On the Consistency of Ordinal Regression Methods

Fabian Pedregosa, Francis Bach, Alexandre Gramfort

On the Consistency of Ordinal Regression Methods

Fabian Pedregosa
INRIA - Parietal Project-Team
Saclay, France

Francis Bach
INRIA - Sierra Project-Team
École Normale Supérieure
Paris, France

Alexandre Gramfort
Institut Mines-Telecom
Telecom ParisTech, CNRS LTCI
Paris, France

Abstract

Many of the ordinal regression algorithms that have been proposed can be viewed as methods that minimize a convex surrogate of the zero-one, absolute, or squared errors. We provide a theoretical analysis of the risk consistency properties of a rich family of surrogate loss functions, including proportional odds and support vector ordinal regression. For all the surrogates considered, we either prove consistency or provide sufficient conditions under which these approaches are consistent. Finally, we illustrate our findings on 8 different datasets.

1 Introduction

In ordinal regression the goal is to learn a rule to predict labels from an ordinal scale, i.e., labels from a discrete but ordered set. This arises often when the target variable consists of human generated ratings. Examples of ordinal scales include ("do-not-bother" $\prec$ "only-if-you-must" $\prec$ "good" $\prec$ "very-good" $\prec$ "run-to-see") in movie ratings [7], ("absent" $\prec$ "mild" $\prec$ "severe") for the symptoms of a physical disease [1] and the NRS-11 numeric rating scale for clinical pain measurement [9].

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space. Let (X, Y) be two random variables with joint probability distribution P, where X takes its values in $\mathcal{X}$ and Y is a random label taking values in a set of *ordered categories* that we will denote $\mathcal{Y} = \{1, 2, \ldots, k\}$. In the ordinal regression problem, we are given a set of n observations $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ drawn i.i.d. from $X \times Y$ and a *loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$. The goal is to learn from the training examples a measurable mapping called a *classifier* $h : \mathcal{X} \to \mathcal{Y}$ so that the *risk* given below is as small as possible:

$$\mathcal{R}_\ell(h) = \mathbb{E}_{X \times Y}(\ell(Y, h(X))) \quad . \tag{1}$$

The setting above looks similar to that of a multiclass classification problem. However, a loss function used for multiclass classification such as the 0-1 loss is not sensitive to the distance among target values. On the other hand, in the ordinal regression setting, because of the order between labels, the loss function becomes lower as the distance among classes decreases. This has been formalized as the *V-shape* property [12]. We will say that a loss function is V-shaped if its forward difference, $\Delta\ell(i, j) = \ell(i, j+1) - \ell(i, j)$, verifies $\Delta\ell(i, j) \leq 0$ for $j \leq i$ and $\Delta\ell(i, j) \geq 0$ for $i < j$.

Commonly used loss functions in ordinal regression verify the V-shape property. Examples of such functions are the *absolute error*, $\ell_{\mathcal{A}}(y, k) = |y - k|$, the *squared error*, $\ell_{\mathcal{S}}(y, k) = (y - k)^2$ and the $0 - 1$ loss.

Attempting to directly minimize Eq. (1) is not feasible in practice for two reasons. First, the probability distribution P is unknown and the risk must be minimized approximately based on the observations. Second, due to the non-convexity and discontinuity of ℓ, the risk is difficult to optimize and can lead to NP-hard problems [3, 8] (note that binary classification can be seen as a particular case of ordinal regression). It is therefore common to approximate ℓ by a function $\psi : \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}$, called a *surrogate loss function*, which has better computational properties. Here d is an integer that depends on the surrogate. For the methods that we consider this will be $1, k-1$ or k. The goal becomes to find the *decision function* f that minimizes instead the ψ-risk, defined as

$$\mathcal{R}_\psi(f) = \mathbb{E}_{X \times Y}(\psi(Y, f(X))) \ . \tag{2}$$

Fisher consistency is a desirable property for surrogate loss functions [14]. It implies that in the population setting, i.e., if the probability distribution P were

available, then optimization of the ψ-risk would yield a function with smallest possible risk, known as the *Bayes predictor* and denoted by h^*.

The paper is organized as follows. In Section 2 we present the ordinal regression models that we will consider for study. These can be broadly separated into *regression-based* and *threshold-based*. Section 3 is divided into several parts. In the first part, we extend results from Ramaswamy and Agarwal [17] and prove consistency of regression-based surrogates. Because of its practical interest, the rest of Section 3 is devoted to investigate the consistency of threshold-based surrogates. Here we present our main results, which gives sufficient conditions under which these surrogates are consistent. We finish with experiments and conclusions in Sections 4 and 5.

1.1 Related work

Fisher consistency of binary and multiclass classification for the zero-one loss has been studied for a variety of surrogate loss functions (see [2, 23] and references therein). Ramaswamy and Agarwal [17] investigated the more general setting of multiclass classification with an arbitrary loss function, a setting that includes ordinal regression. The authors proved Fisher consistency of a surrogate loss function of the absolute error for the case of $k = 3$. However, this work did not prove consistency of this surrogate for $k > 3$, nor did it prove consistency for any squared error surrogate nor for any of the threshold-based surrogates that represent the majority of traditional approaches for ordinal regression.

A related, yet different, notion of consistency is *asymptotic consistency*. A surrogate loss is said to be asymptotically consistent if the minimization of the ψ-risk converges to the optimal risk as the number of samples tends to infinity. It has also been studied in the setting of supervised learning [21, 22]. This work focuses solely on Fisher consistency, and for simplicity we will now use the term consistency to denote Fisher consistency.

Notation. Vectors and vector functions are denoted in boldface. We will denote the sequence of number from one to k as $[k] = \{1, 2, \ldots, k\}$. Through the paper we will use letter k to denote the number of classes in the target space. We denote by H the Heaviside step function, defined such that $H(x) = 1$ if $x \geq 0$, and 0 otherwise.

2 Ordinal regression models

Different methods have been proposed to learn an ordinal regression model. The *regression-based approach* treats the labels as real values. It uses a standard regression algorithm to learn a real-valued function, and then predicts by rounding to the closest label (see, e.g., Kramer et al. [10] for a discussion of this method using regression trees). In this setting we will examine consistency of two different surrogate loss functions, the absolute error (that we will denote $\psi_{\mathcal{A}}$) and the squared error (denoted $\psi_{\mathcal{S}}$), which are convex surrogates of $\ell_{\mathcal{A}}$ and $\ell_{\mathcal{S}}$, respectively. Given $\alpha \in \mathbb{R}$, $y \in [k]$, these are defined as

$$\psi_{\mathcal{A}}(y, \alpha) = |y - \alpha|, \quad \psi_{\mathcal{S}}(y, \alpha) = (y - \alpha)^2 \quad . \tag{3}$$

Note that the loss functions $\ell_{\mathcal{A}}$ and $\ell_{\mathcal{S}}$ have the same expression than their surrogates, however the difference strives in that the surrogates are continuous functions in their second arguments while the loss functions take values in the discrete set $[k]$. The prediction function for these surrogates is given by rounding[1] to the closest integer in $[k]$, i.e., $\text{pred}(\alpha) = \min_{i \in [k]} |i - \alpha|$.

While these approaches may lead to optimal predictors when no constraint is placed on the regressor function space as we will see in Section 3.2, in practice only simple function spaces are explored such as linear or polynomial functions. In these situations, the regression-based approach might lack flexibility. The *threshold-based approaches* [6, 13, 15, 18] provides greater flexibility by seeking for both a mapping $f : \mathcal{X} \to \mathbb{R}$ and a non-decreasing vector $\boldsymbol{\theta} \in \mathbb{R}^{k-1}$, often referred to as *thresholds*, that map the class labels into ordered real values.

Instead of manipulating both the function f and the thresholds, we will find useful to create the auxiliary functions $g_i(x) = \theta_i - f(x)$ and to express the surrogate with them.

In the context of threshold-based functions we will consider two different families of surrogate loss functions. The first family of surrogate loss function that we will consider is the *cumulative link models* of McCullagh [15]. In such models the posterior probability is modeled as $P(Y \leq i | X = x) = \sigma(g_i(x))$, where σ is an appropriate link function. We will prove consistency for the case where σ is the sigmoid function, i.e., $\sigma(t) = 1/(1 + \exp(-t))$. In this case it is known as the *proportional odds* model or *cumulative logit* model. For $x \in \mathcal{X}, y \in [k]$ and $\alpha_i = g_i(x)$, the proportional odds surrogate (denoted $\psi_{\mathcal{C}}$) is defined as

$$\psi_{\mathcal{C}}(y, \boldsymbol{\alpha}) = \begin{cases} -\log(\sigma(\alpha_1)) & \text{if } y = 1 \\ -\log(\sigma(\alpha_y) - \sigma(\alpha_{y-1})) & \text{if } 1 < y < k \\ -\log(1 - \sigma(\alpha_{k-1})) & \text{if } y = k. \end{cases} \tag{4}$$

[1] Although this definition is ambiguous for half-integers, the particular values in a set of null measure are not important for our analysis.

The second family of surrogate loss functions that we will consider are the *margin-based* surrogate loss functions. For appropriate real-valued functions $\phi : \mathbb{R} \to \mathbb{R}$ such as the hinge loss or exponential loss, this surrogate separate target values by the largest margins centered around the thresholds [13]. Given $x \in \mathcal{X}, y \in [k]$ and $\alpha_i = g_i(x)$, the margin-based surrogate (denoted $\psi^\ell_{\mathcal{M}}$) is given by

$$\psi^\ell_{\mathcal{M}}(y, \boldsymbol{\alpha}) = \sum_{i=1}^{y-1} \Delta\ell(y, i)\phi(\alpha_i) - \sum_{i=y}^{k-1} \Delta\ell(y, i)\phi(-\alpha_i) \ .$$

We recall that $\Delta\ell(y, i) = \ell(y, i + 1) - \ell(y, i)$. Note that the V-shape property implies $\Delta\ell(y, i) \geq 0$ for the elements in the first term and $\Delta\ell(y, i) \leq 0$ for elements in the second term, thus this surrogate is convex in its second argument if ϕ is a convex function.

This formulation parametrizes several popular approaches to ordinal regression. For example, let ϕ be the hinge loss and ℓ the zero-one loss, then ψ^T_ℓ coincides with the "Explicit Threshold" formulation in [6]. If instead the mean absolute loss is considered, this approach coincides with the "Implicit Threshold" formulation of the same reference. For other values of ϕ and ℓ this loss includes the approaches proposed in [6, 13, 18, 20]. In section 3.5 we will prove consistency results for arbitrary V-shaped loss function.

Given $\boldsymbol{\alpha} \in \mathbb{R}^{k-1}$, the prediction functions for threshold-based models is

$$\text{pred}(\boldsymbol{\alpha}) = 1 + \sum_{i=1}^{k-1} H(\alpha_i) \ .$$

Since we aim at predicting a finite number of labels with a specific loss functions, it is also possible to use generic multiclass formulations such as the one proposed in [11] which can take into account generic losses. Given ϕ a real-valued function, this formulations considers the following surrogate

$$\psi^\ell_{\mathcal{L}}(y, \boldsymbol{\alpha}) = \sum_{i=1}^{k} \ell(y, i)\phi(-\alpha_i) \tag{5}$$

for $\boldsymbol{\alpha} \in \mathbb{R}^k$ such that $\sum_{i=1}^k \alpha_i = 0$. The prediction function in this case is given by $\text{pred}(\boldsymbol{\alpha}) = \arg\max_{i\in[k]} \alpha_i$. Note however that this method requires the estimation of k decision functions. For this reason, in practical settings threshold-based are often preferred as these only require the estimation of one decision function and $k - 1$ thresholds.

Consistency results of this surrogate was proven by Zhang [24]. We will compare their results to our findings of consistency for threshold-based surrogates in Section 3.6.

Table 1 contains a list of the aforementioned surrogate loss functions, the (non-surrogate) loss function they target and their prediction function.

Table 1: Loss functions and their surrogates.

Loss	**Surrogate**	**Prediction**
Absolute error	$\lvert y - \alpha\rvert$	$\min_{i\in[k]} \lvert i - \alpha\rvert$
Squared error	$(y - \alpha)^2$	$\min_{i\in[k]} \lvert i - \alpha\rvert$
Absolute error	$\psi_{\mathcal{C}}(y, \boldsymbol{\alpha})$	$1 + \sum_{i=1}^{k-1} H(\alpha_i)$
Any V-shaped	$\psi^\ell_{\mathcal{M}}(y, \boldsymbol{\alpha})$	$1 + \sum_{i=1}^{k-1} H(\alpha_i)$
Any	$\psi^\ell_{\mathcal{L}}(y, \boldsymbol{\alpha})$	$\arg\max_{i\in[k]} \alpha_i$

3 Consistency of Surrogate Loss Functions

We will now give a precise definition for the (Fisher) consistency of a surrogate loss function. This notion originates from a classical parameter estimation setting. Suppose that an estimator T of some parameter θ is defined as a functional of the empirical distribution F_n, $T(F_n)$. The estimator is said to be Fisher consistent if its population analog, $T(F)$, coincides with the parameter θ. Adapting this notion to the context of risk minimization (in which the optimal risk is the parameter to estimate) yields the following definition, adapted from [14] to an arbitrary loss ℓ.

Definition (**Consistency**) Given a surrogate loss function $\psi : \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}$, a function space $\mathcal{F}$ and prediction rule pred : $\mathbb{R}^d \to [k]$, we will say that the pair (ψ, pred) is consistent with respect to the loss ℓ if for every probability distribution over $X \times Y$ it is verified that every minimizer of the ψ-risk reaches Bayes optimal risk, that is,

$$f^* \in \arg\min_{f\in\mathcal{F}} \mathcal{R}_\psi(f) \implies \mathcal{R}_\ell(\text{pred} \circ f^*) = \mathcal{R}_\ell(h^*) \quad .$$

By an abuse of notation we will refer to the consistency of a surrogate function ψ to designate the consistency of the pair (ψ, pred).

When the ψ-risk minimization is performed over all measurable functions, it is verified that

$$\inf_f \mathcal{R}_\psi(f) = \inf_f \mathbb{E}_{X\times Y} \left(\psi(Y, f(X))\right) = \mathbb{E}_X \left[\inf_f \mathbb{E}_Y(\psi(Y, f(X)))\right] \quad . \tag{6}$$

Hence in this case in order to compute the decision function with optimal risk it is sufficient to compute the decision function with minimal expected value (over Y) for every $x \in \mathcal{X}$.

3.1 Bayes predictor

In order to prove consistency of a surrogate loss we will find useful to have an explicit form for the Bayes predictor. For example, in the case of binary classification with the zero-one loss, Bayes predictor is known and is given by $\text{sign}(P(y=1|X=x)-1/2)$. In this section we will derive similar results for arbitrary V-shaped loss functions.

We first introduce the following notation. Let $\eta_i(x) = P(Y=i|X=x)$ denote the conditional probability at $X=x$. For $1 \leq i < k$ we also define the functions $u_i, v_i : \mathcal{X} \to \mathbb{R}$ as

$$\begin{aligned} u_i(x) &= \sum_{j=1}^{i} \eta_j(x) \Delta\ell(j,i) \\ v_i(x) &= -\sum_{j=i+1}^{k} \eta_j(x) \Delta\ell(j,i) \quad . \end{aligned} \tag{7}$$

If ℓ is V-shape, then $\Delta\ell(j,i)$ is positive for $j \geq i$ and $(u_1(x), u_2(x), \ldots, u_k(x))$ is a increasing positive sequence. Similarly, $\Delta\ell(j,i) \leq 0$ for $i < j$ and $(v_1(x), v_2(x), \ldots, v_k(x))$ is a decreasing positive sequence.

We now derive a formula for the Bayes predictor of an arbitrary V-shaped loss function.

Theorem 1 (Bayes predictor for an ordinal regression loss)**.** *Let $\ell(i,j)$ be a V-shaped loss function. Then the Bayes predictor is given by*

$$h^*(x) = 1 + \sum_{i=1}^{k-1} H(v_i(x) - u_i(x)) \quad . \tag{8}$$

Proof. Let $x \in \mathcal{X}$ and $r = h^*(x)$, then by the V-shape property we have $(v_i - u_i) > 0$ for $1 \leq i < r$ and $(v_i - u_i) \leq 0$ for $i \geq r$ since $(v_i(x) - u_i(x))$ is a non-increasing sequence of i.

We will first prove $\mathbb{E}_Y(\ell(Y,r)) - \mathbb{E}_Y(\ell(Y,s)) \leq 0$ for any $s \in [k]$. Suppose $s > r$, then we have

$$\begin{aligned} &\mathbb{E}_Y(\ell(Y,r)) - \mathbb{E}_Y(\ell(Y,s)) = \\ &\sum_{i=r}^{s-1} \mathbb{E}_Y(\ell(Y,i) - \ell(Y,i+1)) = \\ &\sum_{i=r}^{s-1} \left(-\sum_{j=1}^{k} \eta_j(x) \Delta\ell(j,i) \right) = \sum_{i=r}^{s-1} (v_i(x) - u_i(x)) \leq 0 \end{aligned}$$

Similarly, for $s < r$

$$\begin{aligned} &\mathbb{E}_Y(\ell(Y,r)) - \mathbb{E}_Y(\ell(Y,s)) = \\ &\sum_{i=s}^{r-1} \mathbb{E}_Y(\ell(Y,i+1)) - \ell(Y,i) = \\ &\sum_{i=s}^{s-1} \left(\sum_{j=1}^{k} \eta_j(x) \Delta\ell(j,i) \right) = -\sum_{i=s}^{r-1} (v_i(x) - u_i(x)) \leq 0 \end{aligned}$$

We have proven that for any classifier h

$$\mathbb{E}_Y(\ell(Y,h^*(x))|X=x) - \mathbb{E}_Y(\ell(Y,h(x))|X=x) \leq 0$$

Integrating both sides with respect to X yields

$$\mathcal{R}(h^*) \leq \mathcal{R}(h) \quad ,$$

that is, h^* is the Bayes predictor. □

An immediate consequence of this theorem is that the Bayes predictor for the mean absolute error and the mean squared error admit the following simple form:

Corollary 1. *. The Bayes predictor for the absolute error loss is given by*

$$h^*(x) = \min_{r \in [k]} \{ r : \sum_{i=1}^{r} \eta_i(x) > \frac{1}{2} \} \quad . \tag{9}$$

Corollary 2. *. The Bayes predictor for the squared error loss is given by*

$$h^*(x) = \min_{r \in [k]} \{ r : \frac{1}{k} \sum_{i=1}^{k} i\eta_i(x) > r \} \quad . \tag{10}$$

3.2 Consistency of regression-based models

We will now examine the consistency of regression-based models. Consistency of the absolute error surrogate was proven by [17] for the case of 3 classes. Here we give an alternate proof that extends beyond $k > 3$. This proof is constructive in the sense that it gives an explicit form for the function that minimizes the ψ-risk. Using similar techniques we also prove consistency for the squared error surrogate.

Lemma 1. *The function with minimal $\psi_{\mathcal{A}}$-risk is $f^*(x) = \text{median}(Y|X=x)$, where* median *represents the median of a random variable (i.e. the value α such that $P(y \leq \alpha|X=x) \geq 1/2$ and $P(y \geq \alpha|X=x) \leq 1/2$). The function with minimal $\psi_{\mathcal{S}}$-risk is $f^*(x) = \mathbb{E}_Y(Y|X=x)$.*

Proof. By the application of optimality properties of the median and mean, the median and the mean are the scalar values that minimize $\mathbb{E}_Y(\psi_{\mathcal{A}}(Y,\alpha)|X=x) = \mathbb{E}_Y(|Y-\alpha||X=x)$ and $\mathbb{E}_Y(\psi_{\mathcal{S}}(Y,\alpha)|X=x) = \mathbb{E}_Y((Y-\alpha)^2|X=x)$, respectively. In light of Eq. (6) this is sufficient to obtain the minimal risk. □

Theorem 2. *The absolute error surrogate $\psi_{\mathcal{A}}$ is consistent with respect to $\ell_{\mathcal{A}}$.*

Proof. Let $x \in \mathcal{X}$, and $\alpha^* = \text{median}(Y|X = x)$. By definition of median, $\sum_{i=1}^{\alpha^*} \eta_i(x) > 1/2$ and $\sum_{\alpha^*}^{k} \eta_i(x) < 1/2$ (note that we can ignore the set of values where $\sum_{i=1}^{\alpha^*} \eta_i(x) = 1/2$).

These inequalities together with Eq. (9) imply that $\alpha^* \in [r - 1/2, r + 1/2]$, where $r = h^*(x)$ is the Bayes predictor from Eq. (9). Hence this surrogate and Bayes predictor have the same prediction except for a set of measure zero, which implies consistency. □

Theorem 3. *The squared error surrogate $\psi_{\mathcal{S}}$ is consistent with respect to $\ell_{\mathcal{S}}$.*

Proof. Let $\alpha^* = \mathbb{E}_Y(Y|X = x) = \sum_{i=1}^{k} i\eta_i(x)/k$. Then $\text{pred}(\alpha^*) = \text{round}(\sum_{i=1}^{k} i\eta_i(x)/k) = \min_{i \in [k]} \sum_{i=1}^{k} i\eta_i(x)/k > i$ which coincides with the Bayes predictor from Eq. (10). □

3.3 Difficulty of consistency in the threshold-based setting

Although the threshold-based setting is of great practical importance, no consistency results exist for these surrogates to the best of our knowledge.

The difficulty of proving such results stems from the fact that within the space of allowed decision functions Eq. (6) is no longer valid. This implies that it is no longer possible to obtain the optimal decision function from the minimization at a fixed $x \in \mathcal{X}$, as we have done in the proof of Theorem 2 and 3.

In section 2, we have defined the decision function $\mathbf{g}(x) = (g_1(x), \dots, g_{k-1}(x))$ to be of the form $g_i(x) = \theta_i - f(x)$, or equivalently to verify the condition that $g_{i+1}(x) - g_i(x)$ is a positive constant (i.e. does not depend on x) for all $1 \leq i < k - 1$. If $\mathbf{g}$ verifies this constraint, we will say that $\mathbf{g}$ is a *threshold-based decision function.*

In order to obtain sufficient conditions for the consistency of threshold-based methods, we will first consider the case in which the decision function $\mathbf{g}$ belongs to the space of all measurable functions. In this case we can construct the optimal decision function by considering each $x \in \mathcal{X}$ separately. Having an explicit form of the minimizer for the ψ-risk in this setting makes it possible to inspect under which conditions does this minimizer belong to the space of threshold-based decision functions.

An interesting relaxation of the threshold-based setting is given in [16] under the name of *partial thresholds.* In this setting, $\mathbf{g}(x) = (g_1(x), \dots, g_k(x))$ is a non-decreasing vector for all $x \in \mathcal{X}$ which does not necessarily verify the constraints of a threshold-based decision function. In this setting, the decision function can represent any real-valued mapping that verifies the order constraints. We will call these decision functions *partial-threshold decision functions.* This setting is rarely used in practice because of the need to estimate $k - 1$ functions.

3.4 Consistency of proportional odds

We begin by proving the strong convexity of proportional odds, whose proof can be found in the appendix. Through this section we will use $\psi_{\mathcal{C}}$ to denote the proportional odds surrogate as defined in Eq. (4).

Lemma 2. *The proportional odds surrogate $\psi_{\mathcal{C}}$ is a convex function of its arguments in the domain of definition.*

For the proportional odds surrogate $\psi_{\mathcal{C}}$ it is possible to find the explicit form of a function that minimizes the $\psi_{\mathcal{C}}$-risk. We will use notation $\mathbf{g}$ to denote the vector-valued function $(g_1(x), \dots, g_{k-1}(x))$.

Theorem 4. *The function $\mathbf{g} : \mathcal{X} \to \mathbb{R}^{k-1}$ given by*

$$g_i^*(x) = \log\left(\frac{u_i(x)}{1 - u_i(x)}\right) \quad ,$$

minimizes the $\psi_{\mathcal{C}}$-risk.

Proof. Let $x \in \mathcal{X}$ and consider the optimization problem

$$\boldsymbol{\alpha}^* \in \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^{k-1}} \mathbb{E}_Y(\psi_{\mathcal{C}}(Y, \boldsymbol{\alpha})|X = x)$$

The KKT conditions associated with this optimization problem are

$$\begin{aligned}
&-\eta_1(x)\frac{1}{\sigma(\alpha_1)} + \eta_2(x)\frac{1}{\sigma(\alpha_2) - \sigma(\alpha_1)} = 0 \\
&-\eta_i(x)\frac{1}{\sigma(\alpha_i) - \sigma(\alpha_{i-1})} + \eta_{i+1}(x)\frac{1}{\sigma(\alpha_{i+1}) - \sigma(\alpha_i)} = 0 \\
&-\eta_{k-1}(x)\frac{1}{\sigma(\alpha_{k-1}) - \sigma(\alpha_{k-2})} + \eta_k(x)\frac{1}{1 - \sigma(\alpha_{k-1})} = 0
\end{aligned}$$

with $1 < i < k - 1$. It is easy to verify that $\sigma(\alpha_i^*) = \sum_{j=1}^{i} \eta_j(x) = u_i(x)$ satisfy the optimality conditions.

Solving for α_i^* results in $\sigma(\alpha_i^*) = \sum_{j=1}^{i} \eta_j(x) \implies \alpha_i^* = \log\left(u_i(x)/(1 - u_i(x))\right)$. By Eq. (6), the function that for all $x \in \mathcal{X}$ returns $\log\left(u_i(x)/(1 - u_i(x))\right)$ is the function that minimizes the ψ-risk. □

Note that for $x \in \mathcal{X}$ fixed, the sequence $(g_1^*(x), \dots, g_{k-1}^*(x))$, with g^* as defined in the previous theorem is non-decreasing since u_i is non-decreasing and due to the monotonicity of the logit function. This

implies that $\mathbf{g}(x) = (g_1^*(x), \ldots, g_{k-1}^*(x))$ is a partial-threshold decision function. Consistency for this class of functions is now immediate since

$$\begin{aligned}\text{pred}(\mathbf{g}^*(x)) &= 1 + \sum_{i=1}^{k-1} H\left(\log\left(\frac{u_i(x)}{1-u_i(x)}\right)\right) = \\ &= 1 + \sum_{i=1}^{k-1} H\left(u_i(x) - 1/2\right) \\ &= \min_{r\in[k]}\{r : \sum_{i=1}^{r} \eta_i(x) > \frac{1}{2}\}\end{aligned} \tag{11}$$

which coincides with the Bayes classifier from Eq. (9). Thus, if the decision function belongs to the space of partial-threshold decision functions, the proportional odds is consistent. For threshold-based decision functions we have the following result:

Corollary 3. *Let P verify the property that the* odds-ratio *is constant, that is,*

$$\frac{\eta_i(x)/(1-\eta_i(x))}{\eta_{i+1}(x)/(1-\eta_{i+1}(x))} \tag{12}$$

is independent of $x \in \mathcal{X}$ for all $i \in [k-1]$. Then the proportional odds surrogate is consistent.

3.5 Consistency of margin-based models

As done in the previous section, we will provide an explicit form of functions that minimize the $\psi^\ell_{\mathcal{M}}$-risk. This will allow to derive conditions under which threshold-based decision functions are consistent.

Theorem 5. *Let ℓ be V-shaped. Then the function $\mathbf{g} : \mathcal{X} \to \mathbb{R}^{k-1}$ minimizes the $\psi^\ell_{\mathcal{M}}$-risk for different values of ϕ:*

- *If ϕ is the hinge loss, i.e., $\phi(t) = \max(1-t, 0)$,*

$$g_i^*(x) = \text{sign}(u_i(x) - v_i(x))$$

- *If ϕ is the logistic loss, i.e., $\phi(t) = 1/(1+\exp(-t))$,*

$$g_i^*(x) = \log(u_i(x)/v_i(x))$$

- *If ϕ is the exponential loss, i.e., $\phi(t) = \exp(-t)$*

$$g_i^*(x) = \frac{1}{2}\log(u_i(x)/v_i(x))$$

- *If ϕ is the squared loss, i.e., $\phi(t) = (1-t)^2$*

$$g_i^*(x) = \frac{u_i(x) + v_i(x)}{u_i(x) - v_i(x)}$$

Proof. Let u_i, v_i be as defined in Eq. (7), $x \in \mathcal{X}$ and $\boldsymbol{\alpha} = (g_1(x), \ldots, g_{k-1}(x))$. Then for any surrogate ψ we can write

$$\begin{aligned}&\mathbb{E}_Y(\psi(Y, \boldsymbol{\alpha})|X = x) = \\ &\sum_{j=1}^{k} \eta_j(x) \left(\sum_{i=1}^{j-1} \Delta\ell(y,i)\phi(\alpha_i) - \sum_{i=j}^{k-1} \Delta\ell(y,i)\phi(-\alpha_i)\right) = \\ &\sum_{i=1}^{k-1} \phi(\alpha_i)v_i(x) + \phi(-\alpha_i)u_i(x) \quad .\end{aligned} \tag{13}$$

If ϕ is the hinge loss, the values of α_i that minimize this expression verify $-1 \leq \alpha_i \leq 1$ for all $i \in [k-1]$, as otherwise truncation of these values at -1 or 1 gives a lower value of the surrogate loss. In this case we have

$$\begin{aligned}&\mathbb{E}_Y(\psi(Y, \boldsymbol{\alpha})|X = x) = \\ &\sum_{i=1}^{k-1}(1-\alpha_i)v_i(x) + (1+\alpha_i)u_i(x) = \\ &\sum_{i=1}^{k-1}\alpha(u_i(x) - v_i(x)) + C\end{aligned}$$

where C are terms that do not depend on α. Therefore, this expression minimized for $\alpha_i^* = \text{sign}(v_i(x) - u_i(x))$.

If ϕ is the logistic loss, the expression from Eq. (7) is differentiable. The derivative with respect to α_i is $(1-\sigma(\alpha_i))v_i - \sigma(\alpha_i)u_i$, where $\sigma(\alpha_i) = 1/(1+\exp(-\alpha_i))$ is the sigmoid function. Equaling this expression to zero and solving for α_i yields the result.

The proof for ψ the rest of surrogates can be found in the appendix. □

In light of this result, it is possible to derive sufficient conditions under which margin-based decision functions are consistent.

Corollary 4. *Under the conditions of Theorem 5, if P is a probability distribution such that*

$$\alpha_i^*(x) - \alpha_{i+1}^*(x)$$

does not depend on x for all $1 \leq i < k$, then the surrogate $\psi^\ell_{\mathcal{M}}$ is consistent.

Proof. The optimal decision functions $\alpha_1^*, \ldots, \alpha_{k-1}^*$ are threshold-based decision functions by assumption. Furthermore, it is easy to verify that all the $\alpha_i^*(x)$ obtained in Theorem 5 verify $H(\alpha_i^*(x)) = H(u_i(x) - v_i(x))$, and thus prediction coincides with the Bayes predictor of Eq. (8). □

The sufficient conditions of Corollary 4 translate into well-known conditions on the probability distribution

for some values of ϕ. For example, let ϕ be the logistic loss and ℓ be the absolute error, the optimal decision function is given by $\alpha_i^*(x) = \log(u_i(x)/v_i(x)) = \log(u_i(x)/(1 - u_i(x)))$. Thus we obtain the function that the optimal decision function for the proportional odds from Theorem 4. This implies (see Corollary 3) that if P verifies that the odds-ratio are constant as defined in (14), then the surrogate is consistent.

As mentioned in Section 2, the surrogate $\psi^\ell_{\mathcal{M}}$ parametrizes several approaches that have appeared in the literature. For the zero-one loss these include:

- "Explicit threshold" from [6] (ϕ = hinge loss),
- "Immediate threshold" from [18] (ϕ = logistic loss),
- "ORBoost with Left-Right margins" from [13] (ϕ = exponential loss),

Likewise, for the mean absolute error these include:

- "Implicit threshold" from [6] (ϕ = hinge loss),
- "All threshold" from [18] (ϕ = logistic loss),
- "ORBoost with All margins" from [13] (ϕ = exponential loss).

Corollary 4 provides sufficient conditions on the probability distribution P for these approaches to be consistent.

In light of these results, it is immediate to show that within the space of partial threshold decision functions, the aforementioned methods are consistent. Furthermore, in this case we can prove a slightly more general result. The following result states consistency while assuming only convexity and a condition of the differential at zero of the function ϕ.

Theorem 6. *Let ℓ be a V-shaped loss function. Given a convex function $\phi : \mathbb{R} \to \mathbb{R}_+$ such that ϕ is differentiable at zero and $\phi'(0) < 0$, then the surrogate loss function $\psi^\ell_{\mathcal{C}}$ is consistent with respect to ℓ if we consider partial-threshold decision functions .*

Proof. See appendix. □

3.6 Relationship with multiclass formulations

Let $\psi^\ell_{\mathcal{L}}$ the surrogate loss function defined in Eq. (5). For a given $x \in \mathcal{X}$, let $f_1^*(x), \dots, f_k^*(x)$ be minimizers of $\mathbb{E}_Y(\psi^\ell_{\mathcal{L}}(Y, f(x)))$. Then it is verified

$$\sum_{i=1}^{k} \left(\sum_{j=1}^{k} \eta_j(x) \ell(j, i) \right) \psi(-f_i(x)) = \sum_{i=1}^{k} (u_i(x) - v_i(x)) \psi(-f_i(x))$$

For the hinge loss, it is shown in Lee et al. [11] that given $x \in \mathcal{X}$, the optimal decision function is of the form $f_i^*(x) = 1$ for $i = \arg\min_i u_i(x) - v_i(x)$, and $-1/(k-1)$ otherwise. Thus, a sufficient condition for consistency is that the k functions above are in the class of functions we are considering for the decision function.

This is to be contrasted with the margin-based formulations, where, for the hinge surrogate, we need the $k - 1$ functions $\text{sign}(u_i(x) - v_i(x)))$ to be in the class of functions of the decision function.

No requirement is stronger than the other. However, for the margin-based formulations, we have developed sufficient conditions under which we may use a single function and fixed thresholds.

4 Experiments

Although the focus of this work is a theoretical investigation of consistency, we have also conducted experiments that study empirical performance of some the methods outlined in this paper. In this section we compare two approaches described earlier in terms of generalization accuracy. The different datasets used are described in [5]. Following [6], we will consider two variants of the margin-based loss function $\psi_{\mathcal{C}}$ for $\ell = \ell_{0-1}$ and $\ell = \ell_{\mathcal{A}}$ with ϕ = hinge loss. Specifically, we compare the "Explicit threshold formulation" (denoted here ET) versus the "Implicit threshold formulation" (denoted IT). Corollary 4 states that under appropriate assumptions on the probability distribution P, ET is consistent with respect to the zero-one loss while AT is consistent with respect to the absolute error loss.

We show in Figure 1 the generalization scores of these two methods using as metric the zero-one loss and the absolute error on 8 different datasets. The generalization accuracy of both models has been computed using 5-fold cross validation. Although consistency results only apply under certain assumptions on the underlying probability distribution, we observe a correlation between consistent surrogates and the best performing model. Our findings provide a theoretical explanation of the poor performance of the ET surrogate compared with the IT surrogate when evaluated using the absolute error loss (since the IT surrogate is consistent w.r.t

the absolute error). Similar results have been observed in the literature for different values of ϕ [6, 13, 20].

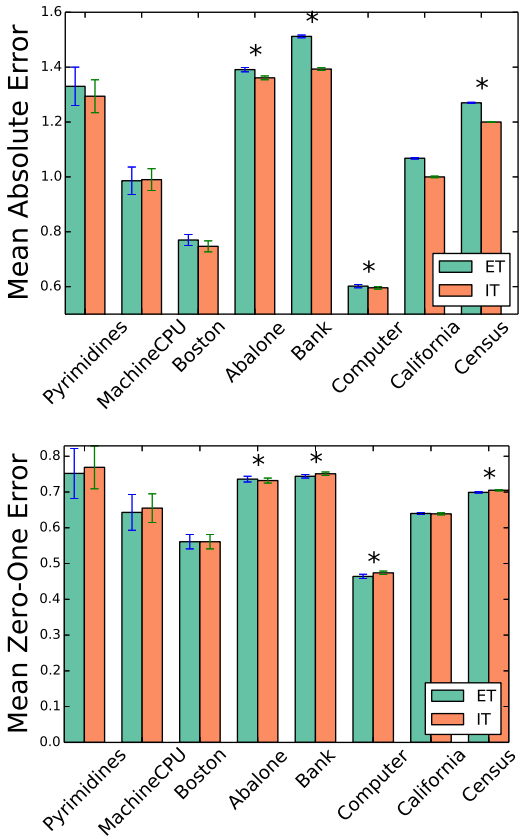

Figure 1: Performance of the "Explicit Threshold" (ET) and "Implicit Threshold" (IT) methods of Chu and Keerthi [6] on 8 different datasets and for two different evaluation metrics. Top: the metric used is the mean absolute error. The IT method is consistent with respect to this loss and performs better on 7 out of 8 datasets. Bottom: the metric used is the mean zero-one loss. The ET method is consistent with respect to this metric and performs better on 6 out of 8 datasets. Datasets for which the difference of performance is significant (Wilcoxon signed-rank test with $p < 0.01$) are denoted with an asterisk ($*$).

5 Conclusion

In this paper we have characterized the consistency for a rich family of surrogate loss functions used for ordinal regression. In the regression-based setting we have extended work from Ramaswamy and Agarwal [17] to prove consistency for the absolute error surrogate as well as the squared error surrogate.

In the threshold-based setting, we studied consistency of the proportional odds model and given sufficient conditions on the underlying probability distribution under which this surrogate is consistent. We also considered formulations such as the Support Vector Ordinal Regression [6], the Ordinal Regression Boosting methods [13] and the Logistic Regression formulation of [18]. We framed these methods under a common formulation that we call *margin-based surrogate*, and derived an explicit form of functions that minimize the ψ-risk. We also gave sufficient conditions for the consistency of the aforementioned approaches.

Since consistency of the threshold-based approach is only proven subject to certain conditions on the underlying probability distribution P, we investigated under which conditions these surrogates are always consistent. Here we show that this is possible by considering an enlarged space for the decision functions that we called partial-threshold decision functions.

Finally, we illustrated our findings on by comparing the performance of two methods on 8 different datasets. Although the conditions for consistency that are required by the underlying probability distribution are not necessarily met, we observed a significantly better performance of the consistent methods versus the non-consistent method.

5.1 Acknowledgments

This work was supported by grants IRMGroup ANR-10-BLAN-0126-02 and BrainPedia ANR-10-JCJC 1408-01. We would like to thank our colleague Guillaume Obozinski for fruitful discussions.

References

[1] Ben G. Armstrong and Margaret Sloan. Ordinal regression models for epidemiologic data. *American Journal of Epidemiology*, 129(1):191–204, 1989.

[2] 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, 2003.

[3] Shai Ben-David, Nadav Eiron, and Philip M. Long. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496–514, 2003.

[4] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[5] Wei Chu and Zoubin Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1–24, 2004.

[6] Wei Chu and S Sathiya Keerthi. New Approaches to Support Vector Ordinal Regression. In *Pro-*

ceedings of the 22th International Conference on Machine Learning (ICML), 2005.

[7] Koby Crammer and Yoram Singer. Pranking with ranking. In *Advances in Neural Information Processing Systems 14*, 2001.

[8] Vitaly Feldman, Venkatesan Guruswami, Prasad Raghavendra, and Yi Wu. Agnostic learning of monomials by halfspaces is hard. *SIAM Journal on Computing*, 41(6):1558–1590, 2012.

[9] Craig T. Hartrick, Juliann P. Kovan, and Sharon Shapiro. The numeric rating scale for clinical pain measurement: A ratio measure? *Pain Practice*, 3 (4):310–316, 2003. ISSN 1533-2500.

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

[11] Yoonkyung Lee, Yi Lin, and Grace Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

[12] Ling Li and Hsuan-tien Lin. Ordinal Regression by Extended Binary Classification. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2007.

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

[14] Yi Lin. A note on margin-based loss functions in classification. *Statistics & Probability Letters*, 68 (1):73 – 82, 2004.

[15] Peter McCullagh. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society*, 42 (2):109–142, 1980.

[16] Bercedis Peterson and Frank E. Harrell. Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society*, 39:205–217, 1990.

[17] Harish G Ramaswamy and Shivani Agarwal. Classification Calibration Dimension for General Multiclass Losses. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2012.

[18] Jason D M Rennie and Nathan Srebro. Loss Functions for Preference Levels : Regression with Discrete Ordered Labels. In *Proceedings of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling*, 2005.

[19] Ralph Rockafellar. On the maximal monotonicity of subdifferential mappings. *Pacific Journal of Mathematics*, 33(1):209–216, 1970.

[20] Amnon Shashua and Anat Levin. Ranking with large margin principle : Two approaches. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2003.

[21] Ingo Steinwart. Support Vector Machines are Universally Consistent. *Journal of Complexity*, 18(3): 768–791, September 2002.

[22] Charles J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, pages 595–620, 1977.

[23] Ambuj Tewari and Peter L. Bartlett. On the Consistency of Multiclass Classification Methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.

[24] Tong Zhang. Statistical Behavior and Consistency of Classification Methods based on Convex Risk Minimization. *The Annals of Statistics*, 32: 56–85, 2004.

Appendix

Lemma 3 (Lemma 2)**.** *The proportional odds surrogate loss is a convex function of its arguments in the domain of definition.*

Proof of Lemma 2. $\psi_{\mathcal{C}}(1,\alpha)$ and $\psi_{\mathcal{C}}(k,\alpha)$ are logistic loss functions, which are convex because they are log-sum-exp functions. We will prove that ψ_i is convex for $1 < i < K$. For convenience we will write this function as $f(a,b) = -\log\left(\frac{1}{1+\exp(a)} - \frac{1}{1+\exp(b)}\right)$, where $a > b$.

By factorizing the fraction inside f to a common denominator, f can equivalently be written as $-\log(\exp(a)-\exp(b))+\log(1+\exp(a))+\log(1+\exp(b))$. The last two terms are convex because they can be written as a log- sum-exp. The convexity of the first term, or equivalently the log- convexity of the function $f(a,b) = \exp(a) - \exp(b)$ can be settled by proving the positive- definiteness of the matrix $Q = f(a,b)\nabla^2 f(a,b) - \nabla f(a,b)\nabla f(a,b)^T$ for all (a,b) in the domain $\{b > a\}$ [4]. In our case,

$$Q = \begin{pmatrix} \exp(a+b) & -\exp(a+b) \\ -\exp(a+b) & \exp(a+b) \end{pmatrix}$$

Let $\lambda_1 \geq \lambda_2$ be the eigenvalues of Q. Since $\det(Q) = \lambda_1\lambda_2 = 0$, one of the eigenvalues is zero. From the identity between the trace and the eigenvalues of a symmetric matrix, $\mathrm{tr}(Q) = \lambda_1 + \lambda_2 = 2\exp(a+b)$. From here we can conclude $\lambda_1 = 2\exp(a+b)$ and $\lambda_2 = 0$. This proves that Q is positive semidefinite and thus the loss function Ψ_i is convex. □

Corollary 5 (Proof of Corollary 3)**.** *Let P verify the property that the* odds-ratio *is constant, that is,*

$$\frac{\eta_i(x)/(1-\eta_i(x))}{\eta_{i+1}(x)/(1-\eta_{i+1}(x))} \tag{14}$$

is independent of $x \in \mathcal{X}$ for all $i \in [k-1]$. Then the proportional odds surrogate is consistent.

Proof. Let $g_i(x) = \log\left(u_i(x)/(1-u_i(x))\right)$ and $g_{i+1}(x) = \log\left(u_{i+1}(x)/(1-u_{i+1}(x))\right)$. Proving is that $g_i(x) - g_{i+1}(x)$ is constant is equivalent to proving that g is of the form $g_i(x) = \theta_i - f(x)$

Then

$$\begin{aligned} g_i(x) - g_{i+1}(x) = \log\left(u_i(x)/(1-u_i(x))\right) - \\ \log\left(u_{i+1}(x)/(1-u_{i+1}(x))\right) = \\ \log\left(\frac{\eta_i(x)/(1-\eta_i(x))}{\eta_{i+1}(x)/(1-\eta_{i+1}(x))}\right) \end{aligned}$$

which is the log of a constant by assumption, hence constant. By Theorem 4 it follows that this function is the minimizer of the $\psi_{\mathcal{C}}$-risk. Consistency is now a consequence of (11). □

Theorem 7 (Theorem 6)**.** *Let ℓ be a V-shaped loss function. Given a convex function $\phi : \mathbb{R} \to \mathbb{R}_+$ such that ϕ is differentiable at zero and $\phi'(0) < 0$, then the surrogate loss function $\psi_{\mathcal{C}}^{\ell}$ is consistent with respect to ℓ if we consider partial-threshold decision functions .*

Proof. Let $x \in \mathcal{X}$ and $r = h^*(x)$. As we did in the proof of Theorem 5, we can write $\mathbb{E}_Y(\psi(y,\boldsymbol{\alpha})|X = x) = \sum_{i=1}^{k-1}\phi(\alpha_i)v_i(x) + \phi(-\alpha_i)u_i(x)$. The KKT conditions for this optimization problem with respect to $\boldsymbol{\alpha}$ (taking into account that $\boldsymbol{\alpha}$ should be non-decreasing) are

$$\begin{aligned} 0 \in \partial F_i(\alpha_i) = \partial\phi(\alpha_i)u_i - \partial\phi(-\alpha_i)v_i - \lambda_{i-1} + \lambda_i, \\ \forall i = 1,\ldots,\ k-1 \quad \lambda_0 = \lambda_k = \lambda_i(\alpha_i - \alpha_{i+1}) = 0, \quad \lambda_i \geq 0 \end{aligned} \tag{15}$$

where ∂ denotes the subgradient operator.

By hypothesis ϕ is differentiable at zero. Thus, $\partial F_r(0) = \phi'(0)(u_r - v_r) - \lambda_{r-1} + \lambda_r$. Let $s \in [k]$ be the largest integer such that $\lambda_r\lambda_{r+1}\ldots\lambda_s > 0$. Because of the slack conditions, this implies that at the optimum $\alpha_r = \alpha_{r+1} = \cdots = \alpha_{s+1}$. The addition of $\partial F_j(0)$ from $j = r$ to s verifies $\sum_{j=r}^{s}\partial F_j(0) = \phi'(0)\left(\sum_{i=r}^{s}u_i - \sum_{i=r}^{s}v_i\right) - \lambda_{i-1} \leq 0$. The expression $\sum_{j=r}^{s}\partial F_j(\alpha_j)$ is the subdifferential of a convex function and is thus a monotone operator [19]. This implies that any zero of $\sum_{j=r}^{s}\partial F(\alpha_j)$ (and thus any solution of the KKT equations) will be located in the region $\alpha_r \geq 0$.

Suppose $r > 1$ and consider $\partial F_{r-1}(0) = \phi'(0)(u_r - v_r) - \lambda_{r-2} + \lambda_{r-1}$. Let t be the smallest integer that verifies $\lambda_t, \lambda_{t+1}, \ldots, \lambda_{r-2} > 0$. This implies that at the optimum $\alpha_t = \alpha_{t+1} = \cdots = \alpha_{r-1}$. The expression $\sum_{j=t}^{r-1}\partial F_j(\alpha_j)$ is again a monotone operator and verifies $\sum_{j=t+1}^{r-1}\partial F(0)_j = \phi'(0)\left(\sum_{j=t}^{r-1}u_i - \sum_{j=t}^{r-1}v_i\right) + \lambda_{r-1} \geq 0$ from where we can conclude that any zero of $\sum_{j=t}^{r-1}\partial F_j(\alpha_j)$ will be located in the region $\alpha_{r-1} \leq 0$.

If $\alpha_r > 0$ and α_{r-1} then our prediction rule would predict class r, thus the approach is consistent. □