On the Consistency of Multi-Label Learning

Wei Gao and Zhi-Hua Zhou*

National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China

Abstract

Multi-label learning has attracted much attention during the past few years. Many multi-label approaches have been developed, mostly working with surrogate loss functions because multi-label loss functions are usually difficult to optimize directly owing to their non-convexity and discontinuity. These approaches are effective empirically, however, little effort has been devoted to the understanding of their consistency, i.e., the convergence of the risk of learned functions to the Bayes risk. In this paper, we present a theoretical analysis on this important issue. We first prove a necessary and sufficient condition for the consistency of multi-label learning based on surrogate loss functions. Then, we study the consistency of two well-known multi-label loss functions, i.e., *ranking loss* and *hamming loss*. For ranking loss, our results disclose that, surprisingly, none of convex surrogate loss is consistent; we present the *partial ranking loss*, with which some surrogate losses are proven to be consistent. We also discuss on the consistency of univariate surrogate losses. For hamming loss, we show that two multi-label learning methods, i.e., *one-vs-all* and *pairwise comparison*, which can be regarded as direct extensions from multi-class learning, are inconsistent in general cases yet consistent under the *dominating setting*, and similar results also hold for some recent multi-label approaches that are variations of one-vs-all. In addition, we discuss on the consistency of learning approaches that address multi-label learning by decomposing into a set of binary classification problems.

Keywords: Machine Learning, Multi-Label Learning, Surrogate Loss, Bayes Consistency, Ranking Loss, Hamming Loss

*Email: zhouzh@nju.edu.cn

1. Introduction

In traditional classification tasks, each instance is associated with a single label in a number of candidate labels, e.g., binary classification and multi-class learning. In real tasks, however, one object is usually relevant to a set of labels simultaneously. For example, in text categorization, a document about national education service may cover several predefined topics, such as government and education, indicating the content in the document (Schapire and Singer, 2000); in bioinformatics, a gene sequence may be relevant to multiple functions, such as metabolism, transcription and protein synthesis, showing the functions of the gene within a cell's life circle (Elisseeff and Weston, 2002); in image annotation, an image may be annotated with a set of words, such as trees and mountains (Carneiro et al., 2007; Qi et al., 2007). To tackle such problems, *multi-label learning* has been explored, and it has attracted much attention during the past decade (Schapire and Singer, 2000; Elisseeff and Weston, 2002; Zhou and Zhang, 2007; Zhang and Zhou, 2007; Hüllermeier et al., 2008; Hsu et al., 2009; Dembczyński et al., 2010, 2012b; Petterson and Caetano, 2010; Zhou et al., 2012).

In multi-label learning, many loss functions (also called *evaluation criteria*) have been utilized to measure the performance of learning algorithms, e.g., *ranking loss*, *hamming loss*, *one-error*, *coverage* and *average precision* (Schapire and Singer, 2000; Zhang and Zhou, 2006); *accuracy*, *precision*, *recall* and F_1 (Godbole and Sarawagi, 2004; Qi et al., 2007); *subset accuracy* (Ghamrawi and McCallum, 2005); etc. It is noteworthy that all of them are non-convex and discontinuous, and directly optimizing such losses often leads to NP-hard problems. To make a compromise for avoiding computational difficulties, surrogate losses that can be optimized more efficiently are usually adopted in practical algorithms, e.g., boosting algorithm AdaBoost.MH (Schapire and Singer, 2000), neural network algorithm BP-MLL (Zhang and Zhou, 2006), SVM-style algorithms (Elisseeff and Weston, 2002; Taskar et al., 2004; Hariharan et al., 2010), etc. Essentially, all these algorithms try to optimize some convex surrogate losses such as the exponential loss and hinge loss. Despite of their efficient computation, there remains an important theoretical problem: Whether the expected risks of the learned functions converge to the Bayes risk? Or in other words, how about their *consistency* (also called *Bayes consistency*)? This paper presents a theoretical study on this important issue.

1.1. Our Contributions

We first prove a necessary and sufficient condition for the consistency of multi-label learning based on surrogate loss functions. Based on this result, we examine the consistency of two well-known multi-label loss functions: *ranking loss* and *hamming loss*.

For ranking loss, our results disclose that none of convex surrogate loss is consistent. So, we present the *partial ranking loss*, with which some surrogate functions, e.g., regularized linear loss and sigmoid-type losses used for neural network, are consistent. We also study the consistency of univariate surrogate loss, and identify a class of consistent univariate losses for partial ranking loss, generalizing the recent results of Dembczyński et al. (2012a).

For hamming loss, we show that two multi-label learning methods, i.e., *one-vs-all* and *pairwise comparison* that can be regarded as direct extensions from multi-class learning, are inconsistent in general cases yet consistent under the *dominating setting*, and similar results also hold for some recent multi-label approaches that are variations of one-vs-all. We also discuss on the consistency of learning approaches that address multi-label learning by decomposing into a set of binary classification problems.

1.2. Related Work

Consistency guarantees that optimizing a surrogate loss function will yield ultimately an optimal function with Bayes risk for the true loss function, and thus proceed in the scope of computationally efficient algorithms. Nowadays, it is well-accepted that a good learner should at least be consistent with large samples. Bühlmann and Yu (2003) discussed on the consistency of boosting algorithms with respect to least square loss, and Breiman (2004) studied the convergence of arcing-style greedy boosting algorithms to the Bayes classifier. The consistency theory on support vector machines was developed in (Lin, 2002; Steinwart, 2005). The most influential and fundamental work (Zhang, 2004b; Bartlett et al., 2006) investigated the consistency for binary classification, in which many popular algorithms (e.g., boosting, logistic regression and SVMs) are proven to be consistent. In addition, McAllester and Keshet (2011) studied the consistency for latent structural probit and ramp loss.

For multi-class learning, the consistent theory has been well-studied in (Zhang, 2004a; Tewari and Bartlett, 2007) and many SVM-type algorithms are proven to be inconsistent. Note that multi-class learning is very different from multi-label learning. Given a set of candidate labels, multi-class learning assumes that there is only one label which is correct for an instance,

whereas multi-label learning accepts the fact that more labels can be correct and is more challenging. For multi-class learning, the 0/1 loss (i.e., accuracy) is naturally the fundamental criterion, whereas for multi-label learning there are many criteria, among which hamming loss is mostly related with accuracy. As will be shown in Section 5.1, decomposing multi-label learning into a series of multi-class learning problems to solve, either by *one-vs-all* or *pairwise comparison*, is inconsistent. It is also possible to decompose multi-label learning into a series of independent binary classification problems to solve, as studied in Section 5.3; such approach completely neglects the interaction between labels (also called label correlations) and could not work well with a large number of labels and/or some labels lacking sufficient training data, and thus rarely adopted in practice.

Much work has been devoted to the analysis of consistency for ranking problems under different learning settings, e.g., subset ranking (Cossock and Zhang, 2008), listwise ranking (Xia et al., 2008), top-k ranking (Xia et al., 2009), etc. Duchi et al. (2010) studied the consistency of general ranking setting where each "instance" consists of a query, a set of inputs and a weighted graph, and the goal is to order the inputs according to the weighted graph. From some sense, multi-label learning contains some behaviors of ranking: It tends to rank relevant labels higher than irrelevant ones. However, multi-label learning requires to estimate the number of relevant labels and is more challenging than a pure ranking. Thus, some results in Section 4.1 may seem similar to those obtained by Duchi et al. (2010) but most are very different.

1.3. Organization

The rest of this paper is organized as follows. Section 2 introduces some preliminaries. Section 3 presents a necessary and sufficient condition for the consistency of multi-label approaches. Sections 4 and 5 study the consistency of ranking loss and hamming loss, respectively. Section 6 presents the detailed proofs of lemmas. Section 7 concludes with future work.

2. Preliminaries

Let $\mathcal{X}$ be an instance space and $\mathcal{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_q\}$ denotes a finite set of candidate labels. An instance $\boldsymbol{x} \in \mathcal{X}$ is associated with a subset of labels $\boldsymbol{y} \subseteq \mathcal{L}$ which are called *relevant* labels, whereas the complement $\mathcal{L} \setminus \boldsymbol{y}$ are called *irrelevant* labels. For convenience, we represent the labels as a binary vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_q)$ where $y_i = +1$ if the label λ_i is relevant to $\boldsymbol{x}$ and

-1 otherwise, and we further denote by $\mathcal{Y} = \{+1, -1\}^q$ the set of all possible labels. For a real number r, $\lfloor r \rfloor$ denotes the largest integer which is no more than r.

Let $\mathcal{D}$ denote an (unknown) underlying distribution over $\mathcal{X} \times \mathcal{Y}$. For an instance $\boldsymbol{x} \in \mathcal{X}$, we denote by $\boldsymbol{p}(\boldsymbol{x})$ a vector of conditional probability over $\boldsymbol{y} \in \mathcal{Y}$, i.e.,

$$\boldsymbol{p}(\boldsymbol{x}) = \left(p_{\boldsymbol{y}}(\boldsymbol{x})\right)_{\boldsymbol{y}\in\mathcal{Y}} = \left(\Pr\left(\boldsymbol{y}|\boldsymbol{x}\right)\right)_{\boldsymbol{y}\in\mathcal{Y}},$$

for some $\boldsymbol{p}(\boldsymbol{x}) \in \Lambda$, where Λ denotes the flat simplex of $\mathbb{R}^{|\mathcal{Y}|}$, that is,

$$\Lambda = \left\{\boldsymbol{p} \in \mathbb{R}^{|\mathcal{Y}|} \colon \sum\nolimits_{\boldsymbol{y}\in\mathcal{Y}} p_{\boldsymbol{y}} = 1 \ \text{ and } \ p_{\boldsymbol{y}} \geq 0\right\}.$$

The formal description of multi-label learning in the probabilistic setting is given as follows. We are given a training sample

$$S = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \ldots, (\boldsymbol{x}_m, \boldsymbol{y}_m)\}$$

drawn independently and identically (i.i.d) according to distribution $\mathcal{D}$, and the objective is to learn a function $h \colon \mathcal{X} \to \mathcal{Y}$, which is able to assign a set of labels to unseen instances. In general, it is not easy to learn h directly, and one instead learns a real-valued vector function

$$\boldsymbol{f} = (f_1, f_2, \ldots, f_K) \colon \mathcal{X} \to \mathbb{R}^K \text{ for some integer } K > 0,$$

where $K = q$ or $K = 2^q$ are common choices for the design of practical algorithms. Based on this vector function $\boldsymbol{f}$, a prediction function $F \colon \mathbb{R}^K \to \mathcal{Y}$ can be attained for assigning the set of relevant labels to an instance. Another popular approach for multi-label learning is to learn a real-valued vector function

$$\boldsymbol{f} = (f_1, f_2, \ldots, f_q) \text{ s.t. } f_i(\boldsymbol{x}) > f_j(\boldsymbol{x}) \text{ if } y_i = +1, y_j = -1,$$

i.e., rank relevant labels higher than irrelevant ones for example $(\boldsymbol{x}, \boldsymbol{y})$, and then, a function should be learned to determine the number of relevant labels.

Essentially, multi-label approaches try to minimize the expected risk of $\boldsymbol{f}$ with respect to some loss L, i.e.,

$$R(\boldsymbol{f}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[L(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y})], \tag{1}$$

where $\boldsymbol{f}$ may be a prediction function or a vector of real-valued functions according to different losses. We further denote the minimal risk (also called the Bayes risk) by

$$R^* = \inf_{\boldsymbol{f}} R(\boldsymbol{f}),$$

where the infimum takes over all measurable functions. Throughout this paper, we mainly focus on *below-bounded* and *distinguishable* loss functions defined as follows:

Definition 1. *A loss function L is said to be below-bounded if $L(\cdot,\cdot) \geq B$ for some constant B.*

Definition 2. *A loss function L is said to be distinguishable if for some $\gamma > 0$, for every $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ and $\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}$, it holds that*

$$L(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = L(\boldsymbol{f}(\boldsymbol{x}'), \boldsymbol{y}') \quad \text{or} \quad |L(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) - L(\boldsymbol{f}(\boldsymbol{x}'), \boldsymbol{y}')| \geq \gamma.$$

Many loss functions are below-bounded and distinguishable, e.g., ranking loss, hamming loss, one-error, average precision, etc., and we will study ranking loss and hamming loss in Sections 4 and 5, respectively.

For notational simplicity, we will suppress dependence of $\boldsymbol{p}(\boldsymbol{x})$ and $\boldsymbol{f}(\boldsymbol{x})$ on the instance $\boldsymbol{x}$ as $\boldsymbol{p}$ and $\boldsymbol{f}$, respectively, when it is clear from the context. For an instance $\boldsymbol{x} \in \mathcal{X}$, we define the conditional risk of $\boldsymbol{f}$ as

$$l(\boldsymbol{p}, \boldsymbol{f}) = \sum\nolimits_{\boldsymbol{y} \in \mathcal{Y}} p_{\boldsymbol{y}} L(\boldsymbol{f}, \boldsymbol{y}) = \sum\nolimits_{\boldsymbol{y} \in \mathcal{Y}} \Pr(\boldsymbol{y}|\boldsymbol{x}) L(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}). \tag{2}$$

It is easy to get the expected risk and the Bayes risk, respectively, as

$$R(\boldsymbol{f}) = \mathbb{E}_{\boldsymbol{x}}[l(\boldsymbol{p}, \boldsymbol{f})] \quad \text{and} \quad R^* = \mathbb{E}_{\boldsymbol{x}}\big[\inf\nolimits_{\boldsymbol{f}}[l(\boldsymbol{p}, \boldsymbol{f})]\big].$$

We further define the *set of Bayes predictions* as

$$\mathcal{A}(\boldsymbol{p}) = \big\{\boldsymbol{f} \colon l(\boldsymbol{p}, \boldsymbol{f}) = \inf\nolimits_{\boldsymbol{f}'} l(\boldsymbol{p}, \boldsymbol{f}')\big\},$$

and it is clear $\mathcal{A}(\boldsymbol{p}) \neq \emptyset$ as L is distinguishable and below-bounded.

Many multi-label loss functions L, as mentioned in Section 1, have been explored to measure the performance of multi-label learning algorithms, whereas it is noteworthy that all of them are non-convex and discontinuous, and directly optimizing such loss functions often yields NP-hard problems.

Therefore, a feasible solution in practice is to consider a convex surrogate loss Ψ in place of L. We define the Ψ-risk and Bayes Ψ-risk of $\boldsymbol{f}$, respectively, as

$$R_\Psi(\boldsymbol{f}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[\Psi(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y})] \quad \text{and} \quad R^*_\Psi = \inf_{\boldsymbol{f}} R_\Psi(\boldsymbol{f}).$$

Similarly, we define the conditional surrogate risk of $\boldsymbol{f}$ and the conditional Bayes surrogate risk, respectively, as

$$W(\boldsymbol{p}, \boldsymbol{f}) = \sum\nolimits_{\boldsymbol{y}\in\mathcal{Y}} p_{\boldsymbol{y}} \Psi(\boldsymbol{f}, \boldsymbol{y}) \quad \text{and} \quad W^*(\boldsymbol{p}) = \inf_{\boldsymbol{f}} W(\boldsymbol{p}, \boldsymbol{f}).$$

It is obvious that

$$R_\Psi(\boldsymbol{f}) = E_{\boldsymbol{x}}[W(\boldsymbol{p}, \boldsymbol{f})] \quad \text{and} \quad R^*_\Psi = E_{\boldsymbol{x}}[W^*(\boldsymbol{p})].$$

3. Multi-Label Consistency

Many notions of consistency have been introduced in the literature, e.g., the Fisher consistency (Lin, 2002), infinite-sample consistency (Zhang, 2004a), classification calibration (Bartlett et al., 2006; Tewari and Bartlett, 2007), edge-consistency (Duchi et al., 2010), etc. In this paper, we introduce the ***multi-label consistency*** as follows:

Definition 3. *Given a below-bounded surrogate loss Ψ where $\Psi(\cdot, \boldsymbol{y})$ is continuous for every $\boldsymbol{y} \in \mathcal{Y}$, Ψ is said to be multi-label consistent w.r.t. the loss L if it holds, for every $\boldsymbol{p} \in \Lambda$, that*

$$W^*(\boldsymbol{p}) < \inf_{\boldsymbol{f}}\{W(\boldsymbol{p}, \boldsymbol{f}) \colon \boldsymbol{f} \notin \mathcal{A}(\boldsymbol{p})\}.$$

The following theorem states that the multi-label consistency is a necessary and sufficient condition for the convergence of Ψ-risk to the Bayes Ψ-risk, implying $R(\boldsymbol{f}) \to R^*$.

Theorem 4. *The surrogate loss Ψ is multi-label consistent w.r.t. the loss L if and only if it holds for any sequence $\{\boldsymbol{f}_n\}_{n\geq 1}$ that*

$$\textit{if} \quad R_\Psi(\boldsymbol{f}_n) \to R^*_\Psi \quad \textit{then} \quad R(\boldsymbol{f}_n) \to R^*.$$

The proof is inspired by the techniques of (Zhang, 2004a; Tewari and Bartlett, 2007). We begin with two useful lemmas, whose proofs are deferred to Sections 6.1 and 6.2, respectively.

Lemma 5. *$W^*(\boldsymbol{p})$ is continuous on Λ.*

Lemma 6. *If the surrogate loss function Ψ is multi-label consistent w.r.t. loss function L, then for any $\epsilon > 0$, there exists $\delta > 0$ such that, for every $\boldsymbol{p} \in \Lambda$,*

$$\textit{if} \quad l(\boldsymbol{p}, \boldsymbol{f}) - \inf_{\boldsymbol{f}'} l(\boldsymbol{p}, \boldsymbol{f}') \geq \epsilon \quad \textit{then} \quad W(\boldsymbol{p}, \boldsymbol{f}) - W^*(\boldsymbol{p}) \geq \delta.$$

Proof of Theorem 4: ("⇒") We first introduce a new notation

$$H(\epsilon) = \inf_{\boldsymbol{p}\in\Lambda, \boldsymbol{f}} \{W(\boldsymbol{p}, \boldsymbol{f}) - \inf_{\boldsymbol{f}'} W(\boldsymbol{p}, \boldsymbol{f}') \colon l(\boldsymbol{p}, \boldsymbol{f}) - \inf_{\boldsymbol{f}'} l(\boldsymbol{p}, \boldsymbol{f}') \geq \epsilon\}.$$

It is obvious that $H(0) = 0$ and $H(\epsilon) > 0$ for $\epsilon > 0$ from Lemma 6. Corollary 26 of (Zhang, 2004a) guarantees the existence of a concave function η on $[0, \infty]$ such that $\eta(0) = 0$ and $\eta(\epsilon) \to 0$ as $\epsilon \to 0$ and

$$R(f) - R^* \leq \eta(R_\Psi(f) - R^*_\Psi).$$

Thus, if $R_\Psi(f) \to R^*_\Psi$ then $R(f) \to R^*$.

("⇐") We proceed by contradiction. Suppose Ψ is not multi-label consistent, and thus there exists some $\boldsymbol{p}$ s.t. $W^*(\boldsymbol{p}) = \inf_{\boldsymbol{f}} \{W(\boldsymbol{p}, \boldsymbol{f}) \colon \boldsymbol{f} \notin \mathcal{A}(\boldsymbol{p})\}$. Let $\boldsymbol{f}^{(n)} \notin \mathcal{A}(\boldsymbol{p})$ be a sequence s.t. $W(\boldsymbol{p}, \boldsymbol{f}^{(n)}) \to W^*(\boldsymbol{p})$. For simplicity, we consider $\mathcal{X} = \{\boldsymbol{x}\}$, i.e., only one instance, and set $\boldsymbol{f}_n(\boldsymbol{x}) = \boldsymbol{f}^{(n)}$. Then,

$$R_\Psi(\boldsymbol{f}_n) = W(\boldsymbol{p}, \boldsymbol{f}^{(n)}) \to W^*(\boldsymbol{p}) = R^*_\Psi,$$

yielding $l(\boldsymbol{p}, \boldsymbol{f}^{(n)}) \to \inf_{\boldsymbol{f}} l(\boldsymbol{p}, \boldsymbol{f})$, whereas it is contrary to

$$l(\boldsymbol{p}, \boldsymbol{f}^{(n)}) \geq \inf_{\boldsymbol{f}} l(\boldsymbol{p}, \boldsymbol{f}) + \gamma(\boldsymbol{p}),$$

where $\gamma(\boldsymbol{p}) = \inf_{\boldsymbol{f} \notin \mathcal{A}(\boldsymbol{p})} l(\boldsymbol{p}, \boldsymbol{f}) - \inf_{\boldsymbol{f}} l(\boldsymbol{p}, \boldsymbol{f}) > 0$, because $\boldsymbol{f}^{(n)} \notin \mathcal{A}(\boldsymbol{p})$ and L is distinguishable. Thus, this theorem follows as desired. □

Owing to the non-convexity and discontinuity of multi-label loss L, there may exist many solutions minimizing the risk $R(h)$, and the set of Bayes predictions $\mathcal{A}(\boldsymbol{p})$ contains all global optimal solutions. For the surrogate loss Ψ, we denote by

$$\mathcal{S}(\boldsymbol{p}) = \{\boldsymbol{f} \colon W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})\}$$

the set of all functions which minimize the surrogate loss Ψ. An intuitive explanation to Theorem 4 is that the surrogate loss Ψ is multi-label consistent w.r.t. the loss L if and only if $\mathcal{S}(\boldsymbol{p}) \subseteq \mathcal{A}(\boldsymbol{p})$.

4. Consistency w.r.t. Ranking Loss

The ranking loss concerns about label pairs that are ordered reversely for an instance. For a real-valued ranking function $\boldsymbol{f} = (f_1, f_2, \ldots, f_q)$, the ranking loss is given by

$$\begin{aligned} L_{\text{rankloss}}(\boldsymbol{f}, (\boldsymbol{x}, \boldsymbol{y})) &= a_{\boldsymbol{y}} \sum\nolimits_{\substack{y_i=-1 \\ y_j=+1}} I[f_i(\boldsymbol{x}) \geq f_j(\boldsymbol{x})] \\ &= a_{\boldsymbol{y}} \sum\nolimits_{y_i<y_j} I[f_i(\boldsymbol{x}) \geq f_j(\boldsymbol{x})], \end{aligned} \tag{3}$$

where $a_{\boldsymbol{y}}$ is a non-negative penalty, and $I[\cdot]$ is the indicator function which returns 1 if the argument is true and 0 otherwise. The most commonly used penalty in multi-label learning is

$$a_{\boldsymbol{y}} = |\{i \in [q] \colon y_i = -1\}|^{-1} \times |\{j \in [q] \colon y_j = +1\}|^{-1}.$$

In this paper we consider the more general penalty, i.e., non-negative penalty. It is clear that the ranking loss is below-bounded from $L_{\text{rankloss}}(\boldsymbol{f}, (\boldsymbol{x}, \boldsymbol{y})) \geq 0$, and it is distinguishable because for each $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ and $\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}$, we have

$$\begin{aligned} &L_{\text{rankloss}}(\boldsymbol{f}, (\boldsymbol{x}, \boldsymbol{y})) = L_{\text{rankloss}}(\boldsymbol{f}, (\boldsymbol{x}', \boldsymbol{y}')) \quad \text{or} \\ &|L_{\text{rankloss}}(\boldsymbol{f}, (\boldsymbol{x}, \boldsymbol{y})) - L_{\text{rankloss}}(\boldsymbol{f}, (\boldsymbol{x}', \boldsymbol{y}'))| \geq \gamma. \end{aligned}$$

Here $\gamma = \min_{\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}, 0 \leq i, j \leq q^2/4} \left\{ |i a_{\boldsymbol{y}} - j a_{\boldsymbol{y}'}| > 0 \right\}$, because Eqn. (3) yields

$$\begin{aligned} &L_{\text{rankloss}}(\boldsymbol{f}, (\boldsymbol{x}, \boldsymbol{y})) \in \{i a_{\boldsymbol{y}},\ 0 \leq i \leq q^2/4\} \quad \text{and} \\ &L_{\text{rankloss}}(\boldsymbol{f}, (\boldsymbol{x}, \boldsymbol{y}')) \in \{j a_{\boldsymbol{y}'},\ 0 \leq j \leq q^2/4\}. \end{aligned}$$

After obtaining ranking function $\boldsymbol{f}$, there are at least two ways to exploit the ranking result to get the actual number of labels. The first way (Elisseeff and Weston, 2002) is to learn another function which is able to predict the number of labels. Another way (Fürnkranz et al., 2008) is to insert a "calibration" label between relevant and irrelevant labels for training examples, and then, after prediction, labels ranked higher than the calibration label will be regarded as relevant ones. More details can be found in (Elisseeff and Weston, 2002; Fürnkranz et al., 2008).

Before further discussion, we introduce the following notations:

$$\Delta = \sum\nolimits_{\boldsymbol{y} \in \mathcal{Y}} a_{\boldsymbol{y}} p_{\boldsymbol{y}}, \quad \Delta_i^- = \sum\nolimits_{\boldsymbol{y} \colon y_i = -1} a_{\boldsymbol{y}} p_{\boldsymbol{y}} \quad \text{and} \quad \Delta_{i,j} = \sum\nolimits_{\boldsymbol{y} \colon y_i < y_j} p_{\boldsymbol{y}} a_{\boldsymbol{y}},$$

for a given vector $\boldsymbol{p} \in \Lambda$ and non-negative vector $(a_{\boldsymbol{y}})_{\boldsymbol{y} \in \mathcal{Y}}$. It is easy to get:

Lemma 7. *For a vector $\boldsymbol{p} \in \Lambda$ and non-negative vector $(a_{\boldsymbol{y}})_{\boldsymbol{y}\in\mathcal{Y}}$, the following properties hold:*

1. $\Delta_{i,i} = 0$;
2. $\Delta_i^- - \Delta_j^- = \Delta_{i,j} - \Delta_{j,i}$;
3. $\Delta_{i,k} + \Delta_{k,j} + \Delta_{j,i} = \Delta_{k,i} + \Delta_{i,j} + \Delta_{j,k}$;
4. $\Delta_{i,k} \leq \Delta_{k,i}$ *if* $\Delta_{i,j} \leq \Delta_{j,i}$ *and* $\Delta_{j,k} \leq \Delta_{k,j}$.

Proof: Property 1 is immediate from definition, and Property 2 holds from

$$\Delta_i^- - \Delta_j^- = \sum_{\boldsymbol{y}:\, y_i=y_j=-1} a_{\boldsymbol{y}} p_{\boldsymbol{y}} + \sum_{\boldsymbol{y}:\, y_i<y_j} a_{\boldsymbol{y}} p_{\boldsymbol{y}} - \sum_{\boldsymbol{y}:\, y_i=y_j=-1} a_{\boldsymbol{y}} p_{\boldsymbol{y}} - \sum_{\boldsymbol{y}:\, y_j<y_i} a_{\boldsymbol{y}} p_{\boldsymbol{y}}.$$

From Property 2, we have

$$\Delta_i^- - \Delta_j^- = \Delta_{i,j} - \Delta_{j,i},\ \Delta_j^- - \Delta_k^- = \Delta_{j,k} - \Delta_{k,j},\ \Delta_k^- - \Delta_i^- = \Delta_{k,i} - \Delta_{i,k};$$

therefore, Property 3 follows. Property 4 holds from Property 3 directly. □

Based on this lemma, we get the set of Bayes predictions for ranking loss as follows:

Lemma 8. *For every $\boldsymbol{p} \in \Lambda$ and non-negative vector $(a_{\boldsymbol{y}})_{\boldsymbol{y}\in\mathcal{Y}}$, the set of Bayes predictions for ranking loss is given by*

$$\mathcal{A}(\boldsymbol{p}) = \{\boldsymbol{f} : \textit{for all } i<j, f_i > f_j \textit{ if } \Delta_{i,j} < \Delta_{j,i};\\ f_i \neq f_j \textit{ if } \Delta_{i,j} = \Delta_{j,i};\ \textit{and}\ \ f_i < f_j \textit{ otherwise}\}.$$

Proof: From the definition of the conditional risk given by Eqn. (2), we have

$$\begin{aligned} l(\boldsymbol{p}, \boldsymbol{f}) &= \sum\nolimits_{\boldsymbol{y}\in\mathcal{Y}} p_{\boldsymbol{y}} L_{\text{rankloss}}(\boldsymbol{f}, (\boldsymbol{x}, \boldsymbol{y})) \\ &= \sum\nolimits_{\boldsymbol{y}\in\mathcal{Y}} p_{\boldsymbol{y}} a_{\boldsymbol{y}} \sum\nolimits_{y_i<y_j} I[f_i \geq f_j] \\ &= \sum\nolimits_{\boldsymbol{y}\in\mathcal{Y}} p_{\boldsymbol{y}} a_{\boldsymbol{y}} \sum\nolimits_{1\leq i,j\leq q} I[f_i \geq f_j] \cdot I[y_i < y_j]. \end{aligned}$$

By swapping the two sums, we get

$$\begin{aligned} l(\boldsymbol{p}, \boldsymbol{f}) &= \sum\nolimits_{1\leq i,j\leq q} I[f_i \geq f_j] \sum\nolimits_{\boldsymbol{y}:\, y_i<y_j} p_{\boldsymbol{y}} a_{\boldsymbol{y}} \\ &= \sum\nolimits_{1\leq i,j\leq q} I[f_i \geq f_j] \Delta_{i,j} \\ &= \sum\nolimits_{1\leq i<j\leq q} I[f_i \geq f_j] \Delta_{i,j} + I[f_i \leq f_j] \Delta_{j,i}. \end{aligned}$$

Hence we complete the proof by combining with Property 4 in Lemma 7. □

For ranking loss, it is natural to consider the following surrogate loss:

$$\Psi(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = a_{\boldsymbol{y}} \sum\nolimits_{y_i<y_j} \phi(f_j(\boldsymbol{x}) - f_i(\boldsymbol{x})), \tag{4}$$

where ϕ is convex and non-increasing such as the hinge loss $\phi(x) = (1-x)_+$ in (Elisseeff and Weston, 2002), exponential loss $\phi(x) = \exp(-x)$ in (Schapire and Singer, 2000; Dekel et al., 2004; Zhang and Zhou, 2006), etc. The following theorem discloses that none of convex surrogate loss is consistent with ranking loss.

Theorem 9. *For any convex function ϕ, the surrogate loss*

$$\Psi(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = \sum\nolimits_{y_i<y_j} a_{\boldsymbol{y}} \phi(f_j(\boldsymbol{x}) - f_i(\boldsymbol{x}))$$

is inconsistent w.r.t. ranking loss.

Proof: For surrogate loss Ψ, we have

$$\begin{aligned} W(\boldsymbol{p}, \boldsymbol{f}) &= \sum\nolimits_{\boldsymbol{y}\in\mathcal{Y}} p_{\boldsymbol{y}} \Psi(\boldsymbol{f}, \boldsymbol{y}) = \sum\nolimits_{\boldsymbol{y}\in\mathcal{Y}} p_{\boldsymbol{y}} a_{\boldsymbol{y}} \sum\nolimits_{y_i<y_j} \phi(f_j - f_i) \\ &= \sum\nolimits_{1\le i<j\le q} \phi(f_j - f_i)\Delta_{i,j} + \phi(f_i - f_j)\Delta_{j,i}. \end{aligned}$$

Consider the probability vector $\boldsymbol{p} = (p_{\boldsymbol{y}})_{\boldsymbol{y}\in\mathcal{Y}}$ and penalty vector $(a_{\boldsymbol{y}})_{\boldsymbol{y}\in\mathcal{Y}}$ s.t. $p_{\boldsymbol{y}_1} = p_{\boldsymbol{y}_2}$ and $a_{\boldsymbol{y}_1} = a_{\boldsymbol{y}_2}$ for every $\boldsymbol{y}_1 \neq \boldsymbol{y}_2$, $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathcal{Y}$. This yields that $\Delta_{i,j} = \Delta_{m,n}$ for every $1 \le i \neq j, m \neq n \le q$, and thus we get

$$W(\boldsymbol{p}, \boldsymbol{f}) = \Delta_{1,2} \sum\nolimits_{1\le i<j\le q} \phi(f_j - f_i) + \phi(f_i - f_j).$$

From the convexity of ϕ, minimizing $W(\boldsymbol{p}, \boldsymbol{f})$ gives

$$W^*(\boldsymbol{p}) = W(\boldsymbol{p}, \hat{\boldsymbol{f}}) = q(q-1)\phi(0)\Delta_{1,2},$$

where $\hat{\boldsymbol{f}} = \{\hat{\boldsymbol{f}} \colon \hat{f}_1 = \hat{f}_2 = \cdots = \hat{f}_q\}$. From Lemma 8, we have $\hat{\boldsymbol{f}} \notin \mathcal{A}(\boldsymbol{p})$, and

$$W^*(\boldsymbol{p}) = \inf\nolimits_{\boldsymbol{f}} \{W(\boldsymbol{p}, \boldsymbol{f}) \colon \boldsymbol{f} \notin \mathcal{A}(\boldsymbol{p})\}$$

which show the inconsistency. This theorem follows as desired. □

Intuitively, Property 4 of Lemma 7 implies that $\{\Delta_{i,j}\}$ defines an order for the label set $\mathcal{L} = \{1, 2, \ldots, q\}$ by $i \succeq j$ if $\Delta_{i,j} \leq \Delta_{j,i}$. Notice that, for $i \succeq j$, $\Delta_{i,j} = \Delta_{j,i}$ is possible. The set of Bayes predictions of a reasonable loss function should include all functions that are compatible with this order, i.e., $\boldsymbol{f}$'s that enable $f_i \geq f_j$ if $i \succeq j$. In the definition of ranking loss given by Eqn. (3), the same penalty term is applied to $f_i < f_j$ and $f_i = f_j$; thus, the set of Bayes predictions with respect to ranking loss does not include some functions that are compatible with the above order as what enforces by ranking loss is $i \succ j$ or $j \succ i$ if $\Delta_{i,j} = \Delta_{j,i}$ for the label set $\mathcal{L}$. For an extreme example, i.e., when all $\Delta_{i,j}$'s are equal for all $i \neq j$, minimizing the convex surrogate loss function Ψ leads to the optimal solution

$$\boldsymbol{f}^* \in \{\boldsymbol{f} \colon f_1 = f_2 = \cdots = f_q\},$$

but $\boldsymbol{f}^* \notin \mathcal{A}(\boldsymbol{p})$ (from Lemma 8). So, the same penalty on $f_i < f_j$ and $f_i = f_j$ encumbers the multi-label consistency.

To overcome the deficiency of ranking loss, we present the *partial ranking loss*

$$L_{\text{p-rankloss}}(\boldsymbol{f}, (\boldsymbol{x}, \boldsymbol{y})) = a_{\boldsymbol{y}} \sum\nolimits_{y_i < y_j} I[f_i(\boldsymbol{x}) > f_j(\boldsymbol{x})] + \frac{1}{2} I[f_i(\boldsymbol{x}) = f_j(\boldsymbol{x})], \quad (5)$$

which has been used for ranking problems. The only difference from ranking loss lies in the use of different penalties for $\sum_{y_i < y_j} I[f_i = f_j]$, where the ranking loss uses $a_{\boldsymbol{y}}$ whereas the partial ranking loss uses $a_{\boldsymbol{y}}/2$. With a proof similar to that of Lemma 8, we can get the set of Bayes predictions with respect to the partial ranking loss:

$$\mathcal{A}(\boldsymbol{p}) = \{\boldsymbol{f} \colon \text{ for all } i < j,\ f_i > f_j \text{ if } \Delta_{i,j} < \Delta_{j,i};\ f_i < f_j \text{ if } \Delta_{i,j} > \Delta_{j,i}\}. \quad (6)$$

Now, consider the previous extreme example in Theorem 9, i.e., $\Delta_{i,j}$'s are equal for all $i \neq j$, again. It is easy to see that by minimizing the surrogate loss function Ψ, the optimal solution

$$\boldsymbol{f}^* \in \{\boldsymbol{f} \colon f_1 = f_2 = \cdots = f_q\} \subseteq \mathcal{A}(\boldsymbol{p})$$

exhibits consistency.

4.1. Consistency of Surrogate Loss of Eqn. (4)

In this section, we consider the surrogate loss of Eqn. (4), i.e.,

$$\Psi(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = a_{\boldsymbol{y}} \sum\nolimits_{y_i<y_j} \phi(f_j(\boldsymbol{x}) - f_i(\boldsymbol{x})),$$

and examine its consistency w.r.t. partial ranking loss. We begin with a sufficient condition for consistency as follows:

Theorem 10. *If $\phi\colon \mathbb{R} \to \mathbb{R}$ is a differentiable and non-increasing function such that*

$$\phi'(0) < 0 \quad \textit{and} \quad \phi(t) + \phi(-t) \equiv 2\phi(0), \tag{7}$$

i.e., $\phi(t) + \phi(-t) = 2\phi(0)$ for every $t \in \mathbb{R}$, then the surrogate loss Ψ of Eqn. (4) *is consistent w.r.t. partial ranking loss.*

Proof: For every probability simplex $\boldsymbol{p} \in \Lambda$ and non-negative vector $(a_{\boldsymbol{y}})_{\boldsymbol{y}\in\mathcal{Y}}$, it suffices to prove that $f_i > f_j$ if $\Delta_{i,j} < \Delta_{j,i}$ for every $\boldsymbol{f}$ such that $W^*(\boldsymbol{p}) = W(\boldsymbol{p}, \boldsymbol{f})$. Without loss of generality, we will prove that $f_1 > f_2$ if $\Delta_{1,2} < \Delta_{2,1}$ by contradiction, i.e., assuming $f_1 \leq f_2$ for some vector $\boldsymbol{f}$ which satisfies $W^*(\boldsymbol{p}) = W(\boldsymbol{p}, \boldsymbol{f})$.

For the case $f_1 < f_2$, we construct another $\boldsymbol{f}'$ by

$$f'_1 = f_2, f'_2 = f_1 \text{ and } f'_k = f_k \text{ for } k \neq 1, 2.$$

From the definition of conditional surrogate risk, we have

$$W(\boldsymbol{p}, \boldsymbol{f}) = \sum\nolimits_{1\leq i<j\leq q} \phi(f_j - f_i)\Delta_{i,j} + \phi(f_i - f_j)\Delta_{j,i},$$

which yields that

$$\begin{aligned} W(\boldsymbol{p}, \boldsymbol{f}) - W(\boldsymbol{p}, \boldsymbol{f}') = {} & (\Delta_{1,2} - \Delta_{2,1})(\phi(f_2 - f_1) - \phi(f_1 - f_2)) \\ & + \sum\nolimits_{i=3}^{q} (\Delta_{1,i} - \Delta_{2,i})(\phi(f_i - f_1) - \phi(f_i - f_2)) \\ & + \sum\nolimits_{i=3}^{q} (\Delta_{i,1} - \Delta_{i,2})(\phi(f_1 - f_i) - \phi(f_2 - f_i)). \end{aligned}$$

From Property 4 of Lemma 7, we have

$$\Delta_{1,i} - \Delta_{2,i} - \Delta_{i,1} + \Delta_{i,2} = \Delta_{1,2} - \Delta_{2,1}. \tag{8}$$

This follows

$$
\begin{aligned}
&W(\boldsymbol{p},\boldsymbol{f}) - W(\boldsymbol{p},\boldsymbol{f}') \\
&\quad = \sum\nolimits_{i=3}^{q} (\Delta_{i,1} - \Delta_{i,2})(\phi(f_1 - f_i) + \phi(f_i - f_1) - \phi(f_i - f_2) - \phi(f_2 - f_i)) \\
&\qquad +(\Delta_{1,2} - \Delta_{2,1})(\phi(f_2 - f_1) - \phi(f_1 - f_2)) \\
&\qquad +(\Delta_{1,2} - \Delta_{2,1}) \sum\nolimits_{i=3}^{q} \big(\phi(f_i - f_1) - \phi(f_i - f_2)\big) \\
&\quad = (\Delta_{1,2} - \Delta_{2,1})\big(\phi(f_2 - f_1) - \phi(f_1 - f_2)\big) \\
&\qquad +(\Delta_{1,2} - \Delta_{2,1}) \sum\nolimits_{i=3}^{q} \big(\phi(f_i - f_1) - \phi(f_i - f_2)\big),
\end{aligned}
$$

where the last equality holds from $\phi(t) + \phi(-t) \equiv 2\phi(0)$. For non-increasing function ϕ with $\phi'(0) < 0$, we have $\phi(t) < \phi(-t)$ for all $t > 0$, and this yields

$$(\Delta_{1,2} - \Delta_{2,1})\big(\phi(f_2 - f_1) - \phi(f_1 - f_2)\big) > 0.$$

Meanwhile, we also have $\phi(f_i - f_1) \leq \phi(f_i - f_2)$, which yields

$$(\Delta_{1,2} - \Delta_{2,1}) \sum\nolimits_{i=3}^{q} \big(\phi(f_i - f_1) - \phi(f_i - f_2)\big) \geq 0.$$

Thus, we prove $W(\boldsymbol{p},\boldsymbol{f}) > W(\boldsymbol{p},\boldsymbol{f}')$ which is contrary to $W(\boldsymbol{p},\boldsymbol{f}) = W^*(\boldsymbol{p})$.

We now consider the case $f_1 = f_2$. The subgradient conditions for optimality of $\frac{\partial}{\partial f_i} W(\boldsymbol{p},\boldsymbol{f}) = 0$ for $i = 1, 2$ give

$$
\begin{aligned}
\sum\nolimits_{i\neq 1} \phi'(f_1 - f_i)\Delta_{i,1} &= \sum\nolimits_{i\neq 1} \phi'(f_i - f_1)\Delta_{1,i}, \\
\sum\nolimits_{i\neq 2} \phi'(f_i - f_2)\Delta_{2,i} &= \sum\nolimits_{i\neq 2} \phi'(f_2 - f_i)\Delta_{i,2}.
\end{aligned}
$$

By combining Eqn. (8), $f_1 = f_2$ and $\phi'(t) = \phi'(-t)$ from Eqn. (7), we have

$$(\Delta_{2,1} - \Delta_{1,2})\Big(2\phi'(0) + \sum\nolimits_{i\neq 1,2} \phi'(f_1 - f_i)\Big) = 0,$$

which is contrary to $\Delta_{1,2} < \Delta_{2,1}$ and $2\phi'(0) + \sum_{i\neq 1,2} \phi'(f_1 - f_i) \leq 2\phi'(0) < 0$. Thus, we complete the proof. □

The condition $\phi(t) + \phi(-t) \equiv 2\phi(0)$ is motivated from linear loss and sigmoid-type losses used for neural networks. Based on this theorem, we can easily get the following consistency for sigmoid-type functions.

Corollary 11. *The surrogate loss Ψ given by Eqn. (4) is consistent w.r.t partial ranking loss for sigmoid-type loss functions $\phi(t) = 1/(1+\exp(t))$, $\phi(t) = -\arctan(t)$, etc.*

It is noteworthy that Theorem 10 cannot be applied directly to $\phi(t) = -ct^{2k+1}$ for some constant $c > 0$ and integer $k \geq 0$, because it is not below-bounded. This problem, however, can be solved by introducing a regularization term as in (Duchi et al., 2010). Here, the regularization can be used to control the model complexity, and guarantee that the linear loss is below-bounded. With a proof similar to that of Theorem 10, we get:

Theorem 12. *The following surrogate loss is consistent w.r.t. partial ranking loss:*

$$\Psi(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = \sum\nolimits_{y_i<y_j} a_{\boldsymbol{y}} \phi(f_j(\boldsymbol{x}) - f_i(\boldsymbol{x})) + \tau \Upsilon(\boldsymbol{f}(\boldsymbol{x})),$$

where $\tau > 0$, $\phi(x) = -cx^{2k+1}$ for some constant $c > 0$ and integer $k \geq 0$, and Υ is symmetric, that is,

$$\Upsilon(\ldots, f_i(\boldsymbol{x}), \ldots, f_j(\boldsymbol{x}), \ldots) = \Upsilon(\ldots, f_j(\boldsymbol{x}), \ldots, f_i(\boldsymbol{x}), \ldots).$$

From this theorem, we can easily construct the following convex surrogate loss function:

$$\Psi(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = \sum\nolimits_{y_i<y_j} -a_{\boldsymbol{y}}(f_j(\boldsymbol{x}) - f_i(\boldsymbol{x})) + \tau \sum\nolimits_{i=1}^{q} f_i^2(\boldsymbol{x}),$$

which is consistent w.r.t. partial ranking loss.

It is worth noting that this does not imply that any convex surrogate loss Ψ given by Eqn. (4) is consistent w.r.t. partial ranking loss. In fact, the following theorem proves that, many non-linear surrogate losses are inconsistent w.r.t. partial ranking loss.

Theorem 13. *If $\phi\colon \mathbb{R} \to \mathbb{R}$ is a convex, differentiable, non-linear and non-increasing function, then the surrogate loss Ψ given by Eqn. (4) is inconsistent w.r.t. partial ranking loss.*

Before giving the detailed proof, we first introduce a lemma, whose proof is deferred to Section 6.3.

Lemma 14. *Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be a convex, differentiable, non-linear and non-increasing function. For $b > a > 0$, if the followings hold:*

$$\frac{\phi'(b-a)(\phi'(a)+\phi'(b))+\phi'(b)\phi'(-a)}{\phi'(a-b)(\phi'(a)+\phi'(b))+\phi'(b)\phi'(a)} > 1, \tag{9}$$

$$0 < \frac{\phi'(-a)}{\phi'(a)} \leq \frac{\phi'(-b)}{\phi'(a)} < \frac{\phi'(-b)}{\phi'(b)}, \tag{10}$$

then there exist some $P_1 > P_2 > 0$, $P_3 > 0$ and $P_4 > 0$ such that

$$P_1\phi'(a-b) - P_2\phi'(b-a) = P_4\phi'(b) - P_3\phi'(-b), \tag{11}$$

$$-P_1\phi'(a) + P_2\phi'(-a) = P_4(\phi'(a)+\phi'(b)) - P_3(\phi'(-a)+\phi'(-b)). \tag{12}$$

Proof of Theorem 13 For convex function ϕ we have $\phi'(t_1) \leq \phi'(t_2)$ for every $t_1 \leq t_2$ from (Rockafellar, 1997), and the derivative function $\phi'(t)$ is continuous for $t \in \mathbb{R}$ if ϕ is differentiable and convex. As ϕ is non-increasing, we have $\phi'(t) \leq 0$ for all $t \in \mathbb{R}$, and without loss of generality, we assume $\phi'(t) < 0$.

We proceed by contradiction. Assume the surrogate loss Ψ is consistent with partial ranking loss for some non-linear function ϕ. Then, from the continuity of $\phi'(t)$, there exists a distinguishable (c, d) for $c < d < 0$ or $0 < c < d$, such that

$$\phi'(t_1) < \phi'(t_2) \text{ for every } t_1 < t_2 \text{ and } t_1, t_2 \in (c, d).$$

In the following, we focus on the case $0 < c < d$, and similar consideration can be made for the case $c < d < 0$.

We first fix $a \in (c, d)$ and introduce a new function

$$G(t) = \big(\phi'(t-a) - \phi'(a-t)\big)\big(\phi'(a)+\phi'(t)\big) + \phi'(t)\big(\phi'(-a) - \phi'(a)\big).$$

It is easy to find that $G(t)$ is continuous and

$$G(a) = \phi'(a)\big(\phi'(-a) - \phi'(a)\big) > 0.$$

Thus, there exists $b > a$ and $b \in (c, d)$ such that

$$G(b) = \big(\phi'(b-a) - \phi'(a-b)\big)\big(\phi'(a)+\phi'(b)\big) + \phi'(b)\big(\phi'(-a) - \phi'(a)\big) > 0,$$

which gives

$$\frac{\phi'(b-a)(\phi'(a)+\phi'(b))+\phi'(b)\phi'(-a)}{\phi'(a-b)(\phi'(a)+\phi'(b))+\phi'(b)\phi'(a)} > 1. \tag{13}$$

Moreover, from $\phi'(a) < \phi'(b) < 0$ and $\phi'(-b) \leq \phi'(-a) < 0$, we have

$$0 < \frac{\phi'(-a)}{\phi'(a)} \leq \frac{\phi'(-b)}{\phi'(a)} < \frac{\phi'(-b)}{\phi'(b)}. \tag{14}$$

We consider the following multi-label task with $q = 3$ labels:

$$\boldsymbol{y}_1 = (-1,+1,+1), \boldsymbol{y}_2 = (+1,-1,-1), \boldsymbol{y}_3 = (+1,+1,-1), \boldsymbol{y}_4 = (-1,-1,+1).$$

Let $\boldsymbol{f} = (f_1, f_2, f_3)$ such that $a = f_3 - f_1$ and $b = f_3 - f_2$, and thus $f_1 > f_2$. For every probability simplex $\boldsymbol{p} = (p_{\boldsymbol{y}_1}, p_{\boldsymbol{y}_2}, p_{\boldsymbol{y}_3}, p_{\boldsymbol{y}_4}) \in \Lambda$ and non-negative penalty vector $(a_{\boldsymbol{y}_1}, a_{\boldsymbol{y}_2}, a_{\boldsymbol{y}_3}, a_{\boldsymbol{y}_4})$, we have

$$\begin{aligned} W(\boldsymbol{p}, \boldsymbol{f}) \;=\; & \Delta_{1,2}\phi(f_2 - f_1) + \Delta_{2,1}\phi(f_1 - f_2) + \Delta_{1,3}\phi(f_3 - f_1) \\ & + \Delta_{3,1}\phi(f_1 - f_3) + \Delta_{2,3}\phi(f_3 - f_2) + \Delta_{3,2}\phi(f_2 - f_3), \end{aligned}$$

where $\Delta_{1,2} = P_1$, $\Delta_{2,1} = P_2$, $\Delta_{1,3} = P_1 + P_4$, $\Delta_{3,1} = P_2 + P_3$, $\Delta_{2,3} = P_4$ and $\Delta_{3,2} = P_3$ with $P_i = p_{\boldsymbol{y}_1} a_{\boldsymbol{y}_1}$ for $i = 1, 2, 3, 4$. In the following, we will construct some $\bar{\boldsymbol{p}}$ and $(\bar{a}_1, \bar{a}_2, \bar{a}_3, \bar{a}_4)$ such that $W^*(\bar{\boldsymbol{p}}) = W(\bar{\boldsymbol{p}}, \boldsymbol{f})$ and $\Delta_{1,2} > \Delta_{2,1}$.

The subgradient conditions for optimality of $\frac{\partial W(\boldsymbol{p},\boldsymbol{f})}{\partial f_i} = 0$ $(i = 1, 2, 3)$ give

$$\begin{aligned} & -P_1\phi'(a-b) + P_2\phi'(b-a) - (P_1+P_4)\phi'(a) + (P_2+P_3)\phi'(-a) = 0, \\ & P_1\phi'(a-b) - P_2\phi'(b-a) - P_4\phi'(b) + P_3\phi'(-b) = 0, \\ & (P_1+P_4)\phi'(a) - (P_2+P_3)\phi'(-a) + P_4\phi'(b) - P_3\phi'(-b) = 0, \end{aligned}$$

which are equivalent to

$$\begin{aligned} & P_1\phi'(a-b) - P_2\phi'(b-a) = P_4\phi'(b) - P_3\phi'(-b), \\ & -P_1\phi'(a) + P_2\phi'(-a) = P_4(\phi'(a)+\phi'(b)) - P_3(\phi'(-a)+\phi'(-b)). \end{aligned}$$

Lemma 14 shows that there exist $\bar{\boldsymbol{p}} = (\bar{p}_{\boldsymbol{y}_1}, \bar{p}_{\boldsymbol{y}_2}, \bar{p}_{\boldsymbol{y}_3}, \bar{p}_{\boldsymbol{y}_4})$ and $(\bar{a}_{\boldsymbol{y}_1}, \bar{a}_{\boldsymbol{y}_2}, \bar{a}_{\boldsymbol{y}_3}, \bar{a}_{\boldsymbol{y}_4})$ s.t. the above hold and $P_1 > P_2$ from Eqns. (13) and (14). Hence this yields $W(\bar{\boldsymbol{p}}, \boldsymbol{f}) = W^*(\bar{\boldsymbol{p}})$. We also have $\boldsymbol{f} \notin \mathcal{A}(\bar{\boldsymbol{p}})$ from $\Delta_{1,2} = P_1 > P_2 = \Delta_{2,1}$ yet $f_1 > f_2$; thus, $W^*(\bar{\boldsymbol{p}}) = \inf_{\boldsymbol{f}}\{W(\bar{\boldsymbol{p}}, \boldsymbol{f}) \colon \boldsymbol{f} \notin \mathcal{A}(\bar{\boldsymbol{p}})\}$, and the theorem holds.□

Based on this theorem, many state-of-the-art multi-label learning approaches (Schapire and Singer, 2000; Dekel et al., 2004; Zhang and Zhou, 2006) are proven to be inconsistent w.r.t. partial ranking loss.

Corollary 15. *The surrogate loss* Ψ *given by Eqn.* (4) *is inconsistent w.r.t. partial ranking loss for exponential loss* $\phi(t) = \exp(-t)$*, logistic loss* $\phi(t) = \ln(1 + \exp(-t))$ *and least square hinge loss* $\phi(t) = (\max(0, 1 - t))^2$.

It is noteworthy that Theorem 13 cannot be used directly to study the consistency of hinge loss because it is non-differentiable, whereas the following theorem shows the inconsistency for hinge loss:

Theorem 16. *For hinge loss* $\phi(t) = \max(0, 1-t)$*, the surrogate loss* Ψ *given by Eqn.* (4) *is inconsistent w.r.t. partial ranking loss.*

Proof: We consider the following multi-label task with $q = 3$ labels:

$$\boldsymbol{y}_1 = (-1, +1, +1), \boldsymbol{y}_2 = (+1, -1, -1), \boldsymbol{y}_3 = (+1, +1, -1), \boldsymbol{y}_4 = (-1, -1, +1),$$

and focus on the probability simplex $\boldsymbol{p} = (p_{\boldsymbol{y}_1}, p_{\boldsymbol{y}_2}, p_{\boldsymbol{y}_3}, p_{\boldsymbol{y}_4}) \in \Lambda$ and non-negative penalty vector $(a_{\boldsymbol{y}_1}, a_{\boldsymbol{y}_2}, a_{\boldsymbol{y}_3}, a_{\boldsymbol{y}_4})$ such that $P_2 < P_1 \leq 2P_2$, $P_1 + P_3 < P_2 + P_4$ and $P_3 < P_4$, where $P_i = p_{\boldsymbol{y}_i} a_{\boldsymbol{y}_i}$ for $1 \leq i \leq 4$. For every $\boldsymbol{f} = (f_1, f_2, f_3)$, we have

$$\begin{aligned} W(\boldsymbol{p}, \boldsymbol{f}) &= \Delta_{1,2}\phi(f_2 - f_1) + \Delta_{2,1}\phi(f_1 - f_2) + \Delta_{1,3}\phi(f_3 - f_1) \\ &\quad + \Delta_{3,1}\phi(f_1 - f_3) + \Delta_{2,3}\phi(f_3 - f_2) + \Delta_{3,2}\phi(f_2 - f_3), \end{aligned}$$

where $\Delta_{1,2} = P_1$, $\Delta_{2,1} = P_2$, $\Delta_{1,3} = P_1 + P_4$, $\Delta_{3,1} = P_2 + P_3$, $\Delta_{2,3} = P_4$ and $\Delta_{3,2} = P_3$. Minimizing $W(\boldsymbol{p}, \boldsymbol{f})$ gives the optimal solution $\boldsymbol{f} = (f_1, f_2, f_3)$ such that $f_1 = f_2 = f_3 - 1$. This gives $\boldsymbol{f} \notin \mathcal{A}(\boldsymbol{p})$ from $\Delta_{1,2} = P_1 > P_2 = \Delta_{2,1}$. Thus, we have $W^*(\boldsymbol{p}) = \inf_{\boldsymbol{f} \notin \mathcal{A}(\boldsymbol{p})} W(\boldsymbol{p}, \boldsymbol{f})$, and this theorem follows. □

Also, the following theorem shows that the least square loss is inconsistent with partial ranking loss:

Theorem 17. *For least square loss* $\phi(t) = (1 - t)^2$*, the surrogate loss* Ψ *given by Eqn.* (4) *is inconsistent w.r.t. partial ranking loss.*

Proof: We consider the following multi-label task with $q = 3$ labels:

$$\boldsymbol{y}_1 = (-1, +1, +1), \boldsymbol{y}_2 = (+1, -1, -1), \boldsymbol{y}_3 = (+1, +1, -1),$$

and focus on the probability simplex $\boldsymbol{p} = (p_{\boldsymbol{y}_1}, p_{\boldsymbol{y}_2}, p_{\boldsymbol{y}_3}) \in \Lambda$ and non-negative penalty vector $(a_{\boldsymbol{y}_1}, a_{\boldsymbol{y}_2}, a_{\boldsymbol{y}_3})$ such that $P_2 = 3P_1/2$ and $P_3 > 5P_1/4$, where $P_i = p_{\boldsymbol{y}_i} a_{\boldsymbol{y}_i} > 0$ for $1 \leq i \leq 3$. For least square loss, we have

$$\begin{aligned} W(\boldsymbol{p}, \boldsymbol{f}) &= P_1(1 - f_2 + f_1)^2 + P_2(1 - f_1 + f_2)^2 \\ &\quad + P_1(1 - f_3 + f_1)^2 + (P_2 + P_3)(1 - f_1 + f_3)^2 + P_3(1 - f_2 + f_3)^2. \end{aligned}$$

The subgradient conditions for optimality of $\frac{\partial W(\boldsymbol{p},\boldsymbol{f})}{\partial f_i} = 0$ $(i = 1, 2, 3)$ give the optimal solution $\boldsymbol{f}^* = (f_1^*, f_2^*, f_3^*)$ such that

$$f_1^* - f_2^* = (P_2^2 - P_1^2 + 2P_3(P_2 - 2P_1))/\kappa = P_1(5/4P_1 - P_3)/\kappa < 0$$

where $\kappa = (P_1 + P_2 + P_3)^2 + P_3(P_1 + P_2)$ and we use $P_2 = 3P_1/2$ and $P_3 > 5P_1/4$. This theorem follows since $\boldsymbol{f} \notin \mathcal{A}(\boldsymbol{p})$ from $\Delta_{1,2} = P_1 < P_2 = \Delta_{2,1}$. $\square$

4.2. Consistency of Univariate Surrogate Loss

Now we consider the univariate surrogate loss as follows:

$$\Psi(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = a_{\boldsymbol{y}} \sum\nolimits_{i=1}^{q} \phi(y_i f_i(\boldsymbol{x})), \tag{15}$$

where ϕ is a convex function, e.g., exponential loss $\phi(t) = \exp(-t)$, logistic loss $\phi(t) = \ln(1+\exp(-t))$, hinge loss $\phi(t) = \max(0, 1-t)$, etc. We have the following sufficient condition for consistency of univariate surrogate loss:

Theorem 18. *If $\phi\colon \mathbb{R} \to \mathbb{R}$ is a convex, non-increasing and differentiable function with $\phi'(0) < 0$, then the univariate surrogate loss Ψ of Eqn.* (15) *is consistent w.r.t. partial ranking loss.*

Proof: For every probability simplex $\boldsymbol{p} = (p_{\boldsymbol{y}})_{\boldsymbol{y}\in\mathcal{Y}}$ and non-negative vector $(a_{\boldsymbol{y}})_{\boldsymbol{y}\in\mathcal{Y}}$, we have

$$W(\boldsymbol{p}, \boldsymbol{f}) = \sum\nolimits_{\boldsymbol{y}\in\mathcal{Y}} p_{\boldsymbol{y}} a_{\boldsymbol{y}} \sum\nolimits_{i=1}^{q} \phi(y_i f_i) = \sum\nolimits_{i=1}^{q} \Delta_i^- \phi(-f_i) + (\Delta - \Delta_i^-)\phi(f_i).$$

It suffices to prove that $f_i > f_j$ if $\Delta_{i,j} < \Delta_{j,i}$ for every $\boldsymbol{f}$ such that $W^*(\boldsymbol{p}) = W(\boldsymbol{p}, \boldsymbol{f})$. Without loss of generality, we will prove that $f_1 < f_2$ if $\Delta_{1,2} > \Delta_{2,1}$.

From Property 2 in Lemma 7, we have $\Delta_1^- - \Delta_2^- = \Delta_{1,2} - \Delta_{2,1}$, and if $\Delta_{1,2} > \Delta_{2,1}$, we have

$$\Delta_1^- > \Delta_2^- \Rightarrow \Delta/\Delta_1^- < \Delta/\Delta_2^- \Rightarrow (\Delta - \Delta_1^-)/\Delta_1^- < (\Delta - \Delta_2^-)/\Delta_2^-.$$

The subgradient conditions for optimality of $\frac{\partial W(\boldsymbol{p},\boldsymbol{f})}{\partial f_i} = 0$ $(i = 1, 2)$ give

$$\Delta_1^- \phi'(-f_1) = (\Delta - \Delta_1^-)\phi'(f_1) \quad \text{and} \quad \Delta_2^- \phi'(-f_2) = (\Delta - \Delta_2^-)\phi'(f_2).$$

This yields that

$$\phi'(-f_1) < 0, \phi'(f_1) < 0, \phi'(-f_2) < 0, \phi'(f_2) < 0,$$

for non-increasing function ϕ with $\phi'(0) < 0$. For convex function ϕ, its derivative $\phi'(t)$ is non-decreasing from (Rockafellar, 1997). Therefore, if $f_1 \geq f_2$ then $\phi'(f_1) \geq \phi'(f_2)$, and we further have

$$\phi'(-f_1) = \frac{\Delta - \Delta_1^-}{\Delta_1^-}\phi'(f_1) > \frac{\Delta - \Delta_2^-}{\Delta_2^-}\phi'(f_1) \geq \frac{\Delta - \Delta_2^-}{\Delta_2^-}\phi'(f_2) = \phi'(-f_2),$$

i.e., $\phi'(-f_1) > \phi'(-f_2)$, leading to $f_1 < f_2$, whereas it is contrary. The theorem follows as desired. □

Based on this theorem, it is easy to get the consistency of the univariate surrogate loss with exponential loss, logistic loss and least square hinge loss.

Corollary 19. *The surrogate loss Ψ of Eqn.* (15) *is consistent w.r.t. partial ranking loss for exponential loss $\phi(t) = \exp(-t)$, logistic loss $\phi(t) = \ln(1 + \exp(-t))$ and least square hinge loss $\phi(t) = (\max(0, 1 - t))^2$.*

It is noteworthy that, after we published our preliminary work (Gao and Zhou, 2011), Dembczyński et al. (2012a) proved that the following surrogate losses are consistent w.r.t. partial ranking loss (the partial ranking loss was referred to as *ranking loss* in (Dembczyński et al., 2012a))

$$\begin{aligned}\Psi(\boldsymbol{f}, \boldsymbol{y}) &= a_{\boldsymbol{y}} \sum\nolimits_{i=1}^{q} \exp(-y_i f_i), \\ \Psi(\boldsymbol{f}, \boldsymbol{y}) &= a_{\boldsymbol{y}} \sum\nolimits_{i=1}^{q} \ln(1 + \exp(-y_i f_i)),\end{aligned}$$

and they further derived their corresponding consistent bounds. It is clear that these results of (Dembczyński et al., 2012a) are special cases of Theorem 18.

Notice that the least square loss $\phi(t) = (1 - t)^2$ increases for $t > 1$; therefore, Theorem 18 cannot be applied to least square loss, whereas we can obtain its consistency with partial ranking loss from the following theorem:

Theorem 20. *For integer $k > 0$ and $\phi(t) = (1 - t)^{2k}$, the univariate surrogate loss Ψ of Eqn.* (15) *is consistent w.r.t. partial ranking loss.*

Proof: Similarly to the proof of Theorem 18, we will prove that $f_1 < f_2$ if $\Delta_{1,2} > \Delta_{2,1}$. The subgradient conditions for optimality of $\frac{\partial W(\boldsymbol{p}, \boldsymbol{f})}{\partial f_i} = 0$ $(i = 1, 2)$ give

$$\Delta_1^- \phi'(-f_1) = (\Delta - \Delta_1^-)\phi'(f_1) \quad \text{and} \quad \Delta_2^- \phi'(-f_2) = (\Delta - \Delta_2^-)\phi'(f_2),$$

which implies

$$\left(\frac{1+f_1}{1-f_1}\right)^{2k-1} = \frac{\Delta}{\Delta_1^-} - 1 \text{ and } \left(\frac{1+f_2}{1-f_2}\right)^{2k-1} = \frac{\Delta}{\Delta_2^-} - 1, \tag{16}$$

respectively. From Property 2 in Lemma 7, we have $\Delta_1^- - \Delta_2^- = \Delta_{1,2} - \Delta_{2,1}$, and this follows that

$$\Delta_1^- > \Delta_2^- \Rightarrow \Delta/\Delta_1^- < \Delta/\Delta_2^-$$

from $\Delta_{1,2} > \Delta_{2,1}$. Therefore, we have $f_1 < f_2$ from Eqn. (16), and this completes the proof. □

It is also noteworthy that the hinge loss $\phi(t) = \max(0, 1-t)$ is not differentiable at $t = 1$; therefore, Theorem 18 cannot be used to study the consistency of hinge loss. The following theorem illustrates the difficulties for consistency without differentiability even if ϕ is a convex and non-increasing function with $\phi'(0) < 0$.

Theorem 21. *For hinge loss $\phi(t) = \max(0, 1-t)$, the univariate surrogate loss Ψ of Eqn.* (15) *is inconsistent w.r.t. partial ranking loss.*

Proof: We consider a multi-label task with $q = 2$ labels:

$$\boldsymbol{y}_1 = (-1,-1), \boldsymbol{y}_2 = (-1,+1), \boldsymbol{y}_3 = (+1,-1), \boldsymbol{y}_4 = (+1,+1),$$

and focus on the probability simplex $\boldsymbol{p} = (p_{\boldsymbol{y}_1}, p_{\boldsymbol{y}_2}, p_{\boldsymbol{y}_3}, p_{\boldsymbol{y}_4})$ and non-negative vector $(a_{\boldsymbol{y}_1}, a_{\boldsymbol{y}_2}, a_{\boldsymbol{y}_3}, a_{\boldsymbol{y}_4})$ such that $p_{\boldsymbol{y}_1} a_{\boldsymbol{y}_1}/2 > p_{\boldsymbol{y}_2} a_{\boldsymbol{y}_2} > p_{\boldsymbol{y}_3} a_{\boldsymbol{y}_3} > p_{\boldsymbol{y}_4} a_{\boldsymbol{y}_4}$. From Eqn. (15), we have

$$\begin{aligned} W(\boldsymbol{p}, \boldsymbol{f}) \;=\; & (p_{\boldsymbol{y}_1} a_{\boldsymbol{y}_1} + p_{\boldsymbol{y}_2} a_{\boldsymbol{y}_2})\phi(-f_1) + (p_{\boldsymbol{y}_3} a_{\boldsymbol{y}_3} + p_{\boldsymbol{y}_4} a_{\boldsymbol{y}_4})\phi(f_1) \\ & + (p_{\boldsymbol{y}_1} a_{\boldsymbol{y}_1} + p_{\boldsymbol{y}_3} a_{\boldsymbol{y}_3})\phi(-f_2) + (p_{\boldsymbol{y}_2} a_{\boldsymbol{y}_2} + p_{\boldsymbol{y}_4} a_{\boldsymbol{y}_4})\phi(f_2). \end{aligned}$$

Minimizing $W(\boldsymbol{p}, \boldsymbol{f})$ gives the optimal solution $\boldsymbol{f} = (f_1, f_2) = (-1, -1)$, i.e., $f_1 = f_2$, yet $\Delta_{1,2} = p_{\boldsymbol{y}_2} a_{\boldsymbol{y}_2} > p_{\boldsymbol{y}_3} a_{\boldsymbol{y}_3} = \Delta_{2,1}$. This implies $\boldsymbol{f} \notin \mathcal{A}(\boldsymbol{p})$, which completes the proof. □

Many univariate surrogate losses are proven to be consistent with partial ranking loss, although they solve the problems by a series of independent (weighted) binary classifications; in other words, the success of approaches optimizing univariate surrogate losses requires that a multi-label learning problem can be learned well by decomposing them into a series of binary classification problems.

5. Consistency w.r.t. Hamming Loss

The hamming loss concerns about how many instance-label pairs are misclassified. For a given vector $\boldsymbol{f}$ and prediction function F, the hamming loss is given by

$$L_{\text{hamloss}}(F(\boldsymbol{f}(\boldsymbol{x})), \boldsymbol{y}) = \frac{1}{q} \sum\nolimits_{i=1}^{q} I[\hat{y}_i \neq y_i],$$

where $\hat{\boldsymbol{y}} = F(\boldsymbol{f}(\boldsymbol{x})) = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_q)$. Hamming loss is below-bounded from $L_{\text{hamloss}}(F(\boldsymbol{f}(\boldsymbol{x})), \boldsymbol{y}) \geq 0$, and distinguishable because it holds that

$$\begin{gathered} L_{\text{hamloss}}(F(\boldsymbol{f}(\boldsymbol{x})), \boldsymbol{y})) = L_{\text{hamloss}}(F(\boldsymbol{f}(\boldsymbol{x}')), \boldsymbol{y}')) \;\; \text{or} \\ |L_{\text{hamloss}}(F(\boldsymbol{f}(\boldsymbol{x})), \boldsymbol{y})) - L_{\text{hamloss}}(F(\boldsymbol{f}(\boldsymbol{x}')), \boldsymbol{y}'))| \geq 1/q, \end{gathered}$$

for every $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ and $\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}$. Further, we have the conditional risk

$$\begin{aligned} l(\boldsymbol{p}, F(\boldsymbol{f}(\boldsymbol{x}))) &= \sum\nolimits_{\boldsymbol{y} \in \mathcal{Y}} p_{\boldsymbol{y}} L_{\text{hamloss}}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{1}{q} \sum\nolimits_{\boldsymbol{y} \in \mathcal{Y}} p_{\boldsymbol{y}} \sum\nolimits_{i=1}^{q} I[\hat{y}_i \neq y_i] \\ &= \frac{1}{q} \sum\nolimits_{i=1}^{q} \Big(\sum\nolimits_{\boldsymbol{y}\colon\, y_i = +1} p_{\boldsymbol{y}} I[\hat{y}_i \neq +1] + \sum\nolimits_{\boldsymbol{y}\colon\, y_i = -1} p_{\boldsymbol{y}} I[\hat{y}_i \neq -1] \Big), \end{aligned}$$

and for hamming loss, the set of Bayes predictions is given by

$$\mathcal{A}(\boldsymbol{p}) = \Big\{ \boldsymbol{f} = \boldsymbol{f}(\boldsymbol{x})\colon \hat{\boldsymbol{y}} = F(\boldsymbol{f}) \text{ with } \hat{y}_i = \operatorname{sgn}\big(\sum\nolimits_{\boldsymbol{y}\colon\, y_i = +1} p_{\boldsymbol{y}} - \frac{1}{2} \big) \Big\}. \quad (17)$$

5.1. Consistency of Multi-Class Extensions

It is possible to solve a multi-label problem by regarding each subset of labels as a meta-class and then try to learn 2^q functions, i.e., $\boldsymbol{f} = (f_{\boldsymbol{y}})_{\boldsymbol{y} \in \mathcal{Y}}$. Then, a prediction function is given by

$$F(\boldsymbol{f}(\boldsymbol{x})) = \max_{\boldsymbol{y} \in \mathcal{Y}} f_{\boldsymbol{y}}(\boldsymbol{x}). \quad (18)$$

Motivated from multi-class learning, it is natural to consider the following surrogate losses.

- One-vs-all:

$$\Psi(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = \phi(f_{\boldsymbol{y}}(\boldsymbol{x}) - \max_{\hat{\boldsymbol{y}} \neq \boldsymbol{y}} f_{\hat{\boldsymbol{y}}}(\boldsymbol{x})), \quad (19)$$

  where ϕ is an appropriately chosen function. This formulation has been used for multi-label learning (Taskar et al., 2004; Hariharan et al., 2010) and multi-class learning (Crammer and Singer, 2001).

- Pairwise comparison:

$$\Psi(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = \sum\nolimits_{\hat{\boldsymbol{y}} \neq \boldsymbol{y}} \phi(f_{\boldsymbol{y}}(\boldsymbol{x}) - f_{\hat{\boldsymbol{y}}}(\boldsymbol{x})), \tag{20}$$

  where ϕ is a convex function such as $\phi(t) = \max(0, 1-t)$ in multi-class learning (Weston and Watkins, 1998).

The following two theorems show that neither the one-vs-all method nor the pairwise comparison method is consistent w.r.t. hamming loss.

Theorem 22. *If ϕ is a non-increasing function with $\phi'(0) < 0$, then the one-vs-all method of Eqn.* (19) *is inconsistent with hamming loss.*

Theorem 23. *If ϕ is a non-increasing function with $\phi'(0) < 0$, then the pairwise comparison method of Eqn.* (20) *is inconsistent with hamming loss.*

Based on Theorems 22 and 23, it is clear that neither the one-vs-all method of Eqn. (19) nor the pairwise comparison method of Eqn. (20) is consistent w.r.t. hamming loss for many commonly-used loss functions, e.g., exponential loss $\phi(t) = \exp(-t)$, logistic loss $\phi(t) = \ln(1 + \exp(-t))$, hinge loss $\phi(t) = \max(0, 1-t)$, least square hinge loss $\phi(t) = (\max(0, 1-t))^2$, etc.

It is also noteworthy that the pairwise comparison method of Eqn. (20) is consistent in multi-class learning (Zhang, 2004a, Theorem 6) for exponential loss, logistic loss, hinge loss, etc., whereas in multi-label learning, Theorem 23 shows their inconsistency.

Proofs of Theorems 22 and 23 We consider a multi-label task with $q = 2$ labels:

$$\boldsymbol{y}_1 = (-1, -1), \boldsymbol{y}_2 = (-1, +1), \boldsymbol{y}_3 = (+1, +1), \boldsymbol{y}_4 = (+1, -1),$$

and focus on the probability simplex $\boldsymbol{p} = (p_{y_1}, p_{y_2}, p_{y_3}, p_{y_4})$ such that $p_{\boldsymbol{y}_1} > p_{\boldsymbol{y}_2} > p_{\boldsymbol{y}_3} > p_{\boldsymbol{y}_4}$, $p_{\boldsymbol{y}_1} + p_{\boldsymbol{y}_4} < p_{\boldsymbol{y}_2} + p_{\boldsymbol{y}_3}$ and $p_{\boldsymbol{y}_1} + p_{\boldsymbol{y}_2} + p_{\boldsymbol{y}_3} + p_{\boldsymbol{y}_4} = 1$. From Eqns. (17) and (18), it is easy to get the set of Bayes predictions

$$\mathcal{A}(\boldsymbol{p}) = \{\boldsymbol{f} \colon f_{\boldsymbol{y}_2} > f_{\boldsymbol{y}_i} \text{ for } i = 1, 3, 4\}.$$

For Theorem 22, we have, from Eqn. (19),

$$W(\boldsymbol{p}, \boldsymbol{f}) = \sum\nolimits_{i=1}^{4} p_{\boldsymbol{y}_i} \phi(f_{\boldsymbol{y}_i} - \max_{j \neq i} f_{\boldsymbol{y}_j}).$$

We complete the proof by showing that $\boldsymbol{f} \notin \mathcal{A}(\boldsymbol{p})$ for every $\boldsymbol{f}$ s.t. $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$. Assume that there exists $\boldsymbol{f} = (f_{\boldsymbol{y}_1}, f_{\boldsymbol{y}_2}, f_{\boldsymbol{y}_3}, f_{\boldsymbol{y}_4}) \in \mathcal{A}(\boldsymbol{p})$ and $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$. We construct another $\hat{\boldsymbol{f}} = (f_{\boldsymbol{y}_2}, f_{\boldsymbol{y}_1}, f_{\boldsymbol{y}_3}, f_{\boldsymbol{y}_4})$, and get

$$W(\boldsymbol{p}, \boldsymbol{f}) - W(\boldsymbol{p}, \hat{\boldsymbol{f}}) = (p_{\boldsymbol{y}_1} - p_{\boldsymbol{y}_2})\big(\phi(f_{\boldsymbol{y}_1} - \max_{i \neq 1} f_{\boldsymbol{y}_i}) - \phi(f_{\boldsymbol{y}_2} - \max_{i \neq 2} f_{\boldsymbol{y}_i})\big). \quad (21)$$

From our assumption $\boldsymbol{f} \in \mathcal{A}(\boldsymbol{p})$, we have $f_{\boldsymbol{y}_2} > \max_{i \neq 2} f_{\boldsymbol{y}_i}$ and

$$f_{\boldsymbol{y}_1} - f_{\boldsymbol{y}_2} < 0 \leq f_{\boldsymbol{y}_2} - \max_{i \neq 2} f_{\boldsymbol{y}_i}.$$

For non-increasing function ϕ with $\phi'(0) < 0$, it holds that

$$\phi(f_{\boldsymbol{y}_1} - f_{\boldsymbol{y}_2}) > \phi(f_{\boldsymbol{y}_2} - \max_{i \neq 2} f_{\boldsymbol{y}_i}).$$

From Eqn. (21) and $p_{\boldsymbol{y}_1} > p_{\boldsymbol{y}_2}$, we have $W(\boldsymbol{p}, \boldsymbol{f}) > W(\boldsymbol{p}, \hat{\boldsymbol{f}})$, which is contrary to $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$. Thus, Theorem 22 follows.

For Theorem 23, we have, from Eqn. (20),

$$W(\boldsymbol{p}, \boldsymbol{f}) = \sum\nolimits_{i=1}^{4} p_{\boldsymbol{y}_i} \sum\nolimits_{j \neq i} \phi(f_{\boldsymbol{y}_i} - f_{\boldsymbol{y}_j}).$$

Therefore, we complete the proof by showing that $\boldsymbol{f} \notin \mathcal{A}(\boldsymbol{p})$ for every $\boldsymbol{f}$ such that $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$. Suppose that there exists $\boldsymbol{f} = (f_{\boldsymbol{y}_1}, f_{\boldsymbol{y}_2}, f_{\boldsymbol{y}_3}, f_{\boldsymbol{y}_4}) \in \mathcal{A}(\boldsymbol{p})$ such that $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$. Then, we can construct another $\hat{\boldsymbol{f}} = (f_{\boldsymbol{y}_2}, f_{\boldsymbol{y}_1}, f_{\boldsymbol{y}_3}, f_{\boldsymbol{y}_4})$, and get

$$\begin{aligned} &W(\boldsymbol{p}, \boldsymbol{f}^*) - W(\boldsymbol{p}, \hat{\boldsymbol{f}}) = (p_{\boldsymbol{y}_1} - p_{\boldsymbol{y}_2}) \times \\ &\Big(\phi(f_{\boldsymbol{y}_1} - f_{\boldsymbol{y}_2}) - \phi(f_{\boldsymbol{y}_2} - f_{\boldsymbol{y}_1}) + \sum\nolimits_{i=3}^{4} \phi(f_{\boldsymbol{y}_1} - f_{\boldsymbol{y}_i}) - \phi(f_{\boldsymbol{y}_2} - f_{\boldsymbol{y}_i})\Big). \quad (22) \end{aligned}$$

From our assumption $\boldsymbol{f} \in \mathcal{A}(\boldsymbol{p})$, we have $f_{\boldsymbol{y}_1} < f_{\boldsymbol{y}_2}$. For non-increasing function ϕ, we have

$$\phi(f_{\boldsymbol{y}_1} - f_{\boldsymbol{y}_i}) \geq \phi(f_{\boldsymbol{y}_2} - f_{\boldsymbol{y}_i}) \text{ for } i = 3, 4;$$

meanwhile, we also have $\phi(f_{\boldsymbol{y}_1} - f_{\boldsymbol{y}_2}) > \phi(f_{\boldsymbol{y}_2} - f_{\boldsymbol{y}_1})$ from $\phi'(0) < 0$. From Eqn. (22), we have $W(\boldsymbol{p}, \boldsymbol{f}) > W(\boldsymbol{p}, \hat{\boldsymbol{f}})$, contrary to $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$. This completes the proof. □

Notice that Theorems 22 and 23 cannot be applied directly to least square loss $\phi(t) = (1 - t)^2$ because it is increasing for $t > 1$, whereas we have:

Theorem 24. *For least square loss $\phi(t) = (1-t)^2$, neither the one-vs-all method of Eqn.* (19) *nor the pairwise comparison method of Eqn.* (20) *is consistent w.r.t. hamming loss.*

Proof: Similarly to the proofs of Theorems 22 and 23, we consider the multi-label task with $q = 2$ labels:

$$\boldsymbol{y}_1 = (-1,-1), \boldsymbol{y}_2 = (-1,+1), \boldsymbol{y}_3 = (+1,+1), \boldsymbol{y}_4 = (+1,-1),$$

with $p_{\boldsymbol{y}_1} > p_{\boldsymbol{y}_2} > p_{\boldsymbol{y}_3} > p_{\boldsymbol{y}_4}$, $p_{\boldsymbol{y}_1} + p_{\boldsymbol{y}_4} < p_{\boldsymbol{y}_2} + p_{\boldsymbol{y}_3}$ and $p_{\boldsymbol{y}_1} + p_{\boldsymbol{y}_2} + p_{\boldsymbol{y}_3} + p_{\boldsymbol{y}_4} = 1$, and it is easy to get the set of Bayes predictions

$$\mathcal{A}(\boldsymbol{p}) = \{\boldsymbol{f} \colon f_{\boldsymbol{y}_2} > f_{\boldsymbol{y}_i} \text{ for } i = 1, 3, 4\}.$$

For the one-vs-all method of Eqn. (19), our proof is rather similar to the proof of Theorem 22. From Eqn. (21), we have

$$W(\boldsymbol{p}, \boldsymbol{f}) - W(\boldsymbol{p}, \hat{\boldsymbol{f}}) = (p_{\boldsymbol{y}_1} - p_{\boldsymbol{y}_2})\big(\phi(f_{\boldsymbol{y}_1} - \max_{i \neq 1} f_{\boldsymbol{y}_i}) - \phi(f_{\boldsymbol{y}_2} - \max_{i \neq 2} f_{\boldsymbol{y}_i})\big),$$

and, for least square loss $\phi(t) = (1-t)^2$, we have $\phi(f_{\boldsymbol{y}_1} - \max_{i \neq 1} f_{\boldsymbol{y}_i}) > \phi(f_{\boldsymbol{y}_2} - \max_{i \neq 2} f_{\boldsymbol{y}_i})$ from the condition $f_{\boldsymbol{y}_2} > \max_{i \neq 2} f_{\boldsymbol{y}_i}$. This is contrary to $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$.

For the pairwise comparison method of Eqn. (20), we have

$$W(\boldsymbol{p}, \boldsymbol{f}) = \sum\nolimits_{i=1}^{4} p_{\boldsymbol{y}_i} \sum\nolimits_{j \neq i} \phi(f_{\boldsymbol{y}_i} - f_{\boldsymbol{y}_j}),$$

and the subgradient conditions for optimality of $\frac{\partial}{\partial f_i} W(\boldsymbol{p}, \boldsymbol{f}) = 0$ $(1 \leq i \leq 4)$ give $\boldsymbol{f}^* = (f_1^*, f_2^*, f_3^*, f_4^*)$ such that

$$f_1^* - f_2^* = (p_{\boldsymbol{y}_1} - p_{\boldsymbol{y}_2})(p_{\boldsymbol{y}_1} + p_{\boldsymbol{y}_2} + 5p_{\boldsymbol{y}_3} + p_{\boldsymbol{y}_4})(p_{\boldsymbol{y}_1} + p_{\boldsymbol{y}_2} + p_{\boldsymbol{y}_3} + 5p_{\boldsymbol{y}_4})/\kappa,$$

where $\kappa = p_{\boldsymbol{y}_1}^3 + p_{\boldsymbol{y}_2}^3 + p_{\boldsymbol{y}_3}^3 + p_{\boldsymbol{y}_4}^3 + 16(p_{\boldsymbol{y}_1}p_{\boldsymbol{y}_2}p_{\boldsymbol{y}_3} + p_{\boldsymbol{y}_1}p_{\boldsymbol{y}_2}p_{\boldsymbol{y}_4} + p_{\boldsymbol{y}_1}p_{\boldsymbol{y}_3}p_{\boldsymbol{y}_4} + p_{\boldsymbol{y}_2}p_{\boldsymbol{y}_3}p_{\boldsymbol{y}_4}) + 5p_{\boldsymbol{y}_1}^2(p_{\boldsymbol{y}_2} + p_{\boldsymbol{y}_3} + p_{\boldsymbol{y}_4}) + 5p_{\boldsymbol{y}_2}^2(p_{\boldsymbol{y}_1} + p_{\boldsymbol{y}_3} + p_{\boldsymbol{y}_4}) + 5p_{\boldsymbol{y}_3}^2(p_{\boldsymbol{y}_1} + p_{\boldsymbol{y}_2} + p_{\boldsymbol{y}_4}) + 5p_{\boldsymbol{y}_4}^2(p_{\boldsymbol{y}_1} + p_{\boldsymbol{y}_2} + p_{\boldsymbol{y}_3})$. This leads to $f_1^* > f_2^*$, implying that $\boldsymbol{f}^* \notin \mathcal{A}(\boldsymbol{p})$. Thus, this theorem follows as desired. □

It is interesting to further understand why those algorithms are inconsistent. Intuitively, the prediction rule $F(\boldsymbol{f}(\boldsymbol{x})) = \max_{\boldsymbol{y} \in \mathcal{Y}} f_{\boldsymbol{y}}(\boldsymbol{x})$ prefers to

choosing $\boldsymbol{y} \in \arg\max\{p_{\boldsymbol{y}} \colon \boldsymbol{y} \in \mathcal{Y}\}$, whereas for hamming loss, Eqn. (17) gives the set of Bayes predictions

$$\mathcal{A}(\boldsymbol{p}) = \Big\{\boldsymbol{f} = \boldsymbol{f}(\boldsymbol{x}) \colon \hat{\boldsymbol{y}} = F(\boldsymbol{f}) \text{ where } \hat{y}_i = \operatorname{sgn}\big(\sum\nolimits_{\boldsymbol{y} \colon y_i = +1} p_{\boldsymbol{y}} - \frac{1}{2}\big)\Big\}.$$

In practice, it does not always hold that

$$\{\boldsymbol{y} \colon \boldsymbol{y} \in \arg\max\{p_{\boldsymbol{y}}\}\} = \Big\{\boldsymbol{y} \colon y_i = \operatorname{sgn}\big(\sum\nolimits_{\boldsymbol{y} \colon y_i = +1} p_{\boldsymbol{y}} - \frac{1}{2}\big)\Big\};$$

this may explain why these algorithms are inconsistent. It leads us to consider other prediction rules such as $F(\boldsymbol{f}(\boldsymbol{x})) = \hat{\boldsymbol{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_q)$ with $\hat{y}_i = \operatorname{sgn}(\sum_{\boldsymbol{y} \in \mathcal{Y}, y_i = +1} f_{\boldsymbol{y}} - \sum_{\boldsymbol{y} \in \mathcal{Y}, y_i = -1} f_{\boldsymbol{y}})$, and we leave it to future work.

Finally, it is interesting to consider the following formulation

$$\Psi(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = \max_{\hat{\boldsymbol{y}} \neq \boldsymbol{y}} \phi(\delta(\hat{\boldsymbol{y}}, \boldsymbol{y}) + f_{\hat{\boldsymbol{y}}}(\boldsymbol{x}) - f_{\boldsymbol{y}}(\boldsymbol{x})), \tag{23}$$

where $\phi(t) = \max(0, t)$ and $\delta(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sum_{i=1}^{q} I[y_i \neq \hat{y}_i]$; this formulation has been used in (Taskar et al., 2004; Hariharan et al., 2010). It is easy to see that these approaches (Taskar et al., 2004; Hariharan et al., 2010) are variations of the one-vs-all method, because Eqn. (23) degenerates to Eqn. (19) by setting $\delta(\boldsymbol{y}, \hat{\boldsymbol{y}}) = I[\boldsymbol{y} \neq \hat{\boldsymbol{y}}]$. We have

Theorem 25. *The surrogate loss Ψ of Eqn. (23) is inconsistent w.r.t. hamming loss for $\phi(t) = \max(0, t)$ and $\delta(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \sum_{i=1}^{q} I[y_i \neq \hat{y}_i]$.*

Proof: We consider the multi-label task with $q = 3$ labels:

$$\begin{aligned}
&\boldsymbol{y}_1 = (1, -1, -1), \boldsymbol{y}_2 = (-1, 1, -1), \boldsymbol{y}_3 = (-1, -1, 1), \boldsymbol{y}_4 = (1, 1, 1),\\
&\boldsymbol{y}_5 = (-1, -1, -1), \boldsymbol{y}_6 = (1, 1, -1), \boldsymbol{y}_7 = (1, -1, 1), \boldsymbol{y}_8 = (-1, 1, 1),
\end{aligned}$$

and probability simplex $\boldsymbol{p} = (p_{y_1}, p_{y_2}, p_{y_3}, p_{\boldsymbol{y}_4})$ such that $p_{\boldsymbol{y}_i} = 0$ $(i \geq 4)$, $p_{\boldsymbol{y}_1} + p_{\boldsymbol{y}_2} + p_{\boldsymbol{y}_3} = 1$, $p_{\boldsymbol{y}_1} < p_{\boldsymbol{y}_2} + p_{\boldsymbol{y}_3}$, $p_{\boldsymbol{y}_2} < p_{\boldsymbol{y}_1} + p_{\boldsymbol{y}_3}$, $p_{\boldsymbol{y}_3} < p_{\boldsymbol{y}_1} + p_{\boldsymbol{y}_2}$. By combining Eqns. (17) and (23), we get the set of Bayes predictions

$$\mathcal{A}(\boldsymbol{p}) = \{\boldsymbol{f} \colon f_{\boldsymbol{y}_5} > f_{\boldsymbol{y}_i} \text{ for } i \neq 5\}.$$

We also have

$$\begin{aligned}
W(\boldsymbol{p}, \boldsymbol{f}) \;=\; & p_{\boldsymbol{y}_1} \max_{\boldsymbol{y} \neq \boldsymbol{y}_1} \{\phi(\delta(\boldsymbol{y}, \boldsymbol{y}_1) + f_{\boldsymbol{y}}(\boldsymbol{x}) - f_{\boldsymbol{y}_1}(\boldsymbol{x}))\}\\
& + p_{\boldsymbol{y}_2} \max_{\boldsymbol{y} \neq \boldsymbol{y}_2} \{\phi(\delta(\boldsymbol{y}, \boldsymbol{y}_2) + f_{\boldsymbol{y}}(\boldsymbol{x}) - f_{\boldsymbol{y}_2}(\boldsymbol{x}))\}\\
& + p_{\boldsymbol{y}_3} \max_{\boldsymbol{y} \neq \boldsymbol{y}_3} \{\phi(\delta(\boldsymbol{y}, \boldsymbol{y}_3) + f_{\boldsymbol{y}}(\boldsymbol{x}) - f_{\boldsymbol{y}_3}(\boldsymbol{x}))\}.
\end{aligned}$$

Minimizing $W(\boldsymbol{p}, \boldsymbol{f})$ gives an optimal solution $\boldsymbol{f} = (f_{\boldsymbol{y}_1}, \ldots, f_{\boldsymbol{y}_8})$ such that $f_{\boldsymbol{y}_1} - 3 = f_{\boldsymbol{y}_2} - 3 = f_{\boldsymbol{y}_3} - 3 = f_{\boldsymbol{y}_5} - 2 = f_{\boldsymbol{y}_4} = f_{\boldsymbol{y}_6} = f_{\boldsymbol{y}_7} = f_{\boldsymbol{y}_8}$. It is obvious that $\boldsymbol{f} \notin \mathcal{A}(\boldsymbol{p})$ and we complete the proof. □

Theorem 25 shows the inconsistency of Eqn. (23) under all measurable functions. Hariharan et al. (2010) considered the special function $\boldsymbol{f} = \boldsymbol{w}^\top(\mu(\boldsymbol{x}) \otimes \nu(\boldsymbol{y}))$ based on the prior label-correlation assumption $\nu(\boldsymbol{y}) = P\boldsymbol{y}$ for some invertible matrix P, where $\otimes$ is the Kronecker product, μ and ν are the feature and label space mappings, respectively. An interesting direction is to study the consistency of Eqn. (23) under specific function space and specific label-correlation assumptions as in (Hariharan et al., 2010), and (Ben-David et al., 2012) may shed some light.

5.2. Consistency of Dominating Setting

Though the previous analysis indicates that the one-vs-all and pairwise comparison methods are inconsistent with hamming loss in general cases, it is noteworthy that they may be used successfully in some practical applications, especially for the formulation (23) as in (Hariharan et al., 2010). This is partly because that such methods may work well in special cases, e.g., the dominating setting:

Definition 26. *A multi-label task is said to be in a dominating setting if for every instance $\boldsymbol{x} \in \mathcal{X}$, there exists a $\boldsymbol{y} \in \mathcal{Y}$ such that $P(\boldsymbol{y}|\boldsymbol{x}) > 0.5$.*

Intuitively, the dominating setting implies that, for every instance, there exists a label subset which dominates other label subsets, and it is sufficient to find the dominating label subset. This learning setting exists in real scenarios where the true label set can definitely be predicted accurately. Under such setting, the following theorem shows that the one-vs-all method is consistent w.r.t. hamming loss.

Theorem 27. *If ϕ is a continuous, convex and non-increasing function with $\phi'(0) < 0$, then the one-vs-all method of Eqn.* (19) *is consistent with hamming loss under the dominating setting.*

Proof: Without loss of generality, we consider a probability simplex $\boldsymbol{p} = (p_{\boldsymbol{y}})_{\boldsymbol{y} \in \mathcal{Y}}$ such that $p_{\boldsymbol{y}_1} > 0.5 > p_{\boldsymbol{y}_k}$ for $k \neq 1$. From Eqns. (17) and (18), it is easy to get the set of Bayes predictions

$$\mathcal{A}(\boldsymbol{p}) = \{\boldsymbol{f} \colon f_{\boldsymbol{y}_1} > f_{\boldsymbol{y}} \text{ for } \boldsymbol{y} \neq \boldsymbol{y}_1\}.$$

It suffices to prove $\boldsymbol{f} \in \mathcal{A}(\boldsymbol{p})$ for every $\boldsymbol{f}$ s.t. $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$. We proceed by contradiction. Suppose that there exists $\boldsymbol{f} \notin \mathcal{A}(\boldsymbol{p})$ and $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$, i.e., there is $\hat{\boldsymbol{y}} \in \arg\max_{\boldsymbol{y}\in\mathcal{Y}}\{f_{\boldsymbol{y}}\}$ such that $\boldsymbol{y}_1 \neq \hat{\boldsymbol{y}}$ and $f_{\hat{\boldsymbol{y}}} \geq f_{\boldsymbol{y}_1}$.

If $f_{\hat{\boldsymbol{y}}} > f_{\boldsymbol{y}_1}$, then we construct another $\boldsymbol{f}'$ by

$$f'_{\boldsymbol{y}_1} = f_{\hat{\boldsymbol{y}}}, \quad f'_{\hat{\boldsymbol{y}}} = f_{\boldsymbol{y}_1}, \quad f'_{\boldsymbol{y}} = f_{\boldsymbol{y}} \text{ for } \boldsymbol{y} \neq \boldsymbol{y}_1, \hat{\boldsymbol{y}},$$

and get

$$W(\boldsymbol{p}, \boldsymbol{f}) - W(\boldsymbol{p}, \boldsymbol{f}') = (p_{\boldsymbol{y}_1} - p_{\hat{\boldsymbol{y}}})\Big(\phi(f_{\boldsymbol{y}_1} - f_{\hat{\boldsymbol{y}}}) - \phi(f_{\hat{\boldsymbol{y}}} - \max_{\boldsymbol{y}\neq\hat{\boldsymbol{y}}} f_{\boldsymbol{y}})\Big). \quad (24)$$

From $\hat{\boldsymbol{y}} \in \arg\max_{\boldsymbol{y}\in\mathcal{Y}}\{f_{\boldsymbol{y}}\}$, we have $f_{\boldsymbol{y}_1} - f_{\hat{\boldsymbol{y}}} < 0 \leq f_{\hat{\boldsymbol{y}}} - \max_{\boldsymbol{y}\neq\hat{\boldsymbol{y}}} f_{\boldsymbol{y}}$, and for non-increasing function ϕ with $\phi'(0) < 0$, we further get

$$\phi(f_{\boldsymbol{y}_1} - f_{\hat{\boldsymbol{y}}}) > \phi(f_{\hat{\boldsymbol{y}}} - \max_{\boldsymbol{y}\neq\hat{\boldsymbol{y}}} f_{\boldsymbol{y}}).$$

From Eqn. (24) and $p_{\boldsymbol{y}_1} > p_{\hat{\boldsymbol{y}}}$, it holds that $W(\boldsymbol{p}, \boldsymbol{f}) > W(\boldsymbol{p}, \boldsymbol{f}')$, which is contrary to the assumption $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$.

We now consider the case $f_{\hat{\boldsymbol{y}}} = f_{\boldsymbol{y}_1}$. For very small $\xi_1 > 0$ and from $\phi'(0) < 0$, we have

$$\phi(\xi_1) \approx \phi(0) + \xi_1\phi'(0), \text{ and } \phi(-\xi_1) \approx \phi(0) - \xi_1\phi'(0). \quad (25)$$

We further denote by $\mathcal{B} = \{\boldsymbol{y} \in \mathcal{Y}\colon f_{\boldsymbol{y}} = f_{\boldsymbol{y}_1}\}$. If $\mathcal{B} = \mathcal{Y}$, then we set $\xi_2 = |f_{\boldsymbol{y}_1}|/2$; otherwise, $\xi_2 = (f_{\boldsymbol{y}_1} - \max_{\boldsymbol{y}\notin\mathcal{B}} f_{\boldsymbol{y}})/2$. Now, we set $\xi = \min(\xi_1, \xi_2)$ and construct another $\boldsymbol{f}'$ by

$$f'_{\boldsymbol{y}_1} = f_{\boldsymbol{y}_1}, \quad f'_{\boldsymbol{y}} = f_{\boldsymbol{y}} \text{ for } \boldsymbol{y} \notin \mathcal{B}, \quad \text{and } f'_{\boldsymbol{y}} = f_{\boldsymbol{y}} - \xi \text{ for } \boldsymbol{y} \neq \boldsymbol{y}_1 \text{ and } \boldsymbol{y} \in \mathcal{B}.$$

This follows that, from Eqn. (25),

$$\begin{aligned} W(\boldsymbol{p}, \boldsymbol{f}) - W(\boldsymbol{p}, \boldsymbol{f}') &= p_{\boldsymbol{y}_1}(\phi(0) - \phi(\xi)) - \sum\nolimits_{\substack{\boldsymbol{y}\in\mathcal{B} \\ \boldsymbol{y}\neq\boldsymbol{y}_1}} p_{\boldsymbol{y}}(\phi(-\xi) - \phi(0)) \\ &\approx \xi\phi'(0)\left(\sum\nolimits_{\substack{\boldsymbol{y}\in\mathcal{B} \\ \boldsymbol{y}\neq\boldsymbol{y}_1}} p_{\boldsymbol{y}} - p_{\boldsymbol{y}_1}\right) > 0, \end{aligned}$$

where the last inequality holds from $\phi'(0) < 0$ and $p_{\boldsymbol{y}_1} > 0.5 > p_{\boldsymbol{y}}$ for $\boldsymbol{y} \neq \boldsymbol{y}_1$. Therefore, we have $W(\boldsymbol{p}, \boldsymbol{f}) > W(\boldsymbol{p}, \boldsymbol{f}')$, which is contrary to the assumption $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$. This theorem follows. □

Based on this theorem, we can see clearly that, under the dominating setting, the one-vs-all method of Eqn. (19) is multi-label consistent w.r.t. hamming loss for exponential loss $\phi(t) = \exp(-t)$, logistic loss $\phi(t) = \ln(1 + \exp(-t))$, hinge loss $\phi(t) = \max(0, 1 - t)$, least square hinge loss $\phi(t) = (\max(0, 1 - t))^2$, etc.

Before discussing on the consistency of the pairwise comparison method, we introduce the following lemma whose proof is deferred to Section 6.4.

Lemma 28. *For the pairwise comparison method of Eqn.* (20)*, if ϕ is a non-increasing function with $\phi'(0) < 0$, then for every $\boldsymbol{p} \in \Lambda$ and $\boldsymbol{f}$ such that $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$, we have $f_{\boldsymbol{y}_i} \geq f_{\boldsymbol{y}_j}$ for $p_{\boldsymbol{y}_i} > p_{\boldsymbol{y}_j}$.*

Based on this lemma, we have:

Theorem 29. *If ϕ is a convex, non-increasing and differentiable function with $\phi'(0) < 0$, then the pairwise comparison method of Eqn.* (20) *is consistent w.r.t. hamming loss under the dominating setting.*

Proof: Without loss of generality, we consider a probability simplex $\boldsymbol{p} = (p_{\boldsymbol{y}})_{\boldsymbol{y} \in \mathcal{Y}}$ such that $p_{\boldsymbol{y}_1} > 0.5 > p_{\boldsymbol{y}_k}$ for $k \neq 1$. From Eqns. (17) and (18), it is easy to get the set of Bayes predictions

$$\mathcal{A}(\boldsymbol{p}) = \{\boldsymbol{f} \colon f_{\boldsymbol{y}_1} > f_{\boldsymbol{y}} \text{ for } \boldsymbol{y} \neq \boldsymbol{y}_1\}.$$

It suffices to prove that $\boldsymbol{f} \in \mathcal{A}(\boldsymbol{p})$ for every $\boldsymbol{f}$ s.t. $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$. From Lemma 28, we have

$$f_{\boldsymbol{y}_1} \geq f_{\boldsymbol{y}} \text{ for } \boldsymbol{y} \neq \boldsymbol{y}_1$$

for every $\boldsymbol{f}$ s.t. $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$. It remains to prove $f_{\boldsymbol{y}_1} \neq f_{\boldsymbol{y}}$ for $\boldsymbol{y} \neq \boldsymbol{y}_1$.

Suppose that there exists $\boldsymbol{f}$ such that $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$ yet $f_{\boldsymbol{y}_1} = f_{\boldsymbol{y}_2}$ for $\boldsymbol{y}_1 \neq \boldsymbol{y}_2$. The subgradient conditions $\frac{\partial W(\boldsymbol{p}, \boldsymbol{f})}{\partial f_i} = 0$ $(i = 1, 2)$ give

$$\begin{aligned} p_{\boldsymbol{y}_1} \sum\nolimits_{\boldsymbol{y} \neq \boldsymbol{y}_1} \phi'(f_{\boldsymbol{y}_1} - f_{\boldsymbol{y}}) - \sum\nolimits_{\boldsymbol{y} \neq \boldsymbol{y}_1} p_{\boldsymbol{y}} \phi'(f_{\boldsymbol{y}} - f_{\boldsymbol{y}_1}) = 0, \\ p_{\boldsymbol{y}_2} \sum\nolimits_{\boldsymbol{y} \neq \boldsymbol{y}_2} \phi'(f_{\boldsymbol{y}_2} - f_{\boldsymbol{y}}) - \sum\nolimits_{\boldsymbol{y} \neq \boldsymbol{y}_2} p_{\boldsymbol{y}} \phi'(f_{\boldsymbol{y}} - f_{\boldsymbol{y}_2}) = 0. \end{aligned}$$

This yields

$$2(p_{\boldsymbol{y}_1} - p_{\boldsymbol{y}_2})\phi'(0) + (p_{\boldsymbol{y}_1} - p_{\boldsymbol{y}_2}) \sum\nolimits_{\boldsymbol{y} \neq \boldsymbol{y}_1, \boldsymbol{y}_2} \phi'(f_{\boldsymbol{y}_1} - f_{\boldsymbol{y}}) = 0,$$

which is contrary to the facts $p_{\boldsymbol{y}_1} > p_{\boldsymbol{y}_2}$, $\phi'(0) < 0$ and $\phi'(f_{\boldsymbol{y}_1} - f_{\boldsymbol{y}}) \leq 0$. $\square$

From this theorem, we can see that, under the dominating setting, the pairwise comparison method of Eqn. (20) is consistent w.r.t. hamming loss for exponential loss $\phi(t) = \exp(-t)$, logistic loss $\phi(t) = \ln(1+\exp(-t))$, least square hinge loss $\phi(t) = (\max(0, 1-t))^2$, etc.

It is also noteworthy that the hinge loss $\phi(t) = \max(0, 1-t)$ is not differentiable at $t = 1$, and Theorem 29 cannot be used to study the consistency of pairwise comparison method w.r.t. hinge loss, whereas we have:

Theorem 30. *For hinge loss $\phi(t) = \max(0, 1-t)$, the pairwise comparison method of Eqn.* (20) *is consistent under the dominating setting.*

Proof: Without loss of generality, we consider the probability simplex $\boldsymbol{p} = (p_{\boldsymbol{y}})_{\boldsymbol{y}\in\mathcal{Y}}$ such that $p_{\boldsymbol{y}_1} > 0.5 > p_{\boldsymbol{y}_k}$ for $k \neq 1$. From Eqns. (17) and (18), it is easy to get the set of Bayes predictions

$$\mathcal{A}(\boldsymbol{p}) = \{\boldsymbol{f} \colon f_{\boldsymbol{y}_1} > f_{\boldsymbol{y}} \text{ for } \boldsymbol{y} \neq \boldsymbol{y}_1\}.$$

It suffices to prove $\boldsymbol{f} \in \mathcal{A}(\boldsymbol{p})$ for every $\boldsymbol{f}$ s.t. $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$. From Lemma 28, we have

$$f_{\boldsymbol{y}_1} \geq f_{\boldsymbol{y}} \text{ for } \boldsymbol{y} \neq \boldsymbol{y}_1,$$

for every $\boldsymbol{f}$ s.t. $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$. It remains to prove $f^*_{\boldsymbol{y}_1} \neq f^*_{\boldsymbol{y}}$ for $\boldsymbol{y} \neq \boldsymbol{y}_1$.

Assume that there exists $\boldsymbol{f} = (f_{\boldsymbol{y}_i})_{\boldsymbol{y}_i\in\mathcal{Y}}$ such that $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$ yet $f_{\boldsymbol{y}_1} = f_{\boldsymbol{y}_2}$ for $\boldsymbol{y}_1 \neq \boldsymbol{y}_2$. For small $\xi \in (0, 1)$, we construct another $\boldsymbol{f}'$ by

$$f'_{\boldsymbol{y}_1} = f_{\boldsymbol{y}_1} + \xi, \quad \text{and} \quad f'_{\boldsymbol{y}} = f_{\boldsymbol{y}} \text{ for } \boldsymbol{y} \neq \boldsymbol{y}_1.$$

For hinge loss $\phi(t) = \max(0, 1-t)$, we have

$$\begin{aligned}
&W(\boldsymbol{p}, \boldsymbol{f}) - W(\boldsymbol{p}, \boldsymbol{f}') \\
&= p_{\boldsymbol{y}_1}(\phi(0) - \phi(\xi)) + p_{\boldsymbol{y}_1} \sum\nolimits_{\substack{\boldsymbol{y}\neq\boldsymbol{y}_1 \\ \boldsymbol{y}\neq\boldsymbol{y}_2}} \big(\phi(f_{\boldsymbol{y}_1} - f_{\boldsymbol{y}}) - \phi(f_{\boldsymbol{y}_1} - f_{\boldsymbol{y}} + \xi)\big) \\
&\quad - \sum\nolimits_{\boldsymbol{y}\neq\boldsymbol{y}_1} p_{\boldsymbol{y}}\big(\phi(f_{\boldsymbol{y}} - f_{\boldsymbol{y}_1} - \xi) - \phi(f_{\boldsymbol{y}} - f_{\boldsymbol{y}_1})\big) \\
&\geq p_{\boldsymbol{y}_1}(\phi(0) - \phi(\xi)) - \sum\nolimits_{\boldsymbol{y}\neq\boldsymbol{y}_1} p_{\boldsymbol{y}}(\phi(f_{\boldsymbol{y}} - f_{\boldsymbol{y}_1} - \xi) - \phi(f_{\boldsymbol{y}} - f_{\boldsymbol{y}_1})) \\
&= \xi(2p_{\boldsymbol{y}_1} - 1) > 0.
\end{aligned}$$

This implies $W(\boldsymbol{p}, \boldsymbol{f}) > W(\boldsymbol{p}, \boldsymbol{f}')$, yet it is contrary to $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$. $\square$

We now consider the formulation Eqn. (23) of (Taskar et al., 2004; Hariharan et al., 2010) under the dominating case:

Theorem 31. *Under the dominating setting, the surrogate loss* Ψ *of Eqn.* (23) *is consistent w.r.t. hamming loss for* $\delta(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \sum_{i=1}^{q} I[y_i \neq \hat{y}_i]$ *and* $\phi(t) = \max(0, t)$.

Proof: Without loss of generality, we consider the probability simplex $\boldsymbol{p} = (p_{\boldsymbol{y}})_{\boldsymbol{y}\in\mathcal{Y}}$ such that $p_{\boldsymbol{y}_1} > 0.5 > p_{\boldsymbol{y}_k}$ for $k \neq 1$. From Eqns. (17) and (18), it is easy to get the set of Bayes predictions

$$\mathcal{A}(\boldsymbol{p}) = \{\boldsymbol{f} \colon f_{\boldsymbol{y}_1} > f_{\boldsymbol{y}} \text{ for } \boldsymbol{y} \neq \boldsymbol{y}_1\}.$$

It suffices to prove that $\boldsymbol{f} \in \mathcal{A}(\boldsymbol{p})$ for every $\boldsymbol{f}$ s.t. $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$.

Suppose that there exists $\boldsymbol{f} \notin \mathcal{A}(\boldsymbol{p})$ and $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$, i.e., it holds that $f_{\boldsymbol{y}_1} \leq f_{\hat{\boldsymbol{y}}}$ for some $\hat{\boldsymbol{y}} \in \mathcal{Y}$ and $\boldsymbol{y}_1 \neq \hat{\boldsymbol{y}}$. Then, we can construct another $\boldsymbol{f}'$ by

$$\boldsymbol{f}'_{\boldsymbol{y}_1} = \boldsymbol{f}_{\boldsymbol{y}_1} + 1, \quad \boldsymbol{f}'_{\boldsymbol{y}} = \boldsymbol{f}_{\boldsymbol{y}} \text{ for } \boldsymbol{y} \neq \boldsymbol{y}_1, \boldsymbol{y} \in \mathcal{Y}.$$

Since $\boldsymbol{y}_1 \neq \hat{\boldsymbol{y}}$ and $f_{\boldsymbol{y}_1} \leq f_{\hat{\boldsymbol{y}}}$, we have $\delta(\boldsymbol{y}_1, \hat{\boldsymbol{y}}) \geq 1$ and

$$p_{\boldsymbol{y}_1} \max_{\boldsymbol{y}\neq\boldsymbol{y}_1}\{\phi(\delta(\boldsymbol{y}, \boldsymbol{y}_1) + f_{\boldsymbol{y}} - f_{\boldsymbol{y}_1})\} - p_{\boldsymbol{y}_1} \max_{\boldsymbol{y}\neq\boldsymbol{y}_1}\{\phi(\delta(\boldsymbol{y}, \boldsymbol{y}_1) + f_{\boldsymbol{y}} - f_{\boldsymbol{y}_1} - 1)\} = p_{\boldsymbol{y}_1}.$$

Further, it holds that

$$\begin{aligned}
&p_{\boldsymbol{y}_i} \max_{\boldsymbol{y}\neq\boldsymbol{y}_i}\{\phi(\delta(\boldsymbol{y}, \boldsymbol{y}_i) + f_{\boldsymbol{y}} - f_{\boldsymbol{y}_i})\} - p_{\boldsymbol{y}_i} \max_{\boldsymbol{y}\neq\boldsymbol{y}_i}\{\phi(\delta(\boldsymbol{y}, \boldsymbol{y}_i) + f'_{\boldsymbol{y}} - f'_{\boldsymbol{y}_i})\} \\
&= p_{\boldsymbol{y}_i} \max\{\phi(\delta(\boldsymbol{y}_1, \boldsymbol{y}_i) + f_{\boldsymbol{y}_1} - f_{\boldsymbol{y}_i}), \max_{\boldsymbol{y}\neq\boldsymbol{y}_i,\boldsymbol{y}_1} \phi(\delta(\boldsymbol{y}, \boldsymbol{y}_i) + f_{\boldsymbol{y}} - f_{\boldsymbol{y}_i})\} \\
&\quad - p_{\boldsymbol{y}_i} \max\{\phi(\delta(\boldsymbol{y}_1, \boldsymbol{y}_i) + f_{\boldsymbol{y}_1} + 1 - f_{\boldsymbol{y}_i}), \max_{\boldsymbol{y}\neq\boldsymbol{y}_i,\boldsymbol{y}_1} \phi(\delta(\boldsymbol{y}, \boldsymbol{y}_i) + f_{\boldsymbol{y}} - f_{\boldsymbol{y}_i})\} \\
&\geq -p_{\boldsymbol{y}_i}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&W(\boldsymbol{p}, \boldsymbol{f}) - W(\boldsymbol{p}, \boldsymbol{f}') \\
&= p_{\boldsymbol{y}_1} \max_{\boldsymbol{y}\neq\boldsymbol{y}_1}\{\phi(\delta(\boldsymbol{y}, \boldsymbol{y}_1) + f_{\boldsymbol{y}} - f_{\boldsymbol{y}_1})\} - p_{\boldsymbol{y}_1} \max_{\boldsymbol{y}\neq\boldsymbol{y}_1}\{\phi(\delta(\boldsymbol{y}, \boldsymbol{y}_1) + f_{\boldsymbol{y}} - f_{\boldsymbol{y}_1} - 1)\} \\
&\quad + \sum_{\boldsymbol{y}_i\neq\boldsymbol{y}_1} p_{\boldsymbol{y}_i} \max_{\boldsymbol{y}\neq\boldsymbol{y}_i}\{\phi(\delta(\boldsymbol{y}, \boldsymbol{y}_i) + f_{\boldsymbol{y}} - f_{\boldsymbol{y}_i})\} - p_{\boldsymbol{y}_i} \max_{\boldsymbol{y}\neq\boldsymbol{y}_i}\{\phi(\delta(\boldsymbol{y}, \boldsymbol{y}_i) + f'_{\boldsymbol{y}} - f'_{\boldsymbol{y}_i})\} \\
&\geq \boldsymbol{p}_{\boldsymbol{y}_1} - \sum\nolimits_{\boldsymbol{y}\neq\boldsymbol{y}_1} \boldsymbol{p}_{\boldsymbol{y}} > 0,
\end{aligned}$$

which is contrary to $W(\boldsymbol{p}, \boldsymbol{f}) = W^*(\boldsymbol{p})$. This completes the proof. □

5.3. Solving Multi-Label Learning by Binary Classifications

It is possible to decompose a multi-label learning task into q independent binary classification tasks (Boutell et al., 2004) when hamming loss is concerned, especially when there are just a few labels. Now the goal is to learn q functions, $\boldsymbol{f} = (f_1, f_2, \ldots, f_q)$, and the prediction function is given by

$$F(\boldsymbol{f}(\boldsymbol{x})) = (\mathrm{sgn}[f_1(\boldsymbol{x})], \mathrm{sgn}[f_2(\boldsymbol{x})], \ldots, \mathrm{sgn}[f_q(\boldsymbol{x})]).$$

A common choice for the surrogate loss is

$$\Psi(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = \sum\nolimits_{i=1}^{q} \phi(y_i f_i(\boldsymbol{x})), \tag{26}$$

where ϕ is a convex function. For example, it was chosen as hinge loss $\phi(t) = (1-t)_+$ in (Elisseeff and Weston, 2002) and exponential loss $\phi(t) = \exp(-t)$ in (Schapire and Singer, 2000), respectively. We have the conditional surrogate loss

$$\begin{aligned} W(\boldsymbol{p}, \boldsymbol{f}) &= \sum\nolimits_{\boldsymbol{y}\in\mathcal{Y}} p_{\boldsymbol{y}} \Psi(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = \sum\nolimits_{i=1}^{q} \sum\nolimits_{\boldsymbol{y}\in\mathcal{Y}} p_{\boldsymbol{y}} \phi(y_i f_i(\boldsymbol{x})) \\ &= \sum\nolimits_{i=1}^{q} p_i^+ \phi(f_i(\boldsymbol{x})) + (1 - p_i^+)\phi(-f_i(\boldsymbol{x})), \end{aligned}$$

where $p_i^+ = \sum_{\boldsymbol{y}:\, y_i=+1} p_{\boldsymbol{y}}$ and $1 - p_i^+ = \sum_{\boldsymbol{y}:\, y_i=-1} p_{\boldsymbol{y}}$. For simplicity, we denote by

$$W_i(p_i^+, f_i) = p_i^+ \phi(f_i) + (1 - p_i^+)\phi(-f_i).$$

This yields that minimizing $W(\boldsymbol{p}, \boldsymbol{f})$ is equivalent to minimizing $W_i(p_i^+, f_i)$ for every $1 \le i \le q$, that is,

$$W^*(\boldsymbol{p}) = \inf_{\boldsymbol{f}} W(\boldsymbol{p}, \boldsymbol{f}) = \sum\nolimits_{i=1}^{q} \inf_{f_i} W_i(p_i^+, f_i).$$

The consistency for binary classification has been well-studied (Zhang, 2004b; Bartlett et al., 2006), and based on the work of Bartlett et al. (2006), we can easily get:

Theorem 32. *If ϕ is a convex function with $\phi'(0) < 0$, then the surrogate loss Ψ given by Eqn.* (26) *is consistent w.r.t. hamming loss.*

It is evident from this theorem that the surrogate loss Ψ given by Eqn. (26) is consistent w.r.t. hamming loss if ϕ is any of the following:

- Exponential loss: $\phi(t) = \exp(-t)$;
- Hinge loss: $\phi(t) = \max(0, 1-t)$;
- Least squares loss: $\phi(t) = (1-t)^2$;
- Logistic loss: $\phi(t) = \ln(1+\exp(-t))$;
- Least squares hinge loss: $\phi(t) = (\max(0, 1-t))^2$.

6. Proofs of Lemmas

In this section, we provide the detailed proofs of lemmas.

6.1. Proof of Lemma 5

From the Heine definition of continuity, it is sufficient to show

$$W^*(\boldsymbol{p}^{(n)}) \to W^*(\boldsymbol{p})$$

for any sequence $\boldsymbol{p}^{(n)} \to \boldsymbol{p}$.

Let B_r be a closed ball with radius r in $\mathbb{R}^K$. Because $|\mathcal{Y}|$ is finite, we have

$$\sum\nolimits_{\boldsymbol{y}\in\mathcal{Y}} p_{\boldsymbol{y}}^{(n)} \Psi(\boldsymbol{f}, \boldsymbol{y}) \to \sum\nolimits_{\boldsymbol{y}\in\mathcal{Y}} p_{\boldsymbol{y}} \Psi(\boldsymbol{f}, \boldsymbol{y})$$

uniformly for every $\boldsymbol{f} \in B_r$ and every sequence $\boldsymbol{p}^{(n)} \to \boldsymbol{p}$, leading to

$$\inf_{\boldsymbol{f}\in B_r} \sum\nolimits_{\boldsymbol{y}\in\mathcal{Y}} p_{\boldsymbol{y}}^{(n)} \Psi(\boldsymbol{f}, \boldsymbol{y}) \to \inf_{\boldsymbol{f}\in B_r} \sum\nolimits_{\boldsymbol{y}\in\mathcal{Y}} p_{\boldsymbol{y}} \Psi(\boldsymbol{f}, \boldsymbol{y}).$$

From

$$W^*(\boldsymbol{p}^{(n)}) \le \inf_{\boldsymbol{f}\in B_r} \sum\nolimits_{\boldsymbol{y}\in\mathcal{Y}} p_{\boldsymbol{y}}^{(n)} \Psi(\boldsymbol{f}, \boldsymbol{y}),$$

and letting $r \to \infty$, we have

$$\limsup_{n\to\infty} W^*(\boldsymbol{p}^{(n)}) \le W^*(\boldsymbol{p}). \tag{27}$$

Denote by $\mathcal{Y}' = \{\boldsymbol{y} | p_{\boldsymbol{y}} > 0 \text{ for } \boldsymbol{y} \in \mathcal{Y}\}$ and assume $\Psi(\cdot,\cdot) \ge C$ for some constant C (since Ψ is below-bounded). We have

$$W^*(\boldsymbol{p}^{(n)}) \ge \inf_{\boldsymbol{f}} \sum\nolimits_{\boldsymbol{y}\in\mathcal{Y}'} p_{\boldsymbol{y}}^{(n)} \Psi(\boldsymbol{f}, \boldsymbol{y}) + C \sum\nolimits_{\boldsymbol{y}\in\mathcal{Y}/\mathcal{Y}'} p_{\boldsymbol{y}}^{(n)},$$

which yields

$$\liminf_{n\to\infty} W^*(\boldsymbol{p}^{(n)}) \ge \liminf_{n\to\infty} \left(\inf_{\boldsymbol{f}} \sum_{\boldsymbol{y}\in\mathcal{Y}'} p_{\boldsymbol{y}}^{(n)} \Psi(\boldsymbol{f}, \boldsymbol{y}) + C \sum_{\boldsymbol{y}\in\mathcal{Y}/\mathcal{Y}'} p_{\boldsymbol{y}}^{(n)} \right) = W^*(\boldsymbol{p}),$$

which completes the proof by combining Eqn. (27). □

6.2. Proof of Lemma 6

We proceed by contradiction. Suppose Ψ is multi-label consistent and there exists $\epsilon > 0$ and a sequence $(\boldsymbol{p}^{(n)}, \boldsymbol{f}^{(n)})$ such that

$$l(\boldsymbol{p}^{(n)}, \boldsymbol{f}^{(n)}) - \inf_{\boldsymbol{f}'} l(\boldsymbol{p}^{(n)}, \boldsymbol{f}') \geq \epsilon \text{ and } W(\boldsymbol{p}^{(n)}, \boldsymbol{f}^{(n)}) \to W^*(\boldsymbol{p}^{(n)}).$$

From the compactness of Λ, there exists a convergence sequence n_k such that $\boldsymbol{p}^{(n_k)} \to \boldsymbol{p}$ for some $\boldsymbol{p} \in \Lambda$. From Lemma 5, we have

$$W(\boldsymbol{p}^{(n_k)}, \boldsymbol{f}^{(n_k)}) \to W^*(\boldsymbol{p}).$$

Similar to the proof of Lemma 5, we set $\mathcal{Y}' = \{\boldsymbol{y} | p_{\boldsymbol{y}} > 0 \text{ for } \boldsymbol{y} \in \mathcal{Y}\}$, and get

$$\begin{aligned}
&\limsup_{n_k} W(\boldsymbol{p}, \boldsymbol{f}^{(n_k)}) \\
&= \limsup_{n_k} \left(C \sum\nolimits_{\boldsymbol{y} \in \mathcal{Y}/\mathcal{Y}'} p_{\boldsymbol{y}}^{(n_k)} + \inf_{\boldsymbol{f}} \sum\nolimits_{\boldsymbol{y} \in \mathcal{Y}'} p_{\boldsymbol{y}}^{(n_k)} \Psi(\boldsymbol{f}^{(n_k)}, \boldsymbol{y}) \right) \\
&\leq \lim_{n_k} W(\boldsymbol{p}^{(n_k)}, \boldsymbol{f}^{(n_k)}) = W^*(\boldsymbol{p}).
\end{aligned}$$

This gives $W(\boldsymbol{p}, \boldsymbol{f}^{(n_k)}) \to W^*(\boldsymbol{p})$ from the definition of $W^*(\boldsymbol{p})$. Since Ψ is multi-label consistent, there exists a sequence $\boldsymbol{f}^{(n_{k_i})}$ such that

$$l(\boldsymbol{p}, \boldsymbol{f}^{(n_{k_i})}) \to \inf_{\boldsymbol{f}'} l(\boldsymbol{p}, \boldsymbol{f}'),$$

which contradicts the assumption $l(\boldsymbol{p}^{(n)}, \boldsymbol{f}^{(n)}) - \inf_{\boldsymbol{f}'} l(\boldsymbol{p}^{(n)}, \boldsymbol{f}') \geq \epsilon$, and the lemma follows. □

6.3. Proof of Lemma 14

From Eqn. (9), we set

$$1 < \frac{P_1}{P_2} < \frac{\phi'(b-a)(\phi'(a) + \phi'(b)) + \phi'(b)\phi'(-a)}{\phi'(a-b)(\phi'(a) + \phi'(b)) + \phi'(b)\phi'(a)}. \tag{28}$$

For $a < b$, we have $\phi'(a-b) \leq \phi'(b-a) < 0$, yielding

$$\frac{P_1}{P_2} > 1 \geq \frac{\phi'(b-a)}{\phi'(a-b)},$$

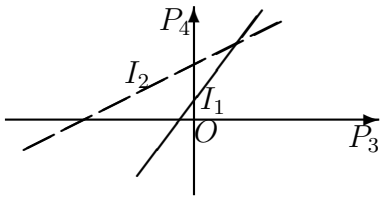

Figure 1: Lines I_1 (solid) and I_2 (dash) corresponding to Eqns. (11) and (12), respectively.

which gives $P_1\phi'(a-b) - P_2\phi'(b-a) < 0$. Thus, Eqn. (11) corresponds to the Line I_1 in Figure 1. From Eqn. (10), we further obtain

$$0 < \frac{\phi'(-a) + \phi'(-b)}{\phi'(a) + \phi'(b)} < \frac{\phi'(-b)}{\phi'(b)}.$$

To guarantee $P_3 > 0$ and $P_4 > 0$ satisfying Eqns. (11) and (12), as shown in Figure 1, we need:

$$\frac{P_1\phi'(a-b) - P_2\phi'(b-a)}{\phi'(b)} < \frac{-P_1\phi'(a) + P_2\phi'(-a)}{\phi'(a) + \phi'(b)}.$$

The above holds obviously from Eqn. (28). Thus, we complete the proof. □

6.4. Proof of Lemma 28

We proceed by contradiction. Suppose there exists a probability simplex $\boldsymbol{p} \in \Lambda$ and $\boldsymbol{f}$ such that $f_{\boldsymbol{y}_1} < f_{\boldsymbol{y}_2}$, $p_{\boldsymbol{y}_1} > p_{\boldsymbol{y}_2}$ and $W^*(\boldsymbol{p}) = W(\boldsymbol{p}, \boldsymbol{f})$. We can construct another $\boldsymbol{f}'$ by

$$f'_{\boldsymbol{y}_1} = f_{\boldsymbol{y}_2}, \quad f'_{\boldsymbol{y}_2} = f_{\boldsymbol{y}_1}, \quad \text{and} \quad f'_{\boldsymbol{y}_k} = f_{\boldsymbol{y}_k} \text{ for } k \neq 1, 2.$$

This follows

$$\begin{aligned} W(\boldsymbol{p}, \boldsymbol{f}) - W(\boldsymbol{p}, \boldsymbol{f}') = (p_{\boldsymbol{y}_1} - p_{\boldsymbol{y}_2})\big(\phi(f_{\boldsymbol{y}_1} - f_{\boldsymbol{y}_2}) - \phi(f_{\boldsymbol{y}_2} - f_{\boldsymbol{y}_1})\big) \\ + (p_{\boldsymbol{y}_1} - p_{\boldsymbol{y}_2}) \sum\nolimits_{\substack{\boldsymbol{y}_k \neq \boldsymbol{y}_1 \\ \boldsymbol{y}_k \neq \boldsymbol{y}_2}} \left(\phi(f_{\boldsymbol{y}_1} - f_{\boldsymbol{y}_k}) - \phi(f_{\boldsymbol{y}_2} - f_{\boldsymbol{y}_k})\right). \quad (29) \end{aligned}$$

For non-increasing function ϕ, we have

$$\phi(f_{\boldsymbol{y}_1} - f_{\boldsymbol{y}_i}) \geq \phi(f_{\boldsymbol{y}_2} - f_{\boldsymbol{y}_i}),$$

and from $\phi'(0) < 0$, we further get

$$\phi(f_{\boldsymbol{y}_1} - f_{\boldsymbol{y}_2}) > \phi(f_{\boldsymbol{y}_2} - f_{\boldsymbol{y}_1}).$$

From Eqn. (29), we have $W(\boldsymbol{p}, \boldsymbol{f}) > W(\boldsymbol{p}, \boldsymbol{f}')$, which is contrary to the assumption $W^*(\boldsymbol{p}) = W(\boldsymbol{p}, \boldsymbol{f})$. The lemma holds as desired. □

Table 1: Summary of consistency result, where $\surd/\times$ indicates consistency/inconsistency.

	ranking loss	partial ranking loss		hamming loss	
loss function ϕ	Eqn. (4)	Eqn. (4)	Eqn. (15)	Eqn. (19)	Eqn. (20)
logistic	$\times$	$\times$	$\surd$	$\times$	$\times$
hinge	$\times$	$\times$	$\times$	$\times$	$\times$
exponential	$\times$	$\times$	$\surd$	$\times$	$\times$
least square	$\times$	$\times$	$\surd$	$\times$	$\times$
least square hinge	$\times$	$\times$	$\surd$	$\times$	$\times$
regularized linear	$\times$	$\surd$	$\surd$	$\times$	$\times$

7. Conclusion and Future Work

During the past decade, multi-label learning has attracted significant attention in the machine learning community. Most studies have been devoted to the algorithm designs and diverse applications. Theoretical analysis, however, remains almost untouched for multi-label learning. This paper extends our preliminary work (Gao and Zhou, 2011), which tries to study the consistency of multi-label learning based on surrogate losses. We present a necessary and sufficient condition for multi-label consistency, and study two well-known loss functions, i.e., ranking loss and hamming loss. Our main results are summarized in Table 1.

The ranking loss is one of the most popularly used multi-label criterion and many approaches (Schapire and Singer, 2000; Elisseeff and Weston, 2002; Dekel et al., 2004; Zhang and Zhou, 2006) try to optimize it under the formulation of Eqn. (4). Our analysis, however, discloses that none of convex surrogate loss is consistent w.r.t. ranking loss; therefore, ranking loss might not be a good criterion for multi-label learning. The partial ranking loss is more reasonable than ranking loss because, on one hand it keeps the nature of ranking loss (by ranking relevant labels higher than irrelevant ones), and on the other hand, it enables many, though not all, convex surrogate losses to be consistent.

The hamming loss is also one of the most popularly used multi-label criterion, based on which it is natural to develop some learning approaches from multi-class learning, e.g., the one-vs-all method and the pairwise comparison

method. Our results disclose that these approaches are inconsistent w.r.t. hamming loss for general cases yet consistent under the dominating setting, and similar results also hold for some multi-label approaches (Taskar et al., 2004; Hariharan et al., 2010) that are variations of the one-vs-all method. In addition, we discuss on the consistency of approaches that address multi-label learning by decomposing the task into a set of binary classification problems.

An important future work is to investigate the convergence rate of consistent surrogate loss functions as in (Bartlett et al., 2006). How to incorporate label correlation into the study of multi-label consistency also remains an open problem. In addition, our work may motivate the consistency researches on other multi-label criteria such as one-error, F_1, etc. It is also interesting to develop some new multi-label learning approaches by minimizing the partial ranking loss. Note that we do not consider how to decide the number of relevant labels for ranking loss and partial ranking loss in this work, whereas in practice this is quite challenging. For hamming loss, it is extremely important to explore new surrogate losses and find new prediction rules for developing consistent approaches, because none of existing algorithms is consistent in general cases. It is also interesting to make a comprehensive empirical study on various approaches and formulations for multi-label learning.

Acknowledgement

The authors want to thank the anonymous reviewers and associate editor for helpful comments and suggestions.

References

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

Ben-David, S., Loker, D., Srebro, N., Sridharan, K., 2012. Minimizing the misclassification error rate using a surrogate convex loss. In: Proceedings of the 29th International Conference on Machine Learning. Edinburgh, Scotland.

Boutell, M. R., Luo, J., Shen, X., Brown, C., 2004. Learning multi-label scene classification. Pattern Recognition 37 (9), 1757–1771.

Breiman, L., 2004. Some infinity theory for predictor ensembles. Annals of Statistics 32 (1), 1–11.

Bühlmann, P., Yu, B., 2003. Boosting with $l-2$-loss: Regression and classification. Journal of the American Statistical Association 98 (462), 324–339.

Carneiro, G., Chan, A., Moreno, P., Vasconcelos, N., 2007. Supervised learning of semantic classes for image annotation and retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (3), 394–410.

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

Crammer, K., Singer, Y., 2001. On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research 2, 265–292.

Dekel, O., Manning, C., Singer, Y., 2004. Log-linear models for label ranking. In: Thrun, S., Saul, L. K., Schölkopf, B. (Eds.), Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA.

Dembczyński, K., Cheng, W. W., Hüllermeier, E., 2010. Bayes optimal multilabel classification via probabilistic classifier chains. In: Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel, pp. 279–286.

Dembczyński, K., Kotlowski, W., Hüllermeier, E., 2012a. Consistent multi-label ranking through univariate loss minimization. In: Proceedings of the 29th International Conference on Machine Learning. New York, NY, pp. 1319–1326.

Dembczyński, K., Waegeman, W., Cheng, W., Hüllermeier, E., 2012b. On label dependence and loss minimization in multi-label classification. Machine Learning 88 (1-2), 5–45.

Duchi, J. C., Mackey, L. W., Jordan, M. I., 2010. On the consistency of ranking algorithms. In: Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel, pp. 327–334.

Elisseeff, A., Weston, J., 2002. A kernel method for multi-labelled classification. In: Dietterich, T. G., Becker, S., Ghahramani, Z. (Eds.), Advances in Neural Information Processing Systems 14. MIT Press, Cambridge, MA, pp. 681–687.

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

Gao, W., Zhou, Z.-H., 2011. On the consistency of multi-label learning. In: Proceedings of the 24th Annual Conference on Learning Theory. Budapest, Hungary, pp. 341–358.

Ghamrawi, N., McCallum, A., 2005. Collective multi-label classification. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management. Bremen, Germany, pp. 195–200.

Godbole, S., Sarawagi, S., 2004. Discriminative methods for multi-labeled classification. In: Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Sydney, Austrialia, pp. 22–30.

Hariharan, B., Zelnik-Manor, L., Vishwanathan, S., Varma, M., 2010. Large scale max-margin multi-label classification with priors. In: Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel, pp. 423–430.

Hsu, D., Kakade, S. M., Langford, J., Zhang, T., 2009. Multi-label prediction via compressed sensing. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 22. MIT Press, Cambridge, MA, pp. 772–780.

Hüllermeier, E., Fürnkranz, J., Cheng, W., Brinker, K., 2008. Label ranking by learning pairwise preferences. Artificial Intelligence 172 (16-17), 1897–1916.

Lin, Y., 2002. Support vector machines and the bayes rule in classification. Data Mining and Knowledge Discovery 6 (3), 259–275.

McAllester, D., Keshet, J., 2011. Generalization bounds and consistency for latent structural probit and ramp loss. In: Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., Weinberger, K. Q. (Eds.), Advances in Neural Information Processing Systems 23. MIT Press, pp. 2205–2212.

Petterson, J., Caetano, T., 2010. Reverse multi-label learning. In: Lafferty, J., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 23. MIT Press, Cambridge, MA, pp. 1912–1920.

Qi, G.-J., Hua, X.-S., Rui, Y., Tang, J., Mei, T., Zhang, H.-J., 2007. Correlative multi-label video annotation. In: Proceedings of the 15th ACM International Conference on Multimedia. Augsburg, Germany, pp. 17–26.

Rockafellar, R. T., 1997. Convex Analysis. Princeton University Press, Princeton, NJ.

Schapire, R. E., Singer, Y., 2000. BoosTexter: A boosting-based system for text categorization. Machine Learning 39 (2), 135–168.

Steinwart, I., 2005. Consistency of support vector machines and other regularized kernel classifiers. IEEE Transactions on Information Theory 51 (1), 128–142.

Taskar, B., Guestrin, C., Koller, D., 2004. Max-margin markov networks. In: Thrun, S., Saul, L., Schölkopf, B. (Eds.), Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA.

Tewari, A., Bartlett, P. L., 2007. On the consistency of multiclass classification methods. Journal of Machine Learning Research 8, 1007–1025.

Weston, J., Watkins, C., 1998. Multi-class support vector machines. Tech. Rep. CSD-TR-98-04, Royal Holloway.

Xia, F., Liu, T. Y., Li, H., 2009. Top-k consistency of learning to rank methods. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 22. MIT Press, Cambridge, MA, pp. 2098–2106.

Xia, F., Liu, T. Y., Wang, J., Zhang, W., Li, H., 2008. Listwise approach to learning to rank: Theory and algorithm. In: Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland, pp. 1192–1199.

Zhang, M.-L., Zhou, Z.-H., 2006. Multi-label neural network with applications to functional genomics and text categorization. IEEE Transactions on Knowledge and Data Engineering 18 (10), 1338–1351.

Zhang, M.-L., Zhou, Z.-H., 2007. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition 40 (7), 2038–2048.

Zhang, T., 2004a. Statistical analysis of some multi-category large margin classification methods. Journal of Machine Learning Research 5, 1225–1251.

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

Zhou, Z.-H., Zhang, M.-L., 2007. Multi-instance multi-label learning with application to scene classification. In: Schölkopf, B., Platt, J., Hofmann, T. (Eds.), Advances in Neural Information Processing Systems 19. MIT Press, Cambridge, MA, pp. 1609–1616.

Zhou, Z.-H., Zhang, M.-L., Huang, S.-J., Li, Y.-F., 2012. Multi-instance multi-label learning. Artificial Intelligence 176 (1), 2291–2320.