On the Consistency of Multi-Label Learning

Wei Gao GAOW@LAMDA.NJU.EDU.CN
Zhi-Hua Zhou ZHOUZH@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China

Editor: Sham Kakade, Ulrike von Luxburg

Abstract

Multi-label learning has attracted much attention during the past few years. Many multi-label learning approaches have been developed, mostly working with surrogate loss functions since multi-label loss functions are usually difficult to optimize directly owing to non-convexity and discontinuity. Though these approaches are effective, to the best of our knowledge, there is no theoretical result on the convergence of risk of the learned functions to the Bayes risk. In this paper, focusing on two well-known multi-label loss functions, i.e., *ranking loss* and *hamming loss*, we prove a necessary and sufficient condition for the consistency of multi-label learning based on surrogate loss functions. Our results disclose that, surprisingly, none convex surrogate loss is consistent with the ranking loss. Inspired by the finding, we introduce the *partial ranking loss*, with which some surrogate functions are consistent. For hamming loss, we show that some recent multi-label learning approaches are inconsistent even for deterministic multi-label classification, and give a surrogate loss function which is consistent for the deterministic case. Finally, we discuss on the consistency of learning approaches which address multi-label learning by decomposing into a set of binary classification problems.

Keywords: Consistency, multi-label learning, surrogate loss, ranking loss, hamming loss

1. Introduction

In traditional supervised learning, each instance is associated with a single label. In real-world applications, however, one object is usually relevant to multiple labels simultaneously. For example, a document about national education service may be categorized into several predefined topics, such as government and education; an image containing forests may be annotated with trees and mountains. For learning with such objects, *multi-label learning* has attracted much attention during the past few years and many effective approaches have been developed (Schapire and Singer, 2000; Elisseeff and Weston, 2002; Zhou and Zhang, 2007; Zhang and Zhou, 2007; Hsu et al., 2009; Dembczyński et al., 2010; Petterson and Caetano, 2010).

The *consistency* (also called Bayes consistency) of learning algorithms concerns if the expected risk of a learned function converges to the Bayes risk as the training sample size increases (Lin, 2002; Zhang, 2004b; Steinwart, 2005; Bartlett et al., 2006; Tewari and Bartlett, 2007; Duchi et al., 2010). Nowadays, it is well-accepted that a good learner should at least be consistent with large samples. It is noteworthy that, though many efforts have been devoted to multi-label learning, few theoretical aspects were explored; in particular, to the best of our knowledge, the consistency of multi-label learning remains untouched although it is a very important theoretical issue.

In this paper, focusing on two well-known multi-label loss functions, i.e., *ranking loss* and *hamming loss*, we present a theoretical study on the consistency of multi-label learning. We prove a necessary and sufficient condition for the consistency of multi-label learning based on surrogate loss functions. Our analysis discloses that, surprisingly, any convex surrogate loss is inconsistent with the ranking loss. Based on this finding, we introduce the *partial ranking loss*, which is consistent with some surrogate loss functions; we also show that many current multi-label learning approaches are even not consistent with the partial ranking loss. As for hamming loss, our analysis shows that some recent multi-label learning approaches are inconsistent even for deterministic multi-label classification, and we give a surrogate loss which is consistent with hamming loss for the deterministic case. Finally, we discuss on the consistency of approaches which address multi-label learning by transforming the problem into a set of binary classification problems.

The rest of this paper is organized as follows. Section 2 briefly introduces the research background. Section 3 presents our necessary and sufficient condition for the consistency of multi-label learning. Sections 4 and 5 study the consistency of multi-label learning approaches with regard to the ranking loss and hamming loss, respectively. Section 6 gives the detailed proofs.

2. Background

Let $\mathcal{X}$ be an instance space and $\mathcal{L} = \{1, 2, \dots, Q\}$ denotes a finite set of possible labels. An instance $X \in \mathcal{X}$ is associated with a subset of labels $Y \subset \mathcal{L}$ which is called *relevant labels*, while the complement $\mathcal{L} \setminus Y$ is called *irrelevant labels*. For convenience of discussion, we represent the labels as a binary vector $Y = (y_1, y_2, \dots, y_Q)$, where $y_i = +1$ if label i is relevant to X and -1 otherwise, and denote by $\mathcal{Y} = \{+1, -1\}^Q$ the set of all possible labels. Let $\mathcal{D}$ denote an unknown underlying probability distribution over $\mathcal{X} \times \mathcal{Y}$. For integer $m > 0$, we denote by $[m] = \{1, 2, \dots, m\}$, and for real r, $\lfloor r \rfloor$ denotes the greatest integer which is no more than r.

The formal description of multi-label learning in the probabilistic setting is given as follows: Given a training sample $S = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_m, Y_m)\}$ drawn i.i.d. according to the distribution $\mathcal{D}$, 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 in practice, one instead learns a real-valued vector function

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

where $K = Q$ or $K = 2^Q$ are common choices. Based on this vector function $\mathbf{f}$, 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 $\mathbf{f} = (f_1, f_2, \dots, f_Q)$ such that $f_i(X) > f_j(X)$ if $y_i = +1$ and $y_j = -1$ for an instance-label pair (X, Y), and a function F should be learned to determine the number of relevant labels.

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

$$R(\mathbf{f}) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[L\left(\mathbf{f}\left(X\right), Y\right)\right]. \tag{1}$$

Notice that $\mathbf{f}$ could be a prediction function or a vector of real-valued functions according to different losses. Further denote by $R^* = \inf_{\mathbf{f}} R(\mathbf{f})$ the minimal risk (e.g., the Bayes risk) over all measurable functions. In this paper, we mainly focus on loss functions which are *below-bounded* and *interval*, defined as follows:

Definition 1 *A loss function L is said to be below-bounded if $L(\cdot,\cdot) \geq B$ holds for some constant B; L is said to be interval if it holds, for some constant $\gamma > 0$, that either*

$$L(\mathbf{f}(X), Y) = L(\mathbf{f}(X'), Y') \quad \textit{or} \quad |L(\mathbf{f}(X), Y) - L(\mathbf{f}(X'), Y')| \geq \gamma,$$

for every $X, X' \in \mathcal{X}$ and $Y, Y' \in \mathcal{Y}$.

There are many multi-label loss functions (also called *evaluation criteria*), 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. In this paper, we focus on two well-known losses, i.e., ranking loss and hamming loss, and leave the discussion on other losses to future work.

Notice that all the above multi-label losses are non-convex and discontinuous. It is difficult, or even impossible, to optimize these losses directly. A feasible method in practice is to consider instead a surrogate loss function which can be optimized by efficient algorithms. Actually, most existing multi-label learning approaches, such as the boosting algorithm AdaBoost.MH (Schapire and Singer, 2000), neural network algorithm BP-MIL (Zhang and Zhou, 2006), SVM-style algorithms (Elisseeff and Weston, 2002; Taskar et al., 2004; Hariharan et al., 2010), etc., in essence, try to optimize some surrogate losses such as the exponential loss and hinge loss.

There are many definitions of consistency, 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., and the consistency of learning algorithms based on optimizing a surrogate loss function has been well-studied for binary classification (Zhang, 2004b; Steinwart, 2005; Bartlett et al., 2006), multi-class classification (Zhang, 2004a; Tewari and Bartlett, 2007), learning to rank (Cossock and Zhang, 2008; Xia et al., 2008; Duchi et al., 2010), etc. The consistency of multi-label learning, however, remains untouched, and to the best of our knowledge, this paper presents the first theoretical analysis on the consistency of multi-label learning.

3. Multi-Label Consistency

Given an instance $X \in \mathcal{X}$, we denote by $\mathbf{p}(X)$ a vector of conditional probability for $Y \in \mathcal{Y}$, i.e.,

$$\mathbf{p}(X) = (p_Y(X))_{Y\in\mathcal{Y}} = (\Pr(Y|X))_{Y\in\mathcal{Y}},$$

for some $\mathbf{p}(X) \in \Lambda$, where Λ is the set of all possible conditional probability distribution vectors, i.e.,

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

In the following, for notational simplicity, we will suppress dependence of $\mathbf{p}(X)$ and $\mathbf{f}(X)$ on the instance X as $\mathbf{p}$ and $\mathbf{f}$, respectively, when it is clear from the context.

For an instance $X \in \mathcal{X}$, we define the conditional risk of $\mathbf{f}$ as

$$l(\mathbf{p}, \mathbf{f}) = \sum\nolimits_{Y \in \mathcal{Y}} p_Y L(\mathbf{f}, Y) = \sum\nolimits_{Y \in \mathcal{Y}} \Pr(Y|X) L(\mathbf{f}(X), Y). \tag{2}$$

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

$$R(\mathbf{f}) = \mathbb{E}_X[l(\mathbf{p}, \mathbf{f})] \quad \text{and} \quad R^* = \mathbb{E}_X\big[\inf_{\mathbf{f}}[l(\mathbf{p}, \mathbf{f})]\big].$$

We further define the *set of Bayes predictors* as

$$A(\mathbf{p}) = \big\{\mathbf{f} \colon l(\mathbf{p}, \mathbf{f}) = \inf_{\mathbf{f}'} l(\mathbf{p}, \mathbf{f}')\big\}.$$

Notice that $A(\mathbf{p}) \neq \emptyset$ since L is interval and below-bounded.

As mentioned above, the loss L in multi-label learning is generally non-convex and discontinuous, and it is difficult, even impossible, to minimize the risk given by Eq.(1) directly. In practice, a surrogate loss function Ψ is usually considered in place of L. We define the Ψ-risk and Bayes Ψ-risk of $\mathbf{f}$, respectively, as

$$R_\Psi(\mathbf{f}) = \mathbb{E}_{X,Y}[\Psi(\mathbf{f}(X), Y)] \quad \text{and} \quad R_\Psi^* = \inf_{\mathbf{f}} R_\Psi(\mathbf{f}).$$

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

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

It is obvious that $R_\Psi(\mathbf{f}) = E_X[W(\mathbf{p}, \mathbf{f})]$ and $R_\Psi^* = E_X[W^*(\mathbf{p})]$.

We now define the ***multi-label consistency*** as follows:

Definition 2 *Given a below-bounded surrogate loss function Ψ where $\Psi(\cdot, Y)$ is continuous for every $Y \in \mathcal{Y}$, Ψ is said to be multi-label consistent w.r.t. the loss L if it holds, for every $\mathbf{p} \in \Lambda$, that*

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

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 $R(\mathbf{f}) \to R^*$.

Theorem 3 *The surrogate loss Ψ is multi-label consistent w.r.t. the loss L if and only if it holds for any sequence $\mathbf{f}_n$ that*

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

We defer the proof of this theorem to Section 6.1, which is inspired by the technique of (Zhang, 2004a) and (Tewari and Bartlett, 2007).

4. Consistency w.r.t. Ranking Loss

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

$$L_{\text{rankloss}}(\mathbf{f}, (X, Y)) = \sum\nolimits_{\substack{y_i=-1\\ y_j=+1}} a_Y I[f_i(X) \geq f_j(X)] = \sum\nolimits_{y_i<y_j} a_Y I[f_i(X) \geq f_j(X)], \quad (3)$$

where a_Y is a non-negative penalty and I is the indicator function, i.e., $I[\pi]$ equals 1 if π holds and 0 otherwise. In multi-label learning, the most common penalty is

$$a_Y = |\{y_i \colon y_i = -1\}|^{-1} \times |\{y_j \colon y_j = +1\}|^{-1}.$$

In this paper we consider the more general penalty, i.e., any non-negative penalty. It is easy to see that the ranking loss is below-bounded and interval since $L_{\text{rankloss}}(\mathbf{f}, (X, Y)) \geq 0$, and for every $X, X' \in \mathcal{X}$ and $Y, Y' \in \mathcal{Y}$, it holds that either

$$L_{\text{rankloss}}(\mathbf{f}, (X, Y)) = L_{\text{rankloss}}(\mathbf{f}, (X', Y')) \text{ or } |L_{\text{rankloss}}(\mathbf{f}, (X, Y)) - L_{\text{rankloss}}(\mathbf{f}, (X', Y'))| \geq \gamma,$$

where $\gamma = \min\left\{|ia_Y - ja_{Y'}| > 0 \colon i, j \in [(Q - \lfloor\frac{Q}{2}\rfloor) \cdot \lfloor\frac{Q}{2}\rfloor]\right\}$. Considering Eq.(3), a natural choice for the surrogate loss is

$$\Psi(\mathbf{f}(X), Y) = \sum\nolimits_{\substack{y_i=-1\\ y_j=+1}} a_Y \phi(f_j(X) - f_i(X)) = \sum\nolimits_{y_i<y_j} a_Y \phi(f_j(X) - f_i(X)), \quad (4)$$

where ϕ is a convex and non-increasing real-valued function, which was chosen as hinge loss $\phi(x) = (1 - x)_+$ in (Elisseeff and Weston, 2002) and exponential loss $\phi(x) = \exp(-x)$ in (Schapire and Singer, 2000; Dekel et al., 2004; Zhang and Zhou, 2006).

Before our discussion, it is necessary to introduce the notations

$$\Delta_{i,j} = \sum\nolimits_{Y \colon y_i<y_j} p_Y a_Y \quad \text{and} \quad \delta(i, j; k, l) = \sum\nolimits_{Y \colon y_i<y_j, y_k<y_l} p_Y a_Y,$$

for a given vector $\mathbf{p} \in \Lambda$ and a non-negative vector $(a_Y)_{Y \in \mathcal{Y}}$. It is easy to get the following:

Lemma 4 *For a vector $\mathbf{p} \in \Lambda$ and a non-negative vector $(a_Y)_{Y \in \mathcal{Y}}$, the following properties hold:*

1. $\Delta_{i,i} = 0$;
2. $\delta(i, j; k, l) = \delta(k, l; i, j)$;
3. $\Delta_{i,j} = \delta(i, j; i, k) + \delta(i, j; k, j)$ *for every* $k \neq i, j$;
4. $\Delta_{i,k} + \Delta_{k,j} + \Delta_{j,i} = \Delta_{k,i} + \Delta_{i,j} + \Delta_{j,k}$;
5. $\Delta_{i,k} \leq \Delta_{k,i}$ *if* $\Delta_{i,j} \leq \Delta_{j,i}$ *and* $\Delta_{j,k} \leq \Delta_{k,j}$.

Proof The Properties 1 and 2 are immediate from the definitions. For Property 3, we have $y_i = -1$ and $y_j = +1$ for every $Y \in \mathcal{Y}$ satisfying $y_i < y_j$. For y_k $(k \neq i, j)$, there are only two choices: $y_k = +1$ or $y_k = -1$, and thus $\Delta_{i,j} = \delta(i, j; i, k) + \delta(i, j; k, j)$. For Property 4, we have, from Property 3:

$$\begin{aligned}
&\Delta_{i,k} = \delta(i, k; j, k) + \delta(i, k; i, j),\ \Delta_{k,i} = \delta(k, i; j, i) + \delta(k, i; k, j),\\
&\Delta_{j,i} = \delta(j, i; k, i) + \delta(j, i; j, k),\ \Delta_{i,j} = \delta(i, j; k, j) + \delta(i, j; i, k),\\
&\Delta_{k,j} = \delta(k, j; k, i) + \delta(k, j; i, j),\ \Delta_{j,k} = \delta(j, k; j, i) + \delta(j, k; i, k).
\end{aligned}$$

Thus, it holds by combining with Property 2. Property 5 follows from Property 4. ■

Lemma 5 *For every* $\mathbf{p} \in \Lambda$ *and non-negative vector* $(a_Y)_{Y\in\mathcal{Y}}$*, the set of Bayes predictors for ranking loss is given by*

$A(\mathbf{p}) = \{\mathbf{f}\colon$ *for all* $i < j, f_i > f_j$ *if* $\Delta_{i,j} < \Delta_{j,i}; f_i \neq f_j$ *if* $\Delta_{i,j} = \Delta_{j,i};$ *and* $f_i < f_j$ *otherwise*$\}$.

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

$$\begin{aligned}
l(\mathbf{p}, \mathbf{f}) &= \sum\nolimits_{Y\in\mathcal{Y}} p_Y L_{\text{rankloss}}(\mathbf{f}, (X, Y)) = \sum\nolimits_{Y\in\mathcal{Y}} p_Y \sum\nolimits_{y_i<y_j} a_Y I[f_i \geq f_j]\\
&= \sum\nolimits_{Y\in\mathcal{Y}} p_Y \sum\nolimits_{1\leq i,j\leq Q} a_Y I[f_i \geq f_j] \cdot I[y_i < y_j].
\end{aligned}$$

By swapping the two sums, we get

$$\begin{aligned}
l(\mathbf{p}, \mathbf{f}) &= \sum\nolimits_{1\leq i,j\leq Q} I[f_i \geq f_j] \sum\nolimits_{Y\,:\,y_i<y_j} p_Y a_Y = \sum\nolimits_{1\leq i,j\leq Q} I[f_i \geq f_j]\Delta_{i,j}\\
&= \sum\nolimits_{1\leq i<j\leq Q} I[f_i \geq f_j]\Delta_{i,j} + I[f_i \leq f_j]\Delta_{j,i}.
\end{aligned}$$

Hence we complete the proof by combining with Property 5 in Lemma 4. ■

The following theorem discloses that none convex surrogate loss is consistent with ranking loss, and the proof is deferred to Section 6.2.

Theorem 6 *For any convex function* ϕ*, the surrogate loss*

$$\Psi(\mathbf{f}(X), Y) = \sum\nolimits_{y_i<y_j} a_Y \phi(f_j(X) - f_i(X))$$

is not multi-label consistent w.r.t. ranking loss.

Intuitively, Property 5 of Lemma 4 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 $i \succeq j$, there is possible that $\Delta_{i,j} = \Delta_{j,i}$. The set of Bayes predictors of a reasonable loss function should include all functions which are compatible with this order, i.e., $\mathbf{f}$'s which enable $f_i \geq f_j$ if $i \succeq j$. In the definition of ranking loss given by Eq.(3), the same penalty term is applied to $f_i < f_j$ and $f_i = f_j$; thus, the set of Bayes predictors with respect to ranking loss does not include some functions which are compatible with the above order since 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 $\mathbf{f}^* \in \{\mathbf{f}\colon f_1 = f_2 = \cdots = f_Q\}$ but $\mathbf{f}^* \notin A(\mathbf{p})$ (from Lemma 5). So, the same penalty on $f_i < f_j$ and $f_i = f_j$ encumbers the multi-label consistency.

To overcome the deficiency, we instead introduce the *partial ranking loss*

$$L_{\text{p-rankloss}}(\mathbf{f}, (X,Y)) = \sum\nolimits_{y_i < y_j} a_Y \left(I[f_i(X) > f_j(X)] + \frac{1}{2} I[f_i(X) = f_j(X)] \right), \qquad (5)$$

which has been commonly 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_Y while the partial ranking loss uses $a_Y/2$. With a proof similar to that of Lemma 5, we can get the set of Bayes predictors with respect to the partial ranking loss:

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

Now, consider the above extreme example, 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 $\mathbf{f}^* \in \{\mathbf{f}\colon f_1 = f_2 = \cdots = f_Q\} \subseteq A(\mathbf{p})$, which exhibits multi-label consistency.

We further study more general cases, and the following theorem provides a sufficient condition for multi-label consistency w.r.t. partial ranking loss:

Theorem 7 *The surrogate loss Ψ given by Eq.(4) is multi-label consistent w.r.t. partial ranking loss if $\phi\colon \mathbb{R} \to \mathbb{R}$ is a differential and non-increasing function, and it holds that*

$$\phi'(0) < 0 \quad \textit{and} \quad \phi(x) + \phi(-x) \equiv 2\phi(0), \qquad (7)$$

i.e., $\phi(x) + \phi(-x) = 2\phi(0)$ for every $x \in \mathbb{R}$.

The proof of Theorem 7 can be found in Section 6.3, and from this theorem we can get:

Corollary 8 *The surrogate loss Ψ given by Eq.(4) is multi-label consistent w.r.t partial ranking loss if*

$$\phi(x) = -\arctan(x) \quad \textit{or} \quad \phi(x) = \frac{1-e^{2x}}{1+e^{2x}}.$$

Notice that Theorem 7 could not be applied directly to $\phi(x) = -cx^{2k+1}$ for some constant $c > 0$ and integer $k \geq 0$, since such setting yields that the surrogate loss Ψ is not below-bounded. This problem, however, can be solved by introducing a regularization term. With a proof similar to that of Theorem 7, we get:

Theorem 9 *The following surrogate loss is multi-label consistent w.r.t. partial ranking loss:*

$$\Psi(\mathbf{f}(X), Y) = \sum\nolimits_{y_i<y_j} a_Y \phi(f_j(X) - f_i(X)) + \tau\Upsilon(\mathbf{f}(X)),$$

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

For example, we can easily construct the following convex surrogate loss

$$\Psi(\mathbf{f}(X), Y) = \sum\nolimits_{y_i < y_j} -a_Y(f_j(X) - f_i(X)) + \tau \sum\nolimits_{i=1}^{Q} f_i^2(X),$$

which is multi-label consistent w.r.t. partial ranking loss.

It is worth noting that it does not mean any convex surrogate loss Ψ given by Eq.(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 10 *If $\phi\colon \mathbb{R} \to \mathbb{R}$ is a convex, differential, non-linear and non-increasing function, the surrogate loss Ψ given by Eq.(4) is not multi-label consistent w.r.t. partial ranking loss.*

The proof is deferred to Section 6.4. The following corollary shows that some state-of-the-art multi-label learning approaches (Schapire and Singer, 2000; Dekel et al., 2004; Zhang and Zhou, 2006) are even not multi-label consistent w.r.t. partial ranking loss.

Corollary 11 *If $\phi(x) = e^{-x}$ or $\phi(x) = \ln(1 + \exp(-x))$, the surrogate loss Ψ given by Eq.(4) is not multi-label consistent w.r.t. partial ranking loss.*

In summary, the ranking loss has been suggested as a standard from early work by Schapire and Singer (2000), and many state-of-the-art multi-label learning approaches work with the formulation given by Eq.(4). However, our analysis shows that none convex surrogate loss functions are consistent w.r.t. ranking loss. Thus, ranking loss might not be a good loss function and evaluation criterion for multi-label learning. The partial ranking loss is more reasonable than ranking loss, since it enables many, though not all, convex surrogate loss functions to have consistency. In future work it would be interesting to develop some new multi-label learning approaches based on minimizing the partial ranking loss, and design other loss functions with better consistency.

5. Consistency w.r.t. Hamming Loss

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

$$L_{\text{hamloss}}(F(\mathbf{f}(X)), Y) = \frac{1}{Q} \sum\nolimits_{i=1}^{Q} I[\hat{y}_i \neq y_i],$$

where $\hat{Y} = F(\mathbf{f}(X)) = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_Q)$.

Hamming loss is obviously below-bounded and interval since $0 \leq L_{\text{hamloss}}(F(\mathbf{f}(X)), Y) \leq 1$, and for every $X, X' \in \mathcal{X}$ and $Y, Y' \in \mathcal{Y}$, it holds either $L_{\text{hamloss}}(F(\mathbf{f}(X)), Y)) = L_{\text{hamloss}}(F(\mathbf{f}(X')), Y'))$ or $|L_{\text{hamloss}}(F(\mathbf{f}(X)), Y)) - L_{\text{hamloss}}(F(\mathbf{f}(X')), Y'))| \geq 1/Q$. We have the conditional risk:

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

and the set of Bayes predictors with regard to hamming loss:

$$A(\mathbf{p}) = \Big\{\mathbf{f} = \mathbf{f}(X) \colon \hat{Y} = F(\mathbf{f}) \text{ with } \hat{y}_i = \operatorname{sgn}\Big(\textstyle\sum_{Y\in\mathcal{Y}, y_i=+1} p_Y - \dfrac{1}{2}\Big)\Big\}. \tag{8}$$

A straightforward multi-label learning approach is to regard each subset of labels as a new class and then try to learn 2^Q functions, i.e., $\mathbf{f} = (f_Y)_{Y\in\mathcal{Y}}$. Such a prediction function is given by

$$F(\mathbf{f}(X)) = \max_{Y\in\mathcal{Y}} f_Y(X). \tag{9}$$

This approach can be viewed as a direct extension of the one-vs-all strategy for multi-class learning. We consider the following formulation:

$$\Psi(\mathbf{f}(X), Y) = \max_{\hat{Y}\neq Y} \phi(\delta(\hat{Y}, Y) + f_{\hat{Y}}(X) - f_Y(X)), \tag{10}$$

where $\phi(x)$ and $\delta(Y, \hat{Y})$ are set as $\phi(x) = \max(0, x)$ and $\delta(Y, \hat{Y}) = \sum_{i=1}^{Q} I[y_i \neq \hat{y}_i]$, respectively, by Taskar et al. (2004); Hariharan et al. (2010).

Before a further discussion, we divide multi-label classification tasks into two categories, i.e., deterministic and non-deterministic, as follows:

Definition 12 *In multi-label classification, if for every instance $X \in \mathcal{X}$ there exists a label $Y \in \mathcal{Y}$ such that $P(Y|X) > 0.5$, the task is deterministic, and non-deterministic otherwise.*

Consistency for the deterministic case is easier than non-deterministic case. For example, many formulations of SVMs are inconsistent for non-deterministic multi-class classification (Zhang, 2004a; Tewari and Bartlett, 2007), but consistent for deterministic case as indicated by Zhang (2004a). The following lemma shows that the approaches of (Taskar et al., 2004; Hariharan et al., 2010) are not multi-label consistent w.r.t. hamming loss, even for the deterministic case.

Lemma 13 *For deterministic multi-label classification, the surrogate loss Ψ given by Eq.*(10) *with $\delta(\hat{Y}, Y) = \sum_{i=1}^{Q} I[y_i \neq \hat{y}_i]$ is not multi-label consistent w.r.t. hamming loss.*

However, if we instead choose $\delta(\hat{Y}, Y) = I[\hat{Y} \neq Y]$, the following theorem guarantees that the surrogate loss is multi-label consistent w.r.t. hamming loss, at least for the deterministic case.

Theorem 14 *For deterministic multi-label classification, the surrogate loss Ψ given by Eq.*(10) *with $\delta(\hat{Y}, Y) = I[Y \neq \hat{Y}]$ is multi-label consistent w.r.t. hamming loss.*

The proofs of Lemma 13 and Theorem 14 can be found in Sections 6.5 and 6.6, respectively.

Alternatively, it is possible to transform a multi-label learning task into Q independent binary classification tasks (Boutell et al., 2004). Now the goal is to learn Q functions, $\mathbf{f} = (f_1, f_2, \ldots, f_Q)$, and the prediction function is given by

$$F(\mathbf{f}(X)) = (\operatorname{sgn}[f_1(X)], \operatorname{sgn}[f_2(X)], \ldots, \operatorname{sgn}[f_Q(X)]).$$

A common choice for the surrogate loss is

$$\Psi(\mathbf{f}(X), Y) = \sum\nolimits_{i=1}^{Q} \phi(y_i f_i(X)), \tag{11}$$

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). We have the conditional surrogate loss

$$\begin{aligned} W(\mathbf{p}, \mathbf{f}) &= \sum\nolimits_{Y \in \mathcal{Y}} p_Y \Psi(\mathbf{f}(X), Y) = \sum\nolimits_{i=1}^{Q} \sum\nolimits_{Y \in \mathcal{Y}} p_Y \phi(y_i f_i(X)) \\ &= \sum\nolimits_{i=1}^{Q} p_i^+ \phi(f_i(X)) + (1 - p_i^+) \phi(-f_i(X)), \end{aligned}$$

where $p_i^+ = \sum_{Y:\, y_i = +1} p_Y$ and $1 - p_i^+ = \sum_{Y:\, y_i = -1} p_Y$. 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 $W(\mathbf{p}, \mathbf{f})$ is equivalent to minimizing $W_i(p_i^+, f_i)$ for all $1 \le i \le Q$, that is

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

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

Theorem 15 *The surrogate loss Ψ given by Eq.*(11) *is consistent w.r.t. hamming loss for convex function ϕ with $\phi'(0) < 0$.*

It is evident from this theorem that the surrogate loss Ψ given by Eq.(11) is consistent w.r.t. hamming loss if ϕ is any of the following:

- Exponential: $\phi(x) = e^{-x}$;
- Hinge: $\phi(x) = \max(0, 1-x)$;
- Least squares: $\phi(x) = (1-x)^2$;
- Logistic regression: $\phi(x) = \ln(1 + \exp(-x))$.

Notice that in this paper, we do not consider how to learn the real-valued functions $\mathbf{f}$ for small data or large number of labels, where many labels or label subsets lack enough training examples, while this is a challenging task and the exploitation of label correlation is therefore needed. In our analysis, we assume sufficient training data and ignore the label correlation. Moreover, we do not consider how to decide the number of relevant labels for ranking loss and partial ranking loss, while in practice this is very challenging. These are important future issues.

6. Proofs

6.1. Proof of Theorem 3

We first introduce some useful lemmas:

Lemma 16 *$W^*(\mathbf{p})$ is continuous on Λ.*

Proof From the Heine definition of continuity, we need to show that $W^*(\mathbf{p}^{(n)}) \to W^*(\mathbf{p})$ for any sequence $\mathbf{p}^{(n)} \to \mathbf{p}$.

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

$$\sum\nolimits_{Y\in\mathcal{Y}} p_Y^{(n)}\Psi(\mathbf{f},Y) \to \sum\nolimits_{Y\in\mathcal{Y}} p_Y\Psi(\mathbf{f},Y)$$

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

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

From the fact $W^*(\mathbf{p}^{(n)}) \leq \inf_{\mathbf{f}\in B_r}\sum_{Y\in\mathcal{Y}} p_Y^{(n)}\Psi(\mathbf{f},Y)$, by letting $r\to\infty$, we have:

$$\limsup_{n\to\infty} W^*(\mathbf{p}^{(n)}) \leq W^*(\mathbf{p}). \tag{12}$$

Denote by $\mathcal{Y}' = \{Y|p_Y > 0 \text{ for } Y\in\mathcal{Y}\}$ and assume $\Psi(\cdot,\cdot)\geq C$ for some constant C (since Ψ is below-bounded). We have $W^*(\mathbf{p}^{(n)}) \geq \inf_{\mathbf{f}}\sum_{Y\in\mathcal{Y}'} p_Y^{(n)}\Psi(\mathbf{f},Y) + C\sum_{Y\in\mathcal{Y}/\mathcal{Y}'} p_Y^{(n)}$, which yields

$$\liminf_{n\to\infty} W^*(\mathbf{p}^{(n)}) \geq \liminf_{n\to\infty}\left(\inf_{\mathbf{f}}\sum\nolimits_{Y\in\mathcal{Y}'} p_Y^{(n)}\Psi(\mathbf{f},Y) + C\sum\nolimits_{Y\in\mathcal{Y}/\mathcal{Y}'} p_Y^{(n)}\right) = W^*(\mathbf{p}),$$

which completes the proof by combining Eq.(12). ■

Lemma 17 *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 $\mathbf{p}\in\Lambda$, $l(\mathbf{p},\mathbf{f}) - \inf_{\mathbf{f}'} l(\mathbf{p},\mathbf{f}') \geq \epsilon$ implies $W(\mathbf{p},\mathbf{f}) - W^*(\mathbf{p}) \geq \delta$.*

Proof We proceed by contradiction. Suppose Ψ is multi-label consistent and there exists $\epsilon > 0$ and a sequence $(\mathbf{p}^{(n)},\mathbf{f}^{(n)})$ such that $l(\mathbf{p}^{(n)},\mathbf{f}^{(n)}) - \inf_{\mathbf{f}'} l(\mathbf{p}^{(n)},\mathbf{f}') \geq \epsilon$ and $W(\mathbf{p}^{(n)},\mathbf{f}^{(n)}) \to W^*(\mathbf{p}^{(n)})$. From the compactness of Λ, there exists a convergence sequence n_k such that $\mathbf{p}^{(n_k)} \to \mathbf{p}$ for some $\mathbf{p}\in\Lambda$. From Lemma 16, we have

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

Similar to the proof of Lemma 16, denoted by $\mathcal{Y}' = \{Y|p_Y > 0 \text{ for } Y\in\mathcal{Y}\}$, we have

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

This gives $W(\mathbf{p}, \mathbf{f}^{(n_k)}) \to W^*(\mathbf{p})$ from the definition of $W^*(\mathbf{p})$. Since Ψ is multi-label consistent, there exists a sequence $\mathbf{f}^{(n_{k_i})}$ satisfying $l(\mathbf{p}, \mathbf{f}^{(n_{k_i})}) \to \inf_{\mathbf{f}'} l(\mathbf{p}, \mathbf{f}')$, which contradicts the assumption $l(\mathbf{p}^{(n)}, \mathbf{f}^{(n)}) - \inf_{\mathbf{f}'} l(\mathbf{p}^{(n)}, \mathbf{f}') \geq \epsilon$, and thus the lemma holds. ■

Proof of Theorem 3:

("⇒") We first introduce a new notation

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

It is obvious that $H(0) = 0$ and $H(\epsilon) > 0$ for $\epsilon > 0$ from Lemma 17. Corollary 26 of (Zhang, 2004a) guarantees there exists 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 $\mathbf{p}$ such that $W^*(\mathbf{p}) = \inf_{\mathbf{f}}\{W(\mathbf{p}, \mathbf{f}) \colon \mathbf{f} \notin A(\mathbf{p})\}$. Let $\mathbf{f}^{(n)} \notin A(\mathbf{p})$ be a sequence such that $W(\mathbf{p}, \mathbf{f}^{(n)}) \to W^*(\mathbf{p})$. For simplicity, we consider $\mathcal{X} = \{x\}$, i.e., only one instance, and set $\mathbf{f}_n(x) = \mathbf{f}^{(n)}$. Then

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

yielding $l(\mathbf{p}, \mathbf{f}^{(n)}) \to \inf_{\mathbf{f}} l(\mathbf{p}, \mathbf{f})$ which is contrary to

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

where $\gamma(\mathbf{p}) = \inf_{\mathbf{f}\notin A(\mathbf{p})} l(\mathbf{p}, \mathbf{f}) - \inf_{\mathbf{f}} l(\mathbf{p}, \mathbf{f}) > 0$, since $\mathbf{f}^{(n)} \notin A(\mathbf{p})$ and L is interval. Thus we complete the proof. ■

6.2. Proof of Theorem 6

We proceed by contradiction. Suppose that the surrogate loss Ψ is multi-label consistent with ranking loss. We have

$$\begin{aligned} W(\mathbf{p}, \mathbf{f}) &= \sum\nolimits_{Y\in\mathcal{Y}} p_Y \Psi(\mathbf{f}, Y) = \sum\nolimits_{Y\in\mathcal{Y}} p_Y a_Y \sum\nolimits_{y_i<y_j} \phi(f_j - f_i) \\ &= \sum\nolimits_{1\leq i<j\leq Q} \phi(f_j - f_i)\Delta_{i,j} + \phi(f_i - f_j)\Delta_{j,i}. \end{aligned}$$

Consider the probability vector $\mathbf{p} = (p_Y)_{Y\in\mathcal{Y}}$ and penalty vector $(a_Y)_{Y\in\mathcal{Y}}$ s.t. $P_{Y_1} = P_{Y_2}$ and $a_{Y_1} = a_{Y_2}$ for every $Y_1 \neq Y_2$, $Y_1, Y_2 \in \mathcal{Y}$. This yields that $\Delta_{i,j} = \Delta_{m,n}$ for every $1 \leq i \neq j, m \neq n \leq Q$, and thus we get

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

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

$$W^*(\mathbf{p}) = W(\mathbf{p}, \hat{\mathbf{f}}) = \Delta_{1,2} \sum\nolimits_{1\leq i<j\leq Q} 2\phi(0) = Q(Q-1)\phi(0)\Delta_{1,2},$$

where $\hat{\mathbf{f}} = \{\hat{\mathbf{f}} \colon \hat{f}_1 = \hat{f}_2 = \ldots = \hat{f}_Q\}$. Notice that $\hat{\mathbf{f}} \notin A(\mathbf{p})$ from Lemma 5, and we have

$$W^*(\mathbf{p}) = \inf_{\mathbf{f}} \{W(\mathbf{p}, \mathbf{f}) \colon \mathbf{f} \notin A(\mathbf{p})\},$$

which completes the proof. ∎

6.3. Proof of Theorem 7

For any $\mathbf{p} \in \Lambda$ and non-negative vector $(a_Y)_{Y \in \mathcal{Y}}$, we will prove that, $f_i > f_j$ if $\Delta_{i,j} < \Delta_{j,i}$ and $f_i < f_j$ if $\Delta_{i,j} > \Delta_{j,i}$, for all $\mathbf{f}$ satisfying $W^*(\mathbf{p}) = W(\mathbf{p}, \mathbf{f})$. Without loss of generality, it is sufficient to prove that $f_1 > f_2$ if $\Delta_{1,2} < \Delta_{2,1}$.

We proceed by contradiction, and assume that there exists a vector $\mathbf{f}$ such that $f_1 \leq f_2$ and $W^*(\mathbf{p}) = W(\mathbf{p}, \mathbf{f})$.

For the case $f_1 < f_2$, we can introduce another vector $\mathbf{f}'$ with $f_1' = f_2$, $f_2' = f_1$ and $f_k' = f_k$ for $k \neq 1, 2$. From the definition of conditional surrogate risk, we have

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

which yields that

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

From Property 4 of Lemma 4, we have

$$\Delta_{1,i} - \Delta_{2,i} - \Delta_{i,1} + \Delta_{i,2} = \Delta_{1,2} - \Delta_{2,1}. \tag{13}$$

It follows that

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

where the last equality holds by the condition $\phi(x) + \phi(-x) \equiv 2\phi(0)$. For non-increasing function ϕ with $\phi'(0) < 0$, we have $\phi(z) < \phi(-z)$ for all $z > 0$, and $\phi(f_i - f_1) \leq \phi(f_i - f_2)$. From $\Delta_{1,2} < \Delta_{2,1}$ we get $W(\mathbf{p}, \mathbf{f}) > W(\mathbf{p}, \mathbf{f}')$, which is contrary to $W(\mathbf{p}, \mathbf{f}) = W^*(\mathbf{p})$.

We now consider the case $f_1 = f_2$. For the optimal solution, we have the first-order condition $\frac{\partial}{\partial f_i} W(\mathbf{p}, \mathbf{f}) = 0$ for $i = 1, 2$:

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

By combining Eq.(13), $f_1 = f_2$ and $\phi'(x) = \phi'(-x)$ from Eq.(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 impossible since $\Delta_{1,2} < \Delta_{2,1}$ and $2\phi'(0) + \sum_{i \neq 1,2} \phi'(f_1 - f_i) \leq 2\phi'(0) < 0$. Thus, we complete the proof. ∎

6.4. Proof of Theorem 10

For convex function ϕ we have $\phi'(x) \le \phi'(y)$ for every $x \le y$ from (Rockafellar, 1997), and the derivative function $\phi'(x)$ is continuous for $x \in \mathbb{R}$ if ϕ is differential and convex. Since ϕ is non-increasing, we have $\phi'(x) \le 0$ for all $x \in \mathbb{R}$, and without loss of generality, we assume $\phi'(x) < 0$.

We proceed by contradiction. Assume the surrogate loss Ψ is multi-label consistent with partial ranking loss for some non-linear function ϕ. Then, from the continuity of $\phi'(x)$, there exists an interval (c, d) for $c < d < 0$ or $0 < c < d$, such that

$$\phi'(x) < \phi'(y) \text{ for every } x < y, x, y \in (c, d).$$

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

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

$$G(x) = \big(\phi'(x-a) - \phi'(a-x)\big)\big(\phi'(a) + \phi'(x)\big) + \phi'(x)\big(\phi'(-a) - \phi'(a)\big).$$

It is easy to find that $G(a) = \phi'(a)\big(\phi'(-a) - \phi'(a)\big) > 0$ and $G(x)$ is continuous. 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

$$\frac{\phi'(b-a)(\phi'(a) + \phi'(b)) + \phi'(b)\phi'(-a)}{\phi'(a-b)(\phi'(a) + \phi'(b)) + \phi'(b)\phi'(a)} > 1. \tag{14}$$

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

$$0 < \frac{\phi'(-a)}{\phi'(a)} \le \frac{\phi'(-b)}{\phi'(a)} < \frac{\phi'(-b)}{\phi'(b)}. \tag{15}$$

We consider the following multi-label classification with $Q = 3$ labels:

$$Y_1 = (-1, +1, +1), Y_2 = (+1, -1, -1), Y_3 = (+1, +1, -1), Y_4 = (-1, -1, +1).$$

Let $\mathbf{f} = (f_1, f_2, f_3)$ such that $a = f_3 - f_1$ and $b = f_3 - f_2$, and thus $f_1 > f_2$. For every probability vector $\mathbf{p} = (p_{Y_1}, p_{Y_2}, p_{Y_3}, p_{Y_4}) \in \Lambda$ and every penalty vector $(a_{Y_1}, a_{Y_2}, a_{Y_3}, a_{Y_4})$, we have

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

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 $P_i = p_{Y_1} a_{Y_1}$ for $i = 1, 2, 3, 4$. In the following, we will construct some $\bar{\mathbf{p}}$ and $\bar{\mathbf{a}}$ such that $W^*(\bar{\mathbf{p}}) = W(\bar{\mathbf{p}}, \mathbf{f})$ and $\Delta_{1,2} > \Delta_{2,1}$.

The subgradient condition for optimality of $W(\mathbf{p}, \mathbf{f})$ gives that

$$\begin{aligned} &\partial W(\mathbf{p}, \mathbf{f})/\partial f_1 = -P_1\phi'(a-b) + P_2\phi'(b-a) - (P_1 + P_4)\phi'(a) + (P_2 + P_3)\phi'(-a) = 0, \\ &\partial W(\mathbf{p}, \mathbf{f})/\partial f_2 = P_1\phi'(a-b) - P_2\phi'(b-a) - P_4\phi'(b) + P_3\phi'(-b) = 0, \\ &\partial W(\mathbf{p}, \mathbf{f})/\partial f_3 = (P_1 + P_4)\phi'(a) - (P_2 + P_3)\phi'(-a) + P_4\phi'(b) - P_3\phi'(-b) = 0, \end{aligned}$$

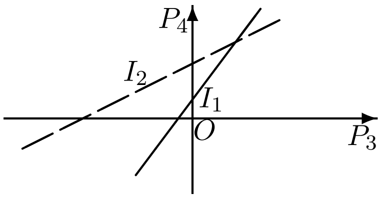

Figure 1: Lines I_1 (solid) and I_2 (dash) corresponding to Eqs. (16) and (17), respectively.

equivalent to

$$\begin{aligned} P_1\phi'(a-b) - P_2\phi'(b-a) &= P_4\phi'(b) - P_3\phi'(-b), & (16)\\ -P_1\phi'(a) + P_2\phi'(-a) &= P_4(\phi'(a)+\phi'(b)) - P_3(\phi'(-a)+\phi'(-b)). & (17)\end{aligned}$$

From Lemma 18, there exists $\bar{\mathbf{p}} = (\bar{p}_{Y_1}, \bar{p}_{Y_2}, \bar{p}_{Y_3}, \bar{p}_{Y_4})$ and $(\bar{a}_{Y_1}, \bar{a}_{Y_2}, \bar{a}_{Y_3}, \bar{a}_{Y_4})$ satisfying Eqs. (16), (17) and $P_1 > P_2$. Hence this yields $W(\bar{\mathbf{p}}, \mathbf{f}) = W^*(\bar{\mathbf{p}})$. Notice that $\mathbf{f} \notin A(\bar{\mathbf{p}})$ from Eq.(6), since $\Delta_{1,2} = P_1 > P_2 = \Delta_{2,1}$ and $f_1 > f_2$, we have

$$W^*(\bar{\mathbf{p}}) = \inf_{\mathbf{f}}\{W(\bar{\mathbf{p}}, \mathbf{f}) \colon \mathbf{f} \notin A(\bar{\mathbf{p}})\},$$

which completes the proof. ■

Lemma 18 *There exist some* $P_1 > P_2 > 0$*,* $P_3 > 0$ *and* $P_4 > 0$ *satisfying Eqs.* (16) *and* (17)*, if Eqs.* (14) *and* (15) *hold for some* $0 < a < b$*.*

Proof From Eq.(14), we set

$$1 < \frac{P_1}{P_2} < \frac{\phi'(b-a)(\phi'(a)+\phi'(b)) + \phi'(b)\phi'(-a)}{\phi'(a-b)(\phi'(a)+\phi'(b)) + \phi'(b)\phi'(a)}. \tag{18}$$

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

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

which gives $P_1\phi'(a-b) - P_2\phi'(b-a) < 0$. Thus, Eq.(16) corresponds to the Line I_1 in Figure 1. From Eq.(15), 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 Eqs. (16) and (17), 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 Eq.(18). Thus, we complete the proof. ■

6.5. Proof of Lemma 13

We consider a multi-label classification task with $Q = 2$ labels, i.e., $\mathcal{L} = \{1, 2\}$ and let $Y_1 = (-1, -1)$, $Y_2 = (-1, 1)$, $Y_3 = (1, 1)$ and $Y_4 = (1, -1)$. Suppose $p_{Y_4} = 0$ and $p_{Y_2} + p_{Y_3} < p_{Y_1} < 2p_{Y_2} + p_{Y_3}$. It is evident that $p_{Y_1} > 0.5$ and thus, this multi-label classification task is deterministic. It is necessary to consider $\mathbf{f} = (f_{Y_1}, f_{Y_2}, f_{Y_3})$. By combining Eqs. (8) and (10), we get

$$A(\mathbf{p}) = \{\mathbf{f}\colon f_{Y_1} \geq f_{Y_2} \text{ and } f_{Y_1} \geq f_{Y_3}\}.$$

We also have

$$\begin{aligned} W(\mathbf{p}, \mathbf{f}) = p_{Y_1} \max\{\phi(1 + f_{Y_2} - f_{Y_1}), \phi(2 + f_{Y_3} - f_{Y_1})\} + p_{Y_2} \max\{\phi(1 + f_{Y_1} - f_{Y_2}), \\ \phi(1 + f_{Y_3} - f_{Y_2})\} + p_{Y_3} \max\{\phi(2 + f_{Y_1} - f_{Y_3}), \phi(1 + f_{Y_2} - f_{Y_3})\} \end{aligned}$$

and

$$W^*(\mathbf{p}) = W(\mathbf{p}, \mathbf{f}^*) = \inf_{\mathbf{f}} W(\mathbf{p}, \mathbf{f}) = 2p_{Y_1} + 2p_{Y_3},$$

where $\mathbf{f}^* = (f^*_{Y_1}, f^*_{Y_2}, f^*_{Y_3})$ such that $f^*_{Y_1} = f^*_{Y_3}$ and $f^*_{Y_1} = f^*_{Y_2} - 1$. Hence $\mathbf{f} \notin A(\mathbf{p})$ and it is not multi-label consistent. ■

6.6. Proof of Theorem 14

We first introduce a useful lemma:

Lemma 19 *For the surrogate loss Ψ given by Eq.(10) with $\delta(\hat{Y}, Y) = I[\hat{Y} \neq Y]$, and for every $\mathbf{p} \in \Lambda$ and $\mathbf{f}$ such that $W(\mathbf{p}, \mathbf{f}) = W^*(\mathbf{p})$, we have $f_{Y_1} \geq f_{Y_2}$ if $p_{Y_1} > p_{Y_2}$.*

Proof We proceed by contradiction. Assume there exist some $\mathbf{p} \in \Lambda$ and $\mathbf{f}$ such that $f_{Y_1} < f_{Y_2}$, $p_{Y_1} > p_{Y_2}$ and $W^*(\mathbf{p}) = W(\mathbf{p}, \mathbf{f})$. We define a new real-valued vector $\mathbf{f}'$ such that $f'_{Y_1} = f_{Y_2}$, $f'_{Y_2} = f_{Y_1}$ and $f'_{Y_i} = f_{Y_i}$ for $i \neq 1, 2$. This follows that

$$W(\mathbf{p}, \mathbf{f}) - W(\mathbf{p}, \mathbf{f}') = (p_{Y_1} - p_{Y_2})\big(\max_{Y \neq Y_1} \phi(1 + f_Y(X) - f_{Y_1}(X)) - \max_{Y \neq Y_2} \phi(1 + f_Y(X) - f_{Y_2}(X))\big).$$

From the assumption $f_{Y_1} < f_{Y_2}$, we have

$$\max_{Y \neq Y_1} \phi(1 + f_Y(X) - f_{Y_1}(X)) > \max_{Y \neq Y_2} \phi(1 + f_Y(X) - f_{Y_2}(X)),$$

leading to $W(\mathbf{p}, \mathbf{f}) - W(\mathbf{p}, \mathbf{f}') > 0$, which is contrary to the assumption $W^*(\mathbf{p}) = W(\mathbf{p}, \mathbf{f})$. Thus, we complete the proof. ■

Proof of Theorem 14: Without loss of generality, we consider a probability vector $\mathbf{p} = (p_Y)_{Y \in \mathcal{Y}}$ satisfying $p_{Y_1} > p_{Y_2} \geq p_{Y_3} \ldots \geq p_{Y_{2^Q}}$ and $p_{Y_1} > 0.5$. It is easy to get

$$A(\mathbf{p}) = \{\mathbf{f}\colon f_{Y_1} > f_Y \text{ for } Y \neq Y_1\},$$

from Eqs. (8) and (9). On the other hand, for each $\mathbf{f}$ satisfying $W(\mathbf{p}, \mathbf{f}) = W^*(\mathbf{p})$, we have

$$f_{Y_1} \geq f_{Y_2} \geq \ldots \geq f_{Y_{2^Q}}$$

from Lemma 19. Thus, from the fact that

$$W(\mathbf{p}, \mathbf{f}) = p_{Y_1} \phi(1 + f_{Y_2} - f_{Y_1}) + \sum\nolimits_{Y \neq Y_1} p_Y \phi(1 + f_{Y_1} - f_Y).$$

and $p_{Y_1} > \sum_{Y \neq Y_1} p_Y$, we complete the proof. ■

Acknowledgments

The authors want to thank the anonymous reviewers for their helpful comments and suggestions, and Tong Zhang for reading a preliminary version. This research was supported by the National Fundamental Research Program of China (2010CB327903), the National Science Foundation of China (61073097) and the Jiangsu Science Foundation (BK2008018).

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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