TCS-TR-A-06-15

TCS Technical Report

Learning-Related Complexity of Linear Ranking Functions

by

Atsuyoshi Nakamura

Division of Computer Science

Report Series A

May 26, 2006

Hokkaido University
Graduate School of
Information Science and Technology

Email: atsu@ist.hokudai.ac.jp

Phone: +81-011-706-6806
Fax: +81-011-706-7832

Learning-Related Complexity of Linear Ranking Functions

May 26, 2006

Abstract

In this paper, we study learning-related complexity of *linear ranking functions* from n-dimensional Euclidean space to $\{1, 2, ..., k\}$. We show that their *graph dimension*, which is considered to indicate complexity of PAC learning of the class in the multiclass classification setting, is $\Theta(n + k)$. This graph dimension is significantly smaller than the graph dimension $\Omega(nk)$ of the class of $\{1, 2, ..., k\}$-valued decision-list functions naturally defined using $k - 1$ linear discrimination functions. We also show a risk bound of learning linear ranking functions in the ordinal regression setting, which is easily shown by the similar technique used in the proof of an upper bound of their graph dimension.

1 Introduction

A *linear ranking function* we study in this paper is a function from the n-dimensional Euclidean space $\Re^n$ to the set of ranks $\{1, 2, ..., k\}$ represented by $k - 1$ parallel hyperplains in $\Re^n$ which separate the domains of consecutive two ranks. This function is a simple one represented by $n+k-1$ real parameters, and the class of linear ranking functions is one of the most popular function classes studied in *ordinal regression.*

Ordinal regression is a kind of multiclass classification problems in which there is a linear ordering among the values of a class attribute. Problems of learning human preference [7, 8] are often formalized as this kind of problems.

Recently, some learning algorithms of linear ranking functions have been developed from the viewpoint of large margin principle [5, 11, 10]. However, few studies on learning-related complexity specific to those functions have been done yet. Only one[1] we know is analysis of *ranking loss* of online learning algorithm derived from perceptron algorithm studied by Crammer and Singer [3], where *ranking loss* of predicted rank $\hat{i}$ for true rank i is $|\hat{i} - i|$.

In this paper, we study learning-related complexity of the class of linear ranking functions. On this issue, Rajaram et al. [10] already proved that *VC dimension* of this

[1] In [11], a certain risk bound was shown using Vapnik's theorem (Theorem 6 in this paper), but there seems to be some problem in their way of application of the theorem. See Remark 2.

class is the same as that of the class of linear discrimination functions. But it seems not to be fair to compare a class of $\{-1, 1\}$-valued functions with a class of $\{1, 2, ..., k\}$-valued functions ($k > 2$) by VC dimension. For such comparison, *graph dimension* [9], which is an extension of VC dimension for $\{1, 2, ..., k\}$-valued functions, appears to be appropriate. In this paper, we show a difference between graph dimensions for the above two classes, concretely, the graph dimension of the class of linear ranking functions is $\Theta(n+k)$ while that of the class of linear discrimination functions is $\Theta(n)$.

For comparison's sake, we also analyze the graph dimension of the class of *decision-list* functions represented by $k-1$ linear discrimination functions $g_1, g_2, ..., g_{k-1}$ that determine the function value for $\mathbf{x}$ by the smallest index i that satisfies $g_i(\mathbf{x}) = 1$. The words that motivated us to study this comparison were the following simple question asked by our colleagues: "Why don't you learn a linear ranking function by decomposing it into $k-1$ learning problems of a linear discrimination function. The separator between rank i and $i+1$ can be learned using all the instances of more than rank i as positive and using others as negative, can't it?" Giving priority to smaller rank for the overlapped domains of different ranks, functions learned by this method can be seen as an above decision list. The answer to this question is that the complexity of the class of linear ranking functions is significantly smaller than that of the class of such decision lists, as a result, the number of instances necessary to learn for the former is smaller than that for the latter. Actually, we show that the graph dimension of the latter is $\Omega(nk)$ while the former is $\Theta(n+k)$ as described above. In the problems of predicting user's ratings for items, initial prediction performance for new users is an important issue, and learners with small sample complexity are preferable for quick acquisition of user's preference.

According to the theorem proved by Ben-David et al. [1], sample complexity of PAC learning of K-valued functions in the multiclass classification setting is linearly upper bounded by graph dimension. Therefore, we can obtain sample complexity upper bounds from our graph dimension upper bounds by applying their theorem.

Actually, their theorem can be implied by Vapnik's more general theorem which includes the settings of not only classification but also regression. The graph dimension of the class LR of linear ranking functions is equal to the VC dimension of the set of *indicators* of the function class defined by LR using the *zero-one loss function*. An upper bound of the VC dimension of the set of *indicators* of the function class defined by LR using the *L1 loss function* instead of the zero-one loss function can be shown by a similar argument used to prove an upper bound of graph dimension of LR. From the upper bound, we can also obtain a risk bound of learning LR in the ordinal regression setting.

This paper is organized as follows. In Section 2, two function classes we analyze in this paper, the class LR of linear ranking functions and the class $\mathrm{DL}_{\mathcal{L}}$ of decision lists with condition function class $\mathcal{L}$, are defined, where $\mathcal{L}$ is the class of linear discrimination functions. Analyses of the two function classes by VC dimension and graph dimension are described in Section 3. In Section 4, risk bounds on learning LR are shown in two settings, the multiclass classification setting and the ordinal regression setting. Section 5 concludes this paper.

2 Linear Ranking Functions

Let K denote the set $\{1, 2, ..., k\}$. In this paper, we consider functions from $\Re^n$ to K. A *linear ranking function* is a function $f : R^n \to K$ defined as

$$f(\mathbf{x}) = \min_{r \in K} \{r : \mathbf{w} \cdot \mathbf{x} - b_r < 0\},$$

where $\mathbf{w} \in \Re^n$, $\mathbf{b} = (b_1, b_2, ..., b_{k-1})$ with $b_1 \leq b_2 \leq ... \leq b_{k-1}$ and $b_k = \infty$. Let LR denote the class of linear ranking functions.

For comparison, we consider a larger function class. Let $\mathcal{B}$ denote a class of functions from $\Re^n$ to $\{-1, 1\}$. For arbitrary $g_1, g_2, ..., g_{k-1} \in \mathcal{B}$ and $g_k \equiv 1$, consider a kind of decision-list $\mathrm{dl}[g_1, g_2, ..., g_{k-1}] : R^n \to K$ defined as

$$\mathrm{dl}[g_1, g_2, ..., g_{k-1}](\mathbf{x}) = \min\{i : g_i(\mathbf{x}) = 1\}.$$

We call functions defined like this *decision lists with condition function class* $\mathcal{B}$. Let $\mathrm{DL}_{\mathcal{B}}$ denote the class of them. As condition function classes, here, we only consider the class $\mathcal{L}$ of *linear discrimination functions* f, namely, a function represented by

$$f(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{w} \cdot \mathbf{x} < b \\ -1 & \text{if } \mathbf{w} \cdot \mathbf{x} \geq b \end{cases}$$

using $\mathbf{w} \in \Re^n$ and $b \in \Re$. We call vector $\mathbf{w}$ a *normal vector* of f. Note that $\mathrm{DL}_{\mathcal{L}} \supseteq \mathrm{LR}$.

3 Complexity of LR

3.1 VC Dimension Analysis

Let $\mathcal{F}$ denote a class of functions from $\Re^n$ to K. Let l denote an arbitrary natural number. For $S = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_l) \in X^l$ and $f \in \mathcal{F}$, define $f_S \in K^l$ as $(f(\mathbf{x}_1), f(\mathbf{x}_2), ..., f(\mathbf{x}_l))$. Function set $\Pi_{\mathcal{F}}(S)$ is defined as the set of functions in $\mathcal{F}$ with the restricted domain S, namely, defined as follows:

$$\Pi_{\mathcal{F}}(S) = \{f_S : f \in \mathcal{F}\}.$$

Set $S \subseteq X^l$ is said to be *shattered* by $\mathcal{F}$ if $|\Pi_{\mathcal{F}}(S)| = k^l$. Furthermore, define $\Pi_{\mathcal{F}}(l)$ as follows:

$$\Pi_{\mathcal{F}}(l) = \max_{S \in X^l} |\Pi_{\mathcal{F}}(S)|,$$

where $|\cdot|$ is the number of elements in a set. *VC dimension* $d_V(\mathcal{F})$ of $\mathcal{F}$ is defined[2] as the maximum number of elements among the sets that are shattered by $\mathcal{F}$ [1, 10], namely, defined as follows:

$$d_V(\mathcal{F}) = \max\{l : |\Pi_{\mathcal{F}}(l)| = k^l\}.$$

Note that this definition of VC dimension coincides with the original definition by Vapnik and Chervonenkis [13] when $k = 2$.

The next proposition holds trivially.

Proposition 1

$$d_V(DL_{\mathcal{B}}) \leq d_V(\mathcal{B})$$

By Proposition 1 and $\mathrm{LR} \subseteq \mathrm{DL}_{\mathcal{L}}$,

$$d_V(\mathrm{LR}) \leq d_V(\mathcal{L}) \tag{1}$$

holds. In [10], it was shown that both sides in Inequality (1) are equal.

Theorem 1 (Rajaram et al. [10])

$$d_V(LR) = d_V(\mathcal{L})$$

Thus, $d_V(\mathrm{LR}) = d_V(\mathrm{DL}_{\mathcal{L}})$ holds, which means that no difference of the two function classes appears by VC dimension analysis.

3.2 Graph Dimension Analysis

As another extension of VC dimension for K-valued function classes, Natarajan considered *graph dimension* defined as follows.

Let function δ denote a $\{-1, 1\}$-valued function on $K \times K$ defined as

$$\delta(i, j) = \begin{cases} 1 & \text{if } i = j \\ -1 & \text{otherwise.} \end{cases}$$

Let l denote an arbitrary natural number. For $S = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_l) \in X^l$, $I = (i_1, i_2, ..., i_l) \in K^l$ and $f \in \mathcal{F}$, define $f_{I,S} \in \{-1, 1\}^l$ as

$$(\delta(f(\mathbf{x}_1), i_1), \delta(f(\mathbf{x}_2), i_2), ..., \delta(f(\mathbf{x}_l), i_l)).$$

Let $\Pi_{I,\mathcal{F}}(S)$ denote $\{f_{I,S} : f \in \mathcal{F}\}$. Then, define $\Pi^G_{\mathcal{F}}(l)$ as follows:

$$\Pi^G_{\mathcal{F}}(l) = \max_{S \in X^l, I \in K^l} |\Pi_{I,\mathcal{F}}(S)|.$$

Graph dimension $d_G(\mathcal{F})$ of K-valued function class $\mathcal{F}$ [9, 1] is defined as follows:

$$d_G(\mathcal{F}) = \max\{l : |\Pi^G_{\mathcal{F}}(l)| = 2^l\}.$$

From the definition of graph dimension, the graph dimension of K-valued function class $\mathcal{F}$ on X can be seen as the VC dimension of $\{-1, 1\}$-valued function class $\mathcal{F}_\delta$ on $X \times K$ defined as follows

$$\mathcal{F}_\delta = \{h_f : h_f(\mathbf{x}, i) \stackrel{def}{=} \delta(f(\mathbf{x}), i), f \in \mathcal{F}\}. \tag{2}$$

[2] $d_V(\mathcal{F})$ is called *ranking dimension* in [10].

3.3 Graph Dimension of $\mathrm{DL}_{\mathcal{L}}$

The following lemma is a slight modification of Lemma 3.2.3 [2] proved by Blumer et al. The proof of Lemma 1 is similar to that of their lemma and is omitted.

Lemma 1 *For $s \geq 1$, let $\mathcal{B}_1, \mathcal{B}_2, ..., \mathcal{B}_s$ be $\{-1, 1\}$-valued function classes on X with $d_V(\mathcal{B}_1) = d_V(\mathcal{B}_2) = \cdots = d_V(\mathcal{B}_s) = d < \infty$. Let $\mathcal{C} = \{\min_{i=1}^{s} f_i : f_i \in \mathcal{B}_i, 1 \leq i \leq s\}$. Then, $d_V(\mathcal{C}) < 2ds \log_2(3s)$.*

Theorem 2 $d_G(DL_{\mathcal{B}}) < d_V(\mathcal{B})k(k+1)\log_2(3k)$

(Proof) Let $I = (i_1, i_2, ..., i_l) \in K^l$ and $S = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_l)$. Let $l_s = |\{j : i_j = s\}|$ and define S_s as $(\mathbf{x}_{j_1}, \mathbf{x}_{j_2}, ..., \mathbf{x}_{j_{l_s}})$, where $j_1, j_2, ..., j_{l_s}$ are distinct elements in $\{j : i_j = s\}$. Let $I_s = (i_{j_1}, i_{j_2}, ..., i_{j_{l_s}})$. We show

$$|\Pi_{I,\mathrm{DL}_{\mathcal{B}}}(S)| = 2^l \Rightarrow l_s < 2d_V(\mathcal{B})s\log_2(3s) \tag{3}$$

for all $1 \leq s \leq k$. Let

$$\mathcal{H}_s = \{h_{f,s} : h_{f,s}(\mathbf{x}) \stackrel{def}{=} \delta(f(\mathbf{x}), s), f \in \mathrm{DL}_{\mathcal{B}}\}.$$

Note that $|\Pi_{I,\mathrm{DL}_{\mathcal{B}}}(S)| = 2^l$ implies $|\Pi_{\mathcal{H}_s}(S_s)| = 2^{l_s}$. Note that $\mathcal{H}_s$ can be represented as follows:

$$\mathcal{H}_s = \begin{cases} \{\min_{i=1}^{s} g_i : g_i \in \overline{\mathcal{B}} \text{ for } 1 \leq i \leq s-1, g_s \in \mathcal{B}\} & \text{for } 1 \leq s < k \\ \{\min_{i=1}^{s-1} g_i : g_i \in \overline{\mathcal{B}}\} & \text{for } s = k, \end{cases}$$

where $\overline{\mathcal{B}} = \{-f : f \in B\}$. By the fact $d_V(\overline{\mathcal{B}}) = d_V(\mathcal{B})$ and Lemma 1,

$$l_s \leq d_V(\mathcal{H}_s) < 2d_V(\mathcal{B})s\log_2(3s).$$

Thus,(3) holds for all $1 \leq s \leq k$. Therefore,

$$d_G(\mathrm{DL}_{\mathcal{B}}) < \sum_{s=1}^{k} 2d_V(\mathcal{B})s\log_2(3s) \leq k(k+1)d_V(\mathcal{B})\log_2(3k).$$

□

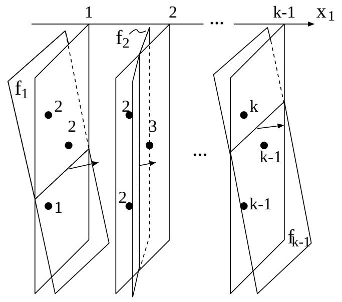
Theorem 3 $n(k-1) \leq d_G(DL_{\mathcal{L}}) < (n+1)k(k+1)\log_2(3k)$

(Proof) By Theorem 2 and the fact that $d_V(\mathcal{L}) = n+1$ [2], $d_G(\mathrm{DL}_{\mathcal{L}}) \leq (n+1)k(k+1)\log_2(3k)$ holds.

Now we prove inequality $(k-1)n \leq d_G(\mathrm{DL}_{\mathcal{L}})$. In $\Re^n = \{(x_1, x_2, ..., x_n) : x_i \in \Re\}$, consider subspace $X_0 = \{(0, x_2, ..., x_n) : x_i \in \Re\}$. Since X_0 is isomorphic to $\Re^{n-1}$, the VC dimension of the class of linear discrimination functions in X_0 is n. Let $S_0 \in X_0^n$ be a list of points shattered by the class of linear discrimination functions. Let S_i be a list of points made by moving S_0 in the direction of x_1-axis by i. Consider list S with length $(k-1)n$ made by concatenating $S_1, S_2, ..., S_{k-1}$ and

$$I = (\underbrace{1, 1, ..., 1}_{n \text{ times}}, \underbrace{2, 2, ..., 2}_{n \text{ times}}, ..., \underbrace{k-1, k-1, ..., k-1}_{n \text{ times}}).$$

We show that $\Pi_{I,\mathrm{DL}_{\mathcal{L}}}(S) = \{-1, 1\}^{(k-1)n}$. Let A be an arbitrary element in $\{-1, 1\}^{(k-1)n}$. In hyperplain $x_1 = i$, which contains S_i, there is a linear discrimination function g_i in $(n-1)$-dimensional Euclidean space such that $g_i(\mathbf{x}) = 1$ for $\mathbf{x} \in S_i$ if and only if the corresponding component in A is 1. Function g_i can be seen as a linear discrimination function $f_i : \Re^n \to \{-1, 1\}$ restricted in hyperplain $x_1 = i$. A normal vector $\mathbf{w}_i$ of such f_i can be as close to $(1, 0, 0, ..., 0)$ as you want, namely, it is possible that $f_i(\mathbf{x}) = -1$ $(\mathbf{w}_i \cdot \mathbf{x} \geq b_i)$ for all elements $\mathbf{x}$ in $S_{i+1}, S_{i+2}, ..., S_{k-1}$. (See the figure below.)



Then, for $f = \mathrm{dl}[f_1, f_2, ..., f_{k-1}]$, $f_{I,S} = A$ holds. □

3.4 Graph Dimension of LR

Theorem 4 $d_G(LR) \geq n + k - 1$

(Proof) Like the proof of Theorem 3, consider a list $S_0 = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ of n points in $\Re^n$ that are contained in hyperplain $x_1 = 1$ and shattered by linear discrimination functions. For $i = 1, 2, ..., k-1$, let $\mathbf{x}_{n+i} = \{(i+1, 0, 0, ..., 0)\}$. Let S be a list $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{n+k-1})$ of $n + k - 1$ points in $\Re^n$ and let

$$I = (\underbrace{1, 1, ..., 1}_{n \text{ times}}, 2, 3, ..., k).$$

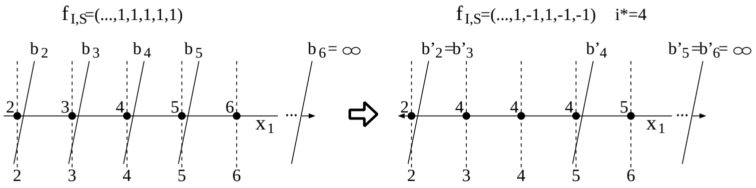
We show that $\Pi_{I,\mathrm{LR}}(S) = \{-1, 1\}^{n+k-1}$. Let $A = (a_1, a_2, ..., a_{n+k-1})$ be an arbitrary element in $\{-1, 1\}^{n+k-1}$. By similar reason argued in the proof of Theorem 3, there exist $\mathbf{w} \in \Re^n$ and $b_1 \leq b_2 \leq ... \leq b_{k-1}$ such that the linear ranking function $f(\mathbf{x}) = \min_{r \in K}\{r : \mathbf{w} \cdot \mathbf{x} - b_r < 0\}$ defined by these parameters $\mathbf{w}, b_1, b_2, ..., b_{k-1}$ satisfies that

$$f(\mathbf{x}_i) = \begin{cases} 1 & \text{if } i \leq n \text{ and } a_i = 1 \\ 2 & \text{if } i \leq n \text{ and } a_i = -1 \\ i - n + 1 & \text{if } n + 1 \leq i \leq n + k - 1. \end{cases}$$

Thus, when $(a_{n+1}, a_{n+2}, ..., a_{n+k-1}) = (1, 1, ..., 1)$, $f_{I,S} = A$ holds. In the case that $(a_{n+1}, a_{n+2}, ..., a_{n+k-1}) \neq (-1, -1, ..., -1)$, $f_{I,S} = A$ holds if thresholds $b_2, b_3, ..., b_{k-1}$ of f are changed to $b'_2, b'_3, ..., b'_{k-1}$ defined as follows:

$$b'_i = \begin{cases} b_i & \text{if } a_{n+i-1} = 1, i \neq i^* \\ b_{i+1} & \text{if } a_{n+i-1} = 1, i = i^* \\ b_{i-1} & \text{if } a_{n+i-1} = -1, i < i^* \\ b_{i+1} & \text{if } a_{n+i-1} = -1, i > i^*, \end{cases}$$

where $i^* = \max\{j : a_{n+j-1} = 1\}$. (See the figure below.)



When $(a_{n+1}, a_{n+2}, ..., a_{n+k-1}) = (-1, -1, ..., -1)$, by choosing a vector $\mathbf{w}$ close to $(-1, 0, 0, ..., 0)$ and using thresholds with $b_1 = b_2 = \cdots = b_{k-1}$, we get a linear ranking function f that satisfies

$$f(\mathbf{x}_i) = \begin{cases} k & \text{if } i \leq n \text{ and } a_i = -1 \\ 1 & \text{otherwise.} \end{cases}$$

Then, $f_{I,S} = A$ holds. $\square$

To prove $d_G(\mathrm{LR}) = O(n + k)$, we use notions and notations in the area of combinatorial geometry [4]. Let H be a finite set of hyperplains in $\Re^d$. The *arrangement* $\mathcal{A}(H)$ of H is a dissection of $\Re^d$ into connected pieces of various dimensions defined by H. Let H be composed of m hyperplains $h_1, h_2, ..., h_m$. Each $h_i \in H$ is represented as $g_i(\mathbf{x}) = 0$ using linear (or, more precisely, affine) function $g_i : \Re^d \to \Re$. For an arbitrary point p, the *face position vector*[3] $u(p)$ of p is defined as $(v_1(p), v_2(p), ..., v_m(p))$, where

$$v_i(p) = \begin{cases} +1 & \text{if } g_i(p) > 0 \\ 0 & \text{if } g_i(p) = 0 \\ -1 & \text{if } g_u(p) < 0. \end{cases}$$

The *binary position vector* of p is a binarized vector of $u(p)$ by replacing 0 with +1. A set of points with the same face position vector is called a *face*. A *k-flat* is defined as the affine hull of $k + 1$ affinely independent points. A face is called a *cell* if it is d-flat. An arrangement $\mathcal{A}(H)$ of $m \geq d$ hyperplains in $\Re^d$ is called *simple* if any d hyperplains of H have a unique point in common and if any $d + 1$ hyperplains have no point in common. When $m < d$, $\mathcal{A}(H)$ is called *simple* if the common intersection of the m hyperplains is a $(d - m)$-*flat*.

[3] In [4], this is simply called a *position vector*. We added word 'face' for distinguishing this from a *binary* position vector we introduced below.

Lemma 2 *Let H be a set of $m \geq d$ hyperplains in $\Re^d$ such that $\mathcal{A}(H)$ is simple. Then, the number of distinct binary position vectors is equal to the number of cells.*

(Proof) Assume that the face position vector $u(p)$ for point $p = (\pi_1, \pi_2, ..., \pi_d)$ has l zero-valued components and that the set of their indices is S. For $i \in S$, there exists a tiny value Δ such that the face position vector for point $p' = (\pi_1, \pi_2, ..., \pi_{i-1}, \pi_i + \Delta, \pi_{i+1}, ..., \pi_d)$ has $+1$ for its ith component and the same values for the other components by 'simple' assumption. By repeating this procedure, the number of zero-valued components can be reduced one by one and finally you can get a point q with $u(q)$ having no zero-valued component, namely, a point in a cell. The binary position vectors of p and q are trivially the same. $\square$

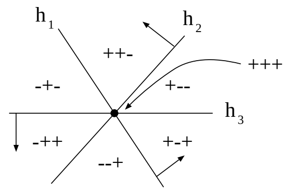


Figure 1: Case that the number of distinct binary vectors is larger than the number of cells. The directions of the arrows going out from hyperplains indicate positive directions.

Remark 1 *Note that Lemma 2 does not hold when $\mathcal{A}(H)$ is* not *simple. See Fig. 1.*

Lemma 3 (A part of Lemma 1.2 in [4]) *Let H be a set of m hyperplains in $\Re^d$ such that $\mathcal{A}(H)$ is simple. Then, the number of cells is*

$$\sum_{i=0}^{d} \binom{m}{i}.$$

Lemma 4 *For $a \geq 2e$, $ax \geq 2^x \Rightarrow x < 2\log_2 a$.*

(Proof)

$$\begin{aligned} 1 + \log_2 e &< e \\ 2\log_2 2e &< 2e \end{aligned}$$

Thus, for $a \geq 2e$,

$$\begin{aligned} 2\log_2 a &< a \\ 2a\log_2 a &< 2^{2log_2 a} \end{aligned}$$

This means $ax < 2^x$ at $x = 2\log_2 a$. Therefore, $x < 2\log_2 a$ must be hold to make $ax \geq 2^x$ hold. $\square$

Theorem 5 $d_G(LR) < 5(n+k-1)$

(Proof) Let $S = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_l) \in X^l$ and $I = (i_1, i_2, ..., i_l) \in K^l$. Then, for an arbitrary $f \in$ LR defined by $\mathbf{w} \in \Re^l$ and $\mathbf{b} = (b_1, b_2, ..., b_{k-1}) \in \Re^{k-1}$,

$$\delta(f(\mathbf{x}_j), i_j) = \begin{cases} 1 & \text{if } \mathbf{w}\cdot\mathbf{x}_j - b_{i_j-1} \geq 0 \text{ and } \mathbf{w}\cdot\mathbf{x}_j - b_{i_j} < 0 \\ -1 & \text{if } \mathbf{w}\cdot\mathbf{x}_j - b_{i_j-1} < 0 \text{ or } \mathbf{w}\cdot\mathbf{x}_j - b_{i_j} \geq 0 \end{cases}$$

when $1 < i_j < k$,

$$\delta(f(\mathbf{x}_j), i_j) = \begin{cases} 1 & \text{if } \mathbf{w}\cdot\mathbf{x}_j - b_{i_j} < 0 \\ -1 & \text{if } \mathbf{w}\cdot\mathbf{x}_j - b_{i_j} \geq 0 \end{cases}$$

when $i_j = 1$, and

$$\delta(f(\mathbf{x}_j), i_j) = \begin{cases} 1 & \text{if } \mathbf{w}\cdot\mathbf{x}_j - b_{i_j-1} \geq 0 \\ -1 & \text{if } \mathbf{w}\cdot\mathbf{x}_j - b_{i_j-1} < 0 \end{cases}$$

when $i_j = k$.

Let

$$h_{j,i} = \{\mathbf{z} : \mathbf{z} \in \Re^{n+k-1}, \mathbf{z}\cdot(\mathbf{x}_j, \mathbf{1}_i) = 0\},$$

where $\mathbf{1}_i$ is a $(k-1)$-dimensional vector of which component values are 0 except the ith one-valued component. Note that $\mathbf{z}$ appeared in the definition of $h_{j,i}$ is a vector that corresponds to $(\mathbf{w}, \mathbf{b})$. Thus, $h_{j,i}$ is a hyperplain in $\Re^{n+k-1}$ with normal vector $(\mathbf{x}_j, \mathbf{1}_i)$, where the space $\Re^{n+k-1}$ can be seen as the functional space corresponding to LR. Consider a set H of hyperplains defined by

$$H = \{h_{j,i_j} : 1 \leq j \leq l, i_j \neq k\} \cup \{h_{j,i_j-1} : 1 \leq j \leq l, i_j \neq 1\}.$$

For $f, g \in$ LR defined by $(\mathbf{w}_f, \mathbf{b}_f), (\mathbf{w}_g, \mathbf{b}_g)$, respectively, if $f_{I,S} \neq g_{I,S}$, then the binary position vectors of $(\mathbf{w}_f, \mathbf{b}_f)$ and $(\mathbf{w}_g, \mathbf{b}_g)$ for H are different. Thus, the number of distinct binary position vectors for H must be at least 2^l in order that equation $|\Pi^G_{\text{LR}}(l)| = 2^l$ holds. The number of binary position vectors is maximized when $\mathcal{A}(H)$ is simple, and then it is equal to the number of cells by Lemma 2. Thus, by Lemma 3,

$$\sum_{i=0}^{d} \binom{m}{i} \geq 2^l$$

must hold, where $d = n+k-1$ and m is the number of hyperplains in H. Define $\Phi_d(m)$ as $\sum_{i=0}^{d} \binom{m}{d-i}$, then Φ_d is increasing, so

$$\Phi_d(2l) \geq 2^l$$

must hold because $m \leq 2l$. Since $\Phi_d(m) \leq (em/d)^d$ by Proposition A2.1 in [2],

$$(2el/d)^d \geq 2^l$$

must be hold. By replacing l with xd, this inequality becomes

$$(2ex)^d \geq 2^{xd},$$

which is equivalent to

$$2ex \geq 2^x. \tag{4}$$

To make (4) hold, x must be less than $2\log_2 2e < 5$ by Lemma 4. Thus,

$$l < 5d = 5(n + k - 1).$$

□

4 Risk Bounds on Learning LR

In this section, we show risk bounds on learning LR in two settings, the multi-class classification setting and the ordinal regression setting.

4.1 Vapnik's Framework of Risk Minimization

Learning framework we adopt here is Vapnik's framework of risk minimization. Vapnik considered the following general setting [12, p.20] of the learning problem, which generalizes all the settings of classification, regression and density estimation.

Problem 1 (Vapnik's general setting of the learning problem) *Let P be a probability measure on the space Z and let $\{Q(\cdot, \alpha) : \alpha \in \Lambda\}$ be a set of real-valued functions on Z parameterized by α. Given an i.i.d. sample $z_1, z_2, ..., z_l$, find $\alpha_0 \in \Lambda$ that minimize the risk functional*

$$R(\alpha) = \int Q(z, \alpha) dP(z)$$

under the condition that the probability measure P is unknown.

In the classification and regression settings, the space Z is the product space of spaces X and Y, and we consider the class of functions $f \in \mathcal{F}$ from the space X to the space Y and real-valued loss function L on $Y \times Y$. Then, Q is defined by

$$Q((x, y), f) = L(f(x), y).$$

The loss function L is the zero-one loss $L_{\text{0-1}}$ defined as

$$L_{\text{0-1}}(y_1, y_2) = \begin{cases} 0 & \text{if } y_1 = y_2 \\ 1 & \text{if } y_1 \neq y_2 \end{cases}$$

in the classification settings, and the L1 loss L_{L1} defined as

$$L_{\text{L1}}(y_1, y_2) = |y_1 - y_2|$$

in our ordinal regression settings.

In order to generalize the result obtained for the set of $\{0,1\}$-valued functions to the set of real-valued functions, Vapnik considered a set of *indicators* $I(\cdot,\alpha,\beta), \alpha \in \Lambda, \beta \in (\inf_{z,\alpha} Q(z,\alpha), \sup_{z,\alpha} Q(z,\alpha))$ of the set of real-valued functions $Q(\cdot,\alpha), \alpha \in \Lambda$:

$$I(z,\alpha,\beta) = \theta(Q(z,\alpha)-\beta) \quad \text{where } \theta(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0. \end{cases}$$

Note that $I(z,\alpha,\beta) = Q(z,\alpha)$ for all $\beta \in (0,1)$ when $\{Q(\cdot,\alpha) : \alpha \in \Lambda\}$ is a set of $\{0,1\}$-valued functions.

Vapnik showed the following theorem for the set of totally bounded nonnegative functions.

Theorem 6 (Vapnik [12, p.84]) *Let $\{Q(\cdot,\alpha) : \alpha \in \Lambda\}$ be a set of nonnegative functions on Z whose range is bounded by B, and let h denote the VC dimension of the set of indicators of the function class. Let $z_1, z_2, ..., z_l$ be an i.i.d. sample drawn from Z according to an arbitrary unknown distribution. Define $R_{\text{emp}}(\alpha)$ as $(\sum_{i=1}^{l} Q(z_i,\alpha))/l$. Then, the following inequality holds with probability at least $1-\delta$ simultaneously for all $\alpha \in \Lambda$:*

$$R(\alpha) \leq R_{\text{emp}}(\alpha) + \frac{B\mathcal{E}}{2}\left(1+\sqrt{1+\frac{4R_{\text{emp}}(\alpha)}{B\mathcal{E}}}\right),$$

where

$$\mathcal{E} = 4\frac{h(\ln(2l/h)+1)-\ln(\delta/4)}{l}.$$

4.2 Error Probability Bounds in the Multi-Class Classification Setting

Let $\mathcal{F}$ be a class of K-valued functions on X and let $\mathcal{F}_\delta$ denote the class of $\{-1,1\}$-valued functions defined by (2). Define $\mathcal{F}_{L_{0\text{-}1}}$ as

$$\mathcal{F}_{L_{0\text{-}1}} = \{Q(\cdot,f) : Q((\mathbf{x},i),f) \stackrel{def}{=} L_{0\text{-}1}(f(\mathbf{x}),i), f \in \mathcal{F}\}.$$

Then, trivially, $d_V(\mathcal{F}_{L_{0\text{-}1}}) = d_V(\mathcal{F}_\delta)$. Since $d_V(\mathcal{F}_\delta) = d_G(\mathcal{F})$, the following corollary of Theorem 6 can be obtained.

Corollary 1 *Let $\mathcal{F}$ be a class of K-valued functions on X. Let $(\mathbf{x}_1,i_1),(\mathbf{x}_2,i_2),...,(\mathbf{x}_l,i_l)$ be an i.i.d. sample drawn from $X \times K$ according to an arbitrary unknown distribution $\mathcal{D}$. Then, the following inequality holds with probability at least $1-\delta$ simultaneously for all $f \in \mathcal{F}$:*

$$\Pr_{(x,i)\sim\mathcal{D}}(f(\mathbf{x}) \neq i) \leq \frac{m}{l} + \frac{\mathcal{E}}{2}\left(1+\sqrt{1+\frac{4m}{l\mathcal{E}}}\right),$$

where m is the number of instances $(\mathbf{x}_j, i_j)$ with $f(\mathbf{x}_j) \neq i_j$ and

$$\mathcal{E} = 4\frac{d_G(\mathcal{F})(\ln(2l/d_G(\mathcal{F}))+1) - \ln(\delta/4)}{l}.$$

For $f : X \to K$, f is said to be an *ϵ-close hypothesis* when f satisfies that

$$\Pr_{(x,i)\sim\mathcal{D}}(f(\mathbf{x}) \neq i) < \epsilon.$$

The above corollary implies the following theorem proved by Ben-David et al. [1].

Theorem 7 (Ben-David et al. [1]) *Let $\mathcal{F}$ be a class of K-valued functions on X. Let $\mathcal{D}$ be an arbitrary unknown probabilistic distribution on $X \times K$. Then, there exists constant value $c > 0$ such that, for given $0 < \epsilon, \delta < 1$, the probability that a hypothesis in $\mathcal{F}$ consistent with*

$$l \geq \frac{c}{\epsilon}\left(d_G(\mathcal{F})\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right) \tag{5}$$

instances randomly drawn according to $\mathcal{D}$ is ϵ-close, is more than $1-\delta$.

By Theorem 7 and the results in Section 3.2, the number of instances necessary to PAC-learn by a consistent hypothesis finder is $O(k+n)$ for LR while $O(nk^2 \log k)$ for $\mathrm{DL}_{\mathcal{L}}$ with respect to parameters k and n.

For given m instances in $\Re^n \times K$, a consistent linear ranking function from $\Re^n$ to K can be obtained in time polynomial to n, l and k by solving a linear programming problem composed of $n+k-1$ variables and at most $2l$ constraints. Note that linear programming problem can be solved in polynomial time by such an algorithm like Karmarkar's one [6]. A consistent function in $\mathrm{DL}_{\mathcal{L}}$ can be also obtained in polynomial time by solving k learning problems of linear discrimination functions, which are known to be solved in polynomial time by linear programming [2].

4.3 Risk Bounds in the Ordinal Regression Setting

The set $\mathcal{F}_\theta$ of indicators of $\{Q(\cdot, f) : Q((\mathbf{x}, i), f) \stackrel{def}{=} L_{L1}(f(\mathbf{x}), i), f \in \mathcal{F}\}$ can be represented as follows:

$$\mathcal{F}_\theta = \{h_{f,\beta} : h_{f,\beta}(\mathbf{x}, i) \stackrel{def}{=} \theta(|f(\mathbf{x}) - i| - \beta), f \in \mathcal{F}, \beta \in (0, k-1)\}.$$

Lemma 5 $d_V(LR_\theta) < 2(\log_2 e(k-1))(n+k-1)$

(Proof) Let $S = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_l) \in X^l$ and $I = (i_1, i_2, ..., i_l) \in K^l$. Then, for an arbitrary $f \in \mathrm{LR}$ defined by $\mathbf{w} \in \Re^l$ and $\mathbf{b} = (b_1, b_2, ..., b_{k-1}) \in \Re^{k-1}$,

$$\theta(|f(\mathbf{x}_j) - i_j| - \beta) = \begin{cases} 0 & \text{if } \mathbf{w}\cdot\mathbf{x}_j - b_{i_j - \lceil\beta\rceil} \geq 0 \text{ and } \mathbf{w}\cdot\mathbf{x}_j - b_{i_j+\lceil\beta\rceil-1} < 0 \\ 1 & \text{if } \mathbf{w}\cdot\mathbf{x}_j - b_{i_j - \lceil\beta\rceil} < 0 \text{ or } \mathbf{w}\cdot\mathbf{x}_j - b_{i_j+\lceil\beta\rceil-1} \geq 0 \end{cases}$$

when $\beta \le i_j - 1$ and $\beta \le k - i_j$,

$$\theta(|f(\mathbf{x}_j) - i_j| - \beta) = \begin{cases} 0 & \text{if } \mathbf{w}\cdot\mathbf{x}_j - b_{i_j+\lceil\beta\rceil-1} < 0 \\ 1 & \text{if } \mathbf{w}\cdot\mathbf{x}_j - b_{i_j+\lceil\beta\rceil-1} \ge 0 \end{cases}$$

when $\beta > i_j - 1$ and $\beta \le k - i_j$,

$$\theta(|f(\mathbf{x}_j) - i_j| - \beta) = \begin{cases} 0 & \text{if } \mathbf{w}\cdot\mathbf{x}_j - b_{i_j-\lceil\beta\rceil} \ge 0 \\ 1 & \text{if } \mathbf{w}\cdot\mathbf{x}_j - b_{i_j-\lceil\beta\rceil} < 0 \end{cases}$$

when $\beta \le i_j - 1$ and $\beta > k - i_j$, and

$$\theta(|f(\mathbf{x}_j) - i_j| - \beta) = 0$$

when $\beta > i_j - 1$ and $\beta > k - i_j$.

Let

$$H = \{h_{j,i} : 1 \le j \le l, 1 \le i \le k-1\}$$

where

$$h_{j,i} = \{\mathbf{z} : \mathbf{z} \in \Re^{n+k-1}, \mathbf{z}\cdot(\mathbf{x}_j, \mathbf{1}_i) = 0\}$$

as defined in the proof of Theorem 5. By a similar discussion as that in the proof of Theorem 5,

$$\Phi_d(m) \ge 2^l$$

must hold, where $d = n + k - 1$ and m is the number of hyperplains in H. Since $m \le (k-1)l$,

$$\Phi_d((k-1)l) \ge 2^l$$

must hold. Thus, a similar argument using Lemma 4 as that in the proof of Theorem 5 leads to

$$l < 2(\log_2 e(k-1))(n+k-1).$$

□

Remark 2 *Note that*[4] $d_V(LR_\theta) \ge n+k-1$ *because* $LR_\theta \supseteq LR_{L_{0\text{-}1}}$ *and* $d_V(LR_{L_{0\text{-}1}}) = d_V(LR_\delta) \ge n+k-1$ *by Theorem 4.*

Corollary 2 *Let* $(\mathbf{x}_1, i_1), (\mathbf{x}_2, i_2), ..., (\mathbf{x}_l, i_l)$ *be an i.i.d. sample drawn from* $X \times K$ *according to an arbitrary unknown distribution* $\mathcal{D}$*. Then, the following inequality holds with probability at least* $1-\delta$ *simultaneously for all* $f \in \mathrm{LR}$*:*

$$E_{(\mathbf{x},i)\sim\mathcal{D}}(|f(\mathbf{x}) - i|) \le R_{\mathrm{emp}}(f) + \frac{(k-1)\mathcal{E}}{2}\left(1 + \sqrt{1 + \frac{4R_{\mathrm{emp}}(f)}{(k-1)\mathcal{E}}}\right),$$

where $R_{\mathrm{emp}}(f) = (\sum_{j=1}^{l} |f(\mathbf{x}_j) - i_j|)/l$,

$$\mathcal{E} = 4\frac{h(\ln(2l/h)+1) - \ln(\delta/4)}{l} \quad \text{and } h = 2(\log_2 e(k-1))(n+k-1).$$

[4]In the process of obtaining a risk bound by applying Theorem 6 to the ordinal regression of LR, $d_V(\mathrm{LR}_\theta)$ was calculated as n in [11], which contradicts our result.

Remark 3 *By Corollary 1, the following inequality holds with probability at least* $1-\delta$ *simultaneously for all* $f \in LR$*:*

$$E_{(\mathbf{x},i)\sim\mathcal{D}}(|f(\mathbf{x})-i|) \le (k-1)\Pr_{(x,i)\sim\mathcal{D}}(f(\mathbf{x}) \ne i) \le (k-1)\left(\frac{m}{l} + \frac{\mathcal{E}}{2}\left(1 + \sqrt{1 + \frac{4m}{l\mathcal{E}}}\right)\right),$$

where m *is the number of instances* $(\mathbf{x}_j, i_j)$ *with* $f(\mathbf{x}_j) \ne i_j$ *and*

$$\mathcal{E} = 4\frac{h(\ln(2l/h)) + 1) - \ln(\delta/4)}{l} \text{ and } h = 5(n+k-1).$$

With respect to k *and* n*, this bound is* $O(k(n+k))$*, which is better than* $O(k(n+k)\log k)$*, the bound obtained in Corollary 2.*

5 Concluding Remarks

We showed that graph dimension of the class of linear ranking functions is $\Theta(n+k)$, which is asymptotically significantly smaller that the graph dimension $\Omega(nk)$ of the class of $\{1, 2, ..., k\}$-valued decision lists naturally defined using $k-1$ linear discrimination functions. This difference induces the difference on sample complexity upper bounds for PAC learning of those classes. However, in order to show that sample complexities of the two learning problems are definitely different, their lower bounds should be also analyzed. Analyses of margin-based risk bounds in both the multiclass classification and the ordinal regression settings also look interesting.

Acknowledgment

The author would like to thank Prof. Mineichi Kudo and Jun Toyama for helpful discussions which led to this research.

References

[1] S. Ben-David, Nicolo Cesa-Bianchi, D. Haussler and P. M. Long. Characterizations of Learnability for Classes of $\{0, ..., n\}$-Valued Functions. *Journal of Computer and System Sciences* 50, 1995, pp.74-86.

[2] A. Blumer, A. Ehrenfeucht, D. Haussler and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis Dimension. *Journal of the ACM* 36(4), 1989, pp.929-965.

[3] K. Crammer and Y. Singer. Pranking with Ranking. *Advances in Neural Information Processing* 14, 2002, pp.641-647.

[4] H. Edelsbrunner, Algorithms in Combinatorial Geometry, *Springer-Verlag*, Berlin Heidelberg, 1987.

[5] R. Herbrich, T. Graepel and K. Obermayer. Large Margin Rank Boundaries for Ordinal Regression. Advances in Large Margin Classifiers, 2000, pp.115-132.

[6] N. Karmarkar. A New Polynomial-Time Algorithm for Linear Programming. *Combinatorica* 4, 1984, PP.373-395.

[7] A. Nakamura and N. Abe. Collaborative Filtering using Weighted Majority Prediction Algorithms. *Proceedings of the 15th International Conference on Machine Learning*, 1998, pp.395-403.

[8] A. Nakamura, M. Kudo and A. Tanaka. Collaborative Filtering using Restoration Operators. *Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2003, pp.339-349.

[9] B. K. Natarajan. On Learning Sets and Functions. *Machine Learning* 4, 1989, pp.67-97.

[10] S. Rajaram, A. Garg, X. S. Zhou and T. S. Huang. Classification Approach towards Ranking and Sorting Problems. *Proceedings of the 14th European Conference on Machine Learning, Lecture Notes in Artificial Intelligence* 2837, 2003, pp.301-312.

[11] A. Shashua and A. Levin. Taxonomy of Large Margin Principle Algorithms for Ordinal Regression Problems. Technical Report 2002-39, Leibniz Center for Research, School of Computer Science and Eng., the Hebrew University of Jerusalem.

[12] V. N. Vapnik. The Nature of Statistical Learning Theory (2nd Edition), *Springer-Verlag* New York, Inc., 1999.

[13] V. N. Vapnik and A. Y. Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory Probab. Appl.* 16(2), 1971, pp.264-280.