Appendix for RankDistil: Knowledge Distillation for Ranking

A Coupled RankDistil Algorithm

We describe the full algorithm for coupled RankDistil in this section. As mentioned earlier, computing the gradient of coupled RankDistil-Loss in Equation 3 can be expensive since it requires iterating over all the permutations. Here, we use a Monte-Carlo approximation to obtain an unbiased estimate of the gradient. Algorithm 2 lists the pseudo-code for the algorithm. Note that the algorithm is with Monte-Carlo approximation on single permutation, however, it practice we use a batch of permutations.

Algorithm 2 Coupled RankDistil

```
Initialization: Initial predictor $f$, Teacher predictor $f^{\mathrm{t}}$, distribution $Q(.|t)$, loss function $\ell_{\textsc{RankDistil}}$, integer $p \geq k$, batch size $m$, mined batch size $b \leq m$, thresholding function $\tau$
for $r = 0, \cdots, R-1$ do
    Uniformly randomly select an example $\{x, y\}$
    Sample index set $B$ of size $m$ using the distribution $Q(.|f^{\mathrm{t}}(x))$
    Compute $P = \textsc{Top}_p(f^{\mathrm{t}}(x))$ and $N = \textsc{Top}_{b,B}(f(x))$
    Compute $s_l = f_l(x)$ for $l \in P \cup N$
    Randomly sample a permutation $\pi$ according to $r$-Plackett's model $\mathbb{P}_{\tau(\alpha t, P)}(.|P \cup N)$
    Compute gradient $g_r = -\nabla \log \mathbb{P}_s(\pi|P \cup N)$ (according to Definition 4)
    Update predictor $f$ using the gradient $g_r$
end for
```

A.1 Proof of Proposition 5

Recall that

$$\mathbb{P}_s(\pi|S) = \frac{1}{(|S|-r)!} \prod_{j=1}^{r} \frac{\exp(s_{\pi(j)})}{\sum_{l=j}^{|S|} \exp(s_{\pi(l)})}.$$

To prove the first part of the result, we observe the following:

$$\sum_{\pi \in \mathcal{P}(S)} \mathbb{P}_s(\pi|S) = \sum_{\pi \in \mathcal{P}(S)} \frac{1}{(|S|-r)!} \prod_{j=1}^{r} \frac{\exp(s_{\pi(j)})}{\sum_{l=j}^{|S|} \exp(s_{\pi(l)})}.$$

Let $\mathcal{G}_l$ be the set of permutations of l elements selected from S. Then, it is easy to see that

$$\sum_{\pi \in \mathcal{P}(S)} \mathbb{P}_s(\pi|S) = \sum_{\pi \in \mathcal{P}(S)} \frac{1}{(|S|-r)!} \prod_{j=1}^{r} \frac{\exp(s_{\pi(j)})}{\sum_{l=j}^{|S|} \exp(s_{\pi(l)})} = \sum_{\pi \in \mathcal{G}_r} \prod_{j=1}^{r} \frac{\exp(s_{\pi(j)})}{\sum_{l=j}^{|S|} \exp(s_{\pi(l)})}$$

This can be shown by using the principle of mathematical induction on size of $\mathcal{G}_r$. It is easy to see that it holds for all subsets of size 1. Suppose it holds for all subsets of size $r-1$. For any permutation π, we use $\{\pi\}$ to denote the set of all items in π. We consider the following:

$$\begin{aligned}\sum_{\pi \in \mathcal{G}_r} \prod_{j=1}^{r} \frac{\exp(s_{\pi(j)})}{\sum_{l=j}^{|S|} \exp(s_{\pi(l)})} &= \sum_{\pi \in \mathcal{G}_{r-1}} \prod_{j=1}^{r-1} \frac{\exp(s_{\pi(j)})}{\sum_{l=j}^{|S|} \exp(s_{\pi(l)})} \sum_{i \in S - \{\pi\}} \frac{\exp(s_i)}{\sum_{l \in S - \{\pi\}}^{|S|} \exp(s_{\pi(l)})} \\ &= \sum_{\pi \in \mathcal{G}_{r-1}} \prod_{j=1}^{r-1} \frac{\exp(s_{\pi(j)})}{\sum_{l=j}^{|S|} \exp(s_{\pi(l)})} = 1.\end{aligned}$$

The second equality follows from the induction hypothesis. Therefore, it also holds for all size $r \leq K$. The base case for $r = 1$ is easy to verify. The result follows by principle of mathematical induction.

A.2 Computation of Coupled RankDistil-Loss

For the purpose of the discussion, let us assume $M = \infty$. Coupled RankDistil-Loss can be computed in $O(p^r r)$. Recall that coupled RankDistil-Loss is

$$\ell_{\text{RankDistil}}(t, s, P, N) = -\sum_{\pi \in \mathcal{P}(P \cup N)} \mathbb{P}_{\tau(\alpha t, P)}(\pi | P \cup N) \log \mathbb{P}_s(\pi | P \cup N).$$

First note that we only have to consider permutations where items in N appear in the last $|N|$ positions. This is due to the fact that $\mathbb{P}_{\tau(\alpha t,P)}(\pi | P \cup N) = 0$ when N appear in the last $|N|$ positions due to the thresholding function τ. Furthermore, only the first r positions of the permutation are important. In particular,

$$\mathbb{P}_{\tau(\alpha t,P)}(\pi_1 | P \cup N) \log \mathbb{P}_s(\pi_1 | P \cup N) = \mathbb{P}_{\tau(\alpha t,P)}(\pi_2 | P \cup N) \log \mathbb{P}_s(\pi_2 | P \cup N)$$

when first r positions of the π_1 and π_2 are the same. The number of such unique permutations is $O(p^r)$. The time complexity for computing the loss for a permutation can be reduced to $O(r)$ if $\sum_{i \in N} e^{s_i}$ is computed beforehand. This can be obtained by computing $\sum_{i \in P \cup N} e^{s_i}$ once and reusing them during the loss computation.

B Example Instantiations of RankDistil

We list a few interesting instantiations of RankDistil-Loss in this section.

Binary RankDistil-Loss

- Sigmoid CE + logistic

$$\ell_{\text{RankDistil}}(t, s, P, N) = \sum_{i \in P} \sum_{z \in \{-1,1\}} \left\{ -\frac{1}{1 + e^{-z t_i}} \log \frac{1}{1 + e^{-z s_i}} \right\} + \sum_{i \in N} \log(1 + e^{s_i})$$

- Softmax CE + Logistic

$$\ell_{\text{RankDistil}}(t, s, P, N) = \sum_{i \in P} \left\{ -\frac{e^{t_i}}{\sum_{i \in P} e^{t_i}} \log \frac{e^{s_i}}{\sum_{i \in P} e^{s_i}} \right\} + \sum_{i \in N} \log(1 + e^{s_i})$$

- 2-Regression + Square Hinge

$$\ell_{\text{RankDistil}}(t, s, P, N) = \|t_P - s_P\|^2 + \sum_{i \in N} \max\{0, \gamma + s_i\}^2$$

Pairwise RankDistil-Loss

- Softmax CE + Pairwise Hinge

$$\ell_{\text{RankDistil}}(t, s, P, N) = \sum_{i \in P} \left\{ -\frac{e^{t_i}}{\sum_{i \in P} e^{t_i}} \log \frac{e^{s_i}}{\sum_{i \in P} e^{s_i}} \right\} + \sum_{i \in N} \sum_{j \in P} \max\{0, \gamma + s_i - s_j\}$$

- Pairwise Hinge + Pairwise Hinge

$$\ell_{\text{RankDistil}}(t, s, P, N) = \sum_{i \neq j \in P} \sum_{z \in \{-1,1\}} \mathbb{1}(z(t_i - t_j)) \max\{0, \gamma + z(s_j - s_i)\} + \sum_{i \in N} \sum_{j \in P} \max\{0, \gamma + s_i - s_j\}$$

- Pairwise Logistic + Pairwise Logistic

$$\ell_{\text{RankDistil}}(t, s, P, N) = \sum_{i \neq j \in P} \sum_{z \in \{-1,1\}} \mathbb{1}(z(t_i - t_j)) \log(1 + e^{z(s_j - s_i)}) + \sum_{i \in N} \sum_{j \in P} \log(1 + e^{s_i - s_j})$$

C Consistency results

C.1 Proof of Claim 3

The proof of Claim 3 follows easily from Definition 2. First note that $[s^*(x)]_{[p+1]} < [s^*(x)]_{[p]}$. Furthermore, $p \geq k$ and $\text{argmax}_i s^*(x) = \text{argmax}_i f^{\text{t}}(x)$ for all $i \in [p]$. Thus, it follows that the items in top-k of $s^*(x)$ and $f^{\text{t}}(x)$ are exactly the same (including the order according to the respective scores). Since f^{t} is k-compatible, the result follows.

C.2 Proof of Theorem 6

We provide the proof for the case where $b = m$ in Algorithm 2. The proof for $b < m$ is similar. First, recall that RankDistil-Loss is

$$\ell_d(t, s) = \mathbb{E}_{B \sim Q(.|t)} \Big[\ell_{\text{RankDistil}}(t, s, \text{Top}_p(t), \text{Top}_{b,B}(s))\Big],$$

where

$$\ell_{\text{RankDistil}}(t, s, P, N) = \mathbb{E}_{\pi \sim \mathbb{P}_{\tau(\alpha t, P)}(.|P \cup N)} \left[-\log \mathbb{P}_s(\pi | P \cup N)\right]$$

Note that P is always $\text{Top}_p(t)$. We choose M such that $[f^{\text{t}}(x)]_i \geq -M$ for all $i \in [K]$. Let us analyze a minimizer $\bar{s} \in \text{argmin}_s \ell_{\text{RankDistil}}(t, s, P, N)$ for a particular a $N \subseteq [K] - \text{Top}_p(t)$. We claim that $\bar{s}_N = -\alpha M$ and $\bar{s}_P = \alpha t_P$. This is obtained from the fact that minimizer of the cross-entropy $\ell_{\text{RankDistil}}(t, s, P, N)$ is when the scores match (this can be easily verified by finding point such that $\nabla \ell_{\text{RankDistil}}(t, s, P, N) = 0$. Let s^* such that $s^*_P = \alpha t_P$ and $s^*_{K - \text{Top}_p(t)} = -\alpha M$ then $s^* \in \text{argmin}_s \ell_{\text{RankDistil}}(t, s, P, N)$ for all $N \subseteq [K]$. Therefore, by Lemma 9, $s^* \in \text{argmin}_s \ell_d(t, s)$, which completes the proof.

C.3 Proof of Theorem 7

We provide the proof for the case where $b = m$ in Algorithm 1. The proof for $b < m$ is similar. Since the $\ell_{\text{RankDistil}}$ in binary RankDistil-Loss case is separable in P and N, RankDistil-Loss can be written as

$$\ell_d(t, s) = \Psi(t, s, \text{Top}_p(t)) + \mathbb{E}_{B \sim Q(.|t)} \sum_{i \in B} \varphi(-s_i)$$

From the above form, it is easy to see that $s^*_i \leq \gamma$ for all $i \notin \text{Top}_p(t)$ since φ is strictly decreasing in $(-\infty, -\gamma]$ and $supp(Q(.|t)) = [K] - \text{Top}_p(t)$. Furthermore, $\text{argmax}_i s^*_P = \text{argmax}_i f^{\text{t}}_P(x)$ for all $x \in \mathcal{X}$, $i \in [p]$ and $[f^{\text{t}}(x)]_{[p]} > \gamma$. Combining these facts gives us the desired result.

C.4 Proof of Theorem 8

The proof is similar to the binary case. Similar to the binary case, we provide the proof for the case where $b = m$ in Algorithm 1. The proof for $b < m$ is similar. In the pairwise case, RankDistil-Loss can be written as

$$\ell_d(t, s) = \Psi(t, s, \text{Top}_p(t)) + \mathbb{E}_{B \sim Q(.|t)} \sum_{i \in B} \sum_{j \in \text{Top}_p(t)} \varphi(s_j - s_i)$$

From the above form, it is easy to see that $s^*_i < s^*_j$ for all $i \notin \text{Top}_p(t)$ and $j \in \text{Top}_p(t)$ since φ is strictly decreasing in $(-\infty, 0]$ and $supp(Q(.|t)) = [K] - \text{Top}_p(t)$. Also, $\text{argmax}_i s^*_P = \text{argmax}_i f^{\text{t}}_P(x)$ for all $x \in \mathcal{X}$, $i \in [p]$. Putting these facts together completes the proof for the result.

Lemma 9. *Suppose $F(u) = \mathbb{E}_z[G(u, z)]$ where $G : \mathbb{R}^K \times \mathcal{Z} \to \mathbb{R}$. If $u^* \in \text{argmin}_u G(u, z)$ for all $z \in \mathcal{Z}$, then $u^* \in \text{argmin}_u F(u)$.*

Proof. The result trivially follows from the definition of argmin. □