} - D_{\mathrm{KL}}\left(p(x_a^{u_i}), p(x_b^{u_i})\right) + \log(2N-1)$$

$$\mathbb{E}_{\mathcal{S}^{u_i} \in \mathcal{B}} - I\left(x_a^{u_i}, x_b^{u_i}\right) + \log(2N-1),$$

$$I\left(x_a^{u_i}, x_b^{u_i}\right) \geq \log(2N-1) - \mathcal{L}_{cl}$$

Limitations of KL-divergence

- Asymetrical Estimation $D_{\mathrm{KL}}\left(p(x_a^{u_i}), p(x_b^{u_i})\right)$ and $D_{\mathrm{KL}}\left(p(x_b^{u_i}), p(x_a^{u_i})\right)$
  - To measure similarity for one pair, we need two distances.
- Exponential Need of Sample Size [1,2]
- Training Instability
  - Infinite KL-divergence When $p(x_b^{u_i}) \approx 0,$
    - The randomness of augmentations is large.
    - User sequences are easily broken.

[1]. David McAllester and Karl Stratos. 2020. Formal limitations on the measurement of mutual information. In International Conference on Artificial Intelligence and Statistics.
[2]. Sherjil Ozair, et, al. 2019. Wasserstein dependency measure for representation learning. Advances in Neural Information Processing Systems 32 (2019).

Propose an alternative loss

Uncertainty consideration.

More robust.

More efficient in the need of number of samples.

Wasserstein Discrepancy Measurement

- Negative 2-Wasserstein Distance as the alternative

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

$$\mathbb{E}_{\mathcal{S}^{u_i}\in\mathcal{B}} \log \frac{\exp\left(-W_2(\mathbf{h}_a^{u_i}, \mathbf{h}_b^{u_i})\right)}{\exp\left(-W_2(\mathbf{h}_a^{u_i}, \mathbf{h}_b^{u_i})\right) + \sum_{j\in S_{\mathcal{B}}^-} \exp\left(-W_2(\mathbf{h}_a^{u_i}, \mathbf{h}^j)\right)},$$

$$-W_2(\mathbf{h}_a^{u_i}, \mathbf{h}_b^{u_i}) = -\left(||\mu_{x_a^{u_i}} - \mu_{x_b^{u_i}}||_2^2 + ||\Sigma_{x_a^{u_i}}^{1/2} - \Sigma_{x_b^{u_i}}^{1/2}||_{\mathrm{F}}^2\right),$$

Exact Optimization of Alignment and Uniformity

- Distribution Alignment

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

- Distribution Uniformity

$$\log \sum \exp\left(||\mu_{x_a^{u_i}} - \mu_{x_b^{u_j}}||_2^2\right) + \log \sum \exp\left(||\Sigma_{x_a^{u_i}}^{1/2} - \Sigma_{x_b^{u_j}}^{1/2}||_\mathrm{F}^2\right)$$

Experiments

- Dataset
  - Amazon Reviews
- Last item as testing, second last item as validation
- User as a sequence (sorted by time)
- All items ranking

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

Overall Comparisons

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%

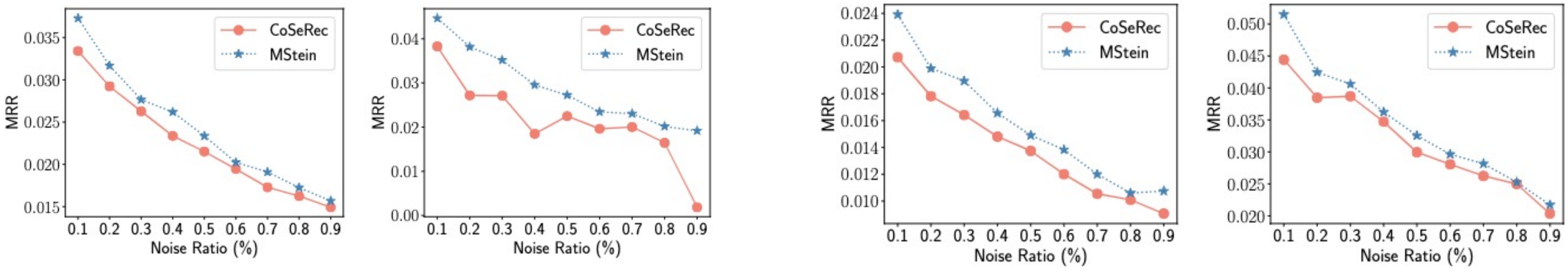
- CoSeRec(WDM) has the output embedding: $[mean\_emb; ELU(cov\_emb) + 1]$.
- Better than CoSeRec with SASRec as backbone.

Overall Comparisons

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%

- MStein uses STOSA as backbone, is the best.
- It shows that Wasserstein Discrepancy Measurement is effective.

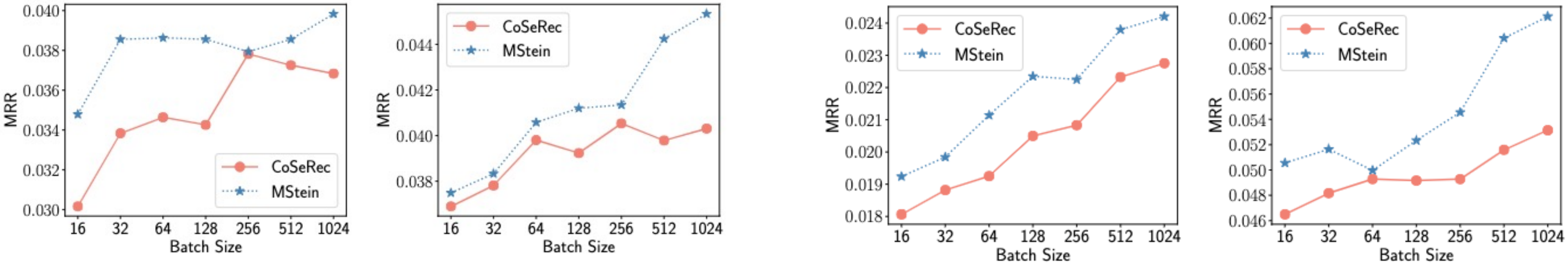
Robustness against Noise Interactions



(a) Beauty (b) Toys (c) Tools (d) Office

- In the Beauty dataset, CoSeRec (0.3 noise ratio)has similar MRR with MStein (0.4 noise ratio) -> MStein is more robust.

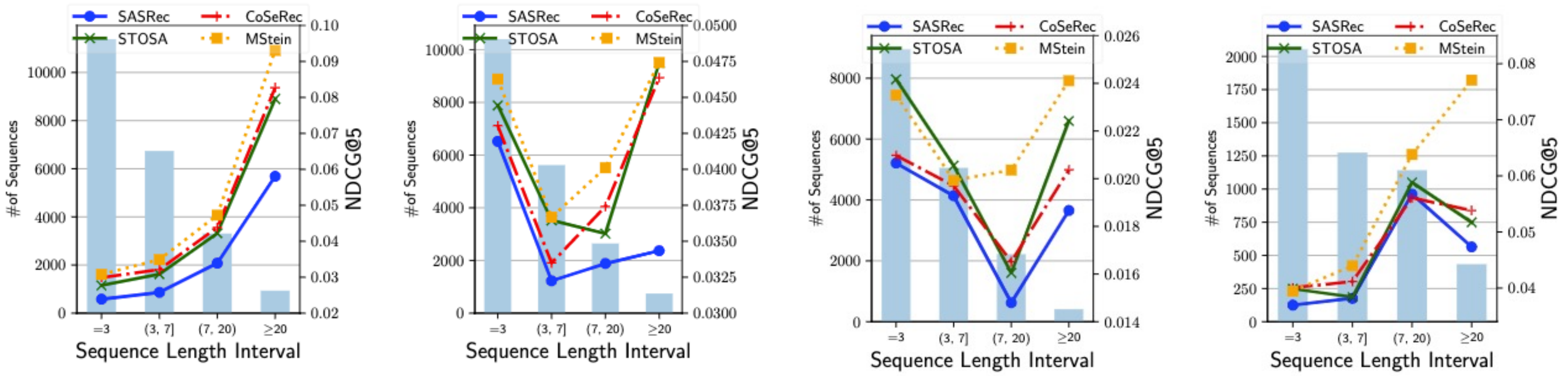
Batch Size Efficiency



(a) Beauty **(b) Toys** **(c) Tools** **(d) Office**

- To achieve similar performances, MStein needs smaller batch sizes.

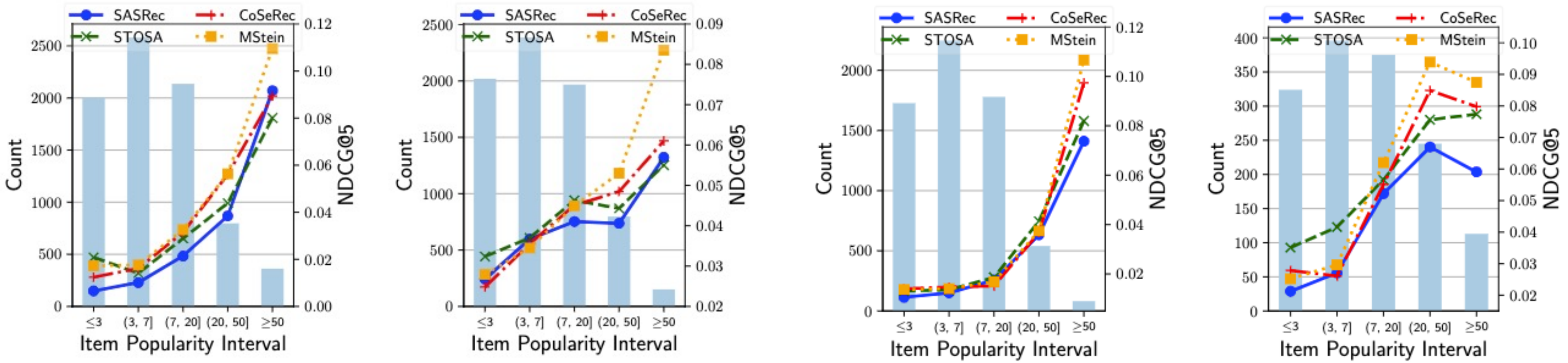
Improvements Analysis (User)



(a) Beauty (b) Toys (c) Tools (d) Office

- Benefits long users

Improvements Analysis (Item)



(a) Beauty **(b) Toys** **(c) Tools** **(d) Office**

- Benefits popular items

Takeaways

- We propose an alternative mutual information measurement based on the Wasserstein distance, with several advantages.
- MStein is more robust and sample efficient.
- MStein improves long users and popular items.

Thanks

Github: https://github.com/zfan20/MStein