Probabilistic Modeling, Learnability and Uncertainty Estimation for Interaction Prediction in Movie Rating Datasets: Supplementary Material

Anonymous Author(s)

A Proof of Mathematical Results

In this appendix, we show that low-rank PMFs are learnable in the sense of L^1 norm and further show that this implies an analogue of an excess risk bound in our implicit feedback context: there exists an algorithm which, consuming fewer than $\widetilde{O}((m+n)r/\epsilon^2)$ samples, picks a low rank distribution whose expected recall at k is guaranteed to be within ϵ of the best possible recall at k achievable.

A.1 Relating the L^1 Loss to the Recall at k

Let $p \in \mathbb{R}^n$ (resp. $\widehat{p} \in \mathbb{R}^n$) be a distribution over $[n]$ (which, as in standard notation, stands for $\{1, 2, \ldots, n\}$). We write $p_{[i]}$ (resp. $\widehat{p}_{[i]}$) for the ith element of p (resp $\widehat{p}$) when written in decreasing order. We also write σ (resp. $\widehat{\sigma}$) for the permutation of $[n]$ such that $p_{[i]} = p_{\sigma(i)}$ (resp.$\widehat{p}_{[i]} = \widehat{p}_{\widehat{\sigma}(i)}$).

If we draw a test (multi) set $\Omega' = \{y_1, \ldots, y_{N'}\} \subset [n]$ consisting of N' i.i.d. samples from p, the *Recall@k* of a scoring function p or $\widehat{p}$ is defined as the number of test samples belong to the top k items as determined by the scoring function p or $\widehat{p}$:

$$R^k_{N'} := \frac{1}{N}\sum_{o=1}^{N'} 1_{y_o \in \sigma^{-1}([k])}, \tag{1}$$

$$\widehat{R}^k_{N'} := \frac{1}{N}\sum_{o=1}^{N'} 1_{y_o \in \widehat{\sigma}^{-1}([k])}. \tag{2}$$

This is a random variable. Note that by the i.i.d. assumption its expectation doesn't depend on N' and is calculated as follows:

$$\mathbb{E}(R^k_{N'}) = \mathbb{E}(R^k_1) = \sum_{i \in \sigma^{-1}([k])} p_i, \tag{3}$$

$$\mathbb{E}(\widehat{R}^k_{N'}) = \mathbb{E}(\widehat{R}^k_1) = \sum_{i \in \widehat{\sigma}^{-1}([k])} p_i. \tag{4}$$

By abuse of notation, we write $\mathbb{E}(R^k)$ for $\mathbb{E}(R^k_1)$ and $\widehat{\mathbb{E}}(\widehat{R}^k)$ for $\widehat{\mathbb{E}}(\widehat{R}^k_1)$.

The quantity $\mathbb{E}(R^k)$ represents the **best possible expected recall**, and is analogous to the Bayes Error in classic Learning Theory. $\mathbb{E}(\widehat{R}^k_{N'})$ is the true expected recall of the trained model $\widehat{p}$, thus, the quantity $\mathbb{E}(\widehat{R}^k_{N'}) - \mathbb{E}(R^k)$ is analogous to the excess risk in learning theory.

We also define the (empirical) estimated recall at k as follows:

$$\widehat{\mathbb{E}}(\widehat{R}^k_1) = \sum_{i \in \widehat{\sigma}^{-1}([k])} \widehat{p}_i = \sum_{i \le k} \widehat{p}_{[i]}. \tag{5}$$

We will now prove the following:

Proposition A.1. *If $\left\|p - \widehat{p}\right\|_1 \le \epsilon$ for some $\epsilon > 0$, then we have:*

$$\mathbb{E}(R^k) - \epsilon \le \widehat{\mathbb{E}}(\widehat{R}^k) \le \mathbb{E}(R^k) + \epsilon. \tag{6}$$

In particular, since we certainly have $\mathbb{E}(\widehat{R}^k) \le \widehat{\mathbb{E}}(\widehat{R}^k) + \epsilon$, we also have the following bound on the excess risk:

$$\mathbb{E}(\widehat{R}^k) - \mathbb{E}(R^k) \le 2\epsilon. \tag{7}$$

Proof. We can rewrite the quantity $\mathbb{E}(R^k) = \sum_{i \le k} p_{[i]}$ as $\max_{\substack{|S|=k \\ S \subset [n]}} \sum_{i \in S} p_i$ (and similarly for $\widehat{\mathbb{E}}(\widehat{R}^k)$).

Thus, we have

$$\begin{aligned}
\widehat{\mathbb{E}}(\widehat{R}^k) &= \sum_{i \le k} \widehat{p}_{[i]} = \max_{\substack{|S|=k \\ S \subset [n]}} \sum_{i \in S} \widehat{p}_i \\
&\le \max_{\substack{|S|=k \\ S \subset [n]}} \sum_{i \in S} p_i + \epsilon \\
&= \sum_{i \le k} p_{[i]} + \epsilon = \mathbb{E}(R^k) + \epsilon, \qquad (8)
\end{aligned}$$

where at the second line (8) we have used the condition $\left\|p - \widehat{p}\right\|_1$.

Similarly, we also have

$$\widehat{\mathbb{E}}(\widehat{R}^k) = \sum_{i \le k} \widehat{p}_{[i]} = \max_{\substack{|S|=k \\ S \subset [n]}} \sum_{i \in S} \widehat{p}_i \tag{9}$$

$$\ge \max_{\substack{|S|=k \\ S \subset [n]}} \sum_{i \in S} p_i - \epsilon \tag{10}$$

$$= \sum_{i \le k} p_{[i]} - \epsilon = \mathbb{E}(R^k) - \epsilon, \tag{11}$$

where at the second line (10) we have used the condition $\left\|p - \widehat{p}\right\|_1$. The result follows. □

We now consider the recommender systems setting, where the recall is defined user-wise and averaged over the users. In this case, we have a ground truth distribution $P \in \mathbb{R}^{m \times n}$ and its estimated version $\widehat{P} \in \mathbb{R}^{m \times n}$. We fix k and define the recall of user i as $R^{k,i}$ via formula (1) where $p \leftarrow p_{i,\cdot} \in \mathbb{R}^n$ is now the normalized version of the ith row of P (i.e. $p_{i,j} = P_{i,j}/p_i$ with $p_i := \sum_{j \in [n]} P_{i,j}$. We can similarly define the quantities relative to $\widehat{P}$. We define the aggregated recall w.r.t. the ranking provided by $\widehat{P}$ (resp. P) by $\widehat{R}^{k,\mathrm{all}} := \frac{1}{m}\sum_{i\leq m}\widehat{R}^{k,i}$ (resp. $R^{k,\mathrm{all}} := \frac{1}{m}\sum_{i\leq m}R^{k,i}$).

Proposition A.2. *Assume that $p_i = \frac{1}{m}$ for all i, and $\left\|\widehat{P} - P\right\|_1 \leq \epsilon$. Then we have the following excess risk bound:*

$$\mathbb{E}(\widehat{R}^{k,all}) \leq \mathbb{E}(R^{k,all}) + 2\epsilon. \tag{12}$$

Proof. Let ϵ_i be defined as $\|p_{i,\cdot} - \widehat{p}_{i,\cdot}\|_1$. Then by Proposition A.1 we certainly have

$$\mathbb{E}(\widehat{R}^{k,\mathrm{all}}) = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\widehat{R}^{k,i}\right] \tag{13}$$

$$= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\widehat{R}^{k,i} \leq \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}R^{k,i} + 2\epsilon_i \tag{14}$$

$$\leq \mathbb{E}(R^{k,\mathrm{all}}) + 2\frac{1}{m}\sum_{i=1}^{m}\epsilon_i \tag{15}$$

$$= \mathbb{E}(R^{k,\mathrm{all}}) + 2\frac{1}{m}\sum_{i=1}^{m}\left\|p_{i,\cdot} - \widehat{p}_{i,\cdot}\right\| \tag{16}$$

$$\leq \mathbb{E}(R^{k,\mathrm{all}}) + 2\frac{1}{m}\sum_{i=1}^{m}\left\|mP_{i,\cdot} - m\widehat{P}_{i,\cdot}\right\| \tag{17}$$

$$= \mathbb{E}(R^{k,\mathrm{all}}) + 2\left\|P - \widehat{P}\right\|_1 \leq \mathbb{E}(R^{k,\mathrm{all}}) + 2\epsilon, \tag{18}$$

where at equation (14) we have used Proposition A.1 and at equation (17) we have used the assumption that $p_i = \frac{1}{m}$ for all i. The result follows.

□

A.2 Bounding the L1 Loss

We have established above that if we can control the $L1$ loss, then we can control the excess risk defined in terms of Recall@k for any k. To control the L^1 loss, we use the following result from [2] which ultimately follows from results on Sheffé tournaments in non-parametric density estimation [1].

Proposition A.3 (Adaptation of Proposition 2.1 in [2])**.** *Let $\mathcal{H}_{m\times n,r}$ denote the set of non-negative rank r distributions over $[m] \times [n]$. Let $p \in \mathcal{H}_{m\times n,r}$ be a probability distribution from which we observe N i.i.d samples. There exists an estimator $\widehat{p} \in \mathcal{H}_{m\times n,r}$ (depending on the N samples) such that for any $\delta > 0$, the following holds with probability $\geq 1 - \delta$:*

$$\left\|p - \widehat{p}\right\|_1 \leq 7\sqrt{\frac{(m+n)r\log(2(m+n)rN)}{N}} + 7\sqrt{\frac{\log(\frac{3}{\delta})}{2N}}.$$

By combining Proposition A.3 with Proposition A.2, we obtain the following:

Theorem A.4. *Let $\mathcal{H}_{m\times n,r}$ denote the set of non-negative rank r distributions over $[m] \times [n]$. Let $p \in \mathcal{H}_{m\times n,r}$ be a probability distribution from which we observe N samples. There exists an estimator $\widehat{p} \in \mathcal{H}_{m\times n,r}$ (depending on the N samples) such that for any $\delta > 0$, the following excess risk bound for the recall at k holds with probability $\geq 1 - \delta$: $\mathbb{E}(\widehat{R}^{k,all}) - \mathbb{E}(R^{k,all}) \leq$*

$$14\sqrt{\frac{(m+n)r\log(2(m+n)rN)}{N}} + 14\sqrt{\frac{\log(\frac{3}{\delta})}{2N}}. \tag{19}$$

References

[1] L. Devroye and G. Lugosi. 2001. *Combinatorial Methods in Density Estimation*. Springer, New York.

[2] Robert A Vandermeulen and Antoine Ledent. 2021. Beyond smoothness: Incorporating low-rank analysis into nonparametric density estimation. *Advances in Neural Information Processing Systems* 34 (2021), 12180–12193.