Ranking with Abstention

Anqi Mao [1] **Mehryar Mohri** [2 1] **Yutao Zhong** [1]

Abstract

We introduce a novel framework of *ranking with abstention*, where the learner can abstain from making prediction at some limited cost c. We present a extensive theoretical analysis of this framework including a series of *$\mathcal{H}$-consistency bounds* for both the family of linear functions and that of neural networks with one hidden-layer. These theoretical guarantees are the state-of-the-art consistency guarantees in the literature, which are upper bounds on the target loss estimation error of a predictor in a hypothesis set $\mathcal{H}$, expressed in terms of the surrogate loss estimation error of that predictor. We further argue that our proposed abstention methods are important when using common equicontinuous hypothesis sets in practice. We report the results of experiments illustrating the effectiveness of ranking with abstention.

1. Introduction

In many applications, ranking is a more appropriate formulation of the learning task than classification, given the crucial significance of the ordering of the items. As an example, for movie recommendation systems, an ordered list of movies is preferable to a comprehensive list of recommended titles, since users are more likely to watch those ranked highest.

The problem of learning to rank has been studied in a large number of publications. Ailon & Mohri (2008; 2010) distinguish two general formulations of the problem: the *score-based setting* and the *preference-based setting*. In the score-based setting, a real-valued function over the input space is learned, whose values determine a total ordering of all input points. In the preference-based setting, a pairwise preference function is first learned, typically by training a classifier over a sample of labeled pairs; next, that function is used to derive an ordering, potentially randomized, of any subset of points.

This paper deals with the score-based ranking formulation both in the general ranking setting, where items are not assigned any specific category, and the bipartite setting, where they are labeled with one of two classes. The evaluation of a ranking solution in this context is based on the average pairwise misranking metric. In the bipartite setting, this metric is directly related to the AUC (Area Under the ROC Curve), which coincides with the average correct pairwise ranking (Hanley & McNeil, 1982; Cortes & Mohri, 2003), also known as the Wilcoxon-Mann-Whitney statistic.

For most hypothesis sets, directly optimizing the pairwise misranking loss is intractable. Instead, ranking algorithms resort to a surrogate loss. As an example, the surrogate loss for RankBoost (Freund et al., 2003; Rudin et al., 2005) is based on the exponential function and that of SVM ranking (Joachims, 2002) on the hinge loss. But, what guarantees can we rely on when minimizing a surrogate loss instead of the original pairwise misranking loss?

The property often invoked in this context is *Bayes consistency*, which has been extensively studied for classification (Zhang, 2004; Bartlett et al., 2006; Tewari & Bartlett, 2007). The Bayes consistency of ranking surrogate losses has been studied in the special case of bipartite ranking: in particular, Uematsu & Lee (2017) proved the inconsistency of the pairwise ranking loss based on the hinge loss and Gao & Zhou (2015) gave excess loss bounds for pairwise ranking losses based on the exponential or the logistic loss (see also (Menon & Williamson, 2014)). A related but distinct consistency question has been studied in several publications (Agarwal et al., 2005; Kotlowski et al., 2011; Agarwal, 2014). It is one with respect to binary classification, that is whether a near minimizer of the surrogate loss of the binary classification loss is a near minimizer of the bipartite misranking loss (Cortes & Mohri, 2003).

However, as recently argued by Awasthi, Mao, Mohri, and Zhong (2022a), Bayes consistency is not a sufficiently informative notion since it only applies to the entire class of measurable functions and does not hold for specific subsets, such as sub-families of linear functions or neural networks. Furthermore, Bayes consistency is solely an asymptotic con-

[1]Courant Institute of Mathematical Sciences, New York, NY; [2]Google Research, New York, NY. Correspondence to: Anqi Mao <aqmao@cims.nyu.edu>, Mehryar Mohri <mohri@google.com>, Yutao Zhong <yutao@cims.nyu.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

cept and does not offer insights into the performance of predictors trained on finite samples. In response, the authors proposed an alternative concept called $\mathcal{H}$*-consistency bounds*, which provide non-asymptotic guarantees tailored to a given hypothesis set $\mathcal{H}$. They proceeded to establish such bounds within the context of classification both in binary and multi-class classification (Awasthi et al., 2022a;b), see also (Mao et al., 2023b;a; Zheng et al., 2023). These are stronger and more informative guarantees than Bayes consistency.

But, can we derive $\mathcal{H}$-consistency bounds guarantees for ranking? We propose a novel framework of *ranking with abstention*, where the learner can abstain from making prediction at some limited cost c, in both the general pairwise ranking scenario and the bipartite ranking scenarios. For surrogate losses of these abstention loss functions, we give a series of $\mathcal{H}$-consistency bounds for both the family of linear functions and that of neural networks with one hidden-layer. A key term appearing in these bounds is the *minimizability gap*, which measures the difference between the best-in-class expected loss and the expected infimum of the pointwise expected loss. This plays a crucial role in these bounds and we give a detailed analysis of these terms.

We will further show that, without abstention, deriving non-trivial $\mathcal{H}$-consistency bounds is not possible for most hypothesis sets used in practice, including the family of constrained linear models or that of the constrained neural networks, or any family of equicontinuous functions with respect to the input. In fact, we will give a relatively simple example where the pairwise misranking error of the RankBoost algorithm remains significant, even after training with relatively large sample sizes. These results further imply the importance of our proposed abstention methods.

We also present the results of experiments illustrating the effectiveness of ranking with abstention.

Technical novelty. The primary technical differences and challenges between the ranking and classification settings (Awasthi et al., 2022a) stem from the fundamental distinction that ranking loss functions take as argument a pair of samples rather than a single one, as is the case for binary classification loss functions. This makes it more challenging to derive $\mathcal{H}$-consistency bounds, as upper bounding the calibration gap of the target loss by that of the surrogate loss becomes technically more difficult.

Additionally, this fundamental difference leads to a negative result for ranking, as $\mathcal{H}$-consistency bounds cannot be guaranteed for most commonly used hypothesis sets, including the family of constrained linear models and that of constrained neural networks, both of which satisfy the equicontinuity property concerning the input. As a result, a natural alternative involves using ranking with abstention, for which $\mathcal{H}$-consistency bounds can be proven. In the abstention setting, an extra challenge lies in carefully monitoring the effect of a threshold γ to relate the calibration gap of the target loss to that of the surrogate loss.

Furthermore, the bipartite ranking setting introduces an added layer of complexity, as each element of a pair of samples has an independent conditional distribution, which results in a more intricate calibration gap.

Structure of the paper. The remaining sections of this paper are organized as follows. In Section 2, we study general pairwise ranking with abstention. We provide a series of explicit $\mathcal{H}$-consistency bounds in the case of the pairwise abstention loss, with multiple choices of the surrogate loss and for both the family of linear functions and that of neural networks with one hidden-layer. We also study bipartite ranking with abstention in Section 3. Here too, we present $\mathcal{H}$-consistency bounds for bipartite abstention loss, for linear hypothesis sets and the family of neural networks with one hidden-layer. In Section 4, we show the importance of our abstention methods by demonstrating that without abstention, there exists no meaningful $\mathcal{H}$-consistency bound for general surrogate loss functions with an equicontinuous hypothesis set $\mathcal{H}$, in both the general pairwise ranking (Section 4.1) and the bipartite ranking (Section 4.2) scenarios. In Section 5, we report the results of experiments illustrating the effectiveness of ranking with abstention.

We give a detailed discussion of related work in Appendix A.

2. General Pairwise Ranking with Abstention

In this section, we introduce a novel framework of *general pairwise ranking with abstention*. We begin by introducing the necessary definitions and concepts.

2.1. Preliminaries

We study the learning scenario of score-based ranking in the *general pairwise ranking* scenario (e.g. see (Mohri et al., 2018)). Let $\mathcal{X}$ denote the input space and $\mathcal{Y} = \{-1, +1\}$ the label space. We denote by $\mathcal{H}$ a hypothesis set of functions mapping from $\mathcal{X}$ to $\mathbb{R}$. The *general pairwise misranking loss* L_{0-1} is defined for all h in $\mathcal{H}$, x, x' in $\mathcal{X}$ and y in $\mathcal{Y}$ by

$$\mathsf{L}_{0-1}(h, x, x', y) = \mathbb{1}_{y \neq \operatorname{sign}(h(x') - h(x))}, \tag{1}$$

where $\operatorname{sign}(u) = \mathbb{1}_{u \geq 0} - \mathbb{1}_{u < 0}$. Thus, h incurs a loss of one on the labeled pair (x, x', y) when it ranks the pair (x, x') opposite to the sign of y, where, by convention, x' is considered as ranked above x when $h(x') \geq h(x)$. Otherwise, the loss incurred is zero.

The framework we propose is that of *general pairwise ranking with abstention*. In this framework, the learner abstains from making a prediction on input pair (x, x') if the distance between x' and x is relatively small, in which case a cost c is incurred. Let $\|\cdot\|$ denote the norm adopted, which is typically an ℓ_p-norm, $p \in [1, +\infty]$. The *pairwise abstention loss* is defined as follows for any $h \in \mathcal{H}$ and $(x, x', y) \in \mathcal{X}\times\mathcal{X}\times\mathcal{Y}$:

$$\mathsf{L}^{\text{abs}}_{0-1}(h, x, x', y) = \mathbb{1}_{y \neq \text{sign}(h(x') - h(x))} \mathbb{1}_{\|x - x'\| > \gamma} + c\,\mathbb{1}_{\|x - x'\| \leq \gamma}, \quad (2)$$

where γ is a given threshold value. For $\gamma = 0$, $\mathsf{L}^{\text{abs}}_{0-1}$ reduces to the pairwise misranking loss L_{0-1} without abstention.

In Section 4, we will show the importance of our proposed abstention methods when using common equicontinuous hypothesis sets in practice. Optimizing the pairwise misranking loss L_{0-1} or pairwise abstention loss $\mathsf{L}^{\text{abs}}_{0-1}$ is intractable for most hypothesis sets. Thus, general ranking algorithms rely on a surrogate loss function L instead of L_{0-1}. The general pairwise ranking surrogate losses widely used in practice admit the following form:

$$\mathsf{L}_\Phi(h, x, x', y) = \Phi\big(y(h(x') - h(x))\big), \quad (3)$$

where Φ is a non-increasing function that is continuous at 0 and upper bounding $u \mapsto \mathbb{1}_{u \leq 0}$ over $\mathbb{R}$. We will analyze the properties of such surrogate loss functions with respect to both L_{0-1} and $\mathsf{L}^{\text{abs}}_{0-1}$. We will specifically consider the hinge loss $\Phi_{\text{hinge}}(t) = \max\{0, 1 - t\}$, the exponential loss $\Phi_{\exp}(t) = e^{-t}$ and the sigmoid loss $\Phi_{\text{sig}}(t) = 1 - \tanh(kt)$, $k > 0$ as auxiliary functions Φ.

Let $\mathcal{D}$ denote a distribution over $\mathcal{X} \times \mathcal{X} \times \mathcal{Y}$. We denote by $\eta(x, x') = \mathcal{D}(Y = +1 \mid (X, X') = (x, x'))$ the conditional probability of $Y = +1$ given $(X, X') = (x, x')$. We also denote by $\mathcal{R}_{\mathsf{L}}(h)$ the *expected* L*-loss* of a hypothesis h and by $\mathcal{R}^*_{\mathsf{L}}(\mathcal{H})$ its infimum over $\mathcal{H}$:

$$\mathcal{R}_{\mathsf{L}}(h) = \mathop{\mathbb{E}}_{(x, x', y) \sim \mathcal{D}}[\mathsf{L}(h, x, x', y)] \quad \mathcal{R}^*_{\mathsf{L}}(\mathcal{H}) = \inf_{h \in \mathcal{H}} \mathcal{R}_{\mathsf{L}}(h)$$

$\mathcal{H}$-consistency bounds. We will analyze the $\mathcal{H}$*-consistency bounds* properties (Awasthi et al., 2022a) of such surrogate loss functions. An $\mathcal{H}$-consistency bound for a surrogate loss L and a target loss $\overline{\mathsf{L}}$ is a guarantee of the form:

$$\forall h \in \mathcal{H}, \quad \mathcal{R}_{\overline{\mathsf{L}}}(h) - \mathcal{R}^*_{\overline{\mathsf{L}}}(\mathcal{H}) \leq f(\mathcal{R}_{\mathsf{L}}(h) - \mathcal{R}^*_{\mathsf{L}}(\mathcal{H})),$$

for some non-decreasing function $f\colon \mathbb{R}_+ \to \mathbb{R}_+$, where $\overline{\mathsf{L}}$ can be taken as L_{0-1} or $\mathsf{L}^{\text{abs}}_{0-1}$. This provides a quantitative relationship between the estimation loss of $\overline{\mathsf{L}}$ and that of the surrogate loss L. The guarantee is stronger and more informative than Bayes consistency, or $\mathcal{H}$-consistency, $\mathcal{H}$-calibration or the excess error bounds (Zhang, 2004; Bartlett et al., 2006; Steinwart, 2007; Mohri et al., 2018) discussed in the literature.

A key quantity appearing in $\mathcal{H}$-consistency bounds is the *minimizability gap*, which is the difference between the best-in-class expected loss and the expected pointwise infimum of the loss:

$$\mathcal{M}_{\mathsf{L}}(\mathcal{H}) = \mathcal{R}^*_{\mathsf{L}}(\mathcal{H}) - \mathop{\mathbb{E}}_{(x, x')}\Big[\inf_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{y}[\mathsf{L}(h, x, x', y) \mid (x, x')]\Big].$$

By the super-additivity of the infimum, the minimizability gap is always non-negative.

We will specifically study the hypothesis set of linear hypotheses, $\mathcal{H}_{\text{lin}} = \{x \mapsto w \cdot x + b \mid \|w\|_q \leq W, |b| \leq B\}$ and the hypothesis set of one-hidden-layer ReLU networks: $\mathcal{H}_{\text{NN}} = \{x \mapsto \sum_{j=1}^n u_j (w_j \cdot x + b_j)_+ \mid \|u\|_1 \leq \Lambda, \|w_j\|_q \leq W, |b_j| \leq B\}$, where $(\cdot)_+ = \max(\cdot, 0)$.

Let $p, q \in [1, +\infty]$ be conjugate numbers, that is $\frac{1}{p} + \frac{1}{q} = 1$. Without loss of generality, we consider $\mathcal{X} = B^d_p(1)$ and $\|\cdot\|$ in (2) to be the ℓ_p norm. The corresponding conjugate ℓ_q norm is adopted in the hypothesis sets $\mathcal{H}_{\text{lin}}$ and $\mathcal{H}_{\text{NN}}$. In the following, we will prove $\mathcal{H}$-consistency bounds for $\mathsf{L} = \mathsf{L}_\Phi$ and $\overline{\mathsf{L}} = \mathsf{L}^{\text{abs}}_{0-1}$ when using as an auxiliary function Φ the hinge loss, the exponential loss, or the sigmoid loss, in the case of the linear hypothesis set $\mathcal{H}_{\text{lin}}$ or that of one-hidden-layer ReLU networks $\mathcal{H}_{\text{NN}}$.

2.2. $\mathcal{H}$-consistency bounds for pairwise abstention loss

Theorem 2.1 shows the $\mathcal{H}$-consistency bounds for L_Φ with respect to $\mathsf{L}^{\text{abs}}_{0-1}$ when using common auxiliary functions. The bounds in Theorem 2.1 depend directly on the threshold value γ, the parameter W in the linear models and parameters of the loss function (e.g., k in sigmoid loss). Different from the bounds in the linear case, all the bounds for one-hidden-layer ReLU networks not only depend on W, but also depend on Λ, which is a parameter appearing in $\mathcal{H}_{\text{NN}}$.

Theorem 2.1 (**$\mathcal{H}$-consistency bounds for pairwise abstention loss**). *Let $\mathcal{H}$ be $\mathcal{H}_{\text{lin}}$ or $\mathcal{H}_{\text{NN}}$. Then, for any $h \in \mathcal{H}$ and any distribution,*

$$\mathcal{R}_{\mathsf{L}^{\text{abs}}_{0-1}}(h) - \mathcal{R}^*_{\mathsf{L}^{\text{abs}}_{0-1}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}^{\text{abs}}_{0-1}}(\mathcal{H}) \leq \Gamma_\Phi\big(\mathcal{R}_{\mathsf{L}_\Phi}(h) - \mathcal{R}^*_{\mathsf{L}_\Phi}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_\Phi}(\mathcal{H})\big),$$

where $\Gamma_\Phi(t) = \frac{t}{\min\{W\gamma, 1\}}$, $\max\Big\{\sqrt{2t}, 2\Big(\frac{e^{2W\gamma}+1}{e^{2W\gamma}-1}\Big)t\Big\}$ and $\frac{t}{\tanh(kW\gamma)}$ for $\Phi = \Phi_{\text{hinge}}$, $\Phi_{\exp}$ and Φ_{sig} respectively. W is replaced by ΛW for $\mathcal{H} = \mathcal{H}_{\text{NN}}$.

As an example, for $\mathcal{H} = \mathcal{H}_{\text{lin}}$ or $\mathcal{H}_{\text{NN}}$, when using as Φ the exponential loss function, modulo the minimizability gaps (which are zero when the best-in-class error coincides with the Bayes error or can be small in some other cases), the bound implies that if the surrogate estimation loss $\mathcal{R}_{\mathsf{L}_{\Phi_{\exp}}}(h) - \mathcal{R}^*_{\mathsf{L}_{\Phi_{\exp}}}(\mathcal{H})$ is reduced to ϵ, then, the target estimation loss $\mathcal{R}_{\mathsf{L}^{\text{abs}}_{0-1}}(h) - \mathcal{R}^*_{\mathsf{L}^{\text{abs}}_{0-1}}(\mathcal{H})$ is upper bounded by $\Gamma_{\Phi_{\exp}}(\epsilon)$. For sufficiently small values of ϵ, the dependence of $\Gamma_{\Phi_{\exp}}$ on ϵ exhibits a square root relationship. However, if this is not the case, the dependence becomes linear, subject to a constant factor depending on the threshold value γ, the parameter W in the linear models and the one-hidden-layer ReLU networks, and an additional parameter Λ in the one-hidden-layer ReLU networks.

The proofs consist of analyzing calibration gaps of the target loss and that of each surrogate loss and seeking a tight lower bound of the surrogate calibration gap in terms of the target one. As an example, for $\Phi = \Phi_{\exp}$, we have the tight lower bound $\Delta\mathcal{C}_{\mathsf{L}_{\Phi_{\exp}},\mathcal{H}}(h,x,x') \geq \Delta\mathcal{C}_{\mathsf{L}_{\Phi_{\exp}},\mathcal{H}}(h_0,x,x') = \Psi_{\exp}\Big(\Delta\mathcal{C}_{\mathsf{L}^{\text{abs}}_{0-1},\mathcal{H}}(h,x,x')\Big)$, where h_0 can be the null hypothesis when $\Delta\mathcal{C}_{\mathsf{L}^{\text{abs}}_{0-1},\mathcal{H}}(h,x,x') \neq 0$ and $\Psi_{\exp}$ is an increasing and piecewise convex function on $[0,1]$ defined by $\Psi_{\exp}(t) = \begin{cases} 1-\sqrt{1-t^2}, & t \leq \frac{e^{2W\gamma}-1}{e^{2W\gamma}+1} \\ 1-\frac{t+1}{2}e^{-W\gamma}-\frac{1-t}{2}e^{W\gamma}, & t > \frac{e^{2W\gamma}-1}{e^{2W\gamma}+1} \end{cases}$, where W is replaced by ΛW for $\mathcal{H} = \mathcal{H}_{\text{NN}}$. The detailed proofs and the expression of the corresponding minimizability gaps are included in Appendix C.

3. Bipartite Ranking with Abstention

As with the general pairwise ranking case, we introduce a novel framework of *bipartite ranking with abstention*. We first introduce the relevant definitions and concepts.

3.1. Preliminaries

In the bipartite setting, each point x admits a label $y \in \{-1,+1\}$. The *bipartite misranking loss* $\widetilde{\mathsf{L}}_{0-1}$ is defined for all h in $\mathcal{H}$, and $(x,y),(x',y')$ in $(\mathcal{X}\times\mathcal{Y})$ by

$$\widetilde{\mathsf{L}}_{0-1}(h,x,x',y,y') = \mathbb{1}_{(y-y')(h(x)-h(x'))<0} + \frac{1}{2}\mathbb{1}_{(h(x)=h(x'))\wedge(y\neq y')}. \quad (4)$$

The framework we propose is that of *bipartite ranking with abstention*. In this framework, the learner can abstain from making prediction on a pair (x,x') with x and x' relatively close. The *bipartite abstention loss* is defined as follows for any $h\in\mathcal{H}$ and $(x,y),(x',y')\in\mathcal{X}\times\mathcal{Y}$:

$$\widetilde{\mathsf{L}}^{\text{abs}}_{0-1}(h,x,x',y,y') = \widetilde{\mathsf{L}}_{0-1}(h,x,x',y,y')\mathbb{1}_{\|x-x'\|>\gamma} + c\,\mathbb{1}_{\|x-x'\|\leq\gamma}, \quad (5)$$

where γ is a given threshold value. When $\gamma = 0$, $\widetilde{\mathsf{L}}^{\text{abs}}_{0-1}$ reduces to bipartite misranking loss $\widetilde{\mathsf{L}}_{0-1}$ without abstention.

Optimizing the bipartite misranking loss $\widetilde{\mathsf{L}}_{0-1}$ or bipartite abstention loss $\widetilde{\mathsf{L}}^{\text{abs}}_{0-1}$ is intractable for most hypothesis sets and bipartite ranking algorithms rely instead on a surrogate loss $\widetilde{\mathsf{L}}$. The bipartite ranking surrogate losses widely used in practice, admit the following form:

$$\widetilde{\mathsf{L}}_{\Phi}(h,x,x',y,y') = \Phi\left(\frac{(y-y')(h(x)-h(x'))}{2}\right)\mathbb{1}_{y\neq y'}, \quad (6)$$

where Φ is a non-increasing function that is continuous at 0 upper bounding $u \mapsto \mathbb{1}_{u\leq 0}$ over $\mathbb{R}$. We will analyze the $\mathcal{H}$-*consistency bounds* properties (Awasthi et al., 2022a) of such surrogate loss functions with respect to both $\widetilde{\mathsf{L}}_{0-1}$ and $\widetilde{\mathsf{L}}^{\text{abs}}_{0-1}$. As with the general pairwise ranking case, we will specifically consider the hinge loss $\Phi_{\text{hinge}}(t) = \max\{0,1-t\}$, the exponential loss $\Phi_{\exp}(t) = e^{-t}$ and the sigmoid loss $\Phi_{\text{sig}}(t) = 1-\tanh(kt)$, $k>0$ as auxiliary functions Φ.

Let $\mathcal{D}$ be a distribution over $\mathcal{X}\times\mathcal{Y}$. We denote by $\eta(x) = \mathcal{D}(Y=+1 \mid X=x)$ the conditional probability of $Y=+1$ given $X=x$. We will use a definition and notation for the expected $\widetilde{\mathsf{L}}$-loss of $h\in\mathcal{H}$, its infimum, and the minimizability gaps similar to what we used in the general pairwise misranking setting:

$$\mathcal{R}_{\widetilde{\mathsf{L}}}(h) = \mathop{\mathbb{E}}_{(x,x',y)\sim\mathcal{D}}\left[\widetilde{\mathsf{L}}(h,x,x',y)\right] \quad \mathcal{R}^*_{\widetilde{\mathsf{L}}}(\mathcal{H}) = \inf_{h\in\mathcal{H}}\mathcal{R}_{\widetilde{\mathsf{L}}}(h)$$

$$\mathcal{M}_{\widetilde{\mathsf{L}}}(\mathcal{H}) = \mathcal{R}^*_{\widetilde{\mathsf{L}}}(\mathcal{H}) - \mathop{\mathbb{E}}_{(x,x')}\left[\inf_{h\in\mathcal{H}}\mathop{\mathbb{E}}_{(y,y')}\left[\mathsf{L}(h,x,x',y,y') \mid (x,x')\right]\right].$$

3.2. $\mathcal{H}$-consistency bounds for bipartite abstention losses

Theorem 3.1 presents a series of $\mathcal{H}$-consistency bounds for $\widetilde{\mathsf{L}}_{\Phi}$ when using as an auxiliary function Φ the hinge loss, the exponential loss, or the sigmoid loss. The bounds in Theorem 3.1 depend directly on the threshold value γ, the parameter W in the linear models and parameters of the loss function (e.g., k in sigmoid loss). Different from the bounds in the linear case, all the bounds for one-hidden-layer ReLU networks not only depend on W, but also depend on Λ, a parameter in $\mathcal{H}_{\text{NN}}$.

Theorem 3.1 (**$\mathcal{H}$-consistency bounds for bipartite abstention losses**). *Let $\mathcal{H}$ be $\mathcal{H}_{\text{lin}}$ or $\mathcal{H}_{\text{NN}}$. Then, for any $h\in\mathcal{H}$ and any distribution,*

$$\mathcal{R}_{\widetilde{\mathsf{L}}^{\text{abs}}_{0-1}}(h) - \mathcal{R}^*_{\widetilde{\mathsf{L}}^{\text{abs}}_{0-1}}(\mathcal{H}) + \mathcal{M}_{\widetilde{\mathsf{L}}^{\text{abs}}_{0-1}}(\mathcal{H}) \leq \Gamma_{\Phi}\Big(\mathcal{R}_{\widetilde{\mathsf{L}}_{\Phi}}(h) - \mathcal{R}^*_{\widetilde{\mathsf{L}}_{\Phi}}(\mathcal{H}) + \mathcal{M}_{\widetilde{\mathsf{L}}_{\Phi}}(\mathcal{H})\Big)$$

where $\Gamma_{\Phi}(t)$ equals $\frac{t}{\min\{W\gamma,1\}}$, $\max\Big\{\sqrt{t},\Big(\frac{e^{2W\gamma}+1}{e^{2W\gamma}-1}\Big)t\Big\}$ and $\frac{t}{\tanh(kW\gamma)}$ for Φ equals Φ_{hinge}, $\Phi_{\exp}$ and Φ_{sig} respectively. W is replaced by ΛW for $\mathcal{H} = \mathcal{H}_{\text{NN}}$.

As an example, for $\mathcal{H} = \mathcal{H}_{\text{lin}}$ or $\mathcal{H}_{\text{NN}}$, when adopting the exponential loss function as Φ, modulo the minimizability gaps (which are zero when the best-in-class error coincides with the Bayes error or can be small in some other cases), the bound implies that if the surrogate estimation loss $\mathcal{R}_{\widetilde{\mathsf{L}}_{\Phi_{\exp}}}(h) - \mathcal{R}^*_{\widetilde{\mathsf{L}}_{\Phi_{\exp}}}(\mathcal{H})$ is reduced to ϵ, then, the target estimation loss $\mathcal{R}_{\widetilde{\mathsf{L}}^{\text{abs}}_{0-1}}(h) - \mathcal{R}^*_{\widetilde{\mathsf{L}}^{\text{abs}}_{0-1}}(\mathcal{H})$ is upper bounded by

Table 1: General pairwise abstention loss for the Rankboost loss on CIFAR-10; mean ± standard deviation over three runs for various γ and cost c.

γ	0	0.3	0.5	0.7	0.9
Cost 0.1	8.33% ± 0.15%	8.33% ± 0.15%	8.33% ± 0.15%	8.25% ± 0.07%	8.54%± 0.07%
Cost 0.3	8.33% ± 0.15%	8.33% ± 0.15%	8.35% ± 0.15%	9.73% ± 0.11%	20.41%± 0.06%
Cost 0.5	8.33% ± 0.15%	8.33% ± 0.15%	8.36% ± 0.14%	11.20% ± 0.14%	32.28% ± 0.07%

$\Gamma_{\Phi_{\rm exp}}(\epsilon)$. For sufficiently small values of ϵ, the dependence of $\Gamma_{\Phi_{\rm exp}}$ on ϵ exhibits a square root relationship. However, if this is not the case, the dependence becomes linear, subject to a constant factor depending on the threshold value γ, the parameter W in the linear models and the one-hidden-layer ReLU networks, and an additional parameter Λ in the one-hidden-layer ReLU networks.

As with the general pairwise ranking setting, the proofs consist of analyzing calibration gaps of the target loss and that of each surrogate loss and seeking a tight lower bound of the surrogate calibration gap in terms of the target one. Additionally, the bipartite ranking setting introduces an added layer of complexity, as x and x' in a pair have independent conditional distributions $\eta(x)$ and $\eta(x')$, which results in a more intricate calibration gap that is harder to address.

As an example, for $\Phi = \Phi_{\rm exp}$ the exponential loss function, we have the lower bound $\Delta\mathcal{C}_{\widetilde{\mathsf{L}}_{\Phi_{\rm exp}},\mathcal{H}}(h,x,x') \geq \Psi_{\rm exp}\Big(\Delta\mathcal{C}_{\mathsf{L}^{\rm abs}_{0-1},\mathcal{H}}(h,x,x')\Big)$, where $\Psi_{\rm exp}$ is an increasing and piece-wise convex function on $[0,2]$ defined by $\Psi_{\rm exp}(t) = \min\Big\{t^2, \Big(\frac{e^{2W\gamma}+1}{e^{2W\gamma}-1}\Big)t\Big\}$, where W is replaced by ΛW for $\mathcal{H} = \mathcal{H}_{\rm NN}$. The detailed proofs and the expression of the corresponding minimizability gaps are included in Appendix D.

4. Importance of abstention

In this section, we show the importance of our abstention methods by demonstrating the impossibility of deriving non-trivial $\mathcal{H}$-consistency bounds with respect to L_{0-1} or $\widetilde{\mathsf{L}}_{0-1}$ for widely used surrogate losses and hypothesis sets.

4.1. Negative Results for General Pairwise Ranking

Here, we give a negative result for standard general pairwise ranking. We will say that a hypothesis set is *regular for general pairwise ranking* if, for any $x \neq x' \in \mathcal{X}$, we have $\{\text{sign}(h(x') - h(x)) : h \in \mathcal{H}\} = \{-1, +1\}$. Hypothesis sets commonly used in practice all admit this property.

The following result shows that the common surrogate losses do not benefit from a non-trivial $\mathcal{H}$-consistency bound when the hypothesis set used is equicontinuous, which includes most hypothesis sets used in practice, in particular the family of linear hypotheses and that of neural networks.

Theorem 4.1 (**Negative results**). *Assume that $\mathcal{X}$ contains an interior point x_0 and that $\mathcal{H}$ is regular for general pairwise ranking, contains 0 and is equicontinuous at x_0. If for some function f that is non-decreasing and continuous at 0, the following bound holds for all $h \in \mathcal{H}$ and any distribution,*

$$\mathcal{R}_{\mathsf{L}_{0-1}}(h) - \mathcal{R}^*_{\mathsf{L}_{0-1}}(\mathcal{H}) \leq f\big(\mathcal{R}_{\mathsf{L}_\Phi}(h) - \mathcal{R}^*_{\mathsf{L}_\Phi}(\mathcal{H})\big),$$

then, $f(t) \geq 1$ for any $t \geq 0$.

Theorem 4.1 shows that for equicontinuous hypothesis sets, any $\mathcal{H}$-consistency bound is vacuous, assuming that f is a non-decreasing function continuous at zero. This is because for any such bound, a small L_Φ-estimation loss does not guarantee a small L_{0-1}-estimation loss, as the right-hand side remains lower-bounded by one.

The proof is given in Appendix E, where we give a simple example on pairs whose distance is relatively small for which the standard surrogate losses including the RankBoost algorithm ($\mathsf{L}_{\rm exp}$) fail (see also Section 5). It is straightforward to see that the assumptions of Theorem 4.1 hold for the case $\mathcal{H} = \mathcal{H}_{\rm lin}$ or $\mathcal{H} = \mathcal{H}_{\rm NN}$. Indeed, we can take $x_0 = 0$ as the interior point and thus for any $h \in \mathcal{H}_{\rm lin}$, $|h(x) - h(x_0)| = |w \cdot x| < \epsilon$ for any $x \in \big\{x \in \mathcal{X} : \|x\|_p < \frac{\epsilon}{W}\big\}$, which implies that $\mathcal{H}_{\rm lin}$ is equicontinuous at x_0. As with the linear hypothesis set, for any $h \in \mathcal{H}_{\rm NN}$, $|h(x) - h(x_0)| = \big|\sum_{j=1}^n u_j(w_j \cdot x + b_j)_+ - \sum_{j=1}^n u_j(b_j)_+\big| = \big|\sum_{j=1}^n u_j[(w_j \cdot x + b_j)_+ - (b_j)_+]\big| \leq \Lambda W \|x\|_p < \epsilon$, for any $x \in \big\{x \in \mathcal{X} : \|x\|_p < \frac{\epsilon}{\Lambda W}\big\}$, which implies that $\mathcal{H}_{\rm NN}$ is equicontinuous at x_0. In fact, Theorem 4.1 holds for any family of Lipschitz constrained neural networks, since a family of functions that share the same Lipschitz constant is equicontinuous.

It is straightforward to verify that the proof of Theorem 4.1 also holds in the deterministic case where $\eta(x, x')$ equals 0 or 1 for any $x \neq x'$, which yields the following corollary.

Corollary 4.2 (**Negative results in the deterministic case**). *In the deterministic case where $\eta(x, x')$ equals 0 or 1 for any $x \neq x'$, the negative result of Theorem 4.1 still holds.*

4.2. Negative Results for Bipartite Ranking

Here, as in the general pairwise misranking scenario, we present a negative result in the standard bipartite setting. We say that a hypothesis set is *regular for bipartite ranking* if, for any $x \neq x' \in \mathcal{X}$, there exists $h_+ \in \mathcal{H}$ such that $h_+(x) < h_+(x')$ and $h_- \in \mathcal{H}$ such that $h_-(x) > h_-(x')$. Hypothesis sets commonly used in practice all admit this property.

As with the general pairwise ranking, we show that common surrogate losses do not benefit from $\mathcal{H}$-consistency bounds when $\mathcal{H}$ is an equicontinuous family.

Theorem 4.3 (**Negative results for bipartite ranking**). *Assume that $\mathcal{X}$ contains an interior point x_0 and that $\mathcal{H}$ is regular for bipartite ranking, contains 0 and is equicontinuous at x_0. If for some function f that is non-decreasing and continuous at 0, the following bound holds for all $h \in \mathcal{H}$ and any distribution,*

$$\mathcal{R}_{\widetilde{\mathsf{L}}_{0-1}}(h) - \mathcal{R}^*_{\widetilde{\mathsf{L}}_{0-1}}(\mathcal{H}) \leq f\Big(\mathcal{R}_{\widetilde{\mathsf{L}}_{\Phi}}(h) - \mathcal{R}^*_{\widetilde{\mathsf{L}}_{\Phi}}(\mathcal{H})\Big),$$

then, $f(t) \geq \frac{1}{2}$ for any $t \geq 0$.

As with Theorem 4.1, Theorem 4.3 shows that in the bipartite ranking setting, any $\mathcal{H}$-consistency bound with an equicontinuous hypothesis set is vacuous, assuming a non-decreasing function f continuous at zero. The proof is given in Appendix F. It is straightforward to verify that the proof holds in the deterministic case where $\eta(x)$ equals 0 or 1 for any $x \in \mathcal{X}$, which yields the following corollary.

Corollary 4.4 (**Negative results in the bipartite deterministic case**). *In the bipartite deterministic case where $\eta(x)$ equals 0 or 1 for any $x \in \mathcal{X}$, the same negative result as in Theorem 4.3 holds.*

The negative results in Section 4.1 and Section 4.2 suggest that without abstention, standard pairwise ranking with theoretical guarantees is difficult with common hypothesis sets. The inherent issue for pairwise ranking is that for equicontinuous hypotheses, when x and x' are arbitrarily close, the confidence value $|h(x) - h(x')|$ can be arbitrary close to zero. These results further imply the importance of ranking with abstention, where the learner can abstain from making prediction on a pair (x, x') with x and x' relatively close, as illustrated in Section 2 and Section 3.

5. Experiments

In this section, we provide empirical results for general pairwise ranking with abstention on the CIFAR-10 dataset (Krizhevsky, 2009).

We used ResNet-34 with ReLU activations (He et al., 2016). Here, ResNet-n denotes a residual network with n convolutional layers. Standard data augmentations, 4-pixel padding with 32×32 random crops and random horizontal flips are applied for CIFAR-10. For training, we used Stochastic Gradient Descent (SGD) with Nesterov momentum (Nesterov, 1983). We set the batch size, weight decay, and initial learning rate to 1,024, 1×10^{-4} and 0.1 respectively. We adopted the cosine decay learning rate schedule (Loshchilov & Hutter, 2016) for a total of 200 epochs. The pairs (x, x', y) are randomly sampled from CIFAR-10 during training, with $y = \pm 1$ indicating if x is ranked above or below x' per the natural ordering of labels of x and x'.

We evaluated the models based on their averaged pairwise abstention loss (2) with γ selected from $\{0.0, 0.3, 0.5, 0.7, 0.9\}$ and the cost c selected from $\{0.1, 0.3, 0.5\}$. We randomly sampled 10,000 pairs (x, x') from the test data for evaluation. The ℓ_∞ distance is adopted in the algorithm. We averaged losses over three runs and report the standard deviation as well.

We used the surrogate loss (3) with $\Phi(t) = \exp(-t)$ the exponential loss, $\mathsf{L}_{\Phi_{\exp}}$, which coincides with the loss function of RankBoost. Table 1 shows that when γ is as small as 0.3, no abstention takes place and the abstention loss coincides with the standard misranking loss ($\gamma = 0$) for any cost c. As γ increases, there are more samples that are abstained. When using a minimal cost c of 0.1 (as demonstrated in the first row of Table 1), abstaining on pairs with a relatively small distance ($\gamma = 0.7$) results in a lower target abstention loss compared to the scenario without abstention ($\gamma = 0$). Conversely, abstaining on pairs with larger distances ($\gamma = 0.9$) led to a higher abstention loss. This can be attributed to the fact that rejected samples at $\gamma = 0.7$ had lower accuracy compared to those at $\gamma = 0.9$. This empirically verifies that the surrogate loss $\mathsf{L}_{\Phi_{\exp}}$ is not favorable on pairs whose distance is relatively small, for equicontinuous hypotheses. When the cost c is larger, the abstention loss, in general, increases with γ, since the number of samples rejected increases with γ.

Overall, the experiment shows that, in practice, for small γ, abstention actually does not take place. Thus, the abstention loss coincides with the standard pairwise misranking loss in those cases, and the surrogate loss is consistent with respect to both of them. Our results also indicate that the surrogate loss $\mathsf{L}_{\Phi_{\exp}}$, a commonly used loss function, for example for RankBoost, is not optimal for pairs with a relatively small distance. Instead, rejecting these pairs at a minimal cost proves to be a more effective strategy.

6. Conclusion

We introduce a novel framework of ranking with abstention, in both the general pairwise ranking and the bipartite ranking scenarios. Our proposed abstention methods are important when using common equicontinuous hypothesis sets in practice. It will be useful to explore alternative non-equicontinuous hypothesis sets that may be of practical use, and to further study the choice of the parameter γ for abstention in practice. We have also initiated the study of randomized ranking solutions with theoretical guarantees.

References

Agarwal, S. Surrogate regret bounds for bipartite ranking via strongly proper losses. *The Journal of Machine Learning Research*, 15(1):1653–1674, 2014.

Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., Roth, D., and Jordan, M. I. Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research*, 6(4), 2005.

Ailon, N. and Mohri, M. An efficient reduction of ranking to classification. In *Conference on Learning Theory*, 2008.

Ailon, N. and Mohri, M. Preference-based learning to rank. *Machine Learning*, 80(2-3):189–211, 2010.

Awasthi, P., Frank, N., Mao, A., Mohri, M., and Zhong, Y. Calibration and consistency of adversarial surrogate losses. In *Advances in Neural Information Processing Systems*, 2021a.

Awasthi, P., Mao, A., Mohri, M., and Zhong, Y. A finer calibration analysis for adversarial robustness. *arXiv preprint arXiv:2105.01550*, 2021b.

Awasthi, P., Mao, A., Mohri, M., and Zhong, Y. H-consistency bounds for surrogate loss minimizers. In *International Conference on Machine Learning*, 2022a.

Awasthi, P., Mao, A., Mohri, M., and Zhong, Y. Multi-class $\mathcal{H}$-consistency bounds. In *Advances in neural information processing systems*, 2022b.

Awasthi, P., Mao, A., Mohri, M., and Zhong, Y. DC-programming for neural network optimizations. *Journal of Global Optimization*, 2023a.

Awasthi, P., Mao, A., Mohri, M., and Zhong, Y. Theoretically grounded loss functions and algorithms for adversarial robustness. In *International Conference on Artificial Intelligence and Statistics*, pp. 10077–10094, 2023b.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Buffoni, D., Calauzenes, C., Gallinari, P., and Usunier, N. Learning scoring functions with order-preserving losses and standardized supervision. In *International Conference on Machine Learning*, pp. 825–832, 2011.

Calauzenes, C., Usunier, N., and Gallinari, P. On the (non-) existence of convex, calibrated surrogate losses for ranking. In *Advances in Neural Information Processing Systems*, 2012.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017.

Clemençon, S., Lugosi, G., and Vayatis, N. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

Cohen, W. W., Schapire, R. E., and Singer, Y. Learning to order things. *Advances in neural information processing systems*, 10, 1997.

Cortes, C. and Mohri, M. AUC optimization vs. error rate minimization. *Advances in neural information processing systems*, 16, 2003.

Cossock, D. and Zhang, T. Statistical analysis of bayes optimal subset ranking. *IEEE Transactions on Information Theory*, 54(11):5140–5154, 2008.

Duchi, J. C., Mackey, L. W., and Jordan, M. I. On the consistency of ranking algorithms. In *International conference on Machine learning*, pp. 327–334, 2010.

Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. An efficient boosting algorithm for combining preferences. *Journal of machine learning research*, 4(Nov):933–969, 2003.

Gao, W. and Zhou, Z.-H. On the consistency of multi-label learning. In *Conference on learning theory*, pp. 341–358, 2011.

Gao, W. and Zhou, Z.-H. On the consistency of AUC pairwise optimization. In *International Joint Conference on Artificial Intelligence*, 2015.

Gao, W., Jin, R., Zhu, S., and Zhou, Z.-H. One-pass auc optimization. In *International conference on machine learning*, pp. 906–914, 2013.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Hanley, J. A. and McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 1982.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Joachims, T. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 133–142, 2002.

Kotlowski, W., Dembczynski, K. J., and Huellermeier, E. Bipartite ranking through minimization of univariate loss. In *International Conference on Machine Learning*, pp. 1113–1120, 2011.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, Toronto University, 2009.

Kuznetsov, V., Mohri, M., and Syed, U. Multi-class deep boosting. In *Advances in Neural Information Processing Systems*, pp. 2501–2509, 2014.

Lan, Y., Guo, J., Cheng, X., and Liu, T.-Y. Statistical consistency of ranking methods in a rank-differentiable probability space. In *Advances in Neural Information Processing Systems*, 2012.

Long, P. and Servedio, R. Consistency versus realizable H-consistency for multiclass classification. In *International Conference on Machine Learning*, pp. 801–809, 2013.

Loshchilov, I. and Hutter, F. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Mao, A., Mohri, M., and Zhong, Y. H-consistency bounds for pairwise misranking loss surrogates. In *International conference on Machine learning*, 2023a.

Mao, A., Mohri, M., and Zhong, Y. Cross-entropy loss functions: Theoretical analysis and applications. *arXiv preprint arXiv:2304.07288*, 2023b.

Menon, A. K. and Williamson, R. C. Bayes-optimal scorers for bipartite ranking. In *Conference on Learning Theory*, pp. 68–106, 2014.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. MIT Press, second edition, 2018.

Nesterov, Y. E. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Dokl. akad. nauk Sssr*, 269:543–547, 1983.

Ramaswamy, H. G. and Agarwal, S. Classification calibration dimension for general multiclass losses. In *Advances in Neural Information Processing Systems*, 2012.

Ramaswamy, H. G., Agarwal, S., and Tewari, A. Convex calibrated surrogates for low-rank loss matrices with applications to subset ranking losses. In *Advances in Neural Information Processing Systems*, 2013.

Ramaswamy, H. G., Babu, B. S., Agarwal, S., and Williamson, R. C. On the consistency of output code based learning algorithms for multiclass learning problems. In *Conference on Learning Theory*, pp. 885–902, 2014.

Ravikumar, P., Tewari, A., and Yang, E. On ndcg consistency of listwise ranking methods. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 618–626, 2011.

Rudin, C., Cortes, C., Mohri, M., and Schapire, R. E. Margin-based ranking meets boosting in the middle. In *Conference on Learning Theory*, pp. 63–78, 2005.

Steinwart, I. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007.

Tewari, A. and Bartlett, P. L. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(36):1007–1025, 2007.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Uematsu, K. and Lee, Y. On theoretically optimal ranking functions in bipartite ranking. *Journal of the American Statistical Association*, 112(519):1311–1322, 2017.

Xia, F., Liu, T.-Y., Wang, J., Zhang, W., and Li, H. Listwise approach to learning to rank: theory and algorithm. In *International conference on Machine learning*, pp. 1192–1199, 2008.

Zhang, M. and Agarwal, S. Bayes consistency vs. H-consistency: The interplay between surrogate loss functions and the scoring function class. In *Advances in Neural Information Processing Systems*, 2020.

Zhang, M., Ramaswamy, H. G., and Agarwal, S. Convex calibrated surrogates for the multi-label f-measure. In *International Conference on Machine Learning*, pp. 11246–11255, 2020.

Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

Zheng, C., Wu, G., Bao, F., Cao, Y., Li, C., and Zhu, J. Revisiting discriminative vs. generative classifiers: Theory and implications. *arXiv preprint arXiv:2302.02334*, 2023.

Contents of Appendix

A. Related work

The notions of Bayes consistency (also known as consistency) and calibration have been extensively studied for classification (Zhang, 2004; Bartlett et al., 2006; Tewari & Bartlett, 2007). The Bayes consistency of ranking surrogate losses has been studied in the special case of bipartite score-based ranking: in particular, Uematsu & Lee (2017) proved the inconsistency of the pairwise ranking loss based on the hinge loss and Gao & Zhou (2015) gave excess loss bounds for pairwise ranking losses based on the exponential or the logistic loss. Later, these results were further generalized by Menon & Williamson (2014). A related but distinct consistency question has been studied in several publications (Agarwal et al., 2005; Kotlowski et al., 2011; Agarwal, 2014). It is one with respect to binary classification, that is whether a near minimizer of the surrogate loss of the binary classification loss is a near minimizer of the bipartite misranking loss (Cortes & Mohri, 2003).

Considerable attention has been devoted to the study of the learning to rank algorithms and their related problems: including one-pass AUC pairwise optimization (Gao et al., 2013), preference-based ranking (Cohen et al., 1997; Clemençon et al., 2008), subset ranking with Discounted Cumulative Gain (DCG) (Cossock & Zhang, 2008; Buffoni et al., 2011), listwise ranking (Xia et al., 2008), subset ranking based on Pairwise Disagreement (PD) (Duchi et al., 2010; Lan et al., 2012), subset ranking using Normalized Discounted Cumulative Gain (NDCG) (Ravikumar et al., 2011), subset ranking with Average Precision (AP) (Calauzenes et al., 2012; Ramaswamy et al., 2013), general multi-class problems (Ramaswamy & Agarwal, 2012; Ramaswamy et al., 2014) and multi-label problems (Gao & Zhou, 2011; Zhang et al., 2020).

Bayes consistency only holds for the full family of measurable functions, which of course is distinct from the more restricted hypothesis set used by a learning algorithm. Therefore, a hypothesis set-dependent notion of $\mathcal{H}$-consistency has been proposed by Long & Servedio (2013) in the realizable setting, which was used by Zhang & Agarwal (2020) for linear models, and generalized by Kuznetsov et al. (2014) to the structured prediction case. Long & Servedio (2013) showed that there exists a case where a Bayes-consistent loss is not $\mathcal{H}$-consistent while inconsistent loss functions can be $\mathcal{H}$-consistent. Zhang & Agarwal (2020) further investigated the phenomenon in (Long & Servedio, 2013) and showed that the situation of loss functions that are not $\mathcal{H}$-consistent with linear models can be remedied by carefully choosing a larger piecewise linear hypothesis set. Kuznetsov et al. (2014) proved positive results for the $\mathcal{H}$-consistency of several multi-class ensemble algorithms, as an extension of $\mathcal{H}$-consistency results in (Long & Servedio, 2013).

Recently, Awasthi et al. (2022a) presented a series of results providing $\mathcal{H}$-consistency bounds in binary classification. These guarantees are significantly stronger than the $\mathcal{H}$-calibration or $\mathcal{H}$-consistency properties studied by Awasthi et al. (2021a;b). Awasthi et al. (2022b) and Mao et al. (2023b) (see also (Zheng et al., 2023)) generalized $\mathcal{H}$-consistency bounds to the scenario of multi-class classification. Awasthi et al. (2023b) proposed a family of loss functions that benefit from such $\mathcal{H}$-consistency bounds guarantees for adversarial robustness (Goodfellow et al., 2014; Madry et al., 2017; Tsipras et al., 2018; Carlini & Wagner, 2017; Awasthi et al., 2023a). Mao et al. (2023a) used $\mathcal{H}$-consistency bounds in the context of ranking. $\mathcal{H}$-consistency bounds are also more informative than similar excess error bounds derived in the literature, which correspond to the special case where $\mathcal{H}$ is the family of all measurable functions (Zhang, 2004; Bartlett et al., 2006; Mohri et al., 2018). Our work significantly generalizes the results of Awasthi et al. (2022a) to the score-based ranking setting, including both the general pairwise ranking and bipartite ranking scenarios.

B. General tools

To begin with the proof, we first introduce some notation. In general pairwise ranking scenario, we denote by $\mathcal{D}$ a distribution over $\mathcal{X} \times \mathcal{X} \times \mathcal{Y}$ and by $\mathcal{P}$ a set of such distributions. We further denote by $\eta(x, x') = \mathcal{D}(Y = 1 \mid (X, X') = (x, x'))$ the conditional probability of $Y = 1$ given $(X, X') = (x, x')$. Without loss of generality, we assume that $\eta(x, x) = 1/2$. The generalization error for a surrogate loss L can be rewritten as $\mathcal{R}_{\mathsf{L}}(h) = \mathbb{E}_X[\mathcal{C}_{\mathsf{L}}(h, x, x')]$, where $\mathcal{C}_{\mathsf{L}}(h, x, x')$ is the conditional L-risk, defined by

$$\mathcal{C}_{\mathsf{L}}(h, x, x') = \eta(x, x')\mathsf{L}(h, x, x', +1) + (1 - \eta(x, x'))\mathsf{L}(h, x, x', -1).$$

We denote by $\mathcal{C}^*_{\mathsf{L}}(\mathcal{H}, x, x') = \inf_{h \in \mathcal{H}} \mathcal{C}_{\mathsf{L}}(h, x, x')$ the minimal conditional L-risk. Then, the minimizability gap can be rewritten as follows:

$$\mathcal{M}_{\mathsf{L}}(\mathcal{H}) = \mathcal{R}^*_{\mathsf{L}}(\mathcal{H}) - \mathbb{E}_X[\mathcal{C}^*_{\mathsf{L}}(\mathcal{H}, x)].$$

We further refer to $\mathcal{C}_{\mathsf{L}}(h, x, x') - \mathcal{C}^*_{\mathsf{L}}(\mathcal{H}, x, x')$ as the calibration gap and denote it by $\Delta\mathcal{C}_{\mathsf{L},\mathcal{H}}(h, x, x')$.

In bipartite ranking scenario, we denote by $\mathcal{D}$ a distribution over $\mathcal{X} \times \mathcal{Y}$ and by $\mathcal{P}$ a set of such distributions. We further denote by $\eta(x) = \mathcal{D}(Y = 1 \mid X = x)$ the conditional probability of $Y = 1$ given $X = x$. The generalization error for a surrogate loss $\widetilde{\mathsf{L}}$ can be rewritten as $\mathcal{R}_{\widetilde{\mathsf{L}}}(h) = \mathbb{E}_X\big[\mathcal{C}_{\widetilde{\mathsf{L}}}(h, x, x')\big]$, where $\mathcal{C}_{\widetilde{\mathsf{L}}}(h, x, x')$ is the conditional $\widetilde{\mathsf{L}}$-risk, defined by

$$\mathcal{C}_{\widetilde{\mathsf{L}}}(h, x, x') = \eta(x)(1 - \eta(x'))\widetilde{\mathsf{L}}(h, x, x', +1, -1) + \eta(x')(1 - \eta(x))\widetilde{\mathsf{L}}(h, x, x' - 1, +1).$$

We denote by $\mathcal{C}^*_{\widetilde{\mathsf{L}}}(\mathcal{H}, x, x') = \inf_{h \in \mathcal{H}} \mathcal{C}_{\widetilde{\mathsf{L}}}(h, x, x')$ the minimal conditional $\widetilde{\mathsf{L}}$-risk. Then, the minimizability gap can be rewritten as follows:

$$\mathcal{M}_{\widetilde{\mathsf{L}}}(\mathcal{H}) = \mathcal{R}^*_{\widetilde{\mathsf{L}}}(\mathcal{H}) - \mathbb{E}_X\big[\mathcal{C}^*_{\widetilde{\mathsf{L}}}(\mathcal{H}, x)\big].$$

We further refer to $\mathcal{C}_{\widetilde{\mathsf{L}}}(h, x, x') - \mathcal{C}^*_{\widetilde{\mathsf{L}}}(\mathcal{H}, x, x')$ as the calibration gap and denote it by $\Delta\mathcal{C}_{\widetilde{\mathsf{L}}, \mathcal{H}}(h, x, x')$. For any $\epsilon > 0$, we will denote by $\langle t \rangle_\epsilon$ the ϵ-truncation of $t \in \mathbb{R}$ defined by $t \mathbb{1}_{t > \epsilon}$.

We first prove two general results, which provide bounds between any loss functions L_1 and L_2 in both general pairwise ranking scenario and bipartite ranking scenario.

Theorem B.1. *Assume that there exists a convex function $\Psi \colon \mathbb{R}_+ \to \mathbb{R}$ with $\Psi(0) \geq 0$ and $\epsilon \geq 0$ such that the following holds for all $h \in \mathcal{H}$, $x \in \mathcal{X}$, $x' \in \mathcal{X}$ and $\mathcal{D} \in \mathcal{P}$:*

$$\Psi\big(\langle \Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x') \rangle_\epsilon\big) \leq \langle \Delta\mathcal{C}_{\mathsf{L}_1, \mathcal{H}}(h, x, x') \rangle_\epsilon. \tag{7}$$

Then, the following inequality holds for any $h \in \mathcal{H}$ and $\mathcal{D} \in \mathcal{P}$:

$$\Psi\big(\mathcal{R}_{\mathsf{L}_2}(h) - \mathcal{R}^*_{\mathsf{L}_2}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_2}(\mathcal{H})\big) \leq \mathcal{R}_{\mathsf{L}_1}(h) - \mathcal{R}^*_{\mathsf{L}_1}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_1}(\mathcal{H}) + \max\{\Psi(0), \Psi(\epsilon)\}. \tag{8}$$

Proof. By the definition of the generalization error and the minimizability gap, for any $h \in \mathcal{H}$ and $\mathcal{D} \in \mathcal{P}$, we can write the left hand side of (8) as

$$\Psi\big(\mathcal{R}_{\mathsf{L}_2}(h) - \mathcal{R}^*_{\mathsf{L}_2}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_2}(\mathcal{H})\big) = \Psi\big(\mathcal{R}_{\mathsf{L}_2}(h) - \mathbb{E}_{(X,X')}\big[\mathcal{C}^*_{\mathsf{L}_2}(\mathcal{H}, x, x')\big]\big) = \Psi\big(\mathbb{E}_{(X,X')}[\Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x')]\big).$$

Since Ψ is convex, by Jensen's inequality, it can be upper bounded by $\mathbb{E}_{(X,X')}[\Psi(\Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x'))]$. Due to the decomposition

$$\Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x') = \langle \Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x') \rangle_\epsilon + \Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x') \mathbb{1}_{\Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x') \leq \epsilon},$$

and the assumption $\Psi(0) \geq 0$, we have the following inequality:

$$\mathbb{E}_{(X,X')}[\Psi(\Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x'))] \leq \mathbb{E}_{(X,X')}\big[\Psi\big(\langle \Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x') \rangle_\epsilon\big)\big] + \mathbb{E}_{(X,X')}\big[\Psi\big(\Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x') \mathbb{1}_{\Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x') \leq \epsilon}\big)\big].$$

By assumption (7), the first term can be bounded as follows:

$$\mathbb{E}_{(X,X')}\big[\Psi\big(\langle \Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x') \rangle_\epsilon\big)\big] \leq \mathbb{E}_{(X,X')}[\Delta\mathcal{C}_{\mathsf{L}_1, \mathcal{H}}(h, x, x')] = \mathcal{R}_{\mathsf{L}_1}(h) - \mathcal{R}^*_{\mathsf{L}_1}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_1}(\mathcal{H}).$$

Since $\Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x') \mathbb{1}_{\Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x') \leq \epsilon} \in [0, \epsilon]$, we can bound $\mathbb{E}_{(X,X')}\big[\Psi\big(\Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x') \mathbb{1}_{\Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x') \leq \epsilon}\big)\big]$ by $\sup_{t \in [0, \epsilon]} \Psi(t)$, which equals $\max\{\Psi(0), \Psi(\epsilon)\}$ due to the convexity of Ψ. □

Theorem B.2. *Assume that there exists a non-decreasing concave function $\Gamma \colon \mathbb{R}_+ \to \mathbb{R}$ and $\epsilon \geq 0$ such that the following holds for all $h \in \mathcal{H}$, $x \in \mathcal{X}$, $x' \in \mathcal{X}$ and $\mathcal{D} \in \mathcal{P}$:*

$$\langle \Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x') \rangle_\epsilon \leq \Gamma\big(\langle \Delta\mathcal{C}_{\mathsf{L}_1, \mathcal{H}}(h, x, x') \rangle_\epsilon\big). \tag{9}$$

Then, the following inequality holds for any $h \in \mathcal{H}$ and $\mathcal{D} \in \mathcal{P}$:

$$\mathcal{R}_{\mathsf{L}_2}(h) - \mathcal{R}^*_{\mathsf{L}_2}(\mathcal{H}) \leq \Gamma\big(\mathcal{R}_{\mathsf{L}_1}(h) - \mathcal{R}^*_{\mathsf{L}_1}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_1}(\mathcal{H})\big) - \mathcal{M}_{\mathsf{L}_2}(\mathcal{H}) + \epsilon. \tag{10}$$

Proof. By the definition of the generalization error and the minimizability gap, for any $h \in \mathcal{H}$ and $\mathcal{D} \in \mathcal{P}$, we can write the left hand side of (10) as

$$\begin{aligned}
&\mathcal{R}_{\mathsf{L}_2}(h) - \mathcal{R}^*_{\mathsf{L}_2}(\mathcal{H})\\
&= \mathbb{E}_{(X,X')}[\Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x')] - \mathcal{M}_{\mathsf{L}_2}(\mathcal{H})\\
&= \mathbb{E}_{(X,X')}\big[\langle \Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x') \rangle_\epsilon\big] + \mathbb{E}_{(X,X')}\big[\Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x') \mathbb{1}_{\Delta\mathcal{C}_{\mathsf{L}_2, \mathcal{H}}(h, x, x') \leq \epsilon}\big] - \mathcal{M}_{\mathsf{L}_2}(\mathcal{H})
\end{aligned}$$

By assumption (9) and that Γ is non-decreasing, the following inequality holds:

$$\mathbb{E}_{(X,X')}\Big[\langle \Delta\mathcal{C}_{\mathsf{L}_2,\mathcal{H}}(h,x,x')\rangle_{\epsilon}\Big] \leq \mathbb{E}_{(X,X')}\big[\Gamma(\Delta\mathcal{C}_{\mathsf{L}_1,\mathcal{H}}(h,x,x'))\big].$$

Since Γ is concave, by Jensen's inequality,

$$\mathbb{E}_{(X,X')}\big[\Gamma(\Delta\mathcal{C}_{\mathsf{L}_1,\mathcal{H}}(h,x,x'))\big] \leq \Gamma\big(\mathbb{E}_{(X,X')}\big[\Delta\mathcal{C}_{\mathsf{L}_1,\mathcal{H}}(h,x,x')\big]\big) = \Gamma\big(\mathcal{R}_{\mathsf{L}_1}(h) - \mathcal{R}^*_{\mathsf{L}_1}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_1}(\mathcal{H})\big).$$

We complete the proof by noting that $\mathbb{E}_{(X,X')}\Big[\Delta\mathcal{C}_{\mathsf{L}_2,\mathcal{H}}(h,x,x')\mathbb{1}_{\Delta\mathcal{C}_{\mathsf{L}_2,\mathcal{H}}(h,x,x')\leq\epsilon}\Big] \leq \epsilon$. □

C. $\mathcal{H}$ - consistency bounds for general pairwise ranking with abstention (Proof of Theorem 2.1)

We first characterize the minimal conditional $\mathsf{L}^{\text{abs}}_{0-1}$-risk and the calibration gap of $\mathsf{L}^{\text{abs}}_{0-1}$ for a broad class of hypothesis sets. We let $\overline{\mathcal{H}}(x,x') = \{h\in\mathcal{H}\colon \operatorname{sign}(h(x')-h(x))(2\eta(x,x')-1)\leq 0\}$ for convenience.

Lemma C.1. *Assume that $\mathcal{H}$ is regular for general pairwise ranking. Then, the minimal conditional $\mathsf{L}^{\text{abs}}_{0-1}$-risk is*

$$\mathcal{C}^*_{\mathsf{L}^{\text{abs}}_{0-1}}(\mathcal{H},x,x') = \min\{\eta(x,x'),1-\eta(x,x')\}\mathbb{1}_{\|x-x'\|>\gamma} + c\,\mathbb{1}_{|x-x'|\leq\gamma}.$$

The calibration gap of $\mathsf{L}^{\text{abs}}_{0-1}$ can be characterized as

$$\Delta\mathcal{C}_{\mathsf{L}^{\text{abs}}_{0-1},\mathcal{H}}(h,x,x') = |2\eta(x,x')-1|\mathbb{1}_{h\in\overline{\mathcal{H}}(x,x')}\mathbb{1}_{\|x-x'\|>\gamma}.$$

Proof. By the definition, the conditional $\mathsf{L}^{\text{abs}}_{0-1}$-risk is

$$\mathcal{C}_{\mathsf{L}^{\text{abs}}_{0-1}}(h,x,x') = \big(\eta(x,x')\mathbb{1}_{h(x')<h(x)} + (1-\eta(x,x'))\mathbb{1}_{h(x')\geq h(x)}\big)\mathbb{1}_{\|x-x'\|>\gamma} + c\,\mathbb{1}_{|x-x'|\leq\gamma}.$$

For any (x,x') such that $\|x-x'\|\leq\gamma$ and $h\in\mathcal{H}$, $\mathcal{C}_{\mathsf{L}^{\text{abs}}_{0-1}}(h,x,x) = \mathcal{C}^*_{\mathsf{L}^{\text{abs}}_{0-1}}(\mathcal{H},x,x) = c$. For any (x,x') such that $\|x-x'\|>\gamma$, by the assumption, there exists $h^*\in\mathcal{H}$ such that $\operatorname{sign}(h^*(x')-h^*(x)) = \operatorname{sign}(2\eta(x,x')-1)$. Therefore, the optimal conditional $\mathsf{L}^{\text{abs}}_{0-1}$-risk can be characterized as for any $x,x'\in\mathcal{X}$,

$$\mathcal{C}^*_{\mathsf{L}^{\text{abs}}_{0-1}}(\mathcal{H},x,x') = \mathcal{C}_{\mathsf{L}^{\text{abs}}_{0-1}}(h^*,x,x') = \min\{\eta(x,x'),1-\eta(x,x')\}\mathbb{1}_{\|x-x'\|>\gamma} + c\,\mathbb{1}_{|x-x'|\leq\gamma}.$$

which proves the first part of lemma. By the definition, for any (x,x') such that $\|x-x'\|\leq\gamma$ and $h\in\mathcal{H}$, $\Delta\mathcal{C}_{\mathsf{L}^{\text{abs}}_{0-1},\mathcal{H}}(h,x,x') = \mathcal{C}_{\mathsf{L}^{\text{abs}}_{0-1}}(h,x,x') - \mathcal{C}^*_{\mathsf{L}^{\text{abs}}_{0-1}}(\mathcal{H},x,x') = 0$. . For any (x,x') such that $\|x-x'\|>\gamma$ and $h\in\mathcal{H}$,

$$\begin{aligned}
\Delta\mathcal{C}_{\mathsf{L}^{\text{abs}}_{0-1},\mathcal{H}}(h,x,x') &= \mathcal{C}_{\mathsf{L}^{\text{abs}}_{0-1}}(h,x,x') - \mathcal{C}^*_{\mathsf{L}^{\text{abs}}_{0-1}}(\mathcal{H},x,x')\\
&= \eta(x,x')\mathbb{1}_{h(x')<h(x)} + (1-\eta(x,x'))\mathbb{1}_{h(x')\geq h(x)} - \min\{\eta(x,x'),1-\eta(x,x')\}\\
&= \begin{cases} |2\eta(x,x')-1|, & h\in\overline{\mathcal{H}}(x,x'),\\ 0, & \text{otherwise.}\end{cases}
\end{aligned}$$

This leads to

$$\Delta\mathcal{C}_{\mathsf{L}^{\text{abs}}_{0-1},\mathcal{H}}(h,x,x') = |2\eta(x,x')-1|\mathbb{1}_{h\in\overline{\mathcal{H}}(x,x')}\mathbb{1}_{\|x-x'\|>\gamma}.$$

□

Theorem 2.1 (**$\mathcal{H}$-consistency bounds for pairwise abstention loss**). *Let $\mathcal{H}$ be $\mathcal{H}_{\text{lin}}$ or $\mathcal{H}_{\text{NN}}$. Then, for any $h\in\mathcal{H}$ and any distribution,*

$$\mathcal{R}_{\mathsf{L}^{\text{abs}}_{0-1}}(h) - \mathcal{R}^*_{\mathsf{L}^{\text{abs}}_{0-1}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}^{\text{abs}}_{0-1}}(\mathcal{H}) \leq \Gamma_{\Phi}\big(\mathcal{R}_{\mathsf{L}_{\Phi}}(h) - \mathcal{R}^*_{\mathsf{L}_{\Phi}}(\mathcal{H}) + \mathcal{M}_{\mathsf{L}_{\Phi}}(\mathcal{H})\big),$$

where $\Gamma_{\Phi}(t) = \frac{t}{\min\{W\gamma,1\}}$, $\max\Big\{\sqrt{2t}, 2\Big(\frac{e^{2W\gamma}+1}{e^{2W\gamma}-1}\Big)t\Big\}$ and $\frac{t}{\tanh(kW\gamma)}$ for $\Phi = \Phi_{\text{hinge}}$, $\Phi_{\exp}$ and Φ_{sig} respectively. W is replaced by ΛW for $\mathcal{H} = \mathcal{H}_{\text{NN}}$.

Proof. Since $\mathcal{H}_{\rm lin}$ and $\mathcal{H}_{\rm NN}$ satisfy the condition of Lemma C.1, by Lemma C.1 the $\left(\mathsf{L}^{\rm abs}_{0-1}, \mathcal{H}_{\rm lin}\right)$-minimizability gap and the $\left(\mathsf{L}^{\rm abs}_{0-1}, \mathcal{H}_{\rm NN}\right)$-minimizability gap can be expressed as follows:

$$\mathcal{M}_{\mathsf{L}^{\rm abs}_{0-1}}(\mathcal{H}_{\rm lin}) = \mathcal{R}^*_{\mathsf{L}^{\rm abs}_{0-1}}(\mathcal{H}_{\rm lin}) - \mathbb{E}_{(X,X')}\left[\min\{\eta(x,x'), 1-\eta(x,x')\}\mathbb{1}_{\|x-x'\|>\gamma} + c\,\mathbb{1}_{|x-x'|\leq\gamma}\right]$$
$$\mathcal{M}_{\mathsf{L}^{\rm abs}_{0-1}}(\mathcal{H}_{\rm NN}) = \mathcal{R}^*_{\mathsf{L}^{\rm abs}_{0-1}}(\mathcal{H}_{\rm NN}) - \mathbb{E}_{(X,X')}\left[\min\{\eta(x,x'), 1-\eta(x,x')\}\mathbb{1}_{\|x-x'\|>\gamma} + c\,\mathbb{1}_{|x-x'|\leq\gamma}.\right].$$

By the definition of $\mathcal{H}_{\rm lin}$ and $\mathcal{H}_{\rm NN}$, for any $(x,x') \in \mathcal{X}\times\mathcal{X}$, $\left\{h(x') - h(x) \mid h \in \mathcal{H}_{\rm lin}\right\} = \left[-W\|x-x'\|_p, W\|x-x'\|_p\right]$ and $\left\{h(x') - h(x) \mid h \in \mathcal{H}_{\rm NN}\right\} = \left[-\Lambda W\|x-x'\|_p, \Lambda W\|x-x'\|_p\right]$. In the following, we will prove the bounds for $\mathcal{H}_{\rm lin}$. Similar proofs with B replaced by ΛB hold for $\mathcal{H}_{\rm NN}$.

Proof for $\mathsf{L}_{\Phi_{\rm hinge}}$. For the hinge loss function $\Phi_{\rm hinge}(u) := \max\{0, 1-u\}$, for all $h \in \mathcal{H}_{\rm lin}$ and (x,x') such that $\|x-x'\|_p > \gamma$,

$$\begin{aligned}
&\mathcal{C}_{\mathsf{L}_{\Phi_{\rm hinge}}}(h,x,x') \\
&= \eta(x,x')\mathsf{L}_{\Phi_{\rm hinge}}(h(x') - h(x)) + (1-\eta(x,x'))\mathsf{L}_{\Phi_{\rm hinge}}(h(x) - h(x')) \\
&= \eta(x,x')\max\{0, 1-h(x')+h(x)\} + (1-\eta(x,x'))\max\{0, 1+h(x')-h(x)\}.
\end{aligned}$$

Then,

$$\mathcal{C}^*_{\mathsf{L}_{\Phi_{\rm hinge}},\mathcal{H}_{\rm lin}}(x,x') = \inf_{h\in\mathcal{H}_{\rm lin}} \mathcal{C}_{\mathsf{L}_{\Phi_{\rm hinge}}}(h,x,x') = 1 - |2\eta(x,x') - 1|\min\left\{W\|x-x'\|_p, 1\right\}.$$

The $\left(\mathsf{L}_{\Phi_{\rm hinge}}, \mathcal{H}_{\rm lin}\right)$-minimizability gap is

$$\begin{aligned}
\mathcal{M}_{\mathsf{L}_{\Phi_{\rm hinge}}}(\mathcal{H}_{\rm lin}) &= \mathcal{R}^*_{\mathsf{L}_{\Phi_{\rm hinge}}}(\mathcal{H}_{\rm lin}) - \mathbb{E}_{(X,X')}\left[\mathcal{C}^*_{\mathsf{L}_{\Phi_{\rm hinge}},\mathcal{H}_{\rm lin}}(x,x')\right] \\
&= \mathcal{R}^*_{\mathsf{L}_{\Phi_{\rm hinge}}}(\mathcal{H}_{\rm lin}) - \mathbb{E}_{(X,X')}\left[1 - |2\eta(x,x') - 1|\min\left\{W\|x-x'\|_p, 1\right\}\right].
\end{aligned} \tag{11}$$

Therefore, $\forall h \in \overline{\mathcal{H}}_{\rm lin}(x,x')$,

$$\begin{aligned}
&\Delta\mathcal{C}_{\mathsf{L}_{\Phi_{\rm hinge}},\mathcal{H}_{\rm lin}}(h,x,x') \\
&\geq \inf_{h\in\overline{\mathcal{H}}_{\rm lin}(x,x')} \mathcal{C}_{\mathsf{L}_{\Phi_{\rm hinge}}}(h,x,x') - \mathcal{C}^*_{\mathsf{L}_{\Phi_{\rm hinge}},\mathcal{H}_{\rm lin}}(x,x') \\
&= \eta(x,x')\max\{0, 1-0\} + (1-\eta(x,x'))\max\{0, 1+0\} - \mathcal{C}^*_{\mathsf{L}_{\Phi_{\rm hinge}},\mathcal{H}_{\rm lin}}(x,x') \\
&= 1 - \left[1 - |2\eta(x,x') - 1|\min\left\{W\|x-x'\|_p, 1\right\}\right] \\
&= |2\eta(x,x') - 1|\min\left\{W\|x-x'\|_p, 1\right\} \\
&\geq |2\eta(x,x') - 1|\min\{W\gamma, 1\}
\end{aligned}$$

which implies that for any $h \in \mathcal{H}_{\rm lin}$ and (x,x') such that $\|x-x'\|_p > \gamma$,

$$\Delta\mathcal{C}_{\mathsf{L}_{\Phi_{\rm hinge}},\mathcal{H}_{\rm lin}}(h,x,x') \geq \min\{W\gamma, 1\}\langle|2\eta(x,x') - 1|\rangle_0 \mathbb{1}_{h\in\overline{\mathcal{H}}_{\rm lin}(x,x')} = \Delta\mathcal{C}_{\mathsf{L}^{\rm abs}_{0-1},\mathcal{H}_{\rm lin}}(h,x,x').$$

Thus, by Theorem B.1 or Theorem B.2, setting $\epsilon = 0$ yields the $\mathcal{H}_{\rm lin}$-consistency bound for $\mathsf{L}_{\Phi_{\rm hinge}}$, valid for all $h \in \mathcal{H}_{\rm lin}$:

$$\mathcal{R}_{\mathsf{L}^{\rm abs}_{0-1}}(h) - \mathcal{R}^*_{\mathsf{L}^{\rm abs}_{0-1}}(\mathcal{H}_{\rm lin}) \leq \frac{\mathcal{R}_{\mathsf{L}_{\Phi_{\rm hinge}}}(h) - \mathcal{R}^*_{\mathsf{L}_{\Phi_{\rm hinge}}}(\mathcal{H}_{\rm lin}) + \mathcal{M}_{\mathsf{L}_{\Phi_{\rm hinge}}}(\mathcal{H}_{\rm lin})}{\min\{W\gamma, 1\}} - \mathcal{M}_{\mathsf{L}^{\rm abs}_{0-1}}(\mathcal{H}_{\rm lin}). \tag{12}$$

Proof for $\mathsf{L}_{\Phi_{\rm exp}}$. For the exponential loss function $\Phi_{\rm exp}(u) := e^{-u}$, for all $h \in \mathcal{H}_{\rm lin}$ and (x,x') such that $\|x-x'\|_p > \gamma$,

$$\begin{aligned}
\mathcal{C}_{\mathsf{L}_{\Phi_{\rm exp}}}(h,x,x') &= \eta(x,x')\mathsf{L}_{\Phi_{\rm exp}}(h(x') - h(x)) + (1-\eta(x,x'))\mathsf{L}_{\Phi_{\rm exp}}(h(x) - h(x')) \\
&= \eta(x,x')e^{-h(x')+h(x)} + (1-\eta(x,x'))e^{h(x')-h(x)}.
\end{aligned}$$

Then,

$$
\begin{aligned}
&\mathcal{C}^*_{\mathsf{L}_{\Phi_{\exp}},\mathcal{H}_{\mathrm{lin}}}(x,x')\\
&= \inf_{h\in\mathcal{H}_{\mathrm{lin}}} \mathcal{C}_{\mathsf{L}_{\Phi_{\exp}}}(h,x,x')\\
&= \begin{cases} 2\sqrt{\eta(x,x')(1-\eta(x,x'))} & \frac{1}{2}\left|\log\frac{\eta(x,x')}{1-\eta(x,x')}\right| \le W\|x-x'\|_p\\ \max\{\eta(x,x'),1-\eta(x,x')\}e^{-W\|x-x'\|_p} + \min\{\eta(x,x'),1-\eta(x,x')\}e^{W\|x-x'\|_p} & \frac{1}{2}\left|\log\frac{\eta(x,x')}{1-\eta(x,x')}\right| > W\|x-x'\|_p. \end{cases}
\end{aligned}
$$

The $\left(\mathsf{L}_{\Phi_{\exp}},\mathcal{H}_{\mathrm{lin}}\right)$-minimizability gap is:

$$
\begin{aligned}
\mathcal{M}_{\mathsf{L}_{\Phi_{\exp}}}(\mathcal{H}_{\mathrm{lin}}) &= \mathcal{R}^*_{\mathsf{L}_{\Phi_{\exp}}}(\mathcal{H}_{\mathrm{lin}}) - \mathbb{E}_{(X,X')}\left[\mathcal{C}^*_{\mathsf{L}_{\Phi_{\exp}},\mathcal{H}_{\mathrm{lin}}}(x,x')\right]\\
&= \mathcal{R}^*_{\mathsf{L}_{\Phi_{\exp}}}(\mathcal{H}_{\mathrm{lin}}) - \mathbb{E}_{(X,X')}\left[2\sqrt{\eta(x,x')(1-\eta(x,x'))}\mathbb{1}_{\frac{1}{2}\left|\log\frac{\eta(x,x')}{1-\eta(x,x')}\right|\le W\|x-x'\|_p}\right]\\
&\quad - \mathbb{E}_{(X,X')}\left[\max\{\eta(x,x'),1-\eta(x,x')\}e^{-W\|x-x'\|_p}\mathbb{1}_{\frac{1}{2}\left|\log\frac{\eta(x,x')}{1-\eta(x,x')}\right|>W\|x-x'\|_p}\right]\\
&\quad - \mathbb{E}_{(X,X')}\left[\min\{\eta(x,x'),1-\eta(x,x')\}e^{W\|x-x'\|_p}\mathbb{1}_{\frac{1}{2}\left|\log\frac{\eta(x,x')}{1-\eta(x,x')}\right|>W\|x-x'\|_p}\right].
\end{aligned}
\tag{13}
$$

Therefore, $\forall h \in \overline{\mathcal{H}}_{\mathrm{lin}}(x,x')$,

$$
\begin{aligned}
&\Delta\mathcal{C}_{\mathsf{L}_{\Phi_{\exp}},\mathcal{H}_{\mathrm{lin}}}(h,x,x')\\
&\ge \inf_{h\in\overline{\mathcal{H}}_{\mathrm{lin}}(x,x')} \mathcal{C}_{\mathsf{L}_{\Phi_{\exp}}}(h,x,x') - \mathcal{C}^*_{\mathsf{L}_{\Phi_{\exp}},\mathcal{H}_{\mathrm{lin}}}(x,x')\\
&= \eta(x,x')e^{-0} + (1-\eta(x,x'))e^{0} - \mathcal{C}^*_{\mathsf{L}_{\Phi_{\exp}},\mathcal{H}_{\mathrm{lin}}}(x,x')\\
&= \begin{cases} 1-2\sqrt{\eta(x,x')(1-\eta(x,x'))} & \frac{1}{2}\left|\log\frac{\eta(x,x')}{1-\eta(x,x')}\right| \le W\|x-x'\|_p\\ 1-\max\{\eta(x,x'),1-\eta(x,x')\}e^{-W\|x-x'\|_p} - \min\{\eta(x,x'),1-\eta(x,x')\}e^{W\|x-x'\|_p} & \frac{1}{2}\left|\log\frac{\eta(x,x')}{1-\eta(x,x')}\right| > W\|x-x'\|_p \end{cases}\\
&\ge \begin{cases} 1-2\sqrt{\eta(x,x')(1-\eta(x,x'))} & \frac{1}{2}\left|\log\frac{\eta(x,x')}{1-\eta(x,x')}\right| \le W\gamma\\ 1-\max\{\eta(x,x'),1-\eta(x,x')\}e^{-W\gamma} - \min\{\eta(x,x'),1-\eta(x,x')\}e^{W\gamma} & \frac{1}{2}\left|\log\frac{\eta(x,x')}{1-\eta(x,x')}\right| > W\gamma \end{cases}\\
&= \Psi_{\exp}(|2\eta(x,x')-1|),
\end{aligned}
$$

where $\Psi_{\exp}$ is the increasing and convex function on $[0,1]$ defined by

$$
\forall t\in[0,1],\quad \Psi_{\exp}(t) = \begin{cases} 1-\sqrt{1-t^2}, & t\le \frac{e^{2W\gamma}-1}{e^{2W\gamma}+1}\\ 1-\frac{t+1}{2}e^{-W\gamma} - \frac{1-t}{2}e^{W\gamma}, & t > \frac{e^{2W\gamma}-1}{e^{2W\gamma}+1} \end{cases}
$$

which implies that for any $h\in\mathcal{H}_{\mathrm{lin}}$ and (x,x') such that $\|x-x'\|_p > \gamma$,

$$
\Delta\mathcal{C}_{\mathsf{L}_{\Phi_{\exp}},\mathcal{H}_{\mathrm{lin}}}(h,x,x') \ge \Psi_{\exp}\left(\Delta\mathcal{C}_{\mathsf{L}^{\mathrm{abs}}_{0-1},\mathcal{H}_{\mathrm{lin}}}(h,x,x')\right).
$$

To simplify the expression, using the fact that

$$
\begin{aligned}
1-\sqrt{1-t^2} &\ge \frac{t^2}{2},\\
1-\frac{t+1}{2}e^{-W\gamma} - \frac{1-t}{2}e^{W\gamma} &= 1-\frac{e^{W\gamma}}{2} - \frac{e^{-W\gamma}}{2} + \frac{e^{W\gamma}-e^{-W\gamma}}{2}t,
\end{aligned}
$$

$\Psi_{\exp}$ can be lower bounded by

$$
\widetilde{\Psi}_{\exp}(t) = \begin{cases} \frac{t^2}{2}, & t\le \frac{e^{2W\gamma}-1}{e^{2W\gamma}+1}\\ \frac{1}{2}\left(\frac{e^{2W\gamma}-1}{e^{2W\gamma}+1}\right)t, & t > \frac{e^{2W\gamma}-1}{e^{2W\gamma}+1}. \end{cases}
$$

Thus, we adopt an upper bound of Ψ^{-1} as follows:

$$\Gamma_{\Phi_{\exp}}(t) = \widetilde{\Psi}_{\exp}^{-1}(t) = \begin{cases} \sqrt{2t}, & t \leq \frac{1}{2}\left(\frac{e^{2W\gamma}-1}{e^{2W\gamma}+1}\right)^2 \\ 2\left(\frac{e^{2W\gamma}+1}{e^{2W\gamma}-1}\right)t, & t > \frac{1}{2}\left(\frac{e^{2W\gamma}-1}{e^{2W\gamma}+1}\right)^2 \end{cases}$$
$$= \max\left\{\sqrt{2t}, 2\left(\frac{e^{2W\gamma}+1}{e^{2W\gamma}-1}\right)t\right\}.$$

Thus, by Theorem B.1 or Theorem B.2, setting $\epsilon = 0$ yields the $\mathcal{H}_{\rm lin}$-consistency bound for $\mathsf{L}_{\Phi_{\exp}}$, valid for all $h \in \mathcal{H}_{\rm lin}$:

$$\mathcal{R}_{\mathsf{L}_{0-1}^{\rm abs}}(h) - \mathcal{R}^*_{\mathsf{L}_{0-1}^{\rm abs}}(\mathcal{H}_{\rm lin}) \leq \Gamma_{\Phi_{\exp}}\left(\mathcal{R}_{\mathsf{L}_{\Phi_{\exp}}}(h) - \mathcal{R}^*_{\mathsf{L}_{\Phi_{\exp}}}(\mathcal{H}_{\rm lin}) + \mathcal{M}_{\mathsf{L}_{\Phi_{\exp}}}(\mathcal{H}_{\rm lin})\right) - \mathcal{M}_{\mathsf{L}_{0-1}^{\rm abs}}(\mathcal{H}_{\rm lin}). \tag{14}$$

where $\Gamma_{\Phi_{\exp}}(t) = \max\left\{\sqrt{2t}, 2\left(\frac{e^{2W\gamma}+1}{e^{2W\gamma}-1}\right)t\right\}$.

Proof for $\mathsf{L}_{\Phi_{\rm sig}}$. For the sigmoid loss function $\Phi_{\rm sig}(u) := 1 - \tanh(ku)$, $k > 0$, for all $h \in \mathcal{H}_{\rm lin}$ and (x, x') such that $\|x - x'\|_p > \gamma$,

$$\begin{aligned}
&\mathcal{C}_{\mathsf{L}_{\Phi_{\rm sig}}}(h, x, x') \\
&= \eta(x, x')\mathsf{L}_{\Phi_{\rm sig}}(h(x') - h(x)) + (1 - \eta(x, x'))\mathsf{L}_{\Phi_{\rm sig}}(h(x) - h(x')) \\
&= \eta(x, x')(1 - \tanh(k[h(x') - h(x)])) + (1 - \eta(x, x'))(1 + \tanh(k[h(x') - h(x)])).
\end{aligned}$$

Then,

$$\mathcal{C}^*_{\mathsf{L}_{\Phi_{\rm sig}}}(\mathcal{H}_{\rm lin})(x, x') = \inf_{h \in \mathcal{H}_{\rm lin}} \mathcal{C}_{\mathsf{L}_{\Phi_{\rm sig}}}(h, x, x') = 1 - |1 - 2\eta(x, x')| \tanh\left(kW\|x - x'\|_p\right).$$

The $\left(\mathsf{L}_{\Phi_{\rm sig}}, \mathcal{H}_{\rm lin}\right)$-minimizability gap is:

$$\begin{aligned}
\mathcal{M}_{\mathsf{L}_{\Phi_{\rm sig}}}(\mathcal{H}_{\rm lin}) &= \mathcal{R}^*_{\mathsf{L}_{\Phi_{\rm sig}}}(\mathcal{H}_{\rm lin}) - \mathbb{E}_{(X,X')}\left[\mathcal{C}^*_{\mathsf{L}_{\Phi_{\rm sig}}, \mathcal{H}_{\rm lin}}(x, x')\right] \\
&= \mathcal{R}^*_{\mathsf{L}_{\Phi_{\rm sig}}}(\mathcal{H}_{\rm lin}) - \mathbb{E}_{(X,X')}\left[1 - |1 - 2\eta(x, x')| \tanh\left(kW\|x - x'\|_p\right)\right].
\end{aligned} \tag{15}$$

Therefore, $\forall h \in \overline{\mathcal{H}}_{\rm lin}(x, x')$,

$$\begin{aligned}
\Delta\mathcal{C}_{\mathsf{L}_{\Phi_{\rm sig}}, \mathcal{H}_{\rm lin}}(h, x, x') &\geq \inf_{h \in \overline{\mathcal{H}}_{\rm lin}(x, x')} \mathcal{C}_{\mathsf{L}_{\Phi_{\rm sig}}}(h, x, x') - \mathcal{C}^*_{\mathsf{L}_{\Phi_{\rm sig}}, \mathcal{H}_{\rm lin}}(x, x') \\
&= 1 - |1 - 2\eta(x, x')| \tanh(0) - \mathcal{C}^*_{\mathsf{L}_{\Phi_{\rm sig}}, \mathcal{H}_{\rm lin}}(x, x') \\
&= |1 - 2\eta(x, x')| \tanh\left(kW\|x - x'\|_p\right) \\
&\geq |1 - 2\eta(x, x')| \tanh(kW\gamma)
\end{aligned}$$

which implies that for any $h \in \mathcal{H}_{\rm lin}$ and (x, x') such that $\|x - x'\|_p > \gamma$,

$$\Delta\mathcal{C}_{\mathsf{L}_{\Phi_{\rm sig}}, \mathcal{H}_{\rm lin}}(h, x, x') \geq \tanh(kW\gamma)\Delta\mathcal{C}_{\mathsf{L}_{0-1}^{\rm abs}, \mathcal{H}_{\rm lin}}(h, x, x').$$

Thus, by Theorem B.1 or Theorem B.2, setting $\epsilon = 0$ yields the $\mathcal{H}_{\rm lin}$-consistency bound for $\mathsf{L}_{\Phi_{\rm sig}}$, valid for all $h \in \mathcal{H}_{\rm lin}$:

$$\mathcal{R}_{\mathsf{L}_{0-1}^{\rm abs}}(h) - \mathcal{R}^*_{\mathsf{L}_{0-1}^{\rm abs}}(\mathcal{H}_{\rm lin}) \leq \frac{\mathcal{R}_{\mathsf{L}_{\Phi_{\rm sig}}}(h) - \mathcal{R}^*_{\mathsf{L}_{\Phi_{\rm sig}}}(\mathcal{H}_{\rm lin}) + \mathcal{M}_{\mathsf{L}_{\Phi_{\rm sig}}}(\mathcal{H}_{\rm lin})}{\tanh(kW\gamma)} - \mathcal{M}_{\mathsf{L}_{0-1}^{\rm abs}}(\mathcal{H}_{\rm lin}). \tag{16}$$

□

D. $\mathcal{H}$ - consistency bounds for bipartite ranking with abstention (Proof of Theorem 3.1)

We first characterize the minimal conditional $\widetilde{\mathsf{L}}_{0-1}^{\rm abs}$-risk and the calibration gap of $\widetilde{\mathsf{L}}_{0-1}^{\rm abs}$ for a broad class of hypothesis sets. We let $\widetilde{\mathcal{H}}(x, x') = \{h \in \mathcal{H} : (h(x) - h(x'))(\eta(x) - \eta(x')) < 0\}$ and $\mathring{\mathcal{H}}(x, x') = \{h \in \mathcal{H} : h(x) = h(x')\}$ for convenience.

Lemma D.1. *Assume that $\mathcal{H}$ is regular for bipartite ranking. Then, the minimal conditional $\widetilde{\mathsf{L}}_{0-1}^{\text{abs}}$-risk is*

$$\mathcal{C}^*_{\widetilde{\mathsf{L}}_{0-1}^{\text{abs}}}(\mathcal{H},x,x') = \min\{\eta(x)(1-\eta(x')),\eta(x')(1-\eta(x))\}\mathbb{1}_{\|x-x'\|>\gamma} + c\,\mathbb{1}_{|x-x'|\leq\gamma}.$$

The calibration gap of $\widetilde{\mathsf{L}}_{0-1}^{\text{abs}}$ can be characterized as

$$\Delta\mathcal{C}_{\widetilde{\mathsf{L}}_{0-1}^{\text{abs}},\mathcal{H}}(h,x,x') = |\eta(x)-\eta(x')|\mathbb{1}_{h\in\widetilde{\mathcal{H}}(x,x')}\mathbb{1}_{\|x-x'\|>\gamma} + \frac{1}{2}|\eta(x)-\eta(x')|\mathbb{1}_{h\in\mathring{\mathcal{H}}(x,x')}\mathbb{1}_{\|x-x'\|>\gamma}.$$

Proof. By the definition, the conditional $\widetilde{\mathsf{L}}_{0-1}^{\text{abs}}$-risk is

$$\begin{aligned}&\mathcal{C}_{\widetilde{\mathsf{L}}_{0-1}^{\text{abs}}}(h,x,x')\\&=\left(\eta(x)(1-\eta(x'))\left[\mathbb{1}_{h(x)-h(x')<0}+\frac{1}{2}\mathbb{1}_{h(x)=h(x')}\right]+\eta(x')(1-\eta(x))\left[\mathbb{1}_{h(x)-h(x')>0}+\frac{1}{2}\mathbb{1}_{h(x)=h(x')}\right]\right)\mathbb{1}_{\|x-x'\|>\gamma}+c\,\mathbb{1}_{|x-x'|\leq\gamma}.\end{aligned}$$

For any (x,x') such that $\|x-x'\|\leq\gamma$ and $h\in\mathcal{H}$, $\mathcal{C}_{\widetilde{\mathsf{L}}_{0-1}^{\text{abs}}}(h,x,x) = \mathcal{C}^*_{\widetilde{\mathsf{L}}_{0-1}^{\text{abs}}}(\mathcal{H},x,x) = c$. For any (x,x') such that $\|x-x'\|>\gamma$, by the assumption, there exists $h^*\in\mathcal{H}$ such that

$$(h^*(x)-h^*(x'))(\eta(x)-\eta(x'))\mathbb{1}_{\eta(x)\neq\eta(x')}>0.$$

Therefore, the optimal conditional $\widetilde{\mathsf{L}}_{0-1}^{\text{abs}}$-risk can be characterized as for any $x,x'\in\mathcal{X}$,

$$\mathcal{C}^*_{\widetilde{\mathsf{L}}_{0-1}^{\text{abs}}}(\mathcal{H},x,x') = \mathcal{C}_{\widetilde{\mathsf{L}}_{0-1}^{\text{abs}}}(h^*,x,x') = \min\{\eta(x)(1-\eta(x')),\eta(x')(1-\eta(x))\}\mathbb{1}_{\|x-x'\|>\gamma}+c\,\mathbb{1}_{|x-x'|\leq\gamma}.$$

which proves the first part of lemma. By the definition, for any (x,x') such that $\|x-x'\|\leq\gamma$ and $h\in\mathcal{H}$, $\Delta\mathcal{C}_{\widetilde{\mathsf{L}}_{0-1}^{\text{abs}},\mathcal{H}}(h,x,x') = \mathcal{C}_{\widetilde{\mathsf{L}}_{0-1}^{\text{abs}}}(h,x,x') - \mathcal{C}^*_{\widetilde{\mathsf{L}}_{0-1}^{\text{abs}}}(\mathcal{H},x,x') = 0.$. For any (x,x') such that $\|x-x'\|>\gamma$ and $h\in\mathcal{H}$,

$$\begin{aligned}\Delta\mathcal{C}_{\widetilde{\mathsf{L}}_{0-1}^{\text{abs}},\mathcal{H}}(h,x,x') &= \mathcal{C}_{\widetilde{\mathsf{L}}_{0-1}^{\text{abs}}}(h,x,x') - \mathcal{C}^*_{\widetilde{\mathsf{L}}_{0-1}^{\text{abs}}}(\mathcal{H},x,x')\\&=\eta(x)(1-\eta(x'))\left[\mathbb{1}_{h(x)-h(x')<0}+\frac{1}{2}\mathbb{1}_{h(x)=h(x')}\right]\\&\quad+\eta(x')(1-\eta(x))\left[\mathbb{1}_{h(x)-h(x')>0}+\frac{1}{2}\mathbb{1}_{h(x)=h(x')}\right]\\&\quad-\min\{\eta(x)(1-\eta(x')),\eta(x')(1-\eta(x))\}\\&=\begin{cases}|\eta(x)(1-\eta(x'))-\eta(x')(1-\eta(x))|, & h\in\widetilde{\mathcal{H}}(x,x'),\\ \frac{1}{2}|\eta(x)(1-\eta(x'))-\eta(x')(1-\eta(x))|, & h\in\mathring{\mathcal{H}}(x,x'),\\ 0, & \text{otherwise.}\end{cases}\\&=\begin{cases}|\eta(x)-\eta(x')|, & h\in\widetilde{\mathcal{H}}(x,x'),\\ \frac{1}{2}|\eta(x)-\eta(x')|, & h\in\mathring{\mathcal{H}}(x,x'),\\ 0, & \text{otherwise.}\end{cases}\end{aligned}$$

This leads to

$$\left\langle\Delta\mathcal{C}_{\widetilde{\mathsf{L}}_{0-1}^{\text{abs}},\mathcal{H}}(h,x,x')\right\rangle_\epsilon = \langle|\eta(x)-\eta(x')|\rangle_\epsilon\mathbb{1}_{h\in\widetilde{\mathcal{H}}(x,x')}\mathbb{1}_{\|x-x'\|>\gamma} + \left\langle\frac{1}{2}|\eta(x)-\eta(x')|\right\rangle_\epsilon\mathbb{1}_{h\in\mathring{\mathcal{H}}(x,x')}\mathbb{1}_{\|x-x'\|>\gamma}.$$

□

Theorem 3.1 (**$\mathcal{H}$-consistency bounds for bipartite abstention losses**). *Let $\mathcal{H}$ be $\mathcal{H}_{\text{lin}}$ or $\mathcal{H}_{\text{NN}}$. Then, for any $h\in\mathcal{H}$ and any distribution,*

$$\mathcal{R}_{\widetilde{\mathsf{L}}_{0-1}^{\text{abs}}}(h) - \mathcal{R}^*_{\widetilde{\mathsf{L}}_{0-1}^{\text{abs}}}(\mathcal{H}) + \mathcal{M}_{\widetilde{\mathsf{L}}_{0-1}^{\text{abs}}}(\mathcal{H}) \leq \Gamma_\Phi\Big(\mathcal{R}_{\widetilde{\mathsf{L}}_\Phi}(h) - \mathcal{R}^*_{\widetilde{\mathsf{L}}_\Phi}(\mathcal{H}) + \mathcal{M}_{\widetilde{\mathsf{L}}_\Phi}(\mathcal{H})\Big)$$

where $\Gamma_\Phi(t)$ equals $\frac{t}{\min\{W\gamma,1\}}$, $\max\left\{\sqrt{t},\left(\frac{e^{2W\gamma}+1}{e^{2W\gamma}-1}\right)t\right\}$ and $\frac{t}{\tanh(kW\gamma)}$ for Φ equals Φ_{hinge}, Φ_{exp} and Φ_{sig} respectively. W is replaced by ΛW for $\mathcal{H}=\mathcal{H}_{\text{NN}}$.

Proof. Since $\mathcal{H}_{\text{lin}}$ and $\mathcal{H}_{\text{NN}}$ satisfy the condition of Lemma D.1, by Lemma D.1 the $\left(\widetilde{\mathsf{L}}^{\text{abs}}_{0-1}, \mathcal{H}_{\text{lin}}\right)$-minimizability gap and the $\left(\widetilde{\mathsf{L}}^{\text{abs}}_{0-1}, \mathcal{H}_{\text{NN}}\right)$-minimizability gap can be expressed as follows:

$$\mathcal{M}_{\widetilde{\mathsf{L}}^{\text{abs}}_{0-1}}(\mathcal{H}_{\text{lin}}) = \mathcal{R}^*_{\widetilde{\mathsf{L}}^{\text{abs}}_{0-1}}(\mathcal{H}_{\text{lin}}) - \mathbb{E}_{(X,X')}\big[\min\{\eta(x)(1-\eta(x')), \eta(x')(1-\eta(x))\}\mathbb{1}_{\|x-x'\|>\gamma} + c\,\mathbb{1}_{|x-x'|\le\gamma}\big]$$
$$\mathcal{M}_{\widetilde{\mathsf{L}}^{\text{abs}}_{0-1}}(\mathcal{H}_{\text{NN}}) = \mathcal{R}^*_{\widetilde{\mathsf{L}}^{\text{abs}}_{0-1}}(\mathcal{H}_{\text{NN}}) - \mathbb{E}_{(X,X')}\big[\min\{\eta(x)(1-\eta(x')), \eta(x')(1-\eta(x))\}\mathbb{1}_{\|x-x'\|>\gamma} + c\,\mathbb{1}_{|x-x'|\le\gamma}\big].$$

By the definition of $\mathcal{H}_{\text{lin}}$ and $\mathcal{H}_{\text{NN}}$, for any $(x,x') \in \mathcal{X}\times\mathcal{X}$, $\left\{h(x') - h(x) \mid h \in \mathcal{H}_{\text{lin}}\right\} = \left[-W\|x-x'\|_p, W\|x-x'\|_p\right]$ and $\left\{h(x') - h(x) \mid h \in \mathcal{H}_{\text{NN}}\right\} = \left[-\Lambda W\|x-x'\|_p, \Lambda W\|x-x'\|_p\right]$. In the following, we will prove the bounds for $\mathcal{H}_{\text{lin}}$. Similar proofs with B replaced by ΛB hold for $\mathcal{H}_{\text{NN}}$.

Proof for $\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}}}$. For the hinge loss function $\Phi_{\text{hinge}}(u) := \max\{0, 1-u\}$, for all $h \in \mathcal{H}_{\text{lin}}$ and (x,x') such that $\|x-x'\|_p > \gamma$,

$$\begin{aligned}
&\mathcal{C}_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}}}}(h,x,x')\\
&= \eta(x)(1-\eta(x'))\Phi_{\text{hinge}}(h(x)-h(x')) + \eta(x')(1-\eta(x))\Phi_{\text{hinge}}(h(x')-h(x))\\
&= \eta(x)(1-\eta(x'))\max\{0, 1-h(x)+h(x')\} + \eta(x')(1-\eta(x))\max\{0, 1+h(x)-h(x')\}.
\end{aligned}$$

Then,

$$\mathcal{C}^*_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}}},\mathcal{H}_{\text{lin}}}(x,x') = \inf_{h\in\mathcal{H}_{\text{lin}}} \mathcal{C}_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}}}}(h,x,x') = \eta(x)(1-\eta(x')) + \eta(x')(1-\eta(x)) - |\eta(x)-\eta(x')|\min\left\{W\|x-x'\|_p, 1\right\}.$$

The $\left(\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}}}, \mathcal{H}_{\text{lin}}\right)$-minimizability gap is

$$\begin{aligned}
&\mathcal{M}_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}}}}(\mathcal{H}_{\text{lin}})\\
&= \mathcal{R}^*_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}}}}(\mathcal{H}_{\text{lin}}) - \mathbb{E}_{(X,X')}\left[\mathcal{C}^*_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}}},\mathcal{H}_{\text{lin}}}(x,x')\right]\\
&= \mathcal{R}^*_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}}}}(\mathcal{H}_{\text{lin}}) - \mathbb{E}_{(X,X')}\left[\eta(x)(1-\eta(x')) + \eta(x')(1-\eta(x)) - |\eta(x)-\eta(x')|\min\left\{W\|x-x'\|_p, 1\right\}\right].
\end{aligned} \tag{17}$$

Therefore, $\forall h \in \widetilde{\mathcal{H}}_{\text{lin}}(x,x') \cup \mathring{\mathcal{H}}_{\text{lin}}(x,x')$,

$$\begin{aligned}
&\Delta\mathcal{C}_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}}},\mathcal{H}_{\text{lin}}}(h,x,x')\\
&\ge \inf_{h\in\widetilde{\mathcal{H}}_{\text{lin}}(x,x')\cup\mathring{\mathcal{H}}_{\text{lin}}(x,x')} \mathcal{C}_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}}}}(h,x,x') - \mathcal{C}^*_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}}},\mathcal{H}_{\text{lin}}}(x,x')\\
&= \eta(x)(1-\eta(x'))\max\{0,1-0\} + \eta(x')(1-\eta(x))\max\{0,1+0\} - \mathcal{C}^*_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}}},\mathcal{H}_{\text{lin}}}(x,x')\\
&= |\eta(x)-\eta(x')|\min\left\{W\|x-x'\|_p, 1\right\}\\
&\ge |\eta(x)-\eta(x')|\min\{W\gamma, 1\}
\end{aligned}$$

which implies that for any $h \in \mathcal{H}_{\text{lin}}$ and (x,x') such that $\|x-x'\|_p > \gamma$,

$$\Delta\mathcal{C}_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}}},\mathcal{H}_{\text{lin}}}(h,x,x') \ge \min\{W\gamma,1\}\Delta\mathcal{C}_{\widetilde{\mathsf{L}}^{\text{abs}}_{0-1},\mathcal{H}}(h,x,x').$$

Thus, by Theorem B.1 or Theorem B.2, setting $\epsilon = 0$ yields the $\mathcal{H}_{\text{lin}}$-consistency bound for $\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}}}$, valid for all $h \in \mathcal{H}_{\text{lin}}$:

$$\mathcal{R}_{\widetilde{\mathsf{L}}^{\text{abs}}_{0-1}}(h) - \mathcal{R}^*_{\widetilde{\mathsf{L}}^{\text{abs}}_{0-1}}(\mathcal{H}_{\text{lin}}) \le \frac{\mathcal{R}_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}}}}(h) - \mathcal{R}^*_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}}}}(\mathcal{H}_{\text{lin}}) + \mathcal{M}_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}}}}(\mathcal{H}_{\text{lin}})}{\min\{W\gamma,1\}} - \mathcal{M}_{\widetilde{\mathsf{L}}^{\text{abs}}_{0-1}}(\mathcal{H}_{\text{lin}}). \tag{18}$$

Proof for $\widetilde{\mathsf{L}}_{\Phi_{\text{exp}}}$. For the exponential loss function $\Phi_{\text{exp}}(u) := e^{-u}$, for all $h \in \mathcal{H}_{\text{lin}}$ and (x,x') such that $\|x-x'\|_p > \gamma$,

$$\begin{aligned}
&\mathcal{C}_{\widetilde{\mathsf{L}}_{\Phi_{\text{exp}}}}(h,x,x')\\
&= \eta(x)(1-\eta(x'))\Phi_{\text{exp}}(h(x)-h(x')) + \eta(x')(1-\eta(x))\Phi_{\text{exp}}(h(x')-h(x))\\
&= \eta(x)(1-\eta(x'))e^{-h(x)+h(x')} + \eta(x')(1-\eta(x))e^{h(x)-h(x')}.
\end{aligned}$$

Then,

$$
\begin{aligned}
&\mathcal{C}^*_{\widetilde{\mathsf{L}}_{\Phi_{\exp}},\mathcal{H}_{\rm lin}}(x,x')\\
&= \inf_{h\in\mathcal{H}_{\rm lin}} \mathcal{C}_{\widetilde{\mathsf{L}}_{\Phi_{\exp}}}(h,x,x')\\
&= \begin{cases}
2\sqrt{\eta(x)\eta(x')(1-\eta(x))(1-\eta(x'))}\\
\text{if } \frac{1}{2}\left|\log \frac{\eta(x)(1-\eta(x'))}{\eta(x')(1-\eta(x))}\right| \le W\|x-x'\|_p\\
\max\{\eta(x)(1-\eta(x')),\eta(x')(1-\eta(x))\}e^{-W\|x-x'\|_p} + \min\{\eta(x)(1-\eta(x')),\eta(x')(1-\eta(x))\}e^{W\|x-x'\|_p}\\
\text{if } \frac{1}{2}\left|\log \frac{\eta(x)(1-\eta(x'))}{\eta(x')(1-\eta(x))}\right| > W\|x-x'\|_p.
\end{cases}
\end{aligned}
$$

The $\left(\widetilde{\mathsf{L}}_{\Phi_{\exp}},\mathcal{H}_{\rm lin}\right)$-minimizability gap is:

$$
\begin{aligned}
\mathcal{M}_{\widetilde{\mathsf{L}}_{\Phi_{\exp}}}(\mathcal{H}_{\rm lin}) &= \mathcal{R}^*_{\widetilde{\mathsf{L}}_{\Phi_{\exp}}}(\mathcal{H}_{\rm lin}) - \mathbb{E}_{(X,X')}\left[\mathcal{C}^*_{\widetilde{\mathsf{L}}_{\Phi_{\exp}},\mathcal{H}_{\rm lin}}(x,x')\right]\\
&= \mathcal{R}^*_{\widetilde{\mathsf{L}}_{\Phi_{\exp}}}(\mathcal{H}_{\rm lin}) - \mathbb{E}_{(X,X')}\left[2\sqrt{\eta(x)\eta(x')(1-\eta(x))(1-\eta(x'))}\mathbb{1}_{\frac{1}{2}\left|\log \frac{\eta(x)(1-\eta(x'))}{\eta(x')(1-\eta(x))}\right|\le W\|x-x'\|_p}\right]\\
&\quad - \mathbb{E}_{(X,X')}\left[[\max\{\eta(x),\eta(x')\}-\eta(x)\eta(x')]e^{-W\|x-x'\|_p}\mathbb{1}_{\frac{1}{2}\left|\log \frac{\eta(x)(1-\eta(x'))}{\eta(x')(1-\eta(x))}\right|> W\|x-x'\|_p}\right]\\
&\quad - \mathbb{E}_{(X,X')}\left[[\min\{\eta(x),\eta(x')\}-\eta(x)\eta(x')]e^{W\|x-x'\|_p}\mathbb{1}_{\frac{1}{2}\left|\log \frac{\eta(x)(1-\eta(x'))}{\eta(x')(1-\eta(x))}\right|> W\|x-x'\|_p}\right].
\end{aligned} \tag{19}
$$

Therefore, $\forall h \in \widetilde{\mathcal{H}}_{\rm lin}(x,x') \cup \mathring{\mathcal{H}}_{\rm lin}(x,x')$,

$$
\begin{aligned}
&\Delta\mathcal{C}_{\widetilde{\mathsf{L}}_{\Phi_{\exp}},\mathcal{H}_{\rm lin}}(h,x,x')\\
&\ge \inf_{h\in\widetilde{\mathcal{H}}_{\rm lin}(x,x')\cup\mathring{\mathcal{H}}_{\rm lin}(x,x')} \mathcal{C}_{\widetilde{\mathsf{L}}_{\Phi_{\exp}}}(h,x,x') - \mathcal{C}^*_{\widetilde{\mathsf{L}}_{\Phi_{\exp}},\mathcal{H}_{\rm lin}}(x,x')\\
&= \eta(x)(1-\eta(x'))e^{-0} + \eta(x')(1-\eta(x))e^{0} - \mathcal{C}^*_{\widetilde{\mathsf{L}}_{\Phi_{\exp}},\mathcal{H}_{\rm lin}}(x,x')\\
&= \begin{cases}
\eta(x)(1-\eta(x')) + \eta(x')(1-\eta(x)) - 2\sqrt{\eta(x)\eta(x')(1-\eta(x))(1-\eta(x'))}\\
\text{if } \frac{1}{2}\left|\log \frac{\eta(x)(1-\eta(x'))}{\eta(x')(1-\eta(x))}\right| \le W\|x-x'\|_p\\
[\max\{\eta(x),\eta(x')\}-\eta(x)\eta(x')]\left(1-e^{-W\|x-x'\|_p}\right) + [\min\{\eta(x),\eta(x')\}-\eta(x)\eta(x')]\left(1-e^{W\|x-x'\|_p}\right)\\
\text{if } \frac{1}{2}\left|\log \frac{\eta(x)(1-\eta(x'))}{\eta(x')(1-\eta(x))}\right| > W\|x-x'\|_p
\end{cases}\\
&\ge \begin{cases}
\eta(x)(1-\eta(x')) + \eta(x')(1-\eta(x)) - 2\sqrt{\eta(x)\eta(x')(1-\eta(x))(1-\eta(x'))}\\
\text{if } \frac{1}{2}\left|\log \frac{\eta(x)(1-\eta(x'))}{\eta(x')(1-\eta(x))}\right| \le W\gamma\\
[\max\{\eta(x),\eta(x')\}-\eta(x)\eta(x')]\left(1-e^{-W\gamma}\right) + [\min\{\eta(x),\eta(x')\}-\eta(x)\eta(x')]\left(1-e^{W\gamma}\right)\\
\text{if } \frac{1}{2}\left|\log \frac{\eta(x)(1-\eta(x'))}{\eta(x')(1-\eta(x))}\right| > W\gamma
\end{cases}\\
&= \begin{cases}
\left(\frac{\eta(x)(1-\eta(x'))-\eta(x')(1-\eta(x))}{\sqrt{\eta(x)(1-\eta(x'))}+\sqrt{\eta(x')(1-\eta(x))}}\right)^2 & \text{if } \frac{1}{2}\left|\log \frac{\eta(x)(1-\eta(x'))}{\eta(x')(1-\eta(x))}\right| \le W\gamma\\
\frac{\eta(x)(1-\eta(x'))+\eta(x')(1-\eta(x))}{2}\left(2-e^{-W\gamma}-e^{W\gamma}\right) + \frac{1}{2}|\eta(x)-\eta(x')|\left(e^{W\gamma}-e^{-W\gamma}\right) & \text{if } \frac{1}{2}\left|\log \frac{\eta(x)(1-\eta(x'))}{\eta(x')(1-\eta(x))}\right| > W\gamma
\end{cases}\\
&\ge \min\left\{(\eta(x)-\eta(x'))^2, \left(\frac{e^{2W\gamma}+1}{e^{2W\gamma}-1}\right)|\eta(x)-\eta(x')|\right\}
\end{aligned}
$$

which implies that for any $h \in \mathcal{H}_{\rm lin}$ and (x,x') such that $\|x-x'\|_p > \gamma$,

$$
\Delta\mathcal{C}_{\widetilde{\mathsf{L}}_{\Phi_{\exp}},\mathcal{H}_{\rm lin}}(h,x,x') \ge \Psi_{\exp}\left(\Delta\mathcal{C}_{\widetilde{\mathsf{L}}^{\rm abs}_{0-1},\mathcal{H}}(h,x,x')\right).
$$

where $\Psi_{\exp}$ is the increasing function on $[0,2]$ defined by

$$
\forall t\in[0,1], \quad \Psi_{\exp}(t) = \min\left\{t^2, \left(\frac{e^{2W\gamma}+1}{e^{2W\gamma}-1}\right)t\right\}.
$$

Thus, by Theorem B.1 or Theorem B.2, setting $\epsilon = 0$ yields the $\mathcal{H}_{\mathrm{lin}}$-consistency bound for $\widetilde{\mathsf{L}}_{\Phi_{\exp}}$, valid for all $h \in \mathcal{H}_{\mathrm{lin}}$:

$$\mathcal{R}_{\widetilde{\mathsf{L}}^{\mathrm{abs}}_{0-1}}(h) - \mathcal{R}^*_{\widetilde{\mathsf{L}}^{\mathrm{abs}}_{0-1}}(\mathcal{H}_{\mathrm{lin}}) \le \Gamma_{\Phi_{\exp}}\Big(\mathcal{R}_{\widetilde{\mathsf{L}}_{\Phi_{\exp}}}(h) - \mathcal{R}^*_{\widetilde{\mathsf{L}}_{\Phi_{\exp}}}(\mathcal{H}_{\mathrm{lin}}) + \mathcal{M}_{\widetilde{\mathsf{L}}_{\Phi_{\exp}}}(\mathcal{H}_{\mathrm{lin}})\Big) - \mathcal{M}_{\widetilde{\mathsf{L}}^{\mathrm{abs}}_{0-1}}(\mathcal{H}_{\mathrm{lin}}). \tag{20}$$

where $\Gamma_{\Phi_{\exp}}(t) = \max\Big\{\sqrt{t}, \Big(\frac{e^{2W\gamma}-1}{e^{2W\gamma}+1}\Big)t\Big\}$.

Proof for $\widetilde{\mathsf{L}}_{\Phi_{\mathrm{sig}}}$. For the sigmoid loss function $\Phi_{\mathrm{sig}}(u) := 1 - \tanh(ku),\ k > 0$, for all $h \in \mathcal{H}_{\mathrm{lin}}$ and (x, x') such that $\|x - x'\|_p > \gamma$,

$$\begin{aligned}
&\mathcal{C}_{\widetilde{\mathsf{L}}_{\Phi_{\mathrm{sig}}}}(h, x, x')\\
&= \eta(x)(1-\eta(x'))\Phi_{\mathrm{sig}}(h(x) - h(x')) + \eta(x')(1-\eta(x))\Phi_{\mathrm{sig}}(h(x') - h(x))\\
&= \eta(x)(1-\eta(x'))(1 - \tanh(k[h(x) - h(x')])) + \eta(x')(1-\eta(x))(1 + \tanh(k[h(x) - h(x')]))
\end{aligned}$$

Then,

$$\mathcal{C}^*_{\widetilde{\mathsf{L}}_{\Phi_{\mathrm{sig}}},\mathcal{H}_{\mathrm{lin}}}(x, x') = \inf_{h \in \mathcal{H}_{\mathrm{lin}}} \mathcal{C}_{\widetilde{\mathsf{L}}_{\Phi_{\mathrm{sig}}}}(h, x, x') = \eta(x)(1-\eta(x')) + \eta(x')(1-\eta(x)) - |\eta(x) - \eta(x')| \tanh\big(kW\|x - x'\|_p\big).$$

The $\big(\widetilde{\mathsf{L}}_{\Phi_{\mathrm{sig}}}, \mathcal{H}_{\mathrm{lin}}\big)$-minimizability gap is

$$\mathcal{M}_{\widetilde{\mathsf{L}}_{\Phi_{\mathrm{sig}}}}(\mathcal{H}_{\mathrm{lin}}) = \mathcal{R}^*_{\widetilde{\mathsf{L}}_{\Phi_{\mathrm{sig}}}}(\mathcal{H}_{\mathrm{lin}}) - \mathbb{E}_{(X,X')}\big[\eta(x)(1-\eta(x')) + \eta(x')(1-\eta(x)) - |\eta(x) - \eta(x')| \tanh\big(kW\|x - x'\|_p\big)\big]. \tag{21}$$

Therefore, $\forall h \in \widetilde{\mathcal{H}}_{\mathrm{lin}}(x, x') \cup \mathring{\mathcal{H}}_{\mathrm{lin}}(x, x')$,

$$\begin{aligned}
&\Delta\mathcal{C}_{\widetilde{\mathsf{L}}_{\Phi_{\mathrm{sig}}},\mathcal{H}_{\mathrm{lin}}}(h, x, x')\\
&\ge \inf_{h \in \widetilde{\mathcal{H}}_{\mathrm{lin}}(x,x') \cup \mathring{\mathcal{H}}_{\mathrm{lin}}(x,x')} \mathcal{C}_{\widetilde{\mathsf{L}}_{\Phi_{\mathrm{sig}}}}(h, x, x') - \mathcal{C}^*_{\widetilde{\mathsf{L}}_{\Phi_{\mathrm{sig}}},\mathcal{H}_{\mathrm{lin}}}(x, x')\\
&= \eta(x)(1-\eta(x')) + \eta(x')(1-\eta(x)) - \mathcal{C}^*_{\widetilde{\mathsf{L}}_{\Phi_{\mathrm{sig}}},\mathcal{H}_{\mathrm{lin}}}(x, x')\\
&= |\eta(x) - \eta(x')| \tanh\big(kW\|x - x'\|_p\big)\\
&\ge |\eta(x) - \eta(x')| \tanh(kW\gamma)
\end{aligned}$$

which implies that for any $h \in \mathcal{H}_{\mathrm{lin}}$ and (x, x') such that $\|x - x'\|_p > \gamma$,

$$\Delta\mathcal{C}_{\widetilde{\mathsf{L}}_{\Phi_{\mathrm{sig}}},\mathcal{H}_{\mathrm{lin}}}(h, x, x') \ge \tanh(kW\gamma)\Delta\mathcal{C}_{\widetilde{\mathsf{L}}^{\mathrm{abs}}_{0-1},\mathcal{H}}(h, x, x').$$

Thus, by Theorem B.1 or Theorem B.2, setting $\epsilon = 0$ yields the $\mathcal{H}_{\mathrm{lin}}$-consistency bound for $\widetilde{\mathsf{L}}_{\Phi_{\mathrm{sig}}}$, valid for all $h \in \mathcal{H}_{\mathrm{lin}}$:

$$\mathcal{R}_{\widetilde{\mathsf{L}}^{\mathrm{abs}}_{0-1}}(h) - \mathcal{R}^*_{\widetilde{\mathsf{L}}^{\mathrm{abs}}_{0-1}}(\mathcal{H}_{\mathrm{lin}}) \le \frac{\mathcal{R}_{\widetilde{\mathsf{L}}_{\Phi_{\mathrm{sig}}}}(h) - \mathcal{R}^*_{\widetilde{\mathsf{L}}_{\Phi_{\mathrm{sig}}}}(\mathcal{H}_{\mathrm{lin}}) + \mathcal{M}_{\widetilde{\mathsf{L}}_{\Phi_{\mathrm{sig}}}}(\mathcal{H}_{\mathrm{lin}})}{\tanh(kW\gamma)} - \mathcal{M}_{\widetilde{\mathsf{L}}^{\mathrm{abs}}_{0-1}}(\mathcal{H}_{\mathrm{lin}}). \tag{22}$$

□

E. Negative results for general pairwise ranking (Proof of Theorem 4.1)

Theorem 4.1 (Negative results). *Assume that $\mathcal{X}$ contains an interior point x_0 and that $\mathcal{H}$ is regular for general pairwise ranking, contains 0 and is equicontinuous at x_0. If for some function f that is non-decreasing and continuous at 0, the following bound holds for all $h \in \mathcal{H}$ and any distribution,*

$$\mathcal{R}_{\mathsf{L}_{0-1}}(h) - \mathcal{R}^*_{\mathsf{L}_{0-1}}(\mathcal{H}) \le f\big(\mathcal{R}_{\mathsf{L}_{\Phi}}(h) - \mathcal{R}^*_{\mathsf{L}_{\Phi}}(\mathcal{H})\big),$$

then, $f(t) \ge 1$ for any $t \ge 0$.

Proof. Assume $x_0 \in \mathcal{X}$ is an interior point and $h_0 = 0 \in \mathcal{H}$. By the assumption that x_0 is an interior point and $\mathcal{H}$ is equicontinuous at x_0, for any $\epsilon > 0$, we are able to take $x' \neq x_0 \in \mathcal{X}$ such that $|h(x') - h(x_0)| < \epsilon$ for all $h \in \mathcal{H}$. Consider the distribution that supports on $\{(x_0, x')\}$ with $\eta(x_0, x') = 0$. Then, for any $h \in \mathcal{H}$,

$$\mathcal{R}_{\mathsf{L}_{0-1}}(h) = \mathcal{C}_{\mathsf{L}_{0-1}}(h, x_0, x') = \mathbb{1}_{h(x') \geq h(x_0)} \geq 0,$$

where the equality can be achieved for some $h \in \mathcal{H}$ since $\mathcal{H}$ is regular for general pairwise ranking. Therefore,

$$\mathcal{R}^*_{\mathsf{L}_{0-1}}(\mathcal{H}) = \mathcal{C}^*_{\mathsf{L}_{0-1}}(\mathcal{H}, x_0, x') = \inf_{h \in \mathcal{H}} \mathcal{C}_{\mathsf{L}_{0-1}}(h, x_0, x') = 0.$$

Note $\mathcal{R}_{\mathsf{L}_{0-1}}(h_0) = 1$. For the surrogate loss L_Φ, for any $h \in \mathcal{H}$,

$$\mathcal{R}_{\mathsf{L}_\Phi}(h) = \mathcal{C}_{\mathsf{L}_\Phi}(h, x_0, x') = \Phi(h(x_0) - h(x')) \in [\Phi(\epsilon), \Phi(-\epsilon)]$$

since $|h(x') - h(x_0)| < \epsilon$ and Φ is non-increasing. Therefore,

$$\mathcal{R}^*_{\mathsf{L}_\Phi}(\mathcal{H}) = \mathcal{C}^*_{\mathsf{L}_\Phi}(\mathcal{H}, x_0, x') \geq \Phi(\epsilon).$$

Note $\mathcal{R}_{\mathsf{L}_\Phi}(h_0) = \Phi(0)$. If for some function f that is non-decreasing and continuous at 0, the bound holds, then, we obtain for any $h \in \mathcal{H}$ and $\epsilon > 0$,

$$\mathcal{R}_{\mathsf{L}_{0-1}}(h) - 0 \leq f\Big(\mathcal{R}_{\mathsf{L}_\Phi}(h) - \mathcal{R}^*_{\mathsf{L}_\Phi}(\mathcal{H})\Big) \leq f(\mathcal{R}_{\mathsf{L}_\Phi}(h) - \Phi(\epsilon)).$$

Let $h = h_0$, then $f(\Phi(0) - \Phi(\epsilon)) \geq 1$ for any $\epsilon > 0$. Take $\epsilon \to 0$, we obtain $f(0) \geq 1$ using the fact that Φ and f are both continuous at 0. Since f is non-decreasing, for any $t \in [0, 1]$, $f(t) \geq 1$. □

F. Negative results for bipartite ranking (Proof of Theorem 4.3)

Theorem 4.3 (Negative results for bipartite ranking). *Assume that $\mathcal{X}$ contains an interior point x_0 and that $\mathcal{H}$ is regular for bipartite ranking, contains 0 and is equicontinuous at x_0. If for some function f that is non-decreasing and continuous at 0, the following bound holds for all $h \in \mathcal{H}$ and any distribution,*

$$\mathcal{R}_{\widetilde{\mathsf{L}}_{0-1}}(h) - \mathcal{R}^*_{\widetilde{\mathsf{L}}_{0-1}}(\mathcal{H}) \leq f\Big(\mathcal{R}_{\widetilde{\mathsf{L}}_\Phi}(h) - \mathcal{R}^*_{\widetilde{\mathsf{L}}_\Phi}(\mathcal{H})\Big),$$

then, $f(t) \geq \frac{1}{2}$ for any $t \geq 0$.

Proof. Assume $x_0 \in \mathcal{X}$ is an interior point and $h_0 = 0 \in \mathcal{H}$. By the assumption that x_0 is an interior point and $\mathcal{H}$ is equicontinuous at x_0, for any $\epsilon > 0$, we are able to take $x' \neq x_0 \in \mathcal{X}$ such that $|h(x') - h(x_0)| < \epsilon$ for all $h \in \mathcal{H}$. Consider the distribution that supports on $\{x_0, x'\}$ with $\eta(x_0) = 1$ and $\eta(x') = 0$. Then, for any $h \in \mathcal{H}$,

$$\mathcal{R}_{\widetilde{\mathsf{L}}_{0-1}}(h) = \mathcal{C}_{\widetilde{\mathsf{L}}_{0-1}}(h, x_0, x') = \mathbb{1}_{h(x_0) < h(x')} + \frac{1}{2}\mathbb{1}_{h(x_0) = h(x')} \geq 0,$$

where the equality can be achieved for some $h \in \mathcal{H}$ since $\mathcal{H}$ is regular for bipartite ranking. Therefore,

$$\mathcal{R}^*_{\mathsf{L}_{0-1}}(\mathcal{H}) = \mathcal{C}^*_{\mathsf{L}_{0-1}}(\mathcal{H}, x_0, x') = \inf_{h \in \mathcal{H}} \mathcal{C}_{\mathsf{L}_{0-1}}(h, x_0, x') = 0.$$

Note $\mathcal{R}_{\mathsf{L}_{0-1}}(h_0) = \frac{1}{2}$. For the surrogate loss L_Φ, for any $h \in \mathcal{H}$,

$$\mathcal{R}_{\mathsf{L}_\Phi}(h) = \mathcal{C}_{\mathsf{L}_\Phi}(h, x_0, x') = \Phi(h(x_0) - h(x')) \in [\Phi(\epsilon), \Phi(-\epsilon)]$$

since $|h(x') - h(x_0)| < \epsilon$ and Φ is non-increasing. Therefore,

$$\mathcal{R}^*_{\mathsf{L}_\Phi}(\mathcal{H}) = \mathcal{C}^*_{\mathsf{L}_\Phi}(\mathcal{H}, x_0, x') \geq \Phi(\epsilon).$$

Note $\mathcal{R}_{\mathsf{L}_\Phi}(h_0) = \Phi(0)$. If for some function f that is non-decreasing and continuous at 0, the bound holds, then, we obtain for any $h \in \mathcal{H}$ and $\epsilon > 0$,

$$\mathcal{R}_{\mathsf{L}_{0-1}}(h) - 0 \leq f\Big(\mathcal{R}_{\mathsf{L}_\Phi}(h) - \mathcal{R}^*_{\mathsf{L}_\Phi}(\mathcal{H})\Big) \leq f(\mathcal{R}_{\mathsf{L}_\Phi}(h) - \Phi(\epsilon)).$$

Let $h = h_0$, then $f(\Phi(0) - \Phi(\epsilon)) \geq \frac{1}{2}$ for any $\epsilon > 0$. Take $\epsilon \to 0$, we obtain $f(0) \geq \frac{1}{2}$ using the fact that Φ and f are both continuous at 0. Since f is non-decreasing, for any $t \in [0, 1]$, $f(t) \geq \frac{1}{2}$. □