Theoretical Characteristics on Scoring Function in Multi-dividing Setting

Linli Zhu, Yu Pan, Mohammad Reza Farahani and Wei Gao

Abstract—The learning and optimizing of the scoring function are widely used in neural network, information retrieval and protein analysis. The multi-dividing ontology algorithms have drawn plenty of attention recent years. In this paper, we consider the Bayes-optimal scoring function in multi-dividing setting. By virtue of conditional risk, proper loss theory and derivative computing, we determine the scoring function in multi-dividing setting for certain special case. The results achieved in our paper illustrate the promising application prospects for multi-dividing ontology algorithm.

Index Terms—About four key words or phrases in alphabetical order, separated by commas, for example, visual-servoing, tracking, biomimetic, redundancy, degrees-of-freedom

I. INTRODUCTION

In computer science application, the goal of a large number of the algorithms is to get a scoring function which maps each object into a real number. The relationship between these objects is represented by their corresponding real numbers. These scoring functions are employed in computer science, biology science, chemical science and pharmaceutical science.

Example 1. In information retrieval, the user inputs a query \( q \), and the computer should return a list in which the items are related to query \( q \). The order of the items in list is determined by the scoring function which returns the information about the similarity between query and object.

Example 2. The goal of ontology mapping is returning a scoring function which maps each vertex in multi-ontology graph into a real number, and the similarities between vertices in different ontologies are reflected by the difference between their scores. At last, the ontology map is constructed based on the score differences.

Example 3. In biology science, scoring function is designed to excavate the relationship between the molecular structure of protein and the disease. In these mathematical settings, a vector with a certain dimension is taken to express the features of the disease and the structure of molecular and protein. The scoring function in high dimension is learned by virtue of the selected sample which maps all objects into a real line. In essence, such scoring function plays active roles in dimensional reducing.


Specially, scoring function learning is widely used in ontology similarity measure and ontology mapping. Lan et al. [11] explored the learning theory approach for ontology similarity computation in a setting when the ontology graph is a tree. He uses the multi-dividing algorithm in which the vertices can be divided into \( k \) parts corresponding to the \( k \) classes of rates. The rate values of all classes are decided by experts. Then, a vertex in a rate \( a \) has larger value than any vertex in rate \( b \) (where \( 1 \leq a < b \leq k \) under ontology scoring function \( f \). Finally, the similarity between two ontology vertices is measured by the difference of two real corresponding numbers. Thus, the multi-dividing algorithm is reasonable to learn a scoring function for an ontology graph with a tree structure. Zhu et. al.,[12] proposed a new criterion for multi-dividing ontology algorithm from AUC standpoint, which was designed to avoid the choice of loss function.

Furthermore, several papers have contributed to the theoretical analysis for different ontology settings with special scoring ontology function. Gao and Xu [13] investigated the uniform stability of multi-dividing ontology
algorithm and gave the generalization bounds for stable multi-dividing ontology algorithms. Gao et al. [14] researched the strong and weak stability of multi-dividing ontology algorithm. Gao and Xu [15] learned some characteristics for such ontology algorithm. Gao et al., [16] studied the multi-dividing ontology algorithm from a theoretical view. It is highlighted that empirical multi-dividing ontology model can be expressed as conditional linear statistical, and an approximation result is achieved based on the projection method. Gao et al. [17] presented the characteristics of the best ontology scoring function among piece constant ontology scoring functions. Gao et al. [18] investigated the upper bound and the lower bound minimax learning rates are obtained based on low noise assumptions. Gao et al. [19] and Yan et al. [20] presented an approach of piecewise constant function approximation for ACR criterion multi-dividing ontology algorithms. For more related results, refer to Lan et al. [21], Gao et al. [22], Gao and Shi [23], Gao et al. [24] and Yu et al. [25].

In this paper, we consider the scoring function learning problem in Bayes-optimal multi-dividing setting. The contribution of this paper is to show the Bayes-optimal multi-dividing scoring function in some special condition. The paper is organized as follows: we introduce the basic setting and algorithm in Section II; then in Section III, we present and prove the main result of this paper. The structure of Section III is designed as follows: first, we show the pair-scorders in multi-dividing setting by using the technology of conditional risk; second, the univariate function is obtained in decomposable case; third, we deal with non-decomposable situation and the result manifested that scoring function can be obtained under some special assumptions; at last, we discuss the scoring function for p-norm push risk in multi-dividing setting, and the results yielded by derivative calculation show that Bayes-optimal multi-dividing scoring function for (l,g)-push can be constructed under proper conditions.

II. SETTING AND MAIN ALGORITHM

Let \( \mathbb{R} \) be the set of real numbers, and \( \mathbb{R}^+ = [0, \infty) \). In the standard supervised multi-dividing setting, we say instance space \( X \) takes its value in a high dimension feature space (often \( \mathbb{R}^n \)), and a label space \( Y = \{1, \cdots, k\} \). An element \( x \in X \) is called an instance, and an element \( y \in \{1, \cdots, k\} \) is called a label. The elements in \( X \) are drawn independently and randomly according to some unknown distribution \( \rho \). For arbitrary sets \( X \) and \( Y \), we denote \( X \setminus Y \) as the set difference. For convenience, slightly confusing different notations, we use \( X, Y \) to denote random variables and also arbitrary sets. Hence, \( E[X] \) is denoted as the expectation of a random variable.

For a given set \( S \), \( \Delta_S \) is denoted as the set of all distributions on \( S \). Let \( \Theta \in [0, 1] \) be a parameter, we use \( \text{Ber}(\Theta) \) to express the Bernoulli distribution. The multi-dividing method is a special kind of scoring function learning approach in which instances come from \( k \) categories and the learner is given examples of instances labeled as the \( k \) classes.

Formally, the settings of multi-dividing scoring function problems can be described as follows. The learner is given a training sample \( (S_1, S_2, \ldots, S_k) \in X^n \times X^n \times \cdots \times X^n \) consisting of a sequence of training sample \( S_a = (x_1^a, \ldots, x_n^a) \) \((1 \leq a \leq k)\). The goal is to get a real-valued scoring function \( f: X \to \mathbb{R} \) that is learned from these samples. Meanwhile, it orders the future \( S_a \) instances to have higher scores than \( S_{\phi} \), where \( a \neq b \). We assume that instances in each \( S_a \) are drawn randomly and independently according to some (unknown) distribution \( \Delta^a \) on the instance space \( X \) respectively.

For any scoring function \( f: X \to \mathbb{R} \), \( \arg \min_{x \in X} f(x) \) is denoted as the set of all \( x \in X \) such that \( f(x) \leq f(x') \) for all \( x' \in X \). If scoring function \( f \) has a unique minimiser, this can be expressed by \( \min_{x \in X} f(x) \). For each pair \( (x, x') \in X \times X \), \( \text{Diff}(f) : X \times X \to \mathbb{R} \) is denoted by the function satisfying \( \text{Diff}(f)(x, x') = f(x) - f(x') \). Let \( \text{Diff}(F) = \{ \text{Diff}(f) : f \in F \} \) for a function set \( F = \{ f : X \to \mathbb{R} \} \).

We use the \( \prod(-) \) to denote the indicator function, whose value is 1 if the argument is true and 0 otherwise. In this way, sign function can be defined as \( \text{sign}(x) = \prod(x \geq 0) - \prod(x \leq 0) \) for any \( x \in \mathbb{R} \). The standard sigmoid function is denoted by \( \sigma(z) = \frac{1}{1 + e^{-z}} \).

Let \( V \subseteq \mathbb{R} \), a scorer (scoring function) \( s: X \to V \). For instance, a classifier is a special scorer with \( V = \{1, \cdots, k\} \), and a class-probability estimator is other kind of particular scorer with \( V = [0, 1] \). A pair-scorer \( s_{\phi} \) for a product space \( X \times X \) \((X^a \times X^b)\) in multi-dividing setting for pair \((a, b)\) with \(1 \leq a < b \leq k\) is certain function \( s_{\phi}: X \times X \to V \). A pair-scorer \( s_{\phi} \) is called decomposable if

\[
s_{\phi} \in S_{\text{Decomp}} = \{ \text{Diff}(s) : s : X \to \mathbb{R} \}.
\]

A loss function (in many references, it called cost function) \( l \) is some measurable function \( f: \{1, \cdots, k\} \times \mathbb{R} \to \mathbb{R} \), which can be used to measure the difference between goal scorer and the scoring function we obtained from the algorithm. We use \( l_a(v) = l(a, v) \) and \( l_b(v) = l(b, v) \) to express the individual partial losses for each pair of \((a, b)\) with \(1 \leq a < b \leq k\). Slightly abusing notation, we sometimes specify a loss via \( l(v) = (l_a(v), l_b(v)) \). A loss function \( l \) is symmetric if \( l_a(v) = l_b(-v) \) holds for each \( v \in V \) and all pair of \((a, b)\) with \(1 \leq a < b \leq k\). We say it is a margin loss if \( l(v, v') = \phi(vv') \) for some \( \phi: \mathbb{R} \to \mathbb{R} \). The conditional l-risk then defined as

\[
\begin{align*}
L_a(\eta, s) &= E_{y \sim \text{Ber}(\eta)}(l(Y, s)) \\
&= \sum_{a=1}^{k-1} \sum_{b=a+1}^{k} \eta^a l_a(s) + (1 - \eta^a) l_b(s).
\end{align*}
\]

Here, \( \eta \) is the posterior distribution, and its restriction on each pair of \((a, b)\) with \(1 \leq a < b \leq k\) is described by
\[ \eta_{a,b} = P[Y = a|Y \in \{a,b\}] . \]  
The 0-1 loss is a kind of special misclassification loss which can be defined as  
\[ l^{0-1}(y,v) = \prod(yv < 0) + \frac{1}{2} \prod(v = 0) . \]  

A probability estimation loss \( \lambda \) is a measurable function  
\[ \lambda : \{1, \ldots, k\} \times \{1, \ldots, k\} \to \mathbb{R}_+ . \]  
A probability estimation loss proper is said minimized by predicting \( \eta \) if for any \( \eta, \eta' \in [0,1] \), we have  
\[ L_\eta(\eta, \eta) \leq L_\eta(\eta, \eta'). \]  

If the inequality is strict, then a loss function is called strict proper. A loss function \( l \) is called (strictly) proper composite if there exists an invertible link function \( \psi : [0,1] \to \mathbb{R} \) such that the probability estimation loss  
\[ \lambda(y,v) = l(y,\psi(v)) \]  
is (strictly) proper. For these (strictly) proper composite losses, we get that  
\[ L_\eta(\eta, \psi(\eta)) \leq L_\eta(\eta,v) \]  
for each \( \eta \in [0,1] \) and \( v \in \mathbb{R} \). If \( l \) is differentiable, then its inverse link can be expressed as  
\[ \psi^{-1}(v) = \sum_{a=0}^{k-1} \sum_{b=1}^{k} \frac{l_b(v)}{l_a(v)} (v-a) . \]  

It’s easy for us to check that the squared hinge loss, exponential loss, squared loss and logistic loss are all proper composite.

Any \( D \in \Delta_{X \times \{1, \ldots, k\}} \) may be specified exactly by the triplet  
\[ (P^{a,b}, Q^{a,b}, \pi^{a,b}) \]  
for each pair of \((a,b)\) with \( 1 \leq a \leq b \leq k \), where for every \( x \in X \)  
\[ (P^{a,b}(x), Q^{a,b}(x), \pi^{a,b}) = \left( P[X = x|y = a, y \in \{a,b\}], P[X = x|y = b, y \in \{a,b\}] \right), \]  
for each \( a \leq b \leq k \), where for every \( x \in X \)  
\[ (M^{a,b}(x), \eta^{a,b}(x)) = \left( P[X = x|y \in \{a,b\}], P[Y = a|X = x, Y \in \{a,b\}] \right) . \]

Here \( P^{a,b}, Q^{a,b} \) are the class conditional densities for each pair of \((a,b)\), and \( \pi^{a,b} \) denoted as the base rate or each pair of \((a,b)\). \( M^{a,b} \) and \( \eta^{a,b} \) are expressed as the observation density and class-conditional density, respectively. In what follows, we use \( P, Q, \pi, M, \eta \) to denote the corresponding objects on the whole multi-dividing domain, and the restriction on pair \((a,b)\) are \( P^{a,b}, Q^{a,b}, \pi^{a,b}, M^{a,b} \) and \( \eta^{a,b} \), respectively. For simplicity consideration, we use \( D_{P,Q,\pi} \) and \( D_{M,\eta} \) to denote the distribution on the whole domain, and its restricted on pair \((a,b)\) are denoted by \( D_{P^{a,b},Q^{a,b},\pi^{a,b}} \) and \( D_{M^{a,b},\eta^{a,b}} \) respectively. If we aim to refer to these densities, we should explicitly parameterise the distribution \( D \in \Delta_{X \times \{1, \ldots, k\}} \) as either \( D_{P,Q,\pi} \) or \( D_{M,\eta} \) as appropriate.

Given any distribution \( D \in \Delta_{X \times \{1, \ldots, k\}} \) and loss function \( l \), the \( l \)-classification risk for a scorer \( s \) is defined as  
\[ L^D_l(s) = E_{X,Y \sim D}[l(Y,s(X))] = E_{X=M}[L_l(\eta(X),s(X))] . \]  

If the infimum is reachable, then the set of Bayes-optimal \( l \)-scorers can comprise those who minimize the risk (see Menon and Williamson [26], Steinwart [27] and Reid and Williamson [28] for more detail):  
\[ S^D_{l^*} = \operatorname{Arg min}_{s \in X \to R} L^D_l(s) . \]  
Given any \( D_{P,Q,\pi} \in \Delta_{X \times \{1, \ldots, k\}} \) and loss function \( l \), the multi-dividing \( l \)-risk for a pair-scorer \( s_{pair} \) is defined by  
\[ L^D_{l^*}(s_{pair}) = \sum_{a=1}^{k-1} \sum_{b=1}^{k} \frac{l_a(s_{pair}(X,X'))}{l_a(s_{pair}(X',X))} \]  
(1)

If we achieve the infimum, then we can define the set of Bayes-optimal multi-dividing pair-scorers as  
\[ S^D_{l^*} = \operatorname{Arg min}_{s \in X \to R} L^D_l(s_{pair}) . \]  
and the set of Bayes-optimal multi-dividing univariate scorers is  
\[ S^{D,\text{Univ}}_{l^*} = \operatorname{Arg min}_{s \in X \to R} L^D_l(s_{pair}) . \]  
In multi-dividing setting, we aim to discover a scorer \( s : X \to \mathbb{R} \) so that \( L^D_{l^{*}}(s_{pair}) \) is (approximately) minimized. Equivalently, it is considered to minimise \( L^D_{l^{*}}(s_{pair}) \) over all \( s_{pair} \in S_{\text{Decomp}} \) in multi-dividing setting. It is verified that minimizing the risk \( L^D_{l^{*}}(s_{pair}) \) equals the area under the multi-dividing ROC curve (AUC) of the scorer \( s \) (the AUC criterion in multi-dividing setting can refer to Gao et. al. [29]):

\[ \operatorname{AUC}^D(s) = \frac{1}{2} \sum_{a=1}^{k-1} \sum_{b=1}^{k} \int_{s(x) > s(x')} P[Y = a|X = x, Y \in \{a,b\}] \]  

Minimising the multi-dividing risk with 0-1 loss function is equivalent to maximising the multi-dividing AUC. There are two tricks to be applied to a scorer \( s \) that approximately minimises \( L^D_{l^{*}}(s_{pair}) \). The first is the pointwise approach, and it minimises \( L^D_{l^{*}}(s) \) for certain kind of surrogate loss function. The Second is the pairwise approach, and it minimises \( L^D_{l^{*}}(s_{pair}) \) for certain surrogate loss function. An essential problem is that whether these approaches are consistent with the task of minimising \( L^D_{l^{*}}(s_{pair}) \) in multi-dividing setting or not. To solve this question, we need to construct the corresponding Bayes-optimal multi-dividing solutions \( S^D_{l^*} \) and \( S^{D,\text{Univ}}_{l^*} \) fall in the set \( S^D_{l^*,\text{Univ}} \). Thus, we aim to characterise \( S^D_{l^*,\text{Univ}} \) for which it will be helpful to

(Advance online publication: 23 February 2017)
build $S^D_{j,*}$. In the following contents, $D = D_{P,Q,x} = D_{M,q}$
$\in \Delta_{X \times \{1, \ldots, k\}}$.

III. MAIN RESULTS AND PROOFS

A. Pair-scors

In multi-dividing setting, we’d like to determine the Bayes-optimal univariate scorers $S^D_{k,l,Univ,*}$. We first determine the Bayes-optimal pair-scors, $S^D_{k,l,*}$ as preparation.

One challenge is to determine a suitable conditional risk by virtue of (1). For this purpose, we use an equivalence of the multi-dividing risk to a pairwise classification risk on a distribution Multi-dividing($D$) which is defined below.

For any $D_{P,Q,x} \in \Delta_{X \times \{1, \ldots, k\}}$, let Multi-dividing($D$) $\in \Delta_{X \times X \times \{1, \ldots, k\}}$ be defined via the triplet $(P_{pair}, Q_{pair}, \pi_{pair})$, where

$$(P_{pair}^{ab}, Q_{pair}^{ab}, \pi_{pair}^{ab}) = (P_{pair}, Q_{pair}, \pi_{pair})_{ab}$$

restricted on pair $(a, b)$ is given by

$$(P_{pair}^{ab}(x, x'), Q_{pair}^{ab}(x, x'), \pi_{pair}^{ab}(x, x')) = (P^{ab}(x)Q^{ab}(x'), P^{ab}(x')Q^{ab}(x), \frac{1}{2}).$$

The classification risk about Multi-dividing($D$) is equivalent to the multi-dividing risk with respect to $D$, as it is well known for loss function $l^{P^{0,1}}$.

Lemma 1. For any $D_{P,Q,x} \in \Delta_{X \times \{1, \ldots, k\}}$, loss function $l$ and

pair-scors $s_{pair}$: $X \times X \rightarrow \mathbb{R}$, we infer

$$I^D_{X,l}(s_{pair}) \leq I^l_{Multi-dividing(D)}(s_{pair}).$$

Proof. According to (1), we deduce

$$I^D_{X,l}(s_{pair}) = \sum_{a=1}^{k-1} \sum_{b=a+1}^{k} \mathbb{E}_{X \sim P_{X}^{ab}, X' \sim Q_{X}^{ab}} [l_a(s_{pair}(X, X')) + l_b(s_{pair}(X', X))].$$

Further, we have

$$I^l_{Multi-dividing(D)}(s_{pair}) = \sum_{a=1}^{k-1} \sum_{b=a+1}^{k} \mathbb{E}_{X \sim P_{X}^{ab}, X' \sim Q_{X}^{ab}} [l_a(s_{pair}(X, X')) + l_b(s_{pair}(X', X))].$$

Lemma 2. For any $D_{M,q} \in \Delta_{X \times \{1, \ldots, k\}}$, Multi-dividing($D$) has observation-conditional density given by

$$\eta_{pair} = \sigma \circ Diff(\sigma^{-1} \circ \eta).$$

Proof. Assume that we have a distribution $D_{P,Q,x} = D_{M,q} \in \Delta_{X \times \{1, \ldots, k\}}$. Let $(X, X', Z)$ be the random variable triplet such that, for any $x, x' \in X$ and $z \in \{1, \ldots, k\}$, we get

$$P[Z = z] \equiv \frac{1}{k}.$$

Furthermore, we assume that $X, X'$ are conditionally independent given $Z$. Thus, the above procedures can be summarized as a distribution Multi-dividing($D$) $\in \Delta_{X \times X \times \{1, \ldots, k\}}$, from which a sample $(x, x', z)$ may be drawn according to the following process:

- Draw $z \sim Ber(1/k)$
- Draw $x \sim \frac{1}{k} \sum_{a=1}^{k} \sum_{b=a+1}^{k} \{\Pi(z = a | z \in \{a, b\})P_{ab}^{ab}(x) + \Pi(z = b | z \in \{a, b\})Q_{ab}^{ab}(x)\}$
- Draw $x' \sim \frac{1}{k} \sum_{a=1}^{k} \sum_{b=a+1}^{k} \{\Pi(z = a | z \in \{a, b\})P_{ab}^{ab}(x') + \Pi(z = b | z \in \{a, b\})Q_{ab}^{ab}(x')\}$.

In terms of the above facts, we derive other marginals and conditionals as follows:

$$P[X = x, X' = x' | Z = z] = \frac{\Pi(z = a | z \in \{a, b\})P_{ab}^{ab}(x) + \Pi(z = b | z \in \{a, b\})Q_{ab}^{ab}(x)}{2}.$$

In the terms of above facts, we derive other marginals and conditionals as follows:

$$P[X = x, X' = x' | Z = z] = \sum_{a=1}^{k} \sum_{b=a+1}^{k} \{\Pi(z = a | z \in \{a, b\})P_{ab}^{ab}(x)Q_{ab}^{ab}(x') + \Pi(z = b | z \in \{a, b\})P_{ab}^{ab}(x')Q_{ab}^{ab}(x)\}.$$

- Draw $x \sim \frac{1}{k} \sum_{a=1}^{k} \sum_{b=a+1}^{k} \{\Pi(z = a | z \in \{a, b\})P_{ab}^{ab}(x)Q_{ab}^{ab}(x') + \Pi(z = b | z \in \{a, b\})P_{ab}^{ab}(x')Q_{ab}^{ab}(x)\}$.

Finally, we have

$$P[X = x, X' = x'] = \sum_{a=1}^{k} \sum_{b=a+1}^{k} \{\Pi(z = a | z \in \{a, b\})P_{ab}^{ab}(x)Q_{ab}^{ab}(x') + \Pi(z = b | z \in \{a, b\})Q_{ab}^{ab}(x')\}.$$
= σ(σ⁻¹(P[Z = a | X = x]) − σ⁻¹(P[Z = a | X' = x']))

= σ(σ⁻¹(P[Y = a | X = x]) − σ⁻¹(P[Y = a | X' = x']))

= σ((diff(σ⁻¹ ◦ ϕ))(x, x')).

The last two identities hold because of each pair of (a, b), we have

\[ σ⁻¹(ϕ^a_b(x)) = σ⁻¹(π^a_b(x)) + log \frac{P^{a,b}(x)}{Q^{a,b}(x)}. \]

Thus, using the conclusion of Lemma 2 and the fact that

\[ sign(2ϕ_{pair}(x, x') − 1) = sign(ϕ(x) − ϕ(x')) \]

we infer

\[ S_{k,l}^{D,*} = \{ s_{pair} : X \times X \to \mathbb{R} : \eta(x) ≠ ϕ(x') \to sign(s_{pair}(x, x')) = sign(ϕ(x) − ϕ(x')) \}. \]

Analogously, if loss function l is proper composite with the link function ψ, then we get

\[ \{ ψ ◦ ϕ_{pair} \} = \{ ψ ◦ σ ◦ Diff(σ⁻¹ ◦ ϕ) \} \subseteq S_{k,l}^{D,*}. \]

Here, \{ ψ ◦ σ ◦ Diff(σ⁻¹ ◦ ϕ) \} \subseteq S_{k,l}^{D,*} if and only if l is strictly proper composite. As with multi-classification, the optimal solution may be trivially transformed to reside in \( S_{k,l}^{D,*} \) for a proper composite loss.

B. Univariate Scorers

Searching the set of scoring functions that minimise \( L^l_{D}(\text{Diff}(s)) \) is equivalent to searching the set of pair-scorers \( s_{pair} \) (in \( S_{\text{Decomp}} \)) that minimise \( L^l_{D}(s_{pair}) \). In general, it is no longer possible to make a pointwise analysis by virtue of the conditional risk since \( S_{\text{Decomp}} \) is innocuous. If the optimal pair-scorder is decomposable, then the restricted function class can be ignored. It is not hard to check that

\[ S_{k,l}^{D,*} \cap S_{\text{Decomp}} \neq \emptyset \]

established for any \( D \in \Delta_{X \times \{1, \ldots, k\}} \) and loss function l. This property simplifies when all Bayes-optimal pair-scorder is decomposable, which is of interest when there is a unique optimal pair-scorder.

We can verify that for any \( D \in \Delta_{X \times \{1, \ldots, k\}} \) and loss function l,

\[ S_{k,l}^{D,*} \subseteq S_{\text{Decomp}} \iff S_{k,l}^{D,*} = \text{Diff}(S_{k,l}^{D,\text{Univ},*}) \].

This is to say, the decomposable Bayes-optimal multi-dividing pair-scorers are exactly the Bayes-optimal multi-dividing univariate scoring function passed through Diff. It implies, if \( S_{k,l}^{D,*} \cap S_{\text{Decomp}} \neq \emptyset \) is true for a loss function l, we automatically obtain the Bayes-optimal multi-dividing scoring function.

Fristly, we deal with the situation there is a decomposable Bayes-optimal multi-dividing pair-scorder, and thus the optimal scoring function can be easily computed. Since \{ Diff(η) \} \subseteq \( S_{k,l}^{D,*} \cap S_{\text{Decomp}} \), we deduce the following property of the optimal univariate scorers for l01.

Lemma 3. For any \( D_{M,1} \in \Delta_{X \times \{1, \ldots, k\}} \),

\[ s_{\text{regret}}^{k,l \rightarrow 0} = \{ s : X \to \mathbb{R} : \eta = ϕ \circ s \} \]

for some monotone increasing \( ϕ : [0,1] \to \mathbb{R} \).

One fact we emphasize here is that \( ϕ \) in Lemma 3 need not to be strictly monotone increasing means that for certain \( x ≠ x' \in X \), we may have \( ϕ(x) ≥ ϕ(x') \) but \( s(x) ≠ s(x') \). Nonetheless, a corollary is immediately obtained that any strictly monotone increasing transformation of \( η \) is necessarily an optimal multi-dividing univariate scoring function.

Lemma 4. Given any strictly monotone increasing \( ϕ : [0,1] \to \mathbb{R} \) and any \( D_{M,1} \in \Delta_{X \times \{1, \ldots, k\}} \), we have

\[ s_{\text{regret}}^{k,l \rightarrow 0} \subseteq S_{k,l}^{D,\text{Univ},*}. \]

By Lemma 4 and \{ ψ ◦ ϕ \} \subseteq \( S_{k,l}^{D,*} \), we find that \( S_{k,l}^{D,*} \subseteq S_{k,l}^{D,\text{Univ},*} \) for a strictly proper composite loss.

When l is a proper composite loss, the subset of proper composite loss functions for which there exists a decomposable pair-scorder is described.

Lemma 5. Given any strictly proper composite loss l with a differentiable, invertible link function ψ, then for any \( D \in \Delta_{X \times \{1, \ldots, k\}} \), we have

\[ S_{k,l}^{D,*} \subseteq S_{\text{Decomp}} \]

\[ \iff (\exists a \in \mathbb{R} \setminus \{0\})(\forall v \in V)ψ^{-1}(v) = \frac{1}{1+e^{av}}. \]

The above result characterizes the decomposability of Bayes-optimal multi-dividing pair-scorder. Furthermore, given any \( D_{M,1} \in \Delta_{X \times \{1, \ldots, k\}} \) and strictly proper composite loss l with inverse link function \( ψ^{-1}(v) = \frac{1}{1+e^{av}} \) for some \( a \in \mathbb{R} \setminus \{0\} \), we infer that

\[ S_{k,l}^{D,\text{Univ},*} = \{ ψ ◦ η + b : b \in \mathbb{R} \} \subseteq S_{k,l}^{D,\text{Univ},*}. \]

Also, surrogate regret bounds from multi-classification to relate the excess pairwise l-risk of a scoring function s: X \( \rightarrow \mathbb{R} \) can be transferred to the excess pairwise \( l^{0,1} \) risk. It reveals that certain pairwise surrogate risks minimizing is consistent with AUC maximization.

Lemma 6. Let \( \text{regret}^{D,\text{Univ},*}_{l^{-1,0}} = \inf_{x \times r_{l^{-1,0}}} L^l_{D}(\text{Diff}(s)) \) given any \( D_{M,1} \in \Delta_{X \times \{1, \ldots, k\}} \) and strictly proper composite loss l with inverse link function \( \psi^{-1}(v) = \frac{1}{1+e^{av}} \) for some \( a \in \mathbb{R} \setminus \{0\} \), and scoring function s: X \( \rightarrow \mathbb{R} \), we can find a convex function \( F : [0,1] \to \mathbb{R}_+ \) so that

\[ F(\text{regret}^{D,\text{Univ},*}_{l^{-1,0}}(s)) \leq \text{regret}^{D,\text{Univ},*}_{l^{-1,0}}. \]

C. Non-decomposable Case

In this subsection, we discuss the situation if the loss l does not have a decomposable Bayes-optimal multi-dividing pair-scorder. We can no longer resort to using the conditional risk in this case, but the risk minimiser can be directly computed by virtue of an appropriate derivative due to the simple structure of \( S_{\text{Decomp}} \). It infers that the Bayes-optimal multi-dividing scoring function is still a strictly monotone.
transform of $\eta$ under some assumptions of the loss, but the transform is distribution dependent rather than given link function $\psi$.

**Lemma 7.** For any $D_{P,Q,\sigma} = D_{M,J} \in \Delta_{X<1, \ldots, k}$ and a margin-based strictly proper composite loss $l(y, v) = \phi(yv)$ with convex $\phi : \mathbb{R} \to \mathbb{R}_+$, for $\forall \nu \in V$, set $f_D^\nu(v) = \sum_{a=1}^{k} \sum_{b=1}^{k} \pi^{ab} \mathbb{E}[l'(v - s^*(X))] \leq \frac{1}{(1 - \pi^{ab})} \pi \mathbb{E}[l'(v - s^*(X))] - \sum_{a=1}^{k} \sum_{b=1}^{k} \pi^{ab} \mathbb{E}[l'(v - s^*(X))]

If $D$ has finite support or $\phi'$ is bounded, we have

$$S_{D,1}^{(D_1, D_2)} = \{ s^* : X \to \mathbb{R} : \eta = f_D^\nu \circ s^* \}.$$  

**Proof.** For the given $D$, set $L(D)$ as the function space for Lebesgue-measurable multi-dividing scorers $s : X \to \mathbb{R}$ which satisfies:

$$L_{\Delta, 1}^{(D_1, D_2)}(s) = \sum_{a=1}^{k} \sum_{b=1}^{k} \mathbb{E}[l'(v - s^*(X))] < \infty.$$  

We verify that $L_{\Delta, 1}^{(D_1, D_2)} : L(D) \to \mathbb{R}$ is a function and its minimizer can be obtained after discussing the derivative of function. For arbitrary $s, t \in L(D)$ and $\varepsilon > 0$, let

$$F_{s,t}(\varepsilon) = L_{\Delta, 1}^{(D_1, D_2)}(s + \varepsilon t) = \sum_{a=1}^{k} \sum_{b=1}^{k} \mathbb{E}[l'(v - s^*(X)) + \varepsilon(t(X) - t(X') - l'(v - s^*(X')))]

Hence, the G-variation of $L_{\Delta, 1}^{(D_1, D_2)}(s)$ at point $s$ and direction of $t$ can be stated as

$$\delta L_{\Delta, 1}^{(D_1, D_2)}(s + t) = \lim_{\varepsilon \to 0} \frac{F_{s,t}(\varepsilon) - F_{s,t}(0)}{\varepsilon},$$

where $F_{s,t}(0)$ is existed. By means of non-negativity and convexity of $\phi$, we infer

$$\frac{\phi((\text{Diff}(s + \varepsilon t))(x,x')) - \phi((\text{Diff}(s))(x,x'))}{\varepsilon} \leq \phi((\text{Diff}(s + \varepsilon t))(x,x')) - \phi((\text{Diff}(s))(x,x')) \leq \phi((\text{Diff}(s))(x,x')) + \phi((\text{Diff}(s))(x,x'))$$

for any $\varepsilon \in (0,1)$ and $x, x' \in X$.

For any $x \in X$, let $r(x) = \sum_{a=0}^{k} \sum_{b=0}^{k} \mathbb{E}[l'(s^*(X)) - l'(s^*(X'))]$

Since both $L_{\Delta, 1}^{(D_1, D_2)}(s + t)$ and $L_{\Delta, 1}^{(D_1, D_2)}(s)$ are finite, and

$$\lim_{\varepsilon \to 0} \frac{\phi((x(X) - s^*(X')) + \varepsilon(t(X) - t(X')) - \phi(s(X) - s^*(X'))}{\varepsilon}\begin{cases} (t(X) - t(X'))\phi(s(X) - s^*(X')) \end{cases},$$

we get

$$F_{s,t}(0) = \sum_{a=0}^{k} \sum_{b=0}^{k} \mathbb{E}[l'(s^*(X)) - l'(s^*(X'))] = \sum_{a=0}^{k} \sum_{b=0}^{k} \mathbb{E}[l'(s^*(X)) - l'(s^*(X'))]$$

Clearly, we have

$$\sum_{a=0}^{k} \sum_{b=0}^{k} \mathbb{E}[l'(s^*(X)) - l'(s^*(X'))] < \infty,$$

We can assume that $\phi'$ is bounded if $X$ is infinite. Thus,

$$\sum_{a=0}^{k} \sum_{b=0}^{k} \mathbb{E}[l'(s^*(X)) - l'(s^*(X'))] < \infty.$$

Moreover, we can check that

$$E_{X \sim Q^b}[l'(X)] < \infty,$$

$$E_{X \sim Q^b}[l'(X)] < \infty.$$  

Let $s^* : X \to \mathbb{R}$ be the minimum of $L_{\Delta, 1}^{(D_1, D_2)}$. According to the convexity of $L_{\Delta, 1}^{(D_1, D_2)}$, for any $t \in L(D)$ we have

$$\int_X t(x)r(x)dx = 0.$$  

It is sufficient and necessary that $r = 0$ holds for almost everywhere. Hence, for $s^*$ is used to minimize the target risk, it is sufficient and necessary that for almost each $x_0 \in X$ and each pair of $(a,b)$,

$$P^{a,b}(x_0)E_{X \sim X^b}[\phi(s^*(x_0) - s^*(X))] = 0.$$

which reveals that for almost each $x_0 \in X$ and each pair of $(a,b)$,

$$\frac{\eta^{a,b}(x_0) - 1 - \pi^{a,b}}{1 - \eta^{a,b}(x_0) - \pi^{a,b}} = \frac{P^{a,b}(x_0)}{Q^{a,b}(x_0)} = E_{X \sim X^b}[\phi(s^*(x_0) - s^*(X))],$$

$$E_{X \sim X^b}[\phi(s^*(x_0) - s^*(X))] = E_{X \sim X^b}[I_{1}(s^*(x_0) - s^*(X')) - I_{s^*(x_0) - s^*(X)}] + I_{s^*(x_0) - s^*(X)}$$
\[
\begin{align*}
E_{X \sim \mathcal{D}}[l_a(s^*(x))-s^*(X)] & = E_{X \sim \mathcal{D}}[l_a(s^*(x)-s^*(X))] = E_{X \sim \mathcal{D}}[l_a(s^*(x))] - s^*(X) \\
E_{X \sim \mathcal{D}}[l_b(s^*(x))-s^*(X)] & = E_{X \sim \mathcal{D}}[l_b(s^*(x)-s^*(X))] = E_{X \sim \mathcal{D}}[l_b(s^*(x))] - s^*(X)
\end{align*}
\]

It further implies that \( \eta = f^{D \ast}_s \circ s^* \) where \( f^{D \ast}_s = \sum_{a=1}^{k} \sum_{b=1}^{k} \pi^{ab} E_{X \sim \mathcal{D}}[l_a(v-s^*(X))] - (1-\pi^{ab}) E_{X \sim \mathcal{D}}[l_b(v-s^*(X))]. \)

Therefore, we get the expected result.

In order to present any optimal multi-dividing scoring function \( s^* \) by virtue of \( \eta \), we have done for the previous scenarios, it still has to check the invertible of \( f^{D \ast}_s \). The following lemma offers sufficient conditions for this to hold.

**Lemma 8.** Suppose \( \phi \) is differentiable, strictly convex, and for any \( v \in V \), it satisfies
\[
\phi'(v) = 0 \iff \phi'(-v) \neq 0.
\]
Assume \( D_{Mg} \in \Delta_{X \times [1,...,k]} \) and \( l(y, v) = \phi(yv) \) is a margin-based strictly proper composite loss. Set \( f^{D \ast}_s \) is defined as in Lemma 7. If \( \phi' \) is bounded or \( D \) has finite support, then \( S_{D, Univ.} = \{ s : X \to \mathbb{R} : s = (f^{D \ast}_s)^{-1} \circ \eta \} \subseteq S_{D, Univ.} \).

**Proof.** We show that \( f^{D \ast}_s \) strictly monotone by virtue of constructing the strict monotonicity of
\[
g(v) = \sum_{a=1}^{k} \sum_{b=1}^{k} E_{X \sim \mathcal{D}}[l_a(v-s^*(X))] - (1-\pi^{ab}) E_{X \sim \mathcal{D}}[l_b(v-s^*(X))].
\]

The derivative of this function is
\[
g'(v) = \sum_{a=1}^{k} \sum_{b=1}^{k} \{(E_{X \sim \mathcal{D}}[l_a(v-s^*(X))] - l_a(v-s^*(X)))/E_{X \sim \mathcal{D}}[l_a(v-s^*(X))]\}^2.
\]

Using the convexity of \( l \), the terms \( l_a(v-s^*(X)) \) and \( l_b(v-s^*(X)) \) are both positive. In addition, by Proposition 15 in Vernet et al., \[30\], \( l_a \) and \( l_b \) are respectively increasing and decreasing, or vice versa. Hence, their derivatives cannot simultaneously be zero by assumption. Furthermore, the expected is always negative or positive for each \( v \), and thus \( g'(v) \) is always strictly negative or positive.

Therefore, \( g \) is strictly monotone, which implies \( f^{D \ast}_s \) is also monotone. In this way, we conclude \( s^* = (f^{D \ast}_s)^{-1} \circ \eta \).

**D. Bayes-optimal Scoring Function for the p-Norm Push Risk**

In this subsection, we discuss the Bayes-optimal solutions of the \( p \)-norm push risk. Next, let \( D_{Mg} \in \Delta_{X \times [1,...,k]} \). For arbitrary loss \( l \) and pair-scoring \( s_{pair} \), the \( (l, g) \)-push multi-dividing risk we defined as
\[
L^{D}_{push, l, g}(s_{pair}) = \sum_{a=1}^{k} \sum_{b=1}^{k} E_{X \sim \mathcal{D}}[g(E_{X \sim \mathcal{D}}[-l_a(s(X, X')) + l_b(s(X', X)))/2] \]

where \( g \) is a nonnegative, monotone increasing function. If \( g(x) = x \), we recover the standard multi-dividing risk to offer a detailed discussion of the selection \( g^0(x) = x^p \) for \( p \geq 1 \), with margin loss function \( l \) and decomposable pair-scoring, leading to the \( p \)-norm multi-dividing push risk:
\[
L^{D}_{push, l, g}(\text{Diff}(s)) = \sum_{a=1}^{k} \sum_{b=1}^{k} E_{X \sim \mathcal{D}}[l_a(s(X, X') - s(X', X))].
\]

For our discussion, let
\[
S^{D, s}_{push, l, g} = \arg \min_{s_{pair} : X \times X \to \mathbb{R}} L^{D}_{push, l, g}(s_{pair}),
\]
\[
S^{D, Univ.}_{push, l, g} = \arg \min_{X \times X \to \mathbb{R}} L^{D}_{push, l, g}(\text{Diff} \circ s).
\]

As with the standard multi-dividing risk, determining the Bayes-optimal scoring function for the \( (l, g) \) push is difficult due to the implicitly restricted function class \( S_{Decomp} \). In fact, this is challenging even for the pair-scoring situation: the \( (l, g) \) multi-dividing push risk is not so expressive by virtue of a conditional risk. Hence, we should compute the derivative of the risk, as in the proof of Lemma 7.

**Lemma 9.** For any \( D_{Mg} \in \Delta_{X \times [1,...,k]} \) and \( l(y, v) = \phi(yv) \) is a differentiable function \( g : X \to \mathbb{R} \), and a strictly proper composite loss \( l \) with link function \( \psi \). Suppose \( l_a, l_b \) are bounded or \( D \) has finite support. Let
\[
G^{D}_{pair}(x, x') = \log \frac{g'(F^{D}_{pair}(x))}{g'(F^{D}_{pair}(x'))}
\]
\[
F^{D}_{pair}(x) = \frac{\sum_{a=1}^{k} \sum_{b=1}^{k} E_{X \sim \mathcal{D}}[l_a(s_{pair}(X, x), x')] + l_b(s_{pair}(X, X'))}{2}
\]

We deduce
\[
S^{D, s}_{push, l, g} = \{ s_{pair} : X \times X \to \mathbb{R} : s_{pair} = \psi \circ \sigma \circ (\text{Diff}(\sigma^{-1} \circ \eta) - G^{D}_{pair}) \}.
\]

For the particular case when \( g: x \to x \), we get the standard multi-dividing risk, \( G^{D} = 0 \) and thus \( s_{pair} = \psi \circ \eta_{pair} \). For general \( (l, g) \) case, unfortunately, we didn’t know how to simplify the term \( G^{D} \), and thus have to settle for the above implicit equation. Interestingly, if loss function \( l \) is the exponential loss and \( g^0(x) = x^p \), the following simple characterization is yielded.

**Lemma 10.** Let \( D_{Mg} \in \Delta_{X \times [1,...,k]} \) and \( l^{exp}(y, v) = \psi \circ \eta_{pair} \) be the exponential loss and \( g^0(x) = x^p \) for some positive \( p \). Then, if \( D \) has finite support, we have
\[
S^{D, s}_{push, l^{exp}, g^p} = \{ \frac{1}{p+1} : \sigma^{-1} \circ \eta_{pair} \} = \{ \frac{1}{p+1} : \text{Diff}((\sigma^{-1} \circ \eta)) \}
\]

We now pay attention to the computation of \( S^{D, s}_{push, l, g} \). It is

(Advance online publication: 23 February 2017)
unsuccessful in computing the optimal multi-dividing pair-scoring for 0-1 loss function $I^{0-1}$. We use a different trick to construct the optimal univariate scoring functions.

Lemma 11. Suppose $\phi : [0, 1] \rightarrow \mathbb{R}$ is strictly monotone increasing. For any given $D_{M,j} \in \Delta_{X \times \{1, \ldots, k\}}$ and nonmonotone, monotone increasing $g$, we have $\phi \circ \eta \in S_{push, D_{M,j}}^{D_{M,j}^*, \phi \circ \eta}$.

It is easy to verify that $S_{push, D_{M,j}}^{D_{M,j}^*, \phi \circ \eta} \cap S_{push, D_{M,j}}^{D_{M,j}^*, \phi} \neq \emptyset$ and so the $(I^{0-1}, g)$-push keeps the optimal solutions for the standard multi-dividing risk.

For a general proper composite loss, it is difficult to appeal to the optimal pair-scoring implicitly obtained in Lemma 11. To our delight, the optimal pair-scoring immediately implies the form of the optimal univariate scoring function for the special case of exponential loss.

Lemma 12. For any $D_{M,j} \in \Delta_{X \times \{1, \ldots, k\}}$. Let $I^{exp}(y, v) = e^{-yv}$ be the exponential loss and $g(x) = x^p$ for any positive $p$. Then, if $D$ has finite support, we have

$$S_{push, I^{exp}, g^p}^{D_{M,j}^*, \phi \circ \eta} = \left\{ \frac{1}{p+1} \cdot (\sigma^{-1} \circ \eta) + b : b \in \mathbb{R} \right\}.

E. Several equivalent risks in Multi-dividing Setting

Now, we discuss the following techniques to obtain an optimal pair-scoring (here we assume that $I$ is a strictly proper composite loss):

- **Approach A**: minimize the classification risk $L^D_I$ (here, there are $k$ classes in total) with loss function $I$ and then deduce the pair-scoring;
- **Approach B**: minimize the multi-dividing risk $L^D_{I,k}$ with loss function $I$ over all decomposable multi-dividing pair-scoring;
- **Approach C**: minimize the multi-dividing risk $L^D_{I,k}$ with loss function $I$ over all multi-dividing pair-scoring;
- **Approach D**: minimize the $p$-norm push risk $L^D_{push, I^{exp}, g^p}$ over all decomposable multi-dividing pair-scoring.

It seems that the above presented versions are very different: Approach D is the special framework which differs away the conventional conditional risk model; Approach C is the unique method to utilize a multi-dividing pair-scoring during optimization; Approach A is really a classification approximation algorithm which is the only one to operate on single sample points not pairs. However, using the conclusion getting in former subsections, we see that all tricks above have the same optimal function which implies that the corresponding risks are equivalent.

Lemma 13. Let $D \in \Delta_{X \times \{1, \ldots, k\}}$, $I$ be a strictly proper composite loss function related on $k$ classes, and $\psi^{-1}(t) = (1 + e^{-at})^{-1}$ be an inverse link function with certain fixed $a \in \mathbb{R} - \{0\}$. Then, we yield

1. **Approach A**, **Approach B** and **Approach C** are equivalent;
2. If $p = a - 1$ for all $a > 1$ and the support of $D$ is finite, then **Approach D** is equivalent to **Approach A**, **Approach B** and **Approach C**.

To explain the Lemma 13, we show that the above mentioned approaches can obtain the same multi-dividing pair-scoring using exponential loss function:

- **Approach A**: $Diff\{\text{arg min } E_{X \rightarrow \mathbb{R}}[I^{exp}(X,Y)]\}$;
- **Approach B**: $Diff\{\text{arg min } \sum_{a=1}^{k} E_{X \rightarrow \mathbb{R}}[I^{exp}(X,Y)]\}$;
- **Approach C**: $Diff\{\text{arg min } \sum_{a=1}^{k} E_{X \rightarrow \mathbb{R}}[I^{exp}(X,Y)]\}$;
- **Approach D**: $Diff\{\text{arg min } \sum_{a=1}^{k} E_{X \rightarrow \mathbb{R}}[I^{exp}(X,Y)]\}$.

IV. CONCLUSIONS

In this paper, we present the Bayes-optimal scoring functions for multi-dividing setting under proper composite family of loss function such as 0-1 loss and exponential loss. The theorem obtained in our paper helps construct the consistency of minimization of multi-dividing risk. To the best of our knowledge, the result achieved in our paper is the first to state in multi-dividing setting and to illustrate the promising application prospects in information retrieval, and the biochemistry field.

ACKNOWLEDGMENT

First, we thank the reviewers for their constructive comments in improving the quality of this paper. This work was supported in part by the National Natural Science Foundation of China (61142007), the Open Research Fund by Jiangsu Key Laboratory of Recycling and Reuse Technology for Mechanical and Electronic Products (RMRME-KF1612) and the Natural Science Foundation of Jiangsu University of Technology in China (KYY14013).

REFERENCES


(Advance online publication: 23 February 2017)


