Exact Passive-Aggressive Algorithms for Ordinal Regression Using Interval Labels

Naresh Manwani and Mohit Chandra

***Abstract*— In this article, we propose exact passive-aggressive (PA) online algorithms for ordinal regression. The proposed algorithms can be used even when we have interval labels instead of actual labels for example. The proposed algorithms solve a convex optimization problem at every trial. We find an exact solution to those optimization problems to determine the updated parameters. We propose a support class algorithm (SCA) that finds the active constraints using the Karush–Kuhn–Tucker (KKT) conditions of the optimization problems. These active constraints form a support set, which determines the set of thresholds that need to be updated. We derive update rules for PA, PA-I, and PA-II. We show that the proposed algorithms maintain the ordering of the thresholds after every trial. We provide the mistake bounds of the proposed algorithms in both ideal and general settings. We also show experimentally that the proposed algorithms successfully learn accurate classifiers using interval labels as well as exact labels. The proposed algorithms also do well compared to other approaches.**

***Index Terms*— Interval labels, online learning, ordinal regression, passive-aggressive (PA).**

I. Introduction

ORDINAL regression is used to learn a model, which can predict labels from a discrete but an ordered set. Ordinal regression is frequently used in settings where it is natural to rank instances. For example, the labels ("do-not-bother" $\prec$ "only-if-you-must" $\prec$ "good" $\prec$ "very-good" $\prec$ "run-to-see") used in movie ratings [5]. Product ratings in online retail stores (e.g., Amazon, eBay, etc.), age of a person from its face image, etc. are other use cases of ordinal regression. Ordinal regression has been successfully used in a wide variety of applications ranging from collaborative filtering [20] to ecology [10] to detect the severity of Alzheimer disease [8] and many more.

An ordinal regression requires a linear (nonlinear) function and a set of $K-1$ thresholds (K be the number of classes). Each threshold corresponds to a class. Thus, the thresholds should have the same order as their corresponding classes. The rank (class) of an example is predicted based on the relative position of the function value concerning different thresholds. Ordinal regression is different from multiclass classification in the sense that there is a natural ordering among the class labels. It is also different from regression as the target values can take only discrete values. Large margin formulations for ordinal regression are proposed in [3] and [23]. Ordering of thresholds can be maintained implicitly or explicitly. In explicit methods [3], [23], explicit ordering constraint is posed in the formulation. On the other hand, implicit methods [3], [14] capture the ordering by posing separability conditions between every pair of classes. Note that ordinal regression problem is different from learning to rank problem [2], [12], [13], [24]. In learning to rank, the goal is to learn ordering among multiple target instances for a given an example (optimizing average precision and normalized discounted cumulative gain).

An incremental algorithm for ordinal regression is proposed in [9]. In the incremental algorithms, given initial training data, the optimal classifier is learned. Then, the algorithm observes a new example. The incremental algorithm adds this new example in the training set and finds an efficient way to learn the new classifier using bigger training set and old classifier.

In the case of big data, they require a huge amount of computation time and memory to solve the optimization problem. In contrast, online learning updates its hypothesis based on a single example at every instant. Thus, online algorithms are even faster than the incremental algorithms. Perceptron-based algorithm for online ordinal regression is proposed in [5] and [11]. Passive-aggressive (PA) [4] is another principled method of learning classifiers in online fashion. The updates made by PA are more aggressive to make the loss incurred on the current example zero. Crammer *et al.* [4] propose two more variants of PA (namely, PA-I and PA-II). This approach can be applied to learning multi-class classification, regression, multitask learning, and so on. A variant of PA learning for the multi-class classifier is proposed in [16].

PA algorithms for ordinal regression have not been well addressed in the literature. Moreover, in the above approaches, it is assumed that the training data contain exact labels for each observation. However, in many situations, we get interval labels instead of the precise label. For example, while predicting product ratings, we can get an entire range of scores (e.g., 1–3 and 4–7) from different customers. Similarly, while learning for predicting human age, we can get a variety of values around the actual age of the person (e.g., 0–9, 10–19, . . ., 90–99). A large margin batch algorithm is proposed in [1] using interval labels. In [15], perceptron-based approach is proposed for learning ordinal regression classifier using interval labels.

In this article, we propose PA algorithms for ordinal regression. These algorithms not only utilize the ordering of the class

Manuscript received November 26, 2018; revised June 1, 2019; accepted August 27, 2019. Date of publication October 14, 2019; date of current version September 1, 2020. *(Corresponding author: Naresh Manwani.)*

The authors are with the Machine Learning Laboratory, KCIS, IIIT Hyderabad, Hyderabad 500032, India (e-mail: naresh.manwani@iiit.ac.in; mohit.chandra@research.iiit.ac.in).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2019.2939861

labels but also are generic enough to accept both exact as well as interval labels in the training data. To the best of authors' knowledge, this is the first work in that direction. Our key contributions are as follows.

A. Main Contributions

1) We derive update rules for all the variants of the PA approach (PA, PA-I, and PA-II) for ordinal regression. At trial t, PA algorithms solve a convex optimization problem. We find the exact solution to these optimization problems. We propose support class algorithm (SCA) which, at any trial, finds active constraints in the Karush–Kuhn–Tucker (KKT) optimality conditions to find the support class set. Support class set describes the thresholds that need to be updated in addition to the weight vector. We show that SCA correctly finds the support classes.
2) We show that the proposed PA algorithms implicitly maintain the ordering of the thresholds after every trial.
3) We provide the mistake bounds for the proposed algorithms in both general and ideal cases.
4) We perform extensive simulations of the proposed algorithms on various data sets and show their effectiveness by comparing the results with different other algorithms.

This article is organized as follows. In Section II, we discuss a generic framework ordinal regression using interval (exact) labels. In Section III, we derive the update rules for PA, PA-I, and PA-II. The order preservation guarantees of the proposed algorithms are discussed in Section IV. In Section V, we discuss the mistake bounds. Experiments are presented in Section VI. We conclude this article with some remarks in Section VII.

II. Ordinal Regression Using Interval (Exact) Labels

Let $\mathcal{X} \subset \mathbb{R}^d$ be the instance space and $\mathcal{Y} = \{1, \ldots, K\}$ be the label space. For every instance $\mathbf{x} \in \mathcal{X}$, an interval label $[y_l, y_r] \in \mathcal{Y} \times \mathcal{Y}$ is given. The exact (actual) label y lie in the interval label. When $y_l = y_r$ for all the examples, it becomes the exact label scenario. Let $S = \{(\mathbf{x}^1, y_l^1, y_r^1), \ldots, (\mathbf{x}^T, y_l^T, y_r^T)\}$ be the training set. Ordinal regression using a function $f : \mathcal{X} \to \mathbb{R}$ and thresholds $\theta_1 \leq \cdots \leq \theta_{K-1}$ is defined as

$$h(\mathbf{x}) = \min_{i \in [K]} \{i \ : \ f(\mathbf{x}) - \theta_i < 0\} \tag{1}$$

where $\theta_K = \infty$ and $[K] = \{1, \ldots, K\}$. Let $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$. We can use the kernel trick to generalize for nonlinear functions. We use mean absolute error (MAE) [1] to capture the discrepancy between the interval label and the predicted label

$$\begin{aligned} L_{\text{MAE}}(f(\mathbf{x}), \boldsymbol{\theta}, y_l, y_r) &= \sum_{i=1}^{y_l - 1} \mathbb{I}_{\{f(\mathbf{x}) < \theta_i\}} + \sum_{i=y_r}^{K-1} \mathbb{I}_{\{f(\mathbf{x}) \geq \theta_i\}} \\ &= (y_l - h(\mathbf{x}))\mathbb{I}_{\{h(\mathbf{x}) < y_l\}} + (h(\mathbf{x}) - y_r)\mathbb{I}_{\{h(\mathbf{x}) > y_r\}} \end{aligned}$$

where $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_K\}$. L_{MAE} takes value 0 whenever $\theta_{y_l} \leq f(\mathbf{x}) < \theta_{y_r}$. When $y_l = y_r = y$ (exact label case), $L_{\text{MAE}} = |y - h(\mathbf{x})|$. Note that L_{MAE} is a discontinuous loss. A convex surrogate of this loss function is as follows [1]:[1]

$$\begin{aligned} &L_{\text{IMC}}(f(\mathbf{x}), \boldsymbol{\theta}, y_l, y_r) \\ &= \sum_{i=1}^{y_l - 1} l_i + \sum_{i=y_r}^{K-1} l_i \\ &= \sum_{i=1}^{y_l - 1} [1 - f(\mathbf{x}) + \theta_i]_+ + \sum_{i=y_r}^{K-1} [1 + f(\mathbf{x}) - \theta_i]_+ \end{aligned} \tag{2}$$

where $\boldsymbol{\theta} = [\theta_1 \ \ldots \ \theta_{K-1}]$ and $[z]_+ = \max(0, z)$. When $y_l = y_r$, then L_{IMC} leads to the implicit formulation described in [3]. L_{IMC} is shown to be the Fisher consistent [18].

III. Exact Passive-Aggressive Algorithms for Ordinal Regression

PA [4] is a principled approach for supervised learning in online fashion. Here, we develop PA algorithms for ordinal regression, which can learn even with interval labels. The proposed method is based on the interval insensitive loss described in (2). We derive the update equations for PA, PA-I, and PA-II separately.

A. PA Algorithm

Let $\mathbf{x}^t$ be the example being observed at trial t. Let $\mathbf{w}^t \in \mathbb{R}^d$ and $\boldsymbol{\theta} \in \mathbb{R}^{K-1}$ be the parameters of the ordinal regression at time t. We now use these parameters to predict the label. Then, we observe the actual label(s). PA algorithm finds $\mathbf{w}^{t+1}$ and $\boldsymbol{\theta}^{t+1}$, which are closest to $\mathbf{w}^t$ and $\boldsymbol{\theta}^t$ such that the loss L_{IMC} (2) becomes zero for the current example. Thus,

$$\begin{aligned} \mathbf{w}^{t+1}, \boldsymbol{\theta}^{t+1} = \arg\min_{\mathbf{w}, \boldsymbol{\theta}} \ & \frac{1}{2}\|\mathbf{w} - \mathbf{w}^t\|^2 + \|\boldsymbol{\theta} - \boldsymbol{\theta}^t\|^2 \\ \text{s.t.} \ & \begin{cases} \mathbf{w} \cdot \mathbf{x}^t - \theta_i \geq 1, & i = 1, \ldots, y_l^t - 1 \\ \mathbf{w} \cdot \mathbf{x}^t - \theta_i \leq -1, & i = y_r^t, \ldots, K - 1. \end{cases} \end{aligned}$$

Geometrically, $(\mathbf{w}^{t+1}, \boldsymbol{\theta}^{t+1})$ are found by projecting $(\mathbf{w}^t, \boldsymbol{\theta}^t)$ onto the half-space of vectors, which attain zero value of L_{IMC} on $\mathbf{x}^t$. The algorithm is passive whenever $L_{\text{IMC}} = 0$ (that is, $\mathbf{w}^{t+1}$ and $\boldsymbol{\theta}^{t+1}$ are same as $\mathbf{w}^t$ and $\boldsymbol{\theta}^t$). In contrast, when $L_{\text{IMC}} > 0$, the algorithm aggressively forces $\mathbf{w}^{t+1}$ and $\boldsymbol{\theta}^{t+1}$ to be such that $L_{\text{IMC}}(\mathbf{w}^{t+1}.\mathbf{x}^t, \boldsymbol{\theta}^{t+1}, y_l^t, y_r^t) = 0$. The KKT optimality conditions are as follows:

$$\begin{aligned} \mathbf{w} &= \mathbf{w}^t + \left(\sum_{i=1}^{y_l^t - 1} \lambda_i^t - \sum_{i=y_r^t}^{K-1} \mu_i^t \right) \mathbf{x}^t \\ \lambda_i \geq 0; \quad \theta_i &= \theta_i^t - \lambda_i^t; \quad i = 1 \ldots y_l^t - 1 \\ \mu_i \geq 0; \quad \theta_i &= \theta_i^t + \mu_i^t, \quad i = y_r^t \ldots K - 1 \\ 1 + \theta_i - \mathbf{w} \cdot \mathbf{x}^t \leq 0; \quad \lambda_i \ &(1 + \theta_i - \mathbf{w} \cdot \mathbf{x}^t) = 0 \\ &i = 1 \ldots y_l^t - 1 \\ 1 + \mathbf{w} \cdot \mathbf{x}^t - \theta_i \leq 0; \quad \mu_i &(1 + \mathbf{w} \cdot \mathbf{x}^t - \theta_i) = 0 \\ &i = y_r^t \ldots K - 1 \end{aligned}$$

[1]Note that $[1 - f(\mathbf{x}) + \theta_i]_+$ is convex loss (hinge loss) for each $i \in \{1, \ldots, y_l^t - 1\}$. Thus, sum of these losses is also convex. Similarly, sum of $[1 + f(\mathbf{x}) - \theta_i]_+$ is also convex. Hence, L_{IMC} is convex. It is also easy to verify that L_{IMC} always upper bounds L_{MAE}.

where $\lambda_i \geq 0\ i \in [y_l^t - 1]$ and $\mu_i \geq 0,\ i = y_r^t, \ldots, K-1$ are Lagrange multipliers. Let $S_l^t = \{1 \leq i \leq y_l^t - 1 | \lambda_i^t > 0\}$ be the left support set and $S_r^t = \{y_r^t \leq i \leq K-1 | \mu_i > 0\}$ be the right support set. Thus, optimal $\mathbf{w}$ can be rewritten as $\mathbf{w} = \mathbf{w}^t + a^t \mathbf{x}^t$ where $a^t = \sum_{i \in S_l^t} \lambda_i^t - \sum_{i \in S_r^t} \mu_i^t$. Also, $\mathbf{w} \cdot \mathbf{x}^t - \theta_i = 1,\ i \in S_l^t$ and $\mathbf{w} \cdot \mathbf{x}^t - \theta_i = -1,\ i \in S_r^t$. Thus, we get $\lambda_i = 1 - \mathbf{w}^t.\mathbf{x}^t - \theta_i^t - a^t \|\mathbf{x}^t\|^2 = l_i^t - a^t \|\mathbf{x}^t\|^2,\ \forall i \in S_l^t$ and $\mu_i = 1 - \theta_i^t + \mathbf{w}^t.\mathbf{x}^t + a^t \|\mathbf{x}^t\|^2 = l_i^t + a^t \|\mathbf{x}^t\|^2,\ \forall i \in S_r^t$. Putting the values of λ_i and μ_i in the expression of a^t, we get, $a^t = ((\sum_{i \in S_l^t} l_i^t - \sum_{i \in S_r^t} l_i^t)/(1 + (|S_l^t| + |S_r^t|)\|\mathbf{x}^t\|^2))$. Note that PA updates assume that at every trial t, sets S_l^t and S_r^t are known. The complete description of the PA algorithm is as given in Algorithm 1.

Algorithm 1 PA Algorithm

Input Training set S
Initialize $\mathbf{w}^0$ and $\boldsymbol{\theta}^0$
for $t = 1, \cdots, T$ **do**
  $\mathbf{x}^t \leftarrow$ *randomly sample an instance from* S
  Predict: $\hat{y}^t = \mathbf{w}^t.\mathbf{x}^t$
  Observe y_l^t, y_r^t
  $l_i^t = \max(0, 1 + \theta_i^t - \mathbf{w}^t.\mathbf{x}^t),\ i = 1 \ldots y_l^t - 1$
  $l_i^t = \max(0, 1 + \mathbf{w}^t.\mathbf{x}^t - \theta_i^t),\ i = y_r^t \ldots K-1$
  $S_l^t, S_r^t = \text{SCA}(y_l^t; y_r^t; \mathbf{x}^t; l_i^t, i \in [K-1])$
  Update:

$$\mathbf{w}^{t+1} = \mathbf{w}^t + a^t \mathbf{x}^t$$
$$\theta_i^{t+1} = \theta_i^t - l_i^t + \|\mathbf{x}^t\|^2 a^t, \quad \forall i \in S_l^t$$
$$\theta_i^{t+1} = \theta_i^t + l_i^t + \|\mathbf{x}^t\|^2 a^t, \quad \forall i \in S_r^t$$

end for

Determining Support Sets S_l^t and S_r^t: L_{IMC} decreases as we move away from the interval label $[y_l^t, y_r^t]$. We initialize with $S_l^t = \{y_l^t - 1\}$ and $S_r^t = \{y_r^t\}$. We can easily verify that with this initialization, $\lambda_{y_l^t - 1}^t, \mu_{y_r^t}^t > 0$. We start with threshold $\theta_{y_l^t - 2}^t$ and find corresponding Lagrange multiplier $\lambda_{y_l^t - 2}^t$. If $\lambda_{y_l^t - 2}^t > 0$, then we add $y_l^t - 2$ to S_l^t. Otherwise, we check if $\mu_{y_r^t + 1}^t$ is positive. If so, we add $y_r^t + 1$ to S_r^t. We repeatedly check this for all the thresholds. The detailed approach for finding support sets is described in Algorithm 2.

The following lemma shows the correctness of the SCA.

Lemma 1: Assume that $S_l^t \neq \phi$. Let, $k \notin S_l^t$ and $k+1 \in S_l^t$. Then, $k' \notin S_l^t,\ \forall k' < k$.

Proof: We are given that $k \notin S_l^t$. Thus,

$$\lambda_k^t = l_k^t - \frac{\|\mathbf{x}^t\|^2 \left(l_k^t + \sum_{j \in S_l^t} l_j^t - \sum_{j \in S_r^t} l_j^t\right)}{1 + \|\mathbf{x}^t\|^2 \left(\left|S_l^t\right| + 1 + \left|S_r^t\right|\right)} \leq 0$$

$\forall k' < k$, we know $l_{k'}^t \leq l_k^t$. Now, if we try to add k' in S_l^t, then

$$\begin{aligned}\lambda_{k'}^t &= l_{k'}^t - \frac{\|\mathbf{x}^t\|^2 \left(l_{k'}^t + \sum_{j \in S_l^t} l_j^t - \sum_{j \in S_r^t} l_j^t\right)}{1 + \|\mathbf{x}^t\|^2 (1 + |S_l^t| + |S_r^t|)} \\ &\leq \frac{l_k^t \left(1 + \|\mathbf{x}^t\|^2 (|S_l^t| + |S_r^t|)\right)}{1 + \|\mathbf{x}^t\|^2 (1 + |S_l^t| + |S_r^t|)} \\ &\quad - \frac{\|\mathbf{x}^t\|^2 \left(\sum_{j \in S_l^t} l_j^t - \sum_{j \in S_r^t} l_j^t\right)}{1 + \|\mathbf{x}^t\|^2 (1 + |S_l^t| + |S_r^t|)}\end{aligned}$$

Algorithm 2 SCA

Input: y_l^t; y_r^t; $\mathbf{x}^t$; $l_i^t,\ i \in [K-1]$
Initialize: $S_l^t = \{y_l^t - 1\}$, $S_r^t = \{y_r^t\}$, flag = 1, p = $y_l^t - 2$, q = $y_r^t + 1$
while flag = 1 **do**
  if $p > 0$ **then**
    if $l_p^t - \frac{\|\mathbf{x}^t\|^2 (l_p^t + \sum_{j \in S_l^t} l_j^t - \sum_{j \in S_r^t} l_j^t)}{1 + \|\mathbf{x}^t\|^2 (1 + |S_l^t| + |S_r^t|)} > 0$ **then**
      $S_l^t = S_l^t \cup \{p\}$; $p = p - 1$; flag = 1
    else
      flag=0
    end if
  end if
  if $q < K$ **then**
    if $l_q^t + \frac{\|\mathbf{x}^t\|^2 (\sum_{j \in S_l^t} l_j^t - l_q^t - \sum_{j \in S_r^t} l_j^t)}{1 + \|\mathbf{x}^t\|^2 (1 + |S_r^t| + |S_l^t|)} > 0$ **then**
      $S_r^t = S_r^t \cup \{q\}$; $q = q + 1$; flag = 1
    else
      flag=0
    end if
  end if
end while

$$= l_k^t - \frac{\|\mathbf{x}^t\|^2 \left(l_k^t + \sum_{j \in S_l^t} l_j^t - \sum_{j \in S_r^t} l_j^t\right)}{1 + \|\mathbf{x}^t\|^2 (1 + |S_l^t| + |S_r^t|)} = \lambda_k \leq 0.$$

Thus, $k' \notin S_l^t$. ∎

Thus, if a threshold does not belong to the left support class S_l^t then all the threshold on its left side also do not belong to S_l^t. If we start adding the classes in the support class set in decreasing order of respective losses, then this would ensure that we end up with only those classes which have positive Lagrange multiplier. Similarly, it can be shown that if $k-1 \in S_r^t$ and $k \notin S_r^t$, then $k' \notin S_r^t,\ \forall k' > k$, which means if a threshold does not belong the right support class S_r^t, then all the threshold on its right side also do not belong to S_r^t.

B. PA-I

PA-I finds new parameters by solving the following optimization problem:

$$\arg\min_{\mathbf{w}, \boldsymbol{\theta}} \frac{1}{2}\|\mathbf{w} - \mathbf{w}^t\|^2 + \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^t\|^2 + C \left(\sum_{i=1}^{y_l^t - 1} \xi_i + \sum_{y_r^t}^{K-1} \xi_i\right)$$
$$\text{s.t.} \begin{cases} \mathbf{w}.\mathbf{x}^t - \theta_i \geq 1 - \xi_i; \quad \xi_i \geq 0, & i = 1, \ldots, y_l^t - 1 \\ \mathbf{w}.\mathbf{x}^t - \theta_i \leq -1 + \xi_i; \quad \xi_i \geq 0, & i = y_r^t, \ldots, K-1 \end{cases}$$

where C is the aggressiveness parameter. We skip the derivation of PA-I updates as it follows the same steps used in case of PA. PA-I updates the parameters as follows:

$$\mathbf{w} = \mathbf{w}^t + \left(\sum_{i \in S_l^t} \lambda_i - \sum_{i \in S_r^t} \mu_i\right) \mathbf{x}^t$$
$$\lambda_i = \min(C, l_i^t - a^t \|\mathbf{x}^t\|^2), \quad i \in S_l^t$$
$$\mu_i = \min(C, l_i^t + a^t \|\mathbf{x}^t\|^2), \quad i \in S_r^t$$

where $S_l^t = \{1 \leq i \leq y_l^t - 1 \mid \lambda_i > 0\}$, $S_r^t = \{y_r^t \leq i \leq K - 1 \mid \mu_i > 0\}$ and $a^t = \sum_{i \in S_l^t} \lambda_i^t - \sum_{i \in S_r^t} \mu_i^t$. PA-I works same as PA except that it uses a different approach to determine the support sets S_l^t and S_r^t. We use an iterative approach to find

the support sets. We first find the values of all λ_i^t and μ_i^t and then compute a^t. We repeat it until all the values get converge. Then, we include i in S_l^t or S_r^t based on whether $\lambda_i > 0$ or $\mu_i > 0$. SCA-I for PA-I is discussed in Algorithm 3.

Algorithm 3 SCA-I

Input: y_l^t; y_r^t; $\mathbf{x}^t$ and l_i^t, $i \in [K-1]$
Initialize: $S_l^t = \{y_l^t - 1\}$, $S_r^t = \{y_r^t\}$, $p = y_l^t - 2$, $q = y_r^t + 1$
while $\lambda_i^1, \ldots, \lambda_{y_l^t-1}^t, \mu_{y_r^t}^t, \ldots, \mu_{K-1}^t$ do not converge **do**
  for $i = \mathrm{p}, \cdots, 1$ **do**
    if $\min(C, l_i^t - a^t\|\mathbf{x}_t\|^2) > 0$ **then**
      $S_l^t = S_l^t \cup \{i\}$
    else
      if $i \in S_l^t$ **then**
        $S_l^t = S_l^t - \{i\}$; $\lambda_i^t = 0$
      end if
    end if
  end for
  for $i = \mathrm{q}, \cdots, K-1$ **do**
    if $\min(C, l_i^t + a^t\|\mathbf{x}_t\|^2) > 0$ **then**
      $S_r^t = S_r^t \cup \{i\}$
    else
      if $i \in S_r^t$ **then**
        $S_r^t = S_r^t - \{i\}$; $\mu_i^t = 0$
      end if
    end if
  end for
end while

C. PA-II

PA-II finds the new parameters by minimizing the following objective function:

$$\mathbf{w}^{t+1}, \boldsymbol{\theta}^{t+1} = \arg\min_{\mathbf{w},\boldsymbol{\theta}} \frac{1}{2}\|\mathbf{w} - \mathbf{w}^t\|^2 + \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^t\|^2 + C\left(\sum_{i=1}^{y_l^t-1} \xi_i^2 + \sum_{i=y_r^t}^{K-1} \xi_i^2\right)$$

$$\text{s.t.} \begin{cases} \mathbf{w}.\mathbf{x}^t - \theta_i \geq 1 - \xi_i, & i = 1, \ldots, y_l^t - 1 \\ \mathbf{w}.\mathbf{x}^t - \theta_i \leq -1 + \xi_i, & i = y_r^t, \ldots, K-1. \end{cases}$$

The PA-II update equations are as follows:

$$\mathbf{w}^{t+1} = \mathbf{w}^t + a^t\mathbf{x}^t$$
$$\theta_i^{t+1} = \theta_i^t - \lambda_i^t \quad \forall i \in S_l^t$$
$$\theta_i^{t+1} = \theta_i^t + \mu_i^t \quad \forall i \in S_r^t$$

where $\lambda_i^t = ((l_i^t - a^t\|\mathbf{x}^t\|^2)/(1 + (1/2C)))$, $\mu_i^t = ((l_i^t + a^t\|\mathbf{x}_t)\|^2)/(1 + (1/2C)))$, and $a^t = ((\sum_{i\in S_l^t} l_i^t - \sum_{i\in S_r^t} l_i^t)/(1 + (1/2C) + \|\mathbf{x}^t\|^2\{|S_l^t| + |S_r^t|\}))$. Thus, a class $i \in \{1, \ldots, y_l^t - 1\}$ lies in the left support set S_l^t if $\lambda_i^t > 0$, and hence, $l_i^t - a^t\|\mathbf{x}^t\|^2 > 0$. Similarly, a class $i \in \{y_r^t, \ldots K-1\}$ lies in the right support set S_r^t if $\mu_i^t > 0$, and hence, $l_i^t + a^t\|\mathbf{x}^t\|^2 > 0$. SCA-II described in Algorithm 4 provides a detailed description of selecting S_l^t and S_r^t.

Algorithm 4 SCA-II

Input: y_l^t; y_r^t; $\mathbf{x}^t$; l_i^t, $i \in [K-1]$
Initialize: $S_l^t = \{y_l^t - 1\}$, $S_r^t = \{y_r^t\}$, flag = 1, p = $y_l^t - 2$, q = $y_r^t + 1$
while flag = 1 **do**
  if $p > 0$ **then**
    if $l_p^t - ((\|\mathbf{x}^t\|^2\left(\sum_{i\in S_l^t} l_i^t - \sum_{i\in S_r^t} l_i^t\right))/((1 + (1/2C) + \|\mathbf{x}^t\|^2\{|S_l^t| + |S_r^t|\} > 0))$ **then**
      $S_l^t = S_l^t \cup \{p\}$; $p = p - 1$; flag = 1
    else
      flag=0
    end if
  end if
  if $q < K$ **then**
    if $l_q^t + ((\|\mathbf{x}^t\|^2\left(\sum_{i\in S_l^t} l_i^t - \sum_{i\in S_r^t} l_i^t\right)))/((1 + \frac{1}{2C} + \|\mathbf{x}^t\|^2\{|S_l^t| + |S_r^t|\})) > 0$ **then**
      $S_r^t = S_r^t \cup \{q\}$; $q = q + 1$; flag = 1
    else
      flag=0
    end if
  end if
end while

IV. Correctness of PA Algorithms

Now, we will show that our approach inherently maintains the ordering of thresholds in each iteration.

Theorem 2 (Order Preservation of Thresholds Using PA Algorithm): Let $\theta_1^t \leq \cdots \leq \theta_{K-1}^t$ be the thresholds at trial t. Let $\theta_1^{t+1}, \ldots, \theta_{K-1}^{t+1}$ be the updated thresholds using PA. Then, $\theta_1^{t+1} \leq \cdots \leq \theta_{K-1}^{t+1}$.

Proof: We need to analyse the following different cases.

1) We know that $\theta_k^{t+1} = \theta_k^t$, $k = y_l^t \ldots y_r^t - 1$. Thus, $\theta_{y_l^t}^{t+1} \leq \ldots \leq \theta_{y_r^t-1}^{t+1}$.
2) $\forall k \in S_l^t$, we see that
$$\theta_k^{t+1} = -1 + \mathbf{w}\cdot\mathbf{x} + \frac{\|\mathbf{x}^t\|^2\left(\sum_{i\in S_l^t} l_i^t - \sum_{i\in S_r^t} l_i^t\right)}{1 + \|\mathbf{x}^t\|^2\left(|S_l^t| + |S_r^t|\right)}.$$
Thus, all the thresholds in the set S_l^t are mapped to the same value, and hence, the ordering is preserved.
3) $\forall k \in S_r^t$, we see that
$$\theta_k^{t+1} = 1 + \mathbf{w}\cdot\mathbf{x} + \frac{\|\mathbf{x}^t\|^2\left(\sum_{i\in S_l^t} l_i^t - \sum_{i\in S_r^t} l_i^t\right)}{1 + \|\mathbf{x}^t\|^2\left(|S_l^t| + |S_r^t|\right)}.$$
All the thresholds in the set S_r^t are mapped to the same value and hence the ordering is preserved.
4) Let $k, k+1 \in [y_l^t - 1]\triangle S_l^t$ where $\triangle$ is symmetric difference between sets. Then, $\theta_{k+1}^{t+1} - \theta_k^{t+1} = \theta_{k+1}^t - \theta_k^t \geq 0$.
5) Let $k \in [y_l^t - 1]\triangle S_l^t$ and $k+1 \in S_l^t$. Then, using Theorem 1, we get
$$l_k^t \leq \frac{\|\mathbf{x}^t\|^2\left(l_k^t + \sum_{i\in S_l^t} l_i^t - \sum_{i\in S_r^t} l_i^t\right)}{1 + \|\mathbf{x}^t\|^2\left(|S_l^t| + 1 + |S_r^t|\right)} \leq \frac{\|\mathbf{x}^t\|^2\left(\sum_{i\in S_l^t} l_i^t - \sum_{i\in S_r^t} l_i^t\right)}{1 + \|\mathbf{x}^t\|^2\left(|S_l^t| + |S_r^t|\right)} = a^t\|\mathbf{x}^t\|^2$$

$\theta_{k+1}^{t+1}-\theta_k^{t+1}=\theta_{k+1}^t-\theta_k^t-l_k^t-a^t\|\mathbf{x}^t\|^2=\theta_{k+1}^t-(l_{k+1}^t-\theta_k^t+\theta_{k+1}^t)-a^2\|\mathbf{x}^t\|^2-\theta_k^t=-l_{k+1}^t-a^t\|\mathbf{x}^t\|^2\geq 0.$

6) Let $k,k+1\in\{y_r^t,\ldots,K-1\}\triangle S_r^t$, then $\theta_{k+1}^{t+1}-\theta_k^{t+1}=\theta_{k+1}^t-\theta_k^t\geq 0$.

7) Let $k+1\in\{y_r^t,\ldots,K-1\}\triangle S_r^t$ and $k\in S_r^t$. Then,

$$l_{k+1}^t\leq -\frac{\|\mathbf{x}^t\|^2\left(\sum_{i\in S_l^t}l_i^t-\sum_{i\in S_r^t}l_i^t-l_{k+1}^t\right)}{1+\|\mathbf{x}^t\|^2\left(|S_l^t|+1+|S_r^t|\right)}\leq -\frac{\|\mathbf{x}^t\|^2\left(\sum_{i\in S_l^t}l_i^t-\sum_{i\in S_r^t}l_i^t\right)}{1+\|\mathbf{x}^t\|^2\left(|S_l^t|+|S_r^t|\right)}=-a^t\|\mathbf{x}^t\|^2$$

$\theta_{k+1}^{t+1}-\theta_k^{t+1}=\theta_{k+1}^t-\theta_k^t-l_k^t-a^t\|\mathbf{x}^t\|^2=\theta_{k+1}^t-(l_{k+1}^t-\theta_k^t+\theta_{k+1}^t)-a^2\|\mathbf{x}^t\|^2-\theta_k^t=-l_{k+1}^t-a^t\|\mathbf{x}^t\|^2\geq 0.$

This completes the proof. ∎

Theorem 3 (Order Preservation of Thresholds Using PA-I): Let $\theta_1^t\leq\cdots\leq\theta_{K-1}^t$ be the thresholds at trial t. Let $\theta_1^{t+1},\ldots,\theta_{K-1}^{t+1}$ be the updated thresholds using PA-I. Then, $\theta_1^{t+1}\leq\cdots\leq\theta_{K-1}^{t+1}$.

Proof: The proof follows in the same manner as PA algorithm. We only consider, here, the following two cases.

1) $k+1\in S_l^t$ and $k\in[y_l^t-1]\triangle S_l^t$. Thus, $\lambda_k^t<0$, which means $l_k^t-a^t\|\mathbf{x}_t\|^2<0$ as $C>0$. Also, $\lambda_{k+1}^t=\min(C,l_{k+1}^t-a^t\|\mathbf{x}^t\|^2)>0$. When $\lambda_{k+1}^t=l_{k+1}^t-a^t|\mathbf{x}^t\|^2$, we see that

$$\begin{aligned}\theta_{k+1}^{t+1}-\theta_k^{t+1}&=\theta_{k+1}^t-l_{k+1}^t+a^t|\mathbf{x}^t\|^2-\theta_k^t\\&=\theta_{k+1}-\left(l_k^t-\theta_k^t+\theta_{k+1}^t\right)+a^t|\mathbf{x}^t\|^2-\theta_k^t\\&=-l_k^t+a^t|\mathbf{x}^t\|^2\geq 0.\end{aligned}$$

When $\lambda_{k+1}^t=C$ ($C\leq l_{k+1}^t-a^t\|\mathbf{x}_t\|^2$), we have $\theta_{k+1}^{t+1}-\theta_k^{t+1}=\theta_{k+1}^t-C-\theta_k^t\geq\theta_{k+1}^t-l_{k+1}^t+a^t\|\mathbf{x}^t\|^2-\theta_k^t\geq 0$.

2) Let $k,k+1\in S_l^t$. Thus, $\theta_{k+1}^{t+1}-\theta_k^{t+1}=\theta_{k+1}^t-\theta_k^t-\lambda_{k+1}^t+\lambda_k^t$. There can be four different cases as follows.

   a) When $\lambda_{k+1}^t=\lambda_k^t=C$. Thus, $\theta_{k+1}^{t+1}-\theta_k^{t+1}=\theta_{k+1}^t-\theta_k^t\geq 0$. Similarly, is the case when $\lambda_k^t=C$, then $\lambda_{k+1}^t=C$ due to the fact that $l_{k+1}^t\geq l_k^t$.

   b) Let $\lambda_k^t=l_k^t-a^t\|\mathbf{x}^t\|^2$ and $\lambda_{k+1}^t=l_{k+1}^t-a^t\|\mathbf{x}^t\|^2$. Thus, $\theta_{k+1}^{t+1}=\theta_k^{t+1}=-1+\mathbf{w}^t\cdot\mathbf{x}^t+a^t\|\mathbf{x}_t\|^2$.

   c) Let $\lambda_k^t=l_k^t-a^t\|\mathbf{x}^t\|^2$ and $\lambda_{k+1}^t=C$. We see that $\theta_k^{t+1}=-1+\mathbf{w}^t\cdot\mathbf{x}^t+a^t\|\mathbf{x}^t\|^2$ and $\theta_{k+1}^{t+1}=\theta_{k+1}^t-C\geq\theta_{k+1}^t-l_{k+1}^t+a^t\|\mathbf{x}^t\|^2=-1+\mathbf{w}^t.\mathbf{x}^t+a^t\|\mathbf{x}^t\|^2$. Thus, $\theta_{k+1}^{t+1}-\theta_k^{t+1}\geq 0$.

Similar arguments can be given for the right support class S_r^t, and hence, we skip the proof for it. ∎

Theorem 4 (Order Preservation of Thresholds Using PA-II): Let $\theta_1^t\leq\cdots\leq\theta_{K-1}^t$ be the thresholds at trial t. Let $\theta_1^{t+1},\ldots,\theta_{K-1}^t$ be the updated thresholds using PA-II. Then, $\theta_1^{t+1}\leq\cdots\leq\theta_{K-1}^{t+1}$.

The order preservation proof for PA-II works similarly as PA algorithm. Thus, PA, PA-I, and PA-II maintain the ordering of the thresholds after every trial.

V. Mistake Bound Analysis

We find the mistake bounds for the proposed PA algorithms under both general and ideal cases. In the ideal case, there exists an ordinal regression function such that for every example, the predicted label lies in the label interval. Let l_i^t be the loss incurred by the algorithm due to ith threshold at trial t. Let l_i^{t*} denote the loss suffered due to ith threshold by the fixed predictor at trial t. The mistake bound of the PA algorithm in general case is as follows.

Theorem 5 (Mistake Bound of PA in General Case): Let $(\mathbf{x}^1,y_l^1,y_r^1)\ldots(\mathbf{x}^T,y_l^T,y_r^T)$ be the sequence of examples to PA algorithm. Let $c=\min_{t\in[T]}(y_r^t-y_l^t)$ and $R^2=\max_{t\in[T]}\|\mathbf{x}^t\|^2$. Let $\mathbf{v}=(\mathbf{u},\mathbf{b})$ be parameters of an arbitrary predictor ($\mathbf{u}\in\mathbb{R}^d$, $\mathbf{b}\in\mathbb{R}^{K-1}$). Let $D=(1+R^2(K-c-1))$, then

$$\sum_{t=1}^T\sum_{i=1}^{K-1}\left(l_i^t\right)^2\leq D^2\left(\|\mathbf{v}\|+4(K-c-1)\sqrt{\sum_{t=1}^T\sum_{i=1}^{K-1}\left(l_i^{t*}\right)^2}\right)^2.$$

Proof: Let l_i^t be the loss incurred by the algorithm due to ith threshold at trial t. Let l_i^{t*} denote the loss suffered due to ith threshold by the fixed predictor at trial t. We define Δ_t as $\Delta_t=\|\mathbf{w}^t-\mathbf{u}\|^2-\|\mathbf{w}^{t+1}-\mathbf{u}\|^2+\|\boldsymbol{\theta}^t-\mathbf{b}\|^2-\|\boldsymbol{\theta}^{t+1}-\mathbf{b}\|^2$. Using the fact that $\mathbf{w}^0=\mathbf{0}$ and $\boldsymbol{\theta}^0=\mathbf{0}$, we get

$$\sum_{i=1}^T\Delta_t=\|\mathbf{w}^0-\mathbf{u}\|^2-\|\mathbf{w}^{T+1}-\mathbf{u}\|^2+\|\boldsymbol{\theta}^0-\mathbf{b}\|^2-\|\boldsymbol{\theta}^{T+1}-\mathbf{b}\|^2\leq\|\mathbf{u}\|^2+\|\mathbf{b}\|^2. \quad (3)$$

This gives an upper bound on the sum of Δ_t. We see that $\theta_i^{t+1}=\theta_i^t,\ \forall i\notin S_l^t\cup S_r^t$. Thus,

$$\Delta_t=-(a^t)^2\|\mathbf{x}^t\|^2-2a^t\mathbf{x}^t\cdot(\mathbf{w}^t-\mathbf{u})-\sum_{i\in S_l^t}\left(\lambda_i^t\right)^2-\sum_{i\in S_r^t}\left(\mu_i^t\right)^2+\sum_{i\in S_l^t}2\lambda_i^t(\theta_i^t-b_i)-\sum_{i\in S_r^t}2\mu_i^t(\theta_i^t-b_i).$$

Note that $\theta_i^t=\mathbf{w}^t.\mathbf{x}^t+l_i^t-1,\ \forall i\in S_l^t$ and $\theta_i^t=1+\mathbf{w}^t.\mathbf{x}^t-l_i^t,\ \forall i\in S_r^t$. Also, note that $-b_i\geq 1-\mathbf{u}.\mathbf{x}^t-l_i^{t*},\forall i\in S_l^t$ and $b_i\geq 1+\mathbf{u}.\mathbf{x}^t-l_i^{t*},\forall i\in S_r^t$. Thus,

$$\Delta_t\geq-(a^t)^2\|\mathbf{x}^t\|^2-\sum_{i\in S_l^t}(\lambda_i^t)^2-\sum_{i\in S_r^t}(\mu_i^t)^2+\sum_{i\in S_l^t}2\lambda_i^t\left(l_i^t-l_i^{t*}\right)+\sum_{i\in S_r^t}2\mu_i^t(l_i^t-l_i^{t*}). \quad (4)$$

Using PA updates and (4), we get

$$\begin{aligned}\Delta_t&=-a_t^2\|\mathbf{x}^t\|^2\left[1+\|\mathbf{x}^t\|^2\left(|S_l^t|+|S_r^t|\right)\right]+\sum_{i\in S_l^t\cup S_r^t}\left(l_i^t\right)^2+\sum_{i\in S_l^t}2\left(a^t\|\mathbf{x}^t\|^2-l_i^t\right)l_i^{t*}-\sum_{i\in S_r^t}2\left(l_i^t+a^t\|\mathbf{x}^t\|^2\right)l_i^{t*}\\&\geq\frac{-\left(\sum_{i\in S_l^t}l_i^t-\sum_{i\in S_r^t}l_i^t\right)^2\|\mathbf{x}^t\|^2}{1+\|\mathbf{x}^t\|^2\left\{|S_l^t|+|S_r^t|\right\}}+\sum_{i\in S_l^t\cup S_r^t}l_i^t\left[l_i^t-2l_i^{t*}\right]-\frac{2\|\mathbf{x}^t\|^2\left(\sum_{i\in S_l^t\cup S_r^t}l_i^t\sum_{j\in S_l^t\cup S_r^t}l_j^{t*}\right)}{1+\|\mathbf{x}^t\|^2\left\{|S_l^t|+|S_r^t|\right\}}\\&\geq-\frac{2\left(1+\|\mathbf{x}^t\|^2\left\{|S_l^t|+|S_r^t|+1\right\}\right)\sum_{i\in S_l^t\cup S_r^t}l_i^t\sum_{j\in S_l^t\cup S_r^t}l_j^{t*}}{1+\|\mathbf{x}^t\|^2\left\{|S_l^t|+|S_r^t|\right\}}+\frac{\sum_{i\in S_l^t\cup S_r^t}\left(l_i^t\right)^2}{1+R^2(K-c-1)}\\&\geq\frac{\sum_{i\in S_l^t\cup S_r^t}\left(l_i^t\right)^2}{D}-4\sum_{i\in S_l^t\cup S_r^t}l_i^t\sum_{j\in S_l^t\cup S_r^t}l_j^{t*}\end{aligned}$$

where $D = 1+R^2(K-c-1)$. Here, we used $\sum_{i\in S_l^t\cup S_r^t} l_i^t l_i^{t*} \le (\sum_{i\in S_l^t\cup S_r^t} l_i^t \sum_{j\in S_l^t\cup S_r^t} l_j^{t*})$ and $(\sum_{i\in S_l^t} l_i^t - \sum_{i\in S_r^t} l_i^t)^2 \le (|S_l^t|+|S_r^t|)\sum_{i\in S_l^t\cup S_r^t}(l_i^t)^2$. Now, using $l_i^t = 0\ \forall i \notin S_l^t \cup S_r^t$ and $\sum_{i\in S_l^t\cup S_r^t} l_i^t \le (|S_l^t|+|S_r^t|)(\sum_{i\in S_l^t\cup S_r^t}(l_i^t)^2)^{1/2}$, we get

$$\Delta_t \ge \frac{\sum_{i=1}^{K-1}(l_i^t)^2}{D} - 4(K-c-1)\sqrt{\sum_{i=1}^{K-1}(l_i^t)^2}\sqrt{\sum_{i=1}^{K-1}(l_i^{t*})^2}.$$

Comparing the upper and lower bounds on $\sum_{t=1}^T \Delta_t$, we get

$$\sum_{t=1}^{T}\sum_{i=1}^{K-1}\left(l_i^t\right)^2 \le D\left(\|\mathbf{v}\|^2 + 4K_1\sum_{t=1}^{T}\sqrt{\sum_{i=1}^{K-1}\left(l_i^t\right)^2}\sqrt{\sum_{i=1}^{K-1}\left(l_i^{t*}\right)^2}\right)$$

where $K_1 = K-c-1$. Using Cauchy–Schwarz inequality, we get $\sum_{t=1}^T(\sum_{i=1}^{K-1}(l_i^t)^2)^{1/2}(\sum_{i=1}^{K-1}(l_i^{t*})^2)^{1/2} \le L_T U_T$ where $L_T = (\sum_{t=1}^T\sum_{i=1}^{K-1}(l_i^t)^2)^{1/2}$ and $U_T = (\sum_{t=1}^T\sum_{i=1}^{K-1}(l_i^{t*})^2)^{1/2}$. Thus, we get $L_T^2 \le D(\|\mathbf{v}\|^2 + 4K_1L_TU_T)$. The upper bound on L_T is obtained by the largest root of the polynomial $L_T^2 - 4K_1DL_TU_T - D\|\mathbf{v}\|^2$, which is $2K_1DU_T + D(4K_1^2U_T^2 + \|\mathbf{v}\|^2)^{1/2}$. Using the fact that $(a+b)^{1/2} \le \sqrt{a}+\sqrt{b}$, we get $L_T \le D\|\mathbf{v}\| + 4K_1DU_T$, which means

$$\sum_{t=1}^{T}\sum_{i=1}^{K-1}\left(l_i^t\right)^2 \le D^2\left(\|\mathbf{v}\| + 4K_1\sqrt{\sum_{t=1}^{T}\sum_{i=1}^{K-1}\left(l_i^{t*}\right)^2}\right)^2.$$

■

We know that $\sum_{i=1}^{K-1}(l_i)^2$ is an upper bound on the MAE. Thus, $\sum_{t=1}^T\sum_{i=1}^{K-1}(l_i^t)^2$ is an upper bound on the number of mistakes in T trials. Thus, $\sum_{t=1}^T L_I^{\text{MAE}}(\mathbf{w}^t\cdot\mathbf{x}^t, \boldsymbol{\theta}^t, y_l^t, y_r^t) \le D^2(\|\mathbf{v}\| + 4K_1(\sum_{t=1}^T\sum_{i=1}^{K-1}(l_i^{t*})^2)^{1/2})^2$. Note that a similar bound is achieved when $c=0$ [5]. Now, we will consider the ideal case.

Corollary 6 (Mistake Bound of PA in Ideal Case): Let $(\mathbf{x}^1, y_l^1, y_r^1)\ldots(\mathbf{x}^T, y_l^T, y_r^T)$ be the sequence of examples presented to PA algorithm. Let $\mathbf{v}^* = (\mathbf{u}^*, \mathbf{b}^*)$ ($\mathbf{u}^* \in \mathbb{R}^d$ and $\mathbf{b}^* \in \mathbb{R}^{K-1}$) be subject to $\mathbf{u}^*\cdot\mathbf{x}^t - b_i^* \ge 1, \forall i \in [y_l^t - 1], \forall t \in [T]$ and $\mathbf{u}^*\cdot\mathbf{x}^t - b_i^* \le -1, \forall i \in \{y_r^t, \ldots, K-1\}, \forall t \in [T]$. Then,

$$\sum_{t=1}^{T}(l_i^t)^2 \le \|\mathbf{v}^*\|^2\left(1 + R^2(K-c-1)\right)$$

where $c = \min_{t\in[T]}(y_r^t - y_l^t)$ and $R^2 = \max_{t\in[T]}\|\mathbf{x}^t\|^2$.

The proof of above can be easily seen by using the bound in Theorem 5 and keeping $l_i^{t*} = 0, \forall t \in [T], \forall i \in [K-1]$. Now, we present the mistake bound for PA-I algorithm.

Theorem 7 (Mistake Bound of PA-I in General Case): Let $(\mathbf{x}^1, y_l^1, y_r^1)\ldots(\mathbf{x}^T, y_l^T, y_r^T)$ be the sequence of examples presented to PA-I algorithm. Let $c = \min_{t\in[T]}(y_r^t - y_l^t)$ and $R^2 = \max_{t\in[T]}\|\mathbf{x}^t\|^2$. Let $\mathbf{v} = (\mathbf{u}, \mathbf{b})$ be the parameters of an arbitrary predictor ($\mathbf{u} \in \mathbb{R}^d$ and $\mathbf{b} \in \mathbb{R}^{K-1}$). Then,

$$\sum_{t=1}^{T}\sum_{i=1}^{K-1} l_i^t \le \sum_{t=1}^{T}\sum_{i=1}^{K-1} l_i^{t*} + \sqrt{DT}\|\mathbf{v}\|$$

where $D = 1 + 2R^2(K-c-1)^2$.

Proof: We use the primal-dual framework proposed in [21], [22] to get the bound. In that framework, online learning is posed as a task of incrementally increasing the dual objective function. The dual optimization problem ($\mathcal{D}$) of the regularized risk under L_I^{IMC} (considering all T examples) is

$$\begin{aligned}\max_{\boldsymbol{\alpha}^1\ldots\boldsymbol{\alpha}^T}\ &\sum_{t=1}^{T}\left(\sum_{j=1}^{y_l^t}\lambda_j^t + \sum_{j=y_r^t}^{K-1}\mu_j^t\right) - \frac{1}{2}\left\|\sum_{t=1}^{T}\left(\sum_{j=1}^{y_l^t}\lambda_j^t - \sum_{j=y_r^t}^{K-1}\mu_j^t\right)\mathbf{x}^t\right\|^2 \\ &- \frac{1}{2}\sum_{j=1}^{K-1}\left(\sum_{t=1}^{T}\left(\mu_j^t\mathbb{I}_{\{j\ge y_r^t\}} - \lambda_j^t\mathbb{I}_{\{j\le y_l^t-1\}}\right)\right)^2 \\ \text{s.t. } &0 \le \lambda_j^t \le C, \quad t\in[T],\ j = 1\ldots y_l^t - 1 \\ &0 \le \mu_j^t \le C, \quad t\in[T],\ j = y_r^t\ldots K-1\end{aligned}$$

where $\boldsymbol{\alpha}^t = [\lambda_1^t\ \ldots\ \lambda_{y_l^t-1}^t\ 0\ \ldots\ 0\ \mu_{y_r^t}\ \ldots\ \mu_{K-1}^t] \in \mathbb{R}^{K-1}$. Let $\Omega = (\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^T)$. PA-I can be viewed as finding a sequence of $\Omega^1, \ldots, \Omega^{T+1}$ where $\Omega^{t+1} = (\boldsymbol{\alpha}_{t+1}^1, \ldots, \boldsymbol{\alpha}_{t+1}^T)$ is the maximizer of the following problem:

$$\max_{\Omega}\ \mathcal{D}(\Omega)\quad \text{s.t. } \boldsymbol{\alpha}^s = \mathbf{0}\quad \forall s > t.$$

PA-I updates are as follows. $\boldsymbol{\alpha}_{t+1}^i = \boldsymbol{\alpha}_t^i,\ \forall i \ne t$. $\boldsymbol{\alpha}_{t+1}^t = [\lambda_1^t\ \ldots\ \lambda_{y_l^t-1}^t\ 0\ \ldots\ 0\ \mu_{y_r^t}^t\ \ldots\ \mu_{K-1}^t]$ where $\lambda_i^t = \min(C, l_i^t - a^t\|\mathbf{x}^t\|^2),\ i = 1\ldots y_l^t - 1$ and $\mu_i^t = \min(C, l_i^t + a^t\|\mathbf{x}^t\|^2),\ i = y_r^t\ldots K-1$. Increment in $\mathcal{D}$ after trial t is

$$\begin{aligned}&\mathcal{D}(\Omega^{t+1}) - \mathcal{D}(\Omega^t) \\ &= -\frac{1}{2}\left(\sum_{i=1}^{y_l^t-1}\lambda_i^t - \sum_{i=y_r^t}^{K-1}\mu_i^t\right)^2\|\mathbf{x}^t\|^2 - \frac{1}{2}\sum_{i=1}^{y_l^t-1}\left(\lambda_i^t\right)^2 \\ &\quad - \frac{1}{2}\sum_{i=y_r^t}^{K-1}\left(\mu_i^t\right)^2 + \sum_{i=1}^{y_l^t-1}\lambda_i^t\left(1 - \mathbf{w}^t\cdot\mathbf{x}^t + \theta_i^t\right) \\ &\quad + \sum_{i=y_r^t}^{K-1}\mu_i^t\left(1 + \mathbf{w}^t\cdot\mathbf{x}^t - \theta_i^t\right)\end{aligned}$$

where $\theta_i^t = \sum_{s=1}^{t-1}(\mu_i^s\mathbb{I}_{\{i\ge y_r^s\}} - \lambda_i^s\mathbb{I}_{\{i\le y_l^s-1\}}),\ i \in [K-1]$ and $\mathbf{w}^t = \sum_{s=1}^{t-1}a^s\mathbf{x}^s$. Note that $\lambda_i^t > 0,\ i \in S_l^t$ and $\mu_i^t > 0,\ i \in S_r^t$. Using $a^t = \sum_{i\in S_l^t}\lambda_i^t - \sum_{i\in S_r^t}\mu_i^t$, we get

$$\begin{aligned}&\mathcal{D}(\Omega^{t+1}) - \mathcal{D}(\Omega^t) \\ &= \sum_{i\in S_l^t}\lambda_i^t\left(l_i^t - a^t\|\mathbf{x}^t\|^2 - \frac{\lambda_i^t}{2}\right) \\ &\quad + \sum_{i\in S_r^t}\mu_i^t\left(l_i^t + a^t\|\mathbf{x}^t\|^2 - \frac{\mu_i^t}{2}\right) \\ &\quad + \frac{1}{2}a^t\|\mathbf{x}^t\|^2\left(\sum_{i\in S_l^t}\lambda_i^t - \sum_{i\in S_r^t}\mu_i^t\right) \\ &\ge C\left[\sum_{i\in S_l^t}\gamma\left(l_i^t - a^t\|\mathbf{x}^t\|^2\right) + \sum_{i\in S_r^t}\gamma\left(l_i^t + a^t\|\mathbf{x}^t\|^2\right)\right] \quad (5)\end{aligned}$$

where $\gamma(z) = (1/C)(\min(z, C)(z - (1/2\min(z, C)))$ [22]. Note that $\mathcal{D}(\Omega^0) = 0$. Summing (5) from $t = 1$ to T,

we get

$$\begin{aligned}
&\mathcal{D}(\Omega^{T+1}) \\
&= \sum_{t=1}^{T} \Big(\mathcal{D}(\Omega^{t+1}) - \mathcal{D}(\Omega^{t}) \Big) \\
&\geq C \sum_{t=1}^{T} \left(\sum_{i \in S_l^t} \gamma \left(l_i^t - a^t \|\mathbf{x}^t\|^2 \right) + \sum_{i \in S_r^t} \gamma \left(l_i^t + a^t \|\mathbf{x}^t\|^2 \right) \right) \\
&\geq CT\gamma \left(\frac{1}{T} \sum_{t=1}^{T} \left(\sum_{i \in S_l^t} \left(l_i^t - a^t \|\mathbf{x}^t\|^2 \right) + \sum_{i \in S_r^t} \left(l_i^t + a^t \|\mathbf{x}^t\|^2 \right) \right) \right)
\end{aligned}$$

where we used the fact that $\gamma(.)$ is a convex function [22]. From the weak duality, we get the following:

$$\mathcal{D}(\Omega^{T+1}) \leq \frac{1}{2} \left(\|\mathbf{w}\|^2 + \|\boldsymbol{\theta}\|^2 \right) + C \sum_{t=1}^{T} \left(\sum_{i=1}^{y_l^t - 1} l_i^{t*} + \sum_{i=y_r^t}^{K-1} l_i^{t*} \right).$$

Comparing the upper and lower bounds on $\mathcal{D}(\Omega^{T+1})$, we get

$$\begin{aligned}
&\frac{1}{T} \sum_{t=1}^{T} \left(\sum_{i \in S_l^t} \left(l_i^t - a^t \|\mathbf{x}^t\|^2 \right) + \sum_{i \in S_r^t} \left(l_i^t + a^t \|\mathbf{x}^t\|^2 \right) \right) \\
&\leq \gamma^{-1} \left(\frac{1}{2TC} \left(\|\mathbf{w}\|^2 + \|\boldsymbol{\theta}\|^2 \right) + \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{i=1}^{y_l^t - 1} l_i^{t*} + \sum_{i=y_r^t}^{K-1} l_i^{t*} \right) \right).
\end{aligned}$$

We note that

$$\begin{aligned}
&\frac{1}{T} \sum_{t=1}^{T} \left(\sum_{i \in S_l^t} \left(l_i^t - a^t \|\mathbf{x}^t\|^2 \right) + \sum_{i \in S_r^t} \left(l_i^t + a^t \|\mathbf{x}^t\|^2 \right) \right) \\
&\quad \geq \frac{1}{T} \sum_{t=1}^{T} \sum_{i \in S_l^t \cup S_r^t} l_i^t - \frac{1}{T} \sum_{t=1}^{T} a^t \|\mathbf{x}^t\|^2 \left(\left| S_l^t \right| + \left| S_r^t \right| \right) \\
&\quad \geq \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{K-1} l_i^t - CR^2 (K - c - 1)^2 \qquad (6)
\end{aligned}$$

where we used the fact that $\|\mathbf{x}^t\|^2 \leq R^2, \ \forall t \in [T]$, $|S_l^t| + |S_r^t| \leq K - c - 1, \ \forall t \in [T]$, and $a^t \leq C(K - c - 1), \ \forall t \in [T]$. From [22], we know that $\gamma^{-1}(z) \leq z + (1/2)C$. Thus,

$$\begin{aligned}
&\gamma^{-1} \left(\frac{1}{2CT} \left(\|\mathbf{w}\|^2 + \|\boldsymbol{\theta}\|^2 \right) + \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{i=1}^{y_l^t - 1} l_i^{t*} + \sum_{i=y_r^t}^{K-1} l_i^{t*} \right) \right) \\
&\leq \frac{1}{2CT} \left(\|\mathbf{w}\|^2 + \|\boldsymbol{\theta}\|^2 \right) + \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{i=1}^{y_l^t - 1} l_i^{t*} + \sum_{i=y_r^t}^{K-1} l_i^{t*} \right) + \frac{C}{2}. \qquad (7)
\end{aligned}$$

Using (6) and (7), we get

$$\sum_{t=1}^{T} \sum_{i=1}^{K-1} \left[l_i^t - l_i^{t*} \right] \leq \frac{1}{2C} \|\mathbf{v}\|^2 + CT \left[\frac{1}{2} + R^2 (K - c - 1)^2 \right].$$

We get least upper bound by putting $C = (\|\mathbf{v}\|)/((T(1 + 2R^2 (K - c - 1)^2))^{1/2})$ as follows:

$$\sum_{t=1}^{T} \sum_{i=1}^{K-1} l_i^t \leq \sum_{t=1}^{T} \sum_{i=1}^{K-1} l_i^{t*} + \sqrt{T(1 + 2R^2 (K - c - 1)^2)} \|\mathbf{v}\|.$$

■

Corollary 8 (Mistake Bound of PA-I in Ideal Case): Let $(\mathbf{x}^1, y_l^1, y_r^1) \ldots (\mathbf{x}^T, y_l^T, y_r^T)$ be the sequence of examples presented to PA-I. Let $c = \min_{t \in [T]} (y_r^t - y_l^t)$ and $R^2 = \max_{t \in [T]} \|\mathbf{x}^t\|^2$. Let $\mathbf{v}^* = (\mathbf{u}^*, \mathbf{b}^*)$ ($\mathbf{u}^* \in \mathbb{R}^d$ and $\mathbf{b}^* \in \mathbb{R}^{K-1}$) be such that $\mathbf{u}^*.\mathbf{x}^t - b_i^* \geq 1, \forall i \in [y_l^t - 1], \forall t \in [T]$ and $\mathbf{u}^*.\mathbf{x}^t - b_i^* \leq -1, \forall i \in \{y_r^t, \ldots, K - 1\}, \forall t \in [T]$. Let $D = 1 + 2R^2 (K - c - 1)^2$, then,

$$\sum_{t=1}^{T} \sum_{i=1}^{K-1} l_i^t \leq \sqrt{DT} \|\mathbf{v}\|.$$

The proof of above Corollary is immediate from Theorem 8 by putting $l_i^{t*} = 0, \ \forall t \in [T], \ \forall i \in [K - 1]$.

Theorem 9 (Mistake Bound of PA-II in General Case): Let $(\mathbf{x}^1, y_l^1, y_r^1), \ldots (\mathbf{x}^T, y_l^T, y_r^T)$ be the sequence of examples presented to PA-II. Let $c = \min_{t \in [T]} (y_r^t - y_l^t)$ and $R^2 = \max_{t \in [T]} \|\mathbf{x}^t\|^2$. Let $\mathbf{v} = (\mathbf{u}, \mathbf{b})(\mathbf{u} \in \mathbb{R}^d, \ \mathbf{b} \in \mathbb{R}^{K-1})$ be the parameters of an arbitrary predictor. Then,

$$\sum_{t=1}^{T} \sum_{i=1}^{K-1} (l_i^t)^2 \leq D \left(\|\mathbf{v}\|^2 + 2C \sum_{t=1}^{T} \sum_{i=1}^{K-1} \left(l_i^{t*} \right)^2 \right)$$

where $D = 1 + (1/2C) + R^2 (K - c - 1)$.

Proof: Let $\alpha = (1/\sqrt{2C})$. Then,

$$\begin{aligned}
\Delta_t &\geq -(a^t)^2 \|\mathbf{x}^t\|^2 - \sum_{i \in S_l^t} \left(\lambda_i^t \right)^2 - \sum_{i \in S_r^t} \left(\mu_i^t \right)^2 \\
&\quad + \sum_{i \in S_l^t} 2\lambda_i^t \left(l_i^t - l_i^{t*} \right) + \sum_{i \in S_r^t} 2\mu_i^t \left(l_i^t - l_i^{t*} \right) \\
&\geq -(a^t)^2 \|\mathbf{x}^t\|^2 - \sum_{i \in S_l^t} \left(\lambda_i^t \right)^2 - \sum_{i \in S_r^t} \left(\mu_i^t \right)^2 \\
&\quad + \sum_{i \in S_l^t} 2\lambda_i^t \left(l_i^t - l_i^{t*} \right) + \sum_{i \in S_r^t} 2\mu_i^t (l_i^t - l_i^{t*} \\
&\quad - \sum_{i \in S_l^t} \left(\alpha \lambda_i^t - \frac{l_i^{t*}}{\alpha} \right)^2 - \sum_{i \in S_r^t} \left(\alpha \mu_i^t - \frac{l_i^{t*}}{\alpha} \right)^2
\end{aligned}$$

where we used the following inequality:

$$\left(\sum_{i \in S_l^t} l_i^t - \sum_{i \in S_r^t} l_i^t \right)^2 \leq \left(\left| S_l^t \right| + |S_r^t| \right) \sum_{i \in S_l^t \cup S_r^t} (l_i^t)^2.$$

By simplifying further, we get

$$\begin{aligned}
\Delta_t &\geq -(a^t)^2 \|\mathbf{x}^t\|^2 - \left(1 + \frac{1}{2C} \right) \left(\sum_{i \in S_l^t} \left(\lambda_i^t \right)^2 + \sum_{i \in S_r^t} \left(\mu_i^t \right)^2 \right) \\
&\quad - 2C \left(\sum_{i \in S_l^t} \left(l_i^{t*} \right)^2 + \sum_{i \in S_r^t} \left(l_i^{t*} \right)^2 \right) \\
&\quad + 2 \left(\sum_{i \in S_l^t} \lambda_i^t l_i^t + \sum_{i \in S_r^t} \mu_i^t l_i^t \right) \\
&= 2 \sum_{i \in S_l^t} \lambda_i^t l_i^t + 2 \sum_{i \in S_r^t} \mu_i^t l_i^t - (a^t)^2 \|\mathbf{x}^t\|^2 \\
&\quad - 2C \sum_{i \in S_l^t \cup S_r^t} \left(l_i^{t*} \right)^2 - \sum_{i \in S_l^t} \frac{\left(a^t \|\mathbf{x}^t\|^2 - l_i^t \right)^2}{1 + \frac{1}{2C}} \\
&\quad - \sum_{i \in S_r^t} \frac{\left(l_i^t + a^t \|\mathbf{x}^t\|^2 \right)^2}{1 + \frac{1}{2C}}
\end{aligned}$$

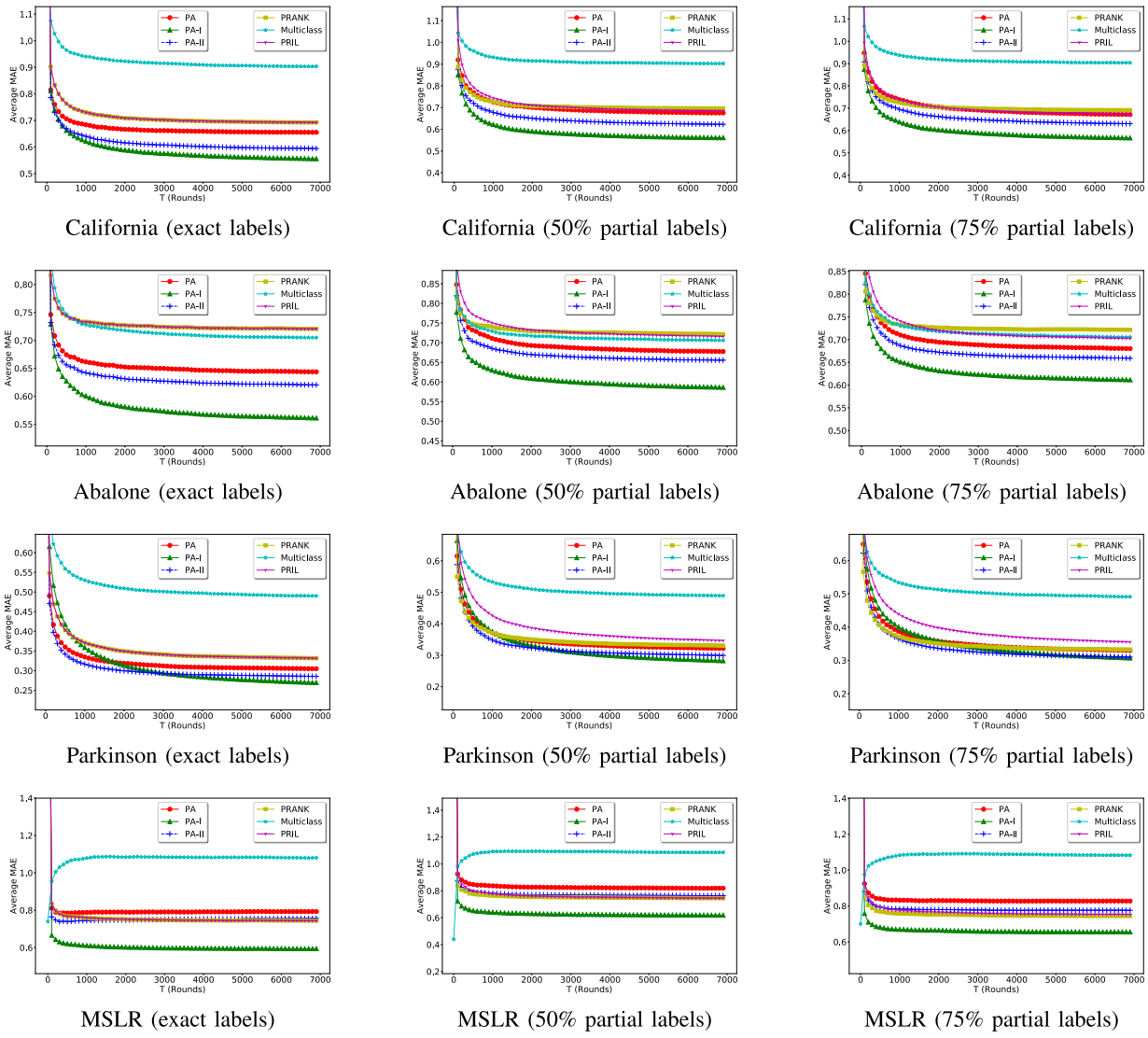


Fig. 1. Comparison results of PA, PA-I, and PA-II with PRIL, MCP, and PRank. The performance measure used is average MAE.

$$\geq \frac{\sum_{i\in S_l^t\cup S_r^t}\left(l_i^t\right)^2}{1+\frac{1}{2C}+R^2(K-c-1)}-2C\sum_{i\in S_l^t\cup S_r^t}\left(l_i^{t*}\right)^2.$$

Let $D=1+(1/2C)+R^2(K-c-1)$, then by comparing the lower and upper bounds on $\sum_{t=1}^{T}\Delta_t$, we get

$$\sum_{t=1}^{T}\sum_{i\in S_l^t\cup S_r^t}\left(l_i^t\right)^2 \leq D\left(\|\mathbf{v}\|^2+2C\sum_{t=1}^{T}\sum_{i\in S_l^t\cup S_r^t}\left(l_i^{t*}\right)^2\right)$$
$$\leq D\left(\|\mathbf{v}\|^2+2C\sum_{t=1}^{T}\sum_{i=1}^{K-1}\left(l_i^{t*}\right)^2\right).$$

We know that $l_i^t=0,\ \forall i\notin S_l^t\cup S_r^t$. Thus,

$$\sum_{t=1}^{T}\sum_{i=1}^{K-1}(l_i^t)^2 \leq D\left(\|\mathbf{v}\|^2+2C\sum_{t=1}^{T}\sum_{i=1}^{K-1}\left(l_i^{t*}\right)^2\right).$$

■

Corollary 10 (Mistake Bound of PA-II in Ideal Case): Let $(\mathbf{x}^1, y_l^1, y_r^1),\ldots(\mathbf{x}^T, y_l^T, y_r^T)$ be the sequence of examples. Let $\mathbf{v}^*=(\mathbf{u}^*,\mathbf{b}^*)$ $(\mathbf{u}^*\in\mathbb{R}^d,\ \mathbf{b}^*\in\mathbb{R}^{K-1})$ be the parameters of an ideal predictor such that $\mathbf{u}^*.\mathbf{x}^t-b_i^*\geq 1,\forall i\in[y_l^t-1],\forall t\in[T]$ and $\mathbf{u}^*\cdot\mathbf{x}^t-b_i^*\leq -1,\forall i\in\{y_r^t,\ldots,K-1\},\forall t\in[T]$. Let $c=\min_{t\in[T]}(y_r^t-y_l^t)$ and $R^2=\max_{t\in[T]}\|\mathbf{x}^t\|^2$. Then, for PA-II updates, we get the following bound:

$$\sum_{t=1}^{T}\sum_{i=1}^{K-1}(l_i^t)^2 \leq \left(1+\frac{1}{2C}+R^2(K-c-1)\right)\|\mathbf{v}\|^2.$$

VI. Experiments

In this section, we describe the experiments performed.

A. Data Sets Used

We perform experiments on the following four data sets. The features in each of the data set are normalized to zero mean and unit variance coordinate wise.

1) *California:* This data set has 20 460 instances with nine features [17]. The target variable "median house value"

ranges over 14999–500001. We create five intervals, i.e., (1–100000), (100001–200000), (200001–300000), (300001–400000), and (400001–500001).

2) *Abalone:* This data set [7] has 4177 instances with eight attributes. The target attribute varies from 1 to 29. We divide it into four intervals as 1–7, 8–9, 10–12, and 13–29.
3) *Parkinson Telemonitoring:* This data set [7] contains 5875 instances with 22 features. The target variable "total_UPDRS" for varies from 7 to 54.992. We divide it into four intervals, i.e., 7–17, 18–27, 28–37, and 38–54.992.
4) *MSLR:* This data set comprises query-url pairs along with the relevance label obtained from Microsoft Bing [19]. We experiment on MSLR-WEB10K data in which we took one of the available fivefolds. There are 723 412 instances, 136 features, and 5 classes.

B. Generating Interval Labels

We choose $m\%$ examples from the training set randomly. Then, for each of the example, we randomly assign one of the following interval: $[y-1, y]$, $[y-1, y+1]$, $[y, y+1]$, $[y-2, y]$, $[y, y+2]$, $[y-2, y+2]$ where y is the actual label. We consider $m = 50\%$ and 75%.

C. Comparison Results With Other Approaches

We compare the performance of the proposed PA algorithms with two approaches. 1) PRank [5], which is an online ranking algorithm using the actual labels; 2) PRIL in [15], which is perceptron-based approach for ordinal regression using interval labels; and 3) multi-class perceptron (MCP) [6] (uses more parameters and ignores the class labels orderings).

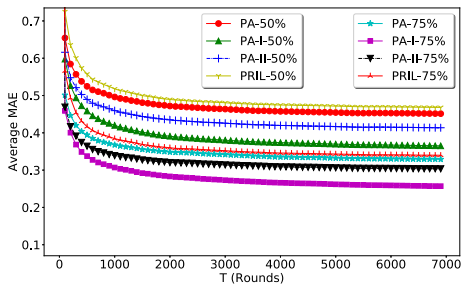
For PRank and MCP, we use only the exact labels for training. For the proposed PA algorithms and PRIL, we use interval labeled data during training. We took three different training sets for the proposed PA algorithms and PRIL. First, with 50% interval labels, second with 75% interval labels and third with actual (exact) labels. To predict the label, we use the ranking function descibed in (1).

We used the exact labels to compute the average L_{MAE} (after every trial) for all the algorithms. We repeat the process 100 times and average the instantaneous losses across the 100 runs. Fig. 1 shows the comparison results.

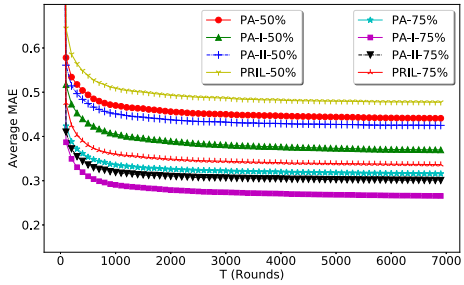
1) For California and Abalone data sets, PA, PA-I, and PA-II trained using exact labels as well as using interval labels outperform the other algorithms.
2) For Parkinson's, PA, PA-I, and PA-II outperform other approaches for exact label case and 50% interval label case. For 75% interval labels, PA-I and PA-II outperform PRIL, PRank, and MCP, while PA performs comparably to PRank and better than PRIL and MCP.
3) For MSLR, PA-I outperforms PRIL, PRank, and MCP in all the cases. Also, PA and PA-II always outperform MCP. PA-II performs comparably to PRIL and PRank.

The major difference between PRIL and the proposed PA algorithms is that PRIL uses a constant step size. On the other hand, PA algorithms choose an appropriate step size to find the new parameters in every trial. In each of the PA variants, the step size is determined by solving an optimization problem. This makes PA algorithms to have better step size selection. Thus, we see that the proposed PA algorithms perform better or comparable to PRIL, PRank, and MCP.

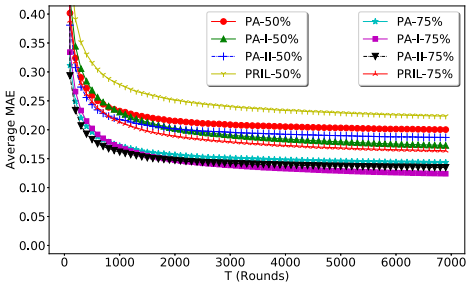
Among the PA variants, PA takes the most aggressive step size to ensure that the loss L_{IMC} on the current example becomes zero. Because of this greedy nature, for some data sets, we see that PA does not perform well compared to other



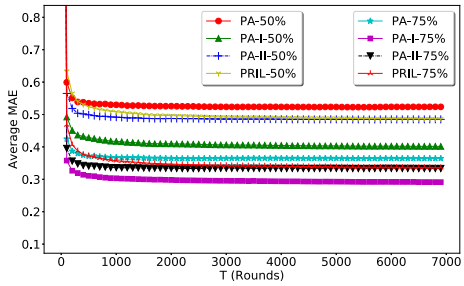
California (50 % and 75% partial labels)



Abalone (50 % and 75% partial labels)



Parkinson (50 % and 75% partial labels)



MSLR (50 % and 75% partial labels)

Fig. 2. Average MAE decreases by increasing the fraction of interval labels. MAE is computed using partial labels.

approaches. On the other hand, PA-I allows small errors and controls the greediness by finding the new parameters closest to the old one while minimizing the error. Due to which PA-I outperforms PA, PA-II also allows some errors but minimizes the square of the error. In general, minimizing the error is better than minimizing the square of the error. Thus, PA-II performs better than PA, and PA-I performs better than both PA and PA-II.

D. Varying the Fraction of Interval Labels

We vary the fraction of partial labels (50% and 75%). We compute the average MAE after every trial with the same interval label used for updating the hypothesis. We repeat the process 100 times and average the instantaneous losses across the 100 runs. The results are shown in Fig. 2. We make the following observations.

1) We see that for all the data sets, the average MAE decreases faster compared to T.
2) Also, the average MAE decreases with the increase in the fraction of interval labels. This happens because the allowed range for predicted rank is more when we use interval labels for computing MAE.
3) For California, Abalone, and Parkinsons data sets, the proposed algorithms such as PA, PA-I, and PA-II achieve smaller values of average MAE compared to PRIL. For MSLR data set, PA-I outperforms PRIL, whereas PA-II performs comparably to PRIL. Thus, overall, PA approaches perform better than PRIL.

VII. Conclusion

We proposed online PA algorithms for ordinal regression, which can also be used when we have interval labels. We presented three algorithms, namely, PA, PA-I, and PA-II. We find the exact solution of the optimization problem at every trial. Our method is based on finding the support classes at each instant using the SCAs. These sets describe the thresholds to be updated at a trial. Advantage of our method is that the ordering of the thresholds is maintained implicitly. We have also given mistake bounds on all the three variants of the algorithm. Practical experiments show that our proposed algorithms perform better than the other algorithms even when we train our algorithms using interval labels.

References

[1] K. Antoniuk, V. Franc, and V. Hlavac, "Interval insensitive loss for ordinal classification," in *Proc. 6th Asian Conf. Mach. Learn.*, Nha Trang, Vietnam, vol. 39, Nov. 2015, pp. 189–204.

[2] W. Chen *et al.*, "Ranking measures and loss functions in learning to rank," in *Proc. Adv. Neural Inf. Process. Syst. 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. 2009, pp. 315–323.

[3] W. Chu and S. S. Keerthi, "New approaches to support vector ordinal regression," in *Proc. 22nd Int. Conf. Mach. Learn. (ICML)*, 2005, pp. 145–152.

[4] K. Crammer *et al.*, "Online passive-aggressive algorithms," *J. Mach. Learn. Res.*, vol. 7, pp. 551–585, Dec. 2006.

[5] K. Crammer and Y. Singer, "Pranking with ranking," in *Proc. 14th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2001, pp. 641–647.

[6] K. Crammer and Y. Singer, "Ultraconservative online algorithms for multiclass problems," *J. Mach. Learn. Res.*, vol. 3, pp. 951–991, Mar. 2003.

[7] D. Dua and C. Graff, "UCI Machine Learning Repository," School Inf. Comput. Sci., Univ. California, Irvine, CA, USA, 2019. [Online]. Available: http://archive.ics.uci.edu/ml

[8] O. M. Doyle *et al.*, "Predicting progression of Alzheimer's disease using ordinal regression," *PLoS ONE*, vol. 9, no. 8, Aug. 2014, Art. no. e105542.

[9] B. Gu *et al.*, "Incremental support vector learning for ordinal regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 7, pp. 1403–1416, Jul. 2015.

[10] A. Guisan and F. E. Harrell, "Ordinal response regression models in ecology," *J. Vegetation Sci.*, vol. 11, no. 5, pp. 617–626, 2000.

[11] E. F. Harrington, "Online ranking/collaborative filtering using the perceptron algorithm," in *Proc. 12th Int. Conf. Int. Conf. Mach. Learn. (ICML)*, 2003, pp. 250–257.

[12] R. Herbrich, T. Graepel, and K. Obermayer, "Support vector learning for ordinal regression," in *Proc. Int. Conf. Artif. Neural Netw.*, 1999, pp. 97–102.

[13] H. Li, "A short introduction to learning to rank," *IEICE Trans. Inf. Syst.*, vol. E94.D, no. 10, pp. 1854–1862, 2011.

[14] L. Li and H.-T. Lin, "Ordinal regression by extended binary classification," in *Proc. 19th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2006, pp. 865–872.

[15] N. Manwani, "PRIL: Perceptron ranking using interval labels," in *Proc. ACM India Joint Int. Conf. Data Sci. Manage. Data (CoDS-COMAD)*, 2019, pp. 78–85.

[16] S. Matsushima *et al.*, "Exact passive-aggressive algorithm for multiclass classification using support class," in *Proc. SDM*, 2010, pp. 303–314.

[17] R. K. Pace and R. Barry, "Sparse spatial autoregressions," *Statist. Probab. Lett.*, vol. 33, pp. 291–297, May 1997.

[18] F. Pedregosa, F. Bach, and A. Gramfort, "On the consistency of ordinal regression methods," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 1769–1803, Jan. 2017.

[19] T. Qin and T.-Y. Liu, "Introducing LETOR 4.0 datasets," 2013, *arXiv:1306.2597*. [Online]. Available: https://arxiv.org/abs/1306.2597

[20] J. D. M. Rennie and N. Srebro, "Loss functions for preference levels: Regression with discrete ordered labels," in *Proc. IJCAI Multidiscipl. Workshop Adv. Preference Handling*, 2005, pp. 180–186.

[21] S. Shalev-Shwartz and Y. Singer, "Online learning meets optimization in the dual," in *Learning Theory*, G. Lugosi and H. U. Simon, Eds. 2006, pp. 423–437.

[22] S. Shalev-Shwartz and Y. Singer, "A primal-dual perspective of online learning algorithms," *Mach. Learn.*, vol. 69, nos. 2–3, pp. 115–142, 2007.

[23] A. Shashua and A. Levin, "Ranking with large margin principle: Two approaches," in *Proc. 15th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2002, pp. 961–968.

[24] H. Valizadegan *et al.*, "Learning to rank by optimizing NDCG measure," in *Proc. 22nd Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2009, pp. 1883–1891.

Naresh Manwani received the B.E. degree in electronics and communication from Rajasthan University, Jaipur, India, in 2003, the M.Tech. degree in information and communication technology from DAIICT, Gandhinagar, India, in 2006, and the Ph.D. degree in machine learning from IISc, Bengaluru, India, in 2013.

After his Ph.D., he was with the Microsoft Research Lab, Bengaluru, the GE Global Research Centre, Bengaluru, and Microsoft India R&D Lab, Bengaluru. In November 2016, he joined the International Institute of Information Technology (IIIT), Hyderabad, India, as an Assistant Professor with the Machine Learning Laboratory. His current research interests include online learning, reinforcement learning, statistical learning theory, and deep learning.

Mohit Chandra is currently pursuing the master's degree in computer science and engineering with the International Institute of Information Technology (IIIT), Hyderabad, India. His ongoing thesis focuses on the analysis of abusive behavior in online communities.

His current research interests include machine learning algorithms, natural language processing with deep learning, and information retrieval and extraction.