On the Learnability of Multilabel Ranking

Vinod Raman
Department of Statistics
University of Michigan
Ann Arbor, MI 48104
vkraman@umich.edu

Unique Subedi
Department of Statistics
University of Michigan
Ann Arbor, MI 48104
subedi@umich.edu

Ambuj Tewari
Department of Statistics
University of Michigan
Ann Arbor, MI 48104
tewaria@umich.edu

Abstract

Multilabel ranking is a central task in machine learning. However, the most fundamental question of learnability in a multilabel ranking setting with relevance-score feedback remains unanswered. In this work, we characterize the learnability of multilabel ranking problems in both batch and online settings for a large family of ranking losses. Along the way, we give two equivalence classes of ranking losses based on learnability that capture most losses used in practice.

1 Introduction

Multilabel ranking is a supervised learning problem where a learner is presented with an instance $x \in \mathcal{X}$ and is required to output a ranking of K different labels in decreasing order of relevance to x. This is in contrast with *multilabel classification* where given an instance $x \in \mathcal{X}$, the learner is tasked with predicting a subset of the K labels without any explicit ordering. Multilabel ranking is a canonical learning problem with a wide range of applications to text categorization, genetics, medical imaging, social networks, and visual object recognition [Joachims, 2005, Schapire and Singer, 2000, McCallum, 1999, Clare and King, 2001, Baltruschat et al., 2019, Wang and Sukthankar, 2013, Bucak et al., 2009, Yang et al., 2016]. Recent years have seen a surge in the development of multilabel ranking methods with strong practical and theoretical guarantees [Schapire and Singer, 2000, Dembczynski et al., 2012, Gong et al., 2013, Bucak et al., 2009, Jung and Tewari, 2018, Gao and Zhou, 2011, Koyejo et al., 2015, Zhang and Zhou, 2013, Korba et al., 2018]. A related line of work has studied consistency for the convex surrogates of natural ranking losses [Duchi et al., 2010, Buffoni et al., 2011, Gao and Zhou, 2011, Ravikumar et al., 2011, Calauzenes et al., 2012, Dembczynski et al., 2012]. Despite this vast literature on multilabel ranking, the fundamental question of when a multilabel ranking problem is *learnable* remains unanswered.

Understanding when a hypothesis class is learnable is a fundamental question in Statistical Learning Theory. For binary classification, the finiteness of the Vapnik–Chervonenkis (VC) dimension is both sufficient and necessary for Probably Approximately Correct (PAC) learning [Vapnik and Chervonenkis, 1974, Valiant, 1984]. Likewise, the finiteness of the Daniely-Shwartz (DS) dimension characterizes multiclass PAC learnability Daniely and Shalev-Shwartz [2014], Brukhim et al. [2022]. In the online setting, the Littlestone dimension [Littlestone, 1987] characterizes the online learnability of a binary hypothesis class and the multiclass Littlestone dimension [Daniely et al., 2011] characterizes online multiclass learnability. Unlike classification, a distinguishing property of multilabel ranking is the mismatch between the predictions the learner makes and the feedback it receives. In particular, a learner is required to produce a permutation that ranks the relevance of the labels but only receives a *relevance-score vector* as feedback. This feedback model is standard in multilabel ranking since obtaining full permutation feedback is generally costly [Liu et al., 2009]. As a result, unlike the 0-1 loss in classification, there is no canonical loss function in ranking. Together, these two issues create barriers for existing techniques used to prove learnability, such as the agnostic-to-realizable reductions from Hopkins et al. [2022] and Raman et al. [2023], to readily extend to ranking.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

In this paper, we characterize the batch and online learnability of a ranking hypothesis class $\mathcal{H} \subseteq \mathcal{S}_K^{\mathcal{X}}$ under relevance-score feedback, where $\mathcal{S}_K$ is the set of all permutations over $[K] = \{1, ..., K\}$. In doing so, we make the following contributions.

- We show that a ranking hypothesis class $\mathcal{H}$ embeds K^2 different *binary* hypothesis classes $\mathcal{H}_i^j$ for $i, j \in [K]$, where hypotheses in $\mathcal{H}_i^j$ answer whether the label i should be ranked in the top j. Our main result relates the learnability of $\mathcal{H}$ to the learnability of $\mathcal{H}_i^j$'s.
- We define two families of ranking loss functions that capture most if not all ranking losses used in practice. We show that these families are actually *equivalence* classes - the same characterization of batch and online learnability holds for every loss in that family.
- By relating the learnability of $\mathcal{H}$ to the learnability of *binary* hypothesis classes $\mathcal{H}_i^j$, we show that existing combinatorial dimensions, like the VC and Littlestone dimension, continue to characterize learnability in the multilabel ranking setting. This allows us to prove that linear ranking hypothesis classes are learnable in the batch setting.

A unifying theme throughout the paper is our ability to *constructively* convert a learning algorithm $\mathcal{A}$ for $\mathcal{H}$ into a learning algorithm $\mathcal{A}_i^j$ for $\mathcal{H}_i^j$ for each $i, j \in [K]$ and vice versa. To do so, our proof techniques involve adapting the agnostic-to-realizable reduction for batch and online classification, proposed by Hopkins et al. [2022] and Raman et al. [2023] respectively, to ranking.

2 Preliminaries and Notation

Let $\mathcal{X}$ denote the instance space, $\mathcal{S}_K$ the set of permutations over labels $[K] := \{1, ..., K\}$, and $\mathcal{Y} = \{0, 1, ..., B\}^K$ the target space for some $K, B \in \mathbb{N}$. We highlight that the set of labels $[K]$ is fixed beforehand and does not depend on the instance $x \in \mathcal{X}$. This is to be contrasted with subset ranking, the set of labels can change depending on the instance $x \in \mathcal{X}$.

We refer to an element $y \in \mathcal{Y}$ as a *relevance-score vector* that indicates the relevance of each of the K labels, where B indicates the highest relevance and 0 indicates the lowest relevance. Throughout the paper, we treat a permutation $\pi \in \mathcal{S}_K$ as a vector in $\{1, ..., K\}^K$ that induces a *ranking* of the K labels in decreasing order of relevance. Accordingly, for an index $i \in [K]$, we let $\pi_i \in [K]$ denote the *rank* of label i. Likewise, given an index $i \in [K]$, we let y^i denote the relevance of label i. In addition, it will be useful to define a mapping from $\mathcal{S}_K$ to $\{0, 1\}^K$. In particular, we define $\text{BinRel}(\cdot, \cdot) : \mathcal{S}_K \times [K] \to \{0, 1\}^K$ as an operator that given a permutation (ranking) $\pi \in \mathcal{S}_K$ and threshold $p \in [K]$, outputs a bit string $b \in \{0, 1\}^K$ s.t. $b_i = \mathbb{1}\{\pi_i \leq p\}$.

Ranking Equivalences. Our construction of ranking loss families in Section 3 requires different notions of equivalence between permutations (rankings) in $\mathcal{S}_K$. To that end, we say that $\pi = \hat{\pi}$ if and only if for all $i \in [K]$, $\pi_i = \hat{\pi}_i$. On the other hand, we say $\pi \overset{p}{=} \hat{\pi}$ if and only if $\{i : \pi_i \leq p\} = \{i : \hat{\pi}_i \leq p\}$. That is, two rankings are p-equivalent if the *set* of labels they rank in the top-p are equal. Finally, we say $\pi \overset{[p]}{=} \hat{\pi}$ if and only if for all $j \in [p]$, $\{i : \pi_i \leq j\} = \{i : \hat{\pi}_i \leq j\}$. That is, two rankings are $[p]$-equivalent if not only the *set* but also the *order* of labels they rank in the top-p are equal.

Ranking Hypothesis. A ranking hypothesis $h \in \mathcal{H} \subseteq \mathcal{S}_K^{\mathcal{X}}$ maps instances in $\mathcal{X}$ to a ranking (permutation) in $\mathcal{S}_K$. Given an instance $x \in \mathcal{X}$, one can think of $h(x)$ as h's ranking of the K different labels in decreasing order of relevance. For any ranking hypothesis h, we let $h_i : \mathcal{X} \to [K]$ denote its restriction to the i'th coordinate output. Accordingly, for an instance $x \in \mathcal{X}$, $h_i(x)$ gives the rank that h assigns to label i. Given a ranking hypothesis class $\mathcal{H} \subseteq \mathcal{S}_K^{\mathcal{X}}$ and any $i, j \in [K]$, we define its binary threshold-restricted hypothesis class $\mathcal{H}_i^j = \{h_i^j : h \in \mathcal{H}\}$ where $h_i^j(x) = \mathbb{1}\{h_i(x) \leq j\}$. We can think of hypotheses in $\mathcal{H}_i^j$ as providing binary responses to queries of the form: "for instance x, should label i ranked in the top j?" These threshold-restricted classes are central to our characterization of learnability in both the batch and online learning settings.

Batch Learnability. In the batch setting, we are interested in characterizing the learnability of a ranking hypothesis class $\mathcal{H}$ under a model similar to the classical PAC model [Valiant, 1984].

Definition 1 (Agnostic Ranking PAC Learnability)**.** *A ranking hypothesis class $\mathcal{H} \subseteq \mathcal{S}_K^{\mathcal{X}}$ is agnostic PAC learnable w.r.t. loss $\ell : \mathcal{S}_K \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, if there exists a function $m : (0,1)^2 \times \mathbb{N} \to \mathbb{N}$ and a learning algorithm $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^\star \to \mathcal{S}_K^{\mathcal{X}}$ with the following property: for every $\epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, running algorithm $\mathcal{A}$ on $n \geq m(\epsilon, \delta, K)$ iid samples from $\mathcal{D}$ outputs a predictor $g = \mathcal{A}(S)$ such that with probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$,*

$$\mathbb{E}_{\mathcal{D}}[\ell(g(x), y)] \leq \inf_{h \in \mathcal{H}} \mathbb{E}_{\mathcal{D}}[\ell(h(x), y)] + \epsilon.$$

If $\mathcal{D}$ is restricted to the class of distributions such that $\inf_{h \in \mathcal{H}} \mathbb{E}_{\mathcal{D}}[\ell(h(x), y)] = 0$, then we say we are in the *realizable* setting. Note that unlike in classification, realizability in the multilabel ranking setting is loss dependent.

Online Learnability. In the online setting, an adversary plays a sequential game with the learner over T rounds. In each round $t \in [T]$, an adversary selects a labeled instance $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ and reveals x_t to the learner. The learner makes a (potentially randomized) prediction $\hat{\pi}_t \in \mathcal{S}_K$. Finally, the adversary reveals the true relevance-score vector y_t, and the learner suffers the loss $\ell(\hat{\pi}_t, y_t)$, where ℓ is some pre-specified ranking loss function. Given a ranking hypothesis class $\mathcal{H} \subseteq \mathcal{S}_K^{\mathcal{X}}$, the goal of the learner is to output predictions $\hat{\pi}_t$ such that its cumulative loss is close to the best possible cumulative loss over hypotheses in $\mathcal{H}$. A hypothesis class is online learnable if there exists an algorithm such that for any sequence of labeled examples $(x_1, y_1), ..., (x_T, y_T)$, the difference in cumulative loss between its predictions and the predictions of the best possible function in $\mathcal{H}$ is small.

Definition 2 (Agnostic Online Ranking Learnability)**.** *A ranking hypothesis class $\mathcal{H} \subseteq \mathcal{S}_K^{\mathcal{X}}$ is agnostic online learnable w.r.t. loss ℓ, if there exists an (potentially randomized) algorithm $\mathcal{A}$ such that for any adaptively chosen sequence of labeled examples $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$, the algorithm outputs $\mathcal{A}(x_t) \in \mathcal{S}_K$ at every iteration $t \in [T]$ such that its expected regret,*

$$R(T, K) := \mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{A}(x_t), y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(x_t), y_t)\right],$$

is a non-decreasing, sub-linear function of T. Here, the expectation is taken with respect to the randomness of the algorithm $\mathcal{A}$.

If it is further guaranteed that there exists a hypothesis $h^\star \in \mathcal{H}$ such that $\sum_{t=1}^{T} \ell(h^\star(x_t), y_t) = 0$, then we say we are in the *realizable* setting. Again, realizability is loss dependent.

3 Ranking Loss Families

In statistical learning theory, we often characterize learnability with respect to a loss function. Unlike the 0-1 loss in classification, there is no canonical loss function in multilabel ranking. Accordingly, we define two general families of ranking loss functions in this section and later characterize learnability with respect to all losses in these families. In Appendix A, we show that many of the ranking metrics used in practice (e.g. Pairwise Rank Loss, Discounted Cumulative Gain, Reciprocal Rank, Average Precision, Precision@p, etc.) fall into one of these two families.

On a high level, we can classify ranking losses into two main groups: (A) those losses that care about both the order and magnitude of the relevance scores within the top-p ranked labels and (B) those losses that only care about the magnitude of the relevance scores within the top-p ranked labels. Our goal will be to define a loss family for both groups A and B. To do so, we start by identifying a canonical ranking loss that lies in each group. For group A, the normalized sum loss@p,

$$\ell_{\text{sum}}^{@p}(\pi, y) = \sum_{i=1}^{K} \min(\pi_i, p+1) y^i - Z_y^p$$

captures both the order and magnitude of the relevance scores only for the top-p ranked labels. Here, Z_y^p is an appropriately chosen normalization factor that only depends on p and y such that $\min_{\pi \in \mathcal{S}_K} \ell_{\text{sum}}^{@p}(\pi, y) = 0$. For Group B, the normalized precision loss@p,

$$\ell_{\text{prec}}^{@p}(\pi, y) = Z_y^p - \sum_{i=1}^{K} \mathbb{1}\{\pi_i \leq p\} y^i$$

cares only about the magnitude of relevance scores in the top-p ranked labels. Again, Z_y^p is an appropriately chosen normalization constant that only depends on p and y such that the minimum loss is 0. The form of $\ell_{\text{prec}}^{@p}$ differs from $\ell_{\text{sum}}^{@p}$ because $\sum_{i=1}^{K} \mathbb{1}\{\pi_i \leq p\} y^i$ is a gain whereas $\sum_{i=1}^{K} \min(\pi_i, p+1) y^i$ is a loss.

Next, we build loss families around $\ell_{\text{sum}}^{@p}$ and $\ell_{\text{prec}}^{@p}$. For $\ell_{\text{sum}}^{@p}$, consider the family:

$$\mathcal{L}(\ell_{\text{sum}}^{@p}) = \{\ell \in \mathbb{R}^{\mathcal{S}_K \times \mathcal{Y}} : \ell = 0 \text{ if and only if } \ell_{\text{sum}}^{@p} = 0\} \cap \{\ell \in \mathbb{R}^{\mathcal{S}_K \times \mathcal{Y}} : \pi \overset{[p]}{=} \hat{\pi} \implies \ell(\pi, y) = \ell(\hat{\pi}, y)\}.$$

By definition, $\mathcal{L}(\ell_{\text{sum}}^{@p})$ contains those ranking losses that are (1) zero-matched with $\ell_{\text{sum}}^{@p}$ and (2) remain unchanged for any two predicted rankings (permutations) that are $[p]$-equivalent. The second constraint is needed to ensure that losses in $\mathcal{L}(\ell_{\text{sum}}^{@p})$ only depend on the order and set of labels that π ranks in the top-p. Likewise, we can construct a similar loss family around $\ell_{\text{prec}}^{@p}$ as follows:

$$\mathcal{L}(\ell_{\text{prec}}^{@p}) = \{\ell \in \mathbb{R}^{\mathcal{S}_K \times \mathcal{Y}} : \ell = 0 \text{ if and only if } \ell_{\text{prec}}^{@p} = 0\} \cap \{\ell \in \mathbb{R}^{\mathcal{S}_K \times \mathcal{Y}} : \pi \overset{p}{=} \hat{\pi} \implies \ell(\pi, y) = \ell(\hat{\pi}, y)\}.$$

The set $\mathcal{L}(\ell_{\text{prec}}^{@p})$ contains those ranking losses that are (1) zero-matched with $\ell_{\text{prec}}^{@p}$ and (2) remain unchanged for any two predicted rankings (permutations) that are *p-equivalent*. The second constraint is needed to ensure that losses in $\mathcal{L}(\ell_{\text{prec}}^{@p})$ only depend on the set of labels that π ranks in the top-p. A major contribution of this paper is showing that both $\mathcal{L}(\ell_{\text{sum}}^{@p})$ and $\mathcal{L}(\ell_{\text{prec}}^{@p})$ are actually *equivalence* classes - the same characterization of learnability holds for every loss in that family.

4 Batch Multilabel Ranking

In this section, we characterize the agnostic PAC learnability of hypothesis classes $\mathcal{H} \subseteq \mathcal{S}_K^{\mathcal{X}}$ with respect to both $\mathcal{L}(\ell_{\text{sum}}^{@p})$ and $\mathcal{L}(\ell_{\text{prec}}^{@p})$. Our main results, stated below as two theorems, relate the learnability of $\mathcal{H}$ to the learnability of the threshold-restricted classes $\mathcal{H}_i^j$.

Theorem 4.1. *A hypothesis class $\mathcal{H} \subseteq \mathcal{S}_K^{\mathcal{X}}$ is agnostic PAC learnable w.r.t $\ell \in \mathcal{L}(\ell_{sum}^{@p})$* if and only if *for all $i \in [K]$ and $j \in [p]$, $\mathcal{H}_i^j$ is agnostic PAC learnable w.r.t the 0-1 loss.*

Theorem 4.2. *A hypothesis class $\mathcal{H} \subseteq \mathcal{S}_K^{\mathcal{X}}$ is agnostic PAC learnable w.r.t $\ell \in \mathcal{L}(\ell_{prec}^{@p})$* if and only if *for all $i \in [K]$, $\mathcal{H}_i^p$ is agnostic PAC learnable w.r.t the 0-1 loss.*

Since VC dimension characterizes the learnability of binary hypothesis classes under the 0-1 loss, an important corollary of Theorems 4.1 and 4.2 is that finiteness of VC($\mathcal{H}_i^j$)'s, for the appropriate $i, j \in [K] \times [p]$, is necessary and sufficient for agnostic ranking PAC learnability. Later on, we use this fact to prove that linear ranking hypothesis classes are agnostic ranking PAC learnable.

We start with the proof of Theorem 4.1, which follows in three steps. First, we show that if for all $(i, j) \in [K] \times [p]$, $\mathcal{H}_i^j$ is agnostic PAC learnable w.r.t 0-1 loss, then Empirical Risk Minimization (ERM) is an agnostic PAC learner for $\mathcal{H}$ w.r.t $\ell_{\text{sum}}^{@p}$. Next, we show that if $\mathcal{H}$ is agnostic PAC learnable w.r.t $\ell_{\text{sum}}^{@p}$, then $\mathcal{H}$ is agnostic PAC learnable w.r.t any loss $\ell \in \mathcal{L}(\ell_{\text{sum}}^{@p})$. Our proof of the latter uses the realizable to agnostic conversion from Hopkins et al. [2022]. Finally, we prove the necessity direction - if $\mathcal{H}$ is agnostic PAC learnable w.r.t an arbitrary $\ell \in \mathcal{L}(\ell_{\text{sum}}^{@p})$, then for all $(i, j) \in [K] \times [p]$, $\mathcal{H}_i^j$ is agnostic PAC learnable w.r.t 0-1 loss. The proof of necessity direction also uses the realizable to agnostic conversion from Hopkins et al. [2022]. The proof of Theorem 4.2 follows exactly the same way as Theorem 4.1 with some minor changes. Thus, we only focus on the proof of Theorem 4.1 in this section and defer all discussion of Theorem 4.2 to Appendix C.3.

We begin with Lemma 4.3, which asserts that if $\mathcal{H}_i^j$ is agnostic PAC learnable for all $(i, j) \in [K] \times [p]$, then ERM is an agnostic PAC learner for $\mathcal{H}$ w.r.t $\ell_{\text{sum}}^{@p}$.

Lemma 4.3. *If for all $i \in [K]$ and $j \in [p]$, $\mathcal{H}_i^j$ is agnostic PAC learnable w.r.t the 0-1 loss, then* ERM *is an agnostic PAC learner for $\mathcal{H} \subseteq \mathcal{S}_K^{\mathcal{X}}$ w.r.t $\ell_{sum}^{@p}$.*

The proof of Lemma 4.3 exploits the nice structure of $\ell_{\text{sum}}^{@p}$ by upperbounding the empirical Rademacher complexity of the loss class $\ell_{\text{sum}}^{@p} \circ \mathcal{H} = \{(x, y) \mapsto \ell_{\text{sum}}^{@p}(h(x), y) : h \in \mathcal{H})\}$ and

showing that it vanishes as the sample size n becomes large. Then, standard uniform convergence arguments outlined in Proposition C.1 imply that ERM is an agnostic PAC learner for $\mathcal{H}$ w.r.t $\ell_{\text{sum}}^{@p}$. The full proof is in Appendix C.

Since arbitrary losses in $\mathcal{L}(\ell_{\text{sum}}^{@p})$ may not have nice analytical forms, Lemma 4.4 relates the learnability of an arbitrary loss $\ell \in \mathcal{L}(\ell_{\text{sum}}^{@p})$ to the learnability of $\ell_{\text{sum}}^{@p}$.

Lemma 4.4. *If $\mathcal{H} \subseteq \mathcal{S}_K^{\mathcal{X}}$ is agnostic PAC learnable w.r.t $\ell_{sum}^{@p}$, then $\mathcal{H}$ is agnostic PAC learnable w.r.t any $\ell \in \mathcal{L}(\ell_{sum}^{@p})$.*

Proof. (of Lemma 4.4) Fix $\ell \in \mathcal{L}(\ell_{\text{sum}}^{@p})$. Let $a = \min_{\pi,y}\{\ell(\pi, y) \mid \ell(\pi, y) \neq 0\}$ and $b = \max_{\pi,y} \ell(\pi, y)$. We need to show that if $\mathcal{H}$ is agnostic PAC learnable w.r.t $\ell_{\text{sum}}^{@p}$, then $\mathcal{H}$ is agnostic PAC learnable w.r.t ℓ. We will do so in two steps. First, we will show that if $\mathcal{A}$ is an agnostic PAC learner for $\ell_{\text{sum}}^{@p}$, then $\mathcal{A}$ is also a *realizable* PAC learner for ℓ. Next, we will show how to convert a realizable PAC learner for ℓ into an agnostic PAC learner for ℓ in a black-box fashion. The composition of these two pieces yields an agnostic PAC learner for $\mathcal{H}$ w.r.t ℓ.

Realizable PAC learnability of $\mathcal{H}$ w.r.t ℓ. If $\mathcal{H}$ is agnostic PAC learnable w.r.t $\ell_{\text{sum}}^{@p}$, then there exists a learning algorithm $\mathcal{A}$ with sample complexity $m(\epsilon, \delta, K)$ s.t. for any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, with probability $1 - \delta$ over a sample $S \sim \mathcal{D}^n$ of size $n \geq m(\epsilon, \delta, K)$, the output predictor $g = \mathcal{A}(S)$ achieves $\mathbb{E}_{\mathcal{D}}\left[\ell_{\text{sum}}^{@p}(g(x), y))\right] \leq \inf_{h\in\mathcal{H}} \mathbb{E}_{\mathcal{D}}\left[\ell_{\text{sum}}^{@p}(h(x), y))\right] + \epsilon$. In the realizable setting, we are further guaranteed that there exists a hypothesis $h^\star \in \mathcal{H}$ s.t. $\mathbb{E}_{\mathcal{D}}\left[\ell(h^\star(x), y))\right] = 0$. Since $\ell \in \mathcal{L}(\ell_{\text{sum}}^{@p})$, this also implies that $\mathbb{E}_{\mathcal{D}}\left[\ell_{\text{sum}}^{@p}(h^\star(x), y))\right] = 0$. Therefore, under realizability and the fact that $\ell \leq b\, \ell_{\text{sum}}^{@p}$, we have $\mathbb{E}_{\mathcal{D}}\left[\ell(g(x), y))\right] \leq b\epsilon$. This completes the first part of the proof as we have shown that $\mathcal{A}$ is also a realizable PAC learner for $\mathcal{H}$ w.r.t ℓ with sample complexity $m(\frac{\epsilon}{b}, \delta, K)$.

Realizable-to-agnostic conversion. Now, we show how to convert the realizable PAC learner $\mathcal{A}$ for ℓ into an agnostic PAC learner for ℓ in a black-box fashion. For this step, we will extend the agnostic-to-realizable reduction proposed by Hopkins et al. [2022] to the ranking setting by accommodating the mismatch between the range space of $\mathcal{H}$ and the label space $\mathcal{Y}$. In particular, we will show that Algorithm 1 below converts a realizable PAC learner for ℓ into an agnostic PAC learner for ℓ. Note that although input $\mathcal{A}$ is a realizable learner, the distribution $\mathcal{D}$ may not be realizable.

Algorithm 1 Agnostic PAC learner for $\mathcal{H}$ w.r.t. ℓ

Input: Realizable PAC learner $\mathcal{A}$ for $\mathcal{H}$, unlabeled and labeled samples $S_U \sim \mathcal{D}_{\mathcal{X}}^n$ and $S_L \sim \mathcal{D}^m$

1 For each $h \in \mathcal{H}_{|S_U}$, construct a dataset

$$S_U^h = \{(x_1, \tilde{y}_1), ..., (x_n, \tilde{y}_n)\} \text{ s.t. } \tilde{y}_i \sim \text{Unif}\{\text{BinRel}(h(x_i), 1), ..., \text{BinRel}(h(x_i), p)\}$$

2 Run $\mathcal{A}$ over all datasets to get $C(S_U) := \left\{\mathcal{A}\left(S_U^h\right) \mid h \in \mathcal{H}_{|S_U}\right\}$

3 Return $\hat{g} \in C(S_U)$ with the lowest empirical error over S_L w.r.t. ℓ.

Let $h^\star = \arg\min_{h\in\mathcal{H}} \mathbb{E}_{\mathcal{D}}\left[\ell(h(x), y)\right]$ denote the optimal predictor in $\mathcal{H}$ w.r.t $\mathcal{D}$. Consider the sample $S_U^{h^\star}$ and let $g = \mathcal{A}(S_U^{h^\star})$. We can think of g as the output of $\mathcal{A}$ run over an i.i.d sample S drawn from $\mathcal{D}^\star$, a joint distribution over $\mathcal{X}\times\mathcal{Y}$ defined procedurally by first sampling $x \sim \mathcal{D}_{\mathcal{X}}$, then independently sampling $j \sim \text{Unif}([p])$, and finally outputting the labeled sample $(x, \text{BinRel}(h^\star(x), j))$. Note that $\mathcal{D}^\star$ is indeed a realizable distribution (realized by $h^\star$) w.r.t both ℓ and $\ell_{\text{sum}}^{@p}$. Recall that $m_{\mathcal{A}}(\frac{\epsilon}{b}, \delta, K)$ is the sample complexity of $\mathcal{A}$. Since $\mathcal{A}$ is a realizable learner for $\mathcal{H}$ w.r.t ℓ, we have that for $n \geq m_{\mathcal{A}}(\frac{a\epsilon}{2b^2p}, \delta/2, K)$, with probability at least $1 - \frac{\delta}{2}$, $\mathbb{E}_{\mathcal{D}^\star}\left[\ell(g(x), y)\right] \leq \frac{a\epsilon}{2bp}$.

Next, by Lemma E.1, we have $\ell(g(x), y) \leq \ell(h^\star(x), y) + \frac{bp}{a}\mathbb{E}_{j\sim\text{Unif}([p])}\left[\ell(g(x), \text{BinRel}(h^\star(x), j))\right]$ pointwise. Taking expectations on both sides of the inequality gives

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[\ell(g(x), y)\right] &\leq \mathbb{E}_{\mathcal{D}}\left[\ell(h^\star(x), y)\right] + \frac{bp}{a}\mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{j\sim\text{Unif}([p])}\left[\ell(g(x), \text{BinRel}(h^\star(x), j))\right]\right] \\ &\leq \mathbb{E}_{\mathcal{D}}\left[\ell(h^\star(x), y)\right] + \frac{\epsilon}{2}.\end{aligned}$$

The last inequality follows from the definition of $\mathcal{D}^\star$, namely $\mathbb{E}_{\mathcal{D}^\star}[\ell(g(x),y)] = \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\mathbb{E}_{j\sim\text{Unif}([p])}[\ell(g(x),\text{BinRel}(h^\star(x),j))]$. This shows that $C(S_U)$ contains a hypothesis g that generalizes well with respect to $\mathcal{D}$. Now we want to show that the predictor $\hat{g}$ returned in step 4 also has good generalization. Crucially, observe that $C(S_U)$ is a finite hypothesis class with cardinality at most 2^{nK}. By standard Chernoff and union bounds, with probability at least $1-\delta/2$, the empirical risk of every hypothesis in $C(S_U)$ on a sample of size $\geq \frac{8}{\epsilon^2}\log\frac{4|C(S_U)|}{\delta}$ is at most $\epsilon/4$ away from its true error. So, if $m = |S_L| \geq \frac{8}{\epsilon^2}\log\frac{4|C(S_U)|}{\delta}$, then with probability at least $1-\delta/2$,

$$\frac{1}{|S_L|}\sum_{(x,y)\in S_L}\ell(g(x),y) \leq \mathbb{E}_\mathcal{D}[\ell(g(x),y)] + \frac{\epsilon}{4} \leq \mathbb{E}_\mathcal{D}[\ell(h^\star(x),y)] + \frac{3\epsilon}{4}.$$

Since $\hat{g}$ is the ERM on S_L over $C(S)$, its empirical risk can be at most $\mathbb{E}_\mathcal{D}[\ell(h^\star(x),y)] + \frac{3\epsilon}{4}$. Given that the population risk of $\hat{g}$ can be at most $\epsilon/4$ away from its empirical risk, we have that

$$\mathbb{E}_\mathcal{D}[\ell(\hat{g}(x),y)] \leq \mathbb{E}_\mathcal{D}[\ell(h^\star(x),y)] + \epsilon.$$

Applying union bounds, the entire process succeeds with probability $1-\delta$. We can upper bound the sample complexity of Algorithm 1, denoted $n(\epsilon,\delta,K)$, as

$$\begin{aligned} n(\epsilon,\delta,K) &\leq m_\mathcal{A}\left(\frac{a\epsilon}{2b^2p},\delta/2,K\right) + O\left(\frac{1}{\epsilon^2}\log\frac{|C(S_U)|}{\delta}\right) \\ &\leq m_\mathcal{A}\left(\frac{a\epsilon}{2b^2p},\delta/2,K\right) + O\left(\frac{Km_\mathcal{A}(\frac{a\epsilon}{2b^2p},\delta/2,K) + \log\frac{1}{\delta}}{\epsilon^2}\right), \end{aligned}$$

where we use $|C(S_U)| \leq 2^{Km_\mathcal{A}(\frac{a\epsilon}{2b^2p},\delta/2,K)}$. This shows that Algorithm 1 is an agnostic PAC learner for $\mathcal{H}$ w.r.t ℓ. □

Finally, Lemma 4.5 gives the necessity direction of Theorem 4.1.

Lemma 4.5. *If a hypothesis class $\mathcal{H} \subseteq \mathcal{S}_K^\mathcal{X}$ is agnostic PAC learnable w.r.t $\ell \in \mathcal{L}(\ell_{sum}^{@p})$, then $\mathcal{H}_i^j$ is agnostic PAC learnable w.r.t the 0-1 loss for all $(i,j) \in [K]\times[p]$.*

Like the sufficiency proofs, the proof of Lemma 4.5 is constructive. Given an agnostic PAC learner for $\mathcal{H}$ w.r.t ℓ, we construct an agnostic PAC learner for $\mathcal{H}_i^j$ w.r.t 0-1 loss using a slight modification of Algorithm 1. We defer the full proof to Appendix C since the analysis is similar to that of Algorithm 1. Together, Lemmas 4.3, 4.4 and 4.5 imply Theorem 4.1.

We conclude this section by giving a concrete application of our characterization. Consider the class of ranking-hypotheses $\mathcal{H} = \{x \mapsto \text{argsort}(Wx) : W \in \mathbb{R}^{K\times d}\}$ that compute rankings by sorting scores, in descending order, obtained from a linear function of the input features. Lemma 4.6, whose proof is in Appendix B, computes the VC dimension of $\mathcal{H}_i^j$ for an arbitrary $i,j \in [K]$.

Lemma 4.6. *Let $\mathcal{H} = \{x \mapsto \text{argsort}(Wx) : W \in \mathbb{R}^{K\times d}\}$ be a linear ranking hypothesis class. Then for all $i,j \in [K]$, $VC(\mathcal{H}_i^j) = \tilde{O}(Kd)$, where $\tilde{O}$ hides logarithmic factors of d and K.*

Combining Lemma 4.6 with Theorems 4.1 and 4.2 shows that linear ranking hypothesis classes are agnostic ranking PAC learnable w.r.t to all losses in $\mathcal{L}(\ell_{\text{sum}}^{@p}) \cup \mathcal{L}(\ell_{\text{prec}}^{@p})$. More generally, in Appendix B we give a dimension-based sufficient condition under which *generic* score-based ranking hypothesis classes are agnostic ranking PAC learnable.

5 Online Multilabel Ranking

We now move to the online setting and characterize the online learnability of hypothesis classes $\mathcal{H} \subseteq \mathcal{S}_K^\mathcal{X}$ with respect to both $\mathcal{L}(\ell_{\text{sum}}^{@p})$ and $\mathcal{L}(\ell_{\text{prec}}^{@p})$. As in the batch setting, our characterization relates the learnability of $\mathcal{H}$ to the learnability of the threshold-restricted classes $\mathcal{H}_i^j$.

Theorem 5.1. *A hypothesis class $\mathcal{H} \subseteq \mathcal{S}_K^\mathcal{X}$ is agnostic online learnable w.r.t $\ell \in \mathcal{L}(\ell_{sum}^{@p})$* if and only if *for all $i \in [K]$ and $j \in [p]$, $\mathcal{H}_i^j$ is agnostic online learnable w.r.t the 0-1 loss.*

Theorem 5.2. *A hypothesis class $\mathcal{H} \subseteq \mathcal{S}_K^{\mathcal{X}}$ is agnostic online learnable w.r.t $\ell \in \mathcal{L}(\ell_{prec}^{@p})$* if and only if *for all $i \in [K]$, $\mathcal{H}_i^p$ is agnostic online learnable w.r.t the 0-1 loss.*

Since the Littlestone dimension characterizes the online learnability of binary hypothesis classes under the 0-1 loss, an important corollary of Theorems 5.1 and 5.2 is is that finiteness of $\text{Ldim}(\mathcal{H}_i^j)$, for the appropriate $i, j \in [K] \times [p]$, is necessary and sufficient for agnostic online ranking learnability.

We now begin the proof of Theorem 5.1. Since the proof of Theorem 5.2 follows a similar trajectory, we defer all discussion of Theorem 5.2 to Appendix D.2. Unlike Theorem 4.1 in the batch setting, we prove the sufficiency and necessity directions of Theorem 5.1 directly. We chose this direct path because, unlike the batch setting, sequential Rademacher analysis does not yield a constructive algorithm [Rakhlin et al., 2015]. On the other hand, our proofs are constructive and use the celebrated Randomized Exponential Weights Algorithm (REWA) [Cesa-Bianchi and Lugosi, 2006]. Moreover, a key ingredient of our proof is the realizable to agnostic conversion from Raman et al. [2023].

Proof. (of sufficiency in Theorem 5.1) Fix $\ell \in \mathcal{L}(\ell_{\text{sum}}^{@p})$. Let $a = \min_{\pi,y}\{\ell(\pi, y) \mid \ell(\pi, y) \neq 0\}$ and $M = \max_{\pi,y} \ell(\pi, y)$. Given online learners for $\mathcal{H}_i^j$ for the 0-1 loss, our goal is to construct an online learner $\mathcal{Q}$ for $\mathcal{H}$ w.r.t ℓ that enjoys sub-linear regret in T. Our strategy will be to construct a set of experts $\mathcal{E}$ using the online learners for $\mathcal{H}_i^j$'s and run REWA using $\mathcal{E}$ and an appropriately scaled version of ℓ. Our proof borrows ideas from the realizable-to-agnostic online conversion from Raman et al. [2023] and so we use the same notation whenever possible.

Let $(x_1, y_1), ..., (x_T, y_T) \in (\mathcal{X} \times \mathcal{Y})^T$ denote the stream of points to be observed by the online learner. We will assume an oblivious adversary and thus the stream is fixed before the game starts. A standard reduction (Chapter 4 in Cesa-Bianchi and Lugosi [2006]) allows us to convert oblivious regret bounds to adaptive regret bounds. Since $\mathcal{H}_i^j \subseteq \{0,1\}^{\mathcal{X}}$ is online learnable w.r.t. 0-1 loss, we are guaranteed the existence of online learners $\mathcal{A}_i^j$ for $\mathcal{H}_i^j$.

Constructing Experts. For any bitstring $b \in \{0,1\}^T$, let $\phi : \{t \in [T] : b_t = 1\} \to \mathcal{S}_K$ denote a function mapping time points where $b_t = 1$ to rankings (permutations). Let $\Phi_b = \mathcal{S}_K^{\{t \in [T] : b_t = 1\}}$ denote all such functions ϕ. For every $h \in \mathcal{H}$, there exists a $\phi_b^h \in \Phi_b$ such that for all $t \in \{t : b_t = 1\}$, $\phi_b^h(t) = h(x_t)$. Let $|b| = |\{t \in [T] : b_t = 1\}|$. For every $b \in \{0,1\}^T$ and $\phi \in \Phi_b$, we will define an Expert $E_{b,\phi}$. Expert $E_{b,\phi}$, formally presented in Algorithm 2, uses $\mathcal{A}_i^j$'s to make predictions in each round. However, $E_{b,\phi}$ only updates the $\mathcal{A}_i^j$'s on those rounds where $b_t = 1$, using ϕ to compute a labeled instance. For every $b \in \{0,1\}^T$, let $\mathcal{E}_b = \bigcup_{\phi \in \Phi_b}\{E_{b,\phi}\}$ denote the set of all Experts parameterized by functions $\phi \in \Phi_b$. If b is the bitstring with all zeros, then $\mathcal{E}_b$ will be empty. Therefore, we will actually define $\mathcal{E}_b = \{E_0\} \cup \bigcup_{\phi \in \Phi_b}\{E_{b,\phi}\}$, where E_0 is the expert that never updates $\mathcal{A}_i^j$'s and only uses them for predictions in all $t \in [T]$. Note that $1 \leq |\mathcal{E}_b| \leq (K!)^{|b|} \leq K^{K|b|}$.

Algorithm 2 Expert (b, ϕ)

Input: Independent copy of realizable learners $\mathcal{A}_i^j$ of $\mathcal{H}_i^j$ for each $(i, j) \in [K] \times [p]$

for $t = 1, ..., T$ **do**
  Receive example x_t
  Define a binary vote matrix $V_t \in \{0,1\}^{K \times p}$ such that $V_t[i, j] = \mathcal{A}_i^j(x_t)$
  Predict $\hat{\pi}_t \in \arg\min_{\pi \in \mathcal{S}_K} \langle \pi, V_t \mathbf{1}_p \rangle$
  if $b_t = 1$ **then**
    Let $\pi = \phi(t)$ and for all $(i, j) \in [K] \times [p]$, update $\mathcal{A}_i^j$ by passing (x_t, π_i^j)
end

Algorithm 3 Agnostic Online Learner $\mathcal{Q}$ for $\mathcal{H}$ w.r.t. ℓ

Input: Parameter $0 < \beta < 1$

Let $B \in \{0,1\}^T$ s.t. $B_t \overset{\text{iid}}{\sim} \text{Bernoulli}(\frac{T^\beta}{T})$
Construct the set of experts $\mathcal{E}_B = \{E_0\} \cup \bigcup_{\phi \in \Phi_B}\{E_{B,\phi}\}$ according to Algorithm 2
Run REWA $\mathcal{P}$ using $\mathcal{E}_B$ and the loss function $\frac{\ell}{M}$ over the stream $(x_1, y_1), ..., (x_T, y_T)$

Using these experts, Algorithm 3 presents our agnostic online learner $\mathcal{Q}$ for $\mathcal{H}$ w.r.t $\ell \in \mathcal{L}(\ell_{\text{sum}}^{@p})$. We now show that $\mathcal{Q}$ enjoys sub-linear regret. We highlight that there are three sources of randomness in online learner $\mathcal{Q}$, namely the randomness of sampling B, the internal randomness of $\mathcal{A}_i^j$'s, and the internal randomness of $\mathcal{P}$. One may think of internal randomness as arising from the sampling step involved in the randomized predictions. Let A be the random variable associated with joint internal randomness of $\mathcal{A}_i^j$ for all $(i,j) \in [K] \times [p]$. Similarly, denote P to be the random variable associated with the internal randomness of $\mathcal{P}$. We begin by using the guarantee of REWA.

REWA Guarantee. Using Theorem 21.11 in Shalev-Shwartz and Ben-David [2014] and the fact that B, A and P are mutually independent, REWA guarantees almost surely that

$$\sum_{t=1}^{T} \mathbb{E}\left[\ell(\mathcal{P}(x_t), y_t) | B, A\right] \leq \inf_{E \in \mathcal{E}_B} \sum_{t=1}^{T} \ell(E(x_t), y_t) + M\sqrt{2T \ln(|\mathcal{E}_B|)}.$$

Taking an outer expectation gives

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{P}(x_t), y_t)\right] \leq \mathbb{E}\left[\inf_{E \in \mathcal{E}_B} \sum_{t=1}^{T} \ell(E(x_t), y_t)\right] + \mathbb{E}\left[M\sqrt{2T \ln(|\mathcal{E}_B|)}\right].$$

Noting that $\mathcal{Q}(x_t) = \mathcal{P}(x_t)$, we obtain

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{Q}(x_t), y_t)\right] &\leq \mathbb{E}\left[\inf_{E \in \mathcal{E}_B} \sum_{t=1}^{T} \ell(E(x_t), y_t)\right] + \mathbb{E}\left[M\sqrt{2T \ln(|\mathcal{E}_B|)}\right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^{T} \ell(E_{B,\phi_B^{h^\star}}(x_t), y_t)\right] + M\mathbb{E}\left[\sqrt{2T \ln(|\mathcal{E}_B|)}\right].\end{aligned}$$

In the last step, we used the fact that for all $b \in \{0,1\}^T$ and $h \in \mathcal{H}$, $E_{b,\phi_b^h} \in \mathcal{E}_b$. Here, $h^\star = \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(x_t), y_t)$ is the optimal function in hindsight. First, note that $\ln(|\mathcal{E}_B|) \leq K|B| \ln(K)$. Using Jensen's inequality gives $\mathbb{E}\left[\sqrt{2T \ln(|\mathcal{E}_B|)}\right] \leq \sqrt{2T^{1+\beta} K \ln K}$. Thus,

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{Q}(x_t), y_t)\right] \leq \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \ell(E_{B,\phi_B^{h^\star}}(x_t), y_t)\right]}_{\text{(I)}} + M\sqrt{2T^{1+\beta} K \ln K}. \tag{1}$$

Upperbounding (I). It now suffices to upperbound $\mathbb{E}\left[\sum_{t=1}^{T} \ell(E_{B,\phi_B^{h^\star}}(x_t), y_t)\right]$. Recall that Lemma E.1 gives pointwise

$$\ell(E_{B,\phi_B^{h^\star}}(x_t), y_t) \leq \ell(h^\star(x_t), y_t) + \frac{pM}{a}\, \mathbb{E}_{j \sim \text{Unif}([p])}[\ell(E_{B,\phi_B^{h^\star}}(x_t), \text{BinRel}(h^\star(x_t), j))] \tag{2}$$

where $M = \max_{\pi,y} \ell(\pi, y)$ and $a = \min_{\pi,y}\{\ell(\pi, y) \mid \ell(\pi, y) \neq 0\}$. Note that, by definition of the constant M, we further get

$$\begin{aligned}\ell(E_{B,\phi_B^{h^\star}}(x_t), \text{BinRel}(h^\star(x_t), j)) &\leq M\, \mathbb{1}\{\ell(E_{B,\phi_B^{h^\star}}(x_t), \text{BinRel}(h^\star(x_t), j)) > 0\} \\ &= M\, \mathbb{1}\{\ell_{\text{sum}}^{@p}(E_{B,\phi_B^{h^\star}}(x_t), \text{BinRel}(h^\star(x_t), j)) > 0\},\end{aligned}$$

where the equality follows from the fact that $\ell \in \mathcal{L}(\ell_{\text{sum}}^{@p})$.

In order to upperbound the indicator above, we need to introduce some more notations. Given the realizable online learner $\mathcal{A}_i^m$ for $(i,m) \in [K] \times [p]$, an instance $x \in \mathcal{X}$, and an ordered finite sequence of labeled examples $L \in (\mathcal{X} \times \{0,1\})^*$, let $\mathcal{A}_i^m(x|L)$ be the random variable denoting the prediction of $\mathcal{A}_i^m$ on the instance x after running and updating on L. For any $b \in \{0,1\}^T$, $h \in \mathcal{H}$, and $t \in [T]$, let $L_{b_{<t}}^h(i,m) = \{(x_s, h_i^m(x_s)) : s < t \text{ and } b_s = 1\}$ denote the *subsequence* of the sequence of labeled instances $\{(x_s, h_i^m(x_s))\}_{s=1}^{t-1}$ where $b_s = 1$. Then, for any $j \in [p]$, we have

$$\mathbb{1}\{\ell_{\text{sum}}^{@p}(E_{B,\phi_B^{h^\star}}(x_t), \text{BinRel}(h^\star(x_t), j)) > 0\} \leq \sum_{i=1}^{K} \sum_{m=1}^{p} \mathbb{1}\{\mathcal{A}_i^m(x_t \mid L_{B_{<t}}^{h^\star}(i,m)) \neq h_i^{\star,m}(x_t)\}.$$

To prove the inequality above, consider the case when $\sum_{i=1}^{K}\sum_{m=1}^{p}\mathbb{1}\{\mathcal{A}_i^m(x_t \mid L_{B_{<t}}^{h^\star}(i,m)) \neq h_i^{\star,m}(x_t)\} = 0$ because the inequality is trivial otherwise. Then, we must have $\mathcal{A}_i^m(x_t \mid L_{B_{<t}}^{h^\star}(i,m)) = h_i^{\star,m}(x_t)$ for all $(i,m) \in [K]\times[p]$. Let $V_t \in \{0,1\}^{K\times p}$ be a binary vote matrix that $E_{B,\phi_B^{h^\star}}$ constructs in round t. Then, we have $V_t[i,m] = \mathcal{A}_i^m(x_t \mid L_{B_{<t}}^{h^\star}(i,m)) = h_i^{\star,m}(x_t)$ for all $(i,m) \in [K]\times[p]$. Since $h^\star(x_t)$ is a permutation, the vote vector $V_t\mathbf{1}_p$ must contain p labels with distinct number of non-zero votes, namely $p, p-1, p-2, \ldots, 2, 1$ votes. Similarly, there must be $K-p$ labels with exactly 0 votes. Thus, every $\hat{\pi}_t \in \arg\min_{\pi\in\mathcal{S}_K}\langle \pi, V_t\mathbf{1}_p\rangle$ must rank label that obtained p votes as 1, label with $p-1$ votes as 2, and so forth. In other words, we must have $\hat{\pi}_t \overset{[p]}{=} h^\star(x_t)$, and thus $\ell_{\text{sum}}^{@p}(\hat{\pi}_t, \text{BinRel}(h^\star(x_t), j)) = 0$ for any $j \in [p]$ by definition of $\ell_{\text{sum}}^{@p}$. Our claim now follows because $E_{B,\phi_B^{h^\star}}(x_t) \in \arg\min_{\pi\in\mathcal{S}_K}\langle \pi, V_t\mathbf{1}_p\rangle$. Using these two inequalities in equation (2), we obtain

$$\ell(E_{B,\phi_B^{h^\star}}(x_t), y_t) \le \ell(h^\star(x_t), y_t) + \frac{pM^2}{a}\sum_{i=1}^{K}\sum_{m=1}^{p}\mathbb{1}\{\mathcal{A}_i^m(x_t \mid L_{B_{<t}}^{h^\star}(i,m)) \neq h_i^{\star,m}(x_t)\},$$

which further implies that

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell(E_{B,\phi_B^{h^\star}}(x_t), y_t)\right] \le \sum_{t=1}^{T}\ell(h^\star(x_t), y_t) + \frac{pM^2}{a}\sum_{i=1}^{K}\sum_{m=1}^{p}\underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\mathcal{A}_i^m(x_t \mid L_{B_{<t}}^{h^\star}(i,m)) \neq h_i^{\star,m}(x_t)\}\right]}_{\text{(II)}}.$$

The first term above is the cumulative loss of the best-fixed hypothesis in hindsight.

Upperbounding (II). It now suffices to show that $\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\mathcal{A}_i^m(x_t \mid L_{B_{<t}}^{h^\star}(i,m)) \neq h_i^{\star,m}(x_t)\}\right]$ is sub-linear for every $(i,m) \in [K]\times[p]$. Note that we can write

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\mathcal{A}_i^m(x_t \mid L_{B_{<t}}^{h^\star}(i,m)) \neq h_i^{\star,m}(x_t)\}\right] &= \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{1}\{\mathcal{A}_i^m(x_t \mid L_{B_{<t}}^{h^\star}(i,m)) \neq h_i^{\star,m}(x_t)\}\right]\frac{\mathbb{E}\left[\mathbb{1}\{B_t=1\}\right]}{\mathbb{E}\left[\mathbb{1}\{B_t=1\}\right]}\\ &= \frac{T}{T^\beta}\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{1}\{\mathcal{A}_i^m(x_t \mid L_{B_{<t}}^{h^\star}(i,m)) \neq h_i^{\star,m}(x_t)\}\mathbb{1}\left\{B_t=1\right\}\right],\end{aligned}$$

where the last equality follows because $\mathbb{E}\left[\mathbb{1}\{B_t=1\}\right] = \frac{T^\beta}{T}$ and the prediction of $\mathcal{A}_i^m(x_t \mid L_{B_{<t}}^{h^\star}(i,m))$ on round t only depends on bitstring $(B_1,\ldots,B_{t-1})$, but is independent of B_t. Next, we can use the regret guarantee of algorithm $\mathcal{A}_i^m$ on the rounds it was updated. That is,

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\left[\mathcal{A}_i^m(x_t \mid L_{B_{<t}}^{h^\star}(i,m))\mathbb{1}\left\{B_t=1\right\}\right] &= \mathbb{E}\left[\sum_{t:B_t=1}\mathcal{A}_i^m(x_t \mid L_{B_{<t}}^{h^\star}(i,m)) \neq h_i^{\star,m}(x_t)\}\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[\sum_{t:B_t=1}\mathcal{A}_i^m(x_t \mid L_{B_{<t}}^{h^\star}(i,m)) \neq h_i^{\star,m}(x_t)\}\right]\middle| B\right] \le \mathbb{E}\left[R_i^m(|B|)\right],\end{aligned}$$

where $R_i^m(|B|)$ is the regret of $\mathcal{A}_i^m$, a sub-linear function of $|B|$. In the last step, we use the fact that $\mathcal{A}_i^m$ is a realizable algorithm for $\mathcal{H}_i^m$ and the feedback that the algorithm received was $(x_t, h_i^{\star,m}(x_t))$ in the rounds whenever $B_t = 1$. Next, Lemma 5.17 from Ceccherini-Silberstein et al. [2017] guarantees that there exists a concave sub-linear function $\tilde{R}_i^m(|B|)$ that upperbounds $R_i^m(|B|)$. Thus, by Jensen's inequality, $\mathbb{E}_B\left[R_i^m(|B|)\right] \le \mathbb{E}_B\left[\tilde{R}_i^m(|B|)\right] \le \tilde{R}_i^m(\mathbb{E}_B\left[|B|\right]) \le \tilde{R}_i^m(T^\beta)$, a sub-linear function of T^β.

Combining (I) and (II) together, we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell(\mathcal{Q}(x_t), y_t)\right] \le \inf_{h\in\mathcal{H}}\sum_{t=1}^{T}\ell(h(x_t), y_t) + \frac{pM^2}{a}\sum_{i=1}^{K}\sum_{m=1}^{p}\frac{T}{T^\beta}\,\tilde{R}_i^m(T^\beta) + M\sqrt{2T^{1+\beta}K\ln K}.$$

Since $\tilde{R}_i^m(T^\beta)$ is a sublinear function of T^β, we have that $\frac{T}{T^\beta}\tilde{R}_i^m(T^\beta)$ is a sublinear function of T. As the sum of sublinear functions is sublinear, the second term above must be a sublinear function of T. The regret is sub-linear for any choice of $\beta \in (0,1)$. This completes our proof as we have shown that the algorithm $\mathcal{Q}$ achieves sub-linear regret in T. □

The proof of the necessity direction of Theorem 5.1 also involves constructing experts and running the REWA algorithm. Since the argument is similar, we defer details to Appendix D.1.

6 Discussion

In this paper, we characterize the learnability of a multilabel ranking hypothesis class in both the batch and online setting for a wide range of practical ranking losses. In all cases, we show that a ranking hypothesis class is learnable if and only if a sufficient number of its binary-valued threshold restrictions are learnable. Our paper studies two families of ranking loss functions and leaves it open to characterize the learnability of other natural ranking loss functions. One loss function not captured by our families is recall@p.

While we do establish quantitative bounds on the sample complexity and regret, our bounds are not optimal. It may be difficult to improve the sample complexity and regret bound at the highest level of generality for all losses in the families considered here. However, for natural losses such as sum loss, it is an interesting future direction to derive the optimal sample complexity and regret bounds in both the realizable and agnostic settings. In addition, our bounds depend on the number of labels K. Recently, K-free bounds have been achieved for multiclass classification problems in both batch and online settings [Brukhim et al., 2022, Hanneke et al., 2023]. An interesting future direction is to study whether K-free bounds are possible for multilabel ranking.

Finally, the focus of this paper is on characterizing learnability, and thus our algorithms are not computationally efficient. A natural future direction is to construct computationally efficient algorithms for multilabel ranking. Along this direction, since ERM is the most common algorithm used in practice, it is an important future direction to tightly quantify the sample complexity of ERM in the batch setting. Moreover, in learning theory, combinatorial dimensions play an important role in providing a tight quantitative characterization of learnability. Thus, it is an interesting future direction to identify combinatorial dimensions that characterize multilabel ranking learnability for specific loss functions.

Acknowledgements

We acknowledge the support of NSF via grant IIS-2007055. VR acknowledges the support of the NSF Graduate Research Fellowship.

References

Noga Alon, Amos Beimel, Shay Moran, and Uri Stemmer. Closure properties for private classification and online prediction. In *Conference on Learning Theory*, pages 119–152. PMLR, 2020.

Ivo M Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. Comparison of deep learning approaches for multi-label chest x-ray classification. *Scientific reports*, 9(1):1–10, 2019.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Nataly Brukhim, Daniel Carmon, Irit Dinur, Shay Moran, and Amir Yehudayoff. A characterization of multiclass learnability, 2022. URL https://arxiv.org/abs/2203.01550.

Serhat S Bucak, Pavan Kumar Mallapragada, Rong Jin, and Anil K Jain. Efficient multi-label ranking for multi-class learning: application to object recognition. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2098–2105. IEEE, 2009.

David Buffoni, Clément Calauzenes, Patrick Gallinari, and Nicolas Usunier. Learning scoring functions with order-preserving losses and standardized supervision. In *The 28th International Conference on Machine Learning (ICML 2011)*, pages 825–832, 2011.

Clément Calauzenes, Nicolas Usunier, and Patrick Gallinari. On the (non-) existence of convex, calibrated surrogate losses for ranking. *Advances in Neural Information Processing Systems*, 25, 2012.

T. Ceccherini-Silberstein, M. Salvatori, and E. Sava-Huss. *Groups, Graphs and Random Walks*. London Mathematical Society Lecture Note Series. Cambridge University Press, 2017. doi: 10.1017/9781316576571.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Amanda Clare and Ross D King. Knowledge discovery in multi-label phenotype data. In *Principles of Data Mining and Knowledge Discovery: 5th European Conference, PKDD 2001, Freiburg, Germany, September 3–5, 2001 Proceedings 5*, pages 42–53. Springer, 2001.

Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In *Conference on Learning Theory*, pages 287–316. PMLR, 2014.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. In Sham M. Kakade and Ulrike von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 207–232, Budapest, Hungary, 09–11 Jun 2011. PMLR.

Krzysztof Dembczynski, Wojciech Kotlowski, and Eyke Hüllermeier. Consistent multilabel ranking through univariate losses. *arXiv preprint arXiv:1206.6401*, 2012.

John C Duchi, Lester W Mackey, and Michael I Jordan. On the consistency of ranking algorithms. In *ICML*, pages 327–334, 2010.

Richard M Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, pages 899–929, 1978.

Wei Gao and Zhi-Hua Zhou. On the consistency of multi-label learning. In *Proceedings of the 24th annual conference on learning theory*, pages 341–358. JMLR Workshop and Conference Proceedings, 2011.

Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*, 2013.

Steve Hanneke, Shay Moran, Vinod Raman, Unique Subedi, and Ambuj Tewari. Multiclass online learning and uniform convergence. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5682–5696. PMLR, 2023.

Max Hopkins, Daniel M. Kane, Shachar Lovett, and Gaurav Mahajan. Realizable learning is all you need. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3015–3069. PMLR, 02–05 Jul 2022.

Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: ECML-98: 10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings*, pages 137–142. Springer, 2005.

Young Hun Jung and Ambuj Tewari. Online boosting algorithms for multi-label ranking. In *International Conference on Artificial Intelligence and Statistics*, pages 279–287. PMLR, 2018.

Anna Korba, Alexandre Garcia, and Florence d'Alché Buc. A structured prediction approach for label ranking. *Advances in Neural Information Processing Systems*, 31, 2018.

Oluwasanmi O Koyejo, Nagarajan Natarajan, Pradeep K Ravikumar, and Inderjit S Dhillon. Consistent multilabel classification. *Advances in Neural Information Processing Systems*, 28, 2015.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1987.

Tie-Yan Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

Andrew Kachites McCallum. Multi-label text classification with a mixture model trained by em. In *AAAI'99 workshop on text learning*, 1999.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability theory and related fields*, 161:111–153, 2015.

Vinod Raman, Unique Subedi, and Ambuj Tewari. A characterization of multioutput learnability. *arXiv cs.LG*, 2023. preprint arXiv:2303.17716.

Pradeep Ravikumar, Ambuj Tewari, and Eunho Yang. On ndcg consistency of listwise ranking methods. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 618–626. JMLR Workshop and Conference Proceedings, 2011.

Robert E Schapire and Yoram Singer. Boostexter: A boosting-based system for text categorization. *Machine learning*, 39:135–168, 2000.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014.

Leslie G. Valiant. A theory of the learnable. In *Symposium on the Theory of Computing*, 1984.

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition [in Russian]*. 1974.

Xi Wang and Gita Sukthankar. Multi-label relational neighbor classification using social context features. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 464–472, 2013.

Hao Yang, Joey Tianyi Zhou, Yu Zhang, Bin-Bin Gao, Jianxin Wu, and Jianfei Cai. Exploit bounding box annotations for multi-label object recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 280–288, 2016.

Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2013.

A Categorizing Popular Ranking Losses

Table 1: Categorizing Popular Ranking Losses.

Loss	Loss Family
Sum Loss@p	$\mathcal{L}(\ell_{\text{sum}}^{@p})$
Precision Loss@p	$\mathcal{L}(\ell_{\text{prec}}^{@p})$
Average Precision	$\mathcal{L}(\ell_{\text{sum}}^{@K})$
Area Under the Curve	$\mathcal{L}(\ell_{\text{sum}}^{@K})$
Reciprocal Rank	$\mathcal{L}(\ell_{\text{prec}}^{@1})$
Pairwise Rank Loss	$\mathcal{L}(\ell_{\text{sum}}^{@K})$
Discounted Cumulative Loss	$\mathcal{L}(\ell_{\text{sum}}^{@K})$
Discounted Cumulative Loss@p	$\mathcal{L}(\ell_{\text{sum}}^{@p})$

In this section, we show that our loss families $\mathcal{L}(\ell_{\text{sum}}^{@p})$ and $\mathcal{L}(\ell_{\text{prec}}^{@p})$ are general and capture many of the popular ranking loss functions used in practice. We summarize the results in Table 1.

Recall that

$$\mathcal{L}(\ell_{\text{sum}}^{@p}) = \{\ell \in \mathbb{R}^{\mathcal{S}_K \times \mathcal{Y}} : \ell = 0 \text{ if and only if } \ell_{\text{sum}}^{@p} = 0\} \cap \{\ell \in \mathbb{R}^{\mathcal{S}_K \times \mathcal{Y}} : \pi \overset{[p]}{=} \hat{\pi} \implies \ell(\pi, y) = \ell(\hat{\pi}, y)\},$$

where

$$\ell_{\text{sum}}^{@p}(\pi, y) = \sum_{i=1}^{K} \min(\pi_i, p+1) y^i - Z_y^p.$$

Note that the normalization constant is defined as $Z_y^p := \min_{\pi \in \mathcal{S}_K} \sum_{i=1}^{K} \min(\pi_i, p+1) y^i$ and thus only depends on y. Furthermore,

$$\mathcal{L}(\ell_{\text{prec}}^{@p}) = \{\ell \in \mathbb{R}^{\mathcal{S}_K \times \mathcal{Y}} : \ell = 0 \text{ if and only if } \ell_{\text{prec}}^{@p} = 0\} \cap \{\ell \in \mathbb{R}^{\mathcal{S}_K \times \mathcal{Y}} : \pi \overset{p}{=} \hat{\pi} \implies \ell(\pi, y) = \ell(\hat{\pi}, y)\}.$$

where

$$\ell_{\text{prec}}^{@p}(\pi, y) = Z_y^p - \sum_{i=1}^{K} \mathbb{1}\{\pi_i \leq p\} y^i.$$

As before, the normalization constant $Z_y^p := \max_{\pi \in \mathcal{S}_K} \sum_{i=1}^{K} \mathbb{1}\{\pi_i \leq p\} y^i$ only depends on y.

In ranking literature, many evaluation metrics are often stated in terms of *gain* functions. However, these can be easily converted into loss functions by subtracting the gain from the maximum possible value of the gain. When relevance scores are restricted to be binary (i.e. $\mathcal{Y} = \{0, 1\}^K$), the **Average Precision** (AP) metric is a *gain* function defined as

$$\text{AP}(\pi, y) = \frac{1}{\|y\|_1} \sum_{i \in \{\pi_m : y^m = 1\}} \frac{\sum_{j=1}^{K} \mathbb{1}\{\pi_j \leq i\} y^j}{i}.$$

Since the maximum value AP can take is 1, we can define its loss function variant as:

$$\ell_{\text{AP}}(\pi, y) = 1 - \text{AP}(\pi, y).$$

Note that $\ell_{\text{AP}}(\pi, y) = 0$ if and only if π ranks all labels where $y_i = 1$ in the top $\|y\|_1$. Therefore, $\ell_{\text{AP}}(\pi, y) \in \mathcal{L}(\ell_{\text{sum}}^{@K})$.

Another useful metric for binary relevance feedback is the **Area Under the Curve** (AUC) loss function:

$$\ell_{\text{AUC}}(\pi, y) = \frac{1}{\|y\|_1 (K - \|y\|_1)} \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbb{1}\{\pi_i < \pi_j\} \mathbb{1}\{y^i < y^j\}.$$

The AUC computes the fraction of "bad pairs" of labels (i.e those pairs of labels where i was more relevant than j, but i was ranked lower than j). Again, note that $\ell_{\text{AUC}}(\pi, y) = 0$ if and only if π ranks all labels where $y^i = 1$ in the top $\|y\|_1$. Therefore, $\ell_{\text{AP}}(\pi, y) \in \mathcal{L}(\ell_{\text{sum}}^{@K})$.

Lastly, the **Reciprocal Rank** (RR) metric is another important *gain* function for binary relevance score feedback,

$$\text{RR}(\pi, y) = \frac{1}{\min_{i:y^i=1} \pi_i}.$$

Its loss equivalent can be written as:

$$\ell_{\text{RR}}(\pi, y) = 1 - \text{RR}(\pi, y).$$

Since $\ell_{\text{RR}}(\pi, y)$ only cares about the relevance of the top-ranked label, we have that $\ell_{\text{RR}}(\pi, y) \in \mathcal{L}(\ell_{\text{prec}}^{@1})$.

Moving onto non-binary relevance scores, we start with the **Pairwise Rank Loss** (PL):

$$\ell_{\text{PL}}(\pi, y) = \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbb{1}\{\pi_i < \pi_j\} \mathbb{1}\{y^i < y^j\}.$$

The Pairwise Ranking loss is the analog of AUC for non-binary relevance scores and thus $\ell_{\text{PL}}(\pi, y) \in \mathcal{L}(\ell_{\text{sum}}^{@K})$.

Finally, we have the **Discounted Cumulative Gain** (DCG) metric, defined as:

$$\text{DCG}(\pi, y) = \sum_{i=1}^{K} \frac{2^{y^i} - 1}{\log_2(1 + \pi_i)}.$$

For an appropriately chosen normalizing constant Z_y, we can define its associated loss:

$$\ell_{\text{DCG}}(\pi, y) = Z_y - \text{DCG}(\pi, y).$$

Like $\ell_{\text{sum}}^{@K}$, $\ell_{\text{DCG}}(\pi, y)$ is 0 if and only if π ranks the K labels in increasing order of relevance, breaking ties arbitrarily. Thus, $\ell_{\text{DCG}}(\pi, y) \in \mathcal{L}(\ell_{\text{sum}}^{@K})$. If one only cares about the top-p ranked results, then the DCG@p loss function evaluates only the top-p ranked labels:

$$\ell_{\text{DCG}}^{@p}(\pi, y) = Z_y^p - \sum_{i=1}^{K} \frac{2^{y^i} - 1}{\log_2(1 + \pi_i)} \mathbb{1}\{\pi_i \leq p\} = Z_y^p - \text{DCG}^{@p}(\pi, y).$$

Analogously, we have that $\ell_{\text{DCG}}^{@p}(\pi, y) \in \mathcal{L}(\ell_{\text{sum}}^{@p})$.

B Agnostic PAC Learnability of Score-based Rankers

In this section, we apply our results in the main paper to give sufficient conditions for the agnostic PAC learnability of score-based ranking hypothesis classes. A score-based ranking hypothesis $h : \mathcal{X} \to \mathcal{S}_K$ first maps an input $x \in \mathcal{X}$ to a vector in $\mathbb{R}^K$ representing the "score" for each label. Then, it outputs a ranking (permutation) over the labels in $[K]$ by sorting the real-valued vector in decreasing order of score.

More formally, let $\mathcal{F} \subseteq (\mathbb{R}^K)^{\mathcal{X}}$ denote a set of functions mapping elements from the input space $\mathcal{X}$ to score-vectors in $\mathbb{R}^K$. For each $f \in \mathcal{F}$, define the score-based ranking hypothesis $h_f(x) = \text{argsort}(f(x))$ which first computes the score-vector $f(x) \in \mathbb{R}^K$, and then outputs a ranking by sorting $f(x)$ in decreasing order, breaking ties by giving the smaller label the higher rank. That is, if $f_1(x) = f_2(x)$, then label 1 will be ranked higher than label 2. Given $\mathcal{F}$, define its induced score-based ranking hypothesis class as $\mathcal{H} = \{h_f : f \in \mathcal{F}\}$. Since our characterization of ranking learnability relates the learnability of $\mathcal{H}$ to the learnability of the *binary* threshold-restricted classes $\mathcal{H}_i^j = \{h_i^j : h \in \mathcal{H}\}$, it suffices to consider an arbitrary threshold-restricted class $\mathcal{H}_i^j$ and bound its VC dimension. Before we do so, we need some more notation regarding $\mathcal{F}$.

For each $k \in [K]$, define the scalar-valued function class $\mathcal{F}_k = \{f_k \mid (f_1, \ldots, f_K) \in \mathcal{F}\}$ by restricting each function in $\mathcal{F}$ to its k^{th} coordinate output. Here, each $\mathcal{F}_k \subseteq \mathbb{R}^{\mathcal{X}}$ and we can write

$\mathcal{F} = (\mathcal{F}_1, \ldots, \mathcal{F}_K)$. For a function $f \in \mathcal{F}$, we will use $f_k(x)$ to denote the k^{th} coordinate output of $f(x)$. For every $(i, j) \in [K] \times [K]$, define the function class $\mathcal{F}_i - \mathcal{F}_j = \{f_i - f_j : f \in \mathcal{F}\}$ where we let $f_i - f_j : \mathcal{X} \to \mathbb{R}$ denote a function such that $(f_i - f_j)(x) = f_i(x) - f_j(x)$. Subsequently, for any $(i, j) \in [K] \times [K]$, define the *binary* hypothesis classes $\mathcal{G}_{i,j} = \{\mathbb{1}\{(f_i - f_j)(x) < 0\} : f_i - f_j \in \mathcal{F}_i - \mathcal{F}_j\}$ and $\tilde{\mathcal{G}}_{i,j} = \{\mathbb{1}\{(f_i - f_j)(x) \leq 0\} : f_i - f_j \in \mathcal{F}_i - \mathcal{F}_j\}$. Finally, let $C_j : \{0,1\}^K \to \{0,1\}$ be the K-wise composition s.t. $C_j(b) = \mathbb{1}\{\sum_{i=1}^K b_i \leq j\}$ and define $C_j(\mathcal{G}_1, ..., \mathcal{G}_K) = \{C_j(g_1, ..., g_K) : (g_1, ..., g_K) \in \mathcal{G}_1 \times ... \times \mathcal{G}_K\}$. In other words, $C_j(\mathcal{G}_1, ..., \mathcal{G}_K)$ is the *binary* hypothesis class constructed by taking all combinations of binary classifiers from $\mathcal{G}_1, ..., \mathcal{G}_K$, summing them up, and thresholding the sum at j. We are now ready to bound the VC dimension of an arbitrary threshold-restricted class $\mathcal{H}_i^j$.

Consider an arbitrary threshold-restricted class $\mathcal{H}_i^j$ and hypothesis $h \in \mathcal{H}$. By definition, $h_i^j \in \mathcal{H}_i^j$. Let $f \in \mathcal{F}$ denote the function associated with h. Given an instance $x \in \mathcal{X}$, recall that $h_i^j(x) = \mathbb{1}\{h_i(x) \leq j\}$ where $h_i(x)$ is the rank that h gives to the label i for instance x. Since $h(x) = \text{argsort}(f(x))$, we have

$$\begin{aligned} h_i(x) &= \text{argsort}(f(x))[i] \\ &= \sum_{m=1}^{i} \mathbb{1}\{f_i(x) \leq f_m(x)\} + \sum_{m=i+1}^{K} \mathbb{1}\{f_i(x) < f_m(x)\} \\ &= \sum_{m=1}^{i} \mathbb{1}\{(f_i - f_m)(x) \leq 0\} + \sum_{m=i+1}^{K} \mathbb{1}\{(f_i - f_m)(x) < 0\} \end{aligned}$$

Thus, we can write:

$$h_i^j(x) = \mathbb{1}\left\{\left(\sum_{m=1}^{i} \mathbb{1}\{(f_i - f_m)(x) \leq 0\} + \sum_{m=i+1}^{K} \mathbb{1}\{(f_i - f_m)(x) < 0\}\right) \leq j\right\}.$$

Note that $h_i^j \in C_j(\tilde{\mathcal{G}}_{i,1}, ..., \tilde{\mathcal{G}}_{i,i}, \mathcal{G}_{i,i+1}, ..., \mathcal{G}_{i,K})$ by construction. Since h, and therefore h_i^j, was arbitrary, it further follows that $\mathcal{H}_i^j \subseteq C_j(\tilde{\mathcal{G}}_{i,1}, ..., \tilde{\mathcal{G}}_{i,i}, \mathcal{G}_{i,i+1}, ..., \mathcal{G}_{i,K})$. Therefore,

$$\text{VC}(\mathcal{H}_i^j) \leq \text{VC}(C_j(\tilde{\mathcal{G}}_{i,1}, ..., \tilde{\mathcal{G}}_{i,i}, \mathcal{G}_{i,i+1}, ..., \mathcal{G}_{i,K})).$$

Since $C_j(\tilde{\mathcal{G}}_{i,1}, ..., \tilde{\mathcal{G}}_{i,i}, \mathcal{G}_{i,i+1}, ..., \mathcal{G}_{i,K})$ is some K-wise composition of binary classes $\tilde{\mathcal{G}}_{i,1}, ..., \tilde{\mathcal{G}}_{i,i}, \mathcal{G}_{i,i+1}, ..., \mathcal{G}_{i,K}$, standard VC composition guarantees that $\text{VC}(C_j(\tilde{\mathcal{G}}_{i,1}, ..., \tilde{\mathcal{G}}_{i,i}, \mathcal{G}_{i,i+1}, ..., \mathcal{G}_{i,K})) = \tilde{O}(\text{VC}(\tilde{\mathcal{G}}_{i,1}) + ... + \text{VC}(\tilde{\mathcal{G}}_{i,i}) + \text{VC}(\mathcal{G}_{i,i+1}) + ... + \text{VC}(\mathcal{G}_{i,K}))$, where we hide log factors of K and the VC dimensions [Dudley, 1978, Alon et al., 2020]. Putting things together, we have that

$$\text{VC}(\mathcal{H}_i^j) \leq \tilde{O}(\text{VC}(\tilde{\mathcal{G}}_{i,1}) + ... + \text{VC}(\tilde{\mathcal{G}}_{i,i}) + \text{VC}(\mathcal{G}_{i,i+1}) + ... + \text{VC}(\mathcal{G}_{i,K})).$$

An identical analysis can also be used to give sufficient conditions for the *online* learnability of score-based rankers in terms of the Littlestone dimensions of $\mathcal{H}_j^i$.

Now, we consider the special class of *linear* score-based ranker and prove Lemma 4.6.

Proof. (of Lemma 4.6) Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{F} = \{f_W : W \in \mathbb{R}^{K \times d}\}$ s.t. $f_W(x) = Wx$. Consider the class of linear score-based rankers $\mathcal{H} = \{h_{f_W} : f_W \in \mathcal{F}\}$ where $h_{f_W}(x) = \text{argsort}(f_W(x)) = \text{argsort}(Wx)$ breaking ties in the same way mentioned above. Note for all $i \in [K]$, $\mathcal{F}_i = \{f_w : w \in \mathbb{R}^d\}$ where $f_w(x) = w^T x$. Furthermore, $\mathcal{F}_i - \mathcal{F}_j = \mathcal{F}_i = \mathcal{F}_j$. Therefore, for any $(i, j) \in [K] \times [K]$,

$$\mathcal{G}_{i,j} = \{\mathbb{1}\{(f_i - f_j)(x) < 0\} : f_i - f_j \in \mathcal{F}_i - \mathcal{F}_j\} = \{\mathbb{1}\{f_w(x) < 0\} : w \in \mathbb{R}^d\}$$

and

$$\tilde{\mathcal{G}}_{i,j} = \{\mathbb{1}\{(f_i - f_j)(x) \leq 0\} : f_i - f_j \in \mathcal{F}_i - \mathcal{F}_j\} = \{\mathbb{1}\{f_w(x) \leq 0\} : w \in \mathbb{R}^d\}$$

are the set of half-space classifiers passing through the origin with dimension d. Since for all $(i,j) \in [K] \times [K]$, $\text{VC}(\tilde{\mathcal{G}}_{i,j}) = \text{VC}(\mathcal{G}_{i,j}) = d$, we get that $\text{VC}(\mathcal{H}_i^j) \leq \tilde{O}(Kd)$. □

C Proofs for Batch Multilabel Ranking

Since many of the ranking losses we consider map to values in $\mathbb{R}$, the *empirical* Rademacher complexity will be a useful tool for proving learnability in the batch setting.

Definition 3 (Empirical Rademacher Complexity of Loss Class)**.** *Let $\ell(\cdot,\cdot)$ be a loss function, $S = \{(x_1,y_1),...,(x_n,y_n)\} \in (\mathcal{X} \times \mathcal{Y})^*$ be a set of examples, and $\ell \circ \mathcal{H} = \{(x,y) \mapsto \ell(h(x),y) : h \in \mathcal{H}\}$ be a loss class. The empirical Rademacher complexity of $\ell \circ \mathcal{H}$ is defined as*

$$\hat{\mathfrak{R}}_n(\ell \circ \mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(h(x_i), y_i) \right) \right]$$

where $\sigma_1, ..., \sigma_n$ are independent Rademacher *random variables.*

In particular, a standard result relates the empirical Rademacher complexity to the generalization error of hypotheses in $\mathcal{H}$ with respect to a real-valued bounded loss function $\ell(h(x),y)$ [Bartlett and Mendelson, 2002].

Proposition C.1 (Rademacher-based Uniform Convergence)**.** *Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$ and $\ell(\cdot,\cdot) \leq c$ be a bounded loss function. With probability at least $1-\delta$ over the sample $S \sim \mathcal{D}^n$, for all $h \in \mathcal{H}$ simultaneously,*

$$\left| \mathbb{E}_{\mathcal{D}}[\ell(h(x),y)] - \hat{\mathbb{E}}_S[\ell(h(x),y)] \right| \leq 2\hat{\mathfrak{R}}_n(\mathcal{F}) + O\left(c\sqrt{\frac{\ln(\frac{1}{\delta})}{n}} \right)$$

where $\hat{\mathbb{E}}_S[\ell(h(x),y)] = \frac{1}{|S|} \sum_{(x,y) \in S} \ell(h(x),y)$ is the empirical average of the loss over S.

When the empirical Rademacher complexity of the loss class $\ell \circ \mathcal{H} = \{(x,y) \mapsto \ell(h(x),y) : h \in \mathcal{H}\}$ is $o(1)$, we state that $\mathcal{H}$ enjoys the uniform convergence property w.r.t ℓ. If $\mathcal{H}$ enjoys the uniform convergence property w.r.t. a loss ℓ, a standard result shows that $\mathcal{H}$ is learnable according to Definition 1 via Empirical Risk Minimization (ERM) (Theorem 26.5 in Shalev-Shwartz and Ben-David [2014]).

C.1 Proof of Lemma 4.3

Proof. Let $\mathcal{H} \subseteq \mathcal{S}_K^{\mathcal{X}}$ be an arbitrary ranking hypothesis class. We need to show that if $\mathcal{H}_i^j$ is agnostic PAC learnable w.r.t to 0-1 loss for all $(i,j) \in [K] \times [p]$, then ERM is an agnostic PAC learnable w.r.t $\ell_{\text{sum}}^{@p}$. By Proposition C.1, it suffices to show that the empirical Rademacher complexity of the loss class $\ell_{\text{sum}}^{@p} \circ \mathcal{H}$ vanishes as n increases. This will imply that $\ell_{\text{sum}}^{@p}$ enjoys the uniform convergence property, and therefore ERM is an agnostic PAC learner for $\mathcal{H}$ w.r.t $\ell_{\text{sum}}^{@p}$. By definition, we have that

$$
\begin{aligned}
\hat{\mathfrak{R}}_n(\ell_{\text{sum}}^{@p} \circ \mathcal{H}) &= \mathbb{E}_{\sigma\sim\{\pm1\}^n}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell_{\text{sum}}^{@p}(h(x_i), y_i))\right] \\
&= \mathbb{E}_{\sigma\sim\{\pm1\}^n}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \left(\sum_{m=1}^{K} \sigma_i \min(h_m(x_i), p+1) y_i^m - \sigma_i Z_{y_i}^p\right)\right] \\
&= \mathbb{E}_{\sigma\sim\{\pm1\}^n}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{K} \sigma_i \min(h_m(x_i), p+1) y_i^m\right] \\
&\leq \sum_{m=1}^{K} \mathbb{E}_{\sigma\sim\{\pm1\}^n}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \min(h_m(x_i), p+1) y_i^m\right] \\
&\leq B\sum_{m=1}^{K} \mathbb{E}_{\sigma\sim\{\pm1\}^n}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \min(h_m(x_i), p+1)\right]
\end{aligned}
$$

where the second inequality follows from the fact that $y_i^m \leq B$ and Talagrand's Contraction Lemma Ledoux and Talagrand [1991].

Next note that $\min(h_m(x_i), p+1) = (p+1) - \sum_{j=1}^{p} \mathbb{1}\{h_m(x_i) \leq j\} = (p+1) - \sum_{j=1}^{p} h_m^j(x_i)$. Substituting and getting rid of constant factors, we have that

$$
\begin{aligned}
\hat{\mathfrak{R}}_n(\ell_{\text{sum}}^{@p} \circ \mathcal{H}) &\leq B\sum_{m=1}^{K} \mathbb{E}_{\sigma\sim\{\pm1\}^n}\left[\sup_{h_m\in\mathcal{H}_m} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \sum_{j=1}^{p} h_m^j(x_i)\right] \\
&\leq B\sum_{m=1}^{K}\sum_{j=1}^{p} \mathbb{E}_{\sigma\sim\{\pm1\}^n}\left[\sup_{h_m\in\mathcal{H}_m} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h_m^j(x_i)\right] \\
&= B\sum_{m=1}^{K}\sum_{j=1}^{p} \hat{\mathfrak{R}}_n(\mathcal{H}_m^j).
\end{aligned}
$$

Since for $\mathcal{H}_m^j$ is agnostic PAC learnable w.r.t 0-1 loss, by Theorem 6.5 in Shalev-Shwartz and Ben-David [2014], $\lim_{n\to\infty} \hat{\mathfrak{R}}_n(\mathcal{H}_m^j) = 0$. Since p, K and B are finite,

$$
\lim_{n\to\infty} \hat{\mathfrak{R}}_n(\ell_{\text{sum}}^{@p} \circ \mathcal{H}) = \lim_{n\to\infty} B\sum_{m=1}^{K}\sum_{j=1}^{p} \hat{\mathfrak{R}}_n(\mathcal{H}_m^j) = 0
$$

.

By Proposition C.1, this implies that $\ell_{\text{sum}}^{@p}$ enjoys the uniform convergence property, and therefore ERM using $\ell_{\text{sum}}^{@p}$ is an agnostic PAC learner for $\mathcal{H}$. □

C.2 Proof of Lemma 4.5

Proof. Fix $\ell \in \mathcal{L}(\ell_{\text{sum}}^{@p})$ and $(i, j) \in [K] \times [p]$. Let $a = \min_{\pi,y}\{\ell(\pi, y) \mid \ell(\pi, y) \neq 0\}$. Let $\mathcal{H}$ be an arbitrary ranking hypothesis class and $\mathcal{A}$ be an agnostic PAC learner for $\mathcal{H}$ w.r.t ℓ. Our goal will be to use $\mathcal{A}$ to construct an agnostic PAC learner for $\mathcal{H}_i^j$.

Let $\mathcal{D}$ be distribution over $\mathcal{X} \times \{0, 1\}$ and $h_i^{\star,j} = \arg\min_{h_i^j\in\mathcal{H}_i^j} \mathbb{E}_{\mathcal{D}}\left[\mathbb{1}\{h_i^j(x) \neq y\}\right]$ be the optimal hypothesis. Let $h^\star \in \mathcal{H}$ be any valid completion of $h_i^{\star,j}$. Our goal will be to show that Algorithm 4 is an agnostic PAC learner for $\mathcal{H}_i^j$ w.r.t 0-1 loss.

Consider the sample $S_U^{h^\star}$ and let $g = \mathcal{A}(S_U^{h^\star})$. We can think of g as the output of $\mathcal{A}$ run over an i.i.d sample S drawn from $\mathcal{D}^\star$, a joint distribution over $\mathcal{X} \times \mathcal{Y}$ defined procedurally by first

Algorithm 4 Agnostic PAC learner for $\mathcal{H}_i^j$ w.r.t. 0-1 loss

Input: Agnostic PAC learner $\mathcal{A}$ for $\mathcal{H}$ w.r.t ℓ, unlabeled samples $S_U \sim \mathcal{D}_{\mathcal{X}}^n$, and labeled samples $S_L \sim \mathcal{D}^m$

1 For each $h \in \mathcal{H}_{|S_U}$, construct a dataset

$$S_U^h = \{(x_1, \tilde{y}_1), ..., (x_n, \tilde{y}_n)\} \text{ s.t. } \tilde{y}_i = \text{BinRel}(h(x_i), j)$$

2 Run $\mathcal{A}$ over all datasets to get $C(S_U) := \left\{\mathcal{A}\left(S_U^h\right) \mid h \in \mathcal{H}_{|S_U}\right\}$

3 Define $C_i^j(S_U) = \{g_i^j | g \in C(S_U)\}$

4 Return $\hat{g}_i^j \in C_i^j(S_U)$ with the lowest empirical error over S_L w.r.t. 0-1 loss.

sampling $x \sim \mathcal{D}_{\mathcal{X}}$ and then outputting the labeled sample $(x, \text{BinRel}(h^\star(x), j))$. Note that $\mathcal{D}^\star$ is a realizable distribution (realized by $h^\star$) w.r.t $\ell_{\text{sum}}^{@p}$ and therefore also ℓ. Let $m_{\mathcal{A}}(\epsilon, \delta, K)$ be the sample complexity of $\mathcal{A}$. Since $\mathcal{A}$ is an agnostic PAC learner for $\mathcal{H}$ w.r.t ℓ, we have that for sample size $n \geq m_{\mathcal{A}}(\frac{a\epsilon}{2}, \delta/2, K)$, with probability at least $1 - \frac{\delta}{2}$,

$$\mathbb{E}_{\mathcal{D}^\star}\left[\ell(g(x), y)\right] \leq \inf_{h \in \mathcal{H}} \mathbb{E}_{\mathcal{D}^\star}\left[\ell(h(x), y)\right] + \frac{a\epsilon}{2} = \frac{a\epsilon}{2}.$$

Furthermore, by definition of $\mathcal{D}^\star$, $\mathbb{E}_{\mathcal{D}^\star}\left[\ell(g(x), y)\right] = \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}}\left[\ell(g(x), \text{BinRel}(h^\star(x), j))\right]$. Therefore, $\mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}}\left[\ell(g(x), \text{BinRel}(h^\star(x), j))\right] \leq \frac{a\epsilon}{2}$. Next, using Lemma E.3, we have pointwise that

$$\begin{aligned}
\mathbb{1}\{g_i^j(x) \neq h_i^{\star,j}(x)\} &\leq \mathbb{1}\{\ell_{\text{sum}}^{@p}(g(x), \text{BinRel}(h^\star(x), j)) > 0\} \\
&= \mathbb{1}\{\ell(g(x), \text{BinRel}(h^\star(x), j)) > 0\} \\
&\leq \frac{1}{a}\,\ell(g(x), \text{BinRel}(h^\star(x), j)).
\end{aligned}$$

Taking expectations on both sides gives,

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{1}\{g_i^j(x) \neq h_i^{\star,j}(x)\}\right] \leq \frac{1}{a}\,\mathbb{E}_{\mathcal{D}}\left[\ell(g(x), \text{BinRel}(h^\star(x), j))\right] \leq \frac{\epsilon}{2},$$

where in the last inequality we use the fact that $\mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}}\left[\ell(g(x), \text{BinRel}(h^\star(x), j))\right] \leq \frac{a\epsilon}{2}$. Finally, using the triangle inequality, we have that

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[\mathbb{1}\{g_i^j(x) \neq y\}\right] &\leq \mathbb{E}_{\mathcal{D}}\left[\mathbb{1}\{h_i^{\star,j}(x) \neq y\}\right] + \mathbb{E}_{\mathcal{D}}\left[\mathbb{1}\{g_i^j(x) \neq h_i^{\star,j}(x\}\right] \\
&\leq \mathbb{E}_{\mathcal{D}}\left[\mathbb{1}\{h_i^{\star,j}(x) \neq y\}\right] + \frac{\epsilon}{2} \\
&= \arg\min_{h_i^j \in \mathcal{H}_i^j} \mathbb{E}_{\mathcal{D}}\left[\mathbb{1}\{h_i^j(x) \neq y\}\right] + \frac{\epsilon}{2}.
\end{aligned}$$

Since $g_i^j \in C_i^j(S_U)$, we have shown that $C_i^j(S_U)$ contains a hypothesis that generalizes well w.r.t $\mathcal{D}$. Now we want to show that the predictor $\hat{g}_i^j$ returned in step 4 also generalizes well. Crucially, observe that $C_i^j(S_U)$ is a finite hypothesis class with cardinality at most K^{jn}. Therefore, by standard Chernoff and union bounds, with probability at least $1 - \delta/2$, the empirical risk of every hypothesis in $C_i^j(S_U)$ on a sample of size $\geq \frac{8}{\epsilon^2}\log\frac{4|C_i^j(S_U)|}{\delta}$ is at most $\epsilon/4$ away from its true error. So, if $m = |S_L| \geq \frac{8}{\epsilon^2}\log\frac{4|C_i^j(S_U)|}{\delta}$, then with probability at least $1 - \delta/2$, we have

$$\frac{1}{|S_L|}\sum_{(x,y) \in S_L} \mathbb{1}\{g_i^j(x) \neq y\} \leq \mathbb{E}_{\mathcal{D}}\left[\mathbb{1}\{g_i^j(x) \neq y\}\right] + \frac{\epsilon}{4} \leq \frac{3\epsilon}{4}.$$

Since $\hat{g}_i^j$ is the ERM on S_L over $C_i^j(S_U)$, its empirical risk can be at most $\frac{3\epsilon}{4}$. Given that the population risk of $\hat{g}_i^j$ can be at most $\epsilon/4$ away from its empirical risk, we have that

$$\mathbb{E}_{\mathcal{D}}[\mathbb{1}\{\hat{g}_i^j(x) \neq y\}] \leq \arg\min_{h_i^j \in \mathcal{H}_i^j} \mathbb{E}_{\mathcal{D}}\left[\mathbb{1}\{h_i^j(x) \neq y\}\right] + \epsilon.$$

Applying union bounds, the entire process succeeds with probability $1-\delta$. We can compute the upper bound on the sample complexity of Algorithm 4, denoted $n(\epsilon, \delta, K)$, as

$$\begin{aligned} n(\epsilon, \delta, K) &\leq m_{\mathcal{A}}(\frac{a\epsilon}{2}, \delta/2, K) + O\left(\frac{1}{\epsilon^2}\log\frac{|C(S_U)|}{\delta}\right) \\ &\leq m_{\mathcal{A}}(\frac{a\epsilon}{2}, \delta/2, K) + O\left(\frac{K m_{\mathcal{A}}(\frac{a\epsilon}{2}, \delta/2, K) + \log\frac{1}{\delta}}{\epsilon^2}\right), \end{aligned}$$

where we use $|C(S_U)| \leq 2^{K m_{\mathcal{A}}(\frac{a\epsilon}{2}, \delta/2, K)}$. This shows that Algorithm 4 is an agnostic PAC learner for $\mathcal{H}_i^j$ w.r.t 0-1 loss. Since our choice of loss $\ell \in \mathcal{L}(\ell_{\text{sum}}^{@p})$ and indices (i, j) were arbitrary, agnostic PAC learnability of $\mathcal{H}$ w.r.t ℓ implies agnostic PAC learnability of $\mathcal{H}_i^j$ w.r.t the 0-1 loss for all $(i, j) \in [K] \times [p]$. □

C.3 Characterizing Batch Learnability of $\mathcal{L}(\ell_{\textbf{prec}}^{@p})$

In this section, we prove Theorem 4.2 which characterizes the agnostic PAC learnability of an arbitrary hypothesis class $\mathcal{H} \subseteq \mathcal{S}_K^{\mathcal{X}}$ w.r.t losses in $\mathcal{L}(\ell_{\text{prec}}^{@p})$. Our proof will again be in three parts. First, we will show that if for all $i \in [K]$, $\mathcal{H}_i^p$ is agnostic PAC learnable w.r.t the 0-1 loss, then ERM is an agnostic PAC learnable w.r.t $\ell_{\text{prec}}^{@p}$. Next, we show that if $\mathcal{H}$ is agnostic PAC learnable w.r.t $\ell_{\text{prec}}^{@p}$, then $\mathcal{H}$ is agnostic PAC learnable w.r.t any loss $\ell \in \mathcal{L}(\ell_{\text{prec}}^{@p})$. Finally, we prove the necessity direction - if $\mathcal{H}$ is agnostic PAC learnable w.r.t an arbitrary $\ell \in \mathcal{L}(\ell_{\text{prec}}^{@p})$, then for all $i \in [K]$, $\mathcal{H}_i^p$ is agnostic PAC learnable w.r.t the 0-1 loss.

We begin with Lemma C.2 which asserts that if for all $i \in [K]$, $\mathcal{H}_i^p$ is agnostic PAC learnable, then ERM is an agnostic PAC learner for $\mathcal{H}$ w.r.t $\ell_{\text{prec}}^{@p}$.

Lemma C.2. *If for all $i \in [K]$, $\mathcal{H}_i^p$ is agnostic PAC learnable w.r.t the 0-1 loss, then ERM is an agnostic PAC learner for $\mathcal{H} \subseteq \mathcal{S}_K^{\mathcal{X}}$ w.r.t $\ell_{prec}^{@p}$*

The proof of Lemma C.2 is similar to the proof of Lemma 4.3 and involves bounding the empirical Rademacher complexity of the loss class $\ell_{\text{prec}}^{@p} \circ \mathcal{H}$. This will imply that $\ell_{\text{prec}}^{@p}$ enjoys the uniform convergence property, and therefore ERM is an agnostic PAC learner for $\mathcal{H}$ w.r.t $\ell_{\text{prec}}^{@p}$. The key insight is that we can write $\ell_{\text{prec}}^{@p}(h(x), y) = Z_y^p - \sum_{i=1}^K \mathbb{1}\{h_i(x) \leq p\} y^i = Z_y^p - \sum_{i=1}^K h_i^p(x) y^i$. Since Z_y^p does not depend on $h(x)$ and $y^i \leq B$, we can upperbound the empirical Rademacher complexity in terms of the empirical Rademacher complexities of $\mathcal{H}_i^p$ using Talagrand's contraction.

Proof. Let $\mathcal{H} \subseteq \mathcal{S}_K^{\mathcal{X}}$ be an arbitrary ranking hypothesis class. Similar to the proof of Lemma 4.3, it suffices to show that the empirical Rademacher complexity of the loss class $\ell_{\text{prec}}^{@p} \circ \mathcal{H}$ vanishes. By Proposition C.1, this will imply that $\ell_{\text{prec}}^{@p}$ enjoys the uniform convergence property, and therefore

ERM is an agnostic PAC learner for $\mathcal{H}$ w.r.t $\ell_{\text{prec}}^{@p}$. By definition, we have that

$$\begin{aligned}
\hat{\mathfrak{R}}_n(\ell_{\text{prec}}^{@p} \circ \mathcal{H}) &= \mathbb{E}_{\sigma\sim\{\pm1\}^n}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell_{\text{prec}}^{@p}(h(x_i), y_i))\right] \\
&= \mathbb{E}_{\sigma\sim\{\pm1\}^n}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \left(\sigma_i Z_{y_i}^p - \sum_{m=1}^{K} \sigma_i \mathbb{1}\{h_m(x_i) \le p\} y_i^m\right)\right] \\
&= \mathbb{E}_{\sigma\sim\{\pm1\}^n}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{K} \sigma_i h_m^p(x_i) y_i^m\right] \\
&\le \sum_{m=1}^{K} \mathbb{E}_{\sigma\sim\{\pm1\}^n}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h_m^p(x_i) y_i^m\right] \\
&\le B\sum_{m=1}^{K} \mathbb{E}_{\sigma\sim\{\pm1\}^n}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h_m^p(x_i)\right] \\
&= B\sum_{m=1}^{K} \hat{\mathfrak{R}}_n(\mathcal{H}_m^p),
\end{aligned}$$

where the second inequality follows from Talagrand's Contraction Lemma and the fact that $y_i^m \le B$ for all i, m. Since for all $m \in [K]$, $\mathcal{H}_m^p$ is agnostic PAC learnable w.r.t 0-1 loss, by Theorem 6.7 in Shalev-Shwartz and Ben-David [2014], $\lim_{n\to\infty} \hat{\mathfrak{R}}_n(\mathcal{H}_m^p) = 0$. Since K and B are finite,

$$\lim_{n\to\infty} \hat{\mathfrak{R}}_n(\ell_{\text{prec}}^{@p} \circ \mathcal{H}) = \lim_{n\to\infty} B\sum_{m=1}^{K} \hat{\mathfrak{R}}_n(\mathcal{H}_m^p) = 0$$

.

By Proposition C.1, this implies that $\ell_{\text{prec}}^{@p}$ enjoys the uniform convergence property, and therefore ERM using $\ell_{\text{prec}}^{@p}$ is an agnostic PAC learner for $\mathcal{H}$. □

Next, Lemma C.3 extends the learnability of $\ell_{\text{prec}}^{@p}$ to the learnability of any loss $\ell \in \mathcal{L}(\ell_{\text{prec}}^{@p})$. In particular, Lemma C.3 asserts that if $\mathcal{H}$ is agnostic PAC learnable w.r.t $\ell_{\text{prec}}^{@p}$ then $\mathcal{H}$ is also agnostic PAC learnable w.r.t any $\ell \in \mathcal{L}(\ell_{\text{prec}}^{@p})$.

Lemma C.3. *If a hypothesis class $\mathcal{H} \subseteq \mathcal{S}_K^{\mathcal{X}}$ is agnostic PAC learnable w.r.t $\ell_{prec}^{@p}$, then $\mathcal{H}$ is agnostic PAC learnable w.r.t any $\ell \in \mathcal{L}(\ell_{prec}^{@p})$.*

The proof of Lemma C.3 follows the same the exact same strategy used in proving Lemma 4.4. More specifically, given an agnostic PAC learner $\mathcal{A}$ for $\mathcal{H}$ w.r.t. $\ell_{\text{prec}}^{@p}$, we first create a *realizable* PAC learner for $\mathcal{H}$ w.r.t $\ell \in \mathcal{L}(\ell_{\text{prec}}^{@p})$. Then, we use a similar realizable-to-agnostic conversion technique as in the proof of Lemma 4.4 to convert the realizable PAC learner into an agnostic PAC learner for $\mathcal{H}$ w.r.t ℓ.

Proof. Fix $\ell \in \mathcal{L}(\ell_{\text{prec}}^{@p})$. Let $a = \min_{\pi,y}\{\ell(\pi, y) \mid \ell(\pi, y) \neq 0\}$ and $b = \max_{\pi,y} \ell(\pi, y)$. We need to show that if $\mathcal{H}$ is agnostic PAC learnable w.r.t $\ell_{\text{prec}}^{@p}$, then $\mathcal{H}$ is agnostic PAC learnable w.r.t ℓ. We will do so in two steps. First, we will show that if $\mathcal{A}$ is an agnostic PAC learner for $\mathcal{H}$ w.r.t. $\ell_{\text{prec}}^{@p}$, then $\mathcal{A}$ is also a *realizable* PAC learner for $\mathcal{H}$ w.r.t ℓ. Next, we will show how to convert the realizable PAC learner w.r.t ℓ into an agnostic PAC learner w.r.t ℓ in a black-box fashion. The composition of these two pieces yields an agnostic PAC learner for $\mathcal{H}$ w.r.t ℓ.

If $\mathcal{H}$ is agnostic PAC learnable w.r.t $\ell_{\text{prec}}^{@p}$, then there exists a learning algorithm $\mathcal{A}$ with sample complexity $m(\epsilon, \delta, K)$ s.t. for any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, with probability $1 - \delta$ over a sample $S \sim \mathcal{D}^n$ of size $n \ge m(\epsilon, \delta, K)$, the output $g = \mathcal{A}(S)$ achieves

$$\mathbb{E}_{\mathcal{D}}\left[\ell^{@p}_{\text{prec}}(g(x),y))\right] \leq \inf_{h\in\mathcal{H}} \mathbb{E}_{\mathcal{D}}\left[\ell^{@p}_{\text{prec}}(h(x),y))\right] + \epsilon.$$

If $\mathcal{D}$ is realizable w.r.t ℓ, then we are guaranteed that there exists a hypothesis $h^\star \in \mathcal{H}$ s.t. $\mathbb{E}_{\mathcal{D}}\left[\ell(h^\star(x),y)\right] = 0$. Since $\ell \in \mathcal{L}(\ell^{@p}_{\text{prec}})$, this also means that $\mathbb{E}_{\mathcal{D}}\left[\ell^{@p}_{\text{prec}}(h^\star(x),y)\right] = 0$. Furthermore, since $\ell \in \mathcal{L}(\ell^{@p}_{\text{prec}})$, $\ell \leq b\ell^{@p}_{\text{prec}}$. Together, this means we have $\mathbb{E}_{\mathcal{D}}\left[\ell(g(x),y)\right] \leq b\epsilon$ showing have that $\mathcal{A}$ is also a realizable PAC learner for $\mathcal{H}$ w.r.t ℓ with sample complexity $m(\frac{\epsilon}{b},\delta,K)$. This completes the first part of the proof.

Now, we show how to convert the realizable PAC learner $\mathcal{A}$ for ℓ into an agnostic PAC learner for ℓ in a black-box fashion. For this step, we will use a similar algorithm as in the proof of Lemma 4.4. That is, we will show that Algorithm 5 below is an agnostic PAC learner for $\mathcal{H}$ w.r.t ℓ.

Algorithm 5 Agnostic PAC learner for $\mathcal{H}$ w.r.t. ℓ

Input: Realizable PAC learner $\mathcal{A}$ for $\mathcal{H}$ w.r.t ℓ, unlabeled samples $S_U \sim \mathcal{D}^n_{\mathcal{X}}$, and labeled samples $S_L \sim \mathcal{D}^m$

For each $h \in \mathcal{H}_{|S_U}$, construct a dataset

$$S^h_U = \{(x_1,\tilde{y}_1),...,(x_n,\tilde{y}_n)\} \text{ s.t. } \tilde{y}_i = \text{BinRel}(h(x_i),p)$$

Run $\mathcal{A}$ over all datasets to get $C(S_U) := \left\{\mathcal{A}\left(S^h_U\right) \mid h \in \mathcal{H}_{|S_U}\right\}$

Return $\hat{g} \in C(S_U)$ with the lowest empirical error over S_L w.r.t. ℓ.

Let $\mathcal{D}$ be any (not necessarily realizable) distribution over $\mathcal{X}\times\mathcal{Y}$. Let $h^\star = \arg\min_{h\in\mathcal{H}} \mathbb{E}_{\mathcal{D}}\left[\ell(h(x),y))\right]$ denote the optimal predictor in $\mathcal{H}$ w.r.t $\mathcal{D}$. Consider the sample $S^{h^\star}_U$ and let $g = \mathcal{A}(S^{h^\star}_U)$. We can think of g as the output of $\mathcal{A}$ run over an i.i.d sample S drawn from $\mathcal{D}^\star$, a joint distribution over $\mathcal{X}\times\mathcal{Y}$ defined procedurally by first sampling $x \sim \mathcal{D}_{\mathcal{X}}$, and then outputting the labeled sample $(x, \text{BinRel}(h^\star(x),p))$. Note that $\mathcal{D}^\star$ is indeed a realizable distribution (realized by $h^\star$) w.r.t both ℓ and $\ell^{@p}_{\text{prec}}$. Recall that $m_{\mathcal{A}}(\frac{\epsilon}{b},\delta,K)$ is the sample complexity of $\mathcal{A}$. Since $\mathcal{A}$ is a realizable learner for $\mathcal{H}$ w.r.t ℓ, we have that for $n \geq m_{\mathcal{A}}(\frac{a\epsilon}{2b^2},\delta/2,K)$, with probability at least $1-\frac{\delta}{2}$,

$$\mathbb{E}_{\mathcal{D}^\star}\left[\ell(g(x),y)\right] \leq \frac{a\epsilon}{2b}.$$

By definition of $\mathcal{D}^\star$, it further follows that $\mathbb{E}_{\mathcal{D}^\star}\left[\ell(g(x),y)\right] = \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\ell(g(x),\text{BinRel}(h^\star(x),p))\right]$. Therefore,

$$\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\ell(g(x),\text{BinRel}(h^\star(x),p))\right] \leq \frac{a\epsilon}{2b}.$$

Next, by Lemma E.2, we have pointwise that:

$$\ell(g(x),y) \leq \ell(h^\star(x),y) + \frac{b}{a}\ell(g(x),\text{BinRel}(h^\star(x),p)).$$

Taking expectations on both sides of the inequality gives:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[\ell(g(x),y)\right] &\leq \mathbb{E}_{\mathcal{D}}\left[\ell(h^\star(x),y)\right] + \mathbb{E}_{\mathcal{D}}\left[\frac{b}{a}\ell(g(x),\text{BinRel}(h^\star(x),p))\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\ell(h^\star(x),y)\right] + \frac{b}{a}\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\ell(g(x),\text{BinRel}(h^\star(x),p))\right] \\
&\leq \mathbb{E}_{\mathcal{D}}\left[\ell(h^\star(x),y)\right] + \frac{\epsilon}{2}.
\end{aligned}$$

Therefore, we have shown that $C(S_U)$ contains a hypothesis g that generalizes well with respect to $\mathcal{D}$. The remaining proof follows exactly as in the proof of Lemma 4.4. We include them here for the sake of completeness.

Now we want to show that the predictor $\hat{g}$ returned in step 4 also has good generalization. Crucially, observe that $C(S_U)$ is a finite hypothesis class with cardinality at most K^{pn}. Therefore, by standard Chernoff and union bounds, with probability at least $1-\delta/2$, the empirical risk of every hypothesis in $C(S_U)$ on a sample of size $\geq \frac{8}{\epsilon^2}\log\frac{4|C(S_U)|}{\delta}$ is at most $\epsilon/4$ away from its true error. So, if $m = |S_L| \geq \frac{8}{\epsilon^2}\log\frac{4|C(S_U)|}{\delta}$, then with probability at least $1-\delta/2$, we have

$$\frac{1}{|S_L|}\sum_{(x,y)\in S_L}\ell(g(x),y) \leq \mathbb{E}_{\mathcal{D}}\left[\ell(g(x),y)\right] + \frac{\epsilon}{4} \leq \mathbb{E}_{\mathcal{D}}\left[\ell(h^\star(x),y)\right] + \frac{3\epsilon}{4}.$$

Since $\hat{g}$ is the ERM on S_L over $C(S)$, its empirical risk can be at most $\mathbb{E}_{\mathcal{D}}\left[\ell(h^\star(x),y)\right] + \frac{3\epsilon}{4}$. Given that the population risk of $\hat{g}$ can be at most $\epsilon/4$ away from its empirical risk, we have that

$$\mathbb{E}_{\mathcal{D}}[\ell(\hat{g}(x),y)] \leq \mathbb{E}_{\mathcal{D}}\left[\ell(h^\star(x),y)\right] + \epsilon.$$

Applying union bounds, the entire process succeeds with probability $1-\delta$. We can upper bound the sample complexity of Algorithm 1, denoted $n(\epsilon,\delta,K)$, as

$$\begin{aligned} n(\epsilon,\delta,K) &\leq m_{\mathcal{A}}(\frac{a\epsilon}{2b^2},\delta/2,K) + O\left(\frac{1}{\epsilon^2}\log\frac{|C(S_U)|}{\delta}\right) \\ &\leq m_{\mathcal{A}}(\frac{a\epsilon}{2b^2},\delta/2,K) + O\left(\frac{p\, m_{\mathcal{A}}(\frac{a\epsilon}{2b^2},\delta/2,K)\log(K) + \log\frac{1}{\delta}}{\epsilon^2}\right), \end{aligned}$$

where we use $|C(S_U)| \leq K^{pm_{\mathcal{A}}(\frac{a\epsilon}{2b^2},\delta/2,K)}$. This shows that Algorithm 1, given as input an realizable PAC learner for $\mathcal{H}$ w.r.t ℓ, is an agnostic PAC learner for $\mathcal{H}$ w.r.t ℓ. Using the realizable learner we constructed before this step as the input completes this proof as we have constructively converted an agnostic PAC learner for $\ell_{\text{prec}}^{@p}$ into an agnostic PAC learner for ℓ. □

Lemma C.2 and C.3 together complete the proof of sufficiency in Theorem 4.2. Finally, Lemma C.4 below shows that the agnostic PAC learnability of $\mathcal{H}_i^p$ for all $i \in [K]$ is necessary for the agnostic PAC learnability of $\mathcal{H}$ w.r.t any $\ell \in \mathcal{L}(\ell_{\text{prec}}^{@p})$. Like before, the proof of Lemma C.4 is constructive and follows exactly the same strategy as Lemma 4.5. That is, given as input a learner for ℓ, we will convert it into an agnostic learner for $\mathcal{H}_i^p$. In fact, the conversion is exactly the same as in the proof of Lemma 4.5 and just requires running Algorithm 4 with an input learner for $\ell \in \mathcal{L}(\ell_{\text{prec}}^{@p})$ and setting $j = p$.

Lemma C.4. *If a function class $\mathcal{H} \subseteq \mathcal{S}_K^{\mathcal{X}}$ is agnostic PAC learnable w.r.t $\ell \in \mathcal{L}(\ell_{prec}^{@p})$, then $\mathcal{H}_i^p$ is agnostic PAC learnable w.r.t the 0-1 loss for all $i \in [K]$.*

Proof. Fix $\ell \in \mathcal{L}(\ell_{\text{prec}}^{@p})$ and $i \in [K]$. Let $a = \min_{\pi,y}\{\ell(\pi,y) \mid \ell(\pi,y) \neq 0\}$. Let $\mathcal{H}$ be an arbitrary ranking hypothesis class and $\mathcal{A}$ be an agnostic PAC learner for $\mathcal{H}$ w.r.t ℓ. Our goal will to be to use $\mathcal{A}$ to construct an agnostic PAC learner for $\mathcal{H}_i^p$.

Let $\mathcal{D}$ be any distribution over $\mathcal{X}\times\{0,1\}$, $h_i^{\star,p} = \arg\min_{h\in\mathcal{H}_i^p}\mathbb{E}_{\mathcal{D}}\left[\mathbb{1}\{h(x)\neq y\}\right]$ the optimal hypothesis, and $h^\star \in \mathcal{H}$ be any valid completion of $h_i^{\star,p}$. We will now show that Algorithm 4 from the proof of Lemma 4.5 is an agnostic PAC learner for $\mathcal{H}_i^p$ if we set $j = p$ and give it as input an agnostic PAC learner $\mathcal{A}$ for $\mathcal{H}$ w.r.t. $\ell \in \mathcal{L}(\ell_{\text{prec}}^{@p})$.

Consider the sample $S_U^{h^\star}$ and let $g = \mathcal{A}(S_U^{h^\star})$. We can think of g as the output of $\mathcal{A}$ run over an i.i.d sample S drawn from $\mathcal{D}^\star$, a joint distribution over $\mathcal{X}\times\mathcal{Y}$ defined procedurally by first sampling $x \sim \mathcal{D}_{\mathcal{X}}$ and then outputting the labeled sample $(x, \text{BinRel}(h^\star(x),p))$. Note that $\mathcal{D}^\star$ is a realizable distribution (realized by $h^\star$) w.r.t $\ell_{\text{prec}}^{@p}$ and therefore also ℓ. Let $m_{\mathcal{A}}(\epsilon,\delta,K)$ be the sample complexity of $\mathcal{A}$.

Since $\mathcal{A}$ is an agnostic PAC learner for $\mathcal{H}$ w.r.t ℓ, we have that for sample size $n \geq m_{\mathcal{A}}(\frac{a\epsilon}{2},\delta/2,K)$, with probability at least $1-\frac{\delta}{2}$,

$$\mathbb{E}_{\mathcal{D}^\star}\left[\ell(g(x),y)\right] \le \inf_{h\in\mathcal{H}} \mathbb{E}_{\mathcal{D}^\star}\left[\ell(h(x),y)\right] + \frac{a\epsilon}{2} = \frac{a\epsilon}{2}.$$

Furthermore, by definition of $\mathcal{D}^\star$, $\mathbb{E}_{\mathcal{D}^\star}\left[\ell(g(x),y)\right] = \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[\ell(g(x),\text{BinRel}(h^\star(x),p))\right]$. Therefore, $\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[\ell(g(x),\text{BinRel}(h^\star(x),p))\right] \le \frac{a\epsilon}{2}$. Next, using Lemma E.4, we have pointwise that

$$\begin{aligned}
\mathbb{1}\{g_i^p(x) \neq h_i^{\star,p}(x)\} &\le \mathbb{1}\{\ell_{\text{prec}}^{@p}(g(x),\text{BinRel}(h^\star(x),p)) > 0\} \\
&= \mathbb{1}\{\ell(g(x),\text{BinRel}(h^\star(x),p)) > 0\} \\
&\le \frac{1}{a}\,\ell(g(x),\text{BinRel}(h^\star(x),p)).
\end{aligned}$$

Taking expectations on both sides gives,

$$\mathbb{E}_\mathcal{D}\left[\mathbb{1}\{g_i^p(x) \neq h_i^{\star,p}(x)\}\right] \le \frac{1}{a}\,\mathbb{E}_\mathcal{D}\left[\ell(g(x),\text{BinRel}(h^\star(x),p))\right] \le \frac{\epsilon}{2},$$

where in the last inequality we use the fact that $\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[\ell(g(x),\text{BinRel}(h^\star(x),p))\right] \le \frac{a\epsilon}{2}$. Finally, using the triangle inequality, we have that

$$\begin{aligned}
\mathbb{E}_\mathcal{D}\left[\mathbb{1}\{g_i^p(x) \neq y\}\right] &\le \mathbb{E}_\mathcal{D}\left[\mathbb{1}\{h_i^{\star,p}(x) \neq y\}\right] + \mathbb{E}_\mathcal{D}\left[\mathbb{1}\{g_i^p(x) \neq h_i^{\star,p}(x\}\right] \\
&\le \mathbb{E}_\mathcal{D}\left[\mathbb{1}\{h_i^{\star,p}(x) \neq y\}\right] + \frac{\epsilon}{2} \\
&= \min_{h_i^p\in\mathcal{H}_i^p} \mathbb{E}_\mathcal{D}\left[\mathbb{1}\{h_i^p(x) \neq y\}\right] + \frac{\epsilon}{2}.
\end{aligned}$$

Since $g_i^p \in C_i^p(S_U)$, we have shown that $C_i^p(S_U)$ contains a hypothesis that generalizes well w.r.t $\mathcal{D}$. Now we want to show that the predictor $\hat{g}_i^p$ returned in step 4 also generalizes well. Crucially, observe that $C_i^p(S_U)$ is a finite hypothesis class with cardinality at most K^{pn}. Therefore, by standard Chernoff and union bounds, with probability at least $1-\delta/2$, the empirical risk of every hypothesis in $C_i^p(S_U)$ on a sample of size $\ge \frac{8}{\epsilon^2}\log\frac{4|C_i^j(S_U)|}{\delta}$ is at most $\epsilon/4$ away from its true error. So, if $m = |S_L| \ge \frac{8}{\epsilon^2}\log\frac{4|C_i^j(S_U)|}{\delta}$, then with probability at least $1-\delta/2$, we have

$$\frac{1}{|S_L|}\sum_{(x,y)\in S_L} \mathbb{1}\{g_i^p(x) \neq y\} \le \mathbb{E}_\mathcal{D}\left[\mathbb{1}\{g_i^p(x) \neq y\}\right] + \frac{\epsilon}{4} \le \frac{3\epsilon}{4}.$$

Since $\hat{g}_i^p$ is the ERM on S_L over $C_i^p(S_U)$, its empirical risk can be at most $\frac{3\epsilon}{4}$. Given that the population risk of $\hat{g}_i^p$ can be at most $\epsilon/4$ away from its empirical risk, we have that

$$\mathbb{E}_\mathcal{D}[\mathbb{1}\{\hat{g}_i^p(x) \neq y\}] \le \min_{h_i^p\in\mathcal{H}_i^p} \mathbb{E}_\mathcal{D}\left[\mathbb{1}\{h_i^p(x) \neq y\}\right] + \epsilon.$$

Applying union bounds, the entire process succeeds with probability $1-\delta$. We can compute the upper bound on the sample complexity of Algorithm 4, denoted $n(\epsilon,\delta,K)$, as

$$\begin{aligned}
n(\epsilon,\delta,K) &\le m_\mathcal{A}(\frac{a\epsilon}{2},\delta/2,K) + O\left(\frac{1}{\epsilon^2}\log\frac{|C(S_U)|}{\delta}\right) \\
&\le m_\mathcal{A}(\frac{a\epsilon}{2},\delta/2,K) + O\left(\frac{p\; m_\mathcal{A}(\frac{a\epsilon}{2},\delta/2,K)\log(K) \;+ \log\frac{1}{\delta}}{\epsilon^2}\right),
\end{aligned}$$

where we use $|C(S_U)| \le K^{pm_\mathcal{A}(\frac{a\epsilon}{2},\delta/2,K)}$. This shows that Algorithm 4 is an agnostic PAC learner for $\mathcal{H}_i^p$ w.r.t 0-1 loss. Since our choice of loss $\ell \in \mathcal{L}(\ell_{\text{prec}}^{@p})$ and index i were arbitrary, agnostic PAC learnability of $\mathcal{H}$ w.r.t ℓ implies agnostic PAC learnability of $\mathcal{H}_i^p$ w.r.t the 0-1 loss for all $i \in [K]$. □

Combining Lemma C.2, C.3 and C.4 gives Theorem 4.2.

D Proofs for Online Multilabel Ranking

D.1 Proof of necessity in Theorem 5.1

Proof. Fix $\ell \in \mathcal{L}(\ell_{\text{sum}}^{@p})$ and $(i,j) \in [K] \times [p]$. Given an online learner $\mathcal{A}$ for $\mathcal{H}$ w.r.t ℓ, our goal is to construct an agnostic online learner $\mathcal{A}_i^j$ for $\mathcal{H}_i^j$. To that end, let $(x_1, y_1), ..., (x_T, y_T) \in (\mathcal{X} \times \{0,1\})^T$ denote a stream of labeled instances. Define $h_i^{\star,j} = \arg\min_{h_i^j \in \mathcal{H}_i^j} \sum_{t=1}^T \mathbb{1}\{h_i^j(x_t) \neq y_t\}$ to be the optimal function in $\mathcal{H}_i^j$ and $h^\star$ be an arbitrary completion of $h_i^{\star,j}$. As in the sufficiency proof, our construction of the online learner for $\mathcal{H}_i^j$ will run REWA over a set of experts we construct below.

For any bitstring $b \in \{0,1\}^T$, let $\phi : \{t \in [T] : b_t = 1\} \to \mathcal{S}_K$ denote a function mapping time points where $b_t = 1$ to permutations. Let $\Phi_b = \mathcal{S}_K^{\{t \in [T] : b_t = 1\}}$ denote all such functions ϕ. For every $h \in \mathcal{H}$, there exists a $\phi_b^h \in \Phi_b$ such that for all $t \in \{t : b_t = 1\}$, $\phi_b^h(t) = h(x_t)$. Let $|b| = |\{t \in [T] : b_t = 1\}|$. For every $b \in \{0,1\}^T$ and $\phi \in \Phi_b$, define an Expert $E_{b,\phi}$. Expert $E_{b,\phi}$, formally presented in Algorithm 6, uses $\mathcal{A}$ to make predictions in each round. For every $b \in \{0,1\}^T$, let $\mathcal{E}_b = \bigcup_{\phi \in \Phi_b} \{E_{b,\phi}\}$ denote the set of all Experts parameterized by functions $\phi \in \Phi_b$. As before, we will actually define $\mathcal{E}_b = \{E_0\} \cup \bigcup_{\phi \in \Phi_b} \{E_{b,\phi}\}$, where E_0 is the expert that never updates $\mathcal{A}$ and only uses it to make predictions in each round. Note that $1 \leq |\mathcal{E}_b| \leq (K!)^{|b|} \leq K^{K|b|}$.

Algorithm 6 Expert (b, ϕ)

Input: Independent copy of online learner $\mathcal{A}$ for $\mathcal{H}$
for $t = 1, ..., T$ **do**
  Receive example x_t
  Predict $\mathbb{1}\{\hat{\pi}_i \leq j\}$ where $\hat{\pi} = \mathcal{A}(x_t)$
  if $b_t = 1$ **then**
    Update $\mathcal{A}$ by passing $(x_t, \text{BinRel}(\phi(t), j))$
end

We are now ready to give the agnostic online learner for $\mathcal{H}_i^j$, henceforth denoted by $\mathcal{Q}$. Our online learner $\mathcal{Q}$ is very similar to Algorithm 3. First, it will sample a $B \in \{0,1\}^T$ s.t. $B_t \sim \text{Bernoulli}(T^\beta/T)$. Then, it will construct a set of experts $\mathcal{E}_B$ using Algorithm 6. Finally, it will run REWA, denoted by $\mathcal{P}$, on the 0-1 loss over the stream $(x_1, y_1), ..., (x_T, y_T)$. As before, let A and P be the random variables denoting internal randomness of the algorithm $\mathcal{A}$ and $\mathcal{P}$. Using REWA guarantees and following exactly the same calculation as in the sufficiency proof, we arrive at

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{\mathcal{Q}(x_t) \neq y_t\}\right] \leq \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{E_{B,\phi_B^{h^\star}}(x_t) \neq y_t\}\right] + \sqrt{2T^{1+\beta} K \ln K}.$$

The inequality above is the adaptation of Equation (1) for this proof. Recall that $h_i^{\star,j}$ is the optimal function in hindsight for the stream and $h^\star$ is a completion of $h_i^{\star,j}$. Since $\mathbb{1}\{E_{B,\phi_B^{h^\star}}(x_t) \neq y_t\} \leq \mathbb{1}\{h_i^{\star,j}(x_t) \neq y_t\} + \mathbb{1}\{E_{B,\phi_B^{h^\star}}(x_t) \neq h_i^{\star,j}(x_t)\}$, the inequality above reduces to

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{\mathcal{Q}(x_t) \neq y_t\}\right] \leq \sum_{t=1}^T \mathbb{1}\{h_i^{\star,j}(x_t) \neq y_t\} + \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{E_{B,\phi_B^{h^\star}}(x_t) \neq h_i^{\star,j}(x_t)\}\right] + \sqrt{2T^{1+\beta} K \ln K}.$$

It now suffices to show that $\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{E_{B,\phi_B^{h^\star}}(x_t) \neq h_i^{\star,j}(x_t)\}\right]$ is sub-linear function of T.

Given an online learner $\mathcal{A}$ for $\mathcal{H}$, an instance $x \in \mathcal{X}$, and an ordered finite sequence of labeled examples $L \in (\mathcal{X} \times \mathcal{Y})^*$, let $\mathcal{A}(x|L)$ be the random variable denoting the prediction of $\mathcal{A}$ on the instance x after running and updating on L. For any $b \in \{0,1\}^T$, $h \in \mathcal{H}$, and $t \in [T]$, let $L_{b_{<t}}^h = \{(x_i, \text{BinRel}(h(x_s), j)) : s < t \text{ and } b_s = 1\}$ denote the *subsequence* of the sequence of

labeled instances $\{(x_s, \text{BinRel}(h(x_s), j))\}_{s=1}^{t-1}$ where $b_s = 1$. Thus, using Lemma E.3, we have

$$\begin{aligned}
\mathbb{1}\{E_{B,\phi_B^{h^\star}}(x_t) \neq h_i^{\star,j}(x_t)\} &\leq \mathbb{1}\{\ell_{\text{sum}}^{@p}(\mathcal{A}(x_t \mid L_{B_{<t}}^{h^\star}), \text{BinRel}(h^\star(x_t), j)) > 0\} \\
&= \mathbb{1}\{\ell(\mathcal{A}(x_t \mid L_{B_{<t}}^{h^\star}), \text{BinRel}(h^\star(x_t), j)) > 0\} \\
&\leq \frac{1}{a}\,\ell(\mathcal{A}(x_t \mid L_{B_{<t}}^{h^\star}), \text{BinRel}(h^\star(x_t), j), \text{BinRel}(h^\star(x_t), j)),
\end{aligned}$$

where equality follows from the fact that $\ell \in \mathcal{L}(\ell_{\text{sum}}^{@p})$. Here, a is the lower bound whenever it is non-zero. Taking expectations of both sides and summing over $t \in [T]$ gives

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{E_{B,\phi_B^{h^\star}}(x_t) \neq h_i^{\star,j}(x_t)\}\right] \leq \frac{1}{a}\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{A}(x_t \mid L_{B_{<t}}^{h^\star}), \text{BinRel}(h^\star(x_t), j))\right].$$

To upperbound the right-hand side, we will again use the fact that the prediction $\mathcal{A}(x_t \mid L_{B_{<t}}^{h^\star})$ only depends on $(B_1, \ldots, B_{t-1})$, but is independent of B_t. The details of this calculation are omitted because they are identical to that of the sufficiency proof. Using independence of $\mathcal{A}(x_t \mid L_{B_{<t}}^{h^\star})$ and B_t, we obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{A}(x_t \mid L_{B_{<t}}^{h^\star}), \text{BinRel}(h^\star(x_t), j))\right] &= \frac{T}{T^\beta}\,\mathbb{E}\left[\sum_{t:B_t=1} \ell(\mathcal{A}(x_t \mid L_{B_{<t}}^{h^\star}), \text{BinRel}(h^\star(x_t), j))\right] \\
&= \frac{T}{T^\beta}\,\mathbb{E}\left[\mathbb{E}\left[\sum_{t:B_t=1} \ell(\mathcal{A}(x_t \mid L_{B_{<t}}^{h^\star}), \text{BinRel}(h^\star(x_t), j)) \,\middle|\, B\right]\right] \\
&\leq \frac{T}{T^\beta}\,\mathbb{E}\left[R(|B|, K)\right],
\end{aligned}$$

where $R(|B|, K)$ is the regret of the algorithm $\mathcal{A}$, a sub-linear function of $|B|$. In the last step, we use the fact that $\mathcal{A}$ is a (realizable) online learner for $\mathcal{H}$ w.r.t. ℓ and the feedback that the algorithm received was $(x_t, \text{BinRel}(h^\star(x_t), j))$ in the rounds whenever $B_t = 1$. Again, Lemma 5.17 from Ceccherini-Silberstein et al. [2017] guarantees an existence of a concave sublinear upperbound $\tilde{R}(|B|, K)$ of $R(|B|, K)$. Then, applying Jensen's inequality yields $\mathbb{E}\left[R(|B|, K)\right] \leq \mathbb{E}\left[\tilde{R}(|B|, K)\right] \leq \tilde{R}(T^\beta, K)$, a concave sub-linear function of T^β. Combining everything, we get

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{Q}(x_t) \neq y_t\}\right] &\leq \sum_{t=1}^{T} \mathbb{1}\{h_i^{\star,j}(x_t) \neq y_t\} + \frac{T}{aT^\beta}\tilde{R}(T^\beta, K) + \sqrt{2T^{1+\beta}K\ln K} \\
&= \arg\min_{h_i^j \in \mathcal{H}_i^j} \sum_{t=1}^{T} \mathbb{1}\{h_i^j(x_t) \neq y_t\} + \frac{T}{aT^\beta}\tilde{R}(T^\beta, K) + \sqrt{2T^{1+\beta}K\ln K}
\end{aligned}$$

For any choice of $\beta \in (0, 1)$, the regret above is a sub-linear function of T. Therefore, we have shown that $\mathcal{Q}$ is an agnostic learner for $\mathcal{H}_i^j$ w.r.t. 0-1 loss. □

D.2 Proof of Theorem 5.2

Proof. (of sufficiency in Theorem 5.2) Fix $\ell \in \mathcal{L}(\ell_{\text{prec}}^{@p})$ and let $M = \max_{\pi,y} \ell(\pi, y)$. This proof is virtually identical to the proof of sufficiency in Theorem 4.1. However, we provide the full details here for completion. Our proof is also based on reduction. That is, given realizable learners $\mathcal{A}_i^p$ of $\mathcal{H}_i^p$'s for $i \in [K]$ w.r.t. 0-1 loss, we will construct an agnostic learner $\mathcal{Q}$ for $\mathcal{H}$ w.r.t. ℓ. We will construct a set of experts $\mathcal{E}$ that uses $\mathcal{A}_i^p$ to make predictions and run the REWA algorithm using these experts.

Let $(x_1, y_1), ..., (x_T, y_T) \in (\mathcal{X} \times \mathcal{Y})^T$ denote the stream of points to be observed by the online learner. As before, we will assume an oblivious adversary. Define $h^\star = \arg\min_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(x_t), y_t)$ to be the optimal hypothesis in hindsight.

For any bitstring $b \in \{0,1\}^T$, let $\phi : \{t \in [T] : b_t = 1\} \to \mathcal{S}_K$ denote a function mapping time points where $b_t = 1$ to permutations. Let $\Phi_b = \mathcal{S}_K^{\{t \in [T]: b_t = 1\}}$ denote all such functions ϕ. For every $h \in \mathcal{H}$, there exists a $\phi_b^h \in \Phi_b$ such that for all $t \in \{t : b_t = 1\}$, $\phi_b^h(t) = h(x_t)$. Let $|b| = |\{t \in [T] : b_t = 1\}|$. For every $b \in \{0,1\}^T$ and $\phi \in \Phi_b$, we will define an Expert $E_{b,\phi}$. Expert $E_{b,\phi}$, formally presented in Algorithm 3, uses $\mathcal{A}_i^p$'s to make predictions in each round. However, $E_{b,\phi}$ only updates the $\mathcal{A}_i^p$'s on those rounds where $b_t = 1$, using ϕ to compute a labeled instance. For every $b \in \{0,1\}^T$, let $\mathcal{E}_b = \bigcup_{\phi \in \Phi_b} \{E_{b,\phi}\}$ denote the set of all Experts parameterized by functions $\phi \in \Phi_b$. If b is the bitstring with all zeros, then $\mathcal{E}_b$ will be empty. Therefore, we will actually define $\mathcal{E}_b = \{E_0\} \cup \bigcup_{\phi \in \Phi_b} \{E_{b,\phi}\}$, where E_0 is the expert that never updates $\mathcal{A}_i^j$'s and only uses them for predictions in all $t \in [T]$. Note that $1 \le |\mathcal{E}_b| \le (K!)^{|b|} \le K^{K|b|}$. Using these experts, Algorithm 3 is our agnostic online learner $\mathcal{Q}$ for $\mathcal{H}$ w.r.t $\ell \in \mathcal{L}(\ell_{\text{prec}}^{@p})$.

Algorithm 7 Expert (b, ϕ)

Input: Independent copy of realizable learners $\mathcal{A}_i^p$ of $\mathcal{H}_i^p$ for $i \in [K]$

for $t = 1, ..., T$ **do**
  Receive example x_t
  Define a binary vote vector $v_t \in \{0,1\}^K$ such that $v_t[i] = \mathcal{A}_i^p(x_t)$
  Predict $\hat{\pi}_t \in \arg\min_{\pi \in \mathcal{S}_K} \langle \pi, v_t \rangle$
  if $b_t = 1$ **then**
    Let $\pi = \phi(t)$ and for each $i \in [K]$, update $\mathcal{A}_i^p$ by passing (x_t, π_i^p)
end

Using REWA guarantees and following exactly the same calculation as in the proof of Theorem 5.1 we immediately arrive at

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{Q}(x_t), y_t)\right] \le \mathbb{E}\left[\sum_{t=1}^{T} \ell(E_{B,\phi_B^{h^\star}}(x_t), y_t)\right] + M\sqrt{2T^{1+\beta} K \ln K},$$

the analog of Equation (1) for this setting. Using Lemma E.2, we have

$$\ell(E_{B,\phi_B^{h^\star}}(x_t), y_t) \le \ell(h^\star(x_t), y_t) + \frac{M}{a}\ell(E_{B,\phi_B^{h^\star}}(x_t), \text{BinRel}(h^\star(x_t), p))$$

pointwise, where $a = \min_{\pi,y}\{\ell(\pi, y) \mid \ell(\pi, y) \neq 0\}$. By definition of M, we further get

$$\begin{aligned}\ell(E_{B,\phi_B^{h^\star}}(x_t), \text{BinRel}(h^\star(x_t), p)) &\le M\, \mathbb{1}\{\ell(E_{B,\phi_B^{h^\star}}(x_t), \text{BinRel}(h^\star(x_t), p)) > 0\} \\ &= M\, \mathbb{1}\{\ell_{\text{prec}}^{@p}(E_{B,\phi_B^{h^\star}}(x_t), \text{BinRel}(h^\star(x_t), p)) > 0\},\end{aligned}$$

where the equality follows from the fact that $\ell \in \mathcal{L}(\ell_{\text{prec}}^{@p})$.

In order to upperbound the indicator above, we need some more notations. Given the realizable online learner $\mathcal{A}_i^p$ for $i \in [K] \times [p]$, an instance $x \in \mathcal{X}$, and an ordered finite sequence of labeled examples $L \in (\mathcal{X} \times \{0,1\})^*$, let $\mathcal{A}_i^p(x|L)$ be the random variable denoting the prediction of $\mathcal{A}_i^p$ on the instance x after running and updating on L. For any $b \in \{0,1\}^T$, $h \in \mathcal{H}$, and $t \in [T]$, let $L_{b_{<t}}^h(i,p) = \{(x_s, h_i^p(x_s)) : s < t \text{ and } b_s = 1\}$ denote the *subsequence* of the sequence of labeled instances $\{(x_s, h_i^p(x_s))\}_{s=1}^{t-1}$ where $b_s = 1$. Then, we have

$$\mathbb{1}\{\ell_{\text{prec}}^{@p}(E_{B,\phi_B^{h^\star}}(x_t), \text{BinRel}(h^\star(x_t), p)) > 0\} \le \sum_{i=1}^{K} \mathbb{1}\{\mathcal{A}_i^p(x_t \mid L_{B_{<t}}^{h^\star}(i,p)) \neq h_i^{\star,p}(x_t)\}.$$

To prove this claimed inequality, consider the case when $\sum_{i=1}^K \mathbb{1}\{\mathcal{A}_i^p(x_t \mid L_{B_{<t}}^{h^\star}(i,p)) \neq h_i^{\star,p}(x_t)\} = 0$ because the inequality is trivial otherwise. Then, we must have $\mathcal{A}_i^p(x_t \mid L_{B_{<t}}^{h^\star}(i,p)) = h_i^{\star,p}(x_t)$ for all $i \in [K]$. Let $v_t \in \{0,1\}^K$ such that $v_t[i] = \mathcal{A}_i^p(x_t \mid L_{B_{<t}}^{h^\star}(i,p))$ be a binary vote vector that the expert $E_{B,\phi_B^{h^\star}}$ constructs in round t. Since $h^\star(x_t)$ is a permutation, the vote vector v_t must contain exactly p labels with 1 vote and $K - p$ labels with 0 votes. Thus, every $\hat{\pi}_t \in \arg\min_{\pi \in \mathcal{S}_K} \langle \pi, v_t \rangle$ must rank labels with 1 vote in top p and labels with 0 votes outside top p.

In other words, we must have $\hat{\pi}_t \overset{p}{=} h^\star(x_t)$, and thus $\ell_{\text{prec}}^{@p}(\hat{\pi}_t, \text{BinRel}(h^\star(x_t), p)) = 0$ by definition of $\ell_{\text{prec}}^{@p}$. Our claim follows because $E_{B,\phi_B^{h^\star}}(x_t) \in \arg\min_{\pi \in \mathcal{S}_K} \langle \pi, v_t \rangle$.

Combining everything, we obtain

$$\ell(E_{B,\phi_B^{h^\star}}(x_t), y_t) \leq \ell(h^\star(x_t), y_t) + \frac{M^2}{a} \sum_{i=1}^{K} \mathbb{1}\{\mathcal{A}_i^p(x_t \mid L_{B_{<t}}^{h^\star}(i,p)) \neq h_i^{\star,p}(x_t)\}.$$

Taking expectations on both sides and summing over all $t \in [T]$ yields

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(E_{B,\phi_B^{h^\star}}(x_t), y_t)\right] \leq \sum_{t=1}^{T} \ell(h^\star(x_t), y_t) + \frac{M^2}{a} \sum_{i=1}^{K} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{A}_i^p(x_t \mid L_{B_{<t}}^{h^\star}(i,p)) \neq h_i^{\star,p}(x_t)\}\right].$$

So, it now suffices to show that $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{A}_i^p(x_t \mid L_{B_{<t}}^{h^\star}(i,p)) \neq h_i^{\star,p}(x_t)\}\right]$ is a sub-linear function of T. Again, using the independence of B_t and the algorithm's prediction in round t, we can write

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{A}_i^p(x_t \mid L_{B_{<t}}^{h^\star}(i,p)) \neq h_i^{\star,p}(x_t)\}\right] &= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\{\mathcal{A}_i^p(x_t \mid L_{B_{<t}}^{h^\star}(i,p)) \neq h_i^{\star,p}(x_t)\}\right] \frac{\mathbb{P}\left[B_t = 1\right]}{\mathbb{P}\left[B_t = 1\right]} \\
&= \frac{T}{T^\beta} \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\{\mathcal{A}_i^p(x_t \mid L_{B_{<t}}^{h^\star}(i,p)) \neq h_i^{\star,p}(x_t)\}\right] \mathbb{E}\left[\mathbb{1}\{B_t = 1\}\right] \\
&= \frac{T}{T^\beta} \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\{\mathcal{A}_i^p(x_t \mid L_{B_{<t}}^{h^\star}(i,p)) \neq h_i^{\star,p}(x_t)\} \mathbb{1}\left\{B_t = 1\right\}\right].
\end{aligned}$$

Next, we can use the regret guarantee of the algorithm $\mathcal{A}_i^p$ on the rounds it was updated. That is,

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\{\mathcal{A}_i^p(x_t \mid L_{B_{<t}}^{h^\star}(i,p)) \neq h_i^{\star,p}(x_t)\} \mathbb{1}\left\{B_t = 1\right\}\right] &= \mathbb{E}\left[\sum_{t:B_t=1} \mathbb{1}\{\mathcal{A}_i^p(x_t \mid L_{B_{<t}}^{h^\star}(i,p)) \neq h_i^{\star,p}(x_t)\}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{t:B_t=1} \mathbb{1}\{\mathcal{A}_i^p(x_t \mid L_{B_{<t}}^{h^\star}(i,p)) \neq h_i^{\star,p}(x_t)\} \,\middle|\, B\right]\right] \\
&\leq \mathbb{E}_B\left[R_i^p(|B|)\right],
\end{aligned}$$

where $R_i^p(|B|)$ is the regret of $\mathcal{A}_i^p$, a sub-linear function of $|B|$. In the last step, we use the fact that $\mathcal{A}_i^p$ is a realizable algorithm for $\mathcal{H}_i^p$ and the feedback that the algorithm received was $(x_t, h_i^{\star,p}(x_t))$ in the rounds whenever $B_t = 1$. By Lemma 5.17 from Ceccherini-Silberstein et al. [2017], there exists a concave sub-linear function $\tilde{R}_i^p(|B|)$ that upperbounds $R_i^p(|B|)$. By Jensen's inequality, $\mathbb{E}_B\left[R_i^p(|B|)\right] \leq \tilde{R}_i^p(T^\beta)$, a sub-linear function of T^β.

Putting everything together, we obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{Q}(x_t), y_t)\right] &\leq \sum_{t=1}^{T} \ell(h^\star(x_t), y_t) + \frac{M^2}{a} \sum_{i=1}^{K} \frac{T}{T^\beta} \tilde{R}_i^p(T^\beta) + M\sqrt{2T^{1+\beta} K \ln K} \\
&= \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(x_t), y_t) + \frac{pM^2}{a} \sum_{i=1}^{K} \frac{T}{T^\beta} \tilde{R}_i^p(T^\beta) + M\sqrt{2T^{1+\beta} K \ln K}.
\end{aligned}$$

Since $\tilde{R}_i^p(T^\beta)$ is a sublinear function of T^β, we have that $\frac{T}{T^\beta}\tilde{R}_i^p(T^\beta)$ is a sublinear function of T. As the sum of sublinear functions is sublinear, the second term above must be a sublinear function of T. Thus, the regret is sub-linear for any choice of $\beta \in (0,1)$. This completes our proof as we have shown that the algorithm $\mathcal{Q}$ achieves sub-linear regret in T. □

We will now show that the online learnability of $\mathcal{H}$ w.r.t ℓ implies that $\mathcal{H}_i^p$ for each $i \in [K]$ is online learnable w.r.t 0-1 loss.

Proof. (of necessity in Theorem 5.2)

Fix $\ell \in \mathcal{L}(\ell_{\text{prec}}^{@p})$ and let $M = \max_{\pi,y} \ell(\pi, y)$. Given an online learner $\mathcal{A}$ for $\mathcal{H}$ w.r.t ℓ, our goal is to construct an agnostic online learner $\mathcal{A}_i^p$ for $\mathcal{H}_i^p$ for a fixed $i \in [K]$. One can construct agnostic online learners for $\mathcal{H}_i^p$ for all $i \in [K]$ by symmetry. Our construction uses the REWA and is similar to the sufficiency proof above.

Let us define function ϕ's, the collection of functions Φ_b for every b in the same way we did before. For every $b \in \{0,1\}^T$ and $\phi \in \Phi_b$, define an Expert $E_{b,\phi}$. Expert $E_{b,\phi}$ is the expert presented in Algorithm 6 after setting $j = p$ and uses $\mathcal{A}$ to make predictions in each round. For every $b \in \{0,1\}^T$, let $\mathcal{E}_b = \bigcup_{\phi \in \Phi_b} \{E_{b,\phi}\}$ denote the set of all Experts parameterized by functions $\phi \in \Phi_b$. As before, we will actually define $\mathcal{E}_b = \{E_0\} \cup \bigcup_{\phi \in \Phi_b} \{E_{b,\phi}\}$, where E_0 is the expert that never updates $\mathcal{A}$ and only uses it to make predictions in each round. Note that $1 \leq |\mathcal{E}_b| \leq (K!)^{|b|} \leq K^{K|b|}$.

The online learner for $\mathcal{H}_i^p$, henceforth denoted by $\mathcal{Q}$, is similar to Algorithm 3. First, it samples a $B \in \{0,1\}^T$ s.t. $B_t \sim \text{Bernoulli}(T^\beta/T)$, constructs a set of experts $\mathcal{E}_B$ using Algorithm 6 and runs REWA, denoted by $\mathcal{P}$, on the 0-1 loss over the stream $(x_1, y_1), ..., (x_T, y_T) \in (\mathcal{X} \times \{0,1\})^T$. Let $h_i^{\star,p} = \arg\min_{h_i^p \in \mathcal{H}_i^p} \sum_{t=1}^T \mathbb{1}\{h_i^p(x_t) \neq y_t\}$ be the optimal function in hindsight and $h^\star$ be any arbitrary completion of $h_i^{\star,p}$.

Using REWA guarantees and following exactly the same calculation as in the sufficiency proof, we arrive at

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{\mathcal{Q}(x_t) \neq y_t\}\right] \leq \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{E_{B,\phi_B^{h^\star}}(x_t) \neq y_t\}\right] + \sqrt{2T^{1+\beta} K \ln K}.$$

The inequality above is the adaptation of Equation (1) for this proof. Since $\mathbb{1}\{E_{B,\phi_B^{h^\star}}(x_t) \neq y_t\} \leq \mathbb{1}\{h_i^{\star,p}(x_t) \neq y_t\} + \mathbb{1}\{E_{B,\phi_B^{h^\star}}(x_t) \neq h_i^{\star,p}(x_t)\}$, the inequality above reduces to

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{\mathcal{Q}(x_t) \neq y_t\}\right] \leq \sum_{t=1}^T \mathbb{1}\{h_i^{\star,p}(x_t) \neq y_t\} + \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{E_{B,\phi_B^{h^\star}}(x_t) \neq h_i^{\star,p}(x_t)\}\right] + \sqrt{2T^{1+\beta} K \ln K}.$$

It now suffices to show that $\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{E_{B,\phi_B^{h^\star}}(x_t) \neq h_i^{\star,p}(x_t)\}\right]$ is sub-linear in T.

Given an online learner $\mathcal{A}$ for $\mathcal{H}$, an instance $x \in \mathcal{X}$, and an ordered finite sequence of labeled examples $L \in (\mathcal{X} \times \mathcal{Y})^*$, let $\mathcal{A}(x|L)$ be the random variable denoting the prediction of $\mathcal{A}$ on the instance x after running and updating on L. For any $b \in \{0,1\}^T$, $h \in \mathcal{H}$, and $t \in [T]$, let $L_{b_{<t}}^h = \{(x_i, \text{BinRel}(h(x_s), p)) : s < t \text{ and } b_s = 1\}$ denote the *subsequence* of the sequence of labeled instances $\{(x_s, \text{BinRel}(h(x_s), p))\}_{s=1}^{t-1}$ where $b_s = 1$. Using Lemma E.4, we have

$$\begin{aligned}
\mathbb{1}\{E_{B,\phi_B^{h^\star}}(x_t) \neq h_i^{\star,p}(x_t)\} &\leq \mathbb{1}\{\ell_{\text{prec}}^{@p}(\mathcal{A}(x_t \mid L_{B_{<t}}^{h^\star}), \text{BinRel}(h^\star(x_t), p)) > 0\} \\
&= \mathbb{1}\{\ell(\mathcal{A}(x_t \mid L_{B_{<t}}^{h^\star}), \text{BinRel}(h^\star(x_t), p)) > 0\} \\
&\leq \frac{1}{a}\, \ell(\mathcal{A}(x_t \mid L_{B_{<t}}^{h^\star}), \text{BinRel}(h^\star(x_t), p)),
\end{aligned}$$

where the equality follows from the definition of the loss class. Here, a is the lower bound on ℓ whenever it is non-zero. Thus, we obtain

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{E_{B,\phi_B^{h^\star}}(x_t) \neq h_i^{\star,p}(x_t)\}\right] \leq \frac{1}{a}\mathbb{E}\left[\sum_{t=1}^T \ell(\mathcal{A}(x_t \mid L_{B_{<t}}^{h^\star}), \text{BinRel}(h^\star(x_t), p))\right]$$

Now, we will again use the fact that the prediction $\mathcal{A}(x_t \mid L^{h^\star}_{B_{<t}})$ only depends on $(B_1, \ldots, B_{t-1})$, but is independent of B_t. Using this independence, we obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{A}(x_t \mid L^{h^\star}_{B_{<t}}), \text{BinRel}(h^\star(x_t), p))\right] &= \frac{T}{T^\beta}\,\mathbb{E}\left[\sum_{t:B_t=1} \ell(\mathcal{A}(x_t \mid L^{h^\star}_{B_{<t}}), \text{BinRel}(h^\star(x_t), p))\right] \\
&= \frac{T}{T^\beta}\,\mathbb{E}\left[\mathbb{E}\left[\sum_{t:B_t=1} \ell(\mathcal{A}(x_t \mid L^{h^\star}_{B_{<t}}), \text{BinRel}(h^\star(x_t), p)) \,\middle|\, B\right]\right] \\
&\le \frac{T}{T^\beta}\,\mathbb{E}\left[R(|B|, K)\right],
\end{aligned}$$

where $R(|B|, K)$ is the regret of the algorithm $\mathcal{A}$ and is a sub-linear function of $|B|$. In the last step, we use the fact that $\mathcal{A}$ is a (realizable) online learner for $\mathcal{H}$ w.r.t. ℓ and the feedback that the algorithm received was $(x_t, \text{BinRel}(h^\star(x_t), p))$ in the rounds whenever $B_t = 1$. Again, using Lemma 5.17 from Ceccherini-Silberstein et al. [2017] and Jensen's inequality yields $\mathbb{E}_B\left[R(|B|, K)\right] \le \tilde{R}(T^\beta, K)$, a concave, sub-linear function of T^β. Combining everything, we get

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{Q}(x_t) \neq h_i^{\star,p}(x_t)\}\right] &\le \sum_{t=1}^{T} \mathbb{1}\{h_i^{\star,p}(x_t) \neq y_t\} + \frac{T}{a\,T^\beta}\tilde{R}(T^\beta, K) + \sqrt{2T^{1+\beta}K\ln K} \\
&\le \inf_{h_i^p \in \mathcal{H}_i^p} \sum_{t=1}^{T} \mathbb{1}\{h_i^p(x_t) \neq y_t\} + \frac{T}{a\,T^\beta}\tilde{R}(T^\beta, K) + \sqrt{2T^{1+\beta}K\ln K}
\end{aligned}$$

For any choice of $\beta \in (0, 1)$, the regret above is a sub-linear function of T. Therefore, we have shown that $\mathcal{Q}$ is an agnostic learner for $\mathcal{H}_i^p$ w.r.t. 0-1 loss. This completes our proof. □

E Technical Lemmas

Throughout this section, for any ranking (permutation) $\pi \in \mathcal{S}_K$, we let $\pi_i^j = \mathbb{1}\{\pi_i \le j\}$ for all $(i, j) \in [K]$.

Lemma E.1. *For any $y \in \mathcal{Y}$, $(\pi, \hat{\pi}) \in \mathcal{S}_k$, and $\ell \in \mathcal{L}(\ell^{@p}_{sum})$*

$$\ell(\pi, y) \le \ell(\hat{\pi}, y) + c\,p\,\mathbb{E}_{j\sim Unif([p])}\left[\ell(\pi, \textit{BinRel}(\hat{\pi}, j))\right].$$

where $c = \frac{\max_{\tilde{\pi},y} \ell(\tilde{\pi},y)}{\min_{\tilde{\pi},y}\{\ell(\tilde{\pi},y) \mid \ell(\tilde{\pi},y)\neq 0\}}$.

Proof. Assume that $\ell(\pi, y) > \ell(\hat{\pi}, y) \ge 0$ (as otherwise the inequality trivially holds). Then, since $\ell \in \mathcal{L}(\ell^{@p}_{\text{sum}})$, it must be the case that $\hat{\pi} \overset{[p]}{\neq} \pi$. That is, $\hat{\pi}$ and π assign different ranks to the labels in the top p. Therefore, there exists $i \in [p]$ s.t. $\ell^{@p}_{\text{sum}}(\pi, \text{BinRel}(\hat{\pi}, i)) > 0$. Since $\ell \in \mathcal{L}(\ell^{@p}_{\text{sum}})$, for this same $i \in [p]$, $\ell(\pi, \text{BinRel}(\hat{\pi}, i)) > 0$. Therefore, we have

$$\begin{aligned}
c\,p\,\mathbb{E}_{j\sim\text{Unif}([p])}\left[\ell(\pi, \text{BinRel}(\hat{\pi}, j))\right] &\ge c\ell(\pi, \text{BinRel}(\hat{\pi}, i)) \\
&= \frac{\max_{\tilde{\pi},y} \ell(\tilde{\pi}, y)}{\min_{\tilde{\pi},y}\{\ell(\tilde{\pi}, y) \mid \ell(\tilde{\pi}, y) \neq 0\}}\ell(\pi, \text{BinRel}(\hat{\pi}, i)) \\
&\ge \max_{\tilde{\pi},y} \ell(\tilde{\pi}, y) \\
&\ge \ell(\pi, y).
\end{aligned}$$

Combining the upperbounds in both cases gives the desired inequality. □

Lemma E.2. *For any $y \in \mathcal{Y}$, $(\pi, \hat{\pi}) \in \mathcal{S}_k$, and $\ell \in \mathcal{L}(\ell^{@p}_{prec})$*

$$\ell(\pi, y) \le \ell(\hat{\pi}, y) + c\,\ell(\pi, \textit{BinRel}(\hat{\pi}, p)).$$

where $c = \frac{\max_{\tilde{\pi},y} \ell(\tilde{\pi},y)}{\min_{\tilde{\pi},y}\{\ell(\tilde{\pi},y) \mid \ell(\tilde{\pi},y)\neq 0\}}$.

Proof. Assume that $\ell(\pi, y) > \ell(\hat{\pi}, y) \geq 0$ (as otherwise the inequality trivially holds). Then, since $\ell \in \mathcal{L}(\ell_{\text{prec}}^{@p})$, it must be the case that $\hat{\pi} \overset{p}{\neq} \pi$. That is, $\hat{\pi}$ and π assign different labels in the top p. Therefore, $\ell_{\text{prec}}^{@p}(\pi, \text{BinRel}(\hat{\pi}, p)) > 0$. Since $\ell \in \mathcal{L}(\ell_{\text{prec}}^{@p})$, $\ell(\pi, \text{BinRel}(\hat{\pi}, p)) > 0$. Therefore, we have

$$\begin{aligned} c\,\ell(\pi, \text{BinRel}(\hat{\pi}, p)) &= \frac{\max_{\tilde{\pi}, y} \ell(\tilde{\pi}, y)}{\min_{\tilde{\pi}, y}\{\ell(\tilde{\pi}, y) \mid \ell(\tilde{\pi}, y) \neq 0\}} \ell(\pi, \text{BinRel}(\hat{\pi}, p)) \\ &\geq \max_{\tilde{\pi}, y} \ell(\tilde{\pi}, y) \\ &\geq \ell(\pi, y). \end{aligned}$$

Combining the upperbounds in both cases gives the desired inequality. □

Lemma E.3. *Let $\pi, \hat{\pi} \in \mathcal{S}_k$. Then, for all $(i, j) \in [K] \times [p]$, $\ell_{sum}^{@p}(\pi, \textit{BinRel}(\hat{\pi}, j)) \geq \mathbb{1}\{\pi_i^j \neq \hat{\pi}_i^j\}$.*

Proof. Fix label $i^\star \in [K]$ and threshold $j^\star \in [p]$. Our goal is to show that $\ell_{\text{sum}}^{@p}(\pi, \text{BinRel}(\hat{\pi}, j^\star)) \geq \mathbb{1}\{\pi_{i^\star}^{j^\star} \neq \hat{\pi}_{i^\star}^{j^\star}\}$. Recall that $\text{BinRel}(\hat{\pi}, j^\star)[i^\star] = \mathbb{1}\{\hat{\pi}_{i^\star} \leq j^\star\}$ by definition. Since $\ell_{\text{sum}}^{@p}(\hat{\pi}, \text{BinRel}(\hat{\pi}, j^\star)) = 0$, we have that

$$\begin{aligned} \ell_{\text{sum}}^{@p}(\pi, \text{BinRel}(\hat{\pi}, j^\star)) &= \ell_{\text{sum}}^{@p}(\pi, \text{BinRel}(\hat{\pi}, j^\star)) - \ell_{\text{sum}}^{@p}(\hat{\pi}, \text{BinRel}(\hat{\pi}, j^\star)) \\ &= \sum_{i=1}^{K} \min(\pi_i, p+1)\text{BinRel}(\hat{\pi}, j^\star)[i] - \sum_{i=1}^{K} \min(\hat{\pi}_i, p+1)\text{BinRel}(\hat{\pi}, j^\star)[i] \\ &= \sum_{i=1}^{K} \min(\pi_i, p+1)\mathbb{1}\{\hat{\pi}_i \leq j^\star\} - \sum_{i=1}^{K} \min(\hat{\pi}_i, p+1)\mathbb{1}\{\hat{\pi}_i \leq j^\star\} \\ &= \sum_{i=1}^{K} \min(\pi_i, p+1)\mathbb{1}\{\hat{\pi}_i \leq j^\star\} - \sum_{i=1}^{K} \hat{\pi}_i\mathbb{1}\{\hat{\pi}_i \leq j^\star\} \end{aligned}$$

Let $\mathcal{I} \subseteq [K]$ s.t. for all $i \in \mathcal{I}$, $\hat{\pi}_i^{j^\star} = \mathbb{1}\{\hat{\pi}_i \leq j^\star\} = 1$. Then, we have that

$$\begin{aligned} \ell_{\text{sum}}^{@p}(\pi, \text{BinRel}(\hat{\pi}, j^\star)) &= \sum_{i \in \mathcal{I}} \min(\pi_i, p+1) - \sum_{i \in \mathcal{I}} \hat{\pi}_i \\ &= \sum_{i \in \mathcal{I}} \min(\pi_i, p+1) - \sum_{i=1}^{j^\star} i \end{aligned}$$

Suppose that $\mathbb{1}\{\pi_{i^\star}^{j^\star} \neq \hat{\pi}_{i^\star}^{j^\star}\} = 1$. It suffices to show that $\ell_{\text{sum}}^{@p}(\pi, \text{BinRel}(\hat{\pi}, j^\star)) \geq 1$. There are two cases to consider. Suppose $i^\star \in \mathcal{I}$. Then, it must be the case that $\mathbb{1}\{\pi_{i^\star} \leq j^\star\} = \pi_{i^\star}^{j^\star} = 0$, implying that $\pi_{i^\star} \geq j^\star + 1$. It then follows that in the best case $\sum_{i \in \mathcal{I}} \min(\pi_i, p+1) \geq \sum_{i=1}^{j^\star - 1} i + (j^\star + 1) > \sum_{i=1}^{j^\star} i$ showcasing that indeed $\ell_{\text{sum}}^{@p}(\pi, \text{BinRel}(\hat{\pi}, j)) \geq 1$. Now, suppose $i^\star \notin \mathcal{I}$. Then, $\mathbb{1}\{\hat{\pi}_{i^\star} \leq j^\star\} = 0$, which means that $\mathbb{1}\{\pi_{i^\star} \leq j^\star\} = 1$. Accordingly, while $\hat{\pi}$ did not rank label $i^\star$ in the top $j^\star$, π *did* rank label $i^\star$ in the top $j^\star$. Since $|\mathcal{I}| = j^\star$, there must exist an label $\hat{i} \in \mathcal{I}$ which π does not rank in the top $j^\star$. That is, there exists $\hat{i} \in \mathcal{I}$ s.t. $\pi_{\hat{i}} \geq j^\star + 1$. Using the same logic, in the best case $\sum_{i \in \mathcal{I}} \min(\pi_i, p+1) \geq \sum_{i=1}^{j-1} i + (j^\star + 1)$ showcasing that again $\ell_{\text{sum}}^{@p}(\pi, \text{BinRel}(\hat{\pi}, j^\star)) \geq 1$. Thus, we have shown that when $\mathbb{1}\{\pi_{i^\star}^{j^\star} \neq \hat{\pi}_{i^\star}^{j^\star}\} = 1$, $\ell_{\text{sum}}^{@p}(\pi, \text{BinRel}(\hat{\pi}, j^\star)) \geq 1$. Since $i^\star$ and $j^\star$ were arbitrary, this must be true for any $(i, j) \in [K] \times [p]$, completing the proof. □

Lemma E.4. *Let $\pi, \hat{\pi} \in \mathcal{S}_k$. Then, for all $i \in [K]$, $\ell_{prec}^{@p}(\pi, \textit{BinRel}(\hat{\pi}, p)) \geq \mathbb{1}\{\pi_i^p \neq \hat{\pi}_i^p\}$.*

Proof. Fix label $i^\star \in [K]$. Our goal is to show that $\ell^{@p}_{\text{prec}}(\pi, \text{BinRel}(\hat{\pi}, p)) \geq \mathbb{1}\{\pi^p_{i^\star} \neq \hat{\pi}^p_{i^\star}\}$. Recall that $\text{BinRel}(\hat{\pi}, p)[i^\star] = \mathbb{1}\{\hat{\pi}_{i^\star} \leq p\}$ by definition. Since $\ell^{@p}_{\text{prec}}(\hat{\pi}, \text{BinRel}(\hat{\pi}, p)) = 0$, we have that

$$\begin{aligned}
\ell^{@p}_{\text{prec}}(\pi, \text{BinRel}(\hat{\pi}, p)) &= \ell^{@p}_{\text{prec}}(\pi, \text{BinRel}(\hat{\pi}, p)) - \ell^{@p}_{\text{prec}}(\hat{\pi}, \text{BinRel}(\hat{\pi}, p)) \\
&= \sum_{i=1}^{K} \mathbb{1}\{\hat{\pi}_i \leq p\}\text{BinRel}(\hat{\pi}, p)[i] - \sum_{i=1}^{K} \mathbb{1}\{\pi_i \leq p\}\text{BinRel}(\hat{\pi}, p)[i] \\
&= p - \sum_{i=1}^{K} \mathbb{1}\{\pi_i \leq p\}\mathbb{1}\{\hat{\pi}_i \leq p\}
\end{aligned}$$

Let $\mathcal{I} \subseteq [K]$ s.t. for all $i \in \mathcal{I}$, $\hat{\pi}^p_i = \mathbb{1}\{\hat{\pi}_i \leq p\} = 1$. Then, we have that

$$\ell^{@p}_{\text{prec}}(\pi, \text{BinRel}(\hat{\pi}, p)) = p - \sum_{i \in \mathcal{I}} \mathbb{1}\{\pi_i \leq p\}.$$

Suppose that $\mathbb{1}\{\pi^p_{i^\star} \neq \hat{\pi}^p_{i^\star}\} = 1$. It suffices to show that $\ell^{@p}_{\text{prec}}(\pi, \text{BinRel}(\hat{\pi}, p)) \geq 1$. There are two cases to consider. Suppose $i^\star \in \mathcal{I}$. Then, it must be the case that $\mathbb{1}\{\pi_{i^\star} \leq p\} = \pi^p_{i^\star} = 0$, implying that $\pi_{i^\star} \geq p + 1$. It then follows that in the best case $\sum_{i \in \mathcal{I}} \mathbb{1}\{\pi_i \leq p\} \leq p - 1 < p$ showcasing that indeed $\ell^{@p}_{\text{sum}}(\pi, \text{BinRel}(\hat{\pi}, p)) \geq 1$. Now, suppose $i^\star \notin \mathcal{I}$. Then, $\mathbb{1}\{\hat{\pi}_{i^\star} \leq p\} = 0$, which means that $\mathbb{1}\{\pi_{i^\star} \leq p\} = 1$. Accordingly, while $\hat{\pi}$ did not rank label $i^\star$ in the top p, π *did* rank label $i^\star$ in the top p. Since $|\mathcal{I}| = p$, there must exist an label $\hat{i} \in \mathcal{I}$ which π does not rank in the top p. That is, there exists $\hat{i} \in \mathcal{I}$ s.t. $\pi_{\hat{i}} \geq p + 1$. Using the same logic, in the best case $\sum_{i \in \mathcal{I}} \mathbb{1}\{\pi_i \leq p\} \leq p - 1 < p$ showcasing that again $\ell^{@p}_{\text{prec}}(\pi, \text{BinRel}(\hat{\pi}, p)) \geq 1$. Thus, we have shown that when $\mathbb{1}\{\pi^p_{i^\star} \neq \hat{\pi}^p_{i^\star}\} = 1$, $\ell^{@p}_{\text{prec}}(\pi, \text{BinRel}(\hat{\pi}, p)) \geq 1$. Since $i^\star$ was arbitrary, this must be true for any $i \in [K]$, completing the proof. □