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.

The study of scoring function has attracted plenty of attention in recent years. Farhadinia [1] presented a series of scoring functions for hesitant fuzzy sets which offer a large amount of new methods for regress and ranking. In terms of a limitation of standard retrievability scoring function, Bashir, K. S. Khattak [2] proposed a normalized retrievability scoring function for information retrieval. Faraggi and Kloczkowsik [3] raised a knowledge-based scoring function to measure protein decoys in view of their similarity to the native structure. Kandel et. al., [4] presented a new scoring function which enhances classification of antibacterial activity. Zhou et. al., [5] developed a novel integrated machine learning scoring function (SVR_CAF) to discriminate native structures from decoys in the protein folding problem. Park et. al., [6] determined the new scoring function to find the novel protein tyrosine phosphatase sigma inhibitors. Huang and Zhang [7] proposed variable choosing procedures rely on penalized score functions which derived for linear measurement error models. Liu et. al., [8] considered the knowledge-based halogen bonding scoring function for the application of protein-ligand interaction predicting. In Yan and Wang [9], optimizing scoring function is used in protein-nucleic acid interactions. Zilian and Sotriffer [10] improved affinity Prediction of protein–ligand complexes using random forest-based scoring functions.

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

Manuscript received April 5, 2016. revised August 20, 2016. This work was supported in part by the National Natural Science Foundation of China(NO.61142007), the Open Research Fund by Jiangsu Key Laboratory of Recycling and Reuse Technology for Mechanical and Electronic Products (RRME-KF1612) and the general program of Jiangsu University of Technology in China (NO.KYY14013).

Linli Zhu is with the school of Computer Engineering, Jiangsu University of Technology, Changzhou, China.(email:zhulinli@jsut.edu.cn).

Yu Pan is with the school of Computer Engineering, Jiangsu University of Technology, Changzhou, China.

Mohammad Reza Farahani is with Department of Applied Mathematics, Iran University of Science and Technology,Narmak, Tehran 16844, Iran.

Wei Gao is with the school of Information and Technology, Yunnan Normal University, Kunming, China.

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 AUC 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-scorers in multi-dividing setting by using the technology of conditional risk; second, the univariate scoring 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 ρ. 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, Δ_S is denoted as the set of all distributions on S. Let $\theta\in[0, 1]$ be a parameter, we use $\mathrm{Ber}(\theta)$ 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 ,…, S_k) $\in X^{n_1}\times X^{n_2}\times\ldots\times X^{n_k}$ consisting of a sequence of training sample S_a=($x_1^a,\ldots,x_{n_a}^a$) ($1\le a\le k$). 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_b where $a<b$. We assume that instances in each S_a are drawn randomly and independently according to some (unknown) distribution Δ^a on the instance space X respectively.

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

We use the $\prod(\cdot)$ 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 sign(x) $=\prod(x\ge 0)-\prod(x\le 0)$ 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 is some 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_{Pair} for a product space $X\times X$ ($X^a\times X^b$ in multi-dividing setting for pair (a, b) with $1\le a<b\le k$) is certain function s_{Pair}: $X\times X\to V$. A pair-scorer s_{Pair} is called decomposable if

$$s_{\mathrm{Pair}}\in S_{\mathrm{Decomp}}=\{\mathrm{Diff}(s): s: X\to\mathbb{R}\}.$$

A loss function (in many references, it called cost function) l is some measurable function l: $\{1,\cdots,k\}\times\mathbb{R}\to\mathbb{R}_+$ 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 $1\le a<b\le k$. 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 $1\le a<b\le k$. We say it is a margin loss if $l(y, v)=\phi(yv)$ for some $\phi:\mathbb{R}\to\mathbb{R}$. The conditional l-risk then defined as

$$L_l(\eta,s)=E_{Y\sim\mathrm{Ber}(\eta)}(l(Y,s))$$
$$=\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\eta^{a,b}l_a(s)+(1-\eta^{a,b})l_b(s).$$

Here, η is the posterior distribution, and its restriction on each pair of (a, b) with $1\le a<b\le k$ is described by

$\eta^{a,b} = \mathrm{P}\{Y=a \mid Y \in \{a,b\}\}$.

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 λ is a measurable function λ: $\{1,\cdots,k\}\times\{1,\cdots,k\}] \to \mathbb{R}_+$. A probability estimation loss proper is said minimized by predicting η if for any $\eta,\eta' \in [0,1]$, we have

$$L_\lambda(\eta,\eta) \le L_\lambda(\eta,\eta').$$

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 ψ :[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 $L_l(\eta,\psi(\eta)) \le L_l(\eta,v)$ for each $\eta \in$ [0, 1] and $v \in \mathbb{R}$. If l is differentiable, then its inverse link can be expressed as

$$\psi^{-1}(v) = \sum_{a=1}^{k-1}\sum_{b=a+1}^{k} \frac{l_b'(v)}{l_b'(v) - l_a'(v)}.$$

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

Any $D \in \Delta_{X\times\{1,\cdots,k\}}$ may be specified exactly by the triplet $(P^{a,b}, Q^{a,b}, \pi^{a,b})$ for each pair of (a, b) with $1 \le a<b \le k$, where for every $x \in X$

$$\begin{aligned}&(P^{a,b}(x), Q^{a,b}(x), \pi^{a,b})\\ &= (\mathrm{P}[X=x \mid y=a, y\in\{a,b\}],\\ &\quad \mathrm{P}[X=x \mid y=b, y\in\{a,b\}], \mathrm{P}[y=a \mid y\in\{a,b\}])\end{aligned}$$

or alternately by the tuple $(M^{a,b}, \eta^{a,b})$ for each pair of (a, b) with $1 \le a<b \le k$, where for every $x \in X$

$$\begin{aligned}&(M^{a,b}(x), \eta^{a,b}(x))\\ &= (\mathrm{P}[X=x \mid Y\in\{a,b\}], \mathrm{P}[Y=a \mid X=x, Y\in\{a,b\}])\end{aligned}$$

.

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, π, M, η 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 $D^{a,b}_{P,Q,\pi}$ and $D^{a,b}_{M,\eta}$ (or, 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 $D \in \Delta_{X\times\{1,\cdots,k\}}$ as either $D_{P,Q,\pi}$ or $D_{M,\eta}$ as appropriate.

Given any distribution $D \in \Delta_{X\times\{1,\cdots,k\}}$ and loss function l, the l-classification risk for a scorer s is defined as

$$L_l^D(s) = \mathrm{E}_{X,Y\sim D}[l(Y,s(X))] = \mathrm{E}_{X\sim M}[L_l(\eta(X),s(X))].$$

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):

$$S_l^{D,*} = \operatorname*{Arg\,min}_{s:X\to\mathbb{R}} L_l^D(s).$$

Given any $D_{P,Q,\pi} \in \Delta_{X\times\{1,\cdots,k\}}$ and loss function l, the multi-dividing l- risk for a pair-scorer s_{Pair} is defined by

$$L_{k,l}^D(s_{\mathrm{Pair}}) \tag{1}$$

$$= \sum_{a=1}^{k-1}\sum_{b=a+1}^{k} \mathrm{E}_{X\sim P^{a,b},X'\sim Q^{a,b}}\left[\frac{l_a(s_{\mathrm{Pair}}(X,X')) + l_b(s_{\mathrm{Pair}}(X',X))}{2}\right].$$

If we achieve the infimum, then we can define the set of Bayes-optimal multi-dividing pair-scorers as

$$S_{k,l}^{D,*} = \operatorname*{Arg\,min}_{s_{\mathrm{Pair}}:X\times X\to\mathbb{R}} L_{k,l}^D(s_{\mathrm{Pair}}),$$

and the set of Bayes-optimal multi-dividing univariate scorers is

$$S_{k,l}^{D,\mathrm{Univ},*} = \operatorname*{Arg\,min}_{S:X\to\mathbb{R}} L_{k,l}^D(\mathrm{Diff}(s)).$$

In multi-dividing setting, we aim to discover a scorer s: $X \to \mathbb{R}$ so that $L_{k,l^{0-1}}^D(\mathrm{Diff}(s))$ is (approximately) minimized. Equivalently, it is considered to minimise $L_{k,l^{0-1}}^D(s_{\mathrm{Pair}})$ over all $s_{\mathrm{Pair}} \in S_{\mathrm{Decomp}}$ in multi-dividing setting. It is verified that minimizing the risk $L_{k,l^{0-1}}^D(\mathrm{Diff}(s))$ 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]):

$$\begin{aligned}\mathrm{AUC}^D(s) = \sum_{a=1}^{k-1}\sum_{b=a+1}^{k} \mathrm{E}_{X\sim P^{a,b},X'\sim Q^{a,b}}[\prod(s(X)>s(X'))\\ + \frac{1}{2}\prod(s(X)=s(X'))].\end{aligned}$$

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 $L_{k,l^{0-1}}^D(\mathrm{Diff}(s))$. The first is the pointwise approach, and it minimises $L_l^D(s)$ for certain kind of surrogate loss function. The Second is the pairwise approach, and it minimises $L_{k,l}^D(\mathrm{Diff}(s))$ for certain surrogate loss function. An essential problem is that whether these approaches are consistent with the task of minimising $L_{k,l^{0-1}}^D(\mathrm{Diff}(s))$ in multi-dividing setting or not. To solve this question, we need to construct the corresponding Bayes-optimal multi-dividing solutions $S_k^{D,*}$ and $S_{k,l}^{D,\mathrm{Univ},*}$ fall in the set $S_{k,l^{0-1}}^{D,\mathrm{Univ},*}$. Thus, we aim to characterise $S_{k,l}^{D,\mathrm{Univ},*}$ for which it will be helpful to build $S_l^{D,*}$. In the following contents, D = $D_{P,Q,\pi}$ = $D_{M,\eta}$ $\in$ $\Delta_{X\times\{1,\cdots,k\}}$.

III. MAIN RESULTS AND PROOFS

A. Pair-scorers

In multi-dividing setting, we'd like to determine the Bayes-optimal univariate scorers $S_{k,l}^{D,\text{Univ},*}$. We first determine the Bayes-optimal pair-scorers, $S_{k,l}^{D,*}$ 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-diving(D) which is defined below.

For any $D_{P,Q,\pi}$ $\in$ $\Delta_{X\times\{1,\cdots,k\}}$, let Multi-diving($D$) $\in$ $\Delta_{X\times X\times\{1,\cdots,k\}}$ be defined via the triplet $(P_{\text{Pair}}, Q_{\text{Pair}}, \pi_{\text{Pair}})$, where $(P_{\text{Pair}}^{a,b}, Q_{\text{Pair}}^{a,b}, \pi_{\text{Pair}}^{a,b})$ = $(P_{\text{Pair}}, Q_{\text{Pair}}, \pi_{\text{Pair}})|_{(a,b)}$ restricted on pair (a, b) is given by

$$(P_{\text{Pair}}^{a,b}(x,x'), Q_{\text{Pair}}^{a,b}(x,x'), \pi_{\text{Pair}}^{a,b})$$
$$=(P^{a,b}(x)Q^{a,b}(x'), P^{a,b}(x')Q^{a,b}(x), \frac{1}{2}).$$

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

Lemma 1. For any $D_{P,Q,\pi}$ $\in$ $\Delta_{X\times\{1,\cdots,k\}}$, loss function l and pair-scorer $s_{\text{Pair}}: X\times X \to \mathbb{R}$, we infer

$$L_{k,l}^D(s_{\text{Pair}}) = L_l^{\text{Multi-dividing}(D)}(s_{\text{Pair}}).$$

Proof. According to (1), we deduce

$$L_{k,l}^D(s_{\text{Pair}})$$
$$=\sum_{a=1}^{k-1}\sum_{b=a+1}^{k} \mathrm{E}_{X\sim P^{a,b},X'\sim Q^{a,b}}[\frac{l_a(s_{\text{Pair}}(X,X'))+l_b(s_{\text{Pair}}(X',X))}{2}].$$
$$=\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\{\frac{1}{2}\mathrm{E}_{X\sim P^{a,b},\,X'\sim Q^{a,b}}[l_a(s_{\text{Pair}}(X,X'))]$$
$$+\frac{1}{2}\mathrm{E}_{X\sim P^{a,b},\,X'\sim Q^{a,b}}[l_b(s_{\text{Pair}}(X',X))]\}$$
$$=\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\{\frac{1}{2}\mathrm{E}_{X\sim P^{a,b},\,X'\sim Q^{a,b}}[l_a(s_{\text{Pair}}(X,X'))]$$
$$+\frac{1}{2}\mathrm{E}_{X\sim P^{a,b},\,X'\sim Q^{a,b}}[l_b(s_{\text{Pair}}(X,X'))]\}$$
$$=\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\{\frac{1}{2}\mathrm{E}_{(X,X')\sim(P^{a,b},Q^{a,b})}[l_a(s_{\text{Pair}}(X,X'))]$$
$$+\frac{1}{2}\mathrm{E}_{(X,X')\sim(P^{a,b},Q^{a,b})}[l_b(s_{\text{Pair}}(X,X'))]\}.$$

In terms of the definition of Multi-diving(D), this is exactly $L_{k,l}^D(s_{\text{Pair}})$. Hence, this conclusion is well-known for the situation of loss function l^{0-1}.

Our first result reveals that $S_{k,l}^{\text{Multi-dividing}(D)} = S_{k,l}^{D,*}$. We now obtain the following elementary character of the observation-conditional density η_{Pair} of Multi-diving(D).

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

$$\eta_{\text{Pair}} = \sigma\circ\text{Diff}(\sigma^{-1}\circ\eta). \quad (2)$$

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

$$\mathrm{P}[\mathrm{Z}=\mathrm{z}]=\frac{1}{k},$$
$$\mathrm{P}[X=x\,|\,Z=z]=\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\{\prod(z=a\,|\,z\in\{a,b\})P^{a,b}(x)$$
$$+\prod(z=b\,|\,z\in\{a,b\})Q^{a,b}(x)\},$$
$$\mathrm{P}[X'=x'\,|\,Z=z]=\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\{\prod(z=a\,|\,z\in\{a,b\})P^{a,b}(x')$$
$$+\prod(z=b\,|\,z\in\{a,b\})Q^{a,b}(x')\}.$$

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

- Draw $z\sim$ Ber(1/k)
- Draw $x\sim\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\{\prod(z=a\,|\,z\in\{a,b\})P^{a,b}(x)$ $+\prod(z=b\,|\,z\in\{a,b\})Q^{a,b}(x)\}$
- Draw $x'\sim\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\{\prod(z=a\,|\,z\in\{a,b\})P^{a,b}(x')$ $+\prod(z=b\,|\,z\in\{a,b\})Q^{a,b}(x')\}$.

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

$$\mathrm{P}[X=x,X'=x'\,|\,Z=z]$$
$$=\mathrm{P}[X=x\,|\,Z=z]\cdot\mathrm{P}[X'=x'\,|\,Z=z]$$
$$=\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\{\prod(z=a\,|\,z\in\{a,b\})P^{a,b}(x)Q^{a,b}(x')$$
$$+\prod(z=b\,|\,z\in\{a,b\})P^{a,b}(x')^{a,b}Q(x)\},$$
$$\mathrm{P}[X=x,X'=x']$$
$$=\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\{\frac{P^{a,b}(x)Q^{a,b}(x')+P^{a,b}(x')Q^{a,b}(x)}{2}\}$$
$$=\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\{\frac{1}{2\pi^{a,b}(1-\pi^{a,b})}M^{a,b}(x)M^{a,b}(x')$$
$$\cdot(\eta^{a,b}(x)(1-\eta^{a,b}(x'))+\eta^{a,b}(x')(1-\eta^{a,b}(x)))\}$$
$$\mathrm{P}[z=a\,|\,X=x,X'=x']$$
$$=\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\frac{P^{a,b}(x)Q^{a,b}(x')}{P^{a,b}(x)Q^{a,b}(x')+P^{a,b}(x')Q^{a,b}(x)}$$
$$=\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\frac{1}{1+\dfrac{Q^{a,b}(x)}{P^{a,b}(x)}\cdot\dfrac{P^{a,b}(x')}{Q^{a,b}(x')}}$$

$$= \sigma(\sigma^{-1}(\mathrm{P}[Z = a \mid X = x]) - \sigma^{-1}(\mathrm{P}[Z = a \mid X' = x']))$$

$$= \sigma(\sigma^{-1}(\mathrm{P}[Y = a \mid X = x]) - \sigma^{-1}(\mathrm{P}[Y = a \mid X' = x']))$$

$$= \sigma((\mathrm{Diff}(\sigma^{-1} \circ \eta))(x, x')) .$$

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

$$\sigma^{-1}(\eta^{a,b}(x)) = \sigma^{-1}(\pi^{a,b}) + \log \frac{P^{a,b}(x)}{Q^{a,b}(x)} . \quad \square$$

Thus, using the conclusion of Lemma 2 and the fact that $\mathrm{sign}(2\eta_{\mathrm{Pair}}(x,x')-1) = \mathrm{sign}(\eta(x)-\eta(x'))$, we infer

$$S^{D,*}_{k,l^{0-1}} = \{s_{\mathrm{Pair}} : X \times X \to \mathbb{R} : \eta(x) \neq \eta(x')$$
$$\to \mathrm{sign}(s_{\mathrm{Pair}}(x,x')) = \mathrm{sign}(\eta(x)-\eta(x'))\} .$$

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

$$\{\psi \circ \eta_{\mathrm{Pair}}\} = \{\psi \circ \sigma \circ \mathrm{Diff}(\sigma^{-1} \circ \eta)\} \subseteq S^{D,*}_{k,l} .$$

Here, $\{\psi \circ \sigma \circ \mathrm{Diff}(\sigma^{-1} \circ \eta)\} = S^{D,*}_{k,l}$ if and only if l is strictly proper composite. As with multi classification, the optimal solution may be trivially transformed to reside in $S^{D,*}_{k,l}$ for a proper composite loss.

B. Univariate Scorers

Searching the set of scoring functions that minimise $L^{D}_{k,l}(\mathrm{Diff}(s))$ is equivalent to searching the set of pair-scorers s_{Pair} (in S_{Decomp}) that minimise $L^{D}_{k,l}(s_{\mathrm{Pair}})$. In general, it is no longer possible to make a pointwise analysis by virtue of the conditional risk since S_{Decomp} is innocuous. If the optimal pair-scorer is decomposable, then the restricted function class can be ignored.

It is not hard to check that

$$S^{D,*}_{k,l} \cap S_{\mathrm{Decomp}} \neq \varnothing$$
$$\leftrightarrow S^{D,*}_{k,l} \cap S_{\mathrm{Decomp}} = \mathrm{Diff}\,(S^{D,\mathrm{Univ},*}_{k,l})$$

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

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

$$S^{D,*}_{k,l} \subseteq S_{\mathrm{Decomp}} \leftrightarrow S^{D,*}_{k,l} = \mathrm{Diff}\,(S^{D,\mathrm{Univ},*}_{k,l}) .$$

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 $S^{D,*}_{k,l} \cap S_{\mathrm{Decomp}} \neq \varnothing$ 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-scorer, and thus the optimal scoring function can be easily computed. Since $\{\mathrm{Diff}\,(\eta)\} \subseteq S^{D,*}_{k,l^{0-1}} \cap S_{\mathrm{Decomp}}$, we deduce the following property of the optimal univariate scorers for $l01$.

Lemma 3. For any $D_{M,\eta} \in \Delta_{X \times \{1,\cdots,k\}}$,

$$S^{D,*}_{k,l^{0-1}} = \{s : X \to \mathbb{R} : \eta = \phi \circ s\}$$

for some monotone increasing $\phi : [0,1] \to \mathbb{R}$.

One fact we emphasize here is that ϕ in Lemma 3 need not to be strictly monotone increasing means that for certain $x \neq x' \in X$, we may have $\eta(x) = \eta(x')$ but $s(x) \neq s(x')$. Nonetheless, a corollary is immediately obtained that any strictly monotone increasing transform 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 $D_{M,\eta} \in \Delta_{X \times \{1,\cdots,k\}}$, we have

$$\phi \circ \eta \in S^{D,\mathrm{Univ},*}_{k,l^{0-1}} .$$

By Lemma 4 and $\{\psi \circ \eta\} \subseteq S^{D,*}_{l}$, we find that $S^{D,*}_{l} \subseteq S^{D,\mathrm{Univ},*}_{k,l^{0-1}}$ 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-scorer is described.

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

$$S^{D,*}_{k,l} \subseteq S_{\mathrm{Decomp}}$$
$$\leftrightarrow (\exists a \in \mathbb{R} \setminus \{0\})(\forall v \in V)\psi^{-1}(v) = \frac{1}{1+e^{-av}} .$$

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

$$S^{D,\mathrm{Univ},*}_{k,l} = \{\psi \circ \eta + b : b \in \mathbb{R}\} \subseteq S^{D,\mathrm{Univ},*}_{k,l^{0-1}} .$$

Also, surrogate regret bounds from multi-classification to relate the excess pairwise l-risk of a scoring function s: $X \to \mathbb{R}$ 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 $\mathrm{regret}^{D,\mathrm{Univ},*}_{k,l^{0-1}} = L^{D}_{k,l}(\mathrm{Diff}(s)) - \inf_{t:X \to \mathbb{R}} L^{D}_{k,l}(\mathrm{Diff}(t))$. Given any $D_{M,\eta} \in \Delta_{X \times \{1,\cdots,k\}}$ 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: $X \to \mathbb{R}$, we can find a convex function F_l :[0,1] $\to \mathbb{R}_{+}$ so that

$$F_l(\mathrm{regret}^{D,\mathrm{Univ},*}_{k,l^{0-1}}(s)) \leq \mathrm{regret}^{D,\mathrm{Univ},*}_{k,l} .$$

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-scorer. 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_{Decomp}. It infers that the Bayes-optimal multi-dividing scoring function is still a strictly monotone transform of η under some assumptions of the loss, but the transform is distribution dependent rather than given link function ψ.

Lemma 7. For any $D_{P,Q,\pi} = D_{M,\eta} \in \Delta_{X\times\{1,\cdots,k\}}$ and a margin-based strictly proper composite loss $l(y, v) = \phi(yv)$ with convex $\phi:\mathbb{R}\to\mathbb{R}_+$. For $\forall v\in V$, set $f_{s^*}^D(v)=$

$$\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\frac{\pi^{a,b}E_{X\sim P^{a,b}}[l_b^{'}(v-s^*(X))]}{\pi^{a,b}E_{X\sim P^{a,b}}[l_b^{'}(v-s^*(X))]-(1-\pi^{a,b})\pi E_{X'\sim Q^{a,b}}[l_a^{'}(v-s^*(X'))]}.$$

If D has finite support or ϕ' is bounded, we have

$$S_{k,l}^{D,\text{Univ},*}=\{s^*:X\to\mathbb{R}:\eta=f_{s^*}^D\circ s^*\}.$$

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:

$$L_{k,l}^{D,\text{Univ}}(s)=\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\mathrm{E}_{X'\sim Q^{a,b},X\sim P^{a,b}}[\phi(s(X)-s(X'))]<\infty.$$

We verify that $L_{k,l}^{D,\text{Univ}}:L(D)\to\mathbb{R}$ 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

$$F_{s,t}(\varepsilon)=L_{k,l}^{D,\text{Univ}}(s+\varepsilon t)=$$
$$\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\mathrm{E}_{X'\sim Q^{a,b},X\sim P^{a,b}}[\phi(s(X)-s(X')+\varepsilon(t(X)-t(X')))].$$

Hence, the G-variation of $L_{k,l}^{D,\text{Univ}}(s)$ at point s and direction of t can be stated as

$$\delta L_{k,l}^{D,\text{Univ}}(s+t)=\lim_{\varepsilon\to0}\frac{L_{k,l}^{D,\text{Univ}}(s+\varepsilon t)-L_{k,l}^{D,\text{Univ}}(s)}{\varepsilon}=F_{s,t}^{'}(0),$$

where $F_{s,t}^{'}(0)$ is existed. By means of non-negativity and convexity of ϕ, we infer

$$\left|\frac{\phi((\mathrm{Diff}(s+\varepsilon t))(x,x'))-\phi((\mathrm{Diff}(s))(x,x'))}{\varepsilon}\right|$$
$$\le\left|\phi((\mathrm{Diff}(s+\varepsilon t))(x,x'))-\phi((\mathrm{Diff}(s))(x,x'))\right|$$
$$\le\phi((\mathrm{Diff}(s+\varepsilon t))(x,x'))+\phi((\mathrm{Diff}(s))(x,x'))$$

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

For any $x\in X$, let $r(x)=$

$$\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\{P^{a,b}(x)\mathrm{E}_{X'\sim Q^{a,b}}[\phi'(s(X)-s(X'))]$$
$$-Q^{a,b}\mathrm{E}_{X\sim P^{a,b}}[\phi'(s(X)-s(X'))]\}.$$

Since both $L_{k,l}^{D,\text{Univ}}(s+\varepsilon t)$ and $L_{k,l}^{D,\text{Univ}}(s)$ are finite, and

$$\lim_{\varepsilon\to0}\frac{\phi(s(x)-s(x')+\varepsilon(t(x)-t(x')))-\phi(s(x)-s(x'))}{\varepsilon}$$
$$=(t(x)-t(x'))\phi'(s(x)-s(x')),$$

we get

$$F_{s,t}^{'}(0)=$$
$$\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\mathrm{E}_{X'\sim Q^{a,b},X\sim P^{a,b}}[(t(X)-t(X'))\phi'(s(X)-s(X'))]$$
$$=\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\mathrm{E}_{X'\sim Q^{a,b},X\sim P^{a,b}}[t(X)\phi'(s(X)-s(X'))]$$
$$-\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\mathrm{E}_{X'\sim P^{a,b},X\sim Q^{a,b}}[t(X')\phi'(s(X')-s(X'))]$$
$$=\int_X t(x)r(x)dx.$$

Clearly, we have

$$\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\mathrm{E}_{X'\sim Q^{a,b},X\sim P^{a,b}}[t(X)\phi'(s(X)-s(X'))]<\infty,$$
$$\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\mathrm{E}_{X'\sim P^{a,b},X\sim Q^{a,b}}[t(X')\phi'(s(X')-s(X'))]<\infty.$$

We can assume that ϕ' is bounded if X is infinite. Thus,

$$\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\mathrm{E}_{X'\sim Q^{a,b},X\sim P^{a,b}}[t(X)\phi'(s(X)-s(X'))]$$
$$<\sup_{z\in\mathbb{R}}|\phi'(z)|E_{X\sim P^{a,b}}[|t(X)|]$$

and

$$\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\mathrm{E}_{X'\sim P^{a,b},X\sim Q^{a,b}}[t(X')\phi'(s(X')-s(X'))]$$
$$<\sup_{z\in\mathbb{R}}|\phi'(z)|E_{X\sim Q^{a,b}}[|t(X)|].$$

Moreover, we can check that

$$E_{X\sim P^{a,b}}[|t(X)|]<\infty,$$
$$E_{X'\sim Q^{a,b}}[|t(X')|]<\infty.$$

Let $s^*:X\to\mathbb{R}$ be the minimum of $L_{k,l}^{D,\text{Univ}}$. According to the convexity of $L_{k,l}^{D,\text{Univ}}$, 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),

$$P^{a,b}(x_0)\mathrm{E}_{X'\sim Q^{a,b}}[\phi'(s^*(x_0)-s^*(X'))]$$
$$=Q^{a,b}(x_0)\mathrm{E}_{X\sim P^{a,b}}[\phi'(s^*(X)-s^*(x_0))]$$

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

$$\frac{\eta^{a,b}(x_0)}{1-\eta^{a,b}(x_0)}\frac{1-\pi^{a,b}}{\pi^{a,b}}=\frac{P^{a,b}(x_0)}{Q^{a,b}(x_0)}$$
$$=\frac{\mathrm{E}_{X\sim P^{a,b}}[\phi'(s^*(X)-s^*(x_0))]}{\mathrm{E}_{X'\sim Q^{a,b}}[\phi'(s^*(x_0)-s^*(X'))]}$$
$$=\frac{\mathrm{E}_{X\sim P^{a,b}}[l_a^{'}(s^*(X)-s^*(x_0))-l_b^{'}(s^*(x_0)-s^*(X))]}{\mathrm{E}_{X'\sim Q^{a,b}}[-l_a^{'}(s^*(x_0)-s^*(X'))+l_b^{'}(s^*(X)-s^*(x_0))]}$$

$$=\frac{\mathrm{E}_{X\sim P^{a,b}}[l_b'(s^*(x_0)-s^*(X))-l_a'(s^*(X)-s^*(x_0))]}{\mathrm{E}_{X'\sim Q^{a,b}}[l_a'(s^*(x_0)-s^*(X'))-l_b'(s^*(X)-s^*(x_0))]}$$

$$=\frac{\mathrm{E}_{X\sim P^{a,b}}[l_b'(s^*(x_0)-s^*(X))]}{\mathrm{E}_{X'\sim Q^{a,b}}[l_a'(s^*(x_0)-s^*(X'))]}.$$

It further implies that $\eta = f_{s^*}^D \circ s^*$ where

$$f_{s^*}^D = \sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\frac{\pi^{a,b}\mathrm{E}_{X\sim P^{a,b}}[l_b'(v-s^*(X))]}{\pi^{a,b}\mathrm{E}_{X\sim P^{a,b}}[l_b'(v-s^*(X))]-(1-\pi^{a,b})\mathrm{E}_{X\sim Q^{a,b}}[l_a'(v-s^*(X'))]}.$$

Therefore, we get the expected result.

In order to present any optimal multi-dividing scoring function s^* by virtue of η, as we have done for the previous scenarios, it still has to check the invertible of $f_{s^*}^D$. The following lemma offers sufficient conditions for this to hold.

Lemma 8. Suppose ϕ is differentiable, strictly convex, and for any $v\in V$, it satisfies

$$\phi'(v)=0 \leftrightarrow \phi'(-v)\neq 0.$$

Assume $D_{M,\eta} \in \Delta_{X\times\{1,\cdots,k\}}$ and $l(y, v) = \phi(yv)$ is a margin-based strictly proper composite loss. Set $f_{s^*}^D$ is defined as in Lemma 7. If ϕ' is bounded or D has finite support, then

$$S_{k,l}^{D,\mathrm{Univ},*}=\{s^*: X\to\mathbb{R}: s^*=(f_{s^*}^D)^{-1}\circ\eta\}\subseteq S_{k,l^{0-1}}^{D,\mathrm{Univ},*}.$$

Proof. We show that $f_{s^*}^D$ strictly monotone by virtue of constructing the strict monotonicity of

$$g(v)=\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\frac{E_{X'\sim Q^{a,b}}[l_b'(v-s^*(X'))]}{E_{X\sim P^{a,b}}[l_a'(v-s^*(X))]}.$$

The derivative of this function is

$$g'(v)=\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\{E_{X\sim P^{a,b},X'\sim Q^{a,b}}[l_a'(v-s^*(X'))l_b''(v-s^*(X'))$$
$$-l_a''(v-s^*(X'))l_b'(v-s^*(X'))]/(E_{X\sim P^{a,b}}[l_a'(v-s^*(X))])^2\}.$$

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 viceversa. 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 $f_{s^*}^D$ is also monotone. In this way, we conclude $s^*=(f_{s^*}^D)^{-1}\circ\eta$. □

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 $D_{M,\eta} \in \Delta_{X\times\{1,\cdots,k\}}$. For arbitrary loss l and pair-scorer s_{Pair}, the (l, g)-push multi-dividing risk we defined as

$$L_{\mathrm{push},l,g}^D(s_{\mathrm{Pair}})=$$
$$\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\mathrm{E}_{X'\sim Q^{a,b}}[g(\mathrm{E}_{X\sim P^{a,b}}[\frac{l_a(s(X,X'))+l_b(s(X',X))}{2}])],$$

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 $g^p(x)=x^p$ for $p\geq 1$, with margin loss function l and decomposable pair-scorer, leading to the p-norm multi-dividing push risk:

$$L_{\mathrm{push},l,g}^D(\mathrm{Diff}(s))$$
$$=\sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\mathrm{E}_{X'\sim Q^{a,b}}[(\mathrm{E}_{X\sim P^{a,b}}[l_a(s(X)-s(X'))])^p].$$

For our discussion, let

$$S_{\mathrm{push},l,g}^{D,*}=\operatorname*{Arg\,min}_{s_{\mathrm{Pair}}:X\times X\to\mathbb{R}} L_{\mathrm{push},l,g}^D(s_{\mathrm{Pair}}),$$

$$S_{\mathrm{push},l,g}^{D,\mathrm{Univ},*}=\operatorname*{Arg\,min}_{s:X\to\mathbb{R}} L_{\mathrm{push},l,g}^D(\mathrm{Diff}\circ s).$$

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-scorer situation: the (l, g) multi-dividing push risk is not so expressible 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 $D_{M,\eta} \in \Delta_{X\times\{1,\cdots,k\}}$, a differentiable function g: $X\to\mathbb{R}$, and a strictly proper composite loss l with link function ψ. Suppose l_a', l_b' are bounded or D has finite support. Let

$$G_{s_{\mathrm{Pair}}^*}^D(x,x')=\log\frac{g'(F_{s_{\mathrm{Pair}}}^D(x))}{g'(F_{s_{\mathrm{Pair}}}^D(x'))}$$

and

$$F_{s_{\mathrm{Pair}}}^D(x) = \sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\mathrm{E}_{X\sim P^{a,b}}[\frac{l_a(s_{\mathrm{Pair}}(X,x))+l_b(s_{\mathrm{Pair}}(x,X))}{2}].$$

We deduce

$$S_{\mathrm{push},l,g}^{D,*}=$$
$$\{s_{\mathrm{Pair}}^*: X\times X\to\mathbb{R}: s_{\mathrm{Pair}}^*=\psi\circ\sigma\circ(\mathrm{Diff}(\sigma^{-1}\circ\eta)-G_{s_{\mathrm{Pair}}^*}^D)\}.$$

For the particular case when g: $x\to x$, we get the standard multi-dividing risk, $G^D\equiv 0$ and thus $s_{\mathrm{Pair}}^*=\psi\circ\eta_{\mathrm{Pair}}$. 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 $g^p(x)=x^p$, the following simple characterization is yielded.

Lemma 10. Let $D_{M,\eta} \in \Delta_{X\times\{1,\cdots,k\}}$, $l^{\exp}(y,v)=\psi\circ\eta_{\mathrm{Pair}}$ be the exponential loss and $g^p(x)=x^p$ for some positive p. Then, if D has finite support, we have

$$S_{\mathrm{push},l^{\exp},g^p}^{D,*} = \{\frac{1}{p+1}\cdot\sigma^{-1}\circ\eta_{\mathrm{Pair}}\} = \{\frac{1}{p+1}\cdot\mathrm{Diff}(\sigma^{-1}\circ\eta)\}.$$

We now pay attention to the computation of $S_{\mathrm{push},l,g}^{D,*}$. It is unsuccessful in computing the optimal multi-dividing pair-scorer for 0-1 loss function l^{0-1}. We use a different trick to construct the optimal univariate scoring functions.

Lemma 11. Suppose ϕ: [0, 1] $\to \mathbb{R}$ is strictly monotone increasing. For any given $D_{M,\eta} \in \Delta_{X\times\{1,\cdots,k\}}$ and nonnegative, monotone increasing g, we have

$$\phi\circ\eta \in S^{D,*}_{\text{push},l^{0-1},g}.$$

It is easy to verify that $S^{D,\text{Univ},*}_{k,l^{0-1}} \cap S^{D,\text{Univ},*}_{\text{push},l^{0-1},g} \neq \varnothing$ and so the (l^{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-scorer implicitly obtained in Lemma 11. To our delight, the optimal pair-scorer immediately implies the form of the optimal univariate scoring function for the special case of exponential loss.

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

$$S^{D,*}_{\text{push},l^{\exp},g^p} = \{\frac{1}{p+1}\cdot(\sigma^{-1}\circ\eta)+b : b\in\mathbb{R}\}.$$

E. Several equivalent risks in Multi-dividing Setting

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

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

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-scorer 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 $D \in \Delta_{X\times\{1,\cdots,k\}}$, l 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-scorer using exponential loss function:

- Approach A: $\text{Diff}\{\arg\min_{s:X\to\mathbb{R}} \text{E}_{(X,Y)\sim D}[e^{-Ys(X)}]\}$;
- Approach B:

$$\text{Diff}\{\arg\min_{s:X\to\mathbb{R}} \sum_{a=1}^{k-1}\sum_{b=a+1}^{k} \text{E}_{X'\sim Q^{a,b},X\sim P^{a,b}}[e^{-(s(X)-s(X'))}]\};$$

- Approach C:

$$\arg\min_{s_{\text{Pair}}:X\times X\to\mathbb{R}} \sum_{a=1}^{k-1}\sum_{b=a+1}^{k} \text{E}_{X'\sim Q^{a,b},X\sim P^{a,b}}[e^{-s_{\text{Pair}}(X,X')}];$$

- Approach D:

$$\text{Diff}\{\arg\min_{s:X\to\mathbb{R}} \sum_{a=1}^{k-1}\sum_{b=a+1}^{k} \text{E}_{X'\sim Q^{a,b}}[(\text{E}_{X\sim P^{a,b}}[e^{-(s(X)-s(X'))}])^p]\}.$$

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 (RRME-KF1612) and the Natural Science Foundation of Jiangsu University of Technology in China (KYY14013).

REFERENCES

[1] B. Farhadinia, "A series of scoring functions for hesitant fuzzy sets", *Information Sciences*, vol. 277, pp. 102-110, 2014.
[2] S. Bashir and K. S. Khattak, "Producing efficient retrievability ranks of documents using normalized retrievability scoring function", *Journal of Intelligent Information Systems*, vol. 42, Iss. 3, pp 457-484, 2014.
[3] E. Faraggi and A.Kloczkowsik, "A global machine learning based scoring function for protein structure prediction", *Proteins: Structure, Function, and Bioinformatics*, vol. 82, Iss. 5, pp. 752–759, 2014.
[4] D. D. Kandel, C. Raychaudhury, and D. Pal, "Two new atom centered fragment descriptors and scoring function enhance classification of antibacterial activity", *J Mol Model*, vol. 20: 2164, DOI 10.1007/ s00894-014-2164-1, 2014.
[5] J. H. Zhou, W. Y. Yan, G. Hu, and B. R. Shen, "SVR_CAF: An integrated score function for detecting native protein structures among decoys", *Proteins: Structure, Function, and Bioinformatics*, vol. 82, iss. 4, pp.556–564, 2014.
[6] H. Park, H. S. Lee, B. Ku, and S. J. Kim, "Discovery of novel protein tyrosine phosphatase sigma inhibitors through the virtual screening with modified scoring function", *Medicinal Chemistry Research*, vol. 23, iss. 2, pp. 1016-1022, 2014.
[7] X. Z. Huang and H. M. Zhang, "Variable selection in linear measurement error models via penalized score functions", *Journal of Statistical Planning and Inference*, vol. 143, iss. 12, pp. 2101-2111, 2013.
[8] Y. T. Liu, Z. J. Xu, Z. Yang, K. X. Chen, and W. L. Zhu, "A knowledge-based halogen bonding scoring function for predicting

protein-ligand interactions", *Journal of Molecular Modeling*, vol. 19, iss. 11, pp 5015-5030, 2013.
[9] Z. Q. Yan and J. Wang, "Optimizing scoring function of protein-nucleic acid interactions with both affinity and specificity", *Plos one*, vol. 8, no. 9, DOI: 10.1371/journal.pone.0074443, 2013.
[10] D. Zilian and C. A. Sotriffer, "SFCscore(RF): a random forest-based scoring function for improved affinity prediction of protein-ligand complexes", *Journal of Chemical Information and Modeling*, vol. 53, no. 8, 1923-1933, 2013.
[11] M. H. Lan, J. Xu, and W. Gao, "Ontology similarity computation using *k*-partite ranking method", *Journal of Computer Applications*, vol. 32, no. 4, pp. 1094-1096, 2012.
[12] L. L. Zhu, W. G. Tao, X. Z. Min, and W.Gao, "Theoretical Characteristics of Ontology Learning Algorithm in Multi-dividing Setting", *IAENG International Journal of Computer Science*, vol. 43, no. 2, pp. 184-191, 2016.
[13] W. Gao and T. W. Xu, "Stability analysis of learning algorithms for ontology similarity computation", *Abstract and Applied Analysis*, Vol. 2013, 9 pages, 2013.
[14] W. Gao, Y. Gao, and Y. G. Zhang, "Strong and weak stability of *k*-partite ranking algorithm", *Information*, vol. 15, no, 11A, pp. 4585-4590, 2012.
[15] W. Gao and T. W. Xu, "Characteristics of optimal function for ontology similarity measure via multi-dividing", *Journal of networks*, vol.7, no. 6, pp. 1251-1259, 2012.
[16] W. Gao, T. Xu, J. Gan, and J. Zhou, "Linear statistical analysis of multi-dividing ontology algorithm", *Journal of Information and Computational Science*, vol. 11, no. 1, pp. 151-159, 2014.
[17] Y. Gao, W. Gao, and L. Liang, "Statistical Characteristics for Multi-dividing Ontology Algorithm in AUC Criterion Setting", *International Journal of Collaborative Intelligence*, vol. 1, no. 3, pp. 178-188, 2016.
[18] W. Gao, Y. Gao, Y. Zhang, and L. Liang, "Minimax learning rate for multi-dividing ontology algorithm", *Journal of Information and Computational Science*, vol. 11, no. 6, pp. 1853-1860, 2014.
[19] W. Gao, L. Yan, and L. Liang, "Piecewise function approximation and vertex partitioning schemes for multi-dividing ontology algorithm in AUC criterion setting (I)", *International Journal of Computer Applications in Technology*, vol. 50, nos. 3/4, pp. 226-231, 2014.
[20] L. Yan, W. Gao, and J. Li, "Piecewise function approximation and vertex partitioning schemes for multi-dividing ontology algorithm in AUC criterion setting (II)", *Journal of Applied Science*, vol. 13, no. 16, pp. 3257-3262, 2013.
[21] M. H. Lan, J. Xu, and W. Gao "Ontology Feature Extraction via Vector Learning Algorithm and Applied to Similarity Measuring and Ontology Mapping", *IAENG International Journal of Computer Science*, vol. 43, no. 1, pp. 10-19, 2016.
[22] W. Gao, L. Shi, and M. R.Farahani, "Distance-Based Indices for Some Families of Dendrimer Nanostars", *IAENG International Journal of Applied Mathematics*, vol. 46, no. 2, pp. 168-186, 2016.
[23] W.Gao and L. Shi, "Szeged related indices of unilateral polyomino chain and unilateral hexagonal chain", *IAENG International Journal of Applied Mathematics*, vol. 45, no. 2, pp. 138-150, 2015.
[24] W. Gao, L. L. Zhu, and Y. Guo, "Multi-dividing infinite push ontology algorithm", *Engineering Letters*, vol. 23, no. 3, pp. 132-139, 2015.
[25] X. Yu, J. Z. Wu, and W. Gao, "Fuse and Divide Technologies for Sparse Vector Learning in Ontology Algorithms", *Engineering Letters*, vol. 24, no.3, pp. 307-316, 2016.
[26] A. K. Menon and R. C. Williamson, "Bayes-optimal scorers for bipartite ranking", *JMLR: Workshop and Conference Proceedings* , vol. 35, pp.1–34, 2014.
[27] I. Steinwart, "How to compare different loss functions and their risks", *Constructive Approximation*, vol. 26, no. 2, pp. 225–287, 2007. doi: 10.1007/s00365-006-0662-3.
[28] M. D. Reid and R. C. Williamson, "Surrogate regret bounds for proper losses", *In International Conference on Machine Learning (ICML)*, pp. 897–904, New York, NY, USA, 2009. ACM. doi: 10.1145/1553374. 1553489.
[29] Y. Gao, W. Gao, and L. Liang, "A new *k*-partite ranking learning algorithm based on AUC metric and application in ontology", *Scientific Journal of Computer Science*, vol. 3, Iss. 5, pp. 136-144, 2013.
[30] E. Vernet, M. D. Reid, and R. C. Williamson. "Composite multiclass losses", *Advances in Neural Information Processing Systems (NIPS) 24*, pp. 1224–1232, 2011.

Linli Zhu, male, was born in the city of Badong, Hubei Province, China on Sep. 20, 1975. He got Master degrees on computer software and theory from Yunnan normal university in 2007. Now, he acts as associate professor in the department of computer engineering, Jiangsu University of Technology. As a researcher in computer science, his interests are covered two disciplines: computer network and artificial intelligence.

Yu Pan, male, was born in the city of Huang Shan, Anhui Province, China on March 31, 1963. He got Bachelor's Degree on Computer and Application from China University of Mining and Technology in 1985. He got Master's Degree on Mining Electrification and Automation from Beijing Graduate School of China University of Mining and Technology in 1990. Now, he acts as professor in the School of computer engineering, Jiangsu University of Technology. As a researcher in computer science, his interests are covered two disciplines: Computer Network and Information Security.

Mohammad Reza Farahani, male, was born in the city of Tehran, Iran on April . 5, 1988. He got his bachelor degree on Applied Mathematics from Iran University of Science and Technology (IUST) in 2010 as best department student. Now, as a researcher in applied mathematics, he work with faculties of this department, in applied mathematics, operation research, mathematics chemistry, Graph theory.

Wei Gao, male, was born in the city of Shaoxing, Zhejiang Province, China on Feb.13, 1981. He got two bachelor degrees on computer science from Zhejiang industrial university in 2004 and mathematics education from College of Zhejiang education in 2006. Then, he was enrolled in department of computer science and information technology, Yunnan normal university, and got Master degree there in 2009. In 2012, he got PhD degree in department of Mathematics, Soochow University, China. He acted as lecturer in the department of information, Yunnan Normal University from July 2012 to December 2015. Now, he acts as associate professor in the department of information, Yunnan Normal University. As a researcher in computer science and mathematics, his interests are covering two disciplines: Graph theory, Statistical learning theory, Information retrieval, and Artificial Intelligence.