On the Consistency of Ordinal Regression Methods

Fabian Pedregosa, Francis Bach, Alexandre Gramfort

On the Consistency of Ordinal Regression Methods

Fabian Pedregosa-Izquierdo[1,2], **Francis Bach**[2,3], **Alexandre Gramfort**[4]
[1] Parietal Project-Team, INRIA Saclay Île-de-France, France
[2] Sierra Project-Team, INRIA Rocquencourt Île-de-France, France
[3] École Normale Supérieure de Paris, Paris, France
[4] Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, Paris, France

Abstract

Ordinal regression is a common supervised learning problem sharing properties with both regression and classification. 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 extend the notion of consistency which has been studied for classification, ranking and some ordinal regression models to the general setting of ordinal regression. We study a rich family of these surrogate loss functions and assess their consistency with both positive and negative results. For arbitrary loss functions that are admissible in the context of ordinal regression, we develop an approach that yields consistent surrogate loss functions. Finally, we illustrate our findings on real-world datasets.

1 Introduction

Supervised learning is commonly applied in situations where the targets to predict are ordered: prediction of the severity level of a disease from medical images [7] or the user rating for a new product [14]. In contrast to regression problems, the labels are discrete and finite. The setting is also different from multiclass classification due to the existence of an ordering among the labels. This is a learning task with target variables of ordinal scale, a setting bridging between metric regression and classification referred to as *ordinal regression*.

Different methods have been proposed for ordinal regression. The *regression-based approach* treats the labels as real values, 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. [8]). While these approaches may lead to optimal predictors when no constraint is placed on the regressor function space, in practice with high-dimensional problems only simple function spaces are explored such as linear functions. In these situations, the regression-based approach might lack flexibility. The *threshold-based approach* is a more flexible approach in which a regression model is learned together with a set of thresholds that model the different labels as intervals on the real line, an idea whose origin dates back to the classical cumulative model of McCullagh [11]. A number of approaches have extended binary-classification models to the ordinal regression setting with the help of thresholds. It is the case for support vector machines [4], logistic regression [13] and boosting [10].

The performance of these models is measured by the output of a given *loss function*. For each pair of labels the loss function will return the disagreement between the two labels. A simple example is the zero-one loss: a function that will output zero if both labels are equal and one otherwise. Since loss functions are typically difficult to optimize directly we will use an approximation called a *surrogate loss function*. Consistency is a desirable property of surrogate loss functions. It implies that at the limit when the number of observations is infinite, optimization of the surrogate loss yields an optimal solution, known as the *Bayes predictor*. The study of consistency properties for binary and

multiclass classification with respect to the zero-one loss have been studied thoroughly in [1, 17–19]. The ranking formulation has also been studied in details [5, 6] where it has been shown that existing pairwise approaches were not consistent with respect to the pairwise disagreement. A more general case was studied in [12] by taking into account general multiclass losses specified by a *loss matrix*. They extend the notion of *classification calibration* to this setting. They prove consistency of a regression model with respect to the absolute error loss, but their analysis extends neither to the squared error nor to the popular threshold-based formulations of ordinal regression.

Our contribution is to characterize the consistency of different surrogate loss functions in the context of ordinal regression, both in the regression-based and threshold-based formulations. More precisely, our contribution is to (**Result 1**) extend the proof in [12] on the consistency of the absolute error for an arbitrary number of classes and provide a proof for the consistency of the squared error loss, (**Result 2**) develop a procedure that in the threshold-based setting yields consistent surrogate loss functions for any admissible loss function and identify models that are consistent in this setting with respect to the zero-one and absolute error, (**Result 3**) prove consistency of the proportional odds model with respect to the mean absolute error, and finally, (**Result 4**) investigate the setting of threshold-based loss functions in which thresholds are constant across samples, a setting of practical importance. We show that results in the general setting do not translate necessarily into this setting. We do this by identifying a surrogate loss function that is consistent in the general threshold-based setting but which becomes inconsistent in the setting of constant thresholds across samples.

Notation. Vectors and vector functions are denoted in boldface. Δ_k denotes the probability simplex in dimension k, that is $\Delta_k = \{\mathbf{p} \in \mathbb{R}^k : \mathbf{p}_i \geq 0 \ \forall i, \ \sum_{i=1}^k \mathbf{p}_i = 1\}$. The set of non-negative number is denoted by $\mathbb{R}_+ = [0, \infty)$. We will denote the sequence of number from one to k as $[k] = \{1, \ 2, \ldots, \ k\}$. Matrices are considered in row-major ordering, that is, given a matrix $\mathbf{A}$, $\mathbf{A}_i$ denotes the i-th row. We will use symbol ∂ to denote the subgradient operator.

2 Problem Setting

In the ordinal regression problem, we are given a set of n training samples $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ drawn i.i.d. from a distribution P on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is an instance space and $\mathcal{Y}$ is a finite set of k *ordered* categories. Without loss of generality, these categories are denoted by $\mathcal{Y} = [k]$. We are also given a *loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ such that $\ell(y, \hat{y})$ returns the penalty incurred when predicting $\hat{y}$ when the true label is y. In this paper we will further assume that the loss function is symmetric, that is, $\ell(i, j) = \ell(j, i)$ for all possible values of i and j.

In ordinal regression, the loss is sensitive to the distance among classes and becomes lower as the distance among classes decreases. For this reason, we consider as *admissible* loss functions those that verify the *V-shape property* [9], that is, $\ell(i, j+1) \geq \ell(i, j)$ for $j \leq i$ and $\ell(i, j+1) \leq \ell(i, j)$ for $i < j$, in which case we will say that ℓ is V-shaped. Commonly used loss functions in ordinal regression that verify the V-shape property include the *absolute error*, $\ell^{\text{AE}}(y, k) = |y - k|$, the *squared error*, $\ell^{\text{SE}}(y, k) = (y - k)^2$ and the *zero-one error*, $\ell^{\text{0-1}}(y, k)$ which is zero when $y = k$ and one otherwise.

The goal is to learn from the training examples a measurable mapping $h : \mathcal{X} \to [k]$ so that the ℓ-risk given below is as small as possible:

$$\mathcal{R}_\ell(h) = \mathbb{E}_{XY} \ell(Y, h(X)) = \mathbb{E}_X \left[\sum_{y=1}^k P(Y|X) \ell(Y, h(X)) \right] = \mathbb{E}_X \left[\boldsymbol{\eta}(X)^T \mathbf{L}_{h(X)} \right], \qquad (1)$$

where $\mathbb{E}_{XY}$ denotes the expectation with respect to the true (but unknown) underlying distribution P, $\boldsymbol{\eta}(X) = (P(1|X), \ldots, P(k|X)) \in \Delta^k$ denotes the conditional probability vector at X and $\mathbf{L}$ is the *loss matrix* $\mathbf{L} \in \mathbb{R}^{k \times k}$ given by $L_{ij} = \ell(i, j)$. The minimum possible risk is called the Bayes risk and the function h that minimizes this risk is usually referred to as the *Bayes predictor*. Estimating the function h that achieves the Bayes risk is typically difficult computationally, consequently one usually employs an approximation to the true loss which is easier to optimize. We will call this function a *surrogate loss function*.

In practical applications, the distribution P is unknown and the expected ℓ-risk is approximated by the empirical ℓ-risk, which replaces the expected value by an average over training samples. Here, as

it is commonplace in consistency studies, we consider the "population" setup in which the expected ℓ-risk is available.

Different approaches have been proposed to learn an ordinal regression model, leading to different surrogate loss functions. In the *regression-based approach*, a mapping $f : \mathcal{X} \to \mathbb{R}$ is learned and prediction is defined by rounding $f(X)$ to the closest discrete label. In this setting we will examine consistency of two different surrogate loss functions, the absolute error ($\Psi^{\text{AE}} : [k] \times \mathbb{R} \to \mathbb{R}$) and the squared error ($\Psi^{\text{SE}} : [k] \times \mathbb{R} \to \mathbb{R}$), which are convex surrogates of the ℓ^{AE} and ℓ^{SE} loss functions respectively

$$\Psi^{\text{AE}}(y, f(X)) = |y - f(X)|, \quad \Psi^{\text{SE}}(y, f(X)) = (y - f(X))^2 \quad . \tag{2}$$

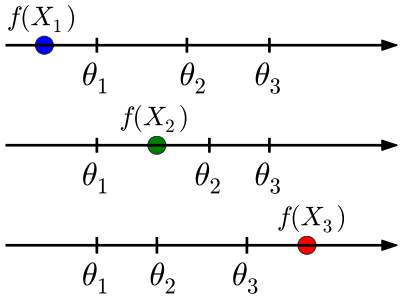

Figure 1: Example of prediction in the threshold-based approach for a 4-class problem. The prediction $f(X)$ for a given sample is denoted by a colored circle and $\theta_1, \theta_2, \theta_3$ are the estimated thresholds for that sample. Prediction in this example would be 1, 2, 4 respectively.

The *threshold-based* approach [4, 11, 13] increases the flexibility of the model by estimating not only a function f but also a set of thresholds that map the labels into ordered real values. More precisely, we will learn k functions $f,\ \theta_1,\ \theta_2, \ldots, \theta_{k-1} : \mathcal{X} \to \mathbb{R}$ such that $\theta_1(X) \leq \theta_2(X) \leq \cdots \leq \theta_{k-1}(X)$ for all $X \in \mathcal{X}$. For convenience we will set $\theta_0(X) = -\infty$ and $\theta_k(X) = \infty$. For a given $X \in \mathcal{X}$, $\boldsymbol{\theta}(X) = (\theta_1(X), \ldots, \theta_{k-1}(X)) \in \mathbb{R}^{k-1}$ will partition the real line into k segments. As illustrated in Figure 1, prediction for a sample $x \in \mathcal{X}$ is given by the cardinality of the interval that includes $f(x)$, that is, $\text{pred}(X) = i$ such that $f(X) \in [\theta_{i-1}, \theta_i)$.

We have considered the general case in which the thresholds can take different values for different samples. However, in practice the values of $\boldsymbol{\theta}$ are often assumed to be constant across samples, that is, $\boldsymbol{\theta}(X)$ does not depend on the value of X. We will first examine consistency in the setting of *free thresholds* where the thresholds are allowed to take different values for different samples. We will see that these results do not always extend to the setting of *common thresholds*, that is, the setting in which thresholds are constant across samples.

In the context of threshold-based models we will consider two different families of surrogate loss functions. The first family of surrogate loss functions can be written as a sum of binary-class loss functions and is denoted *sum-of-loss*. This surrogate $\Psi_{\boldsymbol{\Lambda}} : [k] \times \mathbb{R}^{k-1} \times \mathbb{R} \to \mathbb{R}$ is of the form

$$\Psi_{\boldsymbol{\Lambda}}(y, \boldsymbol{\theta}(X), f(X)) = \sum_{i=1}^{y-1} \Lambda_{y,i} \phi(f(X) - \theta_i(X)) + \sum_{i=y}^{k-1} \Lambda_{y,i} \phi(\theta_i(X) - f(X)) \tag{3}$$

for a given binary-class loss function $\phi : \mathbb{R} \to \mathbb{R}_+$ such as the logistic or hinge loss and a matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{(k-1)\times k}$. The matrix $\boldsymbol{\Lambda}$ is required to have non-negative entries to preserve convexity. This surrogate loss function parametrizes several popular approaches to ordinal regression. For example, let ϕ be the hinge loss and $\boldsymbol{\Lambda}$ be defined as $\Lambda_{y,i} = 1$ when $i \in \{y, y-1\}$ and $\Lambda_{y,i} = 0$ otherwise, then $\Psi_{\boldsymbol{\Lambda}}$ coincides with the "Explicit Threshold" formulation in [4]. If instead $\boldsymbol{\Lambda}$ is constantly 1 then this approach coincides with the "Implicit Threshold" formulation of the same paper. For other values of ϕ and $\boldsymbol{\Lambda}$ this loss includes the approaches proposed in [4, 10, 13, 16].

The second family of surrogate loss functions is the *cumulative link models* of McCullagh [11]. With such models the posterior probability reads $P(y \leq i|X) = \sigma(\theta_i(X) - f(X))$, where σ is an appropriate link function. We will prove consistency for the case where σ is the sigmoid function, in which case this is known as the *proportional odds* model or *cumulative logit* model. This surrogate loss function $\Psi^{\text{CL}} : [k] \times \mathbb{R}^{k-1} \times \mathbb{R}$ (given by the negative log-likelihood) in this case is given by

$$\Psi^{\text{CL}}(y, \boldsymbol{\theta}, f(X)) = -\log(\sigma(\theta_y(X) - f(X)) - \sigma(\theta_{y-1}(X) - f(X))) \quad . \tag{4}$$

In both regression and threshold-based approaches, the optimal f and $\boldsymbol{\theta}$ will be chosen to minimize the Ψ-risk. We will use the following formulation of Ψ-risk for both settings while keeping in mind

that in the regression-based approach the thresholds will simply have no influence in the surrogate loss function. This is given by:

$$\mathcal{R}_\Psi(\boldsymbol{\theta}, f) = \mathbb{E}_{XY}[\Psi(y, \boldsymbol{\theta}(X), f(X))] = \mathbb{E}_X \left[\boldsymbol{\eta}(X)^T \boldsymbol{\Psi}(\boldsymbol{\theta}(X), f(X))\right], \tag{5}$$

where $\boldsymbol{\Psi}(\boldsymbol{\theta}(X), f(X))$ denotes the vector $(\Psi(1, \boldsymbol{\theta}(X), f(X)), \ldots, \Psi(k, \boldsymbol{\theta}(X), f(X)))$.

It is natural to investigate the conditions which guarantee that if the Ψ-risk of f gets close to the optimal then the ℓ-risk of f also approaches the Bayes risk. This is often referred to as the *calibration* or *consistency* of the surrogate loss function Ψ. The main purpose of this paper is to characterize calibrated surrogate loss function in the context of ordinal regression.

Related work. Consistency is well studied for binary and multiclass classification [1, 17–19]. The notion of consistency was extended in [12] to arbitrary loss functions specified by a loss matrix, a setting that includes ordinal regression loss functions. However, Ramaswamy and Agarwal [12] only proved consistency for a specific instance of regression-based surrogate and did not consider the threshold-based approach. Rennie and Srebro [13] identified and systematized some of the approaches in the threshold-based setting based on the optimization of surrogate loss functions.

3 Calibration of Surrogate Loss Functions

In this section we will study calibration in both the regression and the threshold-based approaches to ordinal regression. We will consider the following surrogate loss functions introduced in the previous section. For simplicity we will omit the argument in $\boldsymbol{\theta}(X)$ and $f(X)$ when there is no source of confusion, so that $\boldsymbol{\theta} = \boldsymbol{\theta}(X)$ and $f = f(X)$.

Name	Expression	Prediction
regression-based	$\lvert y - f\rvert$, $(y - f)^2$	$\min_{i\in[k]} \lvert f - i\rvert$
sum-of-loss	$\sum_{i=1}^{y-1} \Lambda_{y,i}\phi(f - \theta_i) + \sum_{i=y}^{k-1} \Lambda_{y,i}\phi(\theta_i - f)$	i such that $f \in [\theta_{i-1}, \theta_i)$
cumulative link	$-\log(\sigma(\theta_y - f) - \sigma(\theta_{y-1} - f))$	i such that $f \in [\theta_{i-1}, \theta_i)$

The surrogate loss function will be of the form $\Psi : [k] \times \mathbb{R}$ in the regression-based setting and $\Psi : [k] \times \mathbb{R}^{k-1} \times \mathbb{R}$ in the threshold-based setting. We will now state the definition of calibration for both settings.

Definition (Calibration [12]) A surrogate loss function $\Psi : [k]\times\mathbb{R} \to \mathbb{R}_+$ (or $\Psi : [k]\times\mathbb{R}^{k-1}\times\mathbb{R} \to \mathbb{R}_+$ in the threshold-based setting) is said to be calibrated with respect to a loss function ℓ whenever there exists a function pred : $\mathbb{R} \to [k]$ (or pred : $\mathbb{R}^{k-1} \times \mathbb{R} \to [k]$ in the threshold-based setting) such that for all distributions P on $\mathcal{X} \times [k]$ and all sequence of random functions $f_n : \mathcal{X} \to \mathcal{T}$,

$$\mathcal{R}_\Psi(f_n) \xrightarrow{P} \mathcal{R}^*_\Psi \quad \text{implies} \quad \mathcal{R}(\text{pred} \circ f_n) \xrightarrow{P} \mathcal{R}^*.$$

When the minimization of the Ψ-risk is done over all measurable functions, the image of f and $\boldsymbol{\theta}$ at a sample $X \in \mathcal{X}$ is essentially independent of the image of f at any other sample. It is thus possible to minimize over all $X \in \mathcal{X}$ independently to obtain the optimal Ψ-risk. That is,

$$\mathcal{R}^*_\Psi = \inf_{\boldsymbol{\theta}, f} \mathcal{R}_\Psi(\boldsymbol{\theta}, f) = \inf_{\boldsymbol{\theta}, f} \mathbb{E}_X \left[\boldsymbol{\eta}(X)^T \boldsymbol{\Psi}(\boldsymbol{\theta}, f)\right] = \mathbb{E}_X \left[\inf_{\boldsymbol{\theta}, f} \boldsymbol{\eta}(X)^T \boldsymbol{\Psi}(\boldsymbol{\theta}, f)\right] \tag{6}$$

where the infimum is taken across all measurable functions for f and across all measurable functions that verify $\theta_1(X) \leq \theta_2(X) \leq \cdots \leq \theta_{k-1}(X)$, $\forall X \in \mathcal{X}$. The property that the image of f and $\boldsymbol{\theta}$ at a sample is independent of the image at any other sample is verified for the regression-based approach and for the free-thresholds approach, that is, the threshold-based approach in which the threshold are allowed to take different values for different samples. However, in the constant-threshold approach the thresholds do not depend on the training samples and must be learned for the entire population. In this case Equation (6) is not verified. We will examine this case in detail in Section 3.4.

Theorem 1. *[12] Let Ψ be a surrogate loss function. Then Ψ is calibrated with respect to the loss function ℓ with loss matrix $\mathbf{L}$ if and only if there exists a function pred : $\mathbb{R} \to [k]$ such that $\forall \mathbf{p} \in \Delta_k$*

$$\inf_{f, \boldsymbol{\theta}} \mathbf{p}^T \boldsymbol{\Psi}(\boldsymbol{\theta}, f) < \inf_{f, \boldsymbol{\theta}} \left\{ \mathbf{p}^T \boldsymbol{\Psi}(\boldsymbol{\theta}, f) : \ \text{pred}(\boldsymbol{\theta}, f) \notin \arg\min_{i\in[k]} \mathbf{p}^T \mathbf{L}_i \right\} . \tag{7}$$

This characterization conveys the message that given $(f^*, \boldsymbol{\theta}^*)$ that minimize the point-wise Ψ-risk, given by $\mathbf{p}^T\boldsymbol{\Psi}(\boldsymbol{\theta}, f)$, then pred$(\boldsymbol{\theta}, f)$ predicts according to Bayes predictor, given by $\arg\min_{i\in[k]} \mathbf{p}^T\mathbf{L}_i$. Calibration in this form has been described as a "pointwise form of Fisher consistency" [1].

3.1 Calibration in the regression-based approach

Our first result states the calibration of surrogate loss functions in the regression-based setting.

Result 1. *The surrogate functions $\Psi^{AE}(y, f) = |y - f|$, $\Psi^{SE}(y, f) = (y - f)^2$, are calibrated with respect to the absolute and squared error, respectively.*

Proof. Consistency of Ψ^{AE} was established by Ramaswamy and Agarwal [12] for the case where the number of classes equals 3. In the appendix we provide both a generalization of that proof to an arbitrary number of classes and an original proof for the consistency for the squared error surrogate. □

3.2 Calibration in the sum-of-loss approach

In this section we will describe a procedure that uses the sum-of-loss approach to generate calibrated surrogate loss functions for any V-shaped loss function. We will first introduce the matrix $\boldsymbol{\Lambda}(\ell)$ and state a property that will be needed in the proof of the results. Result 2 states the main result of this section, that is, the consistency of the surrogate $\Psi_{\boldsymbol{\Lambda}}$ with respect to ℓ.

Definition Let ℓ be a V-shaped loss function. We define the matrix $\boldsymbol{\Lambda}(\ell) \in \mathbb{R}^{(k-1)\times k}$ as $\boldsymbol{\Lambda}(\ell)_{ij} = \ell(i, j+1) - \ell(i, j)$ for $i \geq j$ and $\ell(i, j) - \ell(i, j+1)$ otherwise. The V-shaped condition on ℓ implies that all entries in matrix $\boldsymbol{\Lambda}(\ell)$ are non-negative.

Lemma 1. *Given a V-shaped loss function ℓ, $\mathbf{p} \in \Delta_k$ we define the k-dimensional vectors $\mathbf{u}$ and $\mathbf{v}$ as $u_i = \sum_{j=1}^{i} p_j\Lambda_{ij}$ and $v_i = \sum_{j=i+1}^{k} p_j\Lambda_{ij}$. Let r is such that $r \in \arg\min_{i\in[k]} \mathbf{p}^T\mathbf{L}_i$. Then it is verified $(\sum_{i=r}^{s} u_i \geq \sum_{i=r}^{s} v_i)$ for all s such that $k \geq s > r$, with strict inequality unless $\{r, r+1, \ldots, s\} \in \arg\min_{i\in[k]} \mathbf{p}^T\mathbf{L}_i$. It is also verified $\left(\sum_{i=s}^{j-1} u_i \leq \sum_{i=j}^{r-1} v_i\right)$ for all $r > s \geq 1$, with strict inequality unless $\{s, s+1, \ldots, r\} \in \arg\min_{i\in[k]} \mathbf{p}^T\mathbf{L}_i$*

Result 2. *Let ℓ be a V-shaped loss function. Given a convex function $\phi : \mathbb{R} \to \mathbb{R}_+$ consistent with respect to the zero-one binary loss, i.e., ϕ is differentiable at zero and $\phi'(0) < 0$ [1], the following surrogate loss function*

$$\Psi_{\boldsymbol{\Lambda}}(y, \boldsymbol{\theta}, f) = \sum_{i=1}^{y-1} \Lambda(\ell)_{y,i}\phi(f - \theta_i) + \sum_{i=y}^{k-1} \Lambda(\ell)_{y,i}\phi(\theta_i - f) \quad ,$$

is consistent with respect to ℓ.

Proof. Let $\mathbf{p} \in \Delta_k$, $r \in \arg\min_{i\in[k]} \mathbf{p}^T\mathbf{L}_i$ and $\mathbf{u}, \mathbf{v}$ be as defined in Lemma 1. By the change of variable $g_i = \theta_i - f$ and , the expression $\inf_{f,\boldsymbol{\theta}} \mathbf{p}^T\boldsymbol{\Psi}(\boldsymbol{\theta}, f)$ can be written as

$$\inf_{f,\boldsymbol{\theta}} \mathbf{p}^T\boldsymbol{\Psi}(\theta, f) = \inf_{\boldsymbol{g}} \sum_{j=1}^{k} p_j \left(\sum_{i=1}^{j-1} \Lambda(\ell)_{y,i}\phi(-g_i) + \sum_{i=j}^{k-1} \Lambda(\ell)_{y,i}\phi(g_i) \right) = \inf_{\boldsymbol{g}} \sum_{i=1}^{k-1} F_i(g_i) \tag{8}$$

where $F_i(g_i) = \phi(-g_i)v_i + \phi(g_i)u_i$ and the infimum is taken over the vectors $\boldsymbol{g} \in \mathbb{R}^{k-1}$ that satisify the condition $g_1 \leq g_2 \leq \cdots \leq g_{k-1}$. The KKT conditions for this optimization problem are

$$\begin{aligned} 0 \in \partial F_i(g_i) = \partial\phi(g_i)u_i - \partial\phi(-g_i)v_i - \lambda_{i-1} + \lambda_i, \\ \forall i = 1, \ldots, k-1 \quad \lambda_0 = \lambda_k = \lambda_i(g_i - g_{i+1}) = 0, \quad \lambda_i \geq 0 \end{aligned} \tag{9}$$

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 $g_r = g_{r+1} = \cdots = g_{s+1}$. The addition of $\partial F_j(0)$ from $j = r$ to s verifies $\sum_{j=r}^{s} \partial F_j(0) = \phi'(0)\left(\sum_{i=r}^{s} u_i - \sum_{i=r}^{s} v_i\right) - \lambda_{i-1} \leq 0$ where inequality is strict unless $\{r, r+1, \ldots, s\} \in \arg\min_{i\in[k]} \mathbf{p}^T \mathbf{L}_i$. The expression $\sum_{j=r}^{s} \partial F_j(g_j)$ is the subdifferential of a convex function and is thus a monotone operator [15]. This implies that any zero of $\sum_{j=r}^{s} \partial F(g_j)$ (and thus any solution of the KKT equations) will be located in the region $g_r \geq 0$, with equality only if $r + 1 \in \arg\min_{i\in[k]} \mathbf{p}^T \mathbf{L}_i$.

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 $g_t = g_{t+1} = \cdots = g_{r-1}$. The expression $\sum_{j=t}^{r-1} \partial F_j(g_j)$ 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 $\sum_{j=t}^{r-1} \partial F_j(g_j)$ will be located in the region $g_{r-1} \leq 0$, with equality only if $r - 1 \in \arg\min_{i\in[k]} \mathbf{p}^T \mathbf{L}_i$.

If $g_r > 0$ and $g_{r-1} \leq 0$ then $\theta_{r-1} \leq f < \theta_r$ and our prediction rule would predict class r. In case $g_r = 0$, that is, $f = \theta_r$ it is verified that $r + 1 \in \arg\min_{i\in[k]} \mathbf{p}^T \mathbf{L}_i$.

We have proven that for any solution of $\inf_{f,\boldsymbol{\theta}} \mathbf{p}^T \boldsymbol{\Psi}(\boldsymbol{\theta}(X), f(X))$, $\text{pred}(\boldsymbol{\theta}(X), f(X))$ would predict as the Bayes predictor, that is, $\text{pred}(\boldsymbol{\theta}(X), f(X)) \in \arg\min_{i\in[k]} \mathbf{p}^T \mathbf{L}_i$. By Theorem 1 the surrogate loss $\Psi_{\boldsymbol{\Lambda}}$ is consistent. □

The surrogate loss $\Psi_{\boldsymbol{\Lambda}}$ coincides with several models proposed in the literature for different values of $\boldsymbol{\Lambda}$ and ϕ. By taking $\ell = \ell^{0\text{-}1}$, we obtain the following surrogate that we will denote "Immediate threshold"

$$\Psi^{\text{IT}}(y, \boldsymbol{\theta}, f) = \begin{cases} \phi(\theta_1 - f) & \text{if } y = 1 \\ \phi(-(\theta_{i-1} - f)) + \phi(\theta_i - f) & \text{if } 1 < y < k \\ \phi(-(\theta_{k-1} - f)) & \text{if } y = k \quad . \end{cases} \tag{10}$$

This surrogate loss has appeared in the literature under a variety of names for different values of ϕ. By construction the following methods are consistent with respect to the zero-one loss:

- "Explicit threshold" from [4] (ϕ = hinge loss),
- "Immediate threshold" from [13] (ϕ = logistic loss),
- "ORBoost with Left-Right margins" from [10] (ϕ = exponential loss),

Considering now the mean absolute loss, $\ell = \ell^{AE}$, we obtain a consistent surrogate that we will denote "All threshold",

$$\Psi^{\text{AT}}(y, \boldsymbol{\theta}, f) = \sum_{i=1}^{y-1} \phi(-(\theta_i - f)) + \sum_{i=y}^{k-1} \phi(\theta_i - f) \quad . \tag{11}$$

As for the "Immediate-threshold" loss, this function has also appeared in the literature under a variety of names for different values of ϕ. By construction, the following methods are consistent with respect to the mean absolute error:

- "Implicit threshold" from [4] (ϕ = hinge loss),
- "All threshold" from [13] (ϕ = logistic loss),
- "ORBoost with All margins" from [10] (ϕ = exponential loss).

Note that these consistency results hold only in the case for which the thresholds are functions of the sample (the free thresholds setting). However, often in practical implementations the thresholds are considered constant across samples (the common thresholds setting). We will consider this setting in more detail in Section 3.4.

3.3 Calibration in the cumulative link approach

The cumulative link model estimates f and $\boldsymbol{\theta}$ by modeling the *cumulative posterior probability* for a given label. More precisely, it assumes $P(Y \leq j|X) = \sigma(\theta_j(X) - f(X))$ where σ is an appropriate link function, i.e., a monotone increasing function mapping $\sigma : \mathbb{R} \to (0, 1)$. The conditional probability for a single class can be computed as $P(Y = j|X) = P(Y \leq j|X) - P(Y \leq j - 1|X)$, and the loss function (negative log-likelihood) reads:

$$\Psi_\sigma(y, \boldsymbol{\theta}(X), f(X)) = \begin{cases} -\log(\sigma(\theta_1 - f) & \text{if } y = 1 \\ -\log(\sigma(\theta_y - f) - \sigma(\theta_{y-1} - f)) & \text{if } 1 < y < k \\ -\log(1 - \sigma(\theta_{k-1} - f)) & \text{if } y = k. \end{cases} \tag{12}$$

Albeit a fundamental question, convexity of this model has not been addressed in the literature to the best of our knowledge. We give a positive answer and state consistency of this loss with respect to mean absolute error.

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

Result 3. *The cumulative logit is classification calibrated with respect to the absolute error loss.*

Proof. See supplementary material. □

3.4 Calibration in the setting of common thresholds

Previously we have assumed that thresholds are a function of the data. However, in practice [3, 4, 10, 13] the thresholds are estimated from the input data and *set fixed*, that is, $\theta_i(X)$ is constant for all $X \in \mathcal{X}$. A natural question is whether the consistency results of the previous section extend to this setting. The answer to this question is negative: we present an example of loss function that is consistent in the general setting but that becomes inconsistent in this setting of common thresholds.

Result 4. *The immediate-threshold loss function is not consistent with respect to the absolute or zero-one loss in the setting of common thresholds.*

Proof. Consider the case $k = 3$. Note that given $\boldsymbol{\theta}$ and $f(X)$, setting θ_i to $\theta_i + e$ for all i and $f(X)$ to $f(X) + e$ for $e \in \mathbb{R}$ has the same Ψ-risk. There is thus no loss of generality by assuming $\theta_1 = 0$.

Using the formula for the derivative of logistic regression, $\phi'(t) = (1 - \sigma(t)), \phi'(-t) = \sigma(t)$, the KKT conditions for the minimization of the Ψ-risk where Ψ is the Immediate Threshold loss (ϕ = logistic loss), X takes n values and under the constraint $\theta_2 \geq \theta_1 = 0$ are

$$\begin{aligned} \frac{\partial \mathcal{R}_\Psi}{\partial \theta_2} &= \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\eta}(X_i)_2 \sigma(\theta_2 - f(X_i)) - \boldsymbol{\eta}(X_i)_3 \sigma(f(X_i) - \theta_2)) - \lambda = 0 \\ \frac{\partial \mathcal{R}_\Psi}{\partial f(X_i)} &= -\boldsymbol{\eta}(X_i)_1 \sigma(-f(X_i)) + \boldsymbol{\eta}(X_i)_2 (\sigma(f(X_i)) - \sigma(\theta_2 - f(X_i))) + \boldsymbol{\eta}(X_i)_3 (\sigma(f(X_i) - \theta_2)) \end{aligned} \tag{13}$$

Consider now the random variable that takes two distinct values $\{X_1, X_2\}$ with $\boldsymbol{\eta}(X_1) = (0.233, 0.533, 0.233), \boldsymbol{\eta}(X_2) = (0.08, 0.9, 0.02)$. It can be easily verified that the values $\theta_2 = f(X_1) = 0, f(X_2) = -0.06$ verify the KKT equations. However, to agree with Bayes predictor is must be $f(X_2) \geq \theta_1 = 0$, contradiction. We have thus a probability distribution in which the values that optimize the Ψ-risk do not give an optimal prediction rule (Bayes predictor). We conclude that this method is not consistent (with respect to our prediction function). □

4 Experiments

Although the focus of this work is a theoretical investigation of consistency, we have also conducted experiments that study empirical performance of the methods outlined in this paper. In this section we compare two approaches described earlier in terms of generalization accuracy. Following [4], we will consider the Immediate Threshold (IT) and All Threshold (AT) formulation with ϕ = hinge loss and a Gaussian kernel. For practical reasons we assume the thresholds are constant across

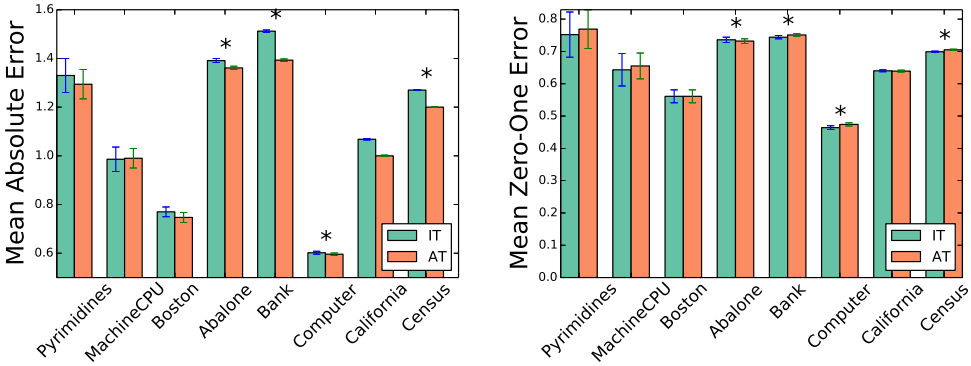

Figure 2: Performance of the Immediate-Threshold (Ψ^{IT}) and All-Threshold (Ψ^{AT}) methods on 8 different datasets and for two different evaluation metrics. Left: the metric used is the mean absolute error. The All-Threshold method is consistent with respect to this loss and performs better on 7 out of 8 datasets. Right: the metric used is the mean zero-one loss. The Immediate-Threshold 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 ($*$).

samples, i.e., we frame the models in the common-threshold setting. Although consistency results do not translate directly into this setting, we have nevertheless observed that surrogates which are consistent in the general setting perform better on most datasets.

The generalization accuracy of both models using a 5-fold cross validation loop are displayed in Figure 2. When considering the mean absolute error as evaluation metric, the All Threshold approach (which is consistent with respect to this metric) performs better on 7 out of 8 datasets. Conversely, when considering the mean zero-one error the Immediate Threshold approach performs better on 6 out of 8 datasets. These empirical findings are therefore inline with our results above.

5 Conclusion and Future Work

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 proven consistency for the absolute and squared error surrogates. In the setting of free thresholds, i.e., when the thresholds are allowed to vary across samples, we have developed a procedure that yields a consistent surrogate for any V-shaped loss function and proven consistency of the proportional odds model. We have also examined these consistency results in the case in which the thresholds are constant across samples (common-threshold setting), a setting of practical importance. By means of a counter-example, we have seen that consistency results do not always translate into consistency in the common-threshold setting.

Several extensions to this work could be considered for future work. For example, Bartlett et al. [1] are able to construct an explicit link function that relates the risk associated with a surrogate function and the risk associated with the zero-one loss. It would be interesting to investigate whether this is feasible in the case of ordinal regression. For threshold-based approaches we have been able to prove consistency only for the case of free thresholds. We believe it would be interesting to investigate the construction of practical algorithms for the setting of free thresholds, and see how these compare in terms of generalization accuracy to classical (common threshold) algorithms.

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] 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.

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

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

[4] Wei Chu and S Sathiya Keerthi. New Approaches to Support Vector Ordinal Regression. In *Proceedings of the 22th International Conference on Machine Learning (ICML)*, 2005.

[5] Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and Empirical Minimization of U -statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

[6] John C. Duchi, Lester W. Mackey, and Michael I. Jordan. On the Consistency of Ranking Algorithms. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010.

[7] C.R. Jack, Ronald C. Petersen, Yue Cheng Xu, Peter C. OBrien, Glenn E. Smith, Robert J. Ivnik, Bradley F. Boeve, Stephen C. Waring, Eric G. Tangalos, and Emre Kokmen. Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. *Neurology*, 52(7): 1397–1397, 1999.

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

[9] Ling Li and Hsuan-tien Lin. Ordinal Regression by Extended Binary Classification. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2006.

[10] 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.

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

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

[13] 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.

[14] Paul Resnick and Hal R Varian. Recommender systems. *Communications of the ACM*, 40(3): 56–58, 1997.

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

[16] Amnon Shashua and Anat Levin. Ranking with Large Margin Principle : Two Approaches, 2003.

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

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

[19] Tong Zhang. Statistical Analysis of Some Multi-Category Large Margin Classification Methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004.

6 Appendix

Proof of result 1, first part. We first extend the proof given in [12] for the consistency of the absolute error loss for the case $k = 3$. We will use the notations from that paper. The surrogate loss function is given by

$$\psi(y, \hat{t}) = |\hat{t} - y| \quad \forall y \in \{1, 2, \ldots, k\} \quad .$$

$\partial\psi_i(t)$ is then given by

$$\partial\psi_y(t) = \begin{cases} \text{conv}(\{+1, -1\}) & \text{if } t = y \\ +1 & \text{if } t > y \\ -1 & \text{if } t < y \end{cases}$$

We fix an arbitrary $1 \leq s \leq k$ and we will compute $\mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}_s)$. In this case the matrix $\mathbf{A}$ is a single-row matrix with $k + 1$ entries. In this case matrices $\mathbf{A} \in \mathbb{R}^{(k+1)\times 1}$ and $\mathbf{B} \in \mathbb{R}^{(k+1)\times k}$ are given by

$$\mathbf{A}_{ij} = \begin{cases} +1 & \text{if } i \leq s \\ -1 & \text{if } i > s \end{cases} \quad , \quad \mathbf{B}_{ij} = \begin{cases} \delta_{i,j} & \text{if } j \leq s \\ \delta_{i,j-1} & \text{if } j > s \end{cases}$$

where δ_{ij} is the Kronecker delta function. In this case $\mathbf{q} \in \text{Null}(A)$ implies that $\sum_{i=1}^{s} q_i = 1/2$. Since $\mathbf{Bq} = (q_1, \ldots, q_s + q_{s+1}, \ldots, q_{k+1})$, $\mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}_s)$ can be written as

$$\begin{aligned} \mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}_s) &= \{\mathbf{p} \in \Delta_k : \mathbf{p} = (q_1, \ldots, q_s + q_{s+1}, \ldots, q_{k+1}) \text{ for some } \mathbf{q} \in \Delta_{k+1}, \sum_{i=1}^{s} q_i = 1/2\} \\ &= \{\mathbf{p} \in \Delta_k : \sum_{i=1}^{s} p_i \geq \frac{1}{2}, \ \sum_{i=1}^{s-1} p_i \leq \frac{1}{2}\} \end{aligned}$$

Consider now Lemma 1. For the case of the absolute error, the matrix $\mathbf{\Lambda}$ is constantly equal to one, $\Lambda_{ij} = 1$ for all i, j. The inequality $u_s \geq v_s$ can be simplified to $\sum_{i=1}^{s} p_i \geq 1/2$. Likewise, $u_{s-1} \leq v_{s-1}$ implies $\sum_{i=1}^{s-1} p_i \leq 1/2$. Since these inequalities are verified for every element of $\mathcal{Q}_s^{\text{ord}}$, we have proven that $\mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}_s) \subseteq \mathcal{Q}_s^{\text{ord}}$, as desired. □

Proof of result 1, second part. We will now proceed to the proof of consistency for the squared error, $\Psi^{\text{SE}}(y, f) = (y - f)^2$. Given $\mathbf{p} \in \Delta_k$ and $X \in \mathcal{X}$, let f^* be a function that minimizes $\mathbf{p}^T \mathbf{\Psi}(f)$. In that case, the KKT conditions for this optimization problem are

$$\frac{\partial}{\partial f} \sum_{i=1}^{k} p_i (f - i)^2 = 0 \implies f^* = \sum_{i=1}^{k} i p_i.$$

Let $r \in \arg\min_{i\in[k]} \mathbf{p}^T \mathbf{L}_i$. Note that for the square loss Lemma 1 implies that $u_r \geq v_r$. Since $\Lambda_i j = 2(i - j) + 1$ when $i \leq j$ and $\Lambda_i j = 2(j - i) - 1$ when $i < j$ for this loss, the inequality adopts the form $\sum_{i=1}^{r} (2(i - r) + 1)p_i \leq \sum_{i=r+1}^{k} (2(r - i) - 1)p_i \implies \sum_{i=1}^{k} i p_i \leq r + 1/2$. Using the KKT equation we obtain $f^*(X) \leq r + 1/2$. Using the same argument with the inequality $u_{r-1} - v_{r-1} \leq 0$ we obtain $f^*(X) \leq r - 1/2$, with equality only if $r - 1 \in \arg\min_{i\in[k]} \mathbf{p}^T \mathbf{L}_i$

We have proven that whenever $r \in \arg\min_{i\in[k]} \mathbf{p}^T \mathbf{L}_i$, $f(X)$ must verify $r - 1/2 \leq f(X) \leq r + 1/2$. If $f(X) = r - 1/2$ then $r - 1 \in \arg\min_{i\in[k]} \mathbf{p}^T \mathbf{L}_i$. This proves that if a function $f^*(X)$ minimizes $\mathbf{p}^T \mathbf{\Psi}(f(X))$, then it leads to an optimal decision rule with respect to the mean squared error when the prediction rule is rounding to the closest integer in $[k]$. Thus, Ψ^{SE} is consistent with respect to the squared error loss. □

Proof of lemma 1. Let $r \in \arg\min_{i\in[k]} \mathbf{p}^T\mathbf{L}_i$. Then it is verified

$$\begin{aligned}
\mathbf{p}^T\mathbf{L}_r &\leq \mathbf{p}^T\mathbf{L}_{r+1} \\
\sum_{i=1}^{k} p_i\ell(i,r) &\leq \sum_{i=1}^{k} p_i\ell(i,r+1) \\
\sum_{i=1}^{k} p_i(\ell(i,r) - \ell(i,r+1)) &\leq 0 \\
\sum_{i=1}^{r} p_i(\ell(i,r) - \ell(i,r+1)) + \sum_{i=r+1}^{k} p_i(\ell(i,r) - \ell(i,r+1)) &\leq 0 \\
\sum_{i=1}^{r} p_i(\ell(i,r) - \ell(i,r+1)) - \sum_{i=r+1}^{k} p_i(\ell(i,r+1) - \ell(i,r)) &\leq 0 \\
\sum_{i=1}^{r} p_i\Lambda_{ir} - \sum_{i=r+1}^{k} p_i\Lambda_{ir} &\leq 0 \\
u_r &\leq v_r
\end{aligned}$$

where we have used the symmetry of the loss. In case $(r+1) \notin \arg\min_{i\in[k]} \mathbf{p}^T\mathbf{L}_i$ the previous inequalities are strict and we have $u_r - v_r < 0$.

It is also verified $\mathbf{p}^T\mathbf{L}_r \leq \mathbf{p}^T\mathbf{L}_{r+2}$ from where

$$\begin{aligned}
\mathbf{p}^T\mathbf{L}_r - \mathbf{p}^T\mathbf{L}_{r+1} &\leq \mathbf{p}^T\mathbf{L}_{r+2} - \mathbf{p}^T\mathbf{L}_{r+1} \\
u_r - v_r &\leq -(u_{r+1} - v_{r+1}) \\
u_r + u_{r+1} &\leq v_r + v_{r+1}
\end{aligned}$$

and so we obtain by induction $\sum_{i=r}^{j} u_i \geq \sum_{i=r}^{j} v_i \quad \forall j \geq r$. Unless $\{r, r+1, \ldots, s\} \in\in \arg\min_{i\in[k]} \mathbf{p}^T\mathbf{L}_i$, the inequality is strict.

Similarly we can prove $\sum_{i=j}^{r-1} u_i \leq \sum_{i=j}^{r-1} v_i$. Under the same hypothesis as before, it is verified that $\mathbf{p}^T\mathbf{L}_r \leq \mathbf{p}^T\mathbf{L}_{r-1}$ and thus

$$\begin{aligned}
\mathbf{p}^T\mathbf{L}_r &\leq \mathbf{p}^T\mathbf{L}_{r-1} \\
\sum_{i=1}^{k} p_i\ell(i,r) &\leq \sum_{i=1}^{k} p_i\ell(i,r-1) \\
\sum_{i=1}^{k} p_i(\ell(i,r) - \ell(i,r-1)) &\leq 0 \\
\sum_{i=1}^{r-1} p_i(\ell(i,r) - \ell(i,r-1)) + \sum_{i=r}^{k} p_i(\ell(i,r) - \ell(i,r-1)) &\leq 0 \\
\sum_{i=1}^{s} p_i(\ell(i,s+1) - \ell(i,s)) + \sum_{i=s+1}^{k} p_i(\ell(i,s+1) - \ell(i,s)) &\leq 0 \\
-\sum_{i=1}^{s} p_i\Lambda_{is} + \sum_{i=s}^{k} p_i\Lambda_{is} &\leq 0 \\
v_s - u_s &\leq 0
\end{aligned}$$

where we have made the change of variable $s = r - 1$. Thus we obtain $v_{r-1} \leq u_{r-1}$. Considering now the inequality $\mathbf{p}^T\mathbf{L}_r \leq \mathbf{p}^T\mathbf{L}_{r-2}$ and proceeding as before we obtain the general inequality $\sum_{i=j}^{r-1} u_i \leq \sum_{i=j}^{r-1} v_i$. □

Proof of Theorem 2. $\Psi_\sigma(1, f, \boldsymbol{\theta})$ and $\Psi_\sigma(k, f, \boldsymbol{\theta})$ are simply 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 $\{b > a\}$ [2]. 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, $\text{tr}(Q) = \lambda_1 + \lambda_2 = 2\exp(a + b)$. 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. □

Proof of Result 3. Let $g_i = \theta_i(X) - f(X)$. The loss function for the proportional odds can be written as

$$\Psi_\sigma(y, \boldsymbol{g}) = \begin{cases} -\log(\sigma(g_1)) & \text{if } y = 1 \\ -\log(\sigma(g_y) - \sigma(g_{y-1})) & \text{if } 1 < y < k \\ -\log(\sigma(-g_{k-1})) & \text{if } y = k \end{cases}$$

Now we consider $\mathbf{z}^* = (\Psi(1, \boldsymbol{g}^*), \ldots, \Psi(k, \boldsymbol{g}^*))$ such that $\mathbf{p}^T\mathbf{z}^* = \inf_{\mathbf{z}\in\mathcal{S}_\Psi} \mathbf{p}^T\mathbf{z}$.

$$\begin{aligned} \mathbf{p}^T\mathbf{z}^* &= \inf_{\mathbf{z}\in\mathcal{S}_\Psi} \mathbf{p}^T\mathbf{z} \\ &= \inf_{\boldsymbol{g}\in\mathcal{T}} -p_1\log(\sigma(g_1)) - \sum_{i=2}^{k} p_i\log(\sigma(g_i) - \sigma(g_{i-1})) - p_k\log(\sigma(-g_{k-1}))) \end{aligned} \tag{14}$$

The KKT conditions associated with this optimization problem are

$$\begin{aligned} &\frac{\partial \mathbf{p}^T\mathbf{z}}{\partial g_1} = G_1(g_1, g_2) = -p_1(1 - \sigma(g_1)) + p_2\frac{\sigma(g_1)(1 - \sigma(g_1))}{\sigma(g_2) - \sigma(g_1)} + \lambda_i = 0 \\ &\frac{\partial \mathbf{p}^T\mathbf{z}}{\partial g_i} = G_i(g_{i-1}, g_i, g_{i+1}) = -p_i\frac{\sigma(g_i)\sigma(-g_i)}{\sigma(g_i) - \sigma(g_{i-1})} + p_{i+1}\frac{\sigma(g_i)\sigma(-g_i)}{\sigma(g_{i+1}) - \sigma(g_i)} - \lambda_{i-1} + \lambda_i = 0 \\ &\frac{\partial \mathbf{p}^T\mathbf{z}}{\partial g_{k-1}} = G_{k-1}(g_{k-2}, g_{k-1}) = -p_{k-1}\frac{\sigma(g_{k-1})\sigma(-g_{k-1})}{\sigma(g_{k-1}) - \sigma(g_{k-2})} + p_k\sigma(g_{k-1}) - \lambda_{i-1} = 0 \\ &\lambda_i(g_i - g_{i+1}) = 0 \\ &\lambda_i \geq 0 \quad \forall i = 1, \ldots, k-1 \end{aligned}$$

Note that all the λ_i must equal to zero because otherwise $\boldsymbol{g}_i^* = \boldsymbol{g}_{i+1}^*$ and the loss function would be infinity. Removing common factors from the above equations we can transform the KKT conditions into the equivalent set of equations:

$$\begin{aligned} \hat{G}_1(g_1, g_2) &= -p_1\frac{1}{\sigma(g_1)} + p_2\frac{1}{\sigma(g_2) - \sigma(g_1)} = 0 \\ \hat{G}_i(g_{i-1}, g_i, g_{i+1}) &= -p_i\frac{1}{\sigma(g_i) - \sigma(g_{i-1})} + p_{i+1}\frac{1}{\sigma(g_{i+1}) - \sigma(g_i)} = 0 \\ G_{k-1}(g_{k-2}, g_{k-1}) &= -p_{k-1}\frac{1}{\sigma(g_{k-1}) - \sigma(g_{k-2})} + p_k\frac{1}{1 - \sigma(g_{k-1})} = 0 \end{aligned}$$

From here it is easy to verify that $\sigma(g_i) = \sum_{j=1}^{i} p_j$ provides a solution to these equations. By convexity this is also the only solution.

Suppose $r \in \arg\min_{i\in[k]} \mathbf{p}^T\mathbf{L}_i$. Then by lemma 1 applied to the absolute error loss we have that $\sum_{j=1}^{r} p_j > 1/2$ from where $\sigma(g_i) > 1/2 \implies g_i > 0$ and similarly $\sum_{j=1}^{r-1} p_j < 1/2 \implies g_{i-1} < 0$, which concludes the proof. $\square$