Optimal Threshold Labeling for Ordinal Regression Methods

Anonymous authors
Paper under double-blind review

Abstract

For an ordinal regression task, a classification task for ordinal data, one-dimensional transformation (1DT)-based methods are often employed since they are considered to capture the ordinal relation of ordinal data well. They learn a 1DT of the observation of the explanatory variables so that an observation with a larger class label tends to have a larger value of the 1DT, and classify the observation by labeling that learned 1DT. In this paper, we study the labeling procedure for 1DT-based methods, which have not been sufficiently discussed in existing studies. While regression-based methods and classical threshold methods conventionally use threshold labelings, which label a learned 1DT according to the rank of the interval to which the 1DT belongs among intervals on the real line separated by threshold parameters, we prove that likelihood-based labeling used in popular statistical 1DT-based methods is also a threshold labeling in typical usages. Moreover, we show that these threshold labelings can be sub-optimal ones depending on the learning result of the 1DT and the task under consideration. On the basis of these findings, we propose to apply empirical optimal threshold labeling, which is a threshold labeling that uses threshold parameters minimizing the empirical task risk for a learned 1DT, to those methods. In experiments with real-world datasets, changing the labeling procedure of existing 1DT-based methods to the proposed one improved the classification performance in many tried cases.

1 Introduction

Ordinal regression (OR) (or called ordinal classification) is the classification of ***ordinal data*** in which the underlying target variable is categorical and labeled from a label set (ordinal scale) that is equipped with a natural ordinal relation for the explanatory variables; see Section 2.1 for a detailed formulation. The ordinal scale is typically formed as a graded (interval) summary of objective indicators like age groups {'0–9', '10–19', ..., '90–99', '100–'} or graded evaluation of subjectivity like human rating {'excellent', 'good', 'average', 'bad', 'terrible'}, and ordinal data appear in various practical applications: age estimation (Niu et al., 2016; Cao et al., 2020), information retrieval (Liu, 2011), movie rating (Yu et al., 2006), and questionnaire survey in social research (Bürkner & Vuorre, 2019).

One-dimensional transformation (1DT)-based methods are often applied to the OR problems as a simple way to capture the ordinal relation of ordinal data: they learn a 1DT of the observation of the explanatory variables so that an observation with a larger class label tends to have a larger value of the 1DT, and classify the observation by labeling that learned 1DT, as we will formalize in Section 2.2. For example, ***regression-based methods*** and ***classical threshold methods*** (or called threshold models) conventionally use a ***threshold labeling***, which labels a learned 1DT according to the rank of the interval to which the 1DT belongs among intervals on the real line separated by ***threshold parameters***. Regression-based methods (Kramer et al., 2001; Agarwal, 2008) learn a 1DT by solving a naïve regression task that infers a class label by the 1DT in a continuous scale, and often apply ***nearest-neighbor threshold (NNT) labeling*** that rounds the learned 1DT to its nearest label (see Section 3.1). Classical threshold methods (Shashua & Levin, 2003; Lin & Li, 2006; Chu & Keerthi, 2007; Lin & Li, 2012; Li & Lin, 2007; Pedregosa et al., 2017) learn a 1DT using an objective function different from the empirical task risk, and commonly use ***minimum threshold (MT)*** and ***summation threshold (ST) labelings*** that apply parameters obtained incidentally in that learning procedure as threshold parameters (see Sections 3.2). Also, statistical 1DT-based methods

(***statistical methods***) (McCullagh, 1980; Williams, 2006; Cao et al., 2020; Yamasaki, 2022), in which the learning procedure of the 1DT can be viewed as statistical modeling of conditional probabilities of the data, can apply ***likelihood-based (LB) labeling*** that is designed to minimize the task risk under the expectation that the assumed likelihood model is correctly specified to the data distribution (see Section 3.3).

We, however, have respective concerns on these existing labeling procedures. First, threshold parameters of NNT, MT, and ST labelings are generally not designed to minimize the task risk. So their threshold parameters can be sub-optimal for the minimization of the task risk, as we will demonstrate in Example 1. On the other hand, the LB labeling is designed to minimize the task risk if the assumed likelihood model is correctly specified to the distribution of the data (see Theorem 2). However, 1DT-based likelihood models have a strongly restricted representation ability and can be mis-specified to the data, and hence the LB labeling can degrade the classification performance depending on the data distribution.

Previous studies have done little theoretical work on the properties of these existing labelings. Therefore, we first study the relationship between these labelings. In particular, we show in Theorem 3 that not only the NNT, MT, and ST labelings but also the LB labeling in typical usages is a threshold labeling. This finding motivates us to search for a better labeling function among the class of threshold labelings. Then, in Section 4, we propose to apply ***empirical optimal threshold (EOT) labeling***, which is a threshold labeling that uses threshold parameters minimizing the empirical task risk for the learned 1DT, to 1DT-based methods, under the expectation that the 1DT is learned successfully and the empirical (training) task risk becomes a good estimate of the (test) task risk. Here, the threshold parameters for the EOT labeling can be computed with a computational complexity of quasi-linear order $O(n \log n)$ regarding the training sample size n using a dynamic programming-based algorithm (Algorithm 1) mentioned in Lin & Li (2006) (see Section 5 for the relation of our proposal to several previous studies including this reference).

In this study, we further performed numerical experiments of the OR task for real-world ordinal data to confirm the practical effectiveness of the EOT labeling (see Section 6). The EOT labeling gave superior generalization performance (more exactly, smaller test task risk) than the NNT, MT, ST, and LB labelings, in many tried cases. Also, a modified 1DT-based method with the EOT labeling outperformed an existing 1DT-based method using the ST labeling that has been declared by Cao et al. (2020) to be state-of-the-art in 2020 for the age estimation from the facial image.

Therefore, in this paper, we propose to change labeling procedures of (existing) 1DT-based OR methods to the EOT labeling, on the ground of the fact (see Example 1 and Theorem 3) that the NNT, MT, and ST labelings and LB labeling in typical usages are possibly sub-optimal threshold labelings, and empirical effectiveness of the EOT labeling.

2 Preliminaries

2.1 Formulation of OR Problem

OR is the classification of ordinal data. The ordinal data have an underlying categorical ***target variable*** $Y \in [K] \coloneqq \{1, \dots, K\}$ that is equipped with an ordinal relation naturally interpretable in the relationship with ***explanatory variables*** $\boldsymbol{X} \in \mathbb{R}^d$, where d and K are supposed to be integers larger than or equal to 1 and 3, respectively.[1] We here suppose that the target class labels are encoded to $1, \dots, K$ in an order-preserving manner, like from 'excellent', ..., 'terrible' to $1, \dots, 5$.

The task of the OR (***OR task***) is basically the same as that of the standard (including cost-sensitive) classification, to obtain a good ***classifier*** $f : \mathbb{R}^d \to [K]$. For a user-specified ***task loss function*** $\ell : [K]^2 \to [0, \infty)$, it is formulated as minimization of the ***task risk*** $\mathbb{E}[\ell(f(\boldsymbol{X}), Y)]$, where the expectation value $\mathbb{E}[\cdot]$ is taken for all random variables in its argument (here $\boldsymbol{X}$ and Y). Popular task losses for OR tasks include not only the ***zero-one task loss*** $\ell_{\mathrm{zo}}(j, k) \coloneqq \mathbb{1}(j \neq k)$, where $\mathbb{1}(c)$ takes 1 if a condition c is true and 0 otherwise, but also V-shaped losses (for cost-sensitive tasks) reflecting one's preference of smaller prediction

[1] For better modeling of the ordinal data, it would be important to provide a mathematical characterization and further discussion of the natural ordinal relation. However, it would be related to learning procedure of the 1DT (defined in Section 2.2) more closely, and its necessity is not so great for the analysis and proposal of this study, so we will not mention it in this paper. Refer to, for example, OR studies (da Costa et al., 2008; Yamasaki, 2022) for the discussion on such characterizations.

errors over larger ones such as the ***absolute deviation task loss*** $\ell_{\mathrm{ad}}(j,k) \coloneqq |j-k|$, ***squared task loss*** $\ell_{\mathrm{sq}}(j,k) \coloneqq (j-k)^2$, and $\ell_{\mathrm{zo},c}(j,k) \coloneqq \mathbb{1}(|j-k| > c)$ with $c \geq 0$.

2.2 Formulation of One-Dimensional Transformation (1DT)-Based Methods and Threshold Methods

In this paper, we discuss only 1DT-based OR methods. We here provide notations and terminologies common for all the 1DT-based OR methods.

Suppose that one has the sample $\mathcal{D}_n \coloneqq \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, each of which is drawn independently from an identical distribution of $(\boldsymbol{X}, Y)$. First, 1DT-based methods learn a ***1DT*** $a : \mathbb{R}^d \to \mathbb{R}$ of an observation of the explanatory variables from a user-specified class $\mathcal{A} \subseteq \{a : \mathbb{R}^d \to \mathbb{R}\}$ (***1DT class***) (e.g., a neural network with a fixed network architecture and learnable weight and bias parameters) possibly together with other objects so that an observation with a larger class label tends to have a larger value of the 1DT; we call this procedure the ***learning procedure*** (of the 1DT). Next, 1DT-based methods build up a classifier as $f = h \circ \bar{a}$ with a learned 1DT $\bar{a}$ and a ***labeling function*** $h : \mathbb{R} \to [K]$; we call this procedure the ***labeling procedure***. Most existing 1DT-based methods can be seen as adopting one of the NNT, MT, ST, and LB labelings, depending on the properties of their learning procedure, as we review later in Section 3.

In particular, we call a labeling function h that can be represented as

$$h(u) = h_{\mathrm{thr}}(u; \boldsymbol{t}) \coloneqq 1 + \sum_{k=1}^{K-1} \mathbb{1}(u \geq t_k) \tag{1}$$

with parameters $\boldsymbol{t} = (t_1, \ldots, t_{K-1}) \in \mathbb{R}^{K-1}$ as the ***threshold labeling***. Also, we call the parameters $\boldsymbol{t}$ as the ***threshold parameters***, and a 1DT-based method using a threshold labeling as the ***threshold method***, in this paper. Note that this formulation of the threshold method is a generalization of the one employed in most previous studies on conventional threshold methods that we will review latter in Section 3.2. The threshold labeling $h_{\mathrm{thr}}(u; \boldsymbol{t})$ has the following properties:

Proposition 1. *The threshold labeling $h_{\mathrm{thr}}(u; \boldsymbol{t})$ is non-decreasing and right-continuous in $u \in \mathbb{R}$ and invariant regarding the permutation of the threshold parameters $t_1, \ldots, t_{K-1}$. Conversely, an arbitrary non-decreasing right-continuous function $h : \mathbb{R} \to [K]$ can be represented by a threshold labeling $h_{\mathrm{thr}}(\cdot; \boldsymbol{t})$ with certain threshold parameters $\boldsymbol{t} \in \mathbb{R}^{K-1}$ (i.e., there exist $\boldsymbol{t} \in \mathbb{R}^{K-1}$ such that $h(u) = h_{\mathrm{thr}}(u; \boldsymbol{t})$ for any $u \in \mathbb{R}$) or their permutation. Also, if $t_1, \ldots, t_{K-1}$ take only L different values, then $h_{\mathrm{thr}}(u; \boldsymbol{t})$ has L change points $u = u_1, \ldots, u_L$ such that $h_{\mathrm{thr}}(u_l - \epsilon; \boldsymbol{t}) \neq h_{\mathrm{thr}}(u_l; \boldsymbol{t})$ with a sufficiently small $\epsilon > 0$ for $l = 1, \ldots, L$.*

The last result implies that the threshold labeling is the simplest as the labeling function in the sense that the resulting classifier has only $(K-1)$ decision boundaries for the learned 1DT at most.

Note that, in the learning procedure, since the ***empirical task risk*** $\frac{1}{n}\sum_{i=1}^n \ell(h(a(\boldsymbol{x}_i)), y_i)$ is discontinuous with respect to the 1DT a and hence difficult to optimize numerically, many methods depend on another objective function. In this paper, we call that objective function the ***empirical surrogate risk***, its population version the ***surrogate risk***, and its data-dependent minimal component the ***surrogate loss (function)***.

2.3 Formulation of Policy of This Study

The goal of this study is to improve the labeling procedure of 1DT-based methods. Thus, regarding the learning procedure of the 1DT, this paper adopts those by existing studies, and we do not discuss the goodness of the learning procedure. Assuming that a learned 1DT $\bar{a}$ and task loss ℓ are given, we will discuss the goodness of the labeling function h with the task risk $\mathbb{E}[\ell(h(\bar{a}(\boldsymbol{X})), Y)]$ or the empirical task risk $\frac{1}{n}\sum_{i=1}^n \ell(h(\bar{a}(\boldsymbol{x}_i)), y_i)$ as a criterion, since the aim of the OR task is to minimize the task risk.

3 Review and Analysis of Existing 1DT-Based Methods and Labeling Functions

3.1 Regression-based Method with Nearest-Neighbor Threshold (NNT) Labeling

Regression-based methods (Kramer et al., 2001; Agarwal, 2008) learn a 1DT a by solving a naïve regression task that infers a class label by the 1DT: for a regression loss function $\phi : \mathbb{R} \times [K] \to [0, \infty)$,

$$\min_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \phi(a(\boldsymbol{x}_i), y_i). \tag{2}$$

For example, Kramer et al. (2001) used the ***squared (SQ) loss*** $\phi_{\mathrm{sq}}(u, y) \coloneqq (u - y)^2$, and Agarwal (2008) used the ***absolute-deviation (AD) loss*** $\phi_{\mathrm{ad}}(u, y) \coloneqq |u - y|$, as a regression loss function ϕ.

As a labeling function h, regression-based methods often use the NNT labeling

$$h_{\mathrm{nnt}}(u) \coloneqq h_{\mathrm{thr}}(u; (1.5, 2.5, \dots, K - 0.5)) \tag{3}$$

that is a threshold labeling $h_{\mathrm{thr}}(\cdot; \boldsymbol{t})$ with the threshold parameters $t_k = k + 0.5$, $k = 1, \dots, K - 1$. The NNT labeling rounds the learned 1DT to its nearest label.

The regression-based methods using the above two surrogate loss functions and NNT labeling have the following optimality guarantee for an OR task with a respective specific task loss:

Theorem 1 (Pedregosa et al. (2017, Theorems 9 and 11))**.** *Let (ℓ, ϕ) be $(\ell_{\mathrm{ad}}, \phi_{\mathrm{ad}})$ or $(\ell_{\mathrm{sq}}, \phi_{\mathrm{sq}})$. Then, for the surrogate risk minimizer $\bar{a} \in \arg\min_{a : \mathbb{R}^d \to \mathbb{R}} \mathbb{E}[\phi(a(\boldsymbol{X}), Y)]$ based on the surrogate loss ϕ, the classifier $f_{\mathrm{nnt}} = h_{\mathrm{nnt}} \circ \bar{a}$ minimizes the task risk $\mathbb{E}[\ell(f(\boldsymbol{X}), Y)]$ for the task loss ℓ: $\mathbb{E}[\ell(f_{\mathrm{nnt}}(\boldsymbol{X}), Y)] = \min_{f : \mathbb{R}^d \to [K]} \mathbb{E}[\ell(f(\boldsymbol{X}), Y)]$.*

According to this theorem, the regression-based methods (Kramer et al., 2001; Agarwal, 2008) would be expectable to work well if the sample size n and 1DT class $\mathcal{A}$ are sufficiently large. On the other hand, if the 1DT class $\mathcal{A}$ is strongly restricted, a classifier based on the NNT labeling may be sub-optimal for the OR task (imagine that $\mathbb{E}[\ell(f_{\mathrm{nnt}}(\boldsymbol{X}), Y)] \neq \mathbb{E}[\ell(\bar{h}(\bar{a}(\boldsymbol{X})), Y)]$ for $f_{\mathrm{nnt}} = h_{\mathrm{nnt}} \circ \bar{a}$ with $\bar{a} \in \arg\min_{a \in \mathcal{A}} \mathbb{E}[\phi(a(\boldsymbol{X}), Y)]$ and $\bar{h} \in \arg\min_{h : \mathbb{R} \to [K]} \mathbb{E}[\ell(h(\bar{a}(\boldsymbol{X})), Y)]$).

3.2 Classical Threshold Method with Minimum and Summation Threshold (MT and ST) Labelings

Classical threshold methods have been studied actively in the machine learning and statistical literature (Shashua & Levin, 2003; Chu & Keerthi, 2007; Li & Lin, 2007; Lin & Li, 2012; Pedregosa et al., 2017; Cao et al., 2020). Many of these methods are formulated with a learning procedure that simultaneously learns $(K - 1)$ parameters in addition to the 1DT a:

$$\min_{a \in \mathcal{A}, \boldsymbol{b} \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^{n} \phi(a(\boldsymbol{x}_i), \boldsymbol{b}, y_i), \tag{4}$$

where we call $\boldsymbol{b} = (b_1, \dots, b_{K-1}) \in \mathbb{R}^{K-1}$ the ***bias parameters***, $\mathcal{B} \subseteq \mathbb{R}^{K-1}$ is a user-specified class for the bias parameters (***BPs class***), and $\phi : \mathbb{R} \times \mathbb{R}^{K-1} \times [K] \to [0, \infty)$ is a surrogate loss function. As the labeling function h, several early works (Shashua & Levin, 2003; Chu & Keerthi, 2007) use the MT labeling

$$h_{\mathrm{mt}}(u; \bar{\boldsymbol{b}}) \coloneqq \min\{k \in [K] \mid u < \bar{b}_k\} \text{ with } \bar{b}_K \coloneqq \infty \tag{5}$$

with learned bias parameters $\bar{\boldsymbol{b}} = (\bar{b}_1, \dots, \bar{b}_{K-1})$, and more recent works (Pedregosa et al., 2017; Cao et al., 2020) use the ST labeling

$$h_{\mathrm{st}}(u; \bar{\boldsymbol{b}}) \coloneqq h_{\mathrm{thr}}(u; \bar{\boldsymbol{b}}). \tag{6}$$

The MT and ST labelings are threshold labelings and have the following relationship:

Proposition 2. *Given $\bar{\boldsymbol{b}} \in \mathbb{R}^{K-1}$ together with $\bar{b}_0 \coloneqq -\infty$, let t_k be $\bar{b}_{i_k}$ with $i_k \coloneqq \min\{j \in \{0, \dots, k\} \mid \bar{b}_k \leq \bar{b}_j\}$ for each $k = 1, \dots, K - 1$. Then, one has that $h_{\mathrm{mt}}(u; \bar{\boldsymbol{b}}) = h_{\mathrm{thr}}(u; \boldsymbol{t})$ with $\boldsymbol{t} = (t_1, \dots, t_{K-1})$. Also, $h_{\mathrm{mt}}(u; \bar{\boldsymbol{b}}) = h_{\mathrm{thr}}(u; \bar{\boldsymbol{b}})$ if $\bar{b}_1 \leq \cdots \leq \bar{b}_{K-1}$.*

We remark that the MT labeling $h_{\mathrm{mt}}(u;\bar{\boldsymbol{b}})$ and ST labeling $h_{\mathrm{st}}(u;\bar{\boldsymbol{b}})$ are different when the learned bias parameters $\bar{\boldsymbol{b}}$ are not ordered.

A surrogate loss function $\phi(u,\boldsymbol{b},k)$ for these threshold methods is often built by replacing step functions $\mathbb{1}(l \geq \cdot)$ and $\mathbb{1}(l+1 \leq \cdot)$ in the expression of the task loss $\ell(\cdot,k) = \ell(k,k) + \sum_{l=1}^{k-1}\{\ell(l,k) - \ell(l+1,k)\}\mathbb{1}(l \geq \cdot) + \sum_{l=k}^{K-1}\{\ell(l+1,k) - \ell(l,k)\}\mathbb{1}(l+1 \leq \cdot)$ to convex surrogates widely-used in binary classification (Lin & Li, 2012; Pedregosa et al., 2017) such as the hinge, logistic, and exponential losses, so that $\phi(\cdot,\boldsymbol{b},k)$ becomes a continuous convex upper bound of $\ell(h_{\mathrm{thr}}(\cdot;\boldsymbol{b}),k)$. For example, in the terminology of Pedregosa et al. (2017), the ***immediate threshold (IT) loss function***

$$\phi(a(\boldsymbol{x}),\boldsymbol{b},y) = \begin{cases} \varphi(b_1 - a(\boldsymbol{x})), & \text{if } y = 1, \\ \varphi(a(\boldsymbol{x}) - b_{K-1}), & \text{if } y = K, \\ \varphi(a(\boldsymbol{x}) - b_{y-1}) + \varphi(b_y - a(\boldsymbol{x})), & \text{otherwise} \end{cases} \tag{7}$$

is an upper bound of zero-one task loss ℓ_{zo}, and the ***all threshold (AT) loss function***

$$\phi(a(\boldsymbol{x}),\boldsymbol{b},y) = \begin{cases} \sum_{k=1}^{K-1} \varphi(b_k - a(\boldsymbol{x})), & \text{if } y = 1, \\ \sum_{k=1}^{K-1} \varphi(a(\boldsymbol{x}) - b_k), & \text{if } y = K, \\ \sum_{k=1}^{y-1} \varphi(a(\boldsymbol{x}) - b_k) + \sum_{k=y}^{K-1} \varphi(b_k - a(\boldsymbol{x})), & \text{otherwise} \end{cases} \tag{8}$$

is an upper bound of absolute deviation task loss ℓ_{ad}. For instance, SVOR-IMC (Chu & Keerthi, 2007) uses the IT loss with $\varphi(u) = \min\{0, 1-u\}$, fixed margin strategy (Shashua & Levin, 2003) and SVOR-EXC (Chu & Keerthi, 2007) use the AT loss with $\varphi(u) = \min\{0, 1-u\}$, ORBoost-LR (Lin & Li, 2006) uses the IT loss with $\varphi(u) = e^{-u}$, ORBoost-ALL (Lin & Li, 2006) uses the AT loss with $\varphi(u) = e^{-u}$, and CORAL (Cao et al., 2020) uses the AT loss with $\varphi(u) = \log(1 + e^{-u})$.

As the BPs class $\mathcal{B}$, the non-restricted (non-ordered) class $\mathbb{R}^{K-1}$ and ordered class $\{\boldsymbol{b} \in \mathbb{R}^{K-1} \mid b_1 \leq \cdots \leq b_{K-1}\}$ are often applied. As a simple implementation of the ordered class, Franses & Paap (2001) mentioned to parametrize the bias parameters $\boldsymbol{b}$ as

$$b_1 = b'_1, \text{ and } b_k = b_{k-1} + {b'_k}^2 \text{ for } k = 2, \ldots, K-1, \tag{9}$$

with other parameters $b'_1, \ldots, b'_{K-1} \in \mathbb{R}$. When the ordered BPs class is applied, the MT and ST labelings bring the same classification results. Also, for many AT loss functions, bias parameters $\bar{\boldsymbol{b}}$ of the surrogate risk minimizer $(\bar{a}, \bar{\boldsymbol{b}}) \in \arg\min_{a \in \mathcal{A}, \boldsymbol{b} \in \mathbb{R}^{K-1}} \mathbb{E}[\phi(a(\boldsymbol{X}),\boldsymbol{b},Y)]$ are ensured to be ordered ($\bar{b}_1 \leq \cdots \leq \bar{b}_{K-1}$) (see Chu & Keerthi (2005, Lemma 1) and Li & Lin (2007, Theorem 2)), and hence the use of the ordered BPs class will be justified.

The use of the surrogate loss φ and the MT or ST labeling h, which make $\phi(\cdot,\bar{\boldsymbol{b}},y)$ (with ordered $\bar{\boldsymbol{b}}$) to be an upper bound of the task loss $\ell(h(\cdot),y)$, is almost like a convention and may facilitate generalization analysis (Li & Lin, 2007; Lin & Li, 2012), but the goodness of selecting the MT or ST labeling is not supported by theoretical discussion. We demonstrate in the following example that the MT or ST labeling may be a negative factor that degrades the classification performance of threshold methods:

Example 1. *We here consider the IT loss* (7) *with* $\varphi(u) = \min\{0, 1-u\}$*, which we denote* $\phi_{\text{hinge-it}}(u,\boldsymbol{b},k)$ *and call Hinge-IT loss. The Hinge-IT loss is a continuous convex upper bound of the zero-one task loss with the MT labeling* $\ell_{\mathrm{zo}}(h_{\mathrm{mt}}(\cdot;\boldsymbol{b}),k)$ *(and ST labeling* $\ell_{\mathrm{zo}}(h_{\mathrm{st}}(\cdot;\boldsymbol{b}),k)$*) when* $\boldsymbol{b}$ *is ordered (i.e.,* $b_1 \leq \cdots \leq b_{K-1}$*):* $\phi_{\text{hinge-it}}(\cdot,\boldsymbol{b},k)$ *is a convex function and* $\phi_{\text{hinge-it}}(\cdot,\boldsymbol{b},k) \geq \ell_{\mathrm{zo}}(h_{\mathrm{mt}}(\cdot;\boldsymbol{b}),k)$*. So one may think that the Hinge-IT loss and the task with the zero-one task loss (**Task-Z**) have friendly compatibility, from an analogy of a well-known result, classification calibration (Bartlett et al., 2006), in binary classification. However, the following demonstration shows that the MT labeling may be sub-optimal in minimizing the task risk as a labeling function for the combination of the Hinge-IT loss and Task-Z.*

We consider a 4-class OR problem (let $K = 4$*), and suppose that the data appear only on 4 different points* $\boldsymbol{x}^{[1]}, \ldots, \boldsymbol{x}^{[4]}$ *in* $\mathbb{R}^d$ *and follow the probability distribution,* $\Pr(\boldsymbol{x}^{[i]}) = 0.25$ *and* $(\Pr(1|\boldsymbol{x}^{[i]}), \ldots, \Pr(K|\boldsymbol{x}^{[i]})) = (0.5, 0.4, 0, 0.1)$*,* $(0.3, 0.5, 0, 0.2)$*,* $(0.2, 0, 0.5, 0.3)$*,* $(0.1, 0, 0.4, 0.5)$ *for* $i = 1, \ldots, 4$*.*[2]

[2]We abbreviate the marginal probability $\Pr(\boldsymbol{X} = \boldsymbol{x})$ to $\Pr(\boldsymbol{x})$ and the conditional probability $\Pr(Y = y|\boldsymbol{X} = \boldsymbol{x})$ to $\Pr(y|\boldsymbol{x})$ (this abbreviation applies to an estimate $\hat{\Pr}$ as well).

It can be shown that the surrogate risk minimizer $(\bar{a}, \bar{\boldsymbol{b}}) \in \arg\min_{a:\mathbb{R}^d\to\mathbb{R}, \boldsymbol{b}\in\mathbb{R}^{K-1}} \mathbb{E}[\phi_{\text{hinge-it}}(a(\boldsymbol{X}), \boldsymbol{b}, Y)]$ *satisfies* $\bar{b}_1 = \bar{b}_2 = \bar{b}_3 = 0$, $\bar{a}(\boldsymbol{x}^{[1]}) = \bar{a}(\boldsymbol{x}^{[2]}) = -1$, *and* $\bar{a}(\boldsymbol{x}^{[3]}) = \bar{a}(\boldsymbol{x}^{[4]}) = 1$ *(ignore the translation invariance) by several simple calculations. The MT labeling predicts a label of the data on* $\boldsymbol{x}^{[i]}$ *as* $h_{\text{mt}}(\bar{a}(\boldsymbol{x}^{[i]}); \bar{\boldsymbol{b}}) = 1, 1, 4, 4$ *for* $i = 1, \dots, 4$ *(*$\mathbb{E}[\ell_{\text{zo}}(h_{\text{mt}}(\bar{a}(\boldsymbol{X}); \bar{\boldsymbol{b}}), Y)] = 0.6$*), despite that using different threshold parameters (say* $\boldsymbol{t} = (-2, 0, 2)$*) can predict it as* $h_{\text{mt}}(\bar{a}(\boldsymbol{x}^{[i]}); \boldsymbol{t}) = 2, 2, 3, 3$ *for* $i = 1, \dots, 4$ *and yield a smaller value of the task risk (*$\mathbb{E}[\ell_{\text{zo}}(h_{\text{mt}}(\bar{a}(\boldsymbol{X}); \boldsymbol{t}), Y)] = 0.55$*).* □

3.3 Statistical Methods with Likelihood-Based (LB) Labeling

In the OR research in statistics, several methods have been developed according to the statistical modeling of the conditional probabilities of the data through a 1DT (McCullagh, 1980; Williams, 2006; Chu & Ghahramani, 2005; Yamasaki, 2022). For example, the cumulative link model (McCullagh, 1980), which is popular in the OR research, models the conditional probabilities $\Pr(y|\boldsymbol{x})$, $(\boldsymbol{x}, y) \in \mathbb{R}^d \times [K]$ by

$$P(y; \sigma, \tilde{a}(\boldsymbol{x}), \tilde{\boldsymbol{b}}) \coloneqq \begin{cases} \sigma(\tilde{b}_y - \tilde{a}(\boldsymbol{x})), & \text{if } y = 1, \\ 1 - \sigma(\tilde{b}_{y-1} - \tilde{a}(\boldsymbol{x})), & \text{if } y = K, \\ \sigma(\tilde{b}_y - \tilde{a}(\boldsymbol{x})) - \sigma(\tilde{b}_{y-1} - \tilde{a}(\boldsymbol{x})), & \text{otherwise,} \end{cases} \tag{10}$$

where $\sigma : \mathbb{R} \to [0, 1]$ is a link function that is non-decreasing and satisfies $\sigma(-\infty) = 0$ and $\sigma(+\infty) = 1$ (i.e., σ is a cumulative distribution (CPD) function), $\tilde{a} : \mathbb{R}^d \to \mathbb{R}$ is an assumed 1DT, and $\tilde{\boldsymbol{b}} = (\tilde{b}_1, \dots, \tilde{b}_{K-1}) \in \mathbb{R}^{K-1}$ are assumed bias parameters that satisfy $\tilde{b}_1 \le \cdots \le \tilde{b}_{K-1}$. As a link function σ, ordinal logistic regression (OLR) (McCullagh, 1980) applies the CPD function of the logistic distribution (a.k.a. the sigmoid function) $\sigma_{\text{logistic}}(u) \coloneqq 1/(1 + e^{-u})$, and cumulative probit model (Agresti, 2010, Section 5.2) and Gaussian process OR (GPOR) proposed by Chu & Ghahramani (2005) use the CPD function of the standard Gaussian distribution (a.k.a. the inverse function of the probit function) $\sigma_{\text{gauss}}(u) \coloneqq \int_{-\infty}^{u} (2\pi)^{-1/2} e^{-v^2/2}\, dv$.

In the learning procedure of the 1DT, statistical methods apply surrogate loss functions associated with their statistical modeling. For the cumulative link model, a surrogate loss function ϕ and the BPs class $\mathcal{B}$ for the learning procedure same as (4) should satisfy

$$(\tilde{u}, \tilde{\boldsymbol{b}}) = \operatorname*{arg\,min}_{u\in\mathbb{R}, \boldsymbol{b}\in\mathcal{B}} \sum_{y=1}^{K} P(y; \sigma, \tilde{u}, \tilde{\boldsymbol{b}}) \phi(u, \boldsymbol{b}, y). \tag{11}$$

A popular instance of the surrogate loss function is the ***negative log likelihood (NLL) loss function*** $\phi_{\text{nll}}(u, \boldsymbol{b}, y; \sigma) \coloneqq -\log\{P(y; \sigma, u, \boldsymbol{b})\}$, for which the learning procedure amounts to the maximum likelihood estimation for the model (10). Also, Cao et al. (2020) used the AT loss (8) with $\varphi(u) = -\log\{\sigma(u)\}$, which we call the ***all negative log cumulative likelihoods (ANLCL) loss function*** and denote $\phi_{\text{anlcl}}(u, \boldsymbol{b}, y; \sigma)$, for $\sigma = \sigma_{\text{logistic}}$. The learning procedure for this loss function can be characterized as the minimization of sum of the NLLs of the models of cumulative conditional probabilities $\Pr(Y \le k|\boldsymbol{X} = \boldsymbol{x})$ for binary classification problems, 'k or less' vs. 'more than k', $k = 1, \dots, K-1$.

The above interpretation on using the surrogate losses ϕ_{nll} and ϕ_{anlcl} under the statistical model (10) can be mathematically understood as follows:

Theorem 2. *Assume that the random variable* $(\boldsymbol{X}, Y)$ *underlying the data has conditional probabilities that can be represented as* (10)*:* $\Pr(y|\boldsymbol{x}) = P(y; \sigma, \tilde{a}(\boldsymbol{x}), \tilde{\boldsymbol{b}})$ *for every* $y \in [K]$ *and any* $\boldsymbol{x} \in \mathbb{R}^d$ *in the support of the distribution of* $\boldsymbol{X}$ *with* σ *that is non-decreasing and satisfies* $\sigma(-\infty) = 0$ *and* $\sigma(+\infty) = 1$ *(and* $\sigma(-\cdot) = 1 - \sigma(\cdot)$ *for* $\phi = \phi_{\text{anlcl}}$*) such as* σ_{logistic} *and* σ_{gauss}*,* $\tilde{a} : \mathbb{R}^d \to \mathbb{R}$*, and* $\tilde{\boldsymbol{b}} \in \mathbb{R}^{K-1}$ *satisfying* $\tilde{b}_1 \le \cdots \le \tilde{b}_{K-1}$*. Let* $\mathcal{A}$ *be* $\{g : \mathbb{R}^d \to \mathbb{R}\}$*, and* $(\phi, \mathcal{B})$ *be* $(\phi_{\text{nll}}, \{\boldsymbol{b} \in \mathbb{R}^{K-1} \mid b_1 \le \cdots \le b_{K-1}\})$*,* $(\phi_{\text{anlcl}}, \mathbb{R}^{K-1})$*, or* $(\phi_{\text{anlcl}}, \{\boldsymbol{b} \in \mathbb{R}^{K-1} \mid b_1 \le \cdots \le b_{K-1}\})$*. Then, any surrogate risk minimizer* $(\bar{a}, \bar{\boldsymbol{b}}) \in \arg\min_{a\in\mathcal{A}, \boldsymbol{b}\in\mathcal{B}} \mathbb{E}[\phi(a(\boldsymbol{X}), \boldsymbol{b}, Y)]$ *satisfies* $P(y; \sigma, \bar{a}(\boldsymbol{x}), \bar{\boldsymbol{b}}) = \Pr(y|\boldsymbol{x})$ *for any* $\boldsymbol{x} \in \mathbb{R}^d$ *in the support of the distribution of* $\boldsymbol{X}$*.*

For such methods, not only the threshold labelings but also the LB labeling that grounds on the assumed statistical model is a widely-used option for the labeling function. Considering Theorem 2 and the equality $\mathbb{E}[\ell(f(\boldsymbol{x}), Y)] = \sum_{y=1}^{K} \Pr(y|\boldsymbol{x}) \ell(f(\boldsymbol{x}), y)$, and aiming to minimize the task risk, $\min_{f:\mathbb{R}^d\to[K]} \mathbb{E}[\ell(f(X), Y)]$, these

methods can predict a label with the LB labeling

$$h_{\mathrm{lb}}(u;\sigma,\bar{\boldsymbol{b}},\ell) \coloneqq \min\left(\operatorname*{arg\,min}_{k\in[K]} \sum_{y=1}^{K} P(y;\sigma,u,\bar{\boldsymbol{b}})\ell(k,y)\right) \tag{12}$$

with learned bias parameters $\bar{\boldsymbol{b}}$, under the expectation that the assumed statistical model (10) correctly represents the actual statistical behavior of the data and it is learned successfully. Note that there can be a tie situation where objective functions with different k of (12) take the same value, and $\arg\min_{k\in[K]}$ operation outputs multiple labels. The overlaid min operation in (12) avoids such a tie situation.

These methods tend to perform better when their assumed statistical model represents the actual statistical behavior of the data well. One can, however, find that the condition in Theorem 2 is very restrictive. Therefore, in many practical situations, their statistical model would deviate from the actual statistical behavior of the data, and then their 1DT model may not be learned appropriately, and the LB labeling $h_{\mathrm{lb}}(\cdot;\sigma,\bar{\boldsymbol{b}},\ell)$ may be sub-optimal for the learned 1DT model $\bar{a}$.

One may still consider that the LB labeling is more flexible, in that it is generally not restricted within the class of non-decreasing threshold labelings, and superior to threshold labelings. However, we found that the LB labeling takes the form of the threshold labeling, for typical statistical models such as ones in OLR and GPOR (i.e., the link function σ such as $\sigma_{\mathrm{logistic}}, \sigma_{\mathrm{gauss}}$) and for typical task losses such as $\ell = \ell_{\mathrm{zo}}, \ell_{\mathrm{zo},c}, \ell_{\mathrm{ad}}, \ell_{\mathrm{sq}}$.

Theorem 3. *Suppose that σ is non-decreasing and satisfies $\sigma(-\infty) = 0$ and $\sigma(+\infty) = 1$ and that $\bar{b}_1 \leq \cdots \leq \bar{b}_{K-1}$. Then, the LB labeling $h_{\mathrm{lb}}(u;\sigma,\bar{\boldsymbol{b}},\ell)$ is*

(i) *a certain threshold labeling $h_{\mathrm{thr}}(u;\boldsymbol{t})$ for some $\boldsymbol{t} \in \mathbb{R}^{K-1}$, if $\ell(k,l)$ at each fixed $k \in [K]$ is non-increasing in l for $l \leq k$ and non-decreasing in l for $l \geq k$, and $\ell_{k,l}(j) \coloneqq \ell(k,j) - \ell(k,j+1) - \ell(l,j) + \ell(l,j+1)$ at each fixed different $k,l \in [K]$ is non-positive (resp. non-negative) for all $j \in [K-1]$ respectively when $k < l$ (resp. $k > l$), such as $\ell = \ell_{\mathrm{ad}}, \ell_{\mathrm{sq}}$,*

(ii) *a certain threshold labeling $h_{\mathrm{thr}}(u;\boldsymbol{t})$ for some $\boldsymbol{t} \in \mathbb{R}^{K-1}$, if $\ell = \ell_{\mathrm{zo}}, \ell_{\mathrm{zo},c}$ with $c \in [0, \lfloor K/2 \rfloor)$, σ is differentiable, $\sigma'(v)$ is even and non-increasing in v if $v > 0$, and $\frac{\sigma'(v_1)-\sigma'(v_2)}{\sigma(v_1)-\sigma(v_2)}$ is non-increasing in v_1 with fixed v_2 and in v_2 with fixed v_1 if $v_1 < v_2$, such as $\sigma_{\mathrm{logistic}}$ and σ_{gauss}, where $\lfloor v \rfloor$ is the greatest integer less than or equal to v,*

(iii) *the threshold labeling $h_{\mathrm{thr}}(u;\bar{\boldsymbol{b}})$ that is same as the MT labeling $h_{\mathrm{mt}}(u;\bar{\boldsymbol{b}})$ and ST labeling $h_{\mathrm{st}}(u;\bar{\boldsymbol{b}})$, if $\ell = \ell_{\mathrm{ad}}$ and $\sigma(0) = 0.5$.*

Here, Theorem 3 (i) assumes that the task loss ℓ is V-shaped, and the condition on $\ell_{k,l}$ in Theorem 3 (i) holds under the convexity of ℓ defined below:

Theorem 4. *$\ell_{k,l}(j)$ at each fixed different $k,l \in [K]$ is non-positive and non-negative for all $j \in [K-1]$ respectively when $k < l$ and $k > l$, if the task risk ℓ is convex in the difference of the two arguments:*

$$\ell(j_3,k_3) \leq \frac{(j_3-k_3)-(j_1-k_1)}{(j_2-k_2)-(j_1-k_1)}\ell(j_1,k_1) + \frac{(j_2-k_2)-(j_3-k_3)}{(j_2-k_2)-(j_1-k_1)}\ell(j_2,k_2) \tag{13}$$

for all $j_1,\ldots,k_3 \in [K]$ such that $j_1 - k_1 \neq j_2 - k_2$ and $j_1 - k_1 \leq j_3 - k_3 \leq j_2 - k_2$.

The condition on σ in Theorem 3 (ii) comes from the consideration for non-convex task losses.

A result similar to Theorem 3 also holds for 1DT-based likelihood models other than the CL model (10); refer to Theorem 5 in Section B. For 1DT-based statistical methods, it may be common that their LB labeling is a threshold labeling.

4 Proposal of 1DT-based Methods with Empirical Optimal Threshold (EOT) Labeling

In typical usages, not only the NNT, MT, and ST labelings, but also the LB labeling is a threshold labeling, as we confirmed in Theorem 3. Thus, we consider that it would be meaningful to aim for a better threshold

Algorithm 1: Calculation of the threshold parameters for the EOT labeling

Input: Task loss ℓ, learned 1DT $\bar{a}$, and training data $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$.

`/* Preparation of` $\{\bar{a}_j\}_{j=1}^N$ `and` $\mathcal{Y}_j, j = 1, \dots, N$. `*/`

1 Let $\{\bar{a}'_j\}_{j=1}^N$ be unique elements of $\{\bar{a}(\boldsymbol{x}_i)\}_{i=1}^n$: $\bar{a}'_{j_1} \neq \bar{a}'_{j_2}$ for all $j_1, j_2 \in [N]$ s.t. $j_1 \neq j_2$, and $\bar{a}(\boldsymbol{x}_i) \in \{\bar{a}'_j\}_{j=1}^N$ for all $i \in [n]$.

2 Sort $\{\bar{a}'_j\}_{j=1}^N$ in the ascending order, and represent the result as $\{\bar{a}_j\}_{j=1}^N$: $\bar{a}_1 \leq \cdots \leq \bar{a}_N$.

3 Create sets $\mathcal{Y}_j = \{y_i \mid \bar{a}(\boldsymbol{x}_i) = \bar{a}_j\}_{i=1}^n$, $j = 1, \dots, N$.

`/* Calculate matrix` $L \in \mathbb{R}^{N \times K}$. `*/`

4 **for** $k = 1, \dots, K$ **do**

5 | $L_{1,k} = \sum_{y_i \in \mathcal{Y}_1} \ell(k, y_i)$.

6 **for** $j = 2, \dots, N$ **do**

7 | **for** $k = 1, \dots, K$ **do**

8 | | $L_{j,k} = \min_{l \in [k]} L_{j-1,l} + \sum_{y_i \in \mathcal{Y}_j} \ell(k, y_i)$.

`/* Calculate threshold parameters` $\bar{\boldsymbol{t}}$. `*/`

9 $I \leftarrow \min(\arg\min_{l \in [K]} L_{N,l})$.

10 **if** $I \neq K$ **then**

| Let $\bar{t}_k$ be a value larger than $\bar{a}_N$ (e.g., $+\infty$) for $k = I, \dots, K-1$.

11 **for** $j = N-1, \dots, 1$ **do**

12 | $J \leftarrow \min(\arg\min_{l \in [I]} L_{j,l})$.

13 | **if** $I \neq J$ **then**

| | $\bar{t}_k = (\bar{a}_j + \bar{a}_{j+1})/2$ for $k = J, \dots, I-1$, and $I \leftarrow J$.

14 **if** $I \neq 1$ **then**

| Let $\bar{t}_k$ be a value smaller than $\bar{a}_1$ (e.g., $-\infty$) for $k = 1, \dots, I-1$.

Output: Threshold parameters $\bar{\boldsymbol{t}} = (\bar{t}_1, \dots, \bar{t}_{K-1})$.

labeling for improving the classification performance of existing 1DT-based methods. Recalling that the final goal is to make the task risk $\mathbb{E}[\ell(f(\boldsymbol{X}), Y)]$ small, and expecting that the 1DT was learned successfully and the empirical (training) task risk becomes a good estimate of the (test) task risk, we propose to apply the EOT labeling

$$h(u; \boldsymbol{t}_{\mathrm{eot}}) \text{ with } \boldsymbol{t}_{\mathrm{eot}} \in \operatorname*{arg\,min}_{\boldsymbol{t} \in \mathbb{R}^{K-1}} \frac{1}{n} \sum_{i=1}^n \ell(h_{\mathrm{thr}}(\bar{a}(\boldsymbol{x}_i); \boldsymbol{t}), y_i) \tag{14}$$

that uses threshold parameters minimizing the empirical task risk for a given learned 1DT model $\bar{a}$.

The threshold parameters for the EOT labeling can be computed with a dynamic programming-based algorithm (Algorithm 1) mentioned in Lin & Li (2006); see also researchers' site (`https://home.work.caltech.edu/~htlin/program/orensemble/`) of Lin & Li (2006), and Section C of our paper for its optimality guarantee. It costs a computational complexity of quasi-linear order $O(n \log n)$ regarding the training sample size n, which stems from the sorting operation in Line 2.

The NNT labeling for regression-based methods (reviewed in Section 3.1) and LB labeling for statistical methods (reviewed in Section 3.3) have optimality guarantees for the task risk minimization under ideal situations; refer to Theorems 1 and 2. Also, many threshold methods employ $(K-1)$ bias parameters, and those bias parameters can be directly used in their labeling procedure like MT and ST labelings, as reviewed in Section 3.2. Presumably, for these reasons, a formulation that allows a threshold labeling with variable threshold parameters have not been considered for these methods. Our formulation in Section 2.2 introduces the threshold labeling with variable threshold parameters and clearly distinguishes the bias and threshold parameters. This is also an important contribution of this paper: due to this formulation, it becomes natural to consider the application of the threshold labeling to 1DT-based methods with a different number of bias parameters than $(K-1)$ such as PO-VS-SL (Yamasaki, 2022). Furthermore, this led to the EOT labeling that has a potential to improve the classification performance.

5 Relationship to Other Previous Studies

Most existing studies on 1DT-based methods discuss the learning procedure. In Section 3, we reviewed ***point-wise 1DT-based methods*** in which the learning procedure is formulated with a surrogate loss function defined for each point $(a(\boldsymbol{x}_i), y_i)$, but some 1DT-based methods employ different formulations. For example, ***pair-wise 1DT-based methods*** including RankSVM (Herbrich et al., 1999) and RankBoost (Lin & Li, 2006) are formulated with a surrogate loss function that is defined for each pair of two points $(a(\boldsymbol{x}_i), y_i)$ and $(a(\boldsymbol{x}_j), y_j)$; see also Lin & Li (2012); Gutierrez et al. (2015) for survey of such methods. The EOT labeling can be applied to those 1DT-based methods as well.

On the other hand, there are several previous works that have studied the labeling procedures different from NNT, MT, ST, LB, and EOT labelings. We review the discussion of those previous works, and describe the relationship between their methods and the EOT labeling, in the following.

Herbrich et al. (1999) considered a pairwise 1DT-based method based on a hinge-type surrogate loss, for the task with the zero-one task loss. Their method (Herbrich et al., 1999, (12)) adopts a threshold labeling with threshold parameters determined by minimizing a unique criterion that emphasizes the shape of the used hinge-type loss and is different from the (empirical) surrogate risk, and has no optimality guarantee in the task risk minimization.

Lin & Li (2006) considered three methods; see RankBoost (4), ORBoost-LR (7), and ORBoost-ALL (8) of this reference. The third method applies the AT loss with $\varphi(u) = e^{-u}$ (ORBoost-ALL), and its labeling procedure is the same as the MT and ST labelings in this paper. Also, one can understand that the second method, ORBoost-LR (7), tries to minimize (an upper bound of) the empirical surrogate risk based on the IT loss function with $\varphi(u) = e^{-u}$ and ordered BPs class in its learning procedure, and its labeling procedure also can be seen to follow the idea of the MT and ST labelings. Finally, the first method, RankBoost, is a pair-wise 1DT-based method, and its objective function (4) for the labeling procedure is defined as the empirical task risk with the absolute deviation task loss. This labeling procedure is similar to the EOT labeling, but Lin & Li (2006) did not mention the relevance between that objective function and the task under consideration. Although Lin & Li (2006) presented a description on the threshold parameters determination based on the zero-one task loss at the part following (4), they tried only the threshold parameters determined by minimizing the empirical task risk with the absolute deviation task loss even when they considered the task with the zero-one task loss in their experiments. The main claim of our consideration in Section 4 is that the threshold parameters should be determined via minimizing the (empirical) task risk for the task under consideration. Therefore, the EOT labeling in Section 4 can be interpreted as a variant of those for Lin & Li (2006, (4)) with respect to the relevance between the objective function to determine the threshold parameters and the task under consideration.

6 Numerical Experiments

Purpose We performed numerical experiments to answer the question, whether a modified 1DT-based method with the EOT labeling can yield better classification performance (i.e., smaller test task risk) for the OR task than existing 1DT-based methods using other labeling functions.

Datasets and Preprocessing In the experiments, we used the five ***various-domain datasets***, LEV (lectures evaluation), ERA (employee rejection/acceptance), SWD (social workers decision), WQR (winequality-red), and CAR (car evaluation) datasets, and the three ***face-age datasets***, MORPH (MORPH Album2), CACD, and AFAD datasets (Ricanek & Tesafaye, 2006; Chen et al., 2014; Niu et al., 2016). The main reason why we used the various-domain and face-age datasets is respectively to experiment with many real-world datasets in various domains and to confirm whether the proposed method achieves the classification performance competitive to the state-of-the-art method in a modern application. For most of the experimental settings, we referred to those of the previous study (Cao et al., 2020).[3]

[3]For the face-age datasets, we used a part of program codes published in `https://github.com/Raschka-research-group/coral-cnn` by Cao et al. (2020), but results of our reproduction of their method differ from theirs mainly because we changed a learning rate from 5×10^{-5} to $10^{-2.5}$. See `https://github.com/Anonymous` for program codes that we used.

Table 1: Dataset properties, the total sample size n_{tot}, the dimension d of the explanatory variables, and the number K of classes of the target variable, of all the used datasets.

	LEV	ERA	SWD	WQR	CAR	MORPH	CACD	AFAD
n_{tot}	1000	1000	1000	1599	1728	55013	159402	164418
d	4	4	10	11	21	$128^2 \times 3$	$128^2 \times 3$	$128^2 \times 3$
K	5	9	4	6	4	55	49	26

For the various-domain datasets, we selected those with the total sample size $n_{\mathrm{tot}} \geq 1000$ from datasets that Gutierrez et al. (2015) used as OR datasets, and used them following the setting of the explanatory and target variables in Gutierrez et al. (2015). One can obtain the various-domain datasets from a researchers' site (`http://www.uco.es/grupos/ayrna/orreview`) of Gutierrez et al. (2015) or our GitHub repository (`https://github.com/Anonymous`). We purchased the MORPH dataset at `https://ebill.uncw.edu/C20231_ustores/web/` and preprocessed it so that the face spanned the whole image with the nose tip, which was located by facial landmark detection (Sagonas et al., 2016), at the center by using `EyepadAlign` function by Raschka (2018). While this dataset contains 55,134 facial images with ages from 16 to 77, we used 55,013 images with ages from 16 to 70. The CACD dataset can be downloaded from `https://bcsiriuschen.github.io/CARC/`. We preprocessed this dataset similarly to the MORPH dataset. Since the CACD dataset collects images from the Internet using computer vision techniques, it includes some facial images inappropriate for our consideration. Excluding images, in which no face or more than two faces were detected in the preprocessing, from the original 163,446 images, we used 159,402 facial images in the age range of 14–62 years. For the AFAD dataset obtainable at `https://github.com/afad-dataset/tarball`, because faces in its images were already centered, we took no further preprocessing, and used its 164,418 images with ages 15–40. For these face-age datasets, we treated the age rank as the target variable. Table 1 summarizes basic dataset properties, the total sample size n_{tot}, the dimension d of the explanatory variables, and the number K of classes of the target variable, of all the used datasets.

For the various-domain datasets, we randomly divided each dataset into 72 % training, 8 % validation, and 20 % test sets. For the face-age datasets, we resized all images to $128 \times 128 \times 3$ pixels (3 stems from RGB channels) and randomly divided each dataset into 72 % training, 8 % validation, and 20 % test sets, and the training phase used images randomly cropped with the size of $120 \times 120 \times 3$ pixels as input to improve the stability of the model against the difference of facial positions, and validation and test phases used images center-cropped to the same size, following procedures by Cao et al. (2020).

Tasks We considered the three popular OR tasks: minimization of the task risk for the zero-one task loss $\ell_{\mathrm{zo}}(j,k) = \mathbb{1}(j \neq k)$ (***Task-Z***), that for the absolute deviation task loss $\ell_{\mathrm{ad}}(j,k) = |j-k|$ (***Task-A***), and that for the squared task loss $\ell_{\mathrm{sq}}(j,k) = (j-k)^2$ (***Task-S***).

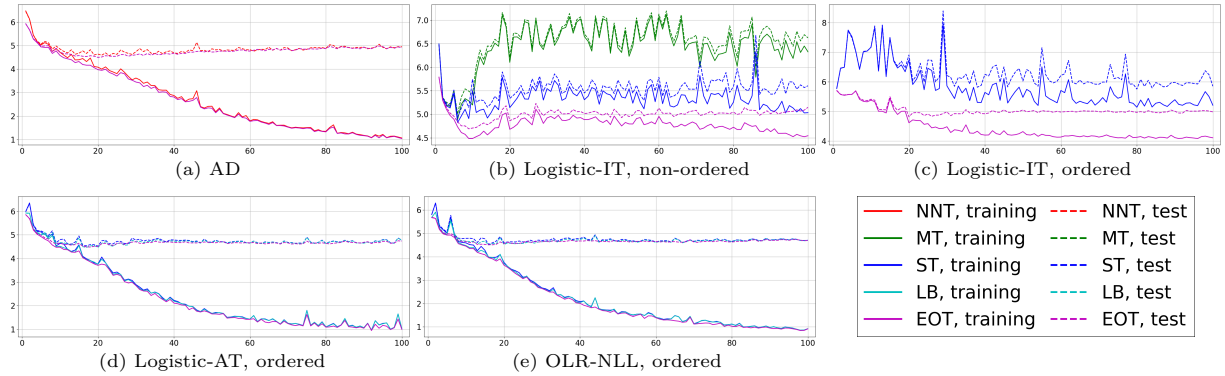
Methods For the various-domain datasets, we applied a 1DT class based on a 4-layer fully-connected neural network, in which every hidden layer has 100 nodes activated with the sigmoid function in addition to bias nodes. Also, for the face-age datasets, we applied a 1DT class based on ResNet-34 (He et al., 2016), a modern CNN architecture, following (Cao et al., 2020)'s implementation. It modifies a fully-connected (the number of classes)-output final layer of the conventional ResNet-34 to a fully-connected 1-output layer.

As the surrogate loss function, we tried the AD loss φ_{ad}; the IT loss (7) with $\varphi(u) = \log(1+e^{-u})$ (***Logistic-IT***); the AT loss (8) with $\varphi(u) = \log(1+e^{-u})$ (***Logistic-AT***); the NLL loss ϕ_{nll} for the statistical model (10) with $\sigma = \sigma_{\mathrm{logistic}}$ (***OLR-NLL***).[4]

As the BPs class, we tried the non-ordered class and the ordered class for the Logistic-IT loss; the ordered class for the Logistic-AT and OLR-NLL losses. Note that the AD loss has no bias parameters.

As the labeling procedure, we tried the NNT and EOT labeling for the AD loss; the MT, ST, and EOT labelings for the Logistic-IT loss; the MT, ST, LB, and EOT labelings for the Logistic-AT and OLR-NLL losses. When using the ordered BPs class, the MT and ST labeling yield the same result (see Proposition 2).

[4]For numerical stability (to avoid log(0)), we used an approximation of the NLL loss in which the logarithmic function $\log(\cdot)$ of ϕ_{nll} is replaced to $\log(\cdot + 10^{-8})$ in the learning procedure.



(a) AD (b) Logistic-IT, non-ordered (c) Logistic-IT, ordered

(d) Logistic-AT, ordered (e) OLR-NLL, ordered

Figure 1: The learning curves of the training and test errors over 100 training epochs in a certain trial. These are for a special case for the AFAD dataset and the RMSE for the Task-S. In Figures (c)–(e), the ST and MT labelings yield the same curves.

Also, for the Logistic-AT and OLR-NLL losses with the ordered BPs class, MT, ST, and LB labelings yield the same result under the Task-A (see Theorem 3).

Cao et al. (2020) declare that their method, which is a combination of the Logistic-AT loss and ST labeling, is the state-of-the-art method for the face-age datasets in 2020. Although they used the non-ordered BPs class, bias parameters of the surrogate risk minimizer are guaranteed to be ordered, and hence using the ordered BPs class would have made little difference to the result. Thus, we treated results for the method with the Logistic-AT loss, ordered BPs class, and ST labeling as their results, in the comparison.

Training and Evaluation During the validation and test phases, models are evaluated based on the mean zero-one error (***MZE***), the mean absolute deviation error (***MAE***), and the root of the mean squared error (***RMSE***), which are defined for a classifier f and m used data points as $\frac{1}{m}\sum_{i=1}^{m}\ell_{\mathrm{zo}}(f(\boldsymbol{x}_i), y_i)$, $\frac{1}{m}\sum_{i=1}^{m}\ell_{\mathrm{ad}}(f(\boldsymbol{x}_i), y_i)$, and $\{\frac{1}{m}\sum_{i=1}^{m}\ell_{\mathrm{sq}}(f(\boldsymbol{x}_i), y_i)\}^{1/2}$, for the Task-Z, Task-A, and Task-S, respectively. Here, the root operation of the RMSE is only for adjusting the scale of the error and does not affect our discussion related to the optimality of the EOT labeling.

We ran 50 trials for the various-domain datasets and 10 trials for the face-age datasets, with randomly-set different divisions of training, validation, and test sets and initial parameters of the network. In each trial, we trained the network using Adam of the learning rate $10^{-2.5}$ and mini-batch size 256 as an optimization procedure for 500 epochs for the various-domain datasets or 100 epochs for the face-age datasets. Additionally, for methods using the EOT labeling, we calculated the threshold parameters according to Algorithm 1 at the end of every training epoch. The above errors were evaluated on the validation set at the end of every training epoch, and then we adopted a model at the timing with the smallest error among the obtained validation error sequences as the test model.

We judge the significance on the classification performance of the labeling function by the one-sided Wilcoxon rank sum test with p-value 0.1 based on errors for all the trials of methods using different labeling functions, in each combination of the dataset, error, surrogate loss function, and BPs class.

Results Figure 1 shows the learning curves of the training and test errors. The EOT labeling results in smaller training errors as its design, which appears to in turn result in smaller test errors. Also, we can find that the EOT labeling may stabilize (or smooth) the learning curve from Figures (c) and (d). Table 2 shows the mean and standard deviation of the errors, for the test model, evaluated on the test set. In many tried cases, the EOT labeling outperformed the NNT, MT, ST, and LB labelings regarding the test task risk. In particular, for the face-age datasets, modified 1DT-based methods using the EOT labeling provided better performance than the method (Logistic-AT, Ordered, ST) in Cao et al. (2020) that was the state-of-the-art in 2020. These results suggest the success of the EOT labeling in the subject (aiming for a better labeling procedure) of our research.

Table 2: It shows the mean (M) and standard deviation (S) of the test errors in the form 'M_S'. The smaller the error, the better that method is for that dataset and that task. In each block specified with the dataset, error, and learning procedure, we highlighted in bold font the best results that were tie each other and superior to all other results with respect to the one-sided Wilcoxon rank sum test with a significance level of 0.1, if they exist. Also, we colored the best results in red for each combination of dataset and error.

	Learning	Labeling	LEV	ERA	SWD	WQR	CAR	MORPH	CACD	AFAD
MZE for Task-Z	AD	NNT	$.385_{.026}$	$.754_{.027}$	$.411_{.028}$	$.392_{.026}$	$.014_{.009}$	$.875_{.005}$	$.927_{.002}$	$.882_{.002}$
		EOT	$.384_{.027}$	$\mathbf{.733_{.032}}$	$.412_{.025}$	$.395_{.023}$	$.013_{.008}$	$.874_{.003}$	$\mathbf{.925_{.001}}$	$\mathbf{.878_{.002}}$
	Logistic-IT non-ordered	MT	$.390_{.030}$	$.744_{.029}$	$.414_{.036}$	$.393_{.023}$	$.017_{.007}$	$.884_{.003}$	$.937_{.001}$	$.887_{.004}$
		ST	$.390_{.030}$	$.744_{.029}$	$.414_{.036}$	$.394_{.021}$	$.018_{.008}$	$.885_{.003}$	$.936_{.002}$	$.895_{.002}$
		EOT	$.387_{.026}$	$\mathbf{.734_{.033}}$	$.414_{.031}$	$\mathbf{.384_{.021}}$	$\mathbf{.010_{.006}}$	$.883_{.004}$	$\mathbf{.936_{.002}}$	$\mathbf{.882_{.002}}$
	Logistic-IT ordered	MT,ST	$.385_{.029}$	$.725_{.031}$	$.416_{.033}$	$.390_{.024}$	$.011_{.006}$	$.893_{.004}$	$.947_{.002}$	$\mathbf{.881_{.002}}$
		EOT	$.386_{.026}$	$.729_{.031}$	$.412_{.030}$	$.389_{.025}$	$.010_{.006}$	$\mathbf{.890_{.002}}$	$\mathbf{.946_{.002}}$	$.882_{.002}$
	Logistic-AT ordered	MT,ST	$.383_{.028}$	$\mathbf{.731_{.026}}$	$.413_{.034}$	$.397_{.022}$	$.012_{.006}$	$.874_{.005}$	$.929_{.004}$	$.883_{.002}$
		LB	$.386_{.029}$	$.758_{.026}$	$.418_{.031}$	$.397_{.022}$	$.011_{.006}$	$.874_{.006}$	$.930_{.005}$	$\mathbf{.880_{.002}}$
		EOT	$.389_{.025}$	$\mathbf{.732_{.030}}$	$.412_{.028}$	$.393_{.020}$	$.011_{.006}$	$.872_{.006}$	$.929_{.004}$	$\mathbf{.879_{.003}}$
	OLR-NLL ordered	MT,ST	$.380_{.025}$	$\mathbf{.725_{.035}}$	$.411_{.033}$	$.394_{.025}$	$.010_{.005}$	$.871_{.005}$	$.926_{.002}$	$.878_{.002}$
		LB	$.383_{.029}$	$.747_{.027}$	$.419_{.030}$	$.392_{.024}$	$.010_{.005}$	$.870_{.005}$	$.926_{.003}$	$\mathbf{.876_{.002}}$
		EOT	$.384_{.026}$	$\mathbf{.733_{.032}}$	$.410_{.031}$	$.394_{.026}$	$.010_{.005}$	$.871_{.004}$	$.925_{.002}$	$\mathbf{.876_{.003}}$
MAE for Task-A	AD	NNT	$.413_{.031}$	$1.236_{.069}$	$.429_{.035}$	$.430_{.029}$	$.014_{.009}$	$2.931_{.025}$	$5.243_{.029}$	$3.352_{.024}$
		EOT	$.413_{.030}$	$1.231_{.072}$	$.428_{.031}$	$.432_{.028}$	$.013_{.009}$	$2.918_{.033}$	$5.238_{.029}$	$3.349_{.020}$
	Logistic-IT non-ordered	MT	$.425_{.036}$	$1.262_{.073}$	$.438_{.037}$	$.430_{.026}$	$.017_{.007}$	$3.189_{.049}$	$5.758_{.101}$	$3.796_{.080}$
		ST	$.425_{.036}$	$1.262_{.073}$	$.438_{.037}$	$.429_{.026}$	$.018_{.009}$	$3.188_{.049}$	$5.757_{.103}$	$3.732_{.038}$
		EOT	$\mathbf{.414_{.031}}$	$\mathbf{1.228_{.070}}$	$.436_{.034}$	$\mathbf{.421_{.026}}$	$\mathbf{.010_{.006}}$	$\mathbf{3.124_{.033}}$	$\mathbf{5.631_{.093}}$	$\mathbf{3.538_{.030}}$
	Logistic-IT ordered	MT,ST	$.416_{.030}$	$1.239_{.076}$	$.440_{.033}$	$.426_{.026}$	$.011_{.006}$	$3.742_{.084}$	$6.685_{.131}$	$4.082_{.043}$
		EOT	$.413_{.026}$	$1.226_{.069}$	$.436_{.031}$	$.428_{.027}$	$.010_{.006}$	$\mathbf{3.540_{.038}}$	$\mathbf{6.359_{.042}}$	$\mathbf{3.619_{.023}}$
	Logistic-AT ordered	MT,ST,LB	$.416_{.032}$	$1.233_{.063}$	$.437_{.034}$	$.431_{.023}$	$.011_{.006}$	$2.874_{.037}$	$5.381_{.138}$	$3.375_{.033}$
		EOT	$.416_{.032}$	$1.233_{.063}$	$.437_{.034}$	$.431_{.023}$	$.011_{.006}$	$2.856_{.030}$	$5.342_{.128}$	$3.367_{.025}$
	OLR-NLL ordered	MT,ST,LB	$.417_{.029}$	$1.234_{.067}$	$.436_{.035}$	$.427_{.026}$	$.010_{.005}$	$2.873_{.043}$	$5.243_{.034}$	$3.345_{.020}$
		EOT	$.412_{.029}$	$1.225_{.068}$	$.433_{.032}$	$.429_{.027}$	$.010_{.005}$	$2.845_{.037}$	$5.228_{.041}$	$3.343_{.017}$
RMSE for Task-S	AD	NNT	$.697_{.036}$	$1.623_{.094}$	$.684_{.035}$	$\mathbf{.710_{.030}}$	$.113_{.044}$	$3.962_{.040}$	$7.395_{.027}$	$4.571_{.032}$
		EOT	$.698_{.036}$	$\mathbf{1.589_{.077}}$	$\mathbf{.676_{.030}}$	$.721_{.033}$	$.108_{.044}$	$\mathbf{3.942_{.039}}$	$7.410_{.035}$	$\mathbf{4.534_{.027}}$
	Logistic-IT non-ordered	MT	$.706_{.041}$	$1.655_{.085}$	$.696_{.035}$	$.713_{.032}$	$.133_{.032}$	$4.257_{.070}$	$7.870_{.109}$	$5.151_{.078}$
		ST	$.706_{.041}$	$1.655_{.085}$	$.696_{.035}$	$.710_{.039}$	$.135_{.036}$	$4.255_{.068}$	$7.869_{.110}$	$5.058_{.065}$
		EOT	$.694_{.034}$	$\mathbf{1.585_{.079}}$	$\mathbf{.684_{.030}}$	$\mathbf{.702_{.026}}$	$\mathbf{.094_{.034}}$	$\mathbf{4.172_{.045}}$	$\mathbf{7.620_{.087}}$	$\mathbf{4.706_{.036}}$
	Logistic-IT ordered	MT,ST	$.698_{.037}$	$1.644_{.090}$	$.694_{.033}$	$.707_{.029}$	$.102_{.032}$	$5.016_{.100}$	$8.902_{.182}$	$5.720_{.062}$
		EOT	$.693_{.036}$	$\mathbf{1.592_{.078}}$	$\mathbf{.684_{.027}}$	$.709_{.031}$	$.095_{.036}$	$\mathbf{4.726_{.064}}$	$\mathbf{8.320_{.055}}$	$\mathbf{4.798_{.035}}$
	Logistic-AT ordered	MT,ST	$.696_{.035}$	$1.585_{.077}$	$.686_{.031}$	$.708_{.027}$	$.102_{.033}$	$3.859_{.043}$	$7.520_{.096}$	$4.589_{.053}$
		LB	$.696_{.038}$	$1.604_{.081}$	$.689_{.034}$	$.710_{.025}$	$.102_{.033}$	$3.858_{.043}$	$7.497_{.080}$	$\mathbf{4.517_{.040}}$
		EOT	$.694_{.036}$	$1.587_{.080}$	$.682_{.032}$	$.713_{.029}$	$.099_{.037}$	$3.841_{.034}$	$7.494_{.053}$	$\mathbf{4.517_{.022}}$
	OLR-NLL ordered	MT,ST	$.697_{.033}$	$\mathbf{1.592_{.081}}$	$.683_{.024}$	$.710_{.028}$	$.100_{.025}$	$3.870_{.037}$	$7.437_{.061}$	$4.574_{.031}$
		LB	$.696_{.032}$	$1.616_{.085}$	$.690_{.034}$	$.710_{.029}$	$.099_{.025}$	$3.869_{.036}$	$7.423_{.060}$	$\mathbf{4.508_{.029}}$
		EOT	$.693_{.035}$	$\mathbf{1.588_{.079}}$	$.682_{.026}$	$.708_{.028}$	$.096_{.030}$	$3.852_{.044}$	$7.417_{.059}$	$\mathbf{4.510_{.025}}$

7 Conclusion and Future Prospect

The NNT, MT, and ST labelings and the LB labeling in typical usages are threshold labelings that may be sub-optimal depending on the learning result of the 1DT, task under consideration, and data distribution. In this study we propose to change the labeling procedure of the existing 1DT-based methods to the EOT labeling that applies the threshold parameters minimizing the empirical task risk for a given 1DT, in order to obtain higher classification performance. Experiments in this paper showed the usefulness of this proposal.

We are also interested in the design of the learning procedure, especially the selection of the surrogate loss function, for the threshold method. One may be able to take systematic discussion on the goodness of the loss function by fixing components of the threshold method other than the loss function to the optimal ones. In such discussion, the EOT labeling will serve as the optimal other components. This is a future prospect.

References

Shivani Agarwal. Generalization bounds for some ordinal regression algorithms. In *Algorithmic Learning Theory*, pp. 7–21, 2008.

Alan Agresti. *Analysis of Ordinal Categorical Data*, volume 656. John Wiley & Sons, 2010.

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

Paul-Christian Bürkner and Matti Vuorre. Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2(1):77–101, 2019.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020.

Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *Proceedings of the European Conference on Computer Vision*, pp. 768–783, 2014.

Wei Chu and Zoubin Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6(Jul):1019–1041, 2005.

Wei Chu and S Sathiya Keerthi. New approaches to support vector ordinal regression. In *Proceedings of the International Conference on Machine Learning*, pp. 145–152, 2005.

Wei Chu and S Sathiya Keerthi. Support vector ordinal regression. *Neural Computation*, 19(3):792–815, 2007.

Joaquim F Pinto da Costa, Hugo Alonso, and Jaime S Cardoso. The unimodal model for the classification of ordinal data. *Neural Networks*, 21(1):78–91, 2008.

Stephen E Fienberg and William M Mason. Identification and estimation of age-period-cohort models in the analysis of discrete archival data. *Sociological Methodology*, 10:1–67, 1979.

Philip Hans Franses and Richard Paap. *Quantitative Models in Marketing Research*. Cambridge University Press, 2001.

Pedro Antonio Gutierrez, Maria Perez-Ortiz, Javier Sanchez-Monedero, Francisco Fernandez-Navarro, and Cesar Hervas-Martinez. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

R Herbrich, T Graepel, and K Obermayer. Support vector learning for ordinal regression. In *International Conference on Artificial Neural Networks*, volume 1, pp. 97–102, 1999.

Stefan Kramer, Gerhard Widmer, Bernhard Pfahringer, and Michael De Groeve. Prediction of ordinal classes using regression trees. *Fundamenta Informaticae*, 47(1-2):1–13, 2001.

Ling Li and Hsuan-Tien Lin. Ordinal regression by extended binary classification. In *Advances in Neural Information Processing Systems*, pp. 865–872, 2007.

Hsuan-Tien Lin and Ling Li. Large-margin thresholded ensembles for ordinal regression: Theory and practice. In *Algorithmic Learning Theory*, pp. 319–333. Springer, 2006.

Hsuan-Tien Lin and Ling Li. Reduction from cost-sensitive ordinal ranking to weighted binary classification. *Neural Computation*, 24(5):1329–1367, 2012.

Tie-Yan Liu. *Learning to Rank for Information Retrieval*. Springer Science & Business Media, 2011.

Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127, 1980.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4920–4928, 2016.

Fabian Pedregosa, Francis Bach, and Alexandre Gramfort. On the consistency of ordinal regression methods. *Journal of Machine Learning Research*, 18(Jan):1769–1803, 2017.

Sebastian Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *Journal of Open Source Software*, 3(24):638, 2018.

Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 341–345, 2006.

Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18, 2016.

Amnon Shashua and Anat Levin. Ranking with large margin principle: Two approaches. In *Advances in Neural Information Processing Systems*, pp. 961–968, 2003.

WA Thompson Jr. On the treatment of grouped observations in life studies. *Biometrics*, pp. 463–470, 1977.

Richard Williams. Generalized ordered logit/partial proportional odds models for ordinal dependent variables. *The Stata Journal*, 6(1):58–82, 2006.

Ryoya Yamasaki. Unimodal likelihood models for ordinal data. *Transactions on Machine Learning Research*, 2022. URL `https://openreview.net/forum?id=1l0sClLiPc`.

Shipeng Yu, Kai Yu, Volker Tresp, and Hans-Peter Kriegel. Collaborative ordinal regression. In *Proceedings of the International Conference on Machine Learning*, pp. 1089–1096, 2006.

A Proof of Consistency of Statistical Methods

We here give proof of the theorem on the interpretation of the learning procedure for statistical methods.

Proof of Theorem 2. We can characterize the surrogate risk minimization for the NLL loss as maximum likelihood estimation for the statistical model (10) for multi-class classification problem through the equation

$$\begin{aligned}\min_{a\in\mathcal{A},\boldsymbol{b}\in\mathcal{B}} \mathbb{E}[\phi_{\mathrm{nll}}(a(\boldsymbol{X}),\boldsymbol{b},Y;\sigma)] &= \min_{a\in\mathcal{A},\boldsymbol{b}\in\mathcal{B}} \mathbb{E}\left[\sum_{y=1}^{K}\Pr(y|\boldsymbol{X})\phi_{\mathrm{nll}}(a(\boldsymbol{X}),\boldsymbol{b},y;\sigma)\right] \\ &= \min_{a\in\mathcal{A},\boldsymbol{b}\in\mathcal{B}} \mathbb{E}\left[-\sum_{y=1}^{K}\Pr(y|\boldsymbol{X})\log P(y;\sigma,a(\boldsymbol{x}),\boldsymbol{b})\right].\end{aligned} \tag{15}$$

According to the method of Lagrange multiplier, one solution of a point-wise (at each $\boldsymbol{X}=\boldsymbol{x}$) minimization problem

$$\min_{\{\hat{\Pr}(k|\boldsymbol{x})\}_k} -\sum_{y=1}^{K}\Pr(y|\boldsymbol{x})\log\hat{\Pr}(y|\boldsymbol{x}), \quad \text{subject to } \sum_{y=1}^{K}\hat{\Pr}(y|\boldsymbol{x})=1 \tag{16}$$

is $\hat{\Pr}(y|\boldsymbol{x})=\Pr(y|\boldsymbol{x})=P(y;\sigma,\tilde{a}(\boldsymbol{x}),\tilde{\boldsymbol{b}})$, $y=1,\dots,K$, where the existence of such $\{\tilde{a}(\boldsymbol{x}),\tilde{\boldsymbol{b}}\}$ is assumed in the statement of the theorem. This solution applies for any $\boldsymbol{x}\in\mathbb{R}^d$, and one can see that a solution of (15) is $\{\tilde{a},\tilde{\boldsymbol{b}}\}$, which completes the proof of the statement for the NLL loss.

Also, for the ANLCL loss, we can provide the following characterization:

$$\begin{aligned}
&\mathbb{E}[\phi_{\mathrm{anlcl}}(a(\boldsymbol{X}), \boldsymbol{b}, Y)] \\
&= \mathbb{E}\left[-\sum_{y=1}^{K} \Pr(y|\boldsymbol{X})\left\{\sum_{k=1}^{y-1} \log\{1 - Q(k; \sigma, a(\boldsymbol{X}), \boldsymbol{b})\} + \sum_{k=y}^{K-1} \log Q(k; \sigma, a(\boldsymbol{X}), \boldsymbol{b})\right\}\right] \\
&= -\sum_{y=1}^{K-1} \mathbb{E}\left[\Pr(Y \leq y|\boldsymbol{X}) \log Q(k; \sigma, a(\boldsymbol{X}), \boldsymbol{b}) + \{1 - \Pr(Y \leq y|\boldsymbol{X})\} \log\{1 - Q(y; \sigma, a(\boldsymbol{X}), \boldsymbol{b})\}\right],
\end{aligned} \tag{17}$$

where $Q(y; \sigma, a(\boldsymbol{x}), \boldsymbol{b}) \coloneqq \sum_{k=1}^{y} P(k; \sigma, a(\boldsymbol{x}), \boldsymbol{b})$ and the expectation value $\mathbb{E}[\cdot]$ is taken for $\boldsymbol{X}$. On the ground of the binary version, 'y or less' vs. 'more than y' ($y = 1, \ldots, K-1$), of (16), one can prove the statement similarly. □

One may consider the IT loss (7) with $\varphi(u) = -\log\{\sigma(u)\}$, which we call ***immediate negative log cumulative likelihoods (INLCL) loss function***. However, it is difficult to characterize the surrogate risk minimization with the INLCL loss as a problem with a known solution unlike those for the NLL and ANLCL losses, and the optimality condition for the INLCL loss is unknown.

B Proof of Relationships between Labeling Functions

This section provides proofs of Theorems 3 and 4 regarding the relationships between the LB and threshold labelings. Propositions 1 and 2 would be trivial, so we omit proofs of them.

First, we prove Theorem 3.

Proof of Theorem 3. We introduce the functions

$$R_j(u) \coloneqq \sum_{k=1}^{K} \{\sigma(\bar{b}_k - u) - \sigma(\bar{b}_{k-1} - u)\}\ell(j,k) = \ell(j,K) + \sum_{k=1}^{K-1} \sigma(\bar{b}_k - u)\{\ell(j,k) - \ell(j,k+1)\} \text{ for } j = 1, \ldots, K, \tag{18}$$

with $\bar{b}_0 = -\infty$ and $\bar{b}_K = +\infty$, where the equation holds, since $\sigma(-\infty) = 0$ and $\sigma(+\infty) = 1$. The classifier based on the LB labeling, $f(\boldsymbol{x}) = \arg\min_{j \in [K]} \sum_{k=1}^{K} P(k; \sigma, \bar{a}(\boldsymbol{x}), \bar{\boldsymbol{b}})\ell(j,k)$, is equal to $\arg\min_{j \in [K]} R_j(\bar{a}(\boldsymbol{x}))$. According to Proposition 1, the LB labeling is a certain threshold labeling if and only if $\arg\min_{j \in [K]}\{R_j(u_1)\}_{j=1}^{K} \leq \arg\min_{j \in [K]}\{R_j(u_2)\}_{j=1}^{K}$ for any $u_1, u_2 \in \mathbb{R}$ such that $u_1 \leq u_2$. The latter condition holds if the situation

$$R_k(u) > R_l(u) \text{ for } u \in (s_1, s_2) \text{ and } R_k(u) < R_l(u) \text{ for } u \in (s_2, s_3) \text{ with } k < l,\ s_1 < s_2 < s_3 \tag{19}$$

does not occur. In the following we assume $k < l$ for the indices $k, l \in [K]$.

Proof of (i). Under the assumption described in the statement of the theorem, the difference

$$R_k(u) - R_l(u) = \underbrace{\{\ell(k,K) - \ell(l,K)\}}_{\substack{\text{non-negative} \\ \text{constant}}} + \sum_{j=1}^{K-1} \underbrace{\sigma(\bar{b}_j - u)}_{\substack{\text{non-negative} \\ \text{non-increasing}}} \underbrace{\{\ell(k,j) - \ell(k,j+1) - \ell(l,j) + \ell(l,j+1)\}}_{\substack{\text{non-positive} \\ \text{constant}}} \tag{20}$$

is non-decreasing with respect to u. Thus, $R_k(u) \leq R_l(u)$ for $u \leq p$ and $R_k(u) \geq R_l(u)$ for $u \geq p$ for a some point p, $R_k(u) \leq R_l(u)$ for any u, or $R_k(u) \geq R_l(u)$ for any u, which implies that the above-mentioned situation (19) does not occur. Note that, for the instances $\ell = \ell_{\mathrm{ad}}, \ell_{\mathrm{sq}}$, one has that

$$\ell_{k,l}(j) = \ell(k,j) - \ell(k,j+1) - \ell(l,j) + \ell(l,j+1) = \begin{cases} -2 \cdot \mathbb{1}(k \leq j \leq l-1) & \text{for } \ell = \ell_{\mathrm{ad}}, \\ 2(k-l), & \text{for } \ell = \ell_{\mathrm{sq}}. \end{cases} \tag{21}$$

This completes the proof of the statement (i).

Proof of (ii). For $\ell = \ell_{\mathrm{zo},c}$ with $c \in [0, \lfloor K/2 \rfloor)$ where $\ell_{\mathrm{zo}} = \ell_{\mathrm{zo},0}$, the function $R_j(u)$ reduces to

$$R_j(u) = 1 - \{\sigma(b_j - u) - \sigma(a_j - u)\}, \tag{22}$$

with $a_j \coloneqq \bar{b}_{\max\{0,j-c\}}$ and $b_j \coloneqq \bar{b}_{\min\{j+c,K\}}$, where $a_j < b_j$. Lemma 1 (described after the proof of Theorem 3) shows the shape of the function $R_j(u)$: Under the assumption of Theorem 3 (ii), $R_j(u)$ is minimized at $u = (a_j + b_j)/2 \coloneqq c_j$, symmetric in u around $u = c_j$, non-increasing in u for $u < c_j$, and non-decreasing in u when $u > c_j$, from Lemma 1 (i) and (ii). Also, assuming that c_j is fixed, then $R_j(u)$ is non-decreasing in $b_j - a_j$, from Lemma 1 (iii).

When $b_k - a_k = b_l - a_l$, the translated two curves $R_k(u)$ and $R_l(u)$ have just one intersection point at $u = (c_k + c_l)/2$, and it holds that $R_k(u) \le R_l(u)$ for $u \le (c_k + c_l)/2$ and $R_k(u) \ge R_l(u)$ for $u \ge (c_k + c_l)/2$. Therefore, the situation (19) does not occur if $b_k - a_k = b_l - a_l$.

Then, assume $b_k - a_k < b_l - a_l$ (the following proof strategy for this setting can be applied to the other setting $b_k - a_k > b_l - a_l$). In this setting, $R_k(u) > R_l(u)$ for $u \ge c_l$ due to the shape of the functions R_k and R_l. Also, within $[c_k, c_l]$, they can have one intersection point p at most such that $R_k(u) \le R_l(u)$ for $u \in [c_k, p]$ and $R_k(u) \ge R_l(u)$ for $u \in [p, c_l]$, since $R_k(u)$ and $R_l(u)$ are respectively non-decreasing and non-increasing in u. Therefore, the situation (19) can be satisfied only in such a situation that there exists a point p satisfying

$$R_k(p) = R_l(p),\ R_k'(p) < R_l'(p), \text{ and } p \le c_k. \tag{23}$$

The existence of such a point p implies that

$$\frac{\sigma'(a_k - p) - \sigma'(b_k - p)}{\sigma(a_k - p) - \sigma(b_k - p)} < \frac{\sigma'(a_l - p) - \sigma'(b_l - p)}{\sigma(a_l - p) - \sigma(b_l - p)} \text{ with } a_k \le a_l,\ b_k \le b_l,\ a_k \le b_k,\ a_l \le b_l,\ p \le c_k. \tag{24}$$

However, the assumption that $\frac{\sigma'(v_1) - \sigma'(v_2)}{\sigma(v_1) - \sigma(v_2)}$ is non-increasing in v_1 with fixed v_2 and in v_2 with fixed v_1 when $v_1 < v_2$ shows that

$$\frac{\sigma'(a_k - p) - \sigma'(b_k - p)}{\sigma(a_k - p) - \sigma(b_k - p)} \ge \frac{\sigma'(a_k - p) - \sigma'(b_l - p)}{\sigma(a_k - p) - \sigma(b_l - p)} \ge \frac{\sigma'(a_l - p) - \sigma'(b_l - p)}{\sigma(a_l - p) - \sigma(b_l - p)}, \tag{25}$$

which contradicts to the equation (24). Therefore, the situation (19) does not occur also when $b_k - a_k < b_l - a_l$.

Note that, especially when $\sigma = \sigma_{\mathrm{logistic}}$, one can show that

$$\begin{aligned} \frac{\sigma'(v_1) - \sigma'(v_2)}{\sigma(v_1) - \sigma(v_2)} &= \frac{\sigma_{\mathrm{logistic}}(v_1)(1 - \sigma_{\mathrm{logistic}}(v_1)) - \sigma_{\mathrm{logistic}}(v_2)(1 - \sigma_{\mathrm{logistic}}(v_2))}{\sigma_{\mathrm{logistic}}(v_1) - \sigma_{\mathrm{logistic}}(v_2)} \\ &= 1 - \{\sigma_{\mathrm{logistic}}(v_1) + \sigma_{\mathrm{logistic}}(v_2)\}, \end{aligned} \tag{26}$$

is decreasing in v_1 with fixed v_2 and in v_2 with fixed v_1. Moreover, when $\sigma = \sigma_{\mathrm{gauss}}$, one has that

$$\frac{\sigma'(v_1) - \sigma'(v_2)}{\sigma(v_1) - \sigma(v_2)} \propto \frac{e^{-v_1^2/2} - e^{-v_2^2/2}}{\sigma_{\mathrm{gauss}}(v_1) - \sigma_{\mathrm{gauss}}(v_2)} \coloneqq f_1(v_1, v_2), \tag{27}$$

that the derivative of $f_1(v_1, v_2)$ with respect to v_1,

$$\frac{\partial}{\partial v_1} f_1(v_1, v_2) = \frac{-v_1 e^{-v_1^2/2}\{\sigma_{\mathrm{gauss}}(v_1) - \sigma_{\mathrm{gauss}}(v_2)\} - \left(e^{-v_1^2/2} - e^{-v_2^2/2}\right)\frac{1}{\sqrt{2\pi}} e^{-v_1^2/2}}{\{\sigma_{\mathrm{gauss}}(v_1) - \sigma_{\mathrm{gauss}}(v_2)\}^2} \tag{28}$$

has the same sign as

$$f_2(v_1, v_2) \coloneqq -v_1\{\sigma_{\mathrm{gauss}}(v_1) - \sigma_{\mathrm{gauss}}(v_2)\} - \left(\frac{1}{\sqrt{2\pi}} e^{-v_1^2/2} - \frac{1}{\sqrt{2\pi}} e^{-v_2^2/2}\right), \tag{29}$$

and that the derivative of $f_2(v_1, v_2)$ with respect to v_2 is

$$\frac{\partial}{\partial v_2} f_2(v_1, v_2) = (v_1 - v_2)\frac{1}{\sqrt{2\pi}} e^{-v_2^2/2}. \tag{30}$$

Since $\frac{\partial}{\partial v_2} f_2(v_1, v_2) < 0$ when $v_1 < v_2$ and $f_2(v_1, v_1) = 0$, it holds that $f_2(v_1, v_2)$, which has the same sign as $\frac{\partial}{\partial v_1}\frac{\sigma'(v_1)-\sigma'(v_2)}{\sigma(v_1)-\sigma(v_2)}$, is negative when $v_1 < v_2$, that is, $\frac{\sigma'(v_1)-\sigma'(v_2)}{\sigma(v_1)-\sigma(v_2)}$ is decreasing in v_1 with fixed v_2 when $v_1 < v_2$; monotonicity in v_2 with fixed v_1 can be proved by the same discussion.

Proof of (iii). Regarding the MT and ST labelings, let $\underline{y} = h_{\mathrm{thr}}(u; \bar{\boldsymbol{b}})$ under the assumption $\bar{b}_1 \le \cdots \le \bar{b}_{K-1}$, which implies that $\bar{b}_1 \le \cdots \le \bar{b}_{y-1} \le u \le \bar{b}_y \le \cdots \le \bar{b}_{K-1}$. Regarding the LB labeling for the likelihood model (10), one has that, with the abbreviations $\sigma_k \coloneqq \sigma(\bar{b}_k - u)$ for $k = 1, \ldots, K$,

$$\begin{aligned}
R_j(u) &= \sum_{k=1}^{K} \{\sigma_k - \sigma_{k-1}\}|j-k|, \\
&= |j-1|\{\sigma_1 - \sigma_0\} + |j-2|\{\sigma_2 - \sigma_1\} + \cdots + 2\{\sigma_{j-2} - \sigma_{j-3}\} + \{\sigma_{j-1} - \sigma_{j-2}\} \\
&\quad + \{\sigma_{j+1} - \sigma_j\} + 2\{\sigma_{j+2} - \sigma_{j+1}\} + \cdots + |j-K+1|\{\sigma_{K-1} - \sigma_{K-2}\} + |j-K|\{\sigma_K - \sigma_{K-1}\} \\
&= -|j-1| \underbrace{\sigma_0}_{0} + \left\{\sum_{k=1}^{j-1} \sigma_k\right\} - \left\{\sum_{k=j}^{K-1} \sigma_k\right\} + |j-K| \underbrace{\sigma_K}_{1} \\
&= \sum_{k=1}^{j-1} \sigma(\bar{b}_k - u) + \sum_{k=j}^{K-1} \{1 - \sigma(\bar{b}_k - u)\},
\end{aligned} \tag{31}$$

for every $j \in [K]$. Simple calculations show that $\sigma(\bar{b}_k - u) \le 0.5$ for $k = 1, \ldots, y-1$ and $\{1 - \sigma(\bar{b}_k - u)\} \le 0.5$ for $k = y, \ldots, K-1$, from $\bar{b}_1 \le \cdots \le \bar{b}_{y-1} \le u \le \bar{b}_y \le \cdots \le \bar{b}_{K-1}$ and the assumption on the shape of σ. One would see that objective function (31) is minimized at $j = y$ because some summands are replaced by ones of 0.5 or more if j deviates from y, which concludes the proof. □

The following is an auxiliary lemma for the above-described proof of Theorem 3.

Lemma 1. *Suppose that σ is non-decreasing and satisfies $\sigma(-\infty) = 0$ and $\sigma(+\infty) = 1$. Define $S(u; a, b) \coloneqq \sigma(b-u) - \sigma(a-u)$ for $a < b$. Then, one has that*

(i) *$S(u; a, b)$ with fixed a and b is symmetric in u around $u = \frac{a+b}{2}$, if $\sigma(-\cdot) = 1 - \sigma(\cdot)$, or if σ is differentiable and σ' is even.*

(ii) *$S(u; a, b)$ with fixed a and b is maximized with respect to u at $u = \frac{a+b}{2}$, non-decreasing in u for $u < \frac{a+b}{2}$, and non-increasing in u for $u > \frac{a+b}{2}$, if σ is differentiable and $\sigma'(u)$ is even and non-increasing in u if $u > 0$.*

(iii) *$S(u; a, b)$ with fixed u and $\frac{a+b}{2}$ is increasing with respect to $(b-a)$.*

Proof of Lemma 1. **Proof of (i).** The assumptions that $\sigma(-\infty) = 0$, $\sigma(+\infty) = 1$, and σ' is even imply that $\sigma(-\cdot) = 1 - \sigma(\cdot)$. On the basis of this result, one then has that

$$\begin{aligned}
S\left(u + \frac{a+b}{2}; a, b\right) &= \sigma\left(b - \left\{u + \frac{a+b}{2}\right\}\right) - \sigma\left(a - \left\{u + \frac{a+b}{2}\right\}\right) \\
&= \sigma\left(\frac{b-a}{2} - u\right) - \sigma\left(-\frac{b-a}{2} - u\right) = \sigma\left(\frac{b-a}{2} - u\right) - 1 + \sigma\left(\frac{b-a}{2} + u\right),
\end{aligned} \tag{32}$$

which implies that

$$S\left(u + \frac{a+b}{2}; a, b\right) = S\left(-u + \frac{a+b}{2}; a, b\right). \tag{33}$$

Proof of (ii). The above equation (32) shows that

$$\frac{\partial}{\partial u}S\left(u+\frac{a+b}{2};a,b\right)=\sigma'\left(\frac{b-a}{2}+u\right)-\sigma'\left(\frac{b-a}{2}-u\right)=0,\quad \text{at } u=0. \tag{34}$$

Also, one can show that

$$\begin{aligned}\frac{\partial}{\partial u}S\left(u+\frac{a+b}{2};a,b\right)&=\sigma'\left(\frac{b-a}{2}+u\right)-\sigma'\left(\frac{b-a}{2}-u\right)\\&=\begin{cases}\sigma'\left(|\frac{b-a}{2}+u|\right)-\sigma'\left(\frac{b-a}{2}-u\right)\geq 0, & \text{for } u<0,\\ \sigma'\left(\frac{b-a}{2}+u\right)-\sigma'\left(|\frac{b-a}{2}-u|\right)\leq 0, & \text{for } u>0.\end{cases}\end{aligned} \tag{35}$$

Here, for $u<0$, we used the fact that σ' is even, which implies that $\sigma'(\frac{b-a}{2}+u)=\sigma'(|\frac{b-a}{2}+u|)$, and $\sigma'(v)$ is non-increasing in v for $v>0$ and $\frac{b-a}{2}-u>|\frac{b-a}{2}+u|>0$; for $u>0$, we used the fact that σ' is even, which implies that $\sigma'(\frac{b-a}{2}-u)=\sigma'(|\frac{b-a}{2}-u|)$, and $\sigma'(v)$ is non-increasing in v for $v>0$ and $\frac{b-a}{2}+u>|\frac{b-a}{2}-u|>0$.

Proof of (iii). With change of variables $t=\frac{b-a}{2}, v=\frac{a+b}{2}$, we introduce a function

$$T(t;u,v)=S(u;v-t,v+t)=\sigma(v-u+t)-\sigma(v-u-t). \tag{36}$$

For this function, one has that

$$\frac{\partial}{\partial t}T(t;u,v)=\sigma'(v-u+t)+\sigma'(v-u-t)\geq 0, \tag{37}$$

since σ is non-decreasing (i.e., $\sigma'(u)\geq 0$ for any u). □

Next, we give a proof of Theorem 4.

Proof of Theorem 4. If $k<l$, the convexity shows that

$$\begin{aligned}\ell(k,j)&\leq\frac{\{k-j\}-\{k-(j+1)\}}{\{l-j\}-\{k-(j+1)\}}\ell(k,j+1)+\frac{\{l-j\}-\{k-j\}}{\{l-j\}-\{k-(j+1)\}}\ell(l,j)\\&=\frac{1}{l-k+1}\ell(k,j+1)+\frac{l-k}{l-k+1}\ell(l,j),\end{aligned} \tag{38}$$

and that

$$\begin{aligned}\ell(l,j+1)&\leq\frac{\{l-(j+1)\}-\{k-(j+1)\}}{\{l-j\}-\{k-(j+1)\}}\ell(k,j+1)+\frac{\{l-j\}-\{l-(j+1)\}}{\{l-j\}-\{k-(j+1)\}}\ell(l,j)\\&=\frac{l-k}{l-k+1}\ell(k,j+1)+\frac{1}{l-k+1}\ell(l,j).\end{aligned} \tag{39}$$

These inequalities imply that $\ell_{k,l}$ is non-positive:

$$\begin{aligned}&\ell_{k,l}(j)\\&=\{\ell(k,j)+\ell(l,j+1)\}-\{\ell(k,j+1)+\ell(l,j)\}\\&=\{\ell(k,j)+\ell(l,j+1)\}-\left[\left\{\frac{1}{l-k+1}\ell(k,j+1)+\frac{l-k}{l-k+1}\ell(l,j)\right\}+\left\{\frac{l-k}{l-k+1}\ell(k,j+1)+\frac{1}{l-k+1}\ell(l,j)\right\}\right]\\&=\left[\ell(k,j)-\left\{\frac{1}{l-k+1}\ell(k,j+1)+\frac{l-k}{l-k+1}\ell(l,j)\right\}\right]+\left[\ell(l,j+1)-\left\{\frac{l-k}{l-k+1}\ell(k,j+1)+\frac{1}{l-k+1}\ell(l,j)\right\}\right]\\&\leq 0.\end{aligned} \tag{40}$$

Similarly, one can show that $\ell_{k,l}$ is non-negative if $k>l$. □

McCullagh (1980, Section 6.1) has proposed the heteroscedastic extension of the cmulative link model (10),

$$P_2(y;\sigma,a(\boldsymbol{x}),\boldsymbol{b},s(\boldsymbol{x})) \coloneqq P(y;\sigma,a(\boldsymbol{x})/s(\boldsymbol{x}),\boldsymbol{b}/s(\boldsymbol{x})) \tag{41}$$

with the scale model $s:\mathbb{R}^d\to(0,\infty)$, and statistical OR studies, Thompson Jr (1977); Fienberg & Mason (1979) and Agresti (2010, Section 4.2), have also considered another model

$$P_3(y;\sigma,a(\boldsymbol{x}),\boldsymbol{b}) \coloneqq \sigma(b_y-a(\boldsymbol{x}))\prod_{k=1}^{y-1}\{1-\sigma(b_{k-1}-a(\boldsymbol{x}))\}. \tag{42}$$

We obtain the following theorem that is similar to Theorem 3 and suggests the efficiency of the EOT labeling for statistical methods adopting these other likelihood models:

Theorem 5. *Suppose that σ is non-decreasing and satisfies $\sigma(-\infty)=0$ and $\sigma(+\infty)=1$ and that $\bar{a}:\mathbb{R}^d\to\mathbb{R}$, $\bar{\boldsymbol{b}}\in\mathbb{R}^{K-1}$, and $\bar{s}:\mathbb{R}^d\to(0,\infty)$.*

(i) $\arg\min_{j\in[K]}\sum_{k=1}^K P_2(k;\sigma,\bar{a}(\boldsymbol{x}),\bar{\boldsymbol{b}},\bar{s}(\boldsymbol{x}))\ell(j,k) = h_{\mathrm{thr}}(\bar{a}(\boldsymbol{x});\bar{\boldsymbol{b}})$ *if* $\ell=\ell_{\mathrm{ad}}$, $\sigma(0)=0.5$, *and* $\bar{b}_1\le\cdots\le\bar{b}_{K-1}$.

(ii) $\arg\min_{j\in[K]}\sum_{k=1}^K P_3(k;\sigma,\bar{a}(\boldsymbol{x}),\bar{\boldsymbol{b}})\ell(j,k) = h_{\mathrm{thr}}(\bar{a}(\boldsymbol{x});\boldsymbol{t})$ *for some* $\boldsymbol{t}\in\mathbb{R}^{K-1}$ *if* $\ell=\ell_{\mathrm{ad}}$.

Proof of Theorem 5. **Proof of (i).** The statement (i) of Theorem 5 is trivial from the statement (iii) of Theorem 3.

Proof of (ii). Regarding the LB labeling for the likelihood model (42), one has that, with the abbreviations $\dot{\sigma}_k\coloneqq 1-\sigma(\bar{b}_k-\bar{a}(\boldsymbol{x}))$ for $k=1,\dots,K$,

$$\begin{aligned}
R_j(\bar{a}(\boldsymbol{x})) &\coloneqq \sum_{k=1}^K P_3(k;\sigma,\bar{a}(\boldsymbol{x}),\bar{\boldsymbol{b}})\ell_{\mathrm{ad}}(j,k)\\
&= \sum_{k=1}^K\left((1-\dot{\sigma}_k)\prod_{l=1}^{k-1}\dot{\sigma}_{l-1}\right)|j-k|,\\
&= |j-1|(1-\dot{\sigma}_1)+|j-2|\dot{\sigma}_1(1-\dot{\sigma}_2)+\cdots+\dot{\sigma}_1\cdots\dot{\sigma}_{j-2}(1-\dot{\sigma}_{j-1})\\
&\quad+\dot{\sigma}_1\cdots\dot{\sigma}_j(1-\dot{\sigma}_{j+1})+\cdots+|j-K+1|\dot{\sigma}_1\cdots\dot{\sigma}_{K-2}(1-\dot{\sigma}_{K-1})+|j-K|\dot{\sigma}_1\cdots\dot{\sigma}_{K-1}\\
&= (j-1)-\left(\sum_{k=1}^{j-1}\prod_{l=1}^{k}\{1-\sigma(\bar{b}_l-\bar{a}(\boldsymbol{x}))\}\right)+\left(\sum_{k=j}^{K-1}\prod_{l=1}^{k}\{1-\sigma(\bar{b}_l-\bar{a}(\boldsymbol{x}))\}\right),
\end{aligned} \tag{43}$$

for every $j\in[K]$. One has that

$$R_{j+1}(\bar{a}(\boldsymbol{x}))-R_j(\bar{a}(\boldsymbol{x})) = 1-2\prod_{l=1}^{j}\{1-\sigma(\bar{b}_l-\bar{a}(\boldsymbol{x}))\}, \tag{44}$$

is non-decreasing in j with fixed $\bar{a}(\boldsymbol{x})$. Therefore, $\arg\min_{j\in[K]}\sum_{k=1}^K P_3(k;\sigma,\bar{a}(\boldsymbol{x}),\bar{\boldsymbol{b}})\ell_{\mathrm{ad}}(j,k)$ is the first index l such that $R_{l+1}(\bar{a}(\boldsymbol{x}))-R_l(\bar{a}(\boldsymbol{x}))\le 0$, or K if $R_{l+1}(\bar{a}(\boldsymbol{x}))-R_l(\bar{a}(\boldsymbol{x}))>0$ for all $l=1,\dots,K-1$. Also, $R_{l+1}(\bar{a}(\boldsymbol{x}))-R_l(\bar{a}(\boldsymbol{x}))$ is non-increasing in $\bar{a}(\boldsymbol{x})$, for each $l=1,\dots,K-1$. These facts show that $\arg\min_{j\in[K]}\sum_{k=1}^K P_3(k;\sigma,\bar{a}(\boldsymbol{x}),\bar{\boldsymbol{b}})\ell_{\mathrm{ad}}(j,k) = h_{\mathrm{thr}}(\bar{a}(\boldsymbol{x});\boldsymbol{t})$ with the threshold parameters t_k, $k=1,\dots,K-1$ satisfying $R_{k+1}(t_k)-R_k(t_k)=0$. □

C Optimality Guarantee of Algorithm for Empirical Optimal Threshold Labeling

Lin & Li (2006) do not describe the optimality guarantee of Algorithm 1 in their paper. As a supplement to their development, we write here the optimality guarantee of Algorithm 1.

Theorem 6. *For any task loss $\ell : [K]^2 \to [0, \infty)$, 1DT $\bar{a}$, and training data $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, the threshold parameters $\bar{\boldsymbol{t}}$ obtained by Algorithm 1 minimize the empirical task risk for a classifier based on the threshold labeling: $\bar{\boldsymbol{t}} \in \arg\min_{\boldsymbol{t} \in \mathbb{R}^{K-1}} \frac{1}{n} \sum_{i=1}^n \ell(h_{\mathrm{thr}}(\bar{a}(\boldsymbol{x}_i); \boldsymbol{t}), y_i)$.*

Proof of Theorem 6. First, we prove 'statement(j)' that, for each $k \in [K]$, $L_{j,k}$ is the minimum task risk for a task such that 1DTs $\{\bar{a}(\boldsymbol{x}_i) \mid \bar{a}(\boldsymbol{x}_i) = \bar{a}_1\}, \ldots, \{\bar{a}(\boldsymbol{x}_i) \mid \bar{a}(\boldsymbol{x}_i) = \bar{a}_{j-1}\}$ are predicted as any of $1, \ldots, k$ in a non-decreasing manner, and 1DTs $\{\bar{a}(\boldsymbol{x}_i) \mid \bar{a}(\boldsymbol{x}_i) = \bar{a}_j\}$ are predicted as k:

$$L_{j,k} = \min_{\substack{h_1, \ldots, h_j \in [k] \\ \text{s.t. } h_1 \le \cdots \le h_j = k}} \sum_{l \in [j]} \sum_{y_m \in \mathcal{Y}_l} \ell(h_l, y_m) \tag{45}$$

The statement(1), which is the starting point for mathematical induction, is trivial. Also, according to the equation,

$$\begin{aligned}
L_{j+1,k} &= \min_{l \in [k]} L_{j,l} + \sum_{y_i \in \mathcal{Y}_{j+1}} \ell(k, y_i) \\
&= \min_{l \in [k]} \left(\min_{\substack{h_1, \ldots, h_j \in [l] \\ \text{s.t. } h_1 \le \cdots \le h_j = l}} \sum_{l \in [j]} \sum_{y_m \in \mathcal{Y}_l} \ell(h_l, y_m) \right) + \left(\sum_{y_i \in \mathcal{Y}_{j+1}} \ell(k, y_i) \right) \\
&= \min_{\substack{h_1, \ldots, h_{j+1} \in [k] \\ \text{s.t. } h_1 \le \cdots \le h_j = l^* \le h_{j+1} = k}} \sum_{l \in [j+1]} \sum_{y_m \in \mathcal{Y}_l} \ell(h_l, y_m) \text{ with } l^* = \arg\min_{l \in [k]} \left(\min_{\substack{h_1, \ldots, h_j \in [l] \\ \text{s.t. } h_1 \le \cdots \le h_j = l}} \sum_{l \in [j]} \sum_{y_m \in \mathcal{Y}_l} \ell(h_l, y_m) \right) \\
&= \min_{\substack{h_1, \ldots, h_{j+1} \in [k] \\ \text{s.t. } h_1 \le \cdots \le h_{j+1} = k}} \sum_{l \in [j+1]} \sum_{y_m \in \mathcal{Y}_l} \ell(h_l, y_m),
\end{aligned} \tag{46}$$

one can find that the statement(j) holds with $j \ge$ as well.

The statement(N) shows that 1DTs $\{\bar{a}(\boldsymbol{x}_i) \mid \bar{a}(\boldsymbol{x}_i) = \bar{a}_N\}$ should be labeled as $\min(\arg\min_{l \in [K]} L_{N,l}) \coloneqq M$. Also, for

$$(\bar{h}_1, \ldots, \bar{h}_N) \in \operatorname*{arg\,min}_{\substack{h_1, \ldots, h_N \in [M] \\ \text{s.t. } h_1 \le \cdots \le h_N = M}} \sum_{l \in [N]} \sum_{y_m \in \mathcal{Y}_l} \ell(h_l, y_m), \tag{47}$$

it will also be clear that 1DTs $\{\bar{a}(\boldsymbol{x}_i) \mid \bar{a}(\boldsymbol{x}_i) = \bar{a}_1\}, \ldots, \{\bar{a}(\boldsymbol{x}_i) \mid \bar{a}(\boldsymbol{x}_i) = \bar{a}_{N-1}\}$ should be labeled as $\bar{h}_1, \ldots, \bar{h}_{N-1}$. The index I or J in Lines 9–14 tracks $\bar{h}_N (= M), \bar{h}_{N-1}, \ldots, \bar{h}_1$. Therefore, it can be found that the obtained threshold parameters are optimal. □