Bipartite Ranking From Multiple Labels: On Loss Versus Label Aggregation

Michal Lukasik [*1] **Lin Chen** [*1] **Harikrishna Narasimhan** [*1] **Aditya Krishna Menon** [*1] **Wittawat Jitkrittum** [1] **Felix X. Yu** [1] **Sashank J. Reddi** [1] **Gang Fu** [1] **Mohammadhossein Bateni** [1] **Sanjiv Kumar** [1]

Abstract

Bipartite ranking is a fundamental supervised learning problem, with the goal of learning a ranking over instances with maximal *area under the ROC curve* (*AUC*) against a *single* binary target label. However, one may often observe *multiple* binary target labels, e.g., from distinct human annotators. How can one synthesize such labels into a *single* coherent ranking? In this work, we formally analyze two approaches to this problem — *loss aggregation* and *label aggregation* — by characterizing their *Bayes-optimal* solutions. We show that while both approaches can yield Pareto-optimal solutions, loss aggregation can exhibit *label dictatorship*: one can inadvertently (and undesirably) favor one label over others. This suggests that label aggregation can be preferable to loss aggregation, which we empirically verify.

1. Introduction

Bipartite ranking is a fundamental supervised learning problem (Freund et al., 2003; Cortes & Mohri, 2003; Agarwal et al., 2005; Clémençon et al., 2008; Kotłowski et al., 2011; Menon & Williamson, 2016), wherein the goal is to learn a ranker that orders "positive" instances over "negative" instances. This is formalized as learning a ranker with maximal *area under the ROC curve* (*AUC*) (Cortes & Mohri, 2003; Agarwal et al., 2005; Krzanowski & Hand, 2009), and is arguably the simplest instantiation of *learning to rank* (Liu, 2009). Bipartite ranking has seen applications in practical problems ranging from medical diagnosis (Swets, 1988; Pepe, 2003) to information retrieval (Ng & Kantor, 2000; Macskassy & Provost, 2001), and is the basis for several other supervised learning problems (Narasimhan & Agarwal, 2013; Reddi et al., 2021; Tang et al., 2022).

[*]Lead authors. [1]Google Research. Correspondence to: Michal Lukasik <mlukasik@google.com>, Lin Chen <linche@google.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

Bipartite ranking conventionally assumes the existence of a *single* binary label denoting whether or not an instance is "positive". However, in practice, there may be *multiple* binary labels, each identifying a different set of "positive" instances. For example, in information retrieval, it is common for there to be distinct label sources (e.g., whether or not a user clicks on a document, whether or not a human rater deems a document to be relevant) (Svore et al., 2011). Similar challenges arise in applications including medical diagnosis (e.g., balancing opinions from multiple experts as to the presence of a certain condition (Verma et al., 2023)), recommendation systems (e.g., balancing click probability with diversity), and computational advertising (e.g., balancing click-through rates with conversion rates); see Appendix E for more discussion of problems involving synthesizing multiple label sources.

Given multiple binary labels, can one produce a *single* coherent ranking over instances? Addressing this requires formalising the *goal* of such a ranking, and then specifying the mechanics of *achieving* this goal. The *multi-objective learning to rank* literature provides guidance on both points (Svore et al., 2011). Typically, the goal here is to find *Pareto optimal* solutions, i.e., solutions that are not dominated across all objectives (Ribeiro et al., 2015). Towards achieving this, previous works have considered two primary approaches: (1) *loss aggregation* (also known as linear scalarization, or weighted sum), where the objective is a weighted combination of per-label objectives (Lin et al., 2019); (2) *label aggregation*, where the objective is formulated on a single target formed from suitable aggregation of the multiple labels, such as a weighted combination (Carmel et al., 2020).

Empirically, both loss and label aggregation have proven successful for multi-objective ranking problems. Theoretically, however, there has been limited analysis providing guidance on the following natural question: *are there reasons to favor label over loss aggregation, or vice-versa*?

In this work, we study this question in the context of bipartite ranking. We study the *Bayes-optimal* scorers for both aggregation approaches, which represent the theoretical minimizers given full access to the underlying statistical distributions. We find that both methods *broadly* yield comparable optimal rankings, which furthermore are Pareto op-

Objective	Defn.	Cost $c_{\bar{y}\bar{y}'}$	Stochastic labels (Any K)	Deterministic labels ($K = 2$)	Pareto-opt	Dictatorship	Prop.
Loss aggregation	(4)	N/A	$\sum_k \alpha^{(k)} \cdot \eta^{(k)}(x)$	$\alpha^{(1)} \cdot \eta^{(1)}(x) + \alpha^{(2)} \cdot \eta^{(2)}(x)$	✓	✓	5.2, 6.1
Label aggregation with $\bar{\mathsf{Y}} = \sum_k \mathsf{Y}^{(k)}$	(5)	1	No closed form	$\eta^{(1)}(x) + \eta^{(2)}(x)$	✗	✗	5.3, 6.2
		$\mathbf{1}(\bar{y} > \bar{y}') \cdot \lvert\bar{y} - \bar{y}'\rvert$	$\sum_k \eta^{(k)}(x)$	$\eta^{(1)}(x) + \eta^{(2)}(x)$	✓	✗	5.4

Table 1. Bayes-optimal scorer for two approaches to bipartite ranking with K binary labels. Here, we assume that we have random variables $(\mathsf{X}, \mathsf{Y}^{(1)}, \ldots, \mathsf{Y}^{(K)})$ representing instances and labels drawn from some joint distribution, with $\eta^{(k)}(x) \doteq \mathbf{P}(\mathsf{Y}^{(k)} = 1 \mid \mathsf{X} = x)$ denoting the marginal class-probability function for each label. Suitable instantiations of both loss aggregation and label aggregation satisfy a notion of Pareto-optimality; however, the loss aggregation approach results in an undesirable "label dictatorship" phenomenon, wherein one label dominates the other (see §6.1). The label aggregation objective in (5) requires the specification of pairwise costs $c_{\bar{y}\bar{y}'}$. The constants $\alpha^{(k)} = a_k / (\pi^{(k)} \cdot (1 - \pi^{(k)}))$ arise in Proposition 6.1, where $\pi^{(k)} \doteq \mathbf{P}(\mathsf{Y}^{(k)} = 1)$ denotes the label priors.

timal. However, we explicate how Pareto optimality – while providing a foundational concept for multi-objective problems – may by itself be insufficient for a desirable practical solution. Indeed, upon closer inspection, we demonstrate that loss aggregation can lead to a *label dictatorship* phenomenon, wherein certain labels are implicitly favored over others. This suggests that label aggregation can be preferred over loss aggregation, which we validate empirically.

In summary, our contributions are (cf. Table 1):

(i) We formalize the problem of bipartite ranking from multiple labels (§3), and the instantiation of loss aggregation and label aggregation in this setting (§4);

(ii) We characterize the *Bayes-optimal solutions* to both loss (§5.1) and label aggregation (§5.2), and establish Pareto-optimality of (suitable instantiations of) both methods;

(iii) We show that loss aggregation can lead to an undesirable *label dictatorship* phenomenon (§6), and empirically validate this on synthetic and real-world datasets (§7).

Our study serves to analyze the theoretical properties of loss and label aggregation, rather than constructing novel algorithms. This analysis, however, yields practical insights for selecting amongst these methods.

2. Background and Notation

Learning to rank problems seek to learn a scorer that orders instances according to an underlying utility score (Liu, 2009). The *area under the ROC curve* (*AUC*) is a traditional metric used to evaluate the efficacy of such a ranking (Cortes & Mohri, 2003; Agarwal et al., 2005; Clémençon et al., 2008; Menon & Williamson, 2016). Below, we define AUC for problems with binary and multi-class labels.

2.1. Bipartite AUC

Consider a supervised learning problem with input space $\mathcal{X}$, binary labels $\mathcal{Y} = \{0, 1\}$, and distribution D over $\mathcal{X} \times \mathcal{Y}$. Let (X, Y) be random variables distributed according to D. Denote by $\eta(x) \doteq \mathbf{P}(\mathsf{Y} = 1 | \mathsf{X} = x)$ the *class-probability function* and $\pi \doteq \mathbf{P}(\mathsf{Y} = 1)$ the positive class *prior*.

Our goal is to learn a scorer $f \colon \mathcal{X} \to \mathbb{R}$ that ranks the positive examples (i.e., those with $\mathsf{Y} = 1$) over the negative ones, as quantified by the area under the ROC curve, or AUC.

Definition 2.1 (Agarwal & Niyogi (2009))**.** For any scorer $f \colon \mathcal{X} \to \mathbb{R}$ and distribution D, the (bipartite) AUC is:

$$\mathrm{AUC}(f; D) \doteq \mathbb{E}_{\mathsf{X},\mathsf{X}'} \left[H(f(\mathsf{X}) - f(\mathsf{X}')) \mid \mathsf{Y} > \mathsf{Y}' \right] \quad (1)$$
$$H(z) \doteq \mathbf{1}(z > 0) + \frac{1}{2} \cdot \mathbf{1}(z = 0).$$

Intuitively, the AUC is the fraction of pairs $(x, x') \in \mathcal{X} \times \mathcal{X}$ with positive and negative labels wherein f scores the positive over the negative, with a reward of 0.5 for ties.

Given some fixed distribution D, a *Bayes-optimal* scorer f^* is one that achieves the *highest possible* AUC; i.e., $\mathrm{AUC}(f^*; D) \geq \mathrm{AUC}(f; D)$ for any $f \colon \mathcal{X} \to \mathbb{R}$. One may employ the Neyman-Pearson lemma (Lehmann & Romano, 2005) to establish these scorers closely follow $\eta(x)$.

Proposition 2.2 (Clémençon et al. (2008); Menon & Williamson (2016))**.** *For any distribution D, any Bayes-optimal AUC scorer f^* satisfies*

$$(\forall x, x' \in \mathcal{X})\, \eta(x) > \eta(x') \implies f^*(x) > f^*(x'),$$

or equally, η is any non-decreasing transformation of f^*.

We remark here that neither the AUC nor its Bayes-optimal scorer depend on the class prior π. Thus, the (population) AUC is invariant to the amount of label skew, supporting its common usage as a metric in problems characterized by label imbalance (Ling & Li, 1998; Menon et al., 2013).

2.2. Multipartite AUC

One may extend bipartite ranking to *ordinal multi-class* labels $\mathcal{Y} = \{1, 2, \ldots, L\}$, wherein higher label values denote higher presence of a certain attribute (e.g., star ratings denoting user's item preferences). In this case, our goal is to learn a scorer that ranks examples with higher labels over examples with lower labels. Further, one may apply variable *costs* on mis-ranking of pairs with different labels.

Formally, let D_{mp} denote a distribution over $\mathcal{X} \times \mathcal{Y}$. As before, we denote the conditional-class probability function by $\eta_y(x) = \mathbf{P}(\mathsf{Y} = y \mid \mathsf{X} = x)$, and the class priors by $\pi_y = \mathbf{P}(\mathsf{Y} = y)$. Further, let $c_{yy'} \geq 0$ denote the cost of scoring an instance with label $y \in \mathcal{Y}$ below an instance with label $y' \in \mathcal{Y}$. The following is an adaptation of the bipartite AUC (Definition 2.1) to multi-class problems.

Definition 2.3 (Uematsu & Lee (2015))**.** For any scorer $f \colon \mathcal{X} \to \mathbb{R}$, distribution D_{mp}, and costs $\{c_{yy'} \geq 0 \colon y, y' \in \mathcal{Y}\}$, the multi-partite AUC is:

$$\mathrm{AUC}_{\mathrm{mp}}(f; D_{\mathrm{mp}}) \doteq \mathbb{E}_{\mathsf{X}}\mathbb{E}_{\mathsf{X}'}\left[c_{\mathsf{Y}\mathsf{Y}'} \cdot H(f(\mathsf{X}) - f(\mathsf{X}')) \mid \mathsf{Y} > \mathsf{Y}'\right]. \tag{2}$$

When $L = 2$, the objective reduces to a scaling of (12). Unlike the bipartite case, the multipartite AUC does *not* admit a tractable Bayes-optimal scorer in general. An exception is when either $L = 3$, or the costs satisfy the following *scale condition*: for suitable constants $w_y, s_y \geq 0$,

$$(\forall y, y' \in \mathcal{Y})\, c_{yy'} = w_y \cdot w_{y'} \cdot (s_y - s_{y'}) \cdot \mathbf{1}(y > y'), \tag{3}$$

For example, (3) holds when $c_{yy'} = (y - y') \cdot \mathbf{1}(y > y')$. Under such a condition, we have the following.

Proposition 2.4 (Uematsu & Lee (2015, Theorem 3))**.** *Suppose $L = 3$, or the costs $\{c_{yy'} \geq 0 \colon y, y' \in \mathcal{Y}\}$ satisfy the scale condition (3). For any distribution D_{mp}, any Bayes-optimal multi-partite AUC scorer f^* satisfies*

$$(\forall x, x' \in \mathcal{X})\, \beta(x) > \beta(x') \implies f^*(x) > f^*(x')$$
$$\beta(x) \doteq \frac{\sum_{i=2}^{L} c_{1,i} \cdot \eta_i(x)}{\sum_{i=1}^{L-1} c_{i,L} \cdot \eta_i(x)},$$

or equally, β is any non-decreasing transformation of f^*.

3. Bipartite Ranking With Multiple Labels

We now formalize our setting of interest — bipartite ranking from *multiple* labels — and the core goal of identifying a *Pareto optimal* solution with respect to these labels.

3.1. Formal setup

Consider a setting where we have access to *multiple* labels for each instance $x \in \mathcal{X}$, and would like to produce a single coherent ranking over instances. For simplicity, we begin by assuming each individual label is *binary*. Formally, for integer $K \geq 2$, $(\mathsf{X}, \mathsf{Y}^{(1)}, \ldots, \mathsf{Y}^{(K)})$ be random variables distributed according to a joint distribution D^{jnt} over $\mathcal{X} \times \{0, 1\}^K$. Let μ denote the marginal distribution of X. For each $k \in [K]$, we have an induced marginal distribution $D^{(k)}$ over $(\mathsf{X}, \mathsf{Y}^{(k)})$. Let $\eta^{(k)}(x) \doteq \mathbf{P}(\mathsf{Y}^{(k)} = 1 \mid \mathsf{X} = x)$ denote the marginal class-probability function of each label $k \in [K]$, and $\pi^{(k)} \doteq \mathbf{P}(\mathsf{Y}^{(k)} = 1)$ the class prior.

Our goal remains to produce a *single* scorer $f \colon \mathcal{X} \to \mathbb{R}$. To do so, we must specify a concrete metric for assessing f.

3.2. Pareto optimal per-label AUC maximisation

One natural summary of f's performance is the *per-label AUC vector*, i.e., $(\mathrm{AUC}(f; D^{(1)}), \ldots, \mathrm{AUC}(f; D^{(K)}))$. The problem of synthesizing such a vector into a single metric falls within the purview of multi-objective optimization (Ehrgott, 2005). Adapting a standard goal from this literature, it is natural to seek scorers that are *Pareto optimal* (Ehrgott, 2005) with respect to the per-label AUCs.

Definition 3.1 (Pareto dominance)**.** We say a scorer $f \colon \mathcal{X} \to \mathbb{R}$ *Pareto dominates* another $g \colon \mathcal{X} \to \mathbb{R}$ with respect to distributions $\{D^{(k)}\}_{k \in [K]}$ if and only if:

(1) $\forall k \in [K] : \mathrm{AUC}(g; D^{(k)}) \geq \mathrm{AUC}(f; D^{(k)})$

(2) $\exists k \in [K] : \mathrm{AUC}(g; D^{(k)}) > \mathrm{AUC}(f; D^{(k)})$.

Definition 3.2 (Pareto optimality)**.** For any $\{D^{(k)}\}_{k \in [K]}$, the set of Pareto optimal scorers $\mathcal{F}_{\mathrm{PFS}}$ comprises scorers that are *not* Pareto dominated by any other scorer.

Essentially, a scorer is Pareto optimal if no other scorer dominates it across at least one AUC objective, while not harming it across all AUC objectives.

3.3. Does Pareto optimality suffice?

Pareto optimality is a necessary, but not sufficient condition for a solution to be practically useful: a model that aggressively optimizes one objective at the complete expense of others, while being Pareto optimal, can be undesirable.

For example, consider a document retrieval problem, with the goal of satisfying various aspects of a user query; e.g., an ideal system should return documents that are both *relevant* to the query and likely to *engage* the user (He et al., 2023). Relevance can be assessed through annotation from a human or machine expert, while engagement can be measured through metrics such as click-through rate or revenue.

In practice, the goals of relevance and engagement may be at odds with each other. To illustrate this point, consider queries from the MS MARCO benchmark dataset (Bajaj et al., 2018). We predict the engagement and relevance of different documents for a query by prompting the Gemini model (Team, 2024) (see Table 5 in Appendix for details). Illustrative queries, documents, and predicted engagement and relevance scores are reported in Table 2. For the query `“What is another name for rust?”`, a document with high relevance but low engagement contains the right answer, albeit with an obscure presentation. Conversely, a document with high engagement but low relevance superficially matches keywords, but lacks the correct answer.

On the other hand, for the query `"How much does it cost`

Query	Low relevance & High engagement document	High relevance & Low engagement document
"What is another name for rust?"	First of all rust is formed when iron is exposed to both oxygen and water/ water vapopurs. The formula for rust is Fe203.xH20. Now the x varies which determines the extent to which rust is formed. Basically Rust is formed throughout the surface of the iron thus preventing rusting of the inner layers.	Rust is the result of the oxidation of iron. The most common cause is prolonged exposure to water. Any metal that contains iron, including steel, will bond with the oxygen atoms found in water to form a layer of iron oxide, or rust. Rust will increase and speed up the corrosion process, so upkeep is important.here are two ways you can use a potato to remove rust: 1 Simply stab the knife into potato and wait a day or overnight. 2 Slice a potato in half, coat the inside with a generous portion of baking soda, and go to town on the rusted surface with the baking soda-coated potato
"How much does it cost to build a deck with a hot tub?"	Once you determine what kind of decking materials you'll need to build the structure, it's time to get down to price. While the average cost to build a deck averages between $4,000 and $10,000, that doesn't account for the materials.Here is the average cost of each decking material, broken down by average price range per board: It's important to know what board sizes you'll need to purchase for your deck.nce you determine what kind of decking materials you'll need to build the structure, it's time to get down to price. While the average cost to build a deck averages between $4,000 and $10,000, that doesn't account for the materials.	1 Pre-fabricated hot tubs cost less, but are still priced between $3,000 and $8,000 (for the smaller-sized models). 2 Although you give up some features and styling with inflatable, you can get a good-quality inflatable hot tub starting at about $500. 3 There is quite a difference in price here.

Table 2. An illustration of the trade-off between engagement and relevance across queries and documents from the MS MARCO dataset (Bajaj et al., 2018). A purely relevance-driven approach recommends the documents from the rightmost column, which contains the correct answer, albeit with an unclear presentation. Conversely, a purely engagement-driven approach recommends the documents from the first column, which may superficially match keywords, but lack the correct answer. While both solutions may be Pareto optimal, in practice, one may favor only one of these solutions; further, the particular choice can vary depending on the query. Therefore, a solution which *globally* favors one signal over the other (i.e., where one label is a *dictator*) may be undesirable.

`to build a deck with a hot tub?"`, a document with high relevance but low engagement addresses the specific question of the hot tub pricing, but ignores the more general context of the price for building the deck. A document with high engagement but low relevance misses the hot tub aspect, but does answer the general question of deck pricing.

These examples highlight a common practical challenge: a trade-off often exists between different objectives (in this case, relevance and engagement). Further, while scorers focusing exclusively on one metric could be Pareto optimal – specifically, if no single document excels at both metrics simultaneously – such solutions may be practically undesirable. Indeed, solely optimizing relevance at the expense of engagement may be desirable for the first query (where the document contains the correct answer), but less desirable for the second query (where the query intent is open-ended).

Therefore, alongside ensuring Pareto optimality, it is crucial to assess whether a solution does not overly favor one of the labels, e.g. via the gap between the per-objective AUC scores (as we consider in §7). This motivates a deeper investigation into the precise forms of Bayes-optimal solutions for different aggregation approaches, allowing us to understand their inherent behaviors and potential biases.

4. The Loss and Label Aggregation Methods

We now formalize two natural approaches for bipartite ranking with multiple binary labels. These aggregate either the *losses* over multiple labels, or the *labels* themselves.

4.1. Loss aggregation

A common strategy in the multi-objective optimization literature to achieve Pareto optimal solutions is to optimize a linear combination of the objectives (Ruchte & Grabocka, 2021). Such an approach is typically referred to as *linear scalarization*, and may be seen as performing *loss aggregation*. In our case, this amounts to maximizing

$$\sum_{k\in[K]} a_k \cdot \mathrm{AUC}(f; D^{(k)}),$$

for mixing coefficients $a_1, \ldots, a_k > 0$. This is equivalent to the following *loss aggregated* AUC objective.

Definition 4.1 (Loss aggregation)**.** For any scorer $f\colon \mathcal{X} \to \mathbb{R}$, distribution D^{jnt} over $(\mathsf{X}, \mathsf{Y}^{(1)}, \ldots, \mathsf{Y}^{(K)})$, and weights $\{a_k > 0\}_{k\in[K]}$, the *loss aggregated AUC* is:

$$\begin{aligned}&\mathrm{AUC}_{\mathrm{LoA}}(f; D^{\mathrm{jnt}}) \\ &\doteq \sum_{k\in[K]} \mathbb{E}_{\mathsf{X},\mathsf{X}'}\Big[a_k \cdot H(f(\mathsf{X}) - f(\mathsf{X}'))\Big|\mathsf{Y}^{(k)} > \mathsf{Y}'^{(k)}\Big].\end{aligned} \tag{4}$$

For brevity, we subsequently omit the dependence of $\mathrm{AUC}_{\mathrm{LoA}}(f)$ on D^{jnt}. Given a finite sample $\{(x^{(i)}, y^{(i,1)}, \ldots, y^{(i,k)})\}_{i\in[N]}$ drawn from D^{jnt}, the empirical loss aggregated AUC is

$$\widehat{\mathrm{AUC}}_{\mathrm{LoA}}(f) \propto \sum_{k\in[K]} \sum_{(i,j)\in P^{(k)}} a_k \cdot H(f(x^{(i)}) - f(x^{(j)})),$$

where $P^{(k)} \doteq \{(i,j) \in [N]\times[N] \colon \mathbf{1}(y^{(i,k)} > y^{(j,k)})\}$.

4.2. Label aggregation

Weighting the individual per-label AUCs is conceptually simple. An alternative approach is to combine the K labels $(\mathsf{Y}^{(1)}, \ldots, \mathsf{Y}^{(K)})$ via some *aggregation function* ψ to obtain a single new *aggregated* label $\bar{\mathsf{Y}}$ (Svore et al., 2011; Agarwal et al., 2011; Dai et al., 2011; Carmel et al., 2020; Wei et al., 2023). Given such an aggregated label, one may then maximize the *multi-partite* AUC (2) on $\bar{\mathsf{Y}}$, as follows:

Definition 4.2 (Label aggregation)**.** For any scorer $f \colon \mathcal{X} \to \mathbb{R}$, distribution D^{jnt} over $(\mathsf{X}, \mathsf{Y}^{(1)}, \ldots, \mathsf{Y}^{(K)})$, aggregation function $\psi : \mathcal{Y}^K \to \bar{\mathcal{Y}}$, and costs $\{c_{\bar{y}\bar{y}'} \geq 0\}_{\bar{y},\bar{y}' \in \bar{\mathcal{Y}}}$, the *label aggregated AUC* is:

$$\begin{aligned} &\mathrm{AUC}_{\mathrm{LaA}}(f; D^{\text{jnt}}) \\ &\doteq \mathbb{E}_{\mathsf{X},\mathsf{X}'}\left[c_{\bar{\mathsf{Y}}\bar{\mathsf{Y}}'} \cdot H(f(\mathsf{X}) - f(\mathsf{X}')) \mid \bar{\mathsf{Y}} > \bar{\mathsf{Y}}'\right] \end{aligned} \tag{5}$$

where $\bar{\mathsf{Y}} = \psi(\mathsf{Y}^{(1)}, \ldots, \mathsf{Y}^{(K)})$ and $\bar{\mathcal{Y}} \subseteq [K]$.

Given a finite sample $\{(x^{(i)}, y^{(i,1)}, \ldots, y^{(i,k)})\}_{i \in [N]}$ drawn from D^{jnt}, and $\bar{y}^{(i)} \doteq \sum_{k \in [K]} y^{(i,k)}$, the empirical label aggregated AUC is

$$\widehat{\mathrm{AUC}}_{\mathrm{LaA}}(f) \propto \sum_{(i,j) \in P} c_{\bar{y}^{(i)},\bar{y}^{(j)}} \cdot H(f(x^{(i)}) - f(x^{(j)})),$$

where $P \doteq \{(i,j) \in [N] \times [N] \colon \mathbf{1}(\bar{y}^{(i)} > \bar{y}^{(j)})\}$.

There are several natural choices of aggregation function. Per §3.1, we consider K individual binary labels $\mathsf{Y}^{(k)} \in \{0,1\}$. One approach is to sum these binary labels: $\psi(\mathsf{Y}^{(1)}, ..., \mathsf{Y}^{(K)}) = \sum_{k \in [K]} \mathsf{Y}^{(k)} \in \{0, 1, ..., K\}$. This results in an integer ordinal value, representing the *count* of positive labels for an instance x. Another natural label aggregation mechanism could be to take the product of the labels: $\psi(\mathsf{Y}^{(1)}, \ldots, \mathsf{Y}^{(K)}) = \prod_{k \in [K]} \mathsf{Y}^{(k)} \in \{0, 1\}$.

5. Bayes-Optimal Scorers Under Aggregation

We now characterize the set of Bayes-optimal scorers for loss aggregation (Definition 4.1) and label aggregation (Definition 4.2). This shall be a step towards understanding the solutions provided by each method.

5.1. Bayes-optimal scorers for loss aggregation

We first show that any scorer that is optimal for loss aggregation is also Pareto optimal; this follows straight-forwardly from Ruchte & Grabocka (2021).

Proposition 5.1. *For any distributions $\{D^{(k)}\}_{k \in [K]}$ and mixing weights $\{a_k \geq 0\}_{k \in [K]}$, a scorer $f^* : \mathcal{X} \to \mathbb{R}$ that maximizes the loss aggregated AUC in (4) is Pareto optimal.*

As noted in Section 3, Pareto optimality alone may not suffice to assess the usefulness of a solution. We therefore take a closer look at the form of the Bayes-optimal scorer.

Proposition 5.2. *For any distributions $\{D^{(k)}\}_{k \in [K]}$ and mixing weights $\{a_k \geq 0\}_{k \in [K]}$, any Bayes-optimal scorer f^* for the loss aggregated AUC satisfies*

$$\gamma(x) > \gamma(x') \implies f^*(x) > f^*(x')$$

$$\gamma(x) \doteq \frac{1}{K} \sum_{k \in [K]} \frac{a_k}{\pi^{(k)} \cdot (1 - \pi^{(k)})} \cdot \eta^{(k)}(x), \tag{6}$$

or equally, γ is some non-decreasing transformation of f^*.

The final expression is intuitive: the optimal scorers involve a weighted sum of individual class-probability functions $\eta^{(k)}$. In fact, when $K = 1$, this reduces to the standard result for the AUC (Lemma 2.2).

5.2. Bayes-optimal scorers for label aggregation

Unlike loss aggregation, label aggregation may *not* always produce Pareto optimal solutions. We show this below with additive label aggregation function and with costs $c_{\bar{y}\bar{y}'} = 1$.

Proposition 5.3. *Suppose $\psi(\mathsf{Y}^{(1)}, \ldots, \mathsf{Y}^{(K)}) = \sum_k \mathsf{Y}^{(k)}$, and $c_{\bar{y}\bar{y}'} = 1, \forall \bar{y}, \bar{y}' \in \bar{\mathcal{Y}}$. There exists a set of distributions $\{D^{(k)}\}_{k \in [K]}$, mixing weights $\{a_k \geq 0\}_{k \in [K]}$, and a scorer $f^* : \mathcal{X} \to \mathbb{R}$ such that f^* maximizes the label aggregated AUC in (5), but is* not *Pareto optimal.*

Interestingly, this issue can be remedied with a simple change to the AUC objective. Specifically, we choose the misranking costs to be $c_{\bar{y}\bar{y}'} = \mathbf{1}(\bar{y} > \bar{y}') \cdot |\bar{y} - \bar{y}'|$ in (5), where the cost of misranking a pair of instances scales linearly with the difference in their aggregated labels. We can then show that the optimal scorers are obtained through simple summation of the class probability functions $\eta^{(k)}(x)$; thus, the Bayes-optimal scorers are also Pareto optimal.

Proposition 5.4. *Suppose $\psi(\mathsf{Y}^{(1)}, \ldots, \mathsf{Y}^{(K)}) = \sum_k \mathsf{Y}^{(k)}$, and $c_{\bar{y}\bar{y}'} = \mathbf{1}(\bar{y} > \bar{y}') \cdot |\bar{y} - \bar{y}'|$. For any distributions $\{D^{(k)}\}_{k \in [K]}$, any Bayes-optimal scorer f^* for the label-aggregated AUC in (5) satisfies*

$$\gamma(x) > \gamma(x') \implies f^*(x) > f^*(x')$$

$$\gamma(x) \doteq \sum_{k \in [K]} \eta^{(k)}(x); \tag{7}$$

or equally, γ is some non-decreasing transformation of f^*. Furthermore, f^* is Pareto optimal *w.r.t. $\{D^{(k)}\}_{k \in [K]}$.*

The simple form of (7) critically relies on choosing costs $c_{\bar{y}\bar{y}'} = \mathbf{1}(\bar{y} > \bar{y}') \cdot |\bar{y} - \bar{y}'|$ in the label aggregated AUC objective (5). As established by Uematsu & Lee (2015) and utilized in our proof (Appendix B), these costs ensure that the optimal scorer follows $\mathbb{E}[\bar{\mathsf{Y}} \mid \mathsf{X} = x]$, which equals $\sum_k \eta^{(k)}(x)$ by linearity of expectation. Other costs (e.g., uniform $c_{\bar{y}\bar{y}'} = 1$, per Proposition 5.3) do not generally lead to this direct summation of $\eta^{(k)}$ as the optimal scorer.

We also note that for an *arbitrary* set of costs $c_{\bar{y}\bar{y}'}$, the Bayes-optimal scorer will in general *not* admit a tractable closed-form solution. This is not surprising: indeed, even when $K = 1$, the multi-partite AUC does not have a tractable closed-form scorer in general (Uematsu & Lee, 2015). However, for the specific choice of *uniform costs* $c_{\bar{y}\bar{y}'} = 1$, we present scorers that are *asymptotically* optimal for $\mathrm{AUC}_{\mathrm{LA}}$ (in the limiting case of $K \to \infty$) under some distributional assumptions; see Theorem C.1 and Corollary C.2 in §C.

5.3. Contrasting the Bayes-optimal scorers

Upon closer inspection, the optimal solution to loss aggregation in (6) reveals a subtlety: for $K > 1$, even for a uniform weighting of the individual label AUCs (i.e., $a_k = 1, \forall k$), the optimal scorers depend on the underlying marginal class priors $\pi^{(k)}$. Conceptually, this is a surprise: indeed, it is in contrast to the standard behaviour when $K = 1$, wherein the class priors do *not* influence the AUC optimizer.

Practically, this suggests that the choice of the mixing weights a_k determines the extent to which the optimal scorer favors one label over the others. More precisely, the optimal scorer *favors labels that are marginally skewed* ($\pi^{(k)}$ is away from 0.5), over labels that are balanced ($\pi^{(k)} \approx 0.5$). As we discuss in §6.1, this can lead to an objective that inadvertently favors certain labels based purely on their dataset-wide imbalance, rather than their instance-specific information content, potentially leading to a "label dictatorship" (see Appendix F for an illustrative example). This is unexpected: one might expect that a uniform weighting ($a_k = 1$) would induce scorers that treat all labels equitably.

It is critical to distinguish the *marginal* probabilities $\pi^{(k)} \doteq \mathbf{P}(\mathsf{Y}^{(k)} = 1)$ from the *instance-conditional* probabilities $\eta^{(k)}(x) \doteq \mathbf{P}(\mathsf{Y}^{(k)} = 1 \mid \mathsf{X} = x)$. A label can be perfectly balanced marginally (i.e., $\pi^{(k)} = 0.5$) while being completely noise-free conditionally (i.e., $\eta^{(k)}(x) \in \{0, 1\}$ for all x). While weighting by the inverse marginal skew $1/[\pi^{(k)}(1 - \pi^{(k)})]$ might appear statistically natural, this may not align with the conditional quality of the signals.

By contrast, under label aggregation, the optimal scorers equally balance each $\eta^{(k)}$. Crucially, the optimal scorers do *not* incorporate additional weighting under the hood, and thus *avoid favoring one set of labels over the others*. Further analysis (see Appendix D) shows this optimal scorer (Proposition 5.4) behaves predictably with (anti-)correlated labels and can be generalized with explicit non-uniform weights, maintaining direct sensitivity to these weights without the hidden prior-dependencies of loss aggregation.

6. The Perils of Loss Aggregation

To better highlight the differences between loss aggregation and label aggregation, we consider the special case of deterministic labels, where $\mathbf{P}(\mathsf{Y}^{(1)} = y_1, \ldots, \mathsf{Y}^{(K)} = y_K) \in \{0, 1\}, \forall x \in \mathcal{X}, \mathbf{y} \in \{0, 1\}^K$. Here, we demonstrate that loss aggregation can exhibit undesirable "label dictatorship" behavior, wherein one label is favored over another.

We emphasize that the assumption of deterministic labels is for illustrative clarity of the "label dictatorship" phenomenon (Proposition 6.1, Figure 1); our main results on Bayes-optimal scorers (Proposition 5.2 and Proposition 5.4) hold for general $\eta^{(k)}(x)$, as validated in §7.

6.1. Loss aggregation can yield label dictatorship

We first show that the optimal scorer for loss aggregation exhibits a certain *dictatorial* behavior under the deterministic label setting. This is easy to see when $K = 2$.

Proposition 6.1. *Suppose* $\eta^{(k)}(x) \in \{0, 1\}, \forall x \in \mathcal{X}$ *and* $K = 2$. *For mixing weights* $a_1, a_2 \geq 0$, *denote* $\alpha^{(k)} = \frac{a_k}{\pi^{(k)} \cdot (1 - \pi^{(k)})}, k \in \{1, 2\}$. *Then any Bayes-optimal scorer* f^* *for the loss aggregated AUC in* (4) *satisfies:*

$$\begin{aligned} &\textit{If } \alpha^{(1)} > \alpha^{(2)}: \\ &\qquad \eta^{(1)}(x) > \eta^{(1)}(x') \implies f^*(x) > f^*(x'); \\ &\textit{If } \alpha^{(2)} > \alpha^{(1)}: \\ &\qquad \eta^{(2)}(x) > \eta^{(2)}(x') \implies f^*(x) > f^*(x'). \end{aligned}$$

When $\alpha^{(1)} > \alpha^{(2)}$, the ranking is predominantly determined by $\eta^{(1)}$, and when $\alpha^{(1)} < \alpha^{(2)}$, the ranking is predominantly determined by $\eta^{(2)}$; i.e., *one of the labels acts as a "dictator"* in determining the optimal solution. This "label dictatorship" behavior can be undesirable in practice. Even if both labels provide perfectly clean conditional signals (i.e., $\eta^{(k)}(x) \in \{0, 1\}$), a label might dominate the ranking due to its marginal skewness, rather than its contextual importance. An example illustrating such an undesirable scenario in information retrieval is provided in Appendix F. In contrast, label aggregation results in scorers that equally balance both labels even under uniform costs $c_{\bar{y}\bar{y}'} = 1$.

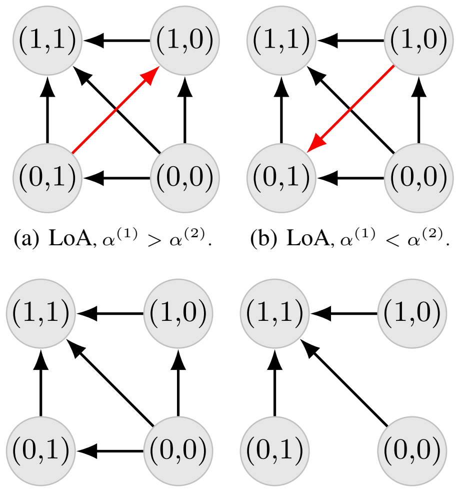
Proposition 6.2. *Suppose* $\eta^{(k)}(x) \in \{0, 1\}, \forall x \in \mathcal{X}$ *and* $K = 2$. *Then for both costs* $c_{\bar{y}\bar{y}'} = 1$ *and* $c_{\bar{y}\bar{y}'} = 1(\bar{y} > \bar{y}') \cdot |\bar{y} - \bar{y}'|$, *any Bayes-optimal scorer* f^* *for the label-aggregated AUC in* (5) *satisfies*

$$\begin{aligned} &\gamma(x) > \gamma(x') \implies f^*(x) > f^*(x') \\ &\gamma(x) \doteq \eta^{(1)}(x) + \eta^{(2)}(x); \end{aligned} \tag{8}$$

or equally, γ *is some non-decreasing transformation of* f^*.

6.2. Illustration of label dictatorship

We illustrate the above in Figure 1 by showing the partial orders induced by the two objectives for two deterministic binary labels $(\mathsf{Y}^{(1)}, \mathsf{Y}^{(2)})$. Each circle denotes an instance with a combination of $(\mathsf{Y}^{(1)}, \mathsf{Y}^{(2)})$, and the arrows indicate that a particular combination is ranked below another. Based on the value of $\alpha^{(1)}$ and $\alpha^{(2)}$, loss aggregation either *always* ranks the label combination $(1, 0)$ above $(0, 1)$, or always ranks $(1, 0)$ below $(0, 1)$. By contrast, label aggregation does not induce any ordering between $(1, 0)$ and $(0, 1)$.



(a) LoA, $\alpha^{(1)} > \alpha^{(2)}$. (b) LoA, $\alpha^{(1)} < \alpha^{(2)}$.

(c) LaA, sum. (d) LaA, product.

Figure 1. Illustration of partial orders among examples for two deterministic binary labels induced by: (a) Loss aggregation with $\alpha^{(1)} > \alpha^{(2)}$; (b) Loss aggregation with $\alpha^{(1)} < \alpha^{(2)}$; (c) Label Aggregation with $\bar{\mathsf{Y}} = \mathsf{Y}^{(1)} + \mathsf{Y}^{(2)}$; (d) Label Aggregation with $\bar{\mathsf{Y}} = \mathsf{Y}^{(1)} \cdot \mathsf{Y}^{(2)}$. As before, $\alpha^{(k)} = \frac{a_k}{\pi^{(k)} \cdot (1-\pi^{(k)})}$. Each circle denotes an example with values for the two binary labels. An arrow from a circle marked $(\mathsf{Y}^{(1)}, \mathsf{Y}^{(2)})$ to a circle marked $(\tilde{\mathsf{Y}}^{(1)}, \tilde{\mathsf{Y}}^{(2)})$ indicates that the label combination $(\mathsf{Y}^{(1)}, \mathsf{Y}^{(2)})$ is ranked below $(\tilde{\mathsf{Y}}^{(1)}, \tilde{\mathsf{Y}}^{(2)})$. The red arrow for the loss aggregation methods depicts the *dictatorial* behavior described in Section 6.1.

To make this point concrete, recall our MS MARCO example (Section 3.3), where $\mathsf{Y}^{(1)}$ represents `relevance` and $\mathsf{Y}^{(2)}$ represents `engagement`. Then, node $(1, 0)$ conceptually corresponds to the document with high relevance but low engagement (the rightmost column in Table 2), while $(0, 1)$ corresponds to the document with low relevance but high engagement (the middle column in Table 2).

The "label dictatorship" behavior of loss aggregation — highlighted by the red arrows in Figures 1(a) and 1(b) — translates to imposing a strict, global preference between high-relevance & low-engagement and low-relevance & high-engagement documents. This preference is determined by the relative values of $\alpha^{(1)} = a_1/(\pi^{(1)}(1-\pi^{(1)}))$ and $\alpha^{(2)} = a_2/(\pi^{(2)}(1-\pi^{(2)}))$ (cf. Proposition 6.1). Specifically, the ranking will favor relevance if $\alpha^{(1)} > \alpha^{(2)}$, and engagement if the inequality is reversed. These α values depend on the mixing weights a_k and the marginal label skews $\pi^{(k)}$, regardless of the specific query or document. As argued in §3.3, such a fixed global trade-off might be undesirable. By contrast, label aggregation avoids imposing a strict order between high-relevance & low-engagement and low-relevance & high-engagement documents.

Before closing, in Figure 1 we contrast label aggregation via product $\bar{\mathsf{Y}} = \mathsf{Y}^{(1)} \cdot \mathsf{Y}^{(2)}$ versus summation. The product aggregation mechanism is more conservative, in that it induces only a subset of the partial orders among label combinations induced by sum aggregation. Formally:

Lemma 6.3. *Suppose $\eta^{(k)}(x) \in \{0, 1\}, \forall x \in \mathcal{X}$. Let f^*_{sum} and f^*_{prod} be Bayes-optimal scorers for the label aggregated AUC in (5) with $\bar{\mathsf{Y}}_{\mathrm{sum}} = \sum_{k \in [K]} \mathsf{Y}^{(k)}$ and $\bar{\mathsf{Y}}_{\mathrm{prod}} = \prod_{k \in [K]} \mathsf{Y}^{(k)}$ respectively. Then for any $x, x' \in \mathcal{X}$:*

(a) $\gamma_{\mathrm{sum}}(x) > \gamma_{\mathrm{sum}}(x') \implies f^*_{\mathrm{sum}}(x) > f^*_{\mathrm{sum}}(x')$;

(b) $\gamma_{\mathrm{prod}}(x) > \gamma_{\mathrm{prod}}(x') \implies f^*_{\mathrm{prod}}(x) > f^*_{\mathrm{prod}}(x')$;

(c) $\gamma_{\mathrm{prod}}(x) > \gamma_{\mathrm{prod}}(x') \implies \gamma_{\mathrm{sum}}(x) > \gamma_{\mathrm{sum}}(x')$,

where $\gamma_{\mathrm{sum}}(x) \doteq \sum_{k \in [K]} \eta^{(k)}(x)$, *and* $\gamma_{\mathrm{prod}}(x) \doteq \mathbf{P}(\mathsf{Y}^{(1)} = 1, \ldots, \mathsf{Y}^{(K)} = 1 \mid \mathsf{X} = x)$.

7. Experimental Results

We present a suite of synthetic and real-world experiments to empirically validate our theoretical findings. Specifically, we aim to verify: (i) the "label dictatorship" phenomenon predicted for loss aggregation (Proposition 6.1 and §6.1), particularly its emergence due to the aforementioned sensitivity to $\pi^{(k)}$; and (ii) the consequent tendency for label aggregation (Proposition 6.2 and §6.2) to offer more balanced handling of multiple labels compared to loss aggregation.

To measure the degree of balanced handling of labels, we report the difference between the per-label AUCs: high differences indicate that one of the two labels is disproportionately favored. Since this metric can be trivially maximized by a scorer achieving 0.5 AUC for both labels, we additionally report the minimum over the per-label AUCs. Note that worst-class performance is a common metric in prior works on fairness (Williamson & Menon, 2019; Sagawa et al., 2020). While our focus is on label and loss aggregation, in future work it will be of interest to study how well directly maximizing the minimum per-label AUC would perform.

We present experiments on a synthetic and 3 real world datasets with training neural models on loss and label aggregation objectives. When optimizing the AUC-based objectives, following prior works we employ a logistic or hinge surrogate function. See Appendix G.2 for details.

Additionally, in Appendix G.1 we empirically verify the derived forms of the Bayes-optimal scorers for loss and label aggregation (Proposition 5.2, Proposition 5.4).

Bayes-optimal solutions on synthetic data We begin by analyzing the "label dictatorship" phenomenon directly on the Bayes-optimal scorers (i.e., *without* any model training) on synthetic data. We consider a 2D prediction task with two labels, and compare loss and label aggregation as we vary the label skewness $\pi^{(1)}, \pi^{(2)}$. This variation in skewness directly probes two contrasting theoretical predictions: recall

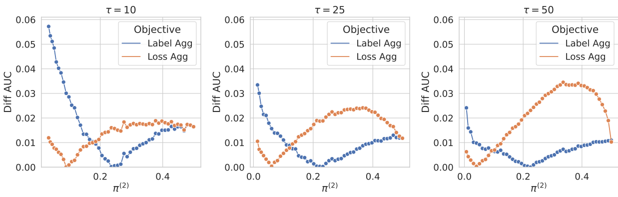


Figure 2. Plot of $|\mathrm{AUC}(\cdot; D^{(1)}) - \mathrm{AUC}(\cdot; D^{(2)})|$ for optimal scorers as a function of skewness in label $\mathsf{Y}^{(2)}$. We compare label aggregation and loss aggregation on the synthetic dataset as described in the main text. We fix $\pi^{(1)} = \mathbf{P}(\mathsf{Y}^{(1)} = 1) = 0.5$ and vary $\pi^{(2)} = \mathbf{P}(\mathsf{Y}^{(2)} = 1)$. Larger values of sigmoid scaling parameter τ make the label distribution closer to deterministic. A lower difference indicates a more balanced label treatment by the optimal solution; loss aggregation leads to a higher difference. Figure 6 reports individual per-label AUC metrics.

that the Bayes-optimal scorer for loss aggregation (Proposition 5.2) involves weighting of individual class-probabilities $\eta^{(k)}(x)$ inversely proportional to $\pi^{(k)} \cdot (1 - \pi^{(k)})$, thus leading to "label dictatorship" (Proposition 6.1). By contrast, the optimal scorer under label aggregation (with costs $c_{\bar{y}\bar{y}'} = |\bar{y} - \bar{y}'|$, Proposition 5.4) does not involve $\pi^{(k)}$.

We generate instances $x \in \mathbb{R}^2$ uniformly from $[-1, 1]^2$. We model the two class probability distributions using a logistic model, with $\eta^{(1)}(x) = \sigma(\tau \cdot w_1^\top x)$ and $\eta^{(2)}(x) = \sigma(\tau \cdot (w_2^\top x - \rho))$, for sigmoid function $\sigma(z) \doteq \frac{1}{1+\exp(-z)}$, $w_1 = \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]^\top$, $w_2 = [0, 1]^\top$, scale parameter $\tau > 0$, and shift parameter $\rho \geq 0$. As $\tau \to +\infty$, the label distributions become more deterministic. As $\rho \to +\infty$, the more skewed is the label distribution $\eta^{(2)}(x)$. Specifically, our choice of w_1 ensures that $\pi^{(1)} = \mathbb{E}[\eta^{(1)}(\mathsf{X})] = 0.5$; we vary the skewness of $\pi^{(2)} = \mathbb{E}[\eta^{(2)}(\mathsf{X})]$ alone by varying ρ.

For each instantiation of the above distributions $D^{(1)}$ and $D^{(2)}$, we compute the Bayes-optimal scorers for both loss aggregation (6) with $a_1 = a_2 = 1$, and label aggregation (7), and evaluate the difference $\Delta_{\mathrm{AUC}} = |\mathrm{AUC}(f; D^{(1)}) - \mathrm{AUC}(f; D^{(2)})|$ on 10^5 samples drawn from $(D^{(1)}, D^{(2)})$. In Figure 2, we plot the differences in the per-label AUC as a function of the second label's skewness $\pi^{(2)}$. Observe that for most skewness values, the optimal scorer for loss aggregation leads to a larger gap in AUCs between the two labels. This is due to the optimal scorer for loss aggregation (6) closely depending on the label priors $\pi^{(1)}$ and $\pi^{(2)}$. In fact, the closer the distributions are to being deterministic, the larger is Δ_{AUC} for loss aggregation. This observation closely aligns with the dictatorial behavior described in Section 6.1 for loss aggregation under deterministic labels.

Banking. We consider the UCI `Banking` dataset composed of information about Bank customers, advertising campaign details, and the success thereof (Moro et al., 2014). We take the *mortgage* and *loan* variables as two targets.

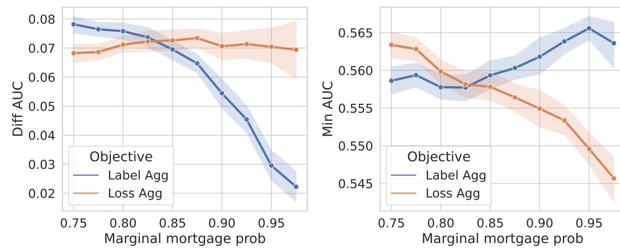


Figure 3. Plots of $|\mathrm{AUC(mortgage)} - \mathrm{AUC(loan)}|$ (*lower* is better) and $\min\{\mathrm{AUC(mortgage)}, \mathrm{AUC(loan)}\}$ (*higher* is better) on the `Banking` dataset. We compare label aggregation and loss aggregation as we vary the marginal probability of the mortgage label. We find label aggregation to fare better in the higher skewness regime.

To study the behavior of label and loss aggregation under label skews, we re-sample this dataset with varying proportions of the positive mortgage label, and average results over 25 trials. We train a linear model on numerical features using Adam for 100 epochs. In Table 3 and Figure 3, we compare the scorers learned by optimizing the loss aggregation objective (4) with $a_1 = a_2 = 1$ and the label aggregation objective (5) with costs $c_{\bar{y}\bar{y}'} = \mathbf{1}(\bar{y} > \bar{y}') \cdot |\bar{y} - \bar{y}'|$. We employ the logistic surrogate loss in both cases. Notice that label aggregation is able to better balance between the two labels, yielding the highest Min AUC metric.

Objective	AUC mortgage ↑	AUC loan ↑	Diff AUC ↓	Min AUC ↑
AUC(mortgage)	0.637 ± 0.004	0.523 ± 0.002	0.113 ± 0.004	0.523 ± 0.002
AUC(loan)	0.550 ± 0.005	0.573 ± 0.002	0.023 ± 0.005	0.550 ± 0.005
$\mathrm{AUC}_{\mathrm{LaA}}$	0.616 ± 0.003	0.562 ± 0.002	0.054 ± 0.005	0.562 ± 0.002
$\mathrm{AUC}^{(1,1)}_{\mathrm{LoA}}$	0.626 ± 0.003	0.555 ± 0.002	0.071 ± 0.005	0.555 ± 0.002

Table 3. Results on the `Banking` dataset with the proportion of positive mortgage label set to 0.9. While optimizing an objective for an individual label maximizes the corresponding AUC, we find label aggregation to strike a balance between the two evaluation metrics, yielding the *highest Min AUC*. Here, $\mathrm{AUC}_{\mathrm{LaA}}$ = AUC(mortgage + loan) and $\mathrm{AUC}^{(1,1)}_{\mathrm{LoA}}$ = AUC(mortgage) + AUC(loan).

HelpSteer. `HelpSteer` (Wang et al., 2023b) consists of evaluations of LLM responses across 5 categories: helpfulness, factuality, correctness, coherence, complexity and verbosity. On each category, responses are rated from 0 to 4, where higher means better. While originally intended for usage in LLM alignment, we repurpose the dataset here to illustrate the tradeoffs when ranking along competing signals.

We consider a task of ranking concatenations of a prompt and candidate response. We binarize each category based on whether or not it equals 4, and fix the first label to coherence (the most balanced category), while iteratively choosing the second label to be each of the remaining categories. Figure 4

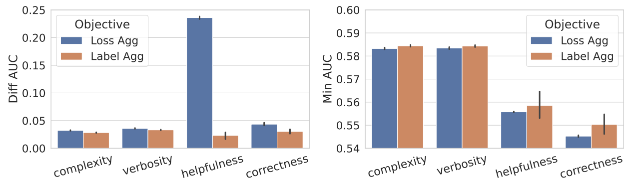


Figure 4. Plot of Diff AUC $|\text{AUC}(\cdot; D^{(1)}) - \text{AUC}(\cdot; D^{(2)})|$ (*lower* is better) and Min AUC $\min\{\text{AUC}(\cdot; D^{(1)}), \text{AUC}(\cdot; D^{(2)})\}$ (*higher* is better) for scorers learned with different labels $\mathsf{Y}^{(2)}$ (sorted according to their skewness, from the most skewed to the least skewed). Experiments on the `HelpSteer` dataset, where the first label is fixed to coherence (which is the closest to balance, and thus, $\pi^{(1)} = \mathbf{P}(\mathsf{Y}^{(1)} = 1) \approx 0.5$), and the second label is chosen over the set of all remaining labels (and so, $\pi^{(2)} = \mathbf{P}(\mathsf{Y}^{(2)} = 1)$ is of varying skewness). We find loss aggregation to lead to a higher difference between per-label AUC metrics.

Objective	AUC Click ↑	AUC Rel ↑	Diff AUC ↓	Min AUC ↑
AUC(Click)	0.74	0.61	0.13	0.61
AUC(Rel)	0.70	0.67	0.03	0.67
AUC_{LaA}	0.73	0.68	0.05	0.68
AUC_{LoA}	0.73	0.67	0.06	0.67

Table 4. Results on the `MSLR` dataset. Learning on both label (click and relevance) helps even when evaluated on each label alone. Label aggregation strikes a better balance between the two objectives, and achieves the largest minimum AUC over the two labels.

shows that label aggregation yields the lowest difference between per-label AUC across all settings.

`MSLR`. We next consider `MSLR Web30k`, a dataset of users' query-document interactions (Qin & Liu, 2013). We follow the methodology from Mahapatra et al. (2023) in constructing the target scoring function by removing Query-URL Click Count (Click), and using it in conjunction with the Relevance label. We train an MLP model over the concatenated features and follow hyper-parameters specified in the TensorFlow ranking library (Pasumarthi et al., 2019).

In Table 4, we report results across different objectives. Since the AUC over clicks is always seen to be higher than AUC over relevance, $\min(\text{AUC click}, \text{AUC relevance}) =$ AUC relevance across all objectives. We find that, compared to the label aggregation solution, loss aggregation solutions tend to over-optimize one metric over the other, depending on the weights. On the other hand, label aggregation yields the highest $\min(\text{AUC click}, \text{AUC relevance})$ and Pareto dominates the loss aggregation with equal weights.

8. Related Work

Bipartite ranking. Bipartite ranking is a fundamental supervised learning problem, with intimate ties to binary classification and class-probability estimation (Narasimhan & Agarwal, 2013). A large body of work has studied various theoretical aspects of the problem, including statistical generalisation (Agarwal et al., 2005; Agarwal, 2014), consistency of suitable surrogate loss minimisation (Clémençon et al., 2008; Uematsu & Lee, 2015; Menon & Williamson, 2016), relation to classic supervised learning problems (Cortes & Mohri, 2003; Kotłowski et al., 2011; Narasimhan & Agarwal, 2013), effective algorithm design (Freund et al., 2003), and extensions to *top-ranking* settings (Rudin, 2009; Agarwal et al., 2011; Li et al., 2014).

Multi-objective ranking. There has been much work on finding Pareto optimal solutions for multi-objective ranking (Mahapatra et al., 2023). A primary focus has been on modifying the LambdaMART objective (Burges, 2010) and evaluating on the nDCG evaluation metric, which the original LambdaMART objective optimizes the upper bound for (Wang et al., 2018). Two prominent lines of works focus on loss aggregation (or linear scalarization) (Ruchte & Grabocka, 2021) and label aggregation (Svore et al., 2011; Agarwal et al., 2011; Dai et al., 2011; Carmel et al., 2020; Wei et al., 2023). To the best of our knowledge, prior works neither considered multi-objective AUC optimization, nor theoretically analyzed the corresponding optimal solutions.

Multi-label ranking. Multi-label classification involves learning to predict *multiple* labels associated with a given instance (Tsoumakas et al., 2010; Read et al., 2009; Dembczyński et al., 2010). Multi-label *ranking* generalizes this to learn a *ranker* over the multiple candidate labels (Brinker et al., 2006; Fürnkranz et al., 2008; Dembczyński et al., 2012). Canonically, such a ranker may be assessed based on an *instance-specific* analogue of the AUC (Wu et al., 2021). While the problem setting is similar to what we consider, the goal is fundamentally different: multi-label ranking seeks to produce a score for *every* candidate label, while we seek to produce a *single* score that synthesizes the labels.

9. Conclusion and Future Work

We have studied the problem of bipartite ranking from multiple labels, with a characterization of the Bayes-optimal solutions for loss and label aggregation techniques. While these optimal scorers have a similar form, we established that loss aggregation can lead to an undesirable "dictatorship" issue. In future work, it would be of interest to study *surrogate* losses for the loss and label aggregation strategies. Another interesting direction would be to move beyond AUC to metrics such as the nDCG (Wang et al., 2013).

Impact Statement

This paper presents work goals of which is to advance the field of machine learning, specifically in the area of bipartite ranking with multiple labels. While there are many potential societal consequences of our work, we don't feel they must be specifically highlighted here. The presented analysis of loss and label aggregation methods provides a deeper understanding of multi-label bipartite ranking, potentially leading to improved algorithms in information retrieval and medical diagnosis applications.

References

Agarwal, D., Chen, B.-C., Elango, P., and Wang, X. Click shaping to optimize multiple objectives. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pp. 132–140, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450308137. doi: 10.1145/2020408.2020435. URL https://doi.org/10.1145/2020408.2020435.

Agarwal, S. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research*, 15(49):1653–1674, 2014. URL http://jmlr.org/papers/v15/agarwal14b.html.

Agarwal, S. and Niyogi, P. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10(16):441–474, 2009. URL http://jmlr.org/papers/v10/agarwal09a.html.

Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., and Roth, D. Generalization bounds for the area under the roc curve. *Journal of Machine Learning Research*, 6 (14):393–425, 2005. URL http://jmlr.org/papers/v6/agarwal05a.html.

Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S., and Wang, T. MS MARCO: A human generated machine reading comprehension dataset, 2018. URL https://arxiv.org/abs/1611.09268.

Brinker, K., Fürnkranz, J., and Hüllermeier, E. A unified model for multilabel classification and ranking. In *Proceedings of the 2006 Conference on ECAI 2006: 17th European Conference on Artificial Intelligence August 29 – September 1, 2006, Riva Del Garda, Italy*, pp. 489–493, NLD, 2006. IOS Press. ISBN 1586036424.

Burges, C. J. C. From RankNet to LambdaRank to LambdaMART: An overview. Technical report, Microsoft Research, 2010. URL http://research.microsoft.com/en-us/um/people/cburges/tech˙reports/MSR-TR-2010-82.pdf.

Carmel, D., Haramaty, E., Lazerson, A., and Lewin-Eytan, L. Multi-objective ranking optimization for product search using stochastic label aggregation. In *Proceedings of The Web Conference 2020*, WWW '20, pp. 373–383, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380122. URL https://doi.org/10.1145/3366423.3380122.

Clémençon, S., Lugosi, G., and Vayatis, N. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, 36(2):844–874, 2008. ISSN 00905364. URL http://www.jstor.org/stable/25464648.

Cortes, C. and Mohri, M. Auc optimization vs. error rate minimization. In Thrun, S., Saul, L., and Schölkopf, B. (eds.), *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL https://proceedings.neurips.cc/paper˙files/paper/2003/file/6ef80bb237adf4b6f77d0700e1255907-Paper.pdf.

Dai, N., Shokouhi, M., and Davison, B. D. Learning to rank for freshness and relevance. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pp. 95–104, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450307574. doi: 10.1145/2009916.2009933. URL https://doi.org/10.1145/2009916.2009933.

Dembczyński, K., Cheng, W., and Hüllermeier, E. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 279–286, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Dembczyński, K., Kotłowski, W., and Hüllermeier, E. Consistent multilabel ranking through univariate loss minimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 1347–1354, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Ehrgott, M. *Multicriteria Optimization*. Springer, 2005.

Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, 4(null):933–969, December 2003. ISSN 1532-4435.

Fürnkranz, J., Hüllermeier, E., Mencía, E. L., and Brinker, K. Multilabel classification via calibrated label ranking. *Machine learning*, 73(2):133–153, 2008.

He, Y., Tian, Y., Wang, M., Chen, F., Yu, L., Tang, M., Chen, C., Zhang, N., Kuang, B., and Prakash, A. Que2engage: Embedding-based retrieval for relevant and engaging products at facebook marketplace. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, pp. 386–390, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394192. doi: 10.1145/3543873.3584633. URL https://doi.org/10.1145/3543873.3584633.

Kotłowski, W., Dembczyński, K., and Hüllermeier, E. Bipartite ranking through minimization of univariate loss. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 1113–1120, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Krzanowski, W. J. and Hand, D. J. *ROC Curves for Continuous Data*. Chapman & Hall/CRC, 1st edition, 2009. ISBN 1439800219.

Lehmann, E. L. and Romano, J. P. *Testing Statistical Hypotheses*. Springer Science & Business Media, 2005.

Li, N., Jin, R., and Zhou, Z.-H. Top rank optimization in linear time. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, pp. 1502–1510, Cambridge, MA, USA, 2014. MIT Press.

Lin, X., Chen, H., Pei, C., Sun, F., Xiao, X., Sun, H., Zhang, Y., Ou, W., and Jiang, P. A pareto-efficient algorithm for multiple objective optimization in e-commerce recommendation. In *Proceedings of the 13th ACM Conference on recommender systems*, pp. 20–28, 2019.

Ling, C. X. and Li, C. Data mining for direct marketing: problems and solutions. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, KDD'98, pp. 73–79. AAAI Press, 1998.

Liu, T.-Y. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331, March 2009. ISSN 1554-0669. doi: 10.1561/1500000016. URL https://doi.org/10.1561/1500000016.

Macskassy, S. A. and Provost, F. Intelligent information triage. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pp. 318–326, New York, NY, USA, 2001. Association for Computing Machinery. ISBN 1581133316. doi: 10.1145/383952.384015. URL https://doi.org/10.1145/383952.384015.

Mahapatra, D., Dong, C., Chen, Y., and Momma, M. Multi-label learning to rank through multi-objective optimization. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4605–4616, 2023.

Menon, A., Narasimhan, H., Agarwal, S., and Chawla, S. On the statistical consistency of algorithms for binary classification under class imbalance. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 603–611, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/menon13a.html.

Menon, A. K. and Williamson, R. C. Bipartite ranking: a risk-theoretic perspective. *Journal of Machine Learning Research*, 17(195):1–102, 2016. URL http://jmlr.org/papers/v17/14-265.html.

Moro, S., Cortez, P., and Rita, P. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014. ISSN 0167-9236. doi: https://doi.org/10.1016/j.dss.2014.03.001. URL https://www.sciencedirect.com/science/article/pii/S016792361400061X.

Narasimhan, H. and Agarwal, S. On the relationship between binary classification, bipartite ranking, and binary class probability estimation. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 2913–2921, Red Hook, NY, USA, 2013. Curran Associates Inc.

Ng, K. B. and Kantor, P. B. Predicting the effectiveness of naïve data fusion on the basis of system characteristics. *Journal of the American Society for Information Science*, 51(13):1177–1189, 2000.

Pasumarthi, R. K., Bruch, S., Wang, X., Li, C., Bendersky, M., Najork, M., Pfeifer, J., Golbandi, N., Anil, R., and Wolf, S. Tf-ranking: Scalable tensorflow library for learning-to-rank. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2970–2978, 2019.

Pepe, M. S. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, 03 2003. ISBN 9780198509844. doi: 10.1093/oso/9780198509844.001.0001. URL https://doi.org/10.1093/oso/9780198509844.001.0001.

Perkio, J., Tuulos, V., Buntine, W., and Tirri, H. Multi-faceted information retrieval system for large scale email archives. In *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, pp. 557–564. IEEE, 2005.

Qin, T. and Liu, T. Introducing LETOR 4.0 datasets. *CoRR*, abs/1306.2597, 2013. URL http://arxiv.org/abs/1306.2597.

Read, J., Pfahringer, B., Holmes, G., and Frank, E. Classifier chains for multi-label classification. In Buntine, W., Grobelnik, M., Mladenić, D., and Shawe-Taylor, J. (eds.), *Machine Learning and Knowledge Discovery in Databases*, pp. 254–269, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-04174-7.

Reddi, S., Pasumarthi, R. K., Menon, A., Rawat, A. S., Yu, F., Kim, S., Veit, A., and Kumar, S. Rankdistil: Knowledge distillation for ranking. In *International Conference on Artificial Intelligence and Statistics*, pp. 2368–2376. PMLR, 2021.

Ribeiro, M. T., Ziviani, N., Moura, E. S. D., Hata, I., Lacerda, A., and Veloso, A. Multiobjective pareto-efficient approaches for recommender systems. *ACM Trans. Intell. Syst. Technol.*, 5(4), December 2015. ISSN 2157-6904. doi: 10.1145/2629350. URL https://doi.org/10.1145/2629350.

Ruchte, M. and Grabocka, J. Scalable pareto front approximation for deep multi-objective learning, 2021. URL https://arxiv.org/abs/2103.13392.

Rudin, C. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list. *Journal of Machine Learning Research*, 10(78):2233–2271, 2009. URL http://jmlr.org/papers/v10/rudin09b.html.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Sun, X., Zhang, B., Zhang, C., Ren, H., and Cai, M. Enhancing personalized ranking with differentiable group auc optimization. *arXiv preprint arXiv:2304.09176*, 2023.

Svore, K. M., Volkovs, M. N., and Burges, C. J. Learning to rank with multiple objective functions. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pp. 367–376, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306324. doi: 10.1145/1963405.1963459. URL https://doi.org/10.1145/1963405.1963459.

Swets, J. A. Measuring the accuracy of diagnostic systems. *Science*, 240(4857):1285–1293, 1988. doi: 10.1126/science.3287615. URL https://www.science.org/doi/abs/10.1126/science.3287615.

Tang, S., Luo, F., and Wu, J. Smooth-auc: Smoothing the path towards rank-based ctr prediction. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2400–2404, 2022.

Team, G. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

Tsoumakas, G., Katakis, I., and Vlahavas, I. *Mining Multi-label Data*, pp. 667–685. Springer US, Boston, MA, 2010. ISBN 978-0-387-09823-4. doi: 10.1007/978-0-387-09823-4_34. URL https://doi.org/10.1007/978-0-387-09823-4˙34.

Uematsu, K. and Lee, Y. Statistical optimality in multipartite ranking and ordinal regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(5):1080–1094, 2015. doi: 10.1109/TPAMI.2014.2360397.

Verma, R., Barrejon, D., and Nalisnick, E. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 11415–11434. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/verma23a.html.

Wang, X., Li, C., Golbandi, N., Bendersky, M., and Najork, M. The lambdaloss framework for ranking metric optimization. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pp. 1313–1322, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360142. doi: 10.1145/3269206.3271784. URL https://doi.org/10.1145/3269206.3271784.

Wang, X., Jin, Q., Huang, S., Zhang, M., Liu, X., Zhao, Z., Chen, Y., Zhang, Z., Yang, J., Wen, E., et al. Towards the better ranking consistency: A multi-task learning framework for early stage ads ranking. In *AdKDD*, 2023a.

Wang, Y., Wang, L., Li, Y., He, D., and Liu, T.-Y. A theoretical analysis of ndcg type ranking measures. In Shalev-Shwartz, S. and Steinwart, I. (eds.), *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pp. 25–54, Princeton, NJ, USA, 12–14 Jun 2013. PMLR. URL https://proceedings.mlr.press/v30/Wang13.html.

Wang, Z., Dong, Y., Zeng, J., Adams, V., Sreedhar, M. N., Egert, D., Delalleau, O., Scowcroft, J. P., Kant, N., Swope, A., and Kuchaiev, O. Helpsteer: Multi-attribute helpfulness dataset for steerlm, 2023b.

Wei, J., Zhu, Z., Luo, T., Amid, E., Kumar, A., and Liu, Y. To aggregate or not? learning with separate noisy labels. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2523–2535, 2023.

Williamson, R. C. and Menon, A. K. Fairness risk measures, 2019. URL https://arxiv.org/abs/1901.08665.

Wu, G., LI, C., Xu, K., and Zhu, J. Rethinking and reweighting the univariate losses for multi-label ranking: Consistency and generalization. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 14332–14344. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper˙files/paper/2021/file/781397bc0630d47ab531ea850bddcf63-Paper.pdf.

Zheng, Y. and Wang, D. X. A survey of recommender systems with multi-objective optimization. *Neurocomputing*, 474:141–153, 2022.

A. Additional helper results

Lemma A.1. *Let $\mathcal{S}$ denote the set of pairwise scorers $s\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ satisfying the* symmetry *condition: $s(x,x') = s(x',x)$. For any distribution μ and weight function $w\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$, let*

$$s^* \in \operatorname*{argmax}_{s\in\mathcal{S}} \mathbb{E}_{\mathsf{X}\sim\mu}\mathbb{E}_{\mathsf{X}'\sim\mu}\left[w(\mathsf{X},\mathsf{X}')\cdot H(s(\mathsf{X},\mathsf{X}'))\right]. \tag{9}$$

Then,

$$s^*(x,x') > 0 \iff w(x,x') - w(x',x) > 0.$$

Proof of Lemma A.1. By symmetry of the expectation, and the definition of $H(\cdot)$, the optimal solution is equivalently

$$\begin{aligned} s^* &\in \operatorname*{argmax}_{s\in\mathcal{S}} \mathbb{E}_{\mathsf{X}\sim\mu}\mathbb{E}_{\mathsf{X}'\sim\mu}\left[w(\mathsf{X},\mathsf{X}')\cdot H(s(\mathsf{X},\mathsf{X}')) + w(\mathsf{X}',\mathsf{X})\cdot H(-s(\mathsf{X},\mathsf{X}'))\right] \\ &= \operatorname*{argmax}_{s\in\mathcal{S}} \mathbb{E}_{\mathsf{X}\sim\mu}\mathbb{E}_{\mathsf{X}'\sim\mu}\left[\begin{cases} w(\mathsf{X},\mathsf{X}') & \text{if } s(\mathsf{X},\mathsf{X}')>0 \\ w(\mathsf{X}',\mathsf{X}) & \text{if } s(\mathsf{X},\mathsf{X}')<0 \\ \frac{w(\mathsf{X},\mathsf{X}')+w(\mathsf{X}',\mathsf{X})}{2} & \text{if } s(\mathsf{X},\mathsf{X}')=0 \end{cases}\right]. \end{aligned}$$

For each fixed $(x,x')\in\mathcal{X}\times\mathcal{X}$, we thus have

$$s^*(x,x') > 0 \iff w(x,x') - w(x',x) > 0.$$

Note that if $w(x,x') = w(x',x)$, then the choice of $s^*(x,x')$ may be arbitrary. □

Lemma A.2. *For any distribution μ and weight function $w\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$, let*

$$f^* \in \operatorname*{argmax}_{f\colon \mathcal{X}\to\mathbb{R}} \mathbb{E}_{\mathsf{X}\sim\mu}\mathbb{E}_{\mathsf{X}'\sim\mu}\left[w(\mathsf{X},\mathsf{X}')\cdot H(f(\mathsf{X})-f(\mathsf{X}'))\right]. \tag{10}$$

Then, if there exists some $g\colon\mathcal{X}\to\mathbb{R}$ such that

$$g(x) > g(x') \iff w(x,x') > w(x',x),$$

we must have

$$f^*(x) > f^*(x') \iff g(x) > g(x'). \tag{11}$$

Proof of Lemma A.2. Let

$$\begin{aligned} R_{\text{pair}}(s) &\doteq \mathbb{E}_{\mathsf{X}\sim\mu}\mathbb{E}_{\mathsf{X}'\sim\mu}\left[w(\mathsf{X},\mathsf{X}')\cdot H(s(\mathsf{X},\mathsf{X}'))\right] \\ R_{\text{diff}}(f) &\doteq \mathbb{E}_{\mathsf{X}\sim\mu}\mathbb{E}_{\mathsf{X}'\sim\mu}\left[w(\mathsf{X},\mathsf{X}')\cdot H(f(\mathsf{X})-f(\mathsf{X}'))\right]. \end{aligned}$$

If there exists a g satisfying the prescribed condition, then by Lemma A.1, any optimal *pairwise* scorer s^* for R_{pair} satisfies $s^*(x,x') > 0 \iff g(x) > g(x')$.

Now let $g_{\text{diff}}(x,x') \doteq g(x) - g(x')$. Then, $g_{\text{diff}}(x,x') > 0 \iff s^*(x,x') > 0 \iff w(x,x') > w(x',x)$, and so g_{diff} is an optimal pairwise scorer for R_{pair}. Thus, for any alternate scorer $\bar{g}\colon\mathcal{X}\to\mathbb{R}$, with $\bar{g}_{\text{diff}}(x,x') \doteq \bar{g}(x) - \bar{g}(x')$, we must have $R_{\text{pair}}(g_{\text{diff}}) \le R_{\text{pair}}(\bar{g}_{\text{diff}})$. This however implies that $R_{\text{diff}}(g) \le R_{\text{diff}}(\bar{g})$, i.e., g is an optimal scorer for the present objective R_{diff} (10).

Now suppose f^* is any candidate optimal solution for R_{diff}. Then, f^* must satisfy the given condition (11), or else (again appealing to Lemma A.1) f^*_{diff} will not be optimal for R_{pair}. Thus, the result follows. □

Lemma A.3 (Agarwal et al. (2005))**.** *For a scorer $f\colon\mathcal{X}\to\mathbb{R}$ and distribution D, the AUC is equally*

$$\begin{aligned} \text{AUC}(f;D) &\doteq \mathbb{E}_{\mathsf{X}\sim q_+}\mathbb{E}_{\mathsf{X}'\sim q_-}\left[H(f(\mathsf{X})-f(\mathsf{X}'))\right] \\ H(z) &\doteq 1(z>0) + \frac{1}{2}\cdot 1(z=0). \end{aligned} \tag{12}$$

where $q_+(x) = \mathbf{P}(x|y=1)$ and $q_-(x) = \mathbf{P}(x|y=0)$ denote the class-conditional distributions.

Proof of Lemma A.3. See proof of Lemma A.5, which is a strict generalisation. ☐

Lemma A.4 (Agarwal (2014)). *For any scorer $f\colon \mathcal{X} \to \mathbb{R}$ and distribution D, the AUC-ROC is equally*

$$\mathrm{AUC}(f; D) = \mathbb{E}_{\mathsf{X}\sim\mu}\mathbb{E}_{\mathsf{X}'\sim\mu}\left[w(\mathsf{X}, \mathsf{X}') \cdot H(f(\mathsf{X}) - f(\mathsf{X}'))\right] \tag{13}$$
$$w(x, x') \doteq \frac{1}{\pi(1-\pi)} \cdot \eta(x) \cdot (1 - \eta(x')).$$

Proof of Lemma A.4. By an application of Bayes' rule to Lemma A.3,

$$\begin{aligned}\mathrm{AUC}(f; D) &= \mathbb{E}_{\mathsf{X}\sim q_+}\mathbb{E}_{\mathsf{X}'\sim q_-}\left[H(f(\mathsf{X}) - f(\mathsf{X}'))\right] \\ &= \frac{1}{\pi \cdot (1-\pi)} \cdot \mathbb{E}_{\mathsf{X}\sim\mu}\mathbb{E}_{\mathsf{X}'\sim\mu}\left[\eta(\mathsf{X}) \cdot (1 - \eta(\mathsf{X}')) \cdot H(f(\mathsf{X}) - f(\mathsf{X}'))\right].\end{aligned}$$

☐

Lemma A.5 (Uematsu & Lee (2015)). *For any scorer $f\colon \mathcal{X} \to \mathbb{R}$, distribution D_{mp}, and costs $\{c_{yy'}\}$,*

$$\mathrm{AUC}(f; D_{\mathrm{mp}}) = \mathbb{E}_{\mathsf{X}\sim\mu}\mathbb{E}_{\mathsf{X}'\sim\mu}\left[w(\mathsf{X}, \mathsf{X}') \cdot H(f(\mathsf{X}) - f(\mathsf{X}'))\right]$$
$$w(x, x') \doteq \frac{1}{v} \cdot \sum_{y\in\mathcal{Y}}\sum_{y'\in\mathcal{Y}} 1(y > y') \cdot c_{yy'} \cdot \eta_y(x) \cdot \eta_{y'}(x')$$
$$v \doteq \sum_{y\in\mathcal{Y}}\sum_{y'\in\mathcal{Y}} 1(y > y') \cdot c_{yy'} \cdot \pi_y \cdot \pi_{y'}.$$

Proof of Lemma A.5. We provide a proof for completeness. By definition of conditional expectation,[1]

$$\begin{aligned}&\mathrm{AUC}(f; D_{\mathrm{mp}}) \\ &= \mathbb{E}_{\mathsf{X}}\mathbb{E}_{\mathsf{X}'}\left[c_{\mathsf{YY}'} \cdot H(f(\mathsf{X}) - f(\mathsf{X}')) \mid \mathsf{Y} > \mathsf{Y}'\right] \\ &= \int_{\mathcal{X}\times\mathcal{X}} \sum_{y,y'} \left[D(x, x', y, y' \mid \mathsf{Y} > \mathsf{Y}') \cdot c_{yy'} \cdot H(f(x) - f(x'))\right] \mathrm{d}x\,\mathrm{d}x' \\ &= \int_{\mathcal{X}\times\mathcal{X}} \sum_{y,y'} \left[\frac{1(y > y') \cdot D(x, x', y, y')}{D(\mathsf{Y} > \mathsf{Y}')} \cdot c_{yy'} \cdot H(f(x) - f(x'))\right] \mathrm{d}x\,\mathrm{d}x' \\ &= \frac{1}{D(\mathsf{Y} > \mathsf{Y}')} \cdot \int_{\mathcal{X}\times\mathcal{X}} \sum_{y,y'} \left[1(y > y') \cdot D(x, x') \cdot D(y, y' \mid x, x') \cdot c_{yy'} \cdot H(f(x) - f(x'))\right] \mathrm{d}x\,\mathrm{d}x' \\ &= \frac{1}{D(\mathsf{Y} > \mathsf{Y}')} \cdot \mathbb{E}_{\mathsf{X}\sim\mu}\mathbb{E}_{\mathsf{X}'\sim\mu}\left[\left[\sum_{y,y'} 1(y > y') \cdot \eta_y(x) \cdot \eta_{y'}(x') \cdot c_{yy'}\right] \cdot H(f(x) - f(x'))\right].\end{aligned}$$

Now observe that

$$\begin{aligned}D(\mathsf{Y} > \mathsf{Y}') &= \sum_{y\in[L]}\sum_{y'\in[L]} 1(y > y') \cdot \pi_y \cdot \pi_{y'} \\ &= \frac{1}{2} \cdot \left[1 - \sum_{y\in\mathcal{Y}} \pi_y^2\right].\end{aligned}$$

☐

Lemma A.6. *For any scorer $f\colon \mathcal{X} \to \mathbb{R}$ and distributions $\{D^{(k)}\}_{k\in[K]}$, the loss aggregated AUC is*

$$\mathrm{AUC}_{\mathrm{LoA}}(f) = \mathbb{E}_{\mathsf{X}\sim\mu}\mathbb{E}_{\mathsf{X}'\sim\mu}\left[w(\mathsf{X}, \mathsf{X}') \cdot H(f(\mathsf{X}) - f(\mathsf{X}'))\right]$$
$$w(x, x') \doteq \frac{1}{K} \sum_{k\in[K]} \frac{\eta^{(k)}(x) \cdot (1 - \eta^{(k)}(x'))}{\pi^{(k)} \cdot (1 - \pi^{(k)})}.$$

[1] We elide again details on the choice of base measure used to define the density on $\mathcal{X}$.

Proof of Lemma A.6. Building on Definition 2.1, we have

$$\begin{aligned}
&\mathrm{AUC}(f; \{D_{\mathrm{bp}}^{(k)}\}_{k\in[K]}) \\
&= \frac{1}{K} \sum_{k\in[K]} \mathbb{E}_{\mathsf{X}} \mathbb{E}_{\mathsf{X}'} \left[H(f(\mathsf{X}) - f(\mathsf{X}')) \mid \mathsf{Y}^{(k)} > \mathsf{Y}'^{(k)} \right] \\
&= \frac{1}{K} \sum_{k\in[K]} \frac{1}{\pi^{(k)} \cdot (1 - \pi^{(k)})} \cdot \mathbb{E}_{\mathsf{X}\sim\mu} \mathbb{E}_{\mathsf{X}'\sim\mu} \left[\eta^{(k)}(\mathsf{X}) \cdot (1 - \eta^{(k)}(\mathsf{X}')) \cdot H(f(\mathsf{X}) - f(\mathsf{X}')) \right].
\end{aligned}$$

The result thus follows. □

B. Proofs of results in body

Proof of Lemma 2.2. First, observe that by symmetry

$$\begin{aligned}
\mathrm{AUC}(f; D) &= \frac{1}{\pi \cdot (1-\pi)} \cdot \mathbb{E}_{\mathsf{X}\sim\mu} \mathbb{E}_{\mathsf{X}'\sim\mu} \left[\eta(\mathsf{X}) \cdot (1 - \eta(\mathsf{X}')) \cdot H(f(\mathsf{X}) - f(\mathsf{X}')) \right] \\
&= \frac{1}{\pi \cdot (1-\pi)} \cdot \mathbb{E}_{\mathsf{X}\sim\mu} \mathbb{E}_{\mathsf{X}'\sim\mu} \left[\frac{1}{2} \cdot [\eta(\mathsf{X}) \cdot (1 - \eta(\mathsf{X}')) \cdot H(f(\mathsf{X}) - f(\mathsf{X}')) + \right. \\
&\quad \left. \eta(\mathsf{X}') \cdot (1 - \eta(\mathsf{X})) \cdot H(f(\mathsf{X}') - f(\mathsf{X}))] \right] \\
&= \mathbb{E}_{\mathsf{X}\sim\mu} \mathbb{E}_{\mathsf{X}'\sim\mu} \left[\frac{1}{2} \cdot [w(\mathsf{X}, \mathsf{X}') \cdot H(f(\mathsf{X}) - f(\mathsf{X}')) + w(\mathsf{X}', \mathsf{X})) \cdot H(f(\mathsf{X}') - f(\mathsf{X}))] \right].
\end{aligned}$$

Now observe that

$$w(x, x') > w(x', x) \iff \eta(x) > \eta(x').$$

Thus, by Lemma A.2, the optimal scorer satisfies $f^*(x) - f^*(x') > 0 \iff \eta(x) - \eta(x') > 0$. □

Proof of Lemma 2.4. We provide a proof for $L = 3$ for completeness. For the general case, see Uematsu & Lee (2015).

Observe that

$$\begin{aligned}
&w(x, x') - w(x', x) \\
&= \frac{1}{v} \cdot \left[\sum_{y\in\mathcal{Y}} \sum_{y'\in\mathcal{Y}} 1(y > y') \cdot c_{yy'} \cdot \eta_y(x) \cdot \eta_{y'}(x') - \sum_{y\in\mathcal{Y}} \sum_{y'\in\mathcal{Y}} 1(y > y') \cdot c_{yy'} \cdot \eta_y(x') \cdot \eta_{y'}(x) \right] \\
&= \frac{1}{v} \cdot \left[\sum_{y\in\mathcal{Y}} \sum_{y'\in\mathcal{Y}} 1(y > y') \cdot c_{yy'} \cdot \eta_y(x) \cdot \eta_{y'}(x') - \sum_{y\in\mathcal{Y}} \sum_{y'\in\mathcal{Y}} 1(y < y') \cdot c_{y'y} \cdot \eta_y(x) \cdot \eta_{y'}(x') \right] \\
&= \frac{1}{v} \cdot \left[\sum_{y\in\mathcal{Y}} \sum_{y'\in\mathcal{Y}} \alpha_{yy'} \cdot \eta_y(x) \cdot \eta_{y'}(x') \right] \\
&= \frac{1}{v} \cdot \left[\sum_{y\in\mathcal{Y}} \beta_y(x') \cdot \eta_y(x) \right],
\end{aligned}$$

where $\alpha_{yy'} \doteq 1(y > y') \cdot c_{yy'} - 1(y < y') \cdot c_{y'y}$, and $\beta_y(x') \doteq \sum_{y'\in\mathcal{Y}} \alpha_{yy'} \cdot \eta_{y'}(x')$. When $L = 3$,

$$\begin{aligned}
\beta_0(x') &= -c_{10} \cdot \eta_1(x') - c_{20} \cdot \eta_2(x') \\
&= -c_{10} \cdot \eta_1(x') - c_{20} + c_{20} \cdot \eta_0(x') + c_{20} \cdot \eta_1(x') \\
&= c_{20} \cdot \eta_0(x') + (c_{20} - c_{10}) \cdot \eta_1(x') - c_{20} \\
\beta_1(x') &= c_{10} \cdot \eta_0(x') - c_{21} \cdot \eta_2(x') \\
&= c_{10} \cdot \eta_0(x') - c_{21} + c_{21} \cdot \eta_0(x') + c_{21} \cdot \eta_1(x')
\end{aligned}$$

$$= (c_{10} + c_{21}) \cdot \eta_0(x') + c_{21} \cdot \eta_1(x') - c_{21}$$
$$\beta_2(x') = c_{20} \cdot \eta_0(x') + c_{21} \cdot \eta_1(x').$$

We may rewrite the first two expressions in terms of β_2:

$$\beta_0(x') = \beta_2(x') + (c_{20} - c_{10} - c_{21}) \cdot \eta_1(x') - c_{20}$$
$$\beta_1(x') = \beta_2(x') + (c_{10} - c_{21} - c_{20}) \cdot \eta_0(x') - c_{21}.$$

The claim is that

$$\begin{aligned} f^*(x) &= \frac{c_{10} \cdot \eta_1(x) + c_{20} \cdot \eta_2(x)}{c_{20} \cdot \eta_0(x) + c_{21} \cdot \eta_1(x)} \\ &= \frac{-c_{20} \cdot \eta_0(x) + (c_{10} - c_{20}) \cdot \eta_1(x) + c_{20}}{c_{20} \cdot \eta_0(x) + c_{21} \cdot \eta_1(x)} \\ &= -\frac{\beta_0(x)}{\beta_2(x)}. \end{aligned}$$

Observe that

$$\begin{aligned} \frac{\beta_0(x')}{\beta_2(x')} - \frac{\beta_0(x)}{\beta_2(x)} &= (c_{20} - c_{10} - c_{21}) \cdot \left[\frac{\eta_1(x')}{\beta_2(x')} - \frac{\eta_1(x)}{\beta_2(x)}\right] - c_{20} \cdot \left[\frac{1}{\beta_2(x')} - \frac{1}{\beta_2(x)}\right] \\ &\propto (c_{20} - c_{10} - c_{21}) \cdot [\beta_2(x) \cdot \eta_1(x') - \beta_2(x') \cdot \eta_1(x)] - c_{20} \cdot [\beta_2(x) - \beta_2(x')], \end{aligned}$$

where the proportionality factor is $\frac{1}{\beta_2(x) \cdot \beta_2(x')} > 0$. Observe that the first term simplifies to:

$$\begin{aligned} &\beta_2(x) \cdot \eta_1(x') - \beta_2(x') \cdot \eta_1(x) \\ &= [c_{20} \cdot \eta_0(x) + c_{21} \cdot \eta_1(x)] \cdot \eta_1(x') - [c_{20} \cdot \eta_0(x') + c_{21} \cdot \eta_1(x')] \cdot \eta_1(x) \\ &= c_{20} \cdot [\eta_0(x) \cdot \eta_1(x') - \eta_0(x') \cdot \eta_1(x)]. \end{aligned}$$

Thus,

$$\frac{\beta_0(x')}{\beta_2(x')} - \frac{\beta_0(x)}{\beta_2(x)} \propto (c_{20} - c_{10} - c_{21}) \cdot [\eta_0(x) \cdot \eta_1(x') - \eta_0(x') \cdot \eta_1(x)] - [\beta_2(x) - \beta_2(x')].$$

Now observe that

$$\begin{aligned} &w(x, x') - w(x', x) \\ &= \beta_0(x') \cdot \eta_0(x) + \beta_1(x') \cdot \eta_1(x) + \beta_2(x') \cdot \eta_2(x) \\ &= [\beta_2(x') + (c_{20} - c_{10} - c_{21}) \cdot \eta_1(x') - c_{20}] \cdot \eta_0(x) + \\ &\quad [\beta_2(x') + (c_{10} - c_{21} - c_{20}) \cdot \eta_0(x') - c_{21}] \cdot \eta_1(x) + \beta_2(x') \cdot \eta_2(x) \\ &= \beta_2(x') \cdot \eta_0(x) + (c_{20} - c_{10} - c_{21}) \cdot \eta_1(x') \cdot \eta_0(x) - c_{20} \cdot \eta_0(x) + \\ &\quad \beta_2(x') \cdot \eta_1(x) + (c_{10} - c_{21} - c_{20}) \cdot \eta_0(x') \cdot \eta_1(x) - c_{21} \cdot \eta_1(x) + \beta_2(x') \cdot \eta_2(x) \\ &= (c_{20} - c_{10} - c_{21}) \cdot [\eta_1(x') \cdot \eta_0(x) - \eta_0(x') \cdot \eta_1(x)] + [\beta_2(x') - \beta_2(x)]. \end{aligned}$$

Thus, $w(x, x') - w(x', x) > 0 \iff \frac{\beta_0(x')}{\beta_2(x')} > \frac{\beta_0(x)}{\beta_2(x)}$. ◻

Proof of Proposition 5.2. Note that

$$w(x, x') = \frac{1}{K} \sum_{k \in [K]} \frac{1}{\pi^{(k)} \cdot (1 - \pi^{(k)})} \cdot \eta^{(k)}(x) - \frac{1}{K} \sum_{k \in [K]} \frac{1}{\pi^{(k)} \cdot (1 - \pi^{(k)})} \cdot \eta^{(k)}(x) \cdot \eta^{(k)}(x'),$$

and so

$$w(x, x') > w(x', x) \iff \frac{1}{K} \sum_{k \in [K]} \frac{1}{\pi^{(k)} \cdot (1 - \pi^{(k)})} \cdot \eta^{(k)}(x) > \frac{1}{K} \sum_{k \in [K]} \frac{1}{\pi^{(k)} \cdot (1 - \pi^{(k)})} \cdot \eta^{(k)}(x').$$

Thus, it follows that the Bayes-optimal scorer takes the form of the average class probabilities. ◻

Proof of Proposition 5.3. Suppose $K = 2$, and let $\bar{\mathsf{Y}} = \sum_{k \in [K]} \mathsf{Y}^{(k)} \in \{0, 1, \ldots, K\}$. Note that $\bar{\mathsf{Y}} = 0 \iff \mathsf{Y}^{(1)} = 0 \wedge \mathsf{Y}^{(2)} = 0$, and similarly $\bar{\mathsf{Y}} = 2 \iff \mathsf{Y}^{(1)} = 1 \wedge \mathsf{Y}^{(2)} = 1$. Further suppose that $\mathsf{Y}^{(1)} \perp\!\!\!\perp \mathsf{Y}^{(2)} \mid \mathsf{X}$. We now have

$$
\begin{aligned}
\bar{\eta}_0(x) &= (1 - \eta^{(1)}(x)) \cdot (1 - \eta^{(2)}(x)) \\
&= 1 - \eta^{(1)}(x) - \eta^{(2)}(x)) + \eta^{(1)}(x) \cdot \eta^{(2)}(x)) \\
\bar{\eta}_1(x) &= \eta^{(1)}(x) \cdot (1 - \eta^{(2)}(x)) + (1 - \eta^{(1)}(x)) \cdot \eta^{(2)}(x) \\
&= \eta^{(1)}(x) + \eta^{(2)}(x) - 2 \cdot \eta^{(1)}(x) \cdot \eta^{(2)}(x) \\
\bar{\eta}_2(x) &= \eta^{(1)}(x) \cdot \eta^{(2)}(x).
\end{aligned}
$$

By Uematsu & Lee (2015, Theorem 3), the Bayes-optimal scorer of AUC with uniform costs against $\bar{\mathsf{Y}}$ will preserve the ordering of

$$
\begin{aligned}
f^*(x) &= \frac{\bar{\eta}_1(x) + \bar{\eta}_2(x)}{\bar{\eta}_0(x) + \bar{\eta}_1(x)} \\
&= \frac{\eta^{(1)}(x) + \eta^{(2)}(x) - \eta^{(1)}(x) \cdot \eta^{(2)}(x)}{1 - \eta^{(1)}(x) \cdot \eta^{(2)}(x)}.
\end{aligned}
$$

Consider a dataset composed of 6 examples with independently distributed binary scoring functions, with the following distribution $D^{(1)}$ for the first signal: $\eta_1(x_1) = 1$, $\eta_1(x_2) = 0.2$, $\eta_1(x_3) = 0.62$, $\eta_1(x_4) = 0.44$, $\eta_1(x_5) = 0.56$, $\eta_1(x_6) = 0.81$, and the following distribution $D^{(2)}$ for the second signal: $\eta_2(x_1) = 0.44$ $\eta_2(x_2) = 0.56$, $\eta_2(x_3) = 0.81$, $\eta_2(x_4) = 1$, $\eta_2(x_5) = 0.2$, $\eta_2(x_6) = 0.62$.

The optimal solution for the label aggregation objective is $f^{\mathrm{LA}}_{\mathrm{opt}}(x_1) = 1.78571$, $f^{\mathrm{LA}}_{\mathrm{opt}}(x_2) = 0.72973$, $f_{\mathrm{opt}}(x_3) = 1.86380$, $f^{\mathrm{LA}}_{\mathrm{opt}}(x_4) = 1.78571$, $f^{\mathrm{LA}}_{\mathrm{opt}}(x_5) = 0.72973$, $f^{\mathrm{LA}}_{\mathrm{opt}}(x_6) = 1.86380$, and yields $\mathrm{AUC}(f^{\mathrm{LA}}_{\mathrm{opt}}; D^{(1)}) = 0.65559$ and $\mathrm{AUC}(f^{\mathrm{LA}}_{\mathrm{opt}}; D^{(2)}) = 0.65559$. That solution is however Pareto dominated by any scoring function following the ordering g of the examples: (4, 0, 2, 5, 1, 3), which yields $\mathrm{AUC}(g; D^{(1)}) = 0.65706$ and $\mathrm{AUC}(g; D^{(2)}) = 0.65862$. □

Proof of Proposition 5.4. Let $\bar{\mathsf{Y}} = \sum_k \mathsf{Y}^{(k)} \in \{0, 1, \ldots, K\}$ have class probability function $\bar{\eta} \colon \mathcal{X} \to \Delta(\{0, 1, \ldots, K\})$. By Uematsu & Lee (2015, Corollary 1), with costs $c_{\bar{y}\bar{y}'} = 1(\bar{y} > \bar{y}') \cdot |\bar{y} - \bar{y}'|$, the Bayes-optimal scorer will preserve the ordering of

$$
f^*(x) = \mathbb{E}\left[\bar{\mathsf{Y}} \mid \mathsf{X} = x\right] \ = \sum_{n=0}^{K} n \cdot \bar{\eta}_n(x) \ = \sum_{n=1}^{K} n \cdot \bar{\eta}_n(x).
$$

We will show that the above evaluates to $\sum_{k=1}^{K} \eta^{(k)}(x)$. We start with the RHS:

$$
\begin{aligned}
&\sum_{k=1}^{K} \eta^{(k)}(x) \\
&= \sum_{k=1}^{K} \sum_{\mathbf{y} \in \{0,1\}^K} \mathbf{P}\left(\mathsf{Y}^{(1)} = y_1, \ldots, \mathsf{Y}^{(K)} = y_K \,\middle|\, \mathsf{X} = x\right) \cdot \mathbf{1}(y_k = 1) \\
&= \sum_{k=1}^{K} \sum_{\mathbf{y} \in \{0,1\}^K} \mathbf{P}\left(\mathsf{Y}^{(1)} = y_1, \ldots, \mathsf{Y}^{(K)} = y_K \,\middle|\, \mathsf{X} = x\right) \cdot \mathbf{1}(y_k = 1) \cdot \mathbf{1}\Big(\textstyle\sum_j y_j \geq 1\Big) \\
&= \sum_{k=1}^{K} \sum_{\mathbf{y} \in \{0,1\}^K} \mathbf{P}\left(\mathsf{Y}^{(1)} = y_1, \ldots, \mathsf{Y}^{(K)} = y_K \,\middle|\, \mathsf{X} = x\right) \cdot \mathbf{1}(y_k = 1) \cdot \sum_{n=1}^{K} \mathbf{1}\Big(\textstyle\sum_j y_j = n\Big) \\
&= \sum_{k=1}^{K} \sum_{n=1}^{K} \sum_{\mathbf{y} \in \{0,1\}^K} \mathbf{P}\left(\mathsf{Y}^{(1)} = y_1, \ldots, \mathsf{Y}^{(K)} = y_K \,\middle|\, \mathsf{X} = x\right) \cdot \mathbf{1}(y_k = 1) \cdot \mathbf{1}\Big(\textstyle\sum_j y_j = n\Big)
\end{aligned}
$$

$$= \sum_{k=1}^{K}\sum_{n=1}^{K} \sum_{\mathbf{y}\in\{0,1\}^K} \mathbf{P}\left(\mathsf{Y}^{(1)} = y_1, \ldots, \mathsf{Y}^{(K)} = y_K \,\middle|\, \mathsf{X} = x\right) \cdot \mathbf{1}\Big(\textstyle\sum_j y_j = n\Big) \cdot \mathbf{1}(y_k = 1)$$
$$= \sum_{n=1}^{K} \sum_{\mathbf{y}\in\{0,1\}^K} \mathbf{P}\left(\mathsf{Y}^{(1)} = y_1, \ldots, \mathsf{Y}^{(K)} = y_K \,\middle|\, \mathsf{X} = x\right) \cdot \mathbf{1}\Big(\textstyle\sum_j y_j = n\Big) \cdot \textstyle\sum_{k=1}^{K} \mathbf{1}(y_k = 1)$$
$$= \sum_{n=1}^{K} \sum_{\mathbf{y}\in\{0,1\}^K} \mathbf{P}\left(\mathsf{Y}^{(1)} = y_1, \ldots, \mathsf{Y}^{(K)} = y_K \,\middle|\, \mathsf{X} = x\right) \cdot \mathbf{1}\Big(\textstyle\sum_j y_j = n\Big) \cdot n$$
$$= \sum_{n=1}^{K} n \cdot \sum_{\mathbf{y}\in\{0,1\}^K} \mathbf{P}\left(\mathsf{Y}^{(1)} = y_1, \ldots, \mathsf{Y}^{(K)} = y_K \,\middle|\, \mathsf{X} = x\right) \cdot \mathbf{1}\Big(\textstyle\sum_j y_j = n\Big)$$
$$= \sum_{n=1}^{K} n \cdot \mathbf{P}\left(\bar{\mathsf{Y}} = n \,\middle|\, \mathsf{X} = x\right)$$
$$= \sum_{n=1}^{K} n \cdot \bar{\eta}_n(x).$$

Since the optimal scorer has the same form as that for the loss aggregated AUC in Proposition 5.2 when $a_k = \pi^{(k)} \cdot (1 - \pi^{(k)}), \forall k$, and we know that the optimal scorers for the loss aggregated AUC are Pareto optimal from Proposition 5.1, it follows that the above solution is also Pareto optimal. □

Proof of Proposition 6.1. We know from Proposition 5.2 that the optimal scorer will preserve the ordering of $f^*(x) = \alpha^{(1)} \cdot \eta^{(1)}(x) + \alpha^{(2)} \cdot \eta^{(2)}(x)$.

We start with the case where $\alpha^{(1)} > \alpha^{(2)}$. Since the labels are deterministic $\eta^{(1)}(x), \eta^{(2)}(x) \in \{0, 1\}$. Hence when $\eta^{(1)}(x) > \eta^{(1)}(x')$, we have that $\eta^{(1)}(x) = 1$ and $\eta^{(1)}(x') = 0$. Therefore

$$\begin{aligned} f^*(x) &= \alpha^{(1)} + \alpha^{(2)} \cdot \eta^{(2)}(x) \geq \alpha^{(1)} > \alpha^{(2)} \geq \alpha^{(2)} \cdot \eta^{(2)}(x') \\ &= \alpha^{(1)} \cdot \eta^{(1)}(x') + \alpha^{(2)} \cdot \eta^{(2)}(x') = f^*(x'). \end{aligned}$$

Similarly, when $\alpha^{(2)} > \alpha^{(1)}$ and $\eta^{(2)}(x) > \eta^{(2)}(x')$, we have $\eta^{(2)}(x) = 1$ and $\eta^{(2)}(x') = 0$, and therefore:

$$\begin{aligned} f^*(x) &= \alpha^{(1)} \cdot \eta^{(1)}(x) + \alpha^{(2)} \geq \alpha^{(2)} > \alpha^{(1)} \geq \alpha^{(1)} \cdot \eta^{(1)}(x') \\ &= \alpha^{(1)} \cdot \eta^{(1)}(x') + \alpha^{(2)} \cdot \eta^{(2)}(x') = f^*(x'), \end{aligned}$$

which completes the proof. □

Proof of Proposition 6.2. The proof for ordinal costs $c_{\bar{y}\bar{y}'} = 1(\bar{y} > \bar{y}') \cdot |\bar{y} - \bar{y}'|$ follows directly from the more general case in Proposition 5.4.

We now provide the proof for costs $c_{\bar{y}\bar{y}'} = 1$. By Lemma 2.4, since $K = 3$ ($\bar{\mathsf{Y}}$ can be $0, 1$ or 2), the Bayes-optimal scorer will preserve the ordering of

$$\begin{aligned} & f^*(x) \\ =\; & \frac{\bar{\eta}_1(x) + \bar{\eta}_2(x)}{\bar{\eta}_0(x) + \bar{\eta}_1(x)} \\ =\; & \frac{\mathbf{P}\big(\mathsf{Y}^{(1)} = 0, \mathsf{Y}^{(2)} = 1 \,\big|\, \mathsf{X} = x\big) + \mathbf{P}\big(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 0 \,\big|\, \mathsf{X} = x\big) + \mathbf{P}\big(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 1 \,\big|\, \mathsf{X} = x\big)}{\mathbf{P}\big(\mathsf{Y}^{(1)} = 0, \mathsf{Y}^{(2)} = 0 \,\big|\, \mathsf{X} = x\big) + \mathbf{P}\big(\mathsf{Y}^{(1)} = 0, \mathsf{Y}^{(2)} = 1 \,\big|\, \mathsf{X} = x\big) + \mathbf{P}\big(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 0 \,\big|\, \mathsf{X} = x\big)}. \end{aligned}$$

We will show that $f^*(x)$ is a strictly monotonic transformation of:

$$\eta^{(1)}(x) + \eta^{(2)}(x) = \mathbf{P}\big(\mathsf{Y}^{(1)} = 0, \mathsf{Y}^{(2)} = 1 \,\big|\, \mathsf{X} = x\big) + \mathbf{P}\big(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 0 \,\big|\, \mathsf{X} = x\big)$$

$$+ 2 \cdot \mathbf{P}\big(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 1 \,\big|\, \mathsf{X} = x\big).$$

Since the labels are deterministic, $\mathbf{P}\big(\mathsf{Y}^{(1)} = i, \mathsf{Y}^{(2)} = j \,\big|\, \mathsf{X} = x\big) = 1$ for exactly one particular (i, j). We consider all four possible cases:

(1) $\mathbf{P}\big(\mathsf{Y}^{(1)} = 0, \mathsf{Y}^{(2)} = 0 \,\big|\, \mathsf{X} = x\big) = 1$. We have: $f^*(x) = 0$ and $\eta^{(1)}(x) + \eta^{(2)}(x) = 0$.

(2) $\mathbf{P}\big(\mathsf{Y}^{(1)} = 0, \mathsf{Y}^{(2)} = 1 \,\big|\, \mathsf{X} = x\big) = 1$. We have: $f^*(x) = 1$ and $\eta^{(1)}(x) + \eta^{(2)}(x) = 1$.

(3) $\mathbf{P}\big(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 0 \,\big|\, \mathsf{X} = x\big) = 1$. We have: $f^*(x) = 1$ and $\eta^{(1)}(x) + \eta^{(2)}(x) = 1$.

(4) $\mathbf{P}\big(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 1 \,\big|\, \mathsf{X} = x\big) = 1$. We have: $f^*(x) = \infty$ and $\eta^{(1)}(x) + \eta^{(2)}(x) = 2$.

Clearly, $f^*(x)$ is a strictly monotonic transformation of $\eta^{(1)}(x) + \eta^{(2)}(x)$. □

Proof of Lemma 6.3. The first statement (a) follows directly from Proposition 5.2.

For the second statement (b), let $\bar{\mathsf{Y}} = \prod_k \mathsf{Y}^{(k)} \in \{0, 1\}$ have class probability function $\bar{\eta}_{\mathrm{prod}} \colon \mathcal{X} \to [0, 1]$. Applying Lemma 2.2, the Bayes-optimal scorer f^*_{prod}ill preserve the ordering of:

$$\gamma_{\mathrm{prod}}(x) = \bar{\eta}_{\mathrm{prod}}(x) = \mathbf{P}(\mathsf{Y}^{(1)} = 1, \ldots, \mathsf{Y}^{(K)} = 1 \,|\, \mathsf{X} = x).$$

For the third statement (c), given that the labels are deterministic, we note that $\gamma_{\mathrm{prod}}(x) > \gamma_{\mathrm{prod}}(x')$ implies that $\mathbf{P}(\mathsf{Y}^{(1)} = 1, \ldots, \mathsf{Y}^{(K)} = 1 \,|\, \mathsf{X} = x) = 1$, while $\mathbf{P}(\mathsf{Y}^{(1)} = 1, \ldots, \mathsf{Y}^{(K)} = 1 \,|\, \mathsf{X} = x') = 0$. This indicates that $\gamma_{\mathrm{sum}}(x) = \sum_k \eta^{(k)}(x) = K$ and $\gamma_{\mathrm{sum}}(x') = \sum_k \eta^{(k)}(x') < K$. As a result, $\gamma_{\mathrm{sum}}(x) > \gamma_{\mathrm{sum}}(x')$. □

C. Additional theoretical results

Theorem C.1. *Suppose that the labels are binary, i.e., $\mathsf{Y}^{(k)} \in \{0, 1\}$ for all $k \in \mathbb{N}_{\geq 1}$, and $a_k > 0$. Assume further that the labels $\{\mathsf{Y}^{(k)}\}_{k \in \mathbb{N}_{\geq 1}}$ are jointly independent conditioned on X. Consider the aggregation function $\psi(\mathsf{Y}^{(1)}, \ldots, \mathsf{Y}^{(K)}) = \sum_{k \in [K]} a_k \mathsf{Y}^{(k)}$. Define*

$$f^* \doteq \underset{f : \mathcal{X} \to \mathbb{R}}{\operatorname{argmax}} \ \mathrm{AUC}_{\mathrm{LA}}\left(f; \{D^{(k)}\}\right),$$

$$\tilde{f}(x) \doteq \phi\left(\sum_{k \in [K]} a_k \eta_k(x)\right),$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a strictly increasing function. Then,

$$\mathrm{AUC}_{\mathrm{LA}}\left(f^*; \{D^{(k)}\}\right) - \mathrm{AUC}_{\mathrm{LA}}\left(\tilde{f}; \{D^{(k)}\}\right) \leq \psi\left(\mathbb{E}_{\mathsf{X}_1, \mathsf{X}_2}\left[\frac{\sum_{k \in [K]} a_k^3 \sum_{i \in \{1,2\}} \eta_k(\mathsf{X}_i)\left(1 - \eta_k(\mathsf{X}_i)\right)}{\left(\sum_{k \in [K]} a_k^2 \sum_{i \in \{1,2\}} \eta_k(\mathsf{X}_i)\left(1 - \eta_k(\mathsf{X}_i)\right)\right)^{3/2}}\right]\right),$$

where $\psi(t) \triangleq \frac{2t}{1-t}$, and X_1 and X_2 are i.i.d. copies of X.

Corollary C.2. *If $a_k = 1$ for all k and $\eta_k(x)$ is bounded away from 0 and 1, i.e., there exists $c \in (0, 1/2)$ such that $\eta_k(x) \in [c, 1 - c]$ for all $k \in \mathbb{N}_{\geq 1}$ and for all $x \in \mathcal{X}$, then the expected AUC difference between the optimal prediction function f^* and $\tilde{f}$ converges to zero with a rate of $O(1/\sqrt{K})$:*

$$\mathrm{AUC}_{\mathrm{LA}}\left(f^*; \{D^{(k)}\}\right) - \mathrm{AUC}_{\mathrm{LA}}\left(\tilde{f}; \{D^{(k)}\}\right) = O\left(\frac{1}{\sqrt{K}}\right).$$

Proof of Theorem C.1. Given a function $f : \mathcal{X} \to \mathbb{R}$, we first rewrite $\mathrm{AUC}\left(f; \{D^{(k)}\}\right)$ as follows:

$$\mathrm{AUC}_{\mathrm{LaA}}\left(f; \{D^{(k)}\}\right) = \mathbb{E}\left[H(f(\mathsf{X}_2) - f(\mathsf{X}_1) \mid \sum_{k \in [K]} a_k \mathsf{Y}_1^{(k)} < \sum_{k \in [K]} a_k \mathsf{Y}_2^{(k)}\right]$$

$$= \frac{\mathbb{E}\left[H(f(\mathsf{X}_2) - f(\mathsf{X}_1)\mathbb{1}_{\{\sum_{k\in[K]} a_k \mathsf{Y}_1^{(k)} < \sum_{k\in[K]} a_k \mathsf{Y}_2^{(k)}\}}\right]}{\mathbb{P}\left(\sum_{k\in[K]} a_k \mathsf{Y}_1^{(k)} < \sum_{k\in[K]} a_k \mathsf{Y}_2^{(k)}\right)}$$
$$= \frac{\mathbb{E}\left[H(f(\mathsf{X}_2) - f(\mathsf{X}_1)\mathbb{1}_{\{\sum_{k\in[K]} a_k \mathsf{Y}_1^{(k)} < \sum_{k\in[K]} a_k \mathsf{Y}_2^{(k)}\}}\right]}{c_K},$$

where

$$c_K \triangleq \mathbb{P}\left(\sum_{k\in[K]} a_k \mathsf{Y}_1^{(k)} < \sum_{k\in[K]} a_k \mathsf{Y}_2^{(k)}\right)$$
$$= \frac{1}{2}\left(1 - \mathbb{P}\left(\sum_{k\in[K]} a_k \mathsf{Y}_1^{(k)} - \sum_{k\in[K]} a_k \mathsf{Y}_2^{(k)} = 0\right)\right).$$

Define

$$\nu(\mathsf{X}_1, \mathsf{X}_2) \triangleq \frac{\sum_{k\in[K]} a_k\left(\eta_k(\mathsf{X}_2) - \eta_k(\mathsf{X}_1)\right)}{\sqrt{\sum_{k\in[K]} a_k^2 \sum_{i\in\{1,2\}} \eta_k(\mathsf{X}_i)\left(1 - \eta_k(\mathsf{X}_i)\right)}}$$

and the functionals

$$F_K(f) \triangleq \mathbb{E}\left[H(f(\mathsf{X}_2) - f(\mathsf{X}_1))\mathbb{1}_{\{\sum_{k\in[K]} a_k \mathsf{Y}_1^{(k)} < \sum_{k\in[K]} a_k \mathsf{Y}_2^{(k)}\}}\right],$$
$$F'_K(f) \triangleq \mathbb{E}\left[H(f(\mathsf{X}_2) - f(\mathsf{X}_1))\Phi\left(\nu\left(\mathsf{X}_1, \mathsf{X}_2\right)\right)\right],$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

Next, we will show that $\tilde{f}$ maximizes $F'_K(\cdot)$. In other words, we aim to prove that for any f,

$$F'_K(f) \le F'_K(\tilde{f}).$$

On one hand, we have

$$F'_K(f) = \frac{1}{2}\mathbb{E}\left[H(f(\mathsf{X}_2) - f(\mathsf{X}_1))\Phi\left(\nu\left(\mathsf{X}_1, \mathsf{X}_2\right)\right) + H(f(\mathsf{X}_1) - f(\mathsf{X}_2))\Phi\left(\nu\left(\mathsf{X}_2, \mathsf{X}_1\right)\right)\right]$$
$$= \frac{1}{2}\mathbb{E}\left[H(f(\mathsf{X}_2) - f(\mathsf{X}_1))\Phi\left(\nu\left(\mathsf{X}_1, \mathsf{X}_2\right)\right) + H(f(\mathsf{X}_1) - f(\mathsf{X}_2))\left(1 - \Phi\left(\nu\left(\mathsf{X}_1, \mathsf{X}_2\right)\right)\right)\right]$$
$$\le \frac{1}{2}\mathbb{E}\left[\max\left\{\Phi\left(\nu\left(\mathsf{X}_1, \mathsf{X}_2\right)\right), 1 - \Phi\left(\nu\left(\mathsf{X}_1, \mathsf{X}_2\right)\right)\right\}\right].$$

On the other hand, let $S_1 = \sum_{k\in[K]} a_k\eta_k(\mathsf{X}_1)$, $S_2 = \sum_{k\in[K]} a_k\eta_k(\mathsf{X}_2)$, and $\Phi_\nu = \Phi\left(\nu\left(\mathsf{X}_1, \mathsf{X}_2\right)\right)$. Since ϕ is strictly increasing, $H(\tilde{f}(\mathsf{X}_2) - \tilde{f}(\mathsf{X}_1)) = H(S_2 - S_1)$ and $H(\tilde{f}(\mathsf{X}_1) - \tilde{f}(\mathsf{X}_2)) = H(S_1 - S_2)$. Then,

$$F'_K(\tilde{f}) = \mathbb{E}\left[H(\tilde{f}(\mathsf{X}_2) - \tilde{f}(\mathsf{X}_1))\Phi\left(\nu\left(\mathsf{X}_1, \mathsf{X}_2\right)\right)\right]$$
$$= \frac{1}{2}\mathbb{E}\left[H(\tilde{f}(\mathsf{X}_2) - \tilde{f}(\mathsf{X}_1))\Phi_\nu + H(\tilde{f}(\mathsf{X}_1) - \tilde{f}(\mathsf{X}_2))\left(1 - \Phi_\nu\right)\right]$$
$$= \frac{1}{2}\mathbb{E}\left[\left(\mathbb{1}_{\{S_2 > S_1\}} + \frac{1}{2}\cdot\mathbb{1}_{\{S_2 = S_1\}}\right)\Phi_\nu + \left(\mathbb{1}_{\{S_1 > S_2\}} + \frac{1}{2}\cdot\mathbb{1}_{\{S_2 = S_1\}}\right)\left(1 - \Phi_\nu\right)\right]$$
$$= \frac{1}{2}\mathbb{E}\left[\max\left\{\Phi_\nu, 1 - \Phi_\nu\right\}\right].$$

Now, we proceed to bound the gap between $F_K(f)$ and $F'_K(f)$. We first rewrite $F_K(f)$ as

$$F_K(f)$$

$$
\begin{aligned}
&= \mathbb{E}\left[H(f(\mathsf{X}_2) - f(\mathsf{X}_1))\mathbb{E}\left[\mathbb{1}_{\{\sum_{k\in[K]} a_k \mathsf{Y}_1^{(k)} < \sum_{k\in[K]} a_k \mathsf{Y}_2^{(k)}\}} \mid \mathsf{X}_1, \mathsf{X}_2\right]\right] \\
&= \mathbb{E}\left[H(f(\mathsf{X}_2) - f(\mathsf{X}_1))\mathbb{P}\left(\sum_{k\in[K]} a_k \mathsf{Y}_1^{(k)} - \sum_{k\in[K]} \mathsf{Y}_2^{(k)} < 0 \mid \mathsf{X}_1, \mathsf{X}_2\right)\right] \\
&= \mathbb{E}\left[H(f(\mathsf{X}_2) - f(\mathsf{X}_1))\mathbb{P}\left(\sum_{k\in[K]} \mathsf{Y}_k + \sum_{k\in[K]} \mathsf{Y}'_k < \sum_{k\in[K]} a_k\left(\eta_k(\mathsf{X}_2) - \eta_k(\mathsf{X}_1)\right) \mid \mathsf{X}_1, \mathsf{X}_2\right)\right] \\
&= \mathbb{E}\left[H(f(\mathsf{X}_2) - f(\mathsf{X}_1))\mathbb{P}\left(\mathsf{S}_K < \nu\left(\mathsf{X}_1, \mathsf{X}_2\right) \mid \mathsf{X}_1, \mathsf{X}_2\right)\right],
\end{aligned}
$$

where

$$
\begin{aligned}
\mathsf{Y}_k &\triangleq a_k\left(\mathsf{Y}_1^{(k)} - \eta_k(\mathsf{X}_1)\right), \\
\mathsf{Y}'_k &\triangleq -a_k\left(\mathsf{Y}_2^{(k)} - \eta_k(\mathsf{X}_2)\right), \\
\mathsf{S}_K &\triangleq \frac{\sum_{k\in[K]} \mathsf{Y}_k + \sum_{k\in[K]} \mathsf{Y}'_k}{\sqrt{\sum_{k\in[K]} a_k^2 \sum_{i\in\{1,2\}} \eta_k(\mathsf{X}_i)\left(1 - \eta_k(\mathsf{X}_i)\right)}}.
\end{aligned}
$$

Note that Y_k and Y'_k have zero mean, i.e., $\mathbb{E}\left[\mathsf{Y}_k \mid \mathsf{X}_1, \mathsf{X}_2\right] = \mathbb{E}\left[\mathsf{Y}'_k \mid \mathsf{X}_1, \mathsf{X}_2\right] = 0$. Moreover, we have

$$
\begin{aligned}
\mathbb{E}\left[|\mathsf{Y}_k|^2 \mid \mathsf{X}_1, \mathsf{X}_2\right] &= a_k^2 \mathbb{E}\left[\left|\mathsf{Y}_1^{(k)} - \eta_k(\mathsf{X}_1)\right|^2 \mid \mathsf{X}_1, \mathsf{X}_2\right] \\
&= a_k^2 \eta_k(\mathsf{X}_1)\left(1 - \eta_k(\mathsf{X}_1)\right), \\
\mathbb{E}\left[|\mathsf{Y}_k|^3 \mid \mathsf{X}_1, \mathsf{X}_2\right] &= a_k^3 \mathbb{E}\left[\left|\mathsf{Y}_1^{(k)} - \eta_k(\mathsf{X}_1)\right|^3 \mid \mathsf{X}_1, \mathsf{X}_2\right] \\
&= a_k^3 \eta_k(\mathsf{X}_1)\left(1 - \eta_k(\mathsf{X}_1)\right)\left[\left(1 - \eta_k(\mathsf{X}_1)\right)^2 + \eta_k(\mathsf{X}_1)^2\right] \\
&\leq \frac{1}{2} a_k^3 \eta_k(\mathsf{X}_1)\left(1 - \eta_k(\mathsf{X}_1)\right).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\mathbb{E}\left[|\mathsf{Y}'_k|^2 \mid \mathsf{X}_1, \mathsf{X}_2\right] &= a_k^2 \eta_k(\mathsf{X}_2)\left(1 - \eta_k(\mathsf{X}_2)\right), \\
\mathbb{E}\left[|\mathsf{Y}'_k|^3 \mid \mathsf{X}_1, \mathsf{X}_2\right] &\leq \frac{1}{2} a_k^3 \eta_k(\mathsf{X}_2)\left(1 - \eta_k(\mathsf{X}_2)\right).
\end{aligned}
$$

By the Berry-Esseen theorem, for any t,

$$
\begin{aligned}
&\left|\mathbb{P}\left(\mathsf{S}_K < t \mid \mathsf{X}_1, \mathsf{X}_2\right) - \Phi\left(t\right)\right| \\
&\leq \left(\sum_{k\in[K]} \mathbb{E}\left[|\mathsf{Y}_k|^2 \mid \mathsf{X}_1, \mathsf{X}_2\right] + \sum_{k\in[K]} \mathbb{E}\left[|\mathsf{Y}'_k|^2 \mid \mathsf{X}_1, \mathsf{X}_2\right]\right)^{-3/2} \cdot \\
&\quad \left(\sum_{k\in[K]} \mathbb{E}\left[|\mathsf{Y}_k|^3 \mid \mathsf{X}_1, \mathsf{X}_2\right] + \sum_{k\in[K]} \mathbb{E}\left[|\mathsf{Y}'_k|^3 \mid \mathsf{X}_1, \mathsf{X}_2\right]\right) \\
&\leq \frac{\sum_{k\in[K]} a_k^3 \sum_{i\in\{1,2\}} \eta_k(\mathsf{X}_i)\left(1 - \eta_k(\mathsf{X}_i)\right)}{2\left(\sum_{k\in[K]} a_k^2 \sum_{i\in\{1,2\}} \eta_k(\mathsf{X}_i)\left(1 - \eta_k(\mathsf{X}_i)\right)\right)^{3/2}}.
\end{aligned}
$$

We can now bound the gap between $F_K(f)$ and $F'_K(f)$:

$$
|F_K(f) - F'_K(f)|
$$

$$
\begin{aligned}
&= \left|\mathbb{E}\left[H(f(\mathsf{X}_2) - f(\mathsf{X}_1)\mathbb{P}\left(\mathsf{S}_K < \nu\left(\mathsf{X}_1, \mathsf{X}_2\right) \mid \mathsf{X}_1, \mathsf{X}_2\right)\right] - \mathbb{E}\left[H(f(\mathsf{X}_2) - f(\mathsf{X}_1))\Phi\left(\nu\left(\mathsf{X}_1, \mathsf{X}_2\right)\right)\right]\right| \\
&\leq \mathbb{E}\left[\left|H(f(\mathsf{X}_2) - f(\mathsf{X}_1)\right| \left|\mathbb{P}\left(\mathsf{S}_K < \nu\left(\mathsf{X}_1, \mathsf{X}_2\right) \mid \mathsf{X}_1, \mathsf{X}_2\right) - \Phi\left(\nu\left(\mathsf{X}_1, \mathsf{X}_2\right)\right)\right|\right] \\
&\leq \mathbb{E}\left[\left|\mathbb{P}\left(\mathsf{S}_K < \nu\left(\mathsf{X}_1, \mathsf{X}_2\right) \mid \mathsf{X}_1, \mathsf{X}_2\right) - \Phi\left(\nu\left(\mathsf{X}_1, \mathsf{X}_2\right)\right)\right|\right] \\
&\leq \mathbb{E}\left[\frac{\sum_{k\in[K]} a_k^3 \sum_{i\in\{1,2\}} \eta_k(\mathsf{X}_i)\left(1-\eta_k(\mathsf{X}_i)\right)}{2\left(\sum_{k\in[K]} a_k^2 \sum_{i\in\{1,2\}} \eta_k(\mathsf{X}_i)\left(1-\eta_k(\mathsf{X}_i)\right)\right)^{3/2}}\right].
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& F_K(f^*) - F_K(\tilde{f}) \\
&= \left(F_K(f^*) - F_K'(f^*)\right) + \left(F_K'(f^*) - F_K'(\tilde{f})\right) + \left(F_K'(\tilde{f}) - F_K(\tilde{f})\right) \\
&\leq \mathbb{E}\left[\frac{\sum_{k\in[K]} a_k^3 \sum_{i\in\{1,2\}} \eta_k(\mathsf{X}_i)\left(1-\eta_k(\mathsf{X}_i)\right)}{\left(\sum_{k\in[K]} a_k^2 \sum_{i\in\{1,2\}} \eta_k(\mathsf{X}_i)\left(1-\eta_k(\mathsf{X}_i)\right)\right)^{3/2}}\right].
\end{aligned}
$$

Next, we establish an upper bound for $\mathbb{P}\left(\sum_{k\in[K]} a_k \mathsf{Y}_1^{(k)} - \sum_{k\in[K]} a_k \mathsf{Y}_2^{(k)} = 0\right)$:

$$
\begin{aligned}
& \mathbb{P}\left(\sum_{k\in[K]} a_k \mathsf{Y}_1^{(k)} - \sum_{k\in[K]} a_k \mathsf{Y}_2^{(k)} = 0\right) \\
&= \mathbb{E}\left[\mathbb{P}\left(\sum_{k\in[K]} a_k \mathsf{Y}_1^{(k)} - \sum_{k\in[K]} a_k \mathsf{Y}_2^{(k)} = 0 \mid \mathsf{X}_1, \mathsf{X}_2\right)\right] \\
&= \mathbb{E}\left[\mathbb{P}\left(\mathsf{S}_K = \nu\left(\mathsf{X}_1, \mathsf{X}_2\right) \mid \mathsf{X}_1, \mathsf{X}_2\right)\right] \\
&= \mathbb{E}\left[\mathbb{P}\left(\mathsf{S}_K \leq \nu\left(\mathsf{X}_1, \mathsf{X}_2\right) \mid \mathsf{X}_1, \mathsf{X}_2\right) - \lim_{t\to\nu(\mathsf{X}_1,\mathsf{X}_2)^-} \mathbb{P}\left(\mathsf{S}_K \leq t \mid \mathsf{X}_1, \mathsf{X}_2\right)\right] \\
&= \mathbb{E}\left[\mathbb{P}\left(\mathsf{S}_K \leq \nu\left(\mathsf{X}_1, \mathsf{X}_2\right) \mid \mathsf{X}_1, \mathsf{X}_2\right) - \Phi\left(\nu\left(\mathsf{X}_1, \mathsf{X}_2\right)\right) - \lim_{t\to\nu(\mathsf{X}_1,\mathsf{X}_2)^-} \left(\mathbb{P}\left(\mathsf{S}_K \leq t \mid \mathsf{X}_1, \mathsf{X}_2\right) - \Phi\left(t\right)\right)\right].
\end{aligned}
$$

Taking the absolute value, we get

$$
\begin{aligned}
& \mathbb{P}\left(\sum_{k\in[K]} a_k \mathsf{Y}_1^{(k)} - \sum_{k\in[K]} a_k \mathsf{Y}_2^{(k)} = 0\right) \\
&\leq \mathbb{E}\left[\left|\mathbb{P}\left(\mathsf{S}_K \leq \nu\left(\mathsf{X}_1, \mathsf{X}_2\right) \mid \mathsf{X}_1, \mathsf{X}_2\right) - \Phi\left(\nu\left(\mathsf{X}_1, \mathsf{X}_2\right)\right)\right| + \lim_{t\to\nu(\mathsf{X}_1,\mathsf{X}_2)^-} \left|\left(\mathbb{P}\left(\mathsf{S}_K \leq t \mid \mathsf{X}_1, \mathsf{X}_2\right) - \Phi\left(t\right)\right)\right|\right] \\
&\leq \mathbb{E}\left[\frac{\sum_{k\in[K]} a_k^3 \sum_{i\in\{1,2\}} \eta_k(\mathsf{X}_i)\left(1-\eta_k(\mathsf{X}_i)\right)}{\left(\sum_{k\in[K]} a_k^2 \sum_{i\in\{1,2\}} \eta_k(\mathsf{X}_i)\left(1-\eta_k(\mathsf{X}_i)\right)\right)^{3/2}}\right].
\end{aligned}
$$

Consequently, we obtain a lower bound for c_K:

$$
c_K \geq \max\left\{\frac{1}{2}\left(1 - \mathbb{E}\left[\frac{\sum_{k\in[K]} a_k^3 \sum_{i\in\{1,2\}} \eta_k(\mathsf{X}_i)\left(1-\eta_k(\mathsf{X}_i)\right)}{\left(\sum_{k\in[K]} a_k^2 \sum_{i\in\{1,2\}} \eta_k(\mathsf{X}_i)\left(1-\eta_k(\mathsf{X}_i)\right)\right)^{3/2}}\right]\right), 0\right\}.
$$

Recall that $\mathrm{AUC}_{\mathrm{LaA}}(f; \{D^{(k)}\}) = \frac{F_K(f)}{c_K}$. We then have

$$
\mathrm{AUC}_{\mathrm{LaA}}\left(f^*; \{D^{(k)}\}\right) - \mathrm{AUC}_{\mathrm{LaA}}\left(\tilde{f}; \{D^{(k)}\}\right)
$$

$$
\begin{aligned}
&= \frac{F_K(f^*) - F_K(\tilde{f})}{c_K} \\
&\leq \frac{2 \cdot \mathbb{E}\left[\frac{\sum_{k\in[K]} a_k^3 \sum_{i\in\{1,2\}} \eta_k(\mathsf{X}_i)(1-\eta_k(\mathsf{X}_i))}{\left(\sum_{k\in[K]} a_k^2 \sum_{i\in\{1,2\}} \eta_k(\mathsf{X}_i)(1-\eta_k(\mathsf{X}_i))\right)^{3/2}}\right]}{\left(1 - \mathbb{E}\left[\frac{\sum_{k\in[K]} a_k^3 \sum_{i\in\{1,2\}} \eta_k(\mathsf{X}_i)(1-\eta_k(\mathsf{X}_i))}{\left(\sum_{k\in[K]} a_k^2 \sum_{i\in\{1,2\}} \eta_k(\mathsf{X}_i)(1-\eta_k(\mathsf{X}_i))\right)^{3/2}}\right]\right)} \\
&= \psi\left(\mathbb{E}\left[\frac{\sum_{k\in[K]} a_k^3 \sum_{i\in\{1,2\}} \eta_k(\mathsf{X}_i)\,(1-\eta_k(\mathsf{X}_i))}{\left(\sum_{k\in[K]} a_k^2 \sum_{i\in\{1,2\}} \eta_k(\mathsf{X}_i)\,(1-\eta_k(\mathsf{X}_i))\right)^{3/2}}\right]\right).
\end{aligned}
$$

□

Proof of Corollary C.2. By Theorem C.1, we have

$$
\begin{aligned}
&\mathrm{AUC}_{\mathrm{LaA}}\left(f^*; \{D^{(k)}\}\right) - \mathrm{AUC}_{\mathrm{LaA}}\left(\tilde{f}; \{D^{(k)}\}\right) \\
&\leq \psi\left(\mathbb{E}\left[\frac{1}{\left(\sum_{k\in[K]} \sum_{i\in\{1,2\}} \eta_k(\mathsf{X}_i)\,(1-\eta_k(\mathsf{X}_i))\right)^{1/2}}\right]\right) \\
&\leq \psi\left(\frac{1}{\sqrt{2c(1-c)K}}\right) \\
&= O\left(\frac{1}{\sqrt{K}}\right).
\end{aligned}
$$

□

D. Further Analysis of Label Aggregation

This section addresses further properties of label aggregation, including its behavior with anti-correlated labels and its sensitivity to mixing weights, expanding on the discussions in Section 5.2.

D.1. Behavior with Anti-correlated Labels

Our theory (Proposition 5.4) shows the optimal scorer for label aggregation (cost $c_{\bar{y}\bar{y}'} = |\bar{y} - \bar{y}'|$) ranks by $f^*(x) \propto \sum_k \eta^{(k)}(x)$. For $K = 2$, this is equivalent to ranking by $\delta(x) = p(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 1 \mid \mathsf{X} = x) - p(\mathsf{Y}^{(1)} = 0, \mathsf{Y}^{(2)} = 0 \mid \mathsf{X} = x)$. The derivation below shows $f^*(x) \propto \delta(x)$.

From Proposition 5.4, the scorer is $\eta^{(1)}(x) + \eta^{(2)}(x)$. We have $\eta^{(1)}(x) = p(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 0 \mid \mathsf{X} = x) + p(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 1 \mid \mathsf{X} = x)$ and $\eta^{(2)}(x) = p(\mathsf{Y}^{(1)} = 0, \mathsf{Y}^{(2)} = 1 \mid \mathsf{X} = x) + p(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 1 \mid \mathsf{X} = x)$. Summing them gives $\eta^{(1)}(x) + \eta^{(2)}(x) = p(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 0 \mid \mathsf{X} = x) + p(\mathsf{Y}^{(1)} = 0, \mathsf{Y}^{(2)} = 1 \mid \mathsf{X} = x) + 2p(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 1 \mid \mathsf{X} = x)$. Since $p(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 0 \mid \mathsf{X} = x) + p(\mathsf{Y}^{(1)} = 0, \mathsf{Y}^{(2)} = 1 \mid \mathsf{X} = x) + p(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 1 \mid \mathsf{X} = x) + p(\mathsf{Y}^{(1)} = 0, \mathsf{Y}^{(2)} = 0 \mid \mathsf{X} = x) = 1$, the sum simplifies to $1 - p(\mathsf{Y}^{(1)} = 0, \mathsf{Y}^{(2)} = 0 \mid \mathsf{X} = x) + p(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 1 \mid \mathsf{X} = x) = 1 + \delta(x)$.

This shows the optimal scorer *does* depend on label correlations via $\delta(x)$. Let's consider the implications:

- **Overlapping ($\mathsf{Y}^{(1)} = \mathsf{Y}^{(2)}$):** $\delta(x) \propto \eta_1(x)$, so the ranking uses $\eta_1(x)$, which is sensible.
- **Anti-correlated ($\mathsf{Y}^{(1)} = 1 - \mathsf{Y}^{(2)}$):** $\delta(x) = 0$. The scorer $f^*(x)$ is constant, yielding no ranking. This indicates reduced robustness when the aggregated label $\mathsf{Y}^{(1)} + \mathsf{Y}^{(2)} = 1$ is constant, making the AUC objective ill-defined.
- **Mild anti-correlation:** Here, both $p(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 1 \mid \mathsf{X} = x)$ and $p(\mathsf{Y}^{(1)} = 0, \mathsf{Y}^{(2)} = 0 \mid \mathsf{X} = x)$ would be small, and the ranking depends on their difference via $\delta(x)$. E.g., when $\delta(x) > 0$, it is more likely that both labels are 1 than 0, resulting in x being ranked higher. The behavior would depend on the exact conditional probabilities.

D.2. Sensitivity of Label Aggregation to Mixing Weights

We next generalize label aggregation by using non-uniform weights $\alpha_1, \alpha_2 > 0$ for the aggregation function, i.e., $\bar{Y} = \alpha_1 \cdot \mathsf{Y}^{(1)} + \alpha_2 \cdot \mathsf{Y}^{(2)}$, again for the $K = 2$ case for simplicity.

Following assumptions of Proposition 5.4 (that the Bayes-optimal scorer for costs $c_{\bar{y}\bar{y}'} = |\bar{y} - \bar{y}'|$ preserves the ordering of $E[\bar{Y}|X = x]$), the Bayes-optimal scorer $\bar{\eta}(x) = E[\bar{Y}|X = x]$ becomes:

$$\begin{aligned}
\bar{\eta}(x) &= 0 \cdot p(\mathsf{Y}^{(1)} = 0, \mathsf{Y}^{(2)} = 0 \mid \mathsf{X} = x) + \alpha_1 \cdot p(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 0 \mid \mathsf{X} = x) \\
&\quad + \alpha_2 \cdot p(\mathsf{Y}^{(1)} = 0, \mathsf{Y}^{(2)} = 1 \mid \mathsf{X} = x) + (\alpha_1 + \alpha_2) \cdot p(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 1 \mid \mathsf{X} = x) \\
&= \alpha_1 \cdot [p(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 0 \mid \mathsf{X} = x) + p(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 1 \mid \mathsf{X} = x)] \\
&\quad + \alpha_2 \cdot [p(\mathsf{Y}^{(1)} = 0, \mathsf{Y}^{(2)} = 1 \mid \mathsf{X} = x) + p(\mathsf{Y}^{(1)} = 1, \mathsf{Y}^{(2)} = 1 \mid \mathsf{X} = x)] \\
&= \alpha_1 \eta^{(1)}(x) + \alpha_2 \eta^{(2)}(x)
\end{aligned}$$

This result shows that the optimal scorer for weighted label aggregation is a direct linear combination of the individual class-probability functions $\eta^{(k)}(x)$, weighted by the *explicitly chosen* aggregation weights α_k.

This contrasts sharply with the optimal scorer for loss aggregation (Proposition 5.4), which is $\sum_k \frac{a_k}{\pi^{(k)} \cdot (1 - \pi^{(k)})} \cdot \eta^{(k)}(x)$. In loss aggregation, the effective weight on $\eta^{(k)}(x)$ depends not only on the chosen mixing weight a_k but also implicitly on the label prior $\pi^{(k)}$.

Therefore, regarding sensitivity:

- The label aggregation optimal scorer (under the assumptions above) is sensitive to the choice of weights α_k in a direct and predictable way: the final scorer is exactly the α_k-weighted sum of the $\eta_k(x)$. It is notably insensitive to the class priors $\pi^{(k)}$.
- The loss aggregation optimal scorer is sensitive to both the chosen weights a_k and the class priors $\pi^{(k)}$ through the $\pi^{(k)}(1 - \pi^{(k)})$ term.

E. Further Discussion on Application Scenarios

The problem of synthesizing multiple binary labels into a single coherent ranking has widespread applications beyond the initial examples of information retrieval and medical diagnosis mentioned in §1. Our analysis of loss versus label aggregation, particularly regarding Pareto optimality and the "label dictatorship" issue, can inform choices in various real-world systems. Below, we elaborate on some potential application scenarios:

- **Multi-faceted Information Retrieval:** Users' information needs can be complex, requiring documents to be relevant across multiple dimensions or interpretations. For example, a search system might need to rank documents based on topical relevance to different aspects of a query, alongside signals for document freshness, geographical relevance for location-sensitive queries, or authoritativeness of the source. Each of these aspects can be represented by a (potentially binary) label (e.g., "`is_fresh`," "`is_geo_relevant`"). Effectively combining these signals is crucial for user satisfaction (Perkio et al., 2005).
- **Recommendation Systems:** Modern recommender systems often optimize for multiple objectives simultaneously. For instance, beyond predicting user engagement (e.g., click-through rate, purchase probability), systems may also aim to promote relevance to users' long-term interests, ensure diversity in recommendations to avoid filter bubbles, or uphold fairness considerations across different item providers or content creators (Zheng & Wang, 2022).
- **Computational Advertising:** Ranking advertisements effectively requires balancing predictions from multiple models. Commonly, platforms consider both the predicted click-through rate (CTR) and the predicted conversion rate (CVR) — the likelihood that a click leads to a desired action (e.g., a sale). The final ranking aims to optimize overall platform revenue or advertiser value, which often involves a synthesis of these distinct predictive labels (Wang et al., 2023a).

In each of these scenarios, understanding how different aggregation strategies (loss versus label) behave, as explored in our work, can lead to more principled system design. The choice of aggregation method can influence how trade-offs between objectives are handled and whether certain objectives inadvertently dominate others.

F. Illustrative Example of Undesirable Dictatorship in Loss Aggregation

§5.1 discusses how the Bayes-optimal scorer for loss aggregation (Proposition 5.2) weights individual class-probability functions $\eta^{(k)}(x)$ by $a_k/[\pi^{(k)}(1-\pi^{(k)})]$. This means that even with uniform explicit weights $a_k = 1$, labels that are more marginally skewed (i.e., $\pi^{(k)}$ is far from 0.5) receive higher effective weighting. We clarify that $\pi^{(k)} = \mathbf{P}(\mathsf{Y}^{(k)} = 1)$ is the marginal prevalence of a label, which is distinct from its conditional probability $\eta^{(k)}(x) = \mathbf{P}(\mathsf{Y}^{(k)} = 1 \mid \mathsf{X} = x)$ for a specific instance x. A label can have a balanced marginal prior ($\pi^{(k)} \approx 0.5$) yet be perfectly deterministic (noise-free) conditionally (e.g., $\eta^{(k)}(x) \in \{0, 1\}$).

Proposition 6.1 formalizes the problem whereby one label can become a "dictator", i.e., the ranking overly favors it. Below is an illustrative example from information retrieval where such dictatorship, driven by marginal skewness, would be undesirable.

Consider ranking documents based on two binary, conditionally deterministic (noise-free, $\eta^{(k)}(x) \in \{0, 1\}$) labels:

- $\mathsf{Y}^{(1)}$: "`is_relevant`" (core topical relevance to the user's query).
- $\mathsf{Y}^{(2)}$: "`is_recent`" (document published in the last 24 hours).

Assume the following marginal priors in the document collection for a typical query:

- For "`is_relevant`" ($\mathsf{Y}^{(1)}$): A good number of documents are relevant, and many are not. Let $\pi^{(1)} = 0.4$. Thus, $\pi^{(1)}(1-\pi^{(1)}) = 0.4 \times 0.6 = 0.24$.
- For "`is_recent`" ($\mathsf{Y}^{(2)}$): Very few documents are extremely recent. Let $\pi^{(2)} = 0.01$. Thus, $\pi^{(2)}(1-\pi^{(2)}) = 0.01 \times 0.99 = 0.0099$.

If we use loss aggregation with uniform explicit weights $a_1 = 1$ and $a_2 = 1$, the $\alpha^{(k)}$ terms from Proposition 6.1 become:

- $\alpha^{(1)} = a_1/[\pi^{(1)}(1-\pi^{(1)})] = 1/0.24 \approx 4.17$.
- $\alpha^{(2)} = a_2/[\pi^{(2)}(1-\pi^{(2)})] = 1/0.0099 \approx 101.01$.

Since $\alpha^{(2)} \gg \alpha^{(1)}$, according to Proposition 6.1, the "`is_recent`" label ($\mathsf{Y}^{(2)}$) will dictate the ranking. The system will primarily rank documents based on whether they are extremely recent. An irrelevant but very recent document would likely be ranked above a highly relevant document that is not from the last 24 hours.

This outcome is undesirable if the user's primary goal is to find relevant information, with recency being a secondary, tie-breaking, or less critical factor. The loss aggregation objective, in this case, inadvertently prioritizes the "`is_recent`" signal heavily, not because its conditional signal is necessarily stronger or more important for the user's core task, but because its marginal distribution is highly skewed. The system optimizes for a property of the label's distribution in the dataset (its rarity) rather than strictly adhering to an equal consideration of the (equally clean) conditional signals for relevance and recency. This highlights why relying on marginal priors for weighting can be problematic. Label aggregation (Proposition 5.4, with appropriate costs) would score by $\eta^{(1)}(x) + \eta^{(2)}(x)$, giving equal intrinsic weight to the conditional signals from $\mathsf{Y}^{(1)}$ and $\mathsf{Y}^{(2)}$ regardless of their marginal priors.

G. Empirical details

In this section we provide details to empirical experiments in the paper.

In Table 5 we summarize the prompts used to obtain relevance and clicks labels for the MS MARCO examples shown in Tables 2.

G.1. Synthetic data generation for the exhaustive enumeration of the optimal solutions

We begin with a synthetic experiment with a goal of verifying our theory pertaining to maximizers of the loss aggregation objectives. We generate a synthetic dataset, and compute the optimal solutions via brute force evaluation;

Label	Input prompt
Engagement	Imagine you are an information retrieval expert. Your aim is to judge whether the following document will be clicked if it is shown for the following query. Query: {query} Document: {document} Do you predict this document will be clicked for this query? Output only yes or no and nothing else.
Relevance	Imagine you are an information retrieval expert. Your aim is to judge whether the following document will be considered relevant by a normal user if it is shown for the following query. Query: {query} Document: {document} Do you predict this document will be considered relevant by a normal user for this query? Output only yes or no and nothing else.

Table 5. Prompts used for predicting engagement and relevance of different documents per query in MS MARCO dataset (Bajaj et al., 2018). We select documents according to the original order from the MS MARCO dataset.

We consider a two dimensional Gaussian distribution with zero mean and a uniformly sampled covariance matrix Σ. We sample N two-dimensional samples and consider the first dimension to correspond to $y_1(x)$ (clicks) and the second dimension to $y_2(x)$ (relevance). We make the two scores bi-level by thresholding the numbers at 0.

We aim to consider an exhaustive set of hypotheses such that it is possible to conduct a brute force evaluation of different objectives. To this end, we consider the following hypothesis class f. We consider the image of f to be $[P] \doteq \{1, 2, \ldots, P\}$. For each sample x and the corresponding relevance scores $y_1(x)$ and $y_2(x)$, if both relevance scores agree, we set $f(x)$ to correspond to $P \cdot y_1(x)$ (i.e., the maximum value if both scores equal 1, and 0 if both scores equal 0). For all remaining x (where $y_1(x) \neq y_2(x)$), we consider all assignments of f from $[P]$. This way, if $y_1(x) \neq y_2(x)$ on M samples, the hypothesis class $\hat{\mathcal{Y}}$ for f consists of P^M hypotheses.

We enumerate different values for the α_1 and α_2 weights in the loss aggregation objective from a cartesian product $[5] \times [5]$, and for each pair of values, we exhaustively enumerate all $f \in \hat{\mathcal{Y}}$ and calculate: the loss aggregation objective $\sum_{k \in [K]} a_k \operatorname{AUC}(f; y_k)$ and both of the per scoring function objectives $\operatorname{AUC}(f; y_1)$ and $\operatorname{AUC}(f; y_2)$. We then do the same for label aggregation and label product objectives.

In the end, we plot the maximizers of loss and label aggregation in Figure 5, and confirm the following characterizations of the optimal solutions:

(a) $\mathcal{Y}^*_{\mathrm{LP}}$ consists of all assignments of scores from $\{0, 1, 2\}$ to the disagreeing examples (they can disagree in any way, and only need to be smaller than 3).

(b) $\mathcal{Y}^*_{\mathrm{LoA},<}$ consists of only a single assignment to f to the disagreeing examples such that: f for examples with $y_1(x) > y_2(x)$ is 1 and otherwise it is 2.

(c) $\mathcal{Y}^*_{\mathrm{LoA},>}$ consists of only a single assignment to f to the disagreeing examples such that: f for examples with $y_1(x) > y_2(x)$ is 2 and otherwise it is 1.

(d) $\mathcal{Y}^*_{\mathrm{LoA},=}$ consists of all assignments to f to the disagreeing examples such that any relation is satisfied between the two sets of disagreeing examples. This case is equivalent to LA.

G.2. Details on AUC optimization

The indicator function in the AUC definition (1) makes it non-differentiable, and thus, direct AUC optimization is challenging. To mitigate this, previous works propose relaxation of the indicator function with different surrogate functions, including the hinge loss function, the logistic loss function and other choices (Sun et al., 2023; Tang et al., 2022). Thus, for a surrogate function ϕ, we arrive at the following formulation:

$$\operatorname{AUC}^{\phi}(f; D) \,=\, \mathbb{E}_{\mathsf{X}^+}\mathbb{E}_{\mathsf{X}^-}\Big[\phi(f(\mathsf{X}^+) - f(\mathsf{X}^-))\Big]$$

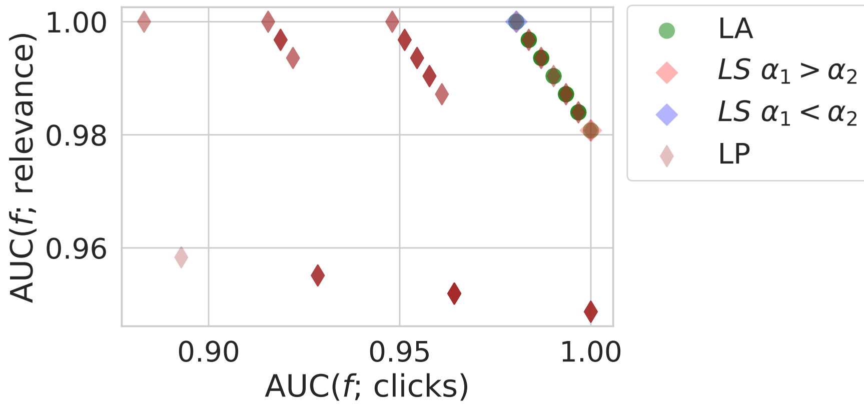


Figure 5. AUC metrics for Bayes-optimal solutions according to different objectives. We confirm the following relations among the per-objective optimal solutions sets: $\mathcal{Y}^*_{\mathrm{LoA},>} \subset \mathcal{Y}^*_{\mathrm{LoA},=} = \mathcal{Y}^*_{\mathrm{LaA}} \subset \mathcal{Y}^*_{\mathrm{LP}}$ and $\mathcal{Y}^*_{\mathrm{LoA},<} \subset \mathcal{Y}^*_{\mathrm{LoA},=} = \mathcal{Y}^*_{\mathrm{LaA}} \subset \mathcal{Y}^*_{\mathrm{LP}}$. Notice the Pareto front is given by a linear function spanned by the solutions corresponding to $\mathcal{Y}^*_{\mathrm{LoA},<}$ and $\mathcal{Y}^*_{\mathrm{LoA},>}$.

Objective	$\pi^{(2)} = 0.5$	$\pi^{(2)} = 0.6$	$\pi^{(2)} = 0.7$	$\pi^{(2)} = 0.8$	$\pi^{(2)} = 0.9$
AUC(y_1)	0.53	0.44	0.52	0.56	0.50
AUC(y_2)	0.49	0.46	0.50	0.54	0.45
$\mathrm{AUC}_{\mathrm{LaA}}$	0.06	0.04	0.08	0.07	0.07
$\mathrm{AUC}^{(1,1)}_{\mathrm{LoA}}$	0.05	0.13	0.09	0.11	0.10

Table 6. $|\mathrm{AUC}(y_1) - \mathrm{AUC}(y_2)|$ on the synthetic dataset from MLP model training. Each column value $(0.5, \ldots, 0.9)$ corresponds to the marginal probability for the second label, $\mathbf{P}(y_2 = 1)$, while the marginal probability for the first label is fixed, $\mathbf{P}(y_1 = 1) = 0.5$. We find label aggregation yields the lowest difference between per-label AUC metrics. Here, $\mathrm{AUC}_{\mathrm{LaA}}$ = AUC(y_1 + y_2) and $\mathrm{AUC}^{(1,1)}_{\mathrm{LoA}}$ = AUC(y_1) + AUC(y_2).

In our work, unless otherwise stated, we assume a logistic surrogate function ϕ.

G.3. Details on synthetic training experiments

We draw uniform two-dimensional random vectors w_1, w_2 over $[-1, 1]^2$ and calculate label distributions using a sigmoid linear model, with $\eta^{(1)}(x) = \sigma(\alpha \cdot w_1^\top x)$ and $\eta^{(2)}(x) = \sigma(\tau \cdot w_2^\top x)$, where $\sigma(z) = \frac{1}{1+\exp(-z)}$ is the sigmoid function. We then sample labels $y_1(x)$ and $y_2(x)$ uniformly from the distributions $\eta^{(1)}(x)$ and $\eta^{(2)}(x)$.

We train a 3-layer MLP model with hidden dimension 256 and ReLU activation over 8K examples for 50 epochs. We evaluate on held out 2K examples.

G.4. Additional results

For completeness, in Figure 6 we report metrics for per-signal AUC corresponding to results in Figure 2.

G.5. Additional synthetic experiments

We generate instances $x \in \mathbb{R}^2$ from a uniform distribution over $[-1, 1]^2$. We next draw uniform random vectors μ_1, μ_2, μ_3, μ_4 over $[0, 1]^2$ controlling the skewness label distributions $\eta^{(1)}(x)$ and $\eta^{(2)}(x)$ (see Appendix G.3 for details of the data generation process). Depending on the sampled values, we end up with different marginal skewness of the distributions. We fix the marginal $\pi^{(1)} = 0.5$ and vary $\pi^{(2)}$ to control the effect label skewness on the performance of different techniques. We train a multi-layer perceptron model with varying level of skewness for the second label, while keeping the first label balanced. We report the results in Table 6. We find that the difference between the per-label AUC is the smallest for label aggregation when the distribution for the second label is imbalanced. Thus, label aggregation better balances between the two objectives, whereas loss aggregation tends to favor one over the other.

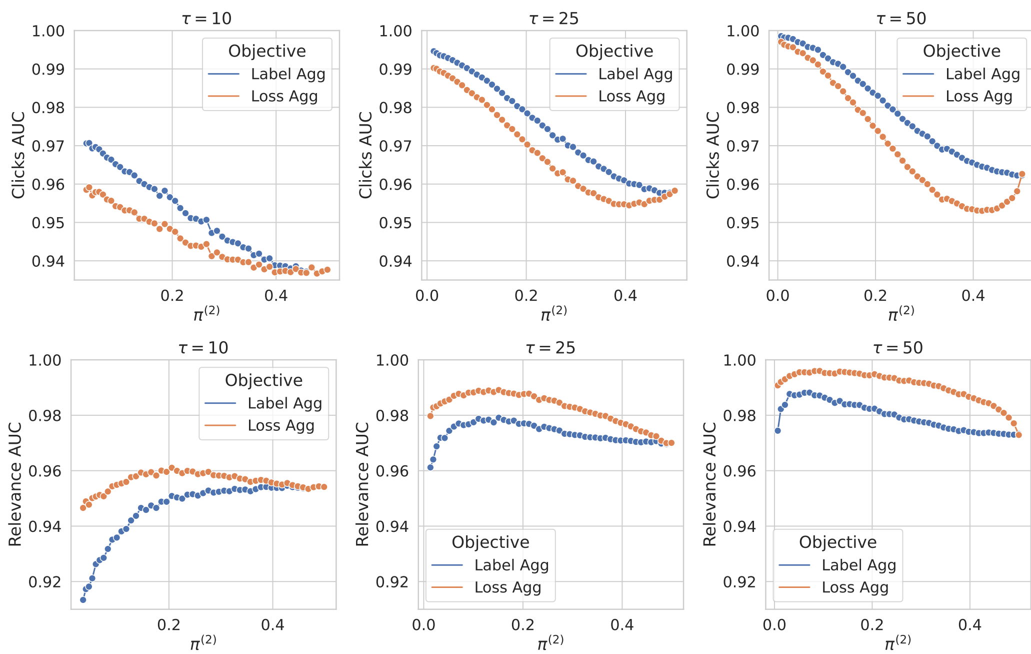


Figure 6. Plots of $\text{AUC}(\cdot; D^{(1)})$ (i.e., clicks) and $\text{AUC}(\cdot; D^{(2)})$ (i.e., relevance) for optimal scorers as a function of skewness in label $\mathsf{Y}^{(2)}$, corresponding to to results in Figure 2. We compare label aggregation and loss aggregation on a synthetic dataset. We fix $\pi^{(1)} = \mathbf{P}(\mathsf{Y}^{(1)} = 1) = 0.5$ and vary $\pi^{(2)} = \mathbf{P}(\mathsf{Y}^{(2)} = 1)$. The sigmoid scaling parameter τ controls how close the label distribution is to a deterministic distribution (the larger τ, the closer it is to a deterministic distribution). We find loss aggregation to lead to a higher difference between per-label AUC metrics.