中国科学技术大学

University of Science and Technology of China

Adap-τ: Adaptively Modulating Embedding Magnitude for Recommendation

Authors: Jiawei Chen*, Junkang Wu*, Jiancan Wu, Sheng Zhou, Xuezhi Cao, Xiangnan He‡

Paper: https://arxiv.org/pdf/2302.04775.pdf

Lab: USTC Lab for Data Science

Code: https://github.com/junkangwu/Adap_tau

Date: April 10, 2023

Outline

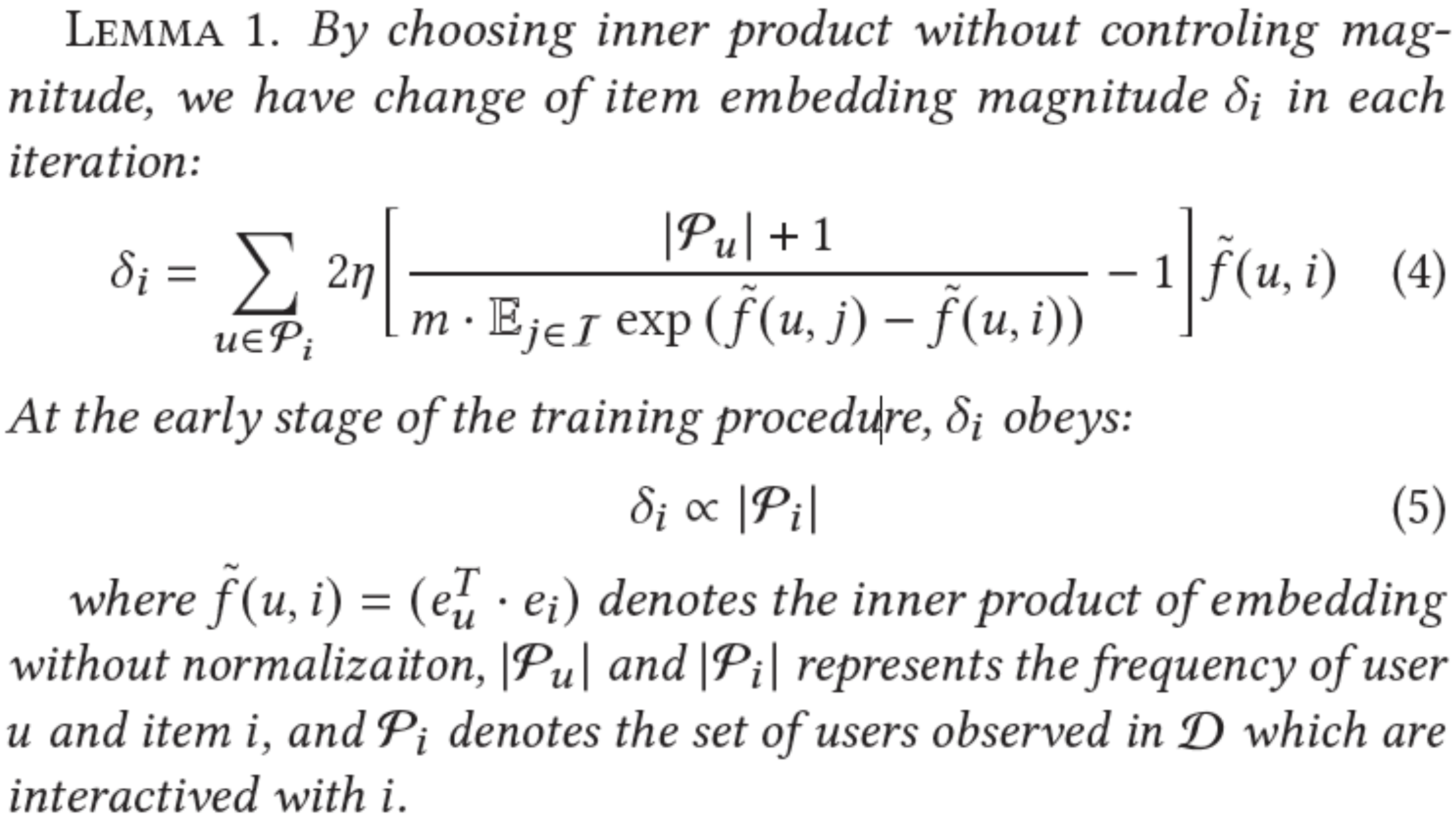
- □ **Background and Motivation**
- □ **Analyses over Embedding Normalization**
- □ **Proposed Method**
- □ **Experiments**
- □ **Summary**

Background and Motivation

- Embedding-based methods achieve competitive performance and support efficient retrieval. **However…**
  - Popular items' embedding **magnitude** grows faster than unpopular ones, which causes excessive contribution to model training and undesirable higher scores due to the potential aggravation of popularity bias from free-varying magnitude.
  - The highly diverse magnitude **prevents model convergence**, even with a proper regularizer, as shown by visual analysis indicating that item embeddings continue to rise instead of converging after numerous epochs.

Analyses over Embedding Normalization

- Necessity of Normalization
  - Theoretical Analysis



Lemma 1. *By choosing inner product without controling magnitude, we have change of item embedding magnitude δ_i in each iteration:*

$$\delta_i = \sum_{u \in \mathcal{P}_i} 2\eta \left[\frac{|\mathcal{P}_u| + 1}{m \cdot \mathbb{E}_{j \in \mathcal{I}} \exp\left(\tilde{f}(u,j) - \tilde{f}(u,i)\right)} - 1 \right] \tilde{f}(u,i) \qquad (4)$$

At the early stage of the training procedure, δ_i obeys:

$$\delta_i \propto |\mathcal{P}_i| \qquad (5)$$

where $\tilde{f}(u,i) = (e_u^T \cdot e_i)$ denotes the inner product of embedding without normalizaiton, $|\mathcal{P}_u|$ and $|\mathcal{P}_i|$ represents the frequency of user u and item i, and $\mathcal{P}_i$ denotes the set of users observed in $\mathcal{D}$ which are interactived with i.

Analyses over Embedding Normalization

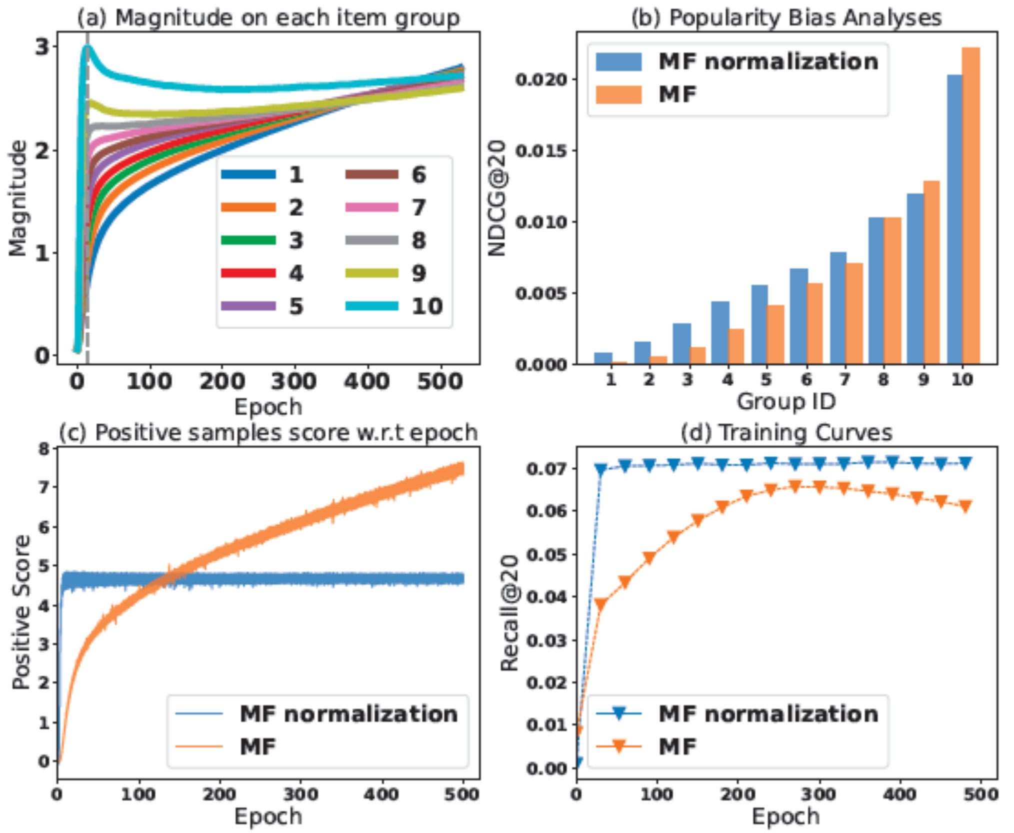
- Necessity of Normalization
  - Empirical Analysis

✓ Free-varying magnitude aggravates popularity bias.

popular items are prone to obtain higher scores as the magnitude directly contributes to model prediction.

✓ Free-varying magnitude hurts convergence

The predicted scores and embedding magnitude of unnormalized methods are still in a state of rising rather than convergence while the performance drop consistently


(a) Magnitude on each item group

(b) Popularity Bias Analyses

(c) Positive samples score w.r.t epoch

(d) Training Curves

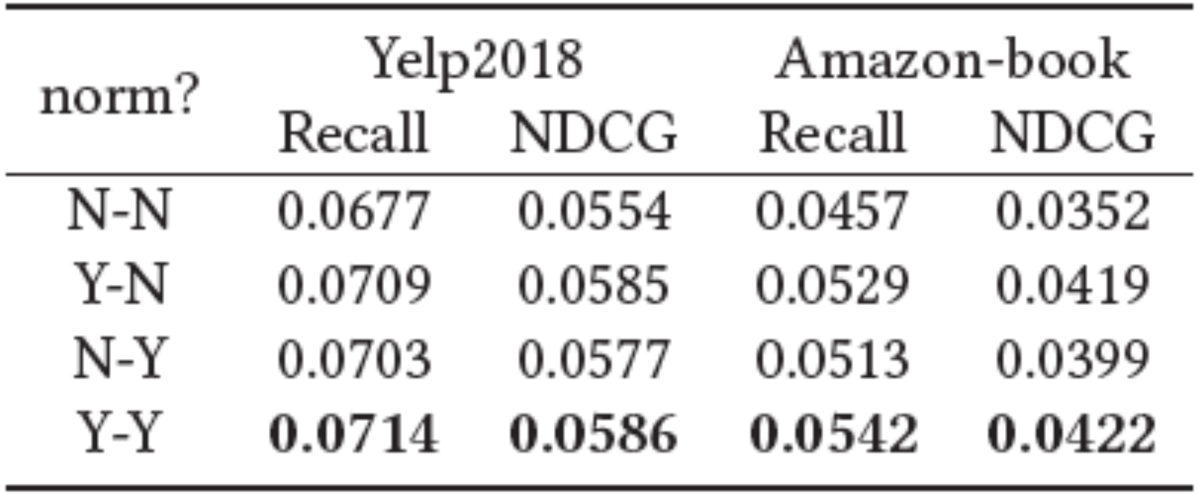
Figure 1: Empirical studies on Yelp2018: Fig. (a) and Fig. (b) represent item embedding magnitude of the different groups across the training procedure and respective performance. The larger GroupID is, the more popular items the group contains. Fig. (c) and Fig. (d) depict the positive samples score and corresponding performance in the training procedure.

Analyses over Embedding Normalization

- Necessity of Normalization
  - Empirical Analysis

✓ Normalization boosts performance.

The model with both-side normalization (i.e., Y-Y) remarkably outperforms the model with one-side normalization (i.e., Y-N or N-Y); and they both surpass the model without normalization (N-N).



norm?	Yelp2018		Amazon-book	
	Recall	NDCG	Recall	NDCG
N-N	0.0677	0.0554	0.0457	0.0352
Y-N	0.0709	0.0585	0.0529	0.0419
N-Y	0.0703	0.0577	0.0513	0.0399
Y-Y	**0.0714**	**0.0586**	**0.0542**	**0.0422**

Analyses over Embedding Normalization

- Limitation of Normalization
  - The performance is highly sensitive to τ
  - Different datasets require rather different τ

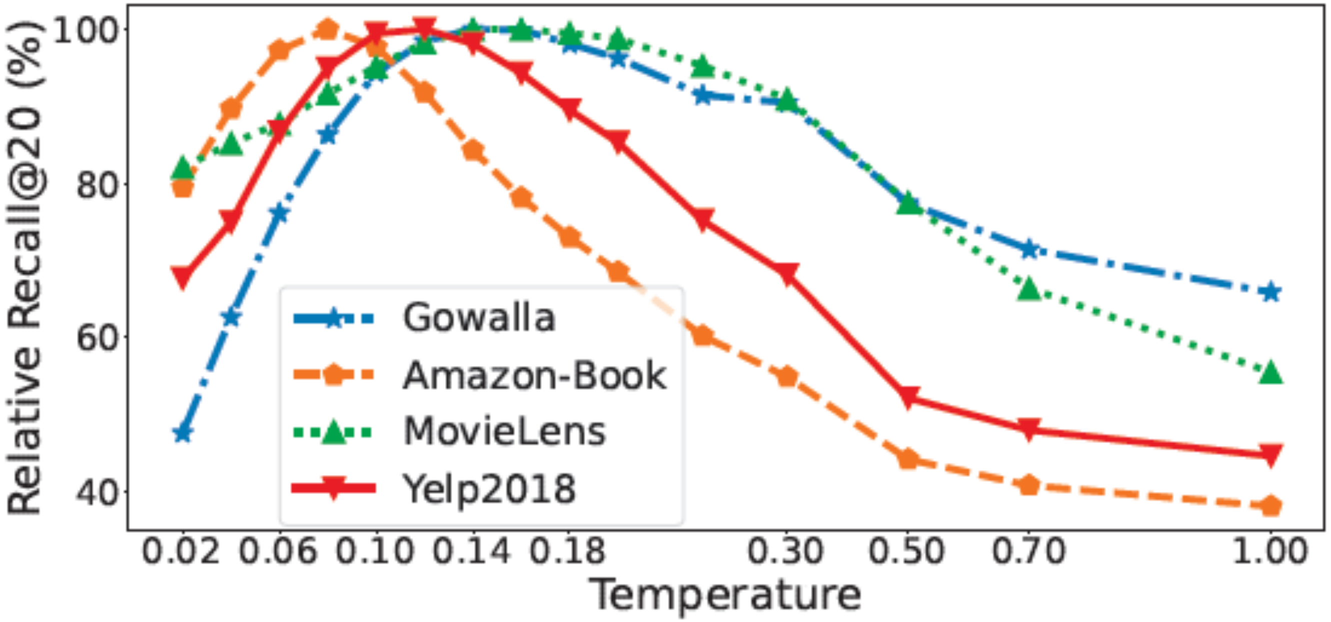


Figure 2: Relative recall@20 over four datasets with τ .

Analyses over Embedding Normalization

- Roles of Temperature
  - Avoiding gradient vanishment.

    Too large or too small τ would cause gradient vanishment.

$$\mathbb{E}_i[|\frac{\partial L}{\partial f(u,i)}|] = \frac{2}{m\tau}\sum_{i\in\mathcal{P}_u} p_{ui}(\tau)(1-\sum_{k\in\mathcal{P}_u} p_{uk}(\tau))$$

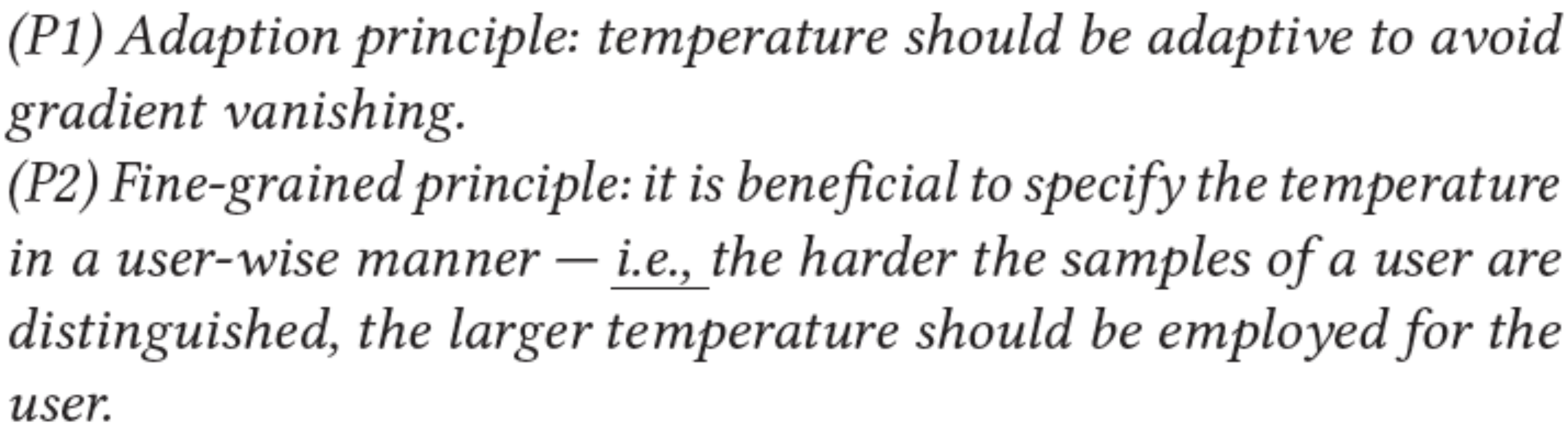
  - Hard-mining [1].

Too small τ would amplify the disparity and focus on hard negative samples.

[1] Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. CVPR

PROPOSED METHOD

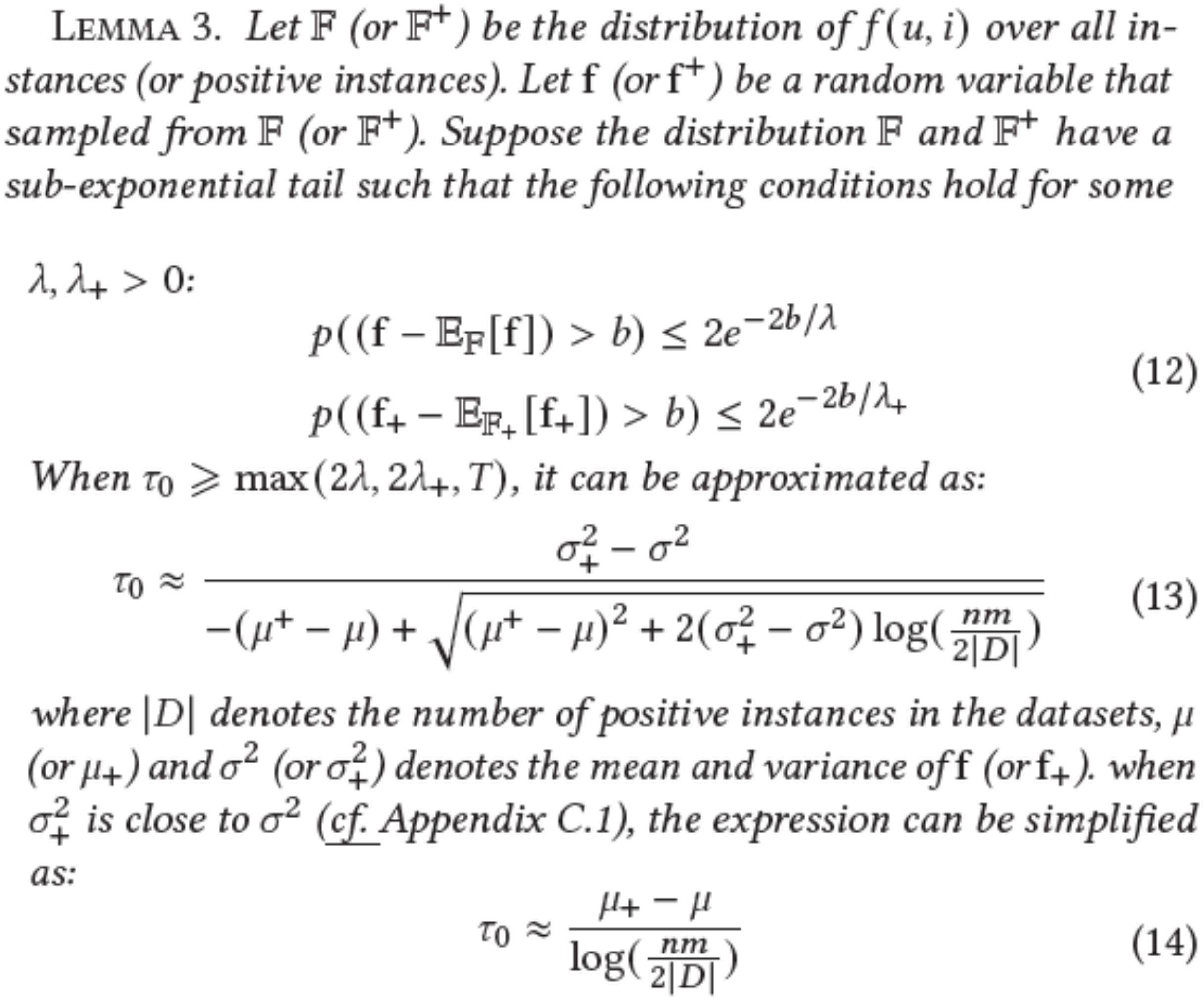
- To adaptively and automatically modulate the embedding magnitude, we propose two principles:



(P1) Adaption principle: temperature should be adaptive to avoid gradient vanishing.
(P2) Fine-grained principle: it is beneficial to specify the temperature in a user-wise manner — i.e., the harder the samples of a user are distinguished, the larger temperature should be employed for the user.

PROPOSED METHOD

❑ Adap-τ_0 : Towards Adaptive Temperature



Lemma 3. *Let $\mathbb{F}$ (or $\mathbb{F}^+$) be the distribution of $f(u, i)$ over all instances (or positive instances). Let f (or f^+) be a random variable that sampled from $\mathbb{F}$ (or $\mathbb{F}^+$). Suppose the distribution $\mathbb{F}$ and $\mathbb{F}^+$ have a sub-exponential tail such that the following conditions hold for some $\lambda, \lambda_+ > 0$:*

$$p((\mathrm{f} - \mathbb{E}_{\mathbb{F}}[\mathrm{f}]) > b) \leq 2e^{-2b/\lambda}$$
$$p((\mathrm{f}_+ - \mathbb{E}_{\mathbb{F}_+}[\mathrm{f}_+]) > b) \leq 2e^{-2b/\lambda_+} \tag{12}$$

When $\tau_0 \geqslant \max(2\lambda, 2\lambda_+, T)$, it can be approximated as:

$$\tau_0 \approx \frac{\sigma_+^2 - \sigma^2}{-(\mu^+ - \mu) + \sqrt{(\mu^+ - \mu)^2 + 2(\sigma_+^2 - \sigma^2)\log(\frac{nm}{2|D|})}} \tag{13}$$

where $|D|$ denotes the number of positive instances in the datasets, μ (or μ_+) and σ^2 (or σ_+^2) denotes the mean and variance of f (or f_+). when σ_+^2 is close to σ^2 (cf. Appendix C.1), the expression can be simplified as:

$$\tau_0 \approx \frac{\mu_+ - \mu}{\log(\frac{nm}{2|D|})} \tag{14}$$

PROPOSED METHOD

❑ Adap-τ : Towards Adaptive Fine-grained Temperature

We introduce personalized temperatures τ_u for each user and leverage a Superloss [2] to supervise their learning.

$$J = \frac{L(u) - m_u}{\tau_u} + \beta(\log \tau_u - \log \tau_0)^2$$

In fact, we have a closed-form solution

$$\tau_u^* = \tau_0 \cdot \exp(\mathbb{W}(\max(-\frac{1}{e}, \frac{L(u) - m_u}{2\beta})))$$

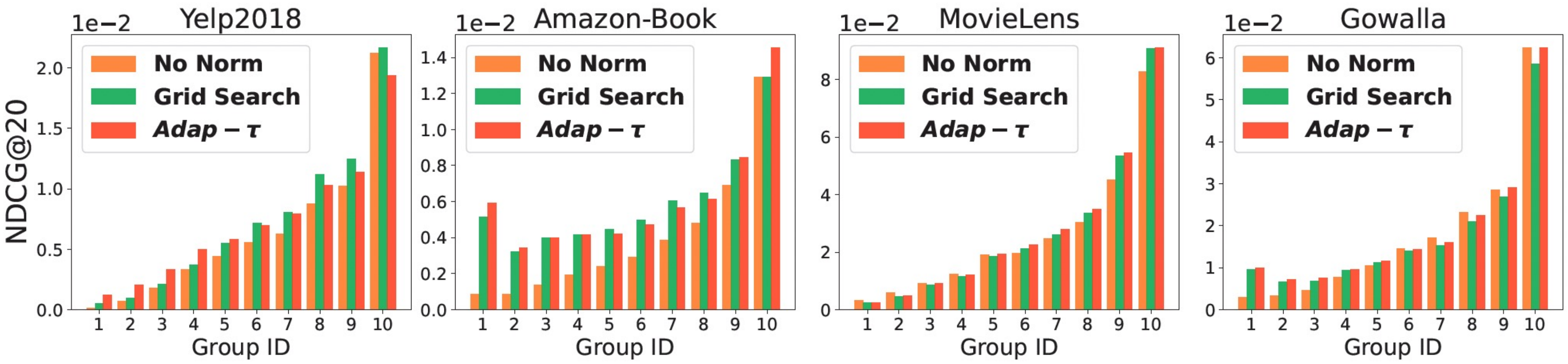
[2] Thibault Castells, Philippe Weinzaepfel, and Jerome Revaud. 2020. SuperLoss: A Generic Loss for Robust Curriculum Learning. In NeurIPS.

Experiments

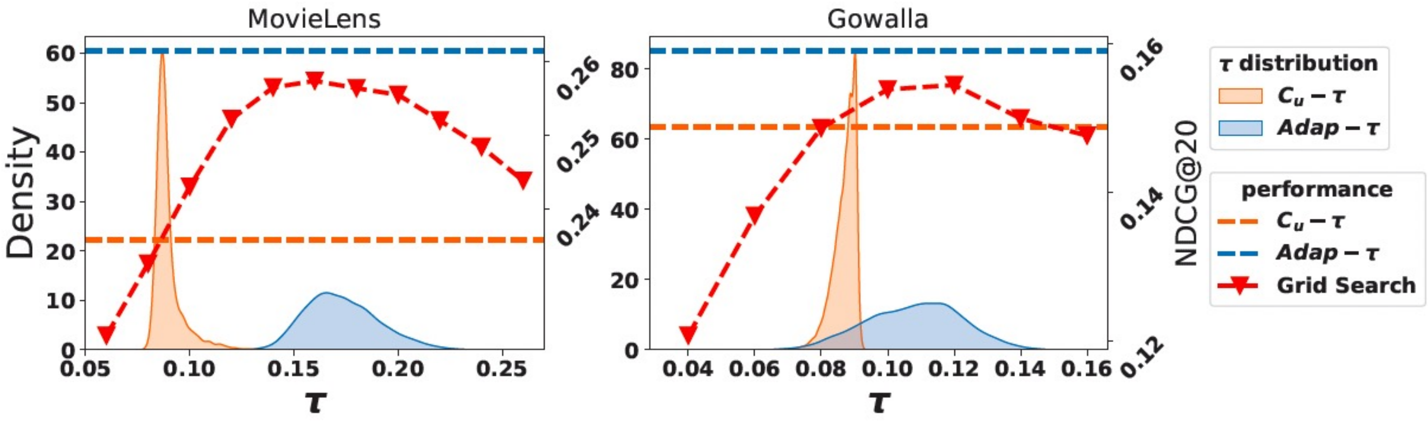
- ❑ How does Adap-τ perform compared with other strategies?
- ❑ Does our Adap-τ adapt to different datasets and users?
- ❑ How does the model equipped with embedding normalization and adaptive τ perform compared with state-of-the-art in terms of both accuracy and efficiency?

Experiments

Backbone	strategy	Yelp2018		Amazon-book		Movielens		Gowalla	
		Recall	NDCG	Recall	NDCG	Recall	NDCG	Recall	NDCG
MF	No Norm	0.0677	0.0554	0.0457	0.0352	0.2721	0.2525	0.1616	0.1366
	Grid Search τ	0.0714	0.0586	0.0542	0.0422	0.2789	0.2624	0.1761	0.1399
	C-τ	0.0647	0.0528	0.0538	0.0418	0.2472	0.2260	0.1723	0.1362
	Cu-τ	0.0691	0.0566	0.0541	0.0421	0.2600	0.2398	0.1751	0.1383
	Adap-τ_0	0.0714	0.0585	0.0549	0.0427	0.2792	0.2638	0.1754	0.1386
	Adap-τ	**0.0721**	**0.0594**	**0.0553**	**0.0430**	**0.2815**	**0.2673**	**0.1838**	**0.1506**
LightGCN	No Norm	0.0649	0.0530	0.0411	0.0315	0.2576	0.2427	0.1830	0.1554
	Grid Search τ	0.0730	0.0605	0.0596	0.0477	0.2767	0.2575	0.1878	0.1577
	C-τ	0.0653	0.0537	0.0571	0.0453	0.2529	0.2282	0.1731	0.1431
	Cu-τ	0.0690	0.0571	0.0586	0.0468	0.2582	0.2357	0.1797	0.1488
	Adap-τ_0	0.0724	0.0603	0.0601	0.0480	0.2744	0.2571	0.1841	0.1526
	Adap-τ	**0.0733**	**0.0612**	**0.0612**	**0.0490**	**0.2787**	**0.2615**	**0.1901**	**0.1590**



Experiments



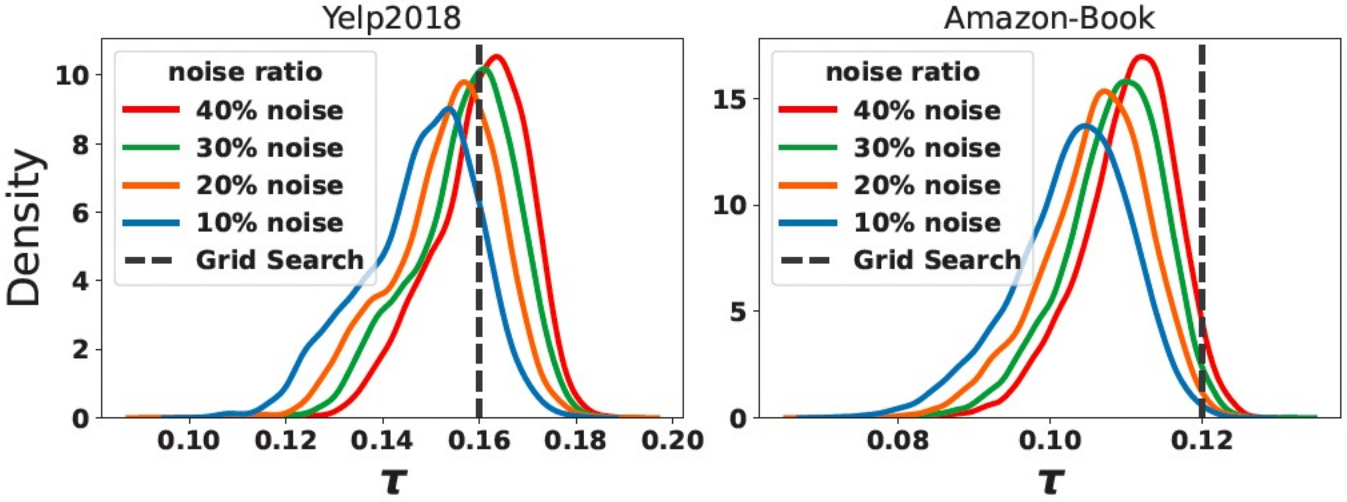
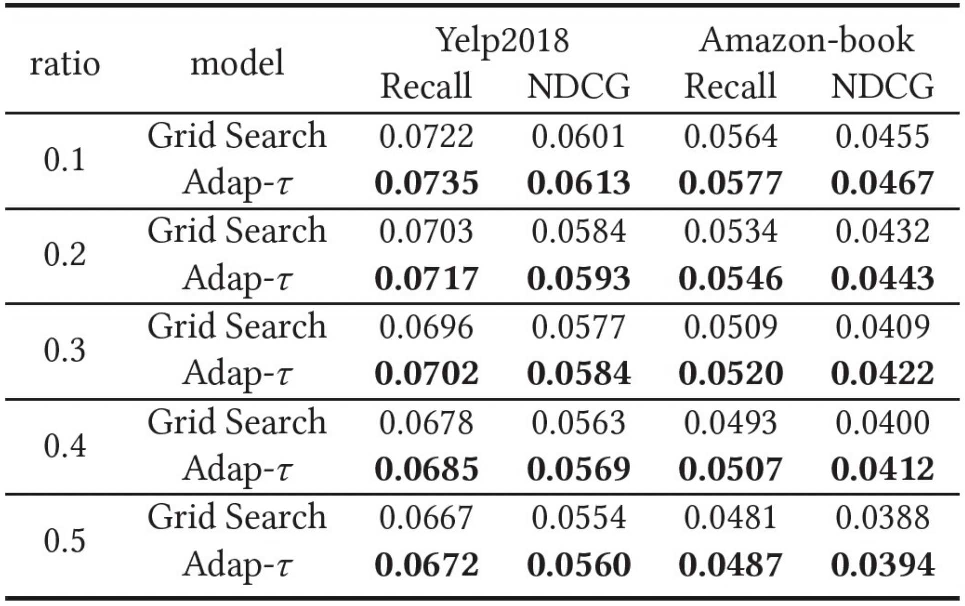


Table 3: Results of MF under different ratio "noisy data"



ratio	model	Yelp2018		Amazon-book	
		Recall	NDCG	Recall	NDCG
0.1	Grid Search	0.0722	0.0601	0.0564	0.0455
	Adap-τ	**0.0735**	**0.0613**	**0.0577**	**0.0467**
0.2	Grid Search	0.0703	0.0584	0.0534	0.0432
	Adap-τ	**0.0717**	**0.0593**	**0.0546**	**0.0443**
0.3	Grid Search	0.0696	0.0577	0.0509	0.0409
	Adap-τ	**0.0702**	**0.0584**	**0.0520**	**0.0422**
0.4	Grid Search	0.0678	0.0563	0.0493	0.0400
	Adap-τ	**0.0685**	**0.0569**	**0.0507**	**0.0412**
0.5	Grid Search	0.0667	0.0554	0.0481	0.0388
	Adap-τ	**0.0672**	**0.0560**	**0.0487**	**0.0394**

Experiments

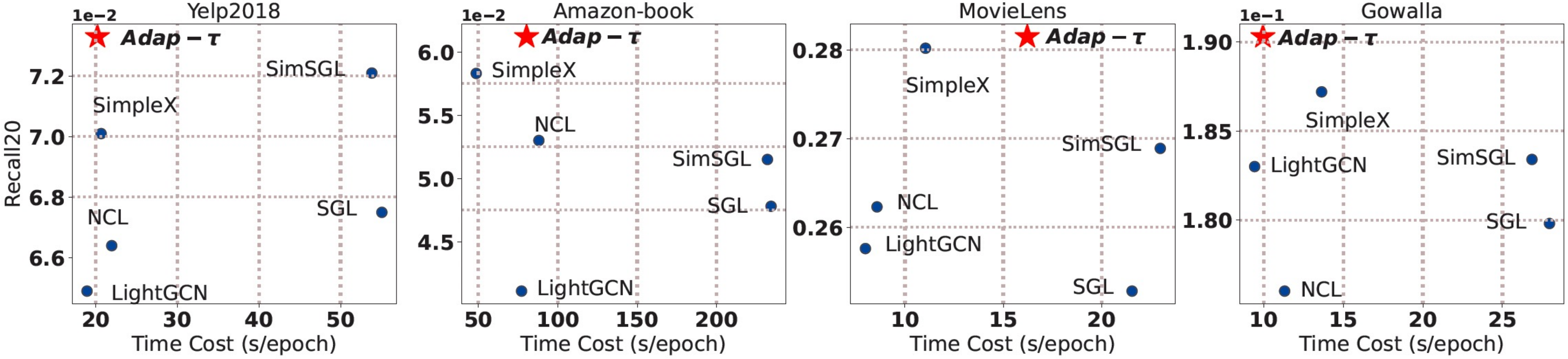


Figure 5: Performance comparisons in terms of both recommendation accuracy and efficiency.

Summary

- Embedding normalization is crucial in RS
  - We verify it from theoretical and empirical analysis
  - High sensitive to the Temperature limits its potential

- We provide two principles to guide the adaptive learning of τ
  - We verify it with different backbones in numerous dataset