SIGIR TAIPEI | TAIWAN 2023

The 46th International ACM SIGIR Conference on Research and Development in Information Retrieval

Collaborative Residual Metric Learning

Tianjun Wei, Jianghong Ma, Tommy W.S. Chow

July 26, 2023
Taipei

香港城市大學
City University of Hong Kong

1920 HIT

Collaborative Filtering

Matrix Completion

- Completing the elements of the user-item interaction matrix that are not 1

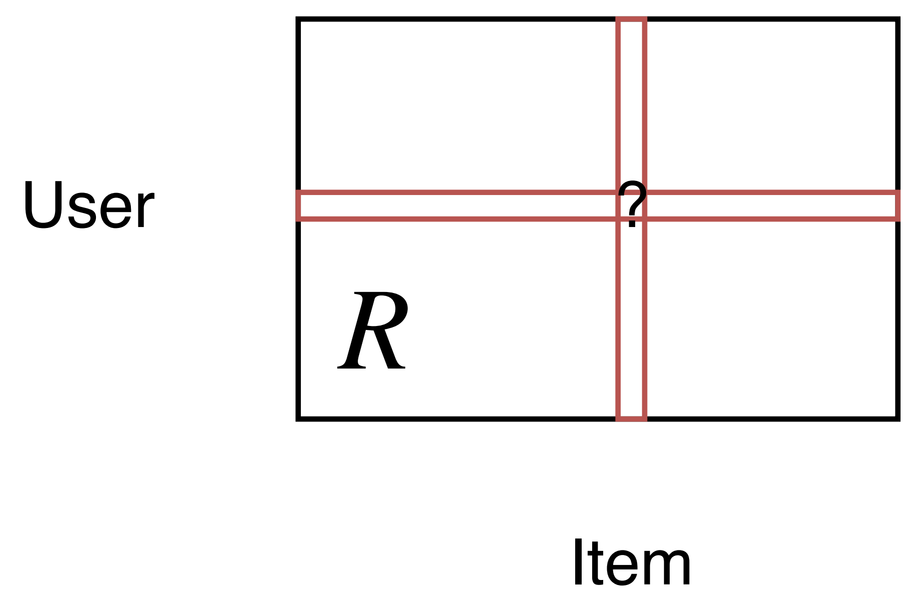
Link Prediction

- Predicting unconnected edges in user-item bipartite graphs

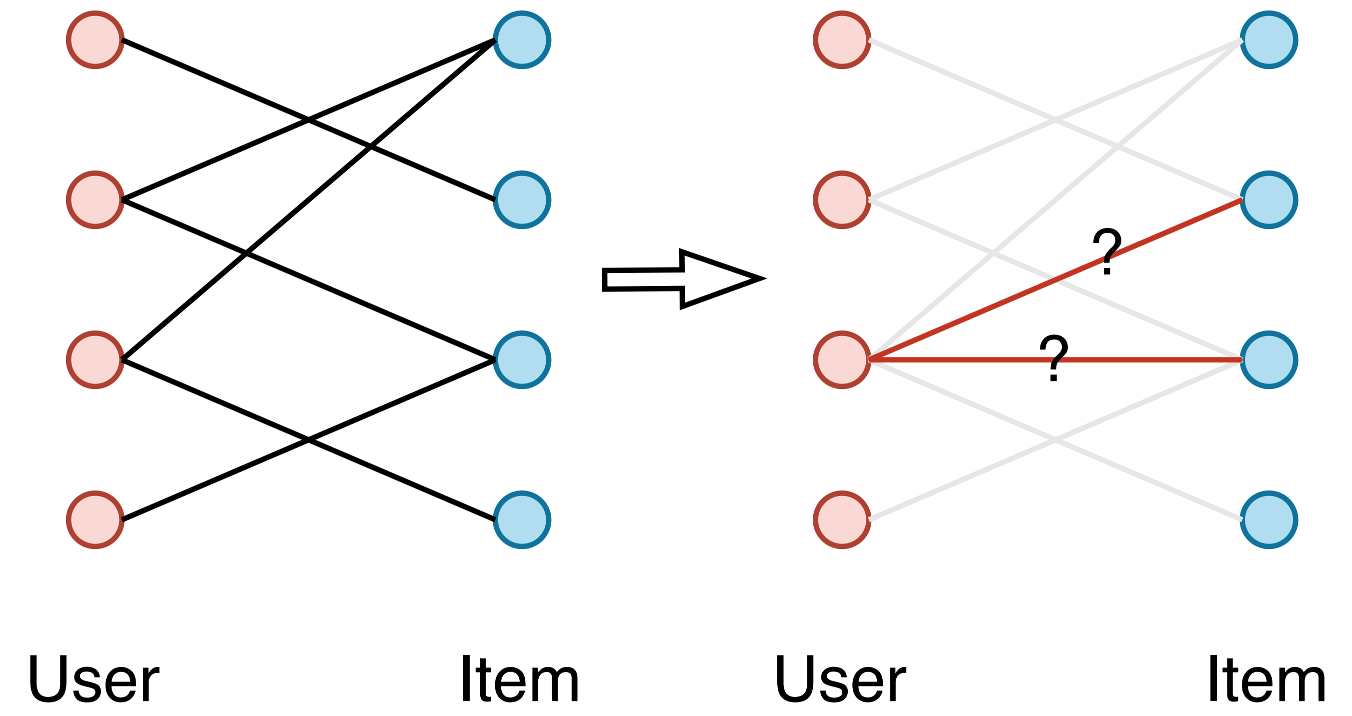
Nature of Recommendation Tasks

- Focus more on ranking of scores
- Focus more on top positions

Collaborative Filtering as a Matrix Completion Problem



Collaborative Filtering as a Link Prediction Problem



Type of Methods

Traditional Matrix Factorization

- Representing users and items with fixed-length vectors
- Both a trainable parameters ($N \propto |U| + |I|$)

Asymmetric Matrix Factorization

- Training only item vectors
- Representing users with the aggregation of item vectors
- No need to train when new users are added

Linear Autoencoder

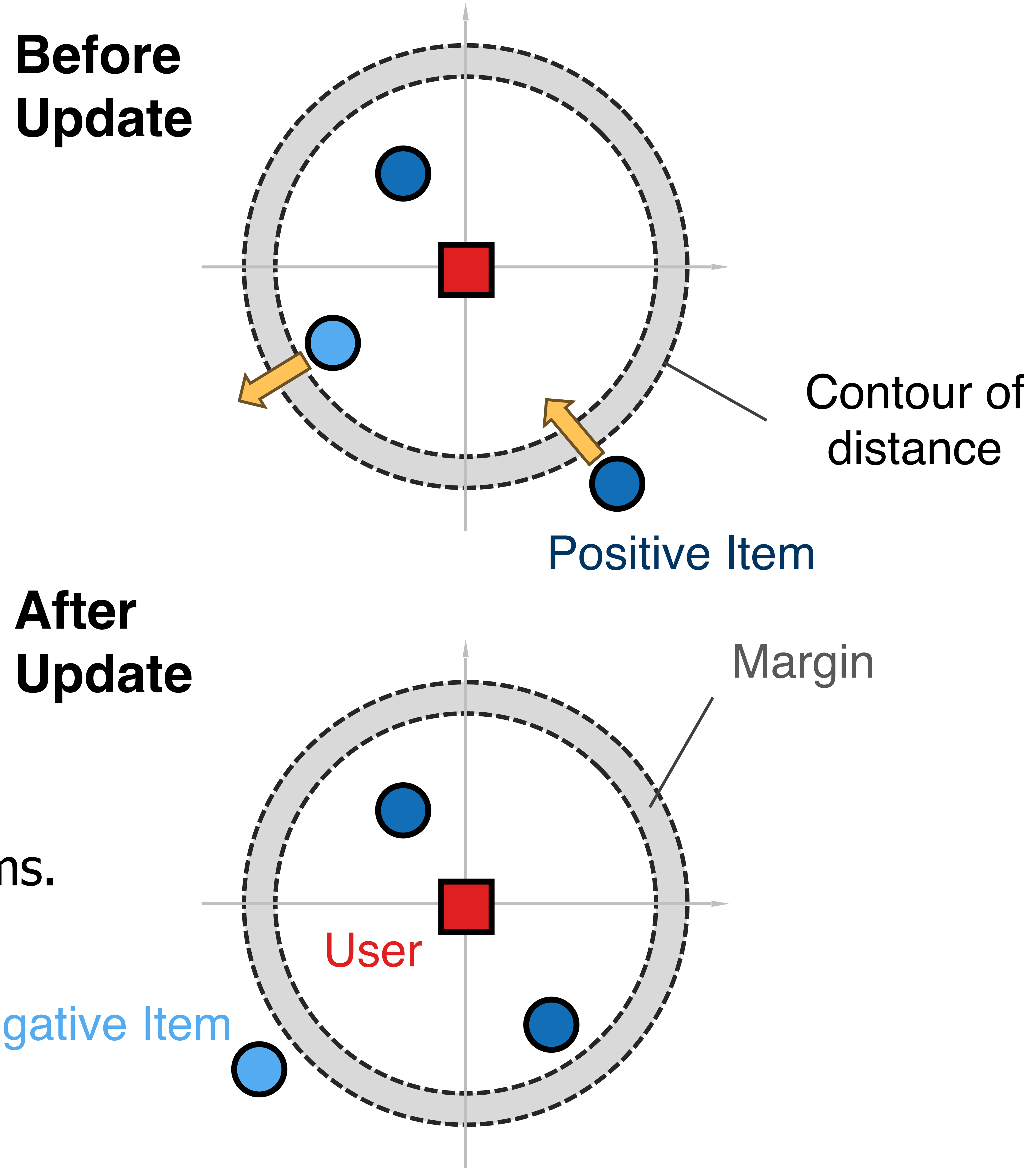
- Full-rank Extension of asymmetric matrix factorization
- Adding sparse constraint to low the cost of storage and inference
- SLIM (ICDM 2011), EDLAE (Netflix, NIPS 2020)

Metric Learning in Collaborative Filtering

Metric Learning

- The objective of metric learning is to learn a **valid distance metric** to pull similar nodes closer, push dissimilar nodes farther away:

- ☑ Non-negativity: $d(\mathbf{x}_i, \mathbf{x}_j) \geq 0$
- ☑ Identity: $d(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \rightarrow i = j$
- ☑ Symmetry: $d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i)$
- ☑ Triangle inequality: $d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_k) + d(\mathbf{x}_k, \mathbf{x}_j)$



Collaborative Metric Learning (CML)

- Updating vectors to decrease the Euclidean distance between users and interacted (similar) items, and the opposite for uninteracted items.
- Adopting triplet hinge loss, with a margin ζ

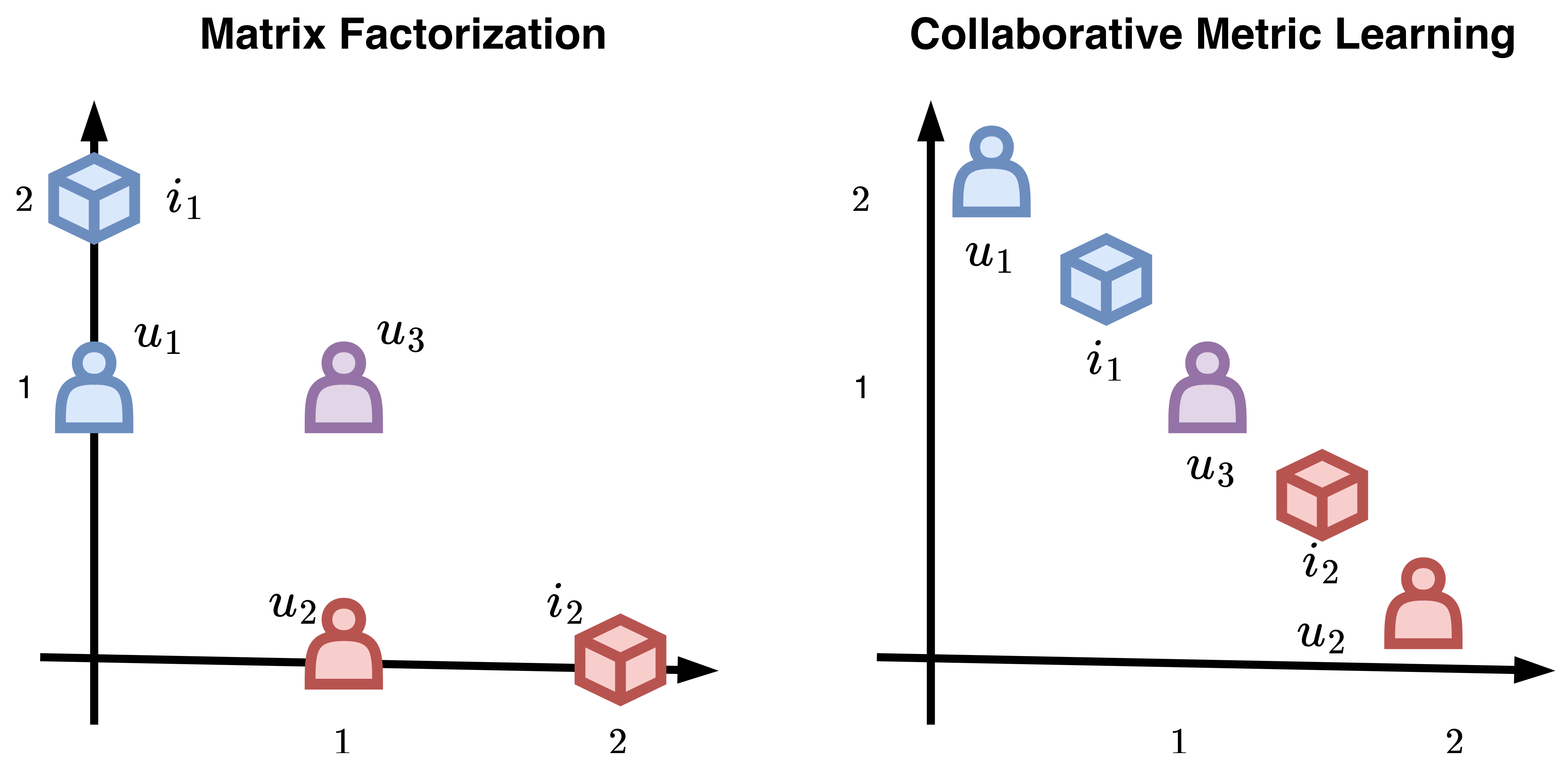
$$L = (d^2(\mathbf{e}_u, \mathbf{e}_i) - d^2(\mathbf{e}_u, \mathbf{e}_j) + \zeta)_+$$

Cheng-Kang Hsieh, Longqi Yang, Yin Cui, Tsung-Yi Lin, Serge Belongie, and Deborah Estrin. 2017. Collaborative Metric Learning. In WWW'17

Metric Learning in Collaborative Filtering

Propagation of Similarity

- MF is not reliable on capturing u-u and i-i similarity
- CML shows the capability of propagating similarities through triangle inequality



Special Case of Metric Learning

- Generalized Mahalanobis (GM) Distance

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W} (\mathbf{x}_i - \mathbf{x}_j)}$$

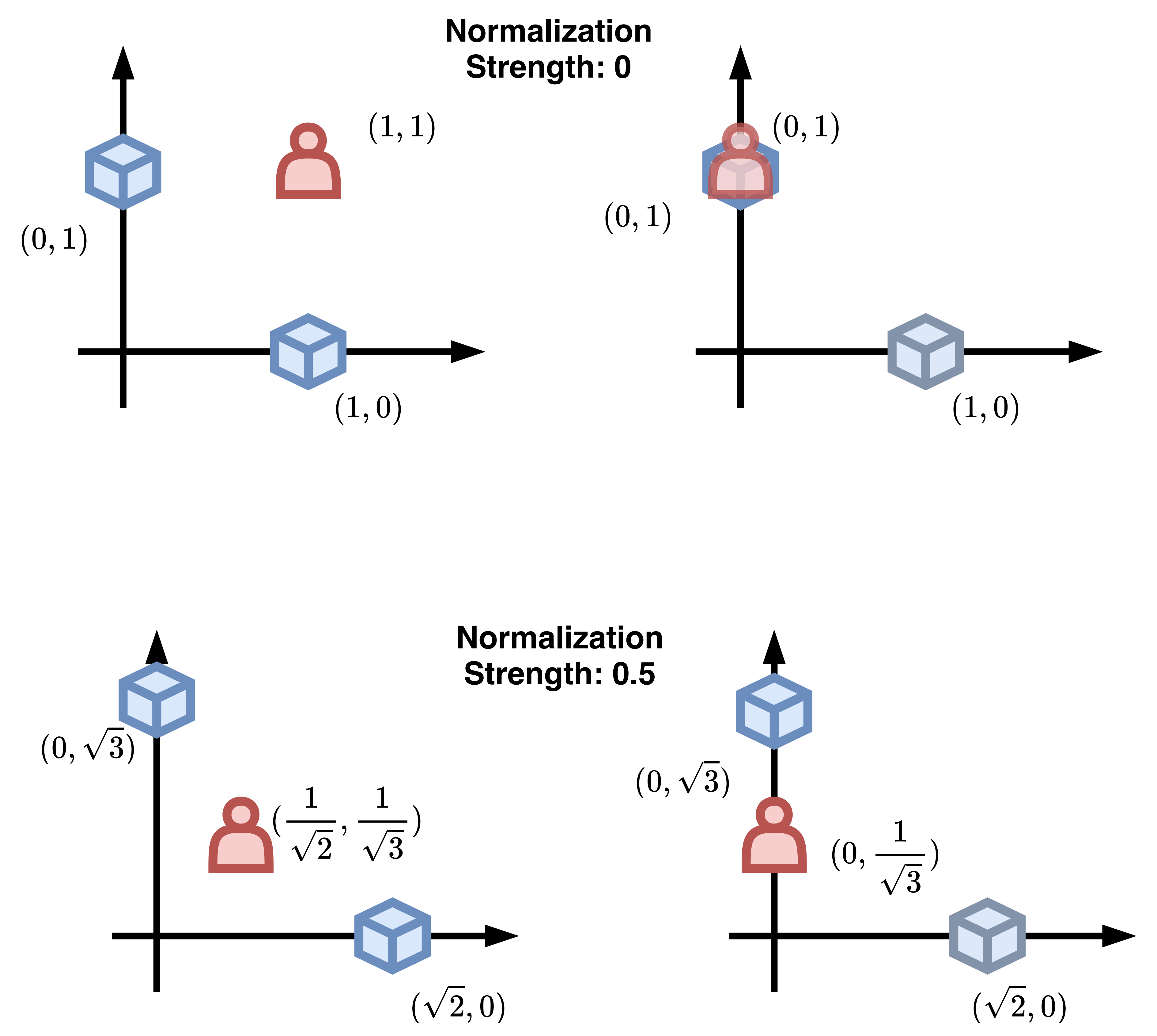
- W must be symmetric positive semidefinite (PSD)
- CML learns a rank-d weight matrix $\mathbf{W} \in \mathbb{R}^{(|U|+|I|)\times(|U|+|I|)}$
- Cannot generalize to methods like asymmetric matrix factorization

Cheng-Kang Hsieh, Longqi Yang, Yin Cui, Tsung-Yi Lin, Serge Belongie, and Deborah Estrin. 2017. Collaborative Metric Learning. In WWW'17

Collaborative Signals

Representing Users with Items

- User feature: $\mathbf{P} \in \mathbb{R}^{|U|\times|I|} = \mathbf{D}_I^{-t}\mathbf{R}$
- Item feature: $\mathbf{Q} \in \mathbb{R}^{|I|\times|I|} = \mathbf{D}_I^{t}$
- Preference score: $y_{ui} = \mathbf{p}_u^T\mathbf{C}\mathbf{q}_i$



Fitting More Methods by Introducing Normalization Strength

Methods	t	C
• Asymmetric Matrix Factorization	0	$\mathbf{E}_I^T\mathbf{E}_I$
• Linear Autoencoder	0	$\mathbf{W}_{sp,diag\_0}$
• Graph Filtering Model	0.5	$\mathbf{V}\mathbf{V}^T$

Incorporate Signal-based Models with ML

Linear Autoencoder not Working Well with ML

☑ Symmetry: can be added as a constraint to the optimization

☐ PSD: almost impossible to satisfy because of the diagonal zero constraint, $diag(\mathbf{W}) = 0$

Focus instead on the difference of distances

- Recommendation task focuses on the relative relationship of preference scores (distance)

$$\begin{aligned}
\Delta D^2 &= d^2(\mathbf{p}_u, \mathbf{q}_i) - d^2(\mathbf{p}_u, \mathbf{q}_j) \\
&= \mathbf{q}_i^T\mathbf{W}\mathbf{q}_i - \mathbf{q}_j^T\mathbf{W}\mathbf{q}_j - 2\mathbf{p}_u^T\mathbf{W}(\mathbf{q}_i - \mathbf{q}_j) + (\mathbf{p}_i^T\mathbf{W}\mathbf{p}_i - \mathbf{p}_i^T\mathbf{W}\mathbf{p}_i) \\
&= W_{ii}(d_i^{2t} - 2R_{ui}) - W_{jj}(d_j^{2t} - 2R_{uj}) - 2\underline{\mathbf{p}_u^T\mathbf{H}(\mathbf{q}_i} - \mathbf{q}_j) \\
&= W_{ii}(d_i^{2t} - 2R_{ui}) - W_{jj}(d_j^{2t} - 2R_{uj}) + \Delta Y
\end{aligned}$$

y_{ui} in signal-based models

where $H = W - diag(W)$ is called the ***Hollow matrix***.

- Eliminate redundant terms in the GM distances
- Separate diagonal and non-diagonal entries

Incorporate Signal-based Models with ML

Finding Alternative Solutions

- $\mathbf{W} = \mathbf{H} + \mathbf{X}$ can always be PSD when $\mathbf{X} = \omega \mathbf{D}_I^{-2t}$, $\mathbf{H}$ is the hollow matrix.
- Derive the relationships between ΔD^2 and ΔY (Preference Residual):

$$\begin{aligned}\Delta D^2 - \Delta Y = &\ W_{ii}(d_i^{2t} - 2R_{ui}) - W_{jj}(d_j^{2t} - 2R_{uj}) \\ = &\ 2\omega(d_i^{-2t}R_{uj} - d_j^{-2t}R_{ui})\end{aligned}$$

A. When $R_{ui} = R_{uj} = 0$

(item i and j are both uninteracted)

$$\Delta D^2 - \Delta Y = 0$$

- Δ Distance = Preference Residual
- Critical to the model inference process

B. When $R_{ui} = 1, R_{uj} = 0$

(item i is interacted, item j is uninteracted)

$$\Delta D^2 - \Delta Y = 2\omega d_i^{-2t}$$

- The bias is always positive
- Useful in the model training process

How does t affect the recommendations?

Non-negativity Constraint

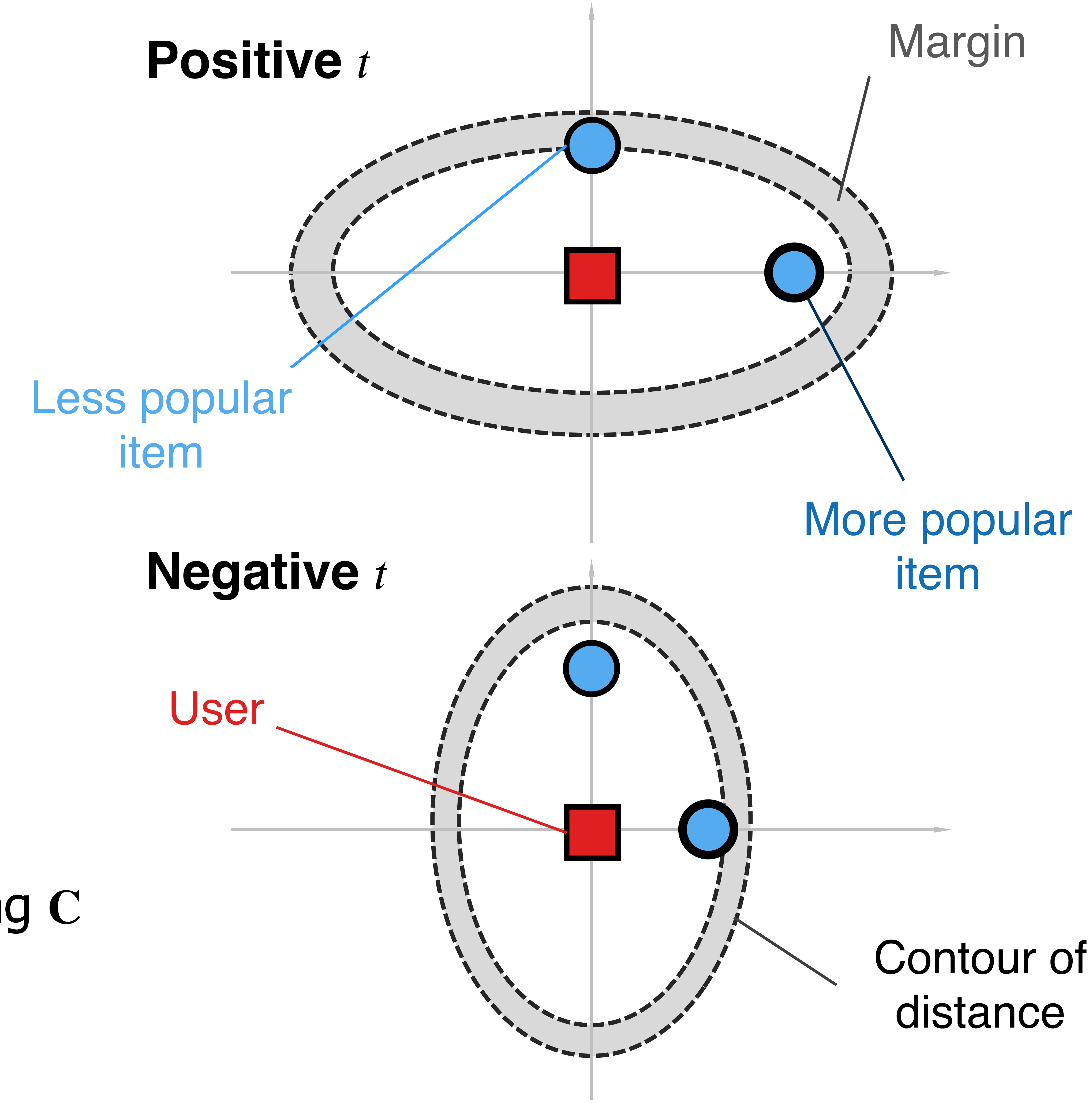
- General form of signal-based model with non-negativity constraint and l_2 regularization:

$$\underset{\mathbf{W}}{\text{minimize}} \;\; \mathscr{L}(\mathbf{W}) + \|\mathbf{W}\|_F^2, \;\; s.t.\, \mathbf{W} \geq 0$$

- Preference score for each (u, i) pair:

$$y_{ui} = \sum_{j \in I_u^+} \left(\frac{d_i}{d_j}\right)^t W_{ji}$$

- Counterfactual results: Larger t will increase the chance of popular items to be recommended
- Improve the **novelty** of the recommendations by decreasing $\mathbf{C}$



Collaborative Residual Metric Learning (CoRML)

Triplet Residual Margin Loss

Triplet Margin Loss

$$L = \sum_{u \in U} \sum_{(i^+, i^-) \in (I_u \times I \backslash I_u)} (d_{ui^+}^2 - d_{ui^-}^2 + \zeta)_+$$

- Replace $d_{ui^+}^2 - d_{ui^-}^2$ (ΔD^2) with ΔY
- Replace ζ with $2\omega d_{i^+}^{-2t}$ as an adaptive margin

Triplet Residual Margin Loss

$$L_{TRM} = \sum_{u \in U} \sum_{(i^+, i^-) \in (I_u \times I \backslash I_u)} (y_{ui^-} - y_{ui^+})_+$$

$$= \sum_{u \in U} \Big(\sum_{i^+ \in I_u} \alpha_{ui^+} y_{ui^+} + \sum_{i^- \notin I_u} \beta_{ui^-} y_{ui^-} \Big)$$

$$\alpha_{ui^+} = \sum_{i^- \notin I_u} - \frac{\delta(y_{ui^-} > y_{ui^+})}{|I| - |I_u|}, \ \beta_{ui^-} = \sum_{i^+ \in I_u} \frac{\delta(y_{ui^-} > y_{ui^+})}{|I_u|}$$

Approximated Ranking Weights

- α and β are are coefficients dynamically updated by the ranking of the value of y_{ui} for user u
- Use numerical value to approximate ranking

$$\tilde{\alpha}_{ui^+} = \phi y_{ui^+} - 1, \tilde{\beta}_{ui^-} = \phi y_{ui^-}$$

Scaling factor

$$\phi_u = \epsilon \Big(\frac{d_u}{max_{u \in U} d_u} \Big)^{-t_u}$$

- Global scaling (ϵ): rescale y_{ui^+} to map $\tilde{\alpha}_{ui^+}$ to negative values, and keep $\tilde{\beta}_{ui^-}$ positive
- User-degree scaling (t_u): reduce the effects of different number of non-zero entries in the collaborative signal of each user

Loss Function

$$L_{CoRML} = \sum_{u \in U} \sum_{i \in I} y_{ui}(\phi_u y_{ui} - R_{ui}) = tr(\mathbf{Y}^T(\mathbf{\Phi Y} - \mathbf{R}))$$

Hybrid Preference Score

$$\mathbf{Y} = \mathbf{R}(\lambda \underbrace{\mathbf{D}_I^{-t}\mathbf{H}\mathbf{d}_I^t}_{} + (1-\lambda)\underbrace{\mathbf{D}_I^{-\frac{1}{2}}\mathbf{G}\mathbf{D}_I^{\frac{1}{2}}}_{})$$

Extension from linear autoencoder with adjustable t

Extension from graph signal model $\mathbf{G} = (\mathbf{V}\mathbf{V}^T - diag(\mathbf{V}\mathbf{V}^T))_+$

Optimization Problem

$$\min_{\boldsymbol{H}} tr\left(\boldsymbol{Y}^T(\boldsymbol{\Phi}\boldsymbol{Y} - \boldsymbol{R})\right),$$

$$s.t. \quad diag(\boldsymbol{H}) = 0, \boldsymbol{H} \geq 0, \boldsymbol{H} = \boldsymbol{H}^T$$

- Optimized through Alternating Directions Method of Multipliers (ADMM)

CoRML

Experiments

Performance Comparison

Dataset

- 4 real-world public datasets

Evaluation Metrics

- NDCG@K
- MRR@K

Baselines

- CML models
- Signal-based Models
- GCN models

Dataset	Metric	CML	L-CML	DPCML	SLIM	EASE	RecVAE	GFCF	UltraGCN	SimGCL	CoRML
Pinterest	NDCG@5	0.0509	0.0594	0.0563	0.0488	0.0558	0.0516	0.0620	0.0572	0.0616	***0.0655**
	NDCG@10	0.0665	0.0766	0.0724	0.0630	0.0704	0.0668	0.0785	0.0729	0.0783	***0.0824**
	NDCG@20	0.0897	0.1021	0.0965	0.0841	0.0921	0.0895	0.1031	0.0962	0.1031	***0.1076**
	MRR@5	0.1018	0.1186	0.1133	0.0957	0.1125	0.1024	0.1239	0.1146	0.1237	***0.1306**
	MRR@10	0.1164	0.1343	0.1283	0.1084	0.1262	0.1164	0.1390	0.1292	0.1390	***0.1458**
	MRR@20	0.1261	0.1444	0.1381	0.1171	0.1353	0.1258	0.1488	0.1387	0.1488	***0.1556**
Gowalla	NDCG@5	0.0853	0.0985	0.0999	0.1100	0.1211	0.0890	0.1174	0.1108	0.1229	***0.1317**
	NDCG@10	0.0953	0.1093	0.1087	0.1156	0.1268	0.0978	0.1257	0.1181	0.1295	***0.1383**
	NDCG@20	0.1125	0.1281	0.1261	0.1302	0.1412	0.1140	0.1440	0.1348	0.1460	***0.1554**
	MRR@5	0.1533	0.1743	0.1811	0.1912	0.2186	0.1613	0.2121	0.2001	0.2235	***0.2334**
	MRR@10	0.1682	0.1899	0.1957	0.2043	0.2323	0.1752	0.2269	0.2144	0.2377	***0.2479**
	MRR@20	0.1768	0.1984	0.2040	0.2118	0.2393	0.1832	0.2352	0.2225	0.2454	***0.2558**
Yelp2018	NDCG@5	0.0483	0.0574	0.0556	0.0535	0.0611	0.0525	0.0587	0.0585	0.0646	***0.0690**
	NDCG@10	0.0521	0.0617	0.0592	0.0554	0.0628	0.0558	0.0617	0.0621	0.0676	***0.0716**
	NDCG@20	0.0629	0.0742	0.0709	0.0644	0.0722	0.0663	0.0731	0.0737	0.0795	***0.0832**
	MRR@5	0.1007	0.1188	0.1156	0.1117	0.1277	0.1106	0.1236	0.1234	0.1349	***0.1435**
	MRR@10	0.1149	0.1345	0.1304	0.1245	0.1413	0.1247	0.1380	0.1385	0.1499	***0.1586**
	MRR@20	0.1241	0.1443	0.1399	0.1327	0.1496	0.1336	0.1472	0.1478	0.1594	***0.1679**
ML-20M	NDCG@5	0.2319	0.2731	0.2620	0.2785	0.3025	0.3045	0.2718	0.2365	0.2675	***0.3189**
	NDCG@10	0.2326	0.2689	0.2588	0.2710	0.2934	0.3033	0.2671	0.2280	0.2644	***0.3103**
	NDCG@20	0.2486	0.2832	0.2725	0.2813	0.3036	0.3204	0.2799	0.2369	0.2794	***0.3212**
	MRR@5	0.3761	0.4341	0.4190	0.4478	0.4829	0.4777	0.4356	0.3919	0.4310	***0.4967**
	MRR@10	0.3932	0.4494	0.4347	0.4621	0.4963	0.4923	0.4506	0.4063	0.4466	***0.5098**
	MRR@20	0.4002	0.4554	0.4409	0.4677	0.5014	0.4976	0.4566	0.4124	0.4527	***0.5149**

Detailed Analysis

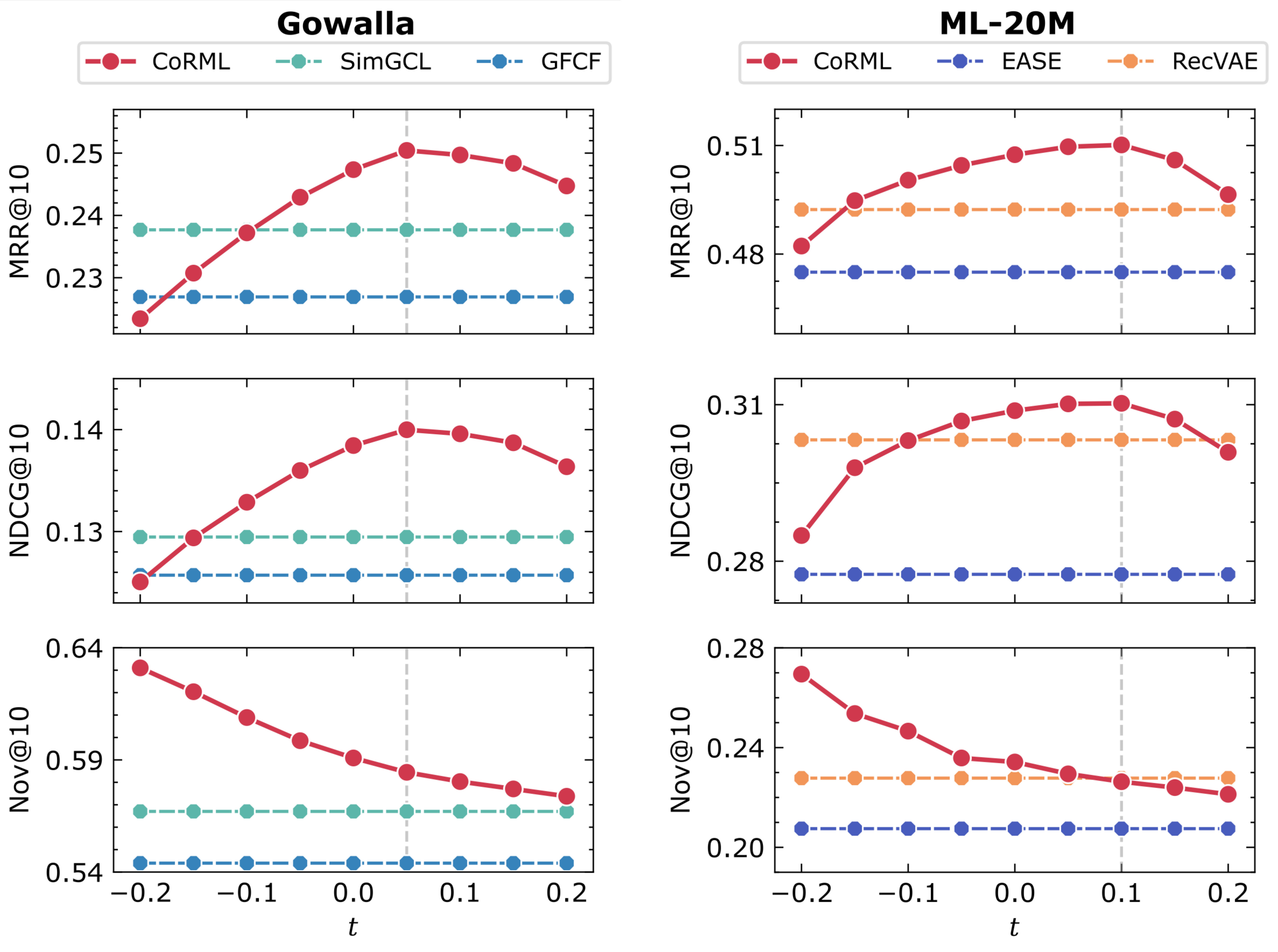
Mitigating Popularity Bias

Novelty

- Measure the popularity of top-K items recommended

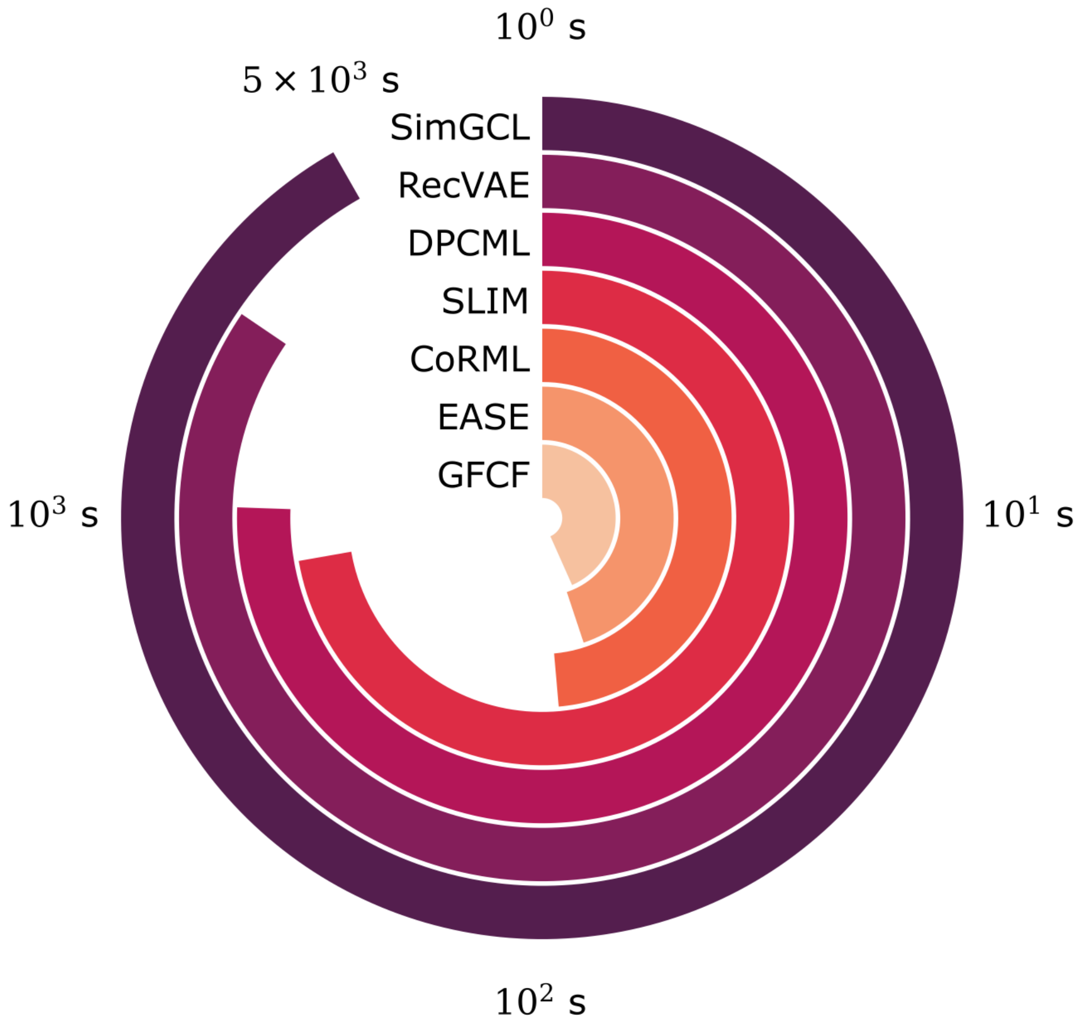
$$Nov@K = \frac{1}{|U|K} \sum_{u}^{|U|} \sum_{i}^{K} -\frac{1}{\log_2 |U|} \log_2 \frac{d_i}{|U|}$$

- Smaller normzalization strength increases the novelty of recommendation
- Accuracy and novelty are not just trade-off

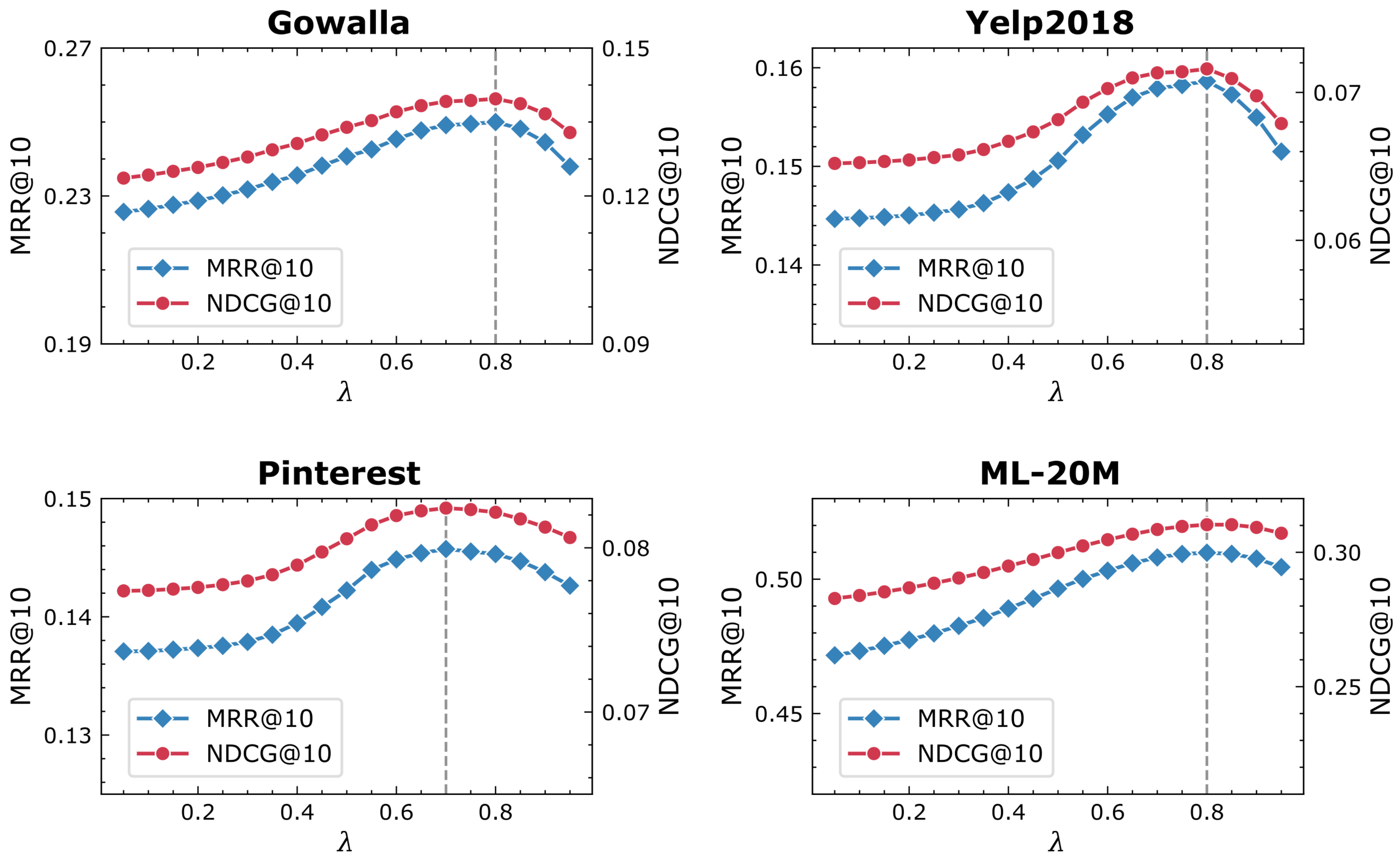


Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. Proceedings of the National Academy of Sciences 107, 10 (2010), 4511–4515.

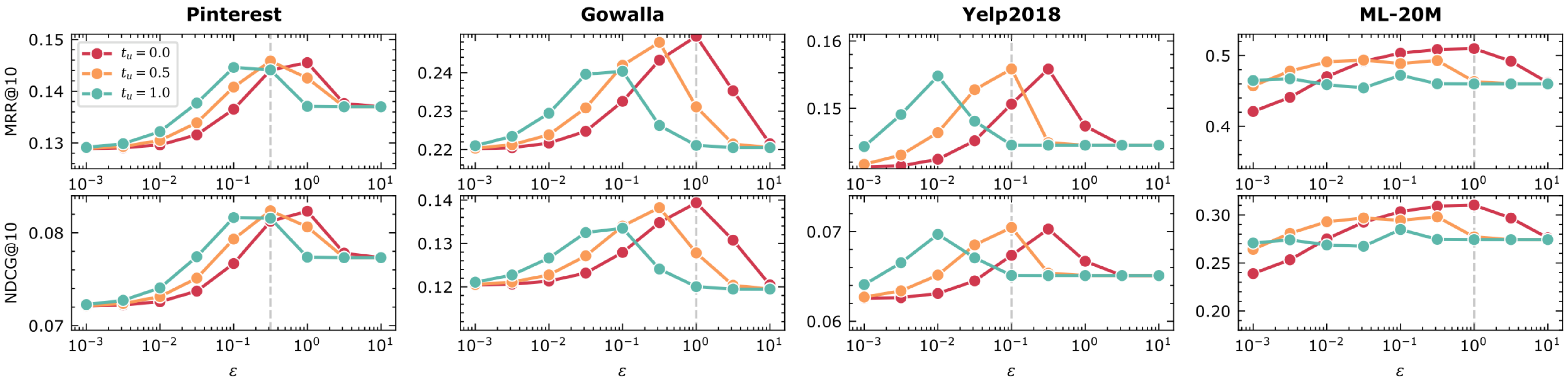
Detailed Analysis



Training Time



Effect of λ



Effect of scaling factors

Thank you!

The code is available at GitHub: *Joinn99/CoRML*

Paper

Presenter:

Tianjun Wei

City University of Hong Kong

tjwei2-c@my.cityu.edu.hk

Code