On Inherited Popularity Bias in Cold-Start Item Recommendation (Supplementary Material)

Gregor Meehan
gregor.meehan@qmul.ac.uk
Queen Mary University of London
London, United Kingdom

Johan Pauwels
j.pauwels@qmul.ac.uk
Queen Mary University of London
London, United Kingdom

1 Inherited Popularity Bias

In Figure 1 we visualize top 20 prediction counts against holdout set counts for the items in the Clothing and Microlens datasets (i.e. the equivalent of Figure 1 in the main paper). Since these datasets have fewer items than Electronics, their outlier behavior is less extreme, but there is still evidence of the same problematic behaviors. In the warm predictions, a small number of items have much larger prediction counts than the rest of the population, and the cold models mirror this pattern. However, the overexposed cold items are often not popular, meaning that these biased behaviors are lowering recommendation quality as well as item fairness. We see this concretely for Clothing and Microlens in the main results in the paper, where our mitigation method actually improves accuracy alongside item fairness in some cases.

2 Statistical Evidence for Proximity to Popular Neighbors

In Section 4.1 of the main paper, we claim that there is a directional dependence between a cold item being overexposed and it having a popular warm item as a close neighbor by content similarity. To support this claim statistically, we conduct one-sided Fisher's exact tests on the proportion of top 1% of cold items by prediction count which have close neighbors in the top 1% of warm items by popularity. To illustrate this process, in Table 1 we include the corresponding contingency tables for Figure 2 in the main paper. We apply this test to all model and dataset combinations, and find that they are significant with $p \ll 1e^{-10}$ and odds ratios ranging from 16.23 up to 61.11. This provides statistical evidence for our claim that the models learn to associate feature patterns of certain items with high exposure; future work may explore why this occurs only in some cases.

3 Impact of Scaling and α

In this section we provide further insight into the effect of our magnitude scaling procedure and the hyperparameter α on model predictive behavior.

In Figure 2, we plot the impact of α on user accuracy and low-end item accuracy. In Clothing and Microlens, both user and item accuracy increase for smaller values of α, although this levels off as α gets larger, and an inverse relationship between the two metrics starts to appear. Nonetheless, by careful selection of α we are able to achieve a material increase in the performance of under-served items for these datasets while preserving recommendation quality from the user perspective. In the Electronics dataset, the inverse relationship between NDCG and MDG-Min80% is stronger, and it seems that the much larger item population makes it significantly more difficult to promote correct items into the top 20. There is therefore a larger compromise between user and item-level performance, and the appropriate α value may depend on the application and other fairness considerations.

In Figure 3 and Figure 4 we include further insight into the impact of our scaling. Figure 3 shows how the cold item prediction counts in Figure 1 of the main paper are impacted after scaling. Figure 4 contains equivalent plots to Figure 4 in the main paper for Clothing and Microlens. We can see that there is a similar balancing effect to model predictions in these datasets.

Table 1: Contingency tables for GAR and GoRec on the Electronics dataset (cf. Figure 2 in main paper). The 'Pop. Close Neighbor' class refers to cold items with a warm item neighbor which is in the top 1% of warm items by popularity. The 'Most Predicted' class refers to cold items which are in the top 1% by prediction count. The odds ratio for the corresponding Fisher exact tests are 40.04 (GAR) and 35.80 (GoRec) with $p \ll 1e^{-10}$ in both cases.

	Pop. Close Neighbor	~(Pop. Close Neighbor)
Most Predicted	105	21
~(Most Predicted)	1,385	11,090

(a) GAR

	Pop. Close Neighbor	~(Pop. Close Neighbor)
Most Predicted	103	23
~(Most Predicted)	1,387	11,088

(b) GoRec

RecSys '25, Prague, Czech Republic

ACM ISBN 979-8-4007-1364-4/2025/09
https://doi.org/XXXXXXX.XXXXXXX

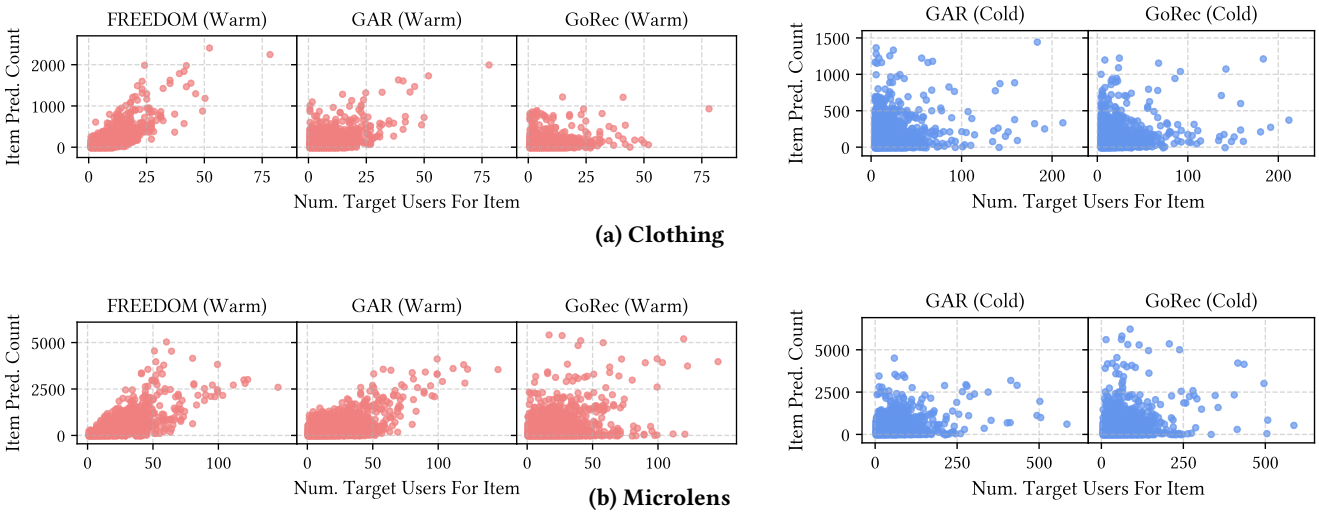


Figure 1: Warm and cold item prediction counts with $k = 20$ against the number of target users (i.e. the number of times an item appears in the validation or test set interactions) for Clothing (top) and Microlens (bottom). Each dot represents an item.

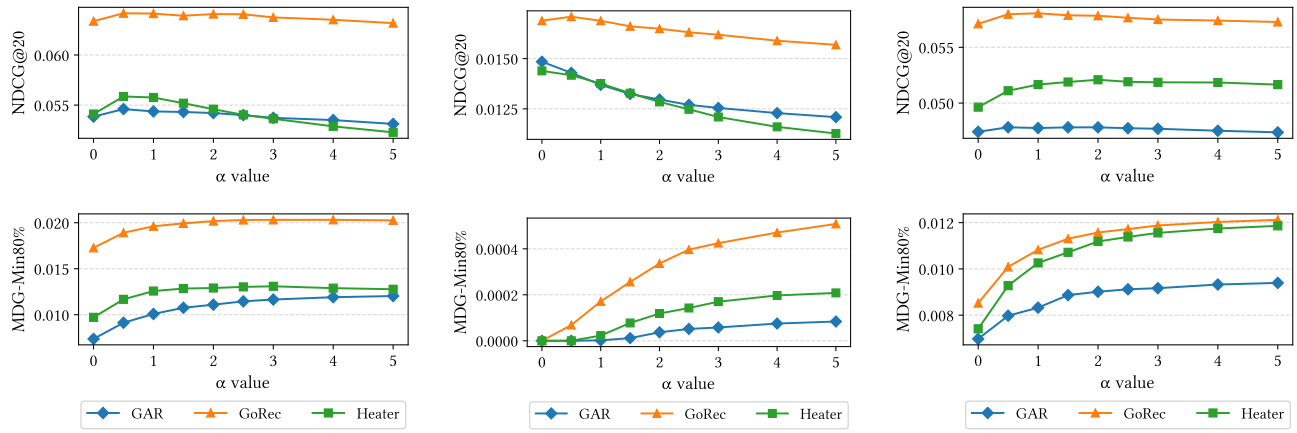


Figure 2: Impact of α on cold test set NDCG and MDG-Min80% with $k = 20$.

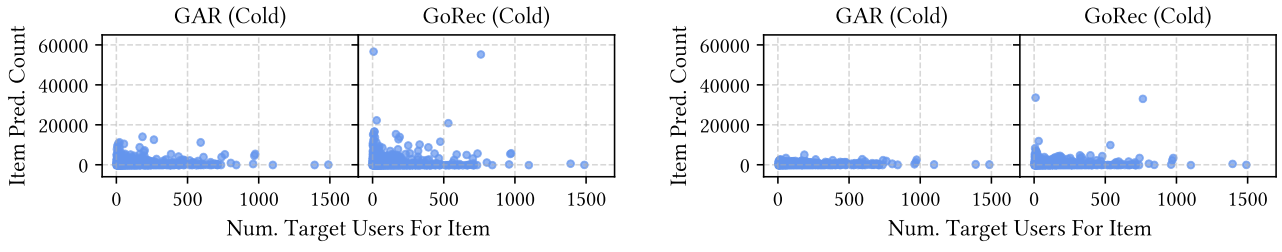
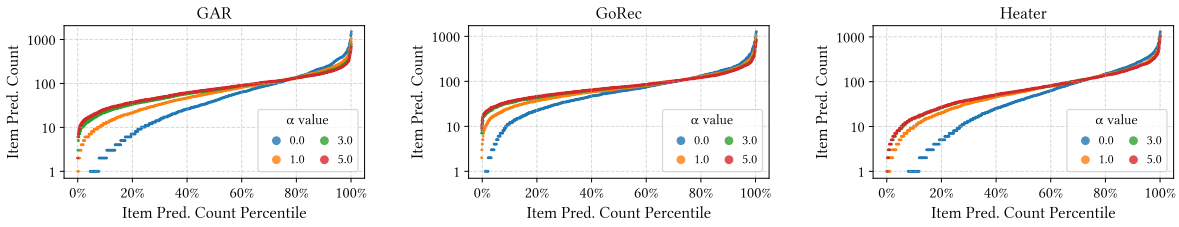
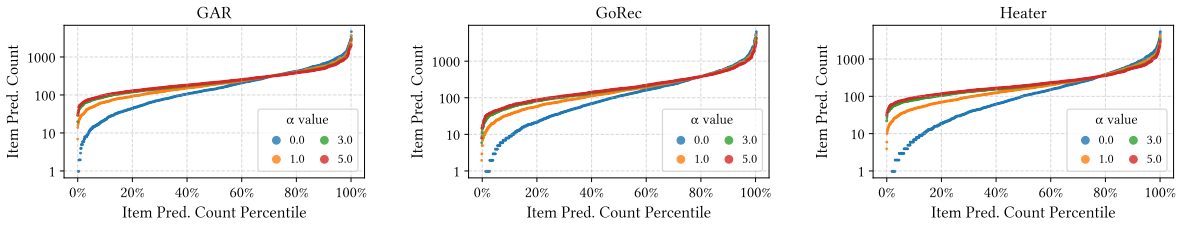


Figure 3: Cold item prediction counts with $k = 20$ against the number of target users for the Electronics dataset before and after magnitude scaling.



(a) Clothing



(b) Microlens

Figure 4: Cold test set item prediction counts at $k = 20$ against item prediction count percentiles (i.e. each item's position in the sorted list of prediction counts) for Clothing and Microlens. Only items predicted at least once are plotted.