Bridging Jensen Gap for Max-Min Group Fairness Optimization in Recommendation

Chen Xu[1] , Yuxin Li[1], Wenjie Wang[2], Liang Pang[3], Jun Xu[1*], Tat–Seng Chua[4]

1 Gaoling School of Artificial Intelligence, Renmin University of China
2 School of Information Science and Technology, University of Science and Technology of China
3 Institute of Computing Technology, Chinese Academy of Sciences
4 NExT++ Research Center, National University of Singapore

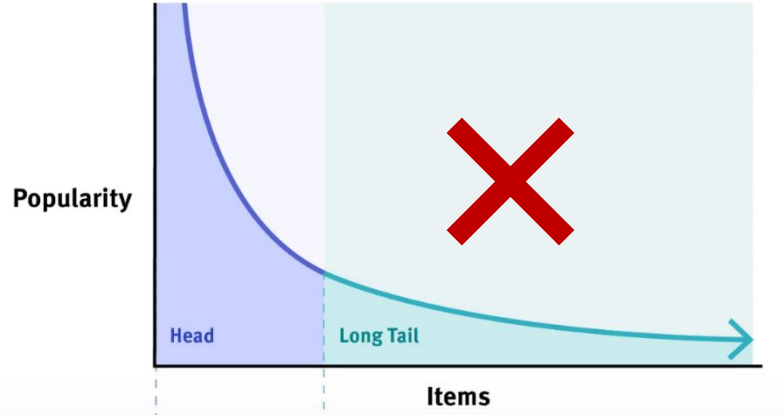
OUTLINE

- Motivation
- Method: FairDual
- Experiments
- Conclusion

Motivation

- Recommendation optimization objective:

$$\mathcal{L} = \min_{\hat{c}_{u,i}} \underbrace{-\sum_{u\in\mathcal{U}}\sum_{i\in\mathcal{I}} c_{u,i}\log(\hat{c}_{u,i})}_{\text{recommendation accuracy loss}}$$



Max-min fairness: Ensure that no group performs exceptionally poorly

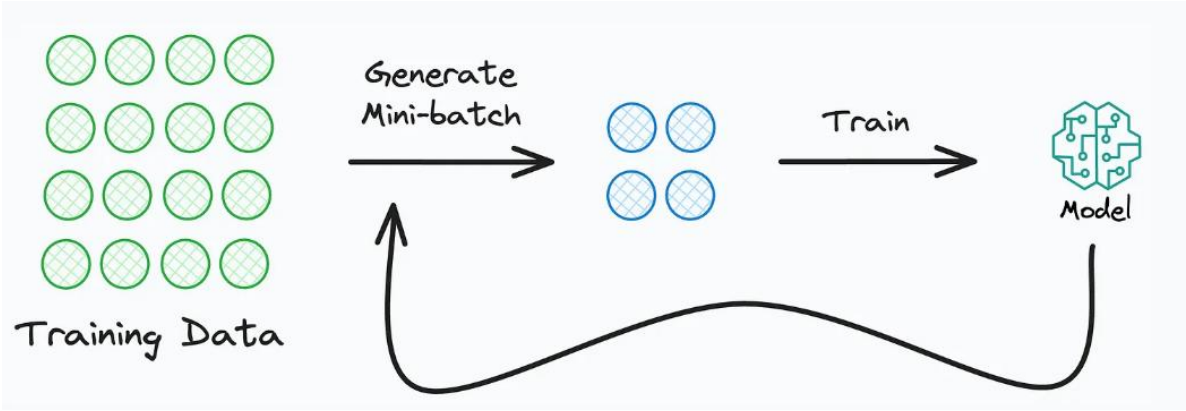
$$\text{s.t.} \quad \underbrace{\max_{g\in\mathcal{G}} \sum_{u\in\mathcal{U}}\sum_{i\in L_K(u)} -\frac{\mathbb{I}(i\in\mathcal{I}_g)}{n_i m_g} c_{u,i}\log(\hat{c}_{u,i}) \le M}_{\text{MMF constraint: loss of worst-off group } g \text{ should at or smaller than } M}, \quad (1)$$

Motivation

- Optimizing the objective **Computational** → **Resource are limited** → Batch training

$$\mathcal{L} = \min_{\hat{c}_{u,i}} \underbrace{-\sum_{u\in\mathcal{U}}\sum_{i\in\mathcal{I}} c_{u,i}\log(\hat{c}_{u,i})}_{\text{recommendation accuracy loss}}$$

$$\text{s.t.} \quad \underbrace{\max_{g\in\mathcal{G}}\sum_{u\in\mathcal{U}}\sum_{i\in L_K(u)} -\frac{\mathbb{I}(i\in\mathcal{I}_g)}{n_i m_g} c_{u,i}\log(\hat{c}_{u,i}) \le M}_{\text{MMF constraint: loss of worst-off group } g \text{ should at or smaller than } M}, \quad (1)$$



No !! It is biased

Can such constrained optimization problem be unbiased when using the batch training?

Motivation

- The constraints make the objective become non-linear additivity
  - Break the independence of samples

$$\mathcal{L} = \min_{\hat{c}_{u,i}} \underbrace{-\sum_{u\in\mathcal{U}}\sum_{i\in\mathcal{I}} c_{u,i}\log(\hat{c}_{u,i})}_{\text{recommendation accuracy loss}}$$

$$\text{s.t.} \quad \underbrace{\max_{g\in\mathcal{G}} \sum_{u\in\mathcal{U}}\sum_{i\in L_K(u)} -\frac{\mathbb{I}(i\in\mathcal{I}_g)}{n_i m_g} c_{u,i}\log(\hat{c}_{u,i}) \le M}_{\text{MMF constraint: loss of worst-off group } g \text{ should at or smaller than } M}, \quad (1)$$

equivalence

$$\mathcal{L} = \min_{w\in\mathcal{W}} \boldsymbol{b}^\top(\hat{\boldsymbol{A}}^\top \boldsymbol{w})^{1+t}$$

$$J(B) = |\mathcal{L}^B - \mathcal{L}| = |\mathcal{L}^B - \min \boldsymbol{b}^\top f(\sum_{j=1}^{|\mathcal{U}|/B} \boldsymbol{e}_j)| \neq 0.$$

There is a bias (Jensen gap) when utilizing batch training!

Motivation

- Factors for influcing the Jensen gap: batch size and group size!

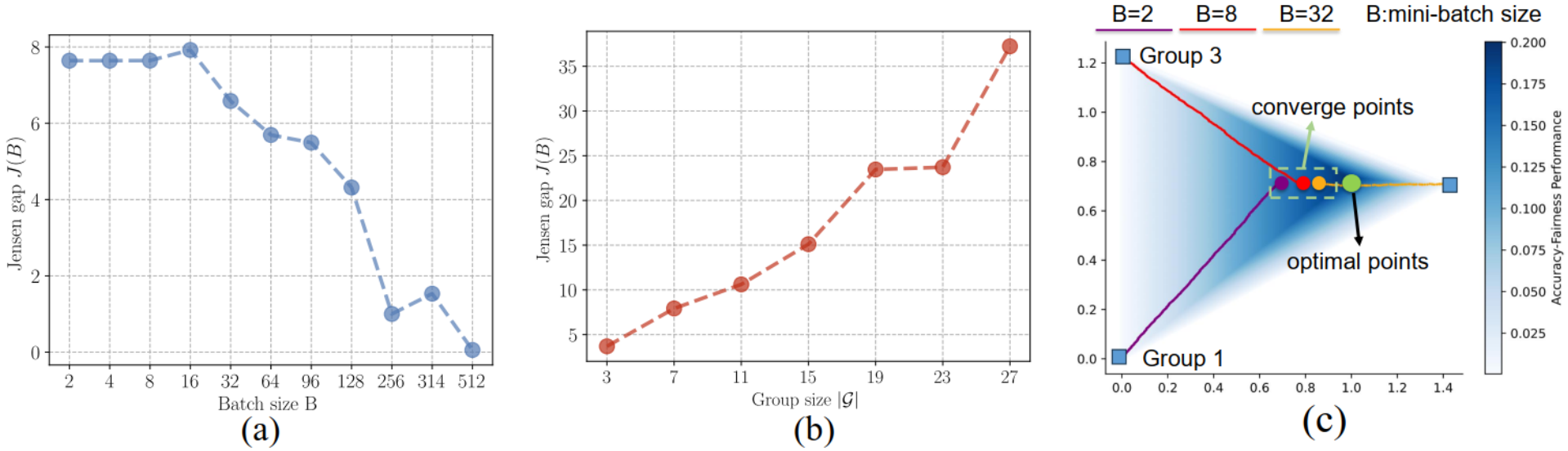


Figure 1: Loss converges simulation with 1000 users and 1000 items. Sub-figure (a) and (b) illustrate the distance away from the optimal point (*i.e.*, Jensen gap) *w.r.t.* mini-batch and group size, respectively. Figure (a) was conducted with the same group size (G=7) under different batch sizes, while Figure (b) was conducted with the same batch size (B=32) under different group sizes. Sub-figure (c) describes the converged trace under different batch sizes.

OUTLINE

- Motivation
- **Method: FairDual**
- Experiments
- Conclusion

FairDual

- Introducing a weighting term to mitigate such a bias

Theorem 3. *By introducing the dual variable $\boldsymbol{\mu}$, the dual form of the Equation (1) is*

$$\mathcal{L}' = \min_{\hat{c}_{u,i}} \quad -\sum_{u\in\mathcal{U}}\sum_{g\in\mathcal{G}} \boldsymbol{s}_g \sum_{i\in\mathcal{I}_g} c_{u,i}\log(\hat{c}_{u,i}), \tag{4}$$

where $\boldsymbol{s}_g = 1-\boldsymbol{\mu}_g$ *and* $\boldsymbol{\mu} = \arg\min_{\boldsymbol{\mu}\in\mathcal{M}} \left(\max \sum_{u\in\mathcal{U}}\sum_{g\in\mathcal{G}} \boldsymbol{s}_g \sum_{i\in\mathcal{I}_g} c_{u,i}\log(\hat{c}_{u,i}) + \lambda r^*(\boldsymbol{\mu})\right)$,
where $r^*(\mu) = \max_{\boldsymbol{w}_g\le m_g} \left(\min_{g\in\mathcal{G}} m_g(\hat{\boldsymbol{A}}\boldsymbol{w})_g + \hat{\boldsymbol{A}}^\top \boldsymbol{w}\boldsymbol{\mu}/\lambda\right) = \sum_g m_g\boldsymbol{\mu}_g/\lambda + 1$, $\mathcal{M} = \left\{\boldsymbol{\mu} \,\middle|\, \sum_{g\in\mathcal{S}} \boldsymbol{\mu}_g m_g \ge -\lambda, \forall \mathcal{S}\in\mathcal{G}_s\right\}$, *where* $\mathcal{G}_s$ *is the set of all subsets of* $\mathcal{G}$ (i.e., *power set*).

Weighting term, be updated each batch utilizing the dual mirror gradient descent!

FairDual

- Introducing a weighting term to mitigate such a bias

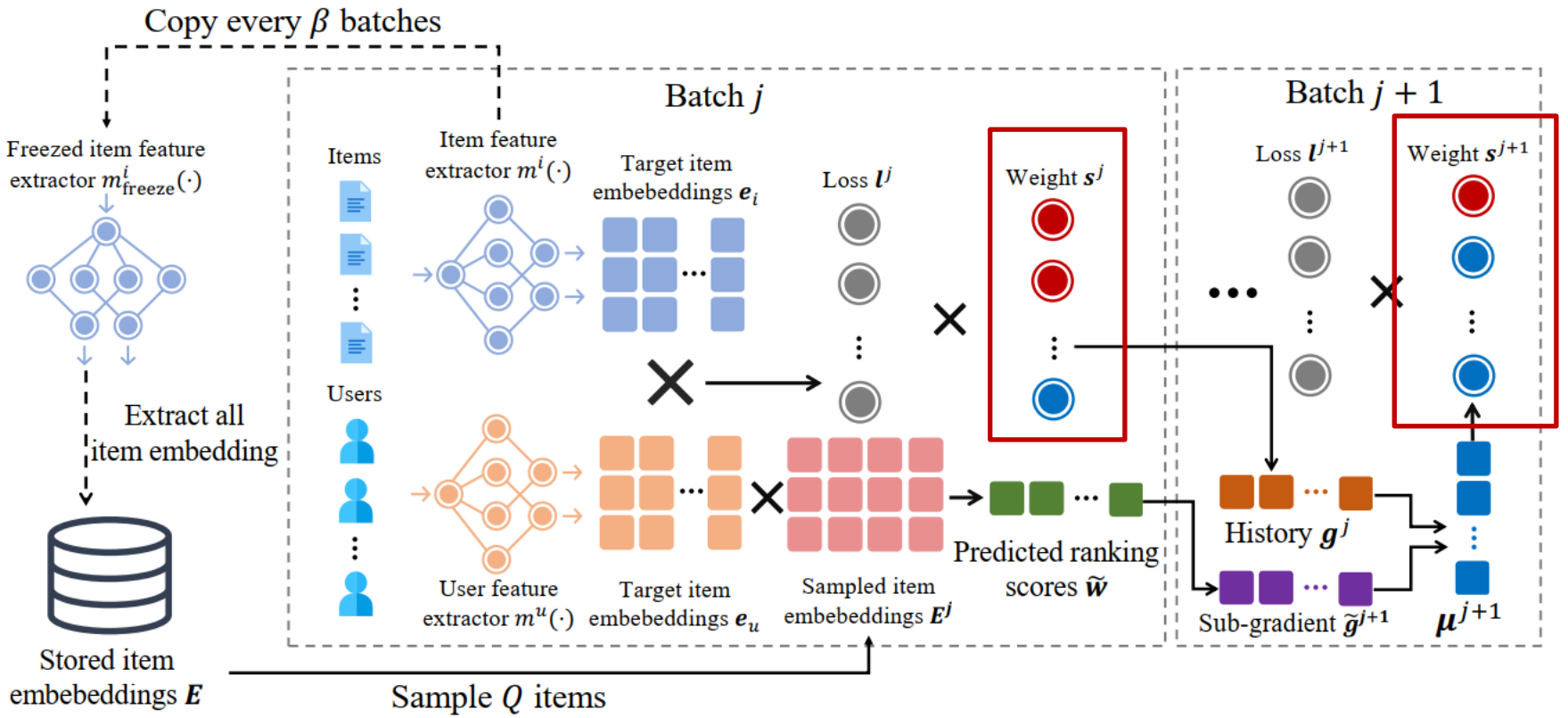


Figure 2: Overall workflow of FairDual under every two batches j and $j+1$.

FairDual

- Algorithm

Algorithm 1: FairDual

Require: Dataset $\mathcal{D} = \{u, i, c_{u,i}\}$, item-group adjacent matrix $\boldsymbol{A}$, dual learning rate η, trade-off coefficient λ, $m^i_{\text{freeze}}(\cdot)$ updating step β, batch size B and sample item number Q and the weight m_g for each group g. $\hat{\boldsymbol{A}} = \text{diag}(\boldsymbol{A}\boldsymbol{1})^{-1}\boldsymbol{A}$.

Ensure: The model parameters of $m^i(\cdot), m^u(\cdot)$.

for $n = 1, \cdots, N$ **do**
  Set $\boldsymbol{\gamma}_{1,g} = m_g, \forall g \in \mathcal{G}$
  for $j = 1, \cdots, |\mathcal{U}|/B$ **do**
    if $(n * |\mathcal{N}|/B + j)\%\beta = 0$ **then**
      Copy parameters from $m^i(\cdot)$ to $m^i_{\text{freeze}}(\cdot)$ and get all item embedding $\boldsymbol{E}$
      Initialize dual solution $\boldsymbol{\mu} = 0$, and momentum gradient $\boldsymbol{g} = 0$ and $t = 0$.
    end if
    Get sub-dataset $\{u, i, c_{u,i}\}_{b=1}^{B}$ and user feature $\boldsymbol{e}_u$ and item feature $\boldsymbol{e}_i$
    $\mathcal{L}^j = [-c_{u,i} \log(\hat{c}_{u,i})]_{b=1}^{B}$, $\boldsymbol{s}^j = \boldsymbol{1} - \hat{\boldsymbol{A}}^j\boldsymbol{\mu}$ → **Weighting term**
    Apply gradient descent based on loss $(\boldsymbol{s}^j)^\top \mathcal{L}^j$
    $\widetilde{\boldsymbol{w}}_b = \sum_{k=1}^{K} (\boldsymbol{e}_{u_b}^\top \boldsymbol{E}^b)_{[k]}, \forall b$
    $\widetilde{\boldsymbol{g}}^j = -(\hat{\boldsymbol{A}}^j)^\top \widetilde{\boldsymbol{w}} + \boldsymbol{\gamma}_j \quad , \boldsymbol{g}^j = \alpha \widetilde{\boldsymbol{g}}^j + (1-\alpha)\mathbf{g}, \quad \boldsymbol{g} = \boldsymbol{g}^j$
    $\boldsymbol{\gamma}_j = \boldsymbol{\gamma}_{j-1} - (\hat{\boldsymbol{A}}^j)^\top \widetilde{\boldsymbol{w}}, \boldsymbol{\mu} = \arg\min_{\boldsymbol{\mu}_0 \in \mathcal{M}} \left[(\boldsymbol{g}^j)^\top \boldsymbol{\mu}_0 + \eta \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}\|_2^2\right]$ → **Mirror SGD for learn the weight**
  end for
end for

FairDual

- The bias can be bounded

Theorem 4 (Bound on Jensen Gap). *There exists $H > 0$ such that $\|\boldsymbol{\mu}^j - \boldsymbol{\mu}\|_2^2 \leq H$ and function $\|\cdot\|_2^2$ is $\sigma-$strongly convex. Then, there exists $L > 0$, the Jensen gap of FairDual can be bounded as:*

$$J(B) \leq \frac{H}{\eta} + \frac{|\mathcal{U}|L|\mathcal{G}|^2}{B(1-\alpha)\sigma}\eta + \frac{L|\mathcal{G}|^2}{2(1-\alpha)^2\sigma\eta}. \tag{10}$$

When setting learning rate $\eta = O(B^{-1/2})$, the bound of Jensen gap is comparable with $O(B^{-1/2})$.

OUTLINE

- Motivation
- Method: FairDual
- **Experiments**
- Conclusion

Experiments

Outperform all the baselines in terms of accuracy and fairness!

Table 1: Performance comparisons between ours and the baselines on three datasets under best-performing BigRec backbones. The $*$ means the improvements are statistically significant (t-tests and p-value < 0.05). The bold number indicates that the accuracy value exceeds that of all the baselines.

Models/Metrics		$K=5$			$K=10$			$K=20$		
		NDCG (%)	MRR (%)	MMF (%)	NDCG (%)	MRR (%)	MMF (%)	NDCG (%)	MRR (%)	MMF (%)
MIND	UNI	1.02	0.79	1.63	1.50	0.98	2.33	2.16	1.16	2.94
	DRO	0.90	0.67	1.81	1.37	0.87	2.51	1.94	1.02	3.21
	Prop	1.11	0.88	1.97	1.62	1.09	2.53	2.14	1.23	3.05
	S-DRO	0.91	0.70	1.87	1.42	0.91	2.41	1.93	1.04	3.02
	IFairLRS	0.87	0.66	2.21	1.27	0.83	2.91	1.78	0.97	2.86
	Maxmin sample	0.98	0.75	2.25	1.49	0.96	1.71	2.19	1.15	3.13
	Ours	**1.15***	**0.88**	**2.82***	**1.69***	**1.11**	**2.99***	**2.28***	**1.27***	**3.39***
	improv.(%)	3.60	0.00	25.33	4.32	1.83	2.75	4.10	3.25	5.61
Book	UNI	2.99	2.79	8.44	3.19	2.87	8.32	3.44	2.94	8.15
	DRO	2.94	2.72	8.39	3.15	2.81	8.29	3.37	2.87	8.10
	Prop	2.64	2.45	8.68	2.83	2.53	8.30	3.05	2.59	8.01
	S-DRO	2.61	2.44	8.37	2.80	2.52	8.21	3.06	2.59	8.07
	IFairLRS	2.30	2.16	8.46	2.51	2.25	8.20	2.76	2.32	8.17
	Maxmin sample	2.49	2.31	6.80	2.72	2.43	6.80	2.97	2.74	7.50
	Ours	**3.11***	**2.88**	**8.90***	**3.31***	**2.96**	**9.00***	**3.60***	**3.04**	**8.89***
	improv.(%)	4.01	3.23	2.53	3.76	3.14	8.17	4.65	3.40	8.81

Experiments

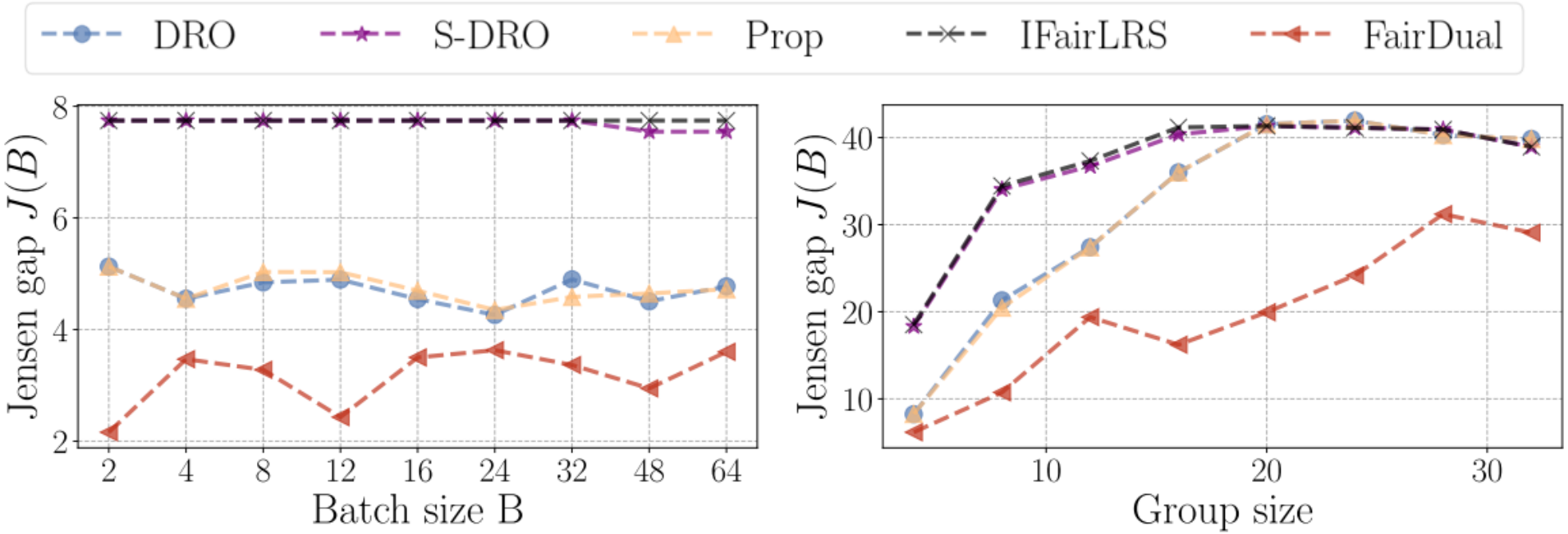
Outperform all the backbones in terms of accuracy and fairness!

Table 2: Performance comparisons between ours under other backbones on MIND dataset. The $*$ means the improvements are statistically significant (t-tests and p-value < 0.05). The bold number indicates that the accuracy value exceeds that of all the baselines.

Models/Metrics		top-5			top-10			top-20		
		NDCG (%)	MRR (%)	MMF (%)	NDCG (%)	MRR (%)	MMF (%)	NDCG (%)	MRR (%)	MMF (%)
NRMS	DRO	0.44	0.32	0.12	0.66	0.42	3.60	1.06	0.50	9.94
	Prop	0.44	0.32	0.12	0.66	0.42	3.49	1.06	0.52	9.94
	S-DRO	0.52	0.34	0.10	0.76	0.40	2.05	1.20	0.52	8.74
	IFairLRS	0.40	0.28	0.69	0.62	0.36	4.20	0.96	0.44	10.58
	Maxmin sample	0.38	0.31	0.20	0.45	0.34	4.00	0.67	0.422	9.99
	Ours	**0.60***	**0.40***	**1.07***	**0.84***	**0.46***	**4.93***	**1.28***	**0.60***	**11.35***
RecFormer	DRO	0.57	0.45	1.08	0.89	0.59	1.08	1.41	0.73	1.52
	Prop	0.57	0.45	1.08	0.89	0.58	1.08	1.41	0.72	1.52
	S-DRO	0.57	0.45	1.20	0.91	0.60	1.15	1.46	0.73	1.62
	IFairLRS	0.46	0.37	1.68	0.76	0.49	1.70	1.29	0.63	2.12
	Maxmin sample	0.51	0.41	0.94	0.85	0.55	1.50	1.37	0.69	2.48
	Ours	**0.59***	**0.45**	**1.88***	**0.99***	**0.60**	**1.94***	**1.55***	**0.75**	**2.58***

Experiments

Smaller Jansen gap compared to other models!



(a) Jensen gap comparision

OUTLINE

- Motivation
- Method: FairDual
- Experiments
- **Conclusion**

Conclusion

- Max-min group fairness contraints will introduce the Jensen gap during the objective optimization in recommendation
- We show batch size and group size are two key factors for the Jensen gap
- We propose a model FairDual, which can efficiently and effectively mitigate the Jensen using the mirror SGD method
- Pay attention to similar bias when apply the contrained optimization!

Thanks!

[paper]

[codes]