On the Consistency of Ordinal Regression Methods

Fabian Pedregosa, Francis Bach, Alexandre Gramfort

▶ **To cite this version:**

Fabian Pedregosa, Francis Bach, Alexandre Gramfort. On the Consistency of Ordinal Regression Methods. 2015. hal-01054942v3

HAL Id: hal-01054942

https://inria.hal.science/hal-01054942v3

Preprint submitted on 29 Sep 2015 (v3), last revised 19 Jun 2017 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Public Domain

Journal of Machine Learning Research 1 (2015) 1–48 Submitted 4/00; Published 10/00

On the Consistency of Ordinal Regression Methods

Fabian Pedregosa F@BIANP.NET
INRIA - Parietal project-team
Chaire Havas-Dauphine Économie des Nouvelles Données
Paris, France

Francis Bach FRANCIS.BACH@ENS.FR
INRIA - SIERRA project-team
Département d'Informatique de l'Ecole Normale Supérieure
Paris, France

Alexandre Gramfort ALEXANDRE.GRAMFORT@TELECOM-PARISTECH.FR
LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay
Paris, France

Editor:

Abstract

Many of the ordinal regression models that have been proposed in the literature can be seen as methods that minimize a convex surrogate of the zero-one, absolute, or squared loss functions. A key property that allows to study the statistical implications of such approximations is that of *Fisher consistency*. In this paper we will characterize the Fisher consistency of a rich family of surrogate loss functions used in the context of ordinal regression, including support vector ordinal regression, ORBoosting and least absolute deviation. We will see that, for a family of surrogate loss functions that subsumes support vector ordinal regression and ORBoosting, consistency can be fully characterized by the derivative of a real-valued function at zero, as happens for convex margin-based surrogates in binary classification. We also derive excess risk bounds for a surrogate of the absolute error that generalize existing risk bounds for binary classification. Finally, our analysis suggests a novel surrogate of the squared error loss. To prove the empirical performance of such surrogate, we benchmarked it in terms of cross-validation error on 9 different datasets, where it outperforms competing approaches on 7 out of 9 datasets.

Keywords: Fisher consistency, ordinal regression, calibration, surrogate loss, excess risk bound.

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, such as ("do-not-bother" $\prec$ "only-if-you-must" $\prec$ "good" $\prec$ "very-good" $\prec$ "run-to-see") in movie ratings (Crammer and Singer, 2001), ("absent" $\prec$ "mild" $\prec$ "severe") for the symptoms of a physical disease (Armstrong and Sloan, 1989) and the NRS-11 numeric rating scale for clinical pain measurement (Hartrick et al., 2003). Ordinal regression models have been successfully applied to fields as diverse as econometrics (Greene,

©2015 Fabian Pedregosa et al..

1997), epidemiology (Ananth and Kleinbaum, 1997), fMRI-based brain decoding (Doyle et al., 2013) and collaborative filtering (Rennie and Srebro, 2005).

Ordinal regression shares properties–and yet is fundamentally different–from both multiclass classification and regression. As in the multiclass classification setting, the target variables consist of discrete values, and as in the regression setting (but unlike the multiclass setting) there is a meaningful order between the classes. If we think of the symptoms of a physical disease, it is clear that if the true label is "severe" it is preferable to predict "mild" than "absent". Ordinal regression models formalize this notion of order by ensuring that predictions farther from the true label incur a greater penalty than those closer to the true label.

The ordinal regression approach also shares properties with the learning-to-rank problem (Liu, 2011), in which the goal is to predict the relative order of a sequence of instances. Hence, this approach focuses on predicting a relative order while ordinal regression focuses on predicting a label for each instance. In this sense, it is possible for a ranking model (but not for an ordinal regression one) that predicts the wrong labels to incur no loss at all, as long as the relative order of those labels are correct, e.g. if the prediction is given by the true label plus an additive bias. Although ordinal regression and ranking are different problems, the distinction between both has not always been clear, generating some confusion between the two problems. For example, in the past some methods presented with the word "ranking" in the title would be considered today ordinal regression methods (Crammer and Singer, 2001; Shashua and Levin, 2003; Crammer and Singer, 2005) and likewise some of the first pairwise ranking methods (Herbrich et al., 1999) featured the word ordinal regression in the title.

Despite its widespread applicability, there exists a relative paucity in the understanding of the theoretical properties behind ordinal regression methods, compared to the ones that already exist for other settings such as binary classification. One such example is that of *Fisher consistency*, a notion that relates the minimization of a given loss to the minimization of a surrogate with better computational properties. The importance of this property stems from the fact that many supervised learning methods, such as support vector machines, boosting and logistic regression for binary classification, can be seen as methods that minimize a convex surrogate on the 0-1 loss. Such results have emerged in recent years for classification (Bartlett et al., 2003; Zhang, 2004a; Tewari and Bartlett, 2007), ranking (Duchi et al., 2010; Calauzenes et al., 2012) and even multiclass classification with an arbitrary loss function (Ramaswamy and Agarwal, 2012, 2014), a setting that subsumes ordinal regression. Despite these recent progress, the Fisher consistency of most surrogates used within the context of ordinal regression remains elusive. The aim of this paper is to bridge the gap by providing an analysis of Fisher consistency for a wide family of ordinal regression methods that parallels the ones that already exist for other multiclass classification and ranking.

Notation. Through the paper, we will use k to denote the number of classes (i.e., labels) in the learning problem. We will denote by $\mathcal{S}$ the subset of $\mathbb{R}^{k-1}$ for which the components are increasing, that is,

$$\mathcal{S} := \left\{ \alpha : \alpha \in \mathbb{R}^{k-1} \text{ and } \alpha_i \leq \alpha_{i+1} \text{ for } 1 \leq i \leq k-2 \right\} \quad .$$

Δ^p denotes the p-dimensional simplex, which is defined as

$$\Delta^p := \left\{ x \in \mathbb{R}^p : x_i \geq 0 \text{ and } \sum_{i=1}^{p} x_i = 1 \right\} \quad .$$

Following Knuth (1992) we use the Iverson bracket $[\![\cdot]\!]$ as

$$[\![q]\!] := \begin{cases} 1 \text{ if q is true} \\ 0 \text{ otherwise} \end{cases} \quad .$$

We will also make reference to loss functions commonly used in binary classification. These are the Hinge loss ($\varphi(t) = \max(1-t, 0)$), the squared Hinge loss ($\varphi(t) = \max(1-t, 0)^2$), the logistic loss ($\varphi(t) = \log(1+e^{-t})$), exponential loss ($\varphi(t) = e^{-t}$) and the squared loss ($\varphi(t) = (1-t)^2$).

1.1 Problem setting

Here we present the formalism that we will be using throughout the paper. 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 finite set of k *ordered categories* that we will denote $\mathcal{Y} = \{1, \ldots, k\}$. 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 the goal is to learn from the observations a measurable mapping called a *decision function* $f : \mathcal{X} \to \mathcal{S} \subseteq \mathbb{R}^{k-1}$ so that the *risk* given below is as small as possible:

$$\mathcal{L}(f) := \mathbb{E}(\ell(Y, f(X))) \quad , \tag{1}$$

where $\ell : \mathcal{Y} \times \mathcal{S}$ is a *loss function* that measures the disagreement between the true label and the prediction. For ease of optimization, the decision function has its image in a subset of $\mathbb{R}^{k-1}$, and the function that converts an element of $\mathcal{S}$ into a class label is called a *prediction function*. The prediction function that we will consider through the paper is given for $\alpha \in \mathcal{S}$ by the number of coordinates below zero plus one, that is,

$$\text{pred}(\alpha) := 1 + \sum_{i=1}^{k-1} [\![\alpha_i < 0]\!] \quad . \tag{2}$$

Note that for the case of two classes $\mathcal{Y} = \{1, 2\}$, the decision function is real-valued and the prediction defaults the common binary classification rule in which prediction depends on the sign of this decision function.

Different loss functions can be used within the context of ordinal regression. The most commonly used one is the absolute error, which measures the absolute difference between the predicted and true labels. For $\alpha \in \mathcal{S}$, this is defined as

$$\ell(y, \alpha) := \left| y - \text{pred}(\alpha) \right| \quad . \tag{3}$$

The absolute error loss is so ubiquitous in ordinal regression that some authors refer to it simply as *the* ordinal regression loss (Agarwal, 2008; Ramaswamy and Agarwal, 2012). For

this reason special emphasis is given in this paper to surrogates of this loss. However, we will also describe methods that minimize the 0-1 loss (i.e., the classification error) and in Section 5 we will see how some results can be generalized beyond these and to general loss functions that verify a certain admissibility criterion.

In order to find the decision function with minimal risk it might seem appropriate to minimize Eq. (1). However, this 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, ℓ is typically discontinuous in its second argument, hence the empirical approximation to the risk is difficult to optimize and can lead to an NP-hard problem (Feldman et al., 2012; Ben-David et al., 2003)[1]. It is therefore common to approximate ℓ by a function $\psi : \mathcal{Y} \times \mathcal{S} \to \mathbb{R}$, called a *surrogate loss function*, which has better computational properties. The goal becomes then to find the decision function that instead minimizes the *surrogate risk*, defined as

$$\mathcal{A}(f) := \mathbb{E}(\psi(Y, f(X))) \ . \tag{4}$$

We are interested by the statistical implications of such approximation. Assuming that we have full knowledge of the probability distribution that generates the data P, what are the consequences of optimizing a convex surrogate of the risk instead of the true risk?

The main property that we will study in order to answer this question is that of *Fisher consistency*. Fisher consistency is a desirable property for surrogate loss functions (Lin, 2004) and implies that in the population setting, i.e., if the probability distribution P were available, then optimization of the surrogate would yield a function with minimal risk. From a computational point of view, this implies that the minimization of the surrogate risk, which is usually a convex optimization problem and hence easier to solve than the minimization of the risk, does not penalize the quality (at least in the population setting) of the obtained solution.

We will use the following notation for the optimal risk and optimal surrogate risk:

$$\mathcal{L}^* := \inf_f \mathcal{L}(f) \quad \text{and} \quad \mathcal{A}^* := \inf_f \mathcal{A}(f) \quad ,$$

where the minimization is done over all measurable functions $\mathcal{X} \to \mathcal{S}$. $\mathcal{L}^*$ is sometimes referred to as the *Bayes risk*, and a decision function (not necessarily unique) that minimizes the risk is called a *Bayes decision function*.

We will now give a precise definition of Fisher consistency. 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 P_n. We denote it $T(P_n)$. The estimator is said to be Fisher consistent if its population analog, $T(P)$, 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 Lin (2004) to an arbitrary loss function ℓ:

Definition 1 *(**Fisher consistency**) Given a surrogate loss function $\psi : \mathcal{Y} \times \mathcal{S} \to \mathbb{R}$, we will say that the surrogate loss function ψ is consistent with respect to the loss $\ell : \mathcal{Y} \times \mathcal{S} \to \mathcal{Y}$*

1. Note that binary classification can be seen as a particular case of ordinal regression.

if for every probability distribution over $X \times Y$ it is verified that every minimizer f of the surrogate risk reaches Bayes optimal risk, that is,

$$\mathcal{A}(f) = \mathcal{A}^* \implies \mathcal{L}(f) = \mathcal{L}^* \quad .$$

For some surrogates we will be able to derive not only Fisher consistency, but also *excess risk bounds*. These are bounds of the form

$$\gamma(\mathcal{L}(f) - \mathcal{L}^*) \leq \mathcal{A}(f) - \mathcal{A}^* \quad ,$$

for some function real-valued function γ with $\gamma(0) = 0$. These inequalities not only imply Fisher consistency, but also allow to bound the excess risk by the excess in surrogate risk. These inequalities play an important role in different areas of learning theory, as they can be used for example to obtain rates of convergence (Bartlett et al., 2003) and oracle inequalities (Boucheron et al., 2005).

1.2 Full and conditional risk

The above definition of Fisher consistency is often replaced by a point-wise version that is easier to verify in practice. A key ingredient of this characterization are the notions of *conditional risk* and *surrogate conditional risk* that we will now define. These are denoted by L and A respectively, and for $\alpha \in \mathcal{S}$ and $p \in \Delta^k$ are defined by

$$L(\alpha, p) := \sum_{i=1}^{k} p_i \ell(i, \alpha) \quad \text{and} \quad A(\alpha, p) := \sum_{i=1}^{k} p_i \psi(i, \alpha) \quad . \tag{5}$$

The full and conditional risk are then related by the equations

$$\begin{aligned} \mathcal{L}(f) &= \mathbb{E}_{X\times Y}(\ell(Y, f(X))) = \mathbb{E}_X \mathbb{E}_{Y|X}(\ell(Y, f(X))) = \mathbb{E}_X(L(f(X), \eta(X))) \\ \mathcal{A}(f) &= \mathbb{E}_{X\times Y}(\psi(Y, f(X))) = \mathbb{E}_X \mathbb{E}_{Y|X}(\psi(Y, f(X))) = \mathbb{E}_X(A(f(X), \eta(X))) \quad , \end{aligned}$$

where $\eta : \mathcal{X} \to \Delta^k$ is the vector of conditional probabilities given by $\eta_i(x) = P(y = i|X = x)$. As for the full risk, we will denote by L^*, A^* the infimum of its value for a given $p \in \Delta^k$, i.e.,

$$L^*(p) = \inf_{\alpha\in\mathcal{S}} L(\alpha, p) \quad \text{and} \quad A^*(p) = \inf_{\alpha\in\mathcal{S}} A(\alpha, p) \quad .$$

When the risk minimization is performed over functions that can be defined independently at every $x \in \mathcal{X}$, it is possible to relate the minimization of the risk with that of the conditional risk since

$$\begin{aligned} \inf_f \mathcal{L}(f) = \inf_f \mathbb{E}_{X\times Y}\left(\ell(Y, f(X))\right) &= \mathbb{E}_X\left[\inf_f \mathbb{E}_{Y|X}(\ell(Y, f(X)))\right] \\ &= \mathbb{E}_X\left[\inf_\alpha L(\alpha, \eta(X))\right] \quad . \end{aligned} \tag{6}$$

This equation implies that the minimal risk can be achieved by minimizing pointwise the conditional risk $L(\cdot)$, which–in general–will be easier that direct minimization of the full risk. The condition for this, i.e., that the functions be estimated independently at every sample

point is verified by the set of measurable functions from the sample space into a subset of $\mathbb{R}^k$ (in this case $\mathcal{S}$), which is the typical setting in studies of Fisher consistency. However, this is no longer true when inter-observation constraints are enforced (e.g. smoothness). As is common in studies of Fisher consistency, we will suppose that the function class verifies the property of Eq. (6) and we will discuss in Section 6 an important family of functions in which this requisite is not met.

We will now present a characterization of Fisher consistency based on the pointwise risk which we will use throughout the paper. Equivalent forms of this characterization have appeared under a variety of names in the literature, such as classification calibration (Bartlett et al., 2003; Ramaswamy and Agarwal, 2012), infinite sample consistency (Zhang, 2004b) and proper surrogates (Buja et al., 2005; Gneiting and Raftery, 2007).

Lemma 2 (Pointwise characterization of Fisher consistency) *Let A and L be defined as in Eq* (5)*. Then ψ is Fisher consistent with respect to ℓ if and only if for all $p \in \Delta^k$ it is verified that*

$$A(\alpha, p) = A^*(p) \implies L(\alpha, p) = L^*(p) \quad . \tag{7}$$

Proof Let $\mathcal{L}$ and $\mathcal{A}$ represent the expected value of ℓ and ψ, as defined in Equations (1) and (4) respectively.

($\implies$) We will first suppose that Eq. (7) is verified and prove Fisher consistency. Let f be such that $\mathcal{A}(f) = \mathcal{A}^*$. Then it is verified that

$$\mathcal{A}(f) - \mathcal{A}^* = \mathbb{E}_X(A(f(X), \eta(X)) - A^*(\eta(X))) = 0 \quad .$$

The value inside the expectation is non-negative by definition of A^*. Since this is verified for all probability distributions over $X \times Y$, then it must be verified that $A(f(x), \eta(x)) = A^*(\eta(x))$ for all $x \in \mathcal{X}$. By assumption $L(f(X), \eta(X)) = L^*(\eta(X))$. Hence the excess risk verifies

$$\mathcal{L}(f) - \mathcal{L}^* = \mathbb{E}_X(L(f(X), \eta(X)) - L^*(\eta(X))) = \mathbb{E}(0) = 0 \quad .$$

and so ψ is Fisher consistent with respect to ℓ.

($\impliedby$) We will prove that Eq. (7) implies Fisher consistency. We do so by contradiction. Let us suppose that there exists a surrogate that is Fisher consistent but Eq. (7) is not verified and arrive to a contradiction. Since Eq. (7) is not verified there exists $\tilde{\alpha} \in \mathcal{S}$ and $\tilde{p} \in \Delta^k$ be such that

$$A(\tilde{\alpha}, \tilde{p}) = A^*(\tilde{p}) \text{ and } L(\tilde{\alpha}, \tilde{p}) > L^*(\tilde{p}) \quad .$$

Now let P be the probability distribution such that $\eta(x) = \tilde{p}$ for all $x \in \mathcal{X}$ and let $f : \mathcal{X} \to \mathcal{S}$ the mapping that is constantly $\tilde{\alpha}$. Then it is verified that

$$\mathcal{A}(f) - \mathcal{A}^* = \mathbb{E}_X(A(f(X), \eta(X)) - A^*(\eta(X))) = \mathbb{E}_X(A(\tilde{\alpha}, \tilde{p}) - A^*(\tilde{p})) = 0 \quad ,$$

and so $\mathcal{A}(f) = \mathcal{A}^*$. Likewise, the excess risk verifies

$$\mathcal{L}(f) - \mathcal{L}^* = \mathbb{E}_X(L(f(X), \eta(X)) - L^*(\eta(X))) = \mathbb{E}_X(L(\tilde{\alpha}, \tilde{p}) - L^*(\tilde{p})) > 0$$

and so ψ cannot be Fisher consistent with respect to ℓ. This is contradiction, and conludes the proof.

■

1.3 Summary of main results

The main contribution of this paper is to characterize the Fisher consistency of a wide family of surrogate loss functions used for the task of ordinal regression. Contrary to known results for multiclass classification and ranking, where One-vs-All and RankSVM have been proven to be inconsistent, in the ordinal regression setting common surrogates such as ORSVM and proportional odds will be proven to be Fisher consistent. One of the most surprising results of this paper is that for a particular class of surrogates that verify a *decomposability* property, it is possible to provide a characterization of Fisher consistency and excess risk bounds that generalize those known for convex margin-based surrogates (loss functions of the form $\varphi(Yf(X))$) in binary classification.

We will introduce the surrogate loss functions that we consider in Section 2. These will be divided between surrogates of the absolute error and surrogate of the 0-1 loss. We organize their study as follows:

- In Section 3 we characterize the **Fisher consistency for surrogates of the absolute error**. The surrogates that we consider in this section are the all threshold (AT), the cumulative link (CL) and the least absolute deviation (LAD). Besides Fisher consistency, a decomposability of the AT loss will allow us to provide excess risk bounds for this surrogate.

- In Section 4 we characterize the **Fisher consistency of the surrogates of the 0-1 loss**. For this loss, denoted immediate threshold (IT), its Fisher consistency will depend on the derivative at zero of a real-valued convex function.

- In Section 5 we **construct a surrogate for an arbitrary loss function** that verifies an admissibility condition. We name this surrogate generalized all threshold (GAT). This loss function generalizes the AT and IT loss functions introduced earlier. We will characterize the Fisher consistency of this surrogate.

- Turning back to one of the topics mentioned in the introduction, we discuss in Section 6 the **implications of inter-observational constraints in Fisher consistency**. Following Shi et al. (2015), we define a restricted notion of consistency known as $\mathcal{F}$-consistency of parametric consistency and give sufficient conditions for the $\mathcal{F}$-consistency of two surrogates.

- In Section 7 we **examine the empirical performance of a novel surrogate**. This novel surrogate is a particular instance of the GAT loss function introduced in Section 5 when considering the squared error as evaluation metric. We compare this novel surrogate against a least squares model on 9 different datasets, where the novel surrogate outperforms the least squares estimate on 7 out of the 9 datasets.

1.4 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 e.g. Bartlett et al. (2003); Zhang (2004a); Tewari and Bartlett (2007); Reid and Williamson (2010)). Some of the results in this paper generalize known results for binary classification to the ordinal regression setting. In

particular, Bartlett et al. (2003) provide a characterization of the Fisher consistency for convex margin-based surrogates that we extend to the all threshold (AT) and immediate threshold (IT) family of surrogate loss functions. The excess error bound that we provide for the AT surrogate also generalizes the excess error bound given in (Bartlett et al., 2003, Section 2.3).

Fisher consistency of arbitrary loss functions (a setting that subsumes ordinal regression) has been studied for some surrogates. Lee et al. (2004) proposed a surrogate that can take into account generic loss functions and for which Fisher consistency was proven by Zhang (2004b). In a more general setting, Ramaswamy and Agarwal (2012, 2014) provide necessary and sufficient conditions for a surrogate to be Fisher consistent with respect to an arbitrary loss function. Among other results, they prove consistency of least absolute deviation (LAD) and an ε-insensitive loss with respect to the absolute error for the case of three classes ($k = 3$). In this paper, we extend the proof of consistency for LAD to an arbitrary number of classes. Unlike previous work, we consider the so-called *threshold-based surrogates* (AT, IT and CL), which rank among the most popular ordinal regression loss functions and for which its Fisher consistency has not been studied previously.

Fisher consistency has also been studied in the pairwise ranking setting, where it has been proven (Duchi et al., 2010; Calauzenes et al., 2012) that some models (such as RankSVM) are not consistent. Despite similarities between ranking and ordinal regression, we will see in this paper that most popular ordinal regression models are Fisher consistent under mild conditions.

There are few studies on the theoretical properties of ordinal regression methods. A notable example comes from Agarwal (2008), where the authors study generalization bounds for some ordinal regression algorithms. Some of the surrogate loss functions used by these models (such as the support vector ordinal regression of Chu and Keerthi (2005)) are analyzed in this paper. In that work, the authors outline the study of consistency properties of ordinal regression models as an important question to be addressed in the future.

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 (Stone, 1977; Steinwart, 2002). This paper focuses solely on Fisher consistency, to whom we will refer simply as consistency from now on.

2. Ordinal regression models

We introduce the different ordinal regression models that we will consider within this paper. Considering first the absolute error, we will write this loss as a sum of 0-1 loss functions[2]. This is a key reformulation of the absolute error that we will use throughout the paper. Let

2. The 0-1 loss, defined as the function that is 1 for negative values and 0 otherwise can be defined in bracket notation as $\ell_{0-1}(t) = [\![\alpha_i \leq 0]\!]$.

$y \in \mathcal{Y}$ and $\alpha \in \mathcal{S}$, then

$$\begin{aligned}\ell(y,\alpha) = \big|y - \text{pred}(\alpha)\big| &= \left|y - 1 - \sum_{i=1}^{k-1} [\![\alpha_i < 0]\!]\right| \\ &= \left|y - 1 - \sum_{i=1}^{y-1} [\![\alpha_i < 0]\!] - \sum_{i=y}^{k-1} [\![\alpha_i < 0]\!]\right| \\ &= \left|\sum_{i=1}^{y-1} [\![\alpha_i \geq 0]\!] - \sum_{i=y}^{k-1} [\![\alpha_i < 0]\!]\right| \quad .\end{aligned} \tag{8}$$

Now, if $\alpha_y \geq 0$ then the second summand equals zero. Otherwise, if $\alpha_y < 0$, then the first summand equals zero. In either case, we have

$$\ell(y,\alpha) = \sum_{i=1}^{y-1} [\![\alpha_i \geq 0]\!] + \sum_{i=y}^{k-1} [\![\alpha_i < 0]\!] \quad . \tag{9}$$

This expression suggests that a natural surrogate can be constructed by replacing the 0-1 loss in the above expression function by a convex surrogate such as the logistic or hinge loss. Denoting by $\varphi : \mathbb{R} \to \mathbb{R}$ such surrogate, we obtain the following loss function that we denote *all threshold (AT)*:

$$\psi_{\text{AT}}(y,\alpha) := \sum_{i=1}^{y-1} \varphi(-\alpha_i) + \sum_{i=y}^{k-1} \varphi(\alpha_i) \quad . \tag{10}$$

This function has appeared under different names in the literature. When φ is the hinge loss, this model is known as support vector ordinal regression with implicit constraints (Chu and Keerthi, 2005) and support vector with sum-of-margins strategy (Shashua and Levin, 2003). When φ is the exponential loss, this model has been described in (Lin and Li, 2006) as ordinal regression boosting with all margins. Finally, Rennie and Srebro (2005) provided a unifying formulation for this approach considering for the hinge, logistic and exponential loss under the name of All-Threshold loss, a name that we will adopt in this paper.

The name *thresholds* comes from the fact that in the aforementioned work, the decision function is of the form $\alpha_i = \theta_i - f(\cdot)$, where $(\theta_1, \ldots, \theta_{k-1})$ is a vector estimated from the data known as the vector of thresholds. We will discuss in Section 6 the implications of such decision function. For the prediction rule to give meaningful results it is important to ensure that the thresholds are ordered, i.e., $\theta_1 \leq \theta_2, \leq \cdots, \leq \theta_{k-1}$ (Chu and Keerthi, 2005). In our setting, we enforce this constraint by $\alpha \in \mathcal{S}$, hence the importance of restricting the problem to this subset of $\mathbb{R}^{k-1}$.

Another family of surrogate loss functions take a more probabilistic approach and model instead the posterior probability. This is the case of the *cumulative link* models of McCullagh (1980). In such models the decision function f is assumed to be such that $\sigma(f_i(x)) = P(Y \leq i | X = x)$, where $\sigma : \mathbb{R} \to [0,1]$ is a function referred to as *link function*. Several functions can be used as link function, although the most used ones are

the sigmoid function and the Gaussian cumulative distribution. The sigmoid function i.e., $\sigma(t) = 1/(1 + \exp(-t))$, leads to a model sometimes referred as proportional odds (McCullagh, 1980) and cumulative logit (Agresti, 2010), although for naming consistency we will refer to it as *logistic cumulative link*. Another important link function is given by the Gaussian cumulative distribution, $\sigma(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2}$, used in the Gaussian process ordinal regression model of Chu and Ghahramani (2004). The cumulative link (CL) loss function is given by its negative likelihood, that is,

$$\psi_{\text{CL}}(y, \alpha) := \begin{cases} -\log(\sigma(\alpha_1)) & \text{if } y = 1 \\ -\log(\sigma(\alpha_y) - \sigma(\alpha_{y-1})) & \text{if } 1 < y < k \\ -\log(1 - \sigma(\alpha_{k-1})) & \text{if } y = k \quad . \end{cases} \tag{11}$$

We will now consider the 0-1 loss. In this case, the loss will be 1 if the prediction is below y, i.e., if $\alpha_{y-1} \geq 0$ or if the prediction is above y, i.e., if $\alpha_y < 0$. Hence, we can write

$$\ell(y, \alpha) = \begin{cases} [\![\alpha_1 < 0]\!] & \text{if } y = 1 \\ [\![\alpha_{y-1} \geq 0]\!] + [\![\alpha_y < 0]\!] & \text{if } 1 < y < k \\ [\![\alpha_{k-1} \geq 0]\!] & \text{if } y = k \quad . \end{cases}$$

As for the absolute error, a natural surrogate is given by replacing the 0-1 loss by a convex surrogate as the hinge or logistic function. Following Rennie and Srebro (2005), we will refer to this loss function as *immediate threshold (IT)*:

$$\psi_{\text{IT}}(y, \alpha) := \begin{cases} \varphi(\alpha_1) & \text{if } y = 1 \\ \varphi(-\alpha_{y-1}) + \varphi(\alpha_y) & \text{if } 1 < y < k \\ \varphi(-\alpha_{k-1}) & \text{if } y = k \quad . \end{cases} \tag{12}$$

As with the AT surrogate, this loss has appeared under a variety of names in the literature. When φ is the hinge loss, this model is known as Support Vector Ordinal Regression with explicit constraints (Chu and Keerthi, 2005) and support vector with fixed-margins strategy (Shashua and Levin, 2003). For $\varphi =$ exponential loss, this model has been described by Lin and Li (2006) as ordinal regression boosting with left-right margins. The construction of the AT and IT surrogates are similar. We will see in Section 5 that both surrogates can be seen as a particular instance of a more general family of loss functions.

The approaches we have seen so far can be seen as methods that extend known binary classification methods to the ordinal regression setting. A different approach consists in treating the labels as real values and use regression algorithms to learn a real-valued mapping between the samples and the labels. This ignores the discrete nature of the labels, thus it is necessary to introduce a prediction function that converts this real value into a label in $\mathcal{Y}$. This prediction function is given by rounding to the closest label (see, e.g., (Kramer et al., 2001) for a discussion of this method using regression trees). This approach is commonly referred to as the *regression-based* approach to ordinal regression. If we are

seeking to minimize the absolute error, a popular loss function is to minimize the least absolute deviation (LAD). For any $\beta \in \mathbb{R}$, this is defined as

$$\psi_{\text{LAD}}(y, \beta) := |y - \beta| \quad ,$$

and prediction is then given by rounding β to the closest label. This setting departs from the approaches introduced earlier by using a different prediction function. However, via a simple transformation it is possible to convert this prediction function (rounding to the closest label) to the prediction function that counts the number of non-zero components defined in Eq. (2). For a given $\beta \in \mathbb{R}$, this transformation is given by

$$\alpha_1 = \frac{3}{2} - \beta, \quad \alpha_2 = \frac{5}{2} - \beta, \quad \ldots, \quad \alpha_i = i + \frac{1}{2} - \beta \quad . \tag{13}$$

It is immediate to see that this vector α belongs to $\mathcal{S}$ and

$$\begin{aligned}
\text{pred}(\alpha) &= 1 + \sum_{i=1}^{k-1} [\![i + \frac{1}{2} < \beta]\!] \\
&= \begin{cases} 1 & \text{if } \beta \leq 1 + \frac{1}{2} \\ i & \text{if } i - \frac{1}{2} \leq \beta < i + \frac{1}{2}, 1 < i < k \\ k & \text{if } \beta \geq k - \frac{1}{2} \end{cases} \\
&= \arg\min_{1 \leq i \leq k} |\beta - i| \quad \text{(rounding to the lower label in case of ties)},
\end{aligned}$$

hence predicting in the transformed vector α is equivalent to the closes label to β. We will adopt this transformation when considering LAD for convenience, in order to consider it within the same framework as the other models. With the aforementioned transformation, the least absolute deviation surrogate is given by

$$\psi_{\text{LAD}}(y, \alpha) = \left| y + \alpha_1 - \frac{3}{2} \right| \tag{14}$$

Although the surrogate loss function LAD and the absolute loss of Eq. (3) look very similar, the fundamental difference being that the LAD surrogate is convex on α, while the absolute error, due to the presence of the function pred is not.

In this section we have introduced some of the most common ordinal regression methods based on the optimization of a convex loss function. These are summarized in Table 1.

3. Absolute error surrogates

In this section we will assume that the loss function is the absolute error, i.e., $\ell(y, \alpha) = \left| y - \text{pred}(\alpha) \right|$ and we will focus on surrogates of this loss. For an arbitrary $\alpha \in \mathcal{S}$, the conditional risk for the absolute error can be reformulated using the development of the

Table 1: Surrogate loss functions considered in this paper.

Model	**Loss Function**	**Also known as**
All thresholds (AT)	$\sum_{i=1}^{y-1} \varphi(-\alpha_i) + \sum_{i=y}^{k-1} \varphi(\alpha_i)$	Implicit constraints (Chu and Keerthi, 2005), all margins (Lin and Li, 2006).
Cumulative link (CL)	$-\log(\sigma(\alpha_y) - \sigma(\alpha_{y-1}))$	Proportional odds (McCullagh, 1980), cumulative logit (Agresti, 2010).
Immediate threshold (IT)	$\varphi(-\alpha_{y-1}) + \varphi(\alpha_y)$	Explicit constraints (Chu and Keerthi, 2005), Fixed-margins (Shashua and Levin, 2003)
Least absolute deviation (LAD)	$\lvert y + \alpha_1 - 0.5 \rvert$	Least absolute error, least absolute residual, Sum of absolute deviations, ℓ_1 regression.

absolute error from Eq. (8):

$$\begin{aligned} L(\alpha, p) &= \sum_{i=1}^{k} p_i \left(\sum_{j=1}^{i-1} [\![\alpha_j \geq 0]\!] + \sum_{j=i}^{k-1} [\![\alpha_j < 0]\!] \right) \\ &= \sum_{i=1}^{k} [\![\alpha_j \geq 0]\!](1 - u_i(p)) + \sum_{j=1}^{k} [\![\alpha_j < 0]\!] u_i(p) \quad , \end{aligned}$$

where u is the vector of cumulative probabilities, i.e., $u_i(p) := \sum_{j=1}^{i} p_j$. Let $r = \text{pred}(\alpha)$. Then $\alpha_{r-1} < 0$ and $\alpha_r \geq 0$, from where the above can be simplified to

$$L(\alpha, p) = \sum_{i=1}^{r-1} u_i(p) + \sum_{i=r}^{k-1} (1 - u_i(p)) \quad . \tag{15}$$

Using this expression, we will now derive a minimizer of the conditional risk:

Lemma 3 *For any $p \in \Delta^k$, let $\underline{\alpha}(p)$ be defined as*

$$\underline{\alpha}(p) = (2u_1(p) - 1, \ldots, 2u_{k-1}(p) - 1) \quad .$$

Then, $L(\cdot, p)$ achieves its minimum at $\underline{\alpha}(p)$, that is,

$$\underline{\alpha}(p) \in \arg\min L(\alpha, p) \quad .$$

Proof We will prove that $L(\alpha, p) \geq L(\underline{\alpha}(p), p)$ for any $\alpha \in \mathcal{S}$ and any $p \in \Delta^k$. Let p and α be fixed and we denote $r^* = \text{pred}(\underline{\alpha}(p))$ and $r = \text{pred}(\alpha)$. We distinguish three cases, $r < r^*$, $r > r^*$ and $r = r^*$.

- $r < r^*$. In this case, using Eq. (15) it is verified that

$$L(\alpha, p) - L(\underline{\alpha}(p), p) = -\sum_{i=r}^{r^*-1} u_i(p) + \sum_{i=r}^{r^*-1} (1 - u_i(p)) = -\sum_{i=r}^{r^*-1} 2u_i(p) - 1 \quad .$$

  Now, by the definition of prediction function, $2u_i(p) - 1 < 0$ for $i < r^*$, and so we have

$$L(\alpha, p) - L(\underline{\alpha}(p), p) = \sum_{i=r}^{r^*-1} \left|2u_i(p) - 1\right| \quad .$$

- $r > r^*$. By same reasoning, it is verified that

$$L(\alpha, p) - L(\underline{\alpha}(p), p) = \sum_{i=r^*}^{r-1} u_i(p) - \sum_{i=r^*}^{r-1} (1 - u_i(p)) = \sum_{i=r^*}^{r-1} 2u_i(p) - 1 \quad .$$

  Since by definition of prediction function $2u_i(p) - 1 \geq 0$ for $i \geq r^*$, it is verified that

$$L(\alpha, p) - L(\underline{\alpha}(p), p) = \sum_{i=r^*}^{r-1} \left|2u_i(p) - 1\right| \quad .$$

- $r = r^*$. In this case, Eq. (15) yields

$$L(\alpha, p) - L(\underline{\alpha}(p), p) = 0 \quad .$$

Let I denote the set of indices for which α disagrees in sign with $\underline{\alpha}$, that is, $I = \{i : \alpha_i(2u_i(p) - 1) < 0\}$. Then, combining the three cases we have the following formula for the excess in conditional risk

$$L(\alpha, p) - L(\underline{\alpha}(p), p) = \sum_{i \in I} \left|2u_i(p) - 1\right| \quad , \tag{16}$$

which is always non-negative and hence $L^*(p) = L(\underline{\alpha}(p), p)$. ■

3.1 All Threshold (AT)

We will now consider the AT surrogate. We will prove that many properties known for binary classification are inherited by this loss function. More precisely, we will provide a characterization of consistency for convex φ in Theorem 5 and excess risk bounds in Theorem 6 that parallel those of Bartlett et al. (2003) for binary classification.

Through this section A will represent the conditional risk of the AT surrogate, which can be expressed as:

$$\begin{aligned} A(\alpha, p) &= \sum_{j=1}^{k} p_j \psi_{\text{AT}}(j, \alpha) = \sum_{j=1}^{k} p_j \left(\sum_{i=1}^{j-1} \varphi(-\alpha_i) + \sum_{i=j}^{k-1} \varphi(\alpha_i) \right) \\ &= \sum_{i=1}^{k-1} (1 - u_i(p))\varphi(-\alpha_i) + u_i(p)\varphi(\alpha_i) \quad , \end{aligned} \tag{17}$$

where as in the previous section $u_i(p) = \sum_{j=1}^{i} p_i$, $\alpha \in \mathcal{S}$ and $p \in \Delta^k$. This surrogate verifies a decomposable property that will be key to further analysis. The property that we are referring to is that the above conditional risk it can be expressed as the sum of $k-1$ binary classification conditional risks. For $\beta \in \mathbb{R}, q \in [0, 1]$, we define C as follows

$$C(\beta, q) = q\varphi(\beta) + (1 - q)\varphi(-\beta) \quad ,$$

where C can be seen as the conditional risk associated with the binary classification loss function φ. Using this notation, the conditional risk A can be expressed in terms of C as:

$$A(\alpha, p) = \sum_{i=1}^{k-1} C(\alpha_i, u_i(p)) \quad .$$

Our aim is to compute A^* in terms of the infimum of C, denoted $C^*(q) := \inf_\beta C(q, \beta)$. Since C is the conditional risk of a binary classification problem, this would yield a link between the optimal risk for the AT surrogate and the optimal risk for a binary classification surrogate. However, this is in general not possible because of the monotonicity constraints in $\mathcal{S}$: the infimum over $\mathcal{S}$, need not equal the infimum over the superset $\mathbb{R}^{k-1}$. We will now present a result that states sufficient conditions under which the infimum over $\mathcal{S}$ and over $\mathbb{R}^{k-1}$ do coincide. This implies that A^* can be estimated as the sum of $k-1$ different surrogate conditional risks, each one corresponding to a binary classification surrogate with different probability distributions. A similar result was proven in the empirical approximation setting by Chu and Keerthi (2005, Lemma 1). In this work, the authors consider the hinge loss and show that a minimizer of this loss necessarily verifies the monotonicity constraints in $\mathcal{S}$.

In the following lemma we provide minimal conditions on φ under which the monotonicity constraints can be ignored when computing A^*. This is an important step towards obtaining an explicit expression for A^*:

Lemma 4 *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a function such that $\varphi(\beta) - \varphi(-\beta)$ is a decreasing function of $\beta \in \mathbb{R}$. Then for all $p \in \Delta^k$, it is verified that*

$$A^*(p) = \sum_{i=1}^{k-1} C^*(u_i(p)) \quad .$$

Proof Let $p \in \Delta^k$ be fixed and let $\alpha^* \in \arg\min_{\alpha \in \mathbb{R}^{k-1}} A(\alpha_i, p)$. If $\alpha^* \in \mathcal{S}$, then the result is immediate since

$$\sum_{i=1}^{k-1} C^*(u_i(p)) = A(\alpha^*, p) = \inf_{\alpha \in \mathcal{S}} A(\alpha, p) = A^*(p) \ .$$

Suppose now $\alpha^* \notin \mathcal{S}$, i.e., there exists a $1 \leq i \leq k-2$ for which the monotonicity conditions in $\mathcal{S}$ are not verified, that is, $\alpha_{i+1} < \alpha_i$. Since $(u_1(p), \ldots, u_{k-1}(p))$ is an increasing sequence, for a fixed p it is possible to write $u_{i+1}(p) = u_i(p) + \varepsilon$, with $\varepsilon \geq 0$. Then it is verified that

$$\begin{aligned} C(\alpha_i^*, u_{i+1}(p)) &= (1 - u_i(p) - \varepsilon)\varphi(-\alpha_i^*) + (u_i(p) + \varepsilon)\varphi(\alpha_i^*) \\ &= C(\alpha_i^*, u_i(p)) + \varepsilon(\varphi(\alpha_i^*) - \varphi(-\alpha_i^*)) \quad . \end{aligned}$$

By assumption $\varepsilon(\varphi(\alpha_i^*) - \varphi(-\alpha_i^*))$ is a decreasing function of α_i^* and so $\alpha_{i+1}^* < \alpha_i^* \implies C(\alpha_i, u_{i+1}(p)) \leq C(\alpha_{i+1}, u_{i+1}(p))$. By the optimality of α_{i+1}^*, it must be $C(\alpha_i^*, u_{i+1}(p)) = C(\alpha_{i+1}^*, u_{i+1}(p))$. This implies that the vector in which α_{i+1}^* is replaced by α_i^* has the same conditional risk and hence suggest a procedure to construct a vector that satisfies the constraints in $\mathcal{S}$ and achieves the minimal risk in $\mathbb{R}^{k-1}$. More formally, we define $\tilde{\alpha} \in \mathcal{S}$ as:

$$\tilde{\alpha}_i = \begin{cases} \alpha_1^* \text{ if } i = 1 \\ \alpha_i^* \text{ if } \alpha_{i-1}^* \leq \alpha_i^* \\ \alpha_{i-1}^* \text{ if } \alpha_{i-1}^* > \alpha_i^* \end{cases} \quad .$$

Then by the above $C(\alpha_i^*, u_i(p)) = C(\tilde{\alpha}_i, u_i(p))$ for all i and so $A(\alpha^*, p) = A(\tilde{\alpha}, p)$, which completes the proof.

■

It is easy to verify that the condition on φ of this theorem is satisfied by all the binary losses that we consider: Hinge loss, the squared Hinge loss, the logistic loss, exponential loss and the squared loss. With this result, if α_i^* is a minimizer of $C(u_i(p))$, then $(\alpha_1^*, \ldots, \alpha_{k-1}^*)$ will be a minimizer of $A(p)$. Hence, the optimal decision function for the aforementioned values of φ is simply the concatenation of known results for binary classification. These are compiled in the following table, where we show α^* and A^* for different values of φ:

- *Hinge AT*, : $\alpha_i^*(p) = \text{sign}(2u_i(p) - 1)$, $\quad A^*(p) = \sum_{i=1}^{k-1}\{1 - \big|2u_i(p) - 1\big|\}$.
- *Squared Hinge AT*, : $\alpha_i^*(p) = (2u_i(p) - 1)$, $\quad A^*(p) = \sum_{i=1}^{k-1} 4u_i(p)(1 - u_i(p))$.
- *Logistic AT*: $\alpha_i^*(p) = \log\left(\frac{u_i(p)}{1-u_i(p)}\right)$, $A^*(p) = \sum_{i=1}^{k-1}\{E(u_i(p)) - E(1 - u_i(p))\}$, where $E(t) = t\log(t)$.
- *Exponential AT*: $\alpha_i^*(p) = \frac{1}{2}\log\left(\frac{u_i(p)}{1-u_i(p)}\right)$, $\quad A^*(p) = \sum_{i=1}^{k-1} 2\sqrt{u_i(p)(1 - u_i(p))}$.
- *Squared AT*, $\alpha_i^*(p) = 2u_i(p) - 1$, $\quad A^*(p) = \sum_{i=1}^{k-1}(2 - 2u_i(p))^2$.

It is immediate to check that the models mentioned above are consistent since the decision functions coincides in sign with the minimizer of the risk defined in Lemma 3. Note that the sign of $\alpha_i^*(p)$ at $u_i(p) = \frac{1}{2}$ is irrelevant, since by Eq. (15) both signs have equal risk. We now provide a result that characterizes consistency for a convex φ:

Theorem 5 *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be convex. Then the AT surrogate is consistent if and only if φ is differentiable at 0 and $\varphi'(0) < 0$.*

Proof We postpone the proof until Section 5, where this will follow as a particular case of Theorem 11. ■

We will now derive excess risk bounds for AT. These are inequalities that relate the excess conditional risk $L(\alpha) - L^*$, to the excess in surrogate conditional risk $A(\alpha) - A^*$. For this, we will make use of the γ-transform[3] of a loss function (Bartlett et al., 2003). For a convex function φ this is defined as

$$\gamma(\theta) = \varphi(0) - C^*\left(\frac{1+\theta}{2}\right) \quad . \tag{18}$$

We will now state the excess risk bound of the AT surrogate in terms of the γ-transform:

Theorem 6 (Excess risk bounds) *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a function that verifies the following conditions:*

- *φ is convex.*
- *φ is differentiable at 0 and $\varphi'(0) < 0$.*
- *$\varphi(\beta) - \varphi(-\beta)$ is a decreasing function of β.*

Then for any $\alpha \in \mathcal{S}, p \in \Delta^k$, the following excess risk bound is verified:

$$\gamma\left(\frac{L(\alpha, p) - L^*(p)}{k-1}\right) \leq \frac{A(\alpha, p) - A^*(p)}{k-1} \quad . \tag{19}$$

Proof Let I denote the set of indices in which the sign of α does not coincide with $\underline{\alpha}$, that is, $I = \{i : \alpha_i(2u_i(p) - 1) < 0\}$. From Bartlett et al. (2003, Lemma 7), we know that if φ is convex and consistent (in the context of binary classification), then ψ is convex and we can write

$$\begin{aligned}
\psi\left(\frac{L(\alpha, p) - L^*(p)}{k-1}\right) &= \psi\left(\frac{\sum_{i\in I} |2u_i(p) - 1|}{k-1}\right) && \text{(by Eq. (16))} \\
&\leq \frac{\sum_{i\in I} \psi(|2u_i(p) - 1|)}{k-1} && \text{(by Jensen inequality)} \\
&= \sum_{i\in I} \frac{\psi(2u_i(p) - 1)}{k-1} && \text{(by symmetry of } \psi\text{)} \\
&= \sum_{i\in I} \frac{\varphi(0) - C^*\left(u_i(p)\right)}{k-1} && \text{(by definition of } \psi\text{).}
\end{aligned} \tag{20}$$

Let $q \in [0, 1], \beta \in \mathbb{R}$. If we can further show that $\beta(2q-1) \leq 0$ implies $\varphi(0) \leq C(\beta, q)$, then

$$\begin{aligned}
\sum_{i\in I} \left(\varphi(0) - C^*(u_i(p))\right) &\leq \sum_{i\in I} \left(C(\alpha_i, u_i(p)) - C^*(u_i(p))\right) \\
&\leq \sum_{i=1}^{k-1} \left(C(\alpha_i, u_i(p)) - C^*(u_i(p))\right) \\
&= A(\alpha, p) - A^*(p) \quad \text{(by Lemma 4).}
\end{aligned}$$

3. Bartlett et al. (2003) define this as the ψ-transform. However, since we already use ψ to denote the surrogate loss functions we will use letter γ in this case.

Combining this inequality with Eq. (20), we obtain the theorem. Therefore we only need to prove that $\beta(2q-1) \leq 0$ implies $\varphi(0) \leq C(\beta, q)$. Suppose $\beta(2q-1) \leq 0$. Then by Jensen inequality

$$C(\beta, q) = q\varphi(\beta) + (1-q)\varphi(-\beta) \geq \varphi(q\beta - (1-q)\beta) = \varphi(\beta(2q-1)) \quad .$$

Now, by convexity of φ we have

$$\varphi(\beta(2q-1)) \geq \varphi(0) + \beta(2q-1)\varphi'(0) \geq \varphi(0) \quad ,$$

where the last inequality follows from the fact that $\varphi'(0) < 0$ and $\beta(2q-1) \leq 0$. This concludes the proof. ■

Note that we have given the excess risk bounds in terms of the conditional risk. These can also be expressed in terms of the full risk, as done for example by Bartlett et al. (2003); Zhang (2004a). Within the conditions of the theorem, ψ is convex and because of Jensen inequality, it is verified that

$$\gamma\left(\mathbb{E}_X\left[L(f(X), \eta(X)) - L^*(\eta(X))\right]\right) \leq \mathbb{E}_X\left[\gamma(L(f(X), \eta(X))\right] \quad .$$

This, together with Eq. 19 yields the following bound in terms of the full risk

$$\begin{aligned}
\gamma\left(\frac{\mathcal{L}(f) - \mathcal{L}}{k-1}\right) &\leq \mathbb{E}_X\left[\gamma\left(\frac{L(f(X), \eta(X)) - L^*(\eta(X))}{k-1}\right)\right] \\
&\leq \mathbb{E}_X\left[\frac{A(f(X), \eta(X)) - A^*(\eta(X))}{k-1}\right] \\
&= \frac{\mathcal{A}(f) - \mathcal{A}^*}{k-1}
\end{aligned}$$

Examples of excess risk bounds. We will now derive excess bounds for different instances of the AT loss function. The values of γ only depend on φ, so we refer the reader to Bartlett et al. (2003) on the estimation of γ for the Hinge, squared Hinge and Exponential loss and to (Zhang, 2004a) for the logistic loss. Here, we will merely apply the known form of γ to the aforementioned surrogates.

- *Hinge AT*, : $\gamma(\theta) = |\theta| \implies L(\alpha, p) - L^*(p) \leq A(\alpha, p) - A^*$.
- *Squared Hinge AT*, : $\gamma(\theta) = \theta^2 \implies \frac{(L(\alpha,p) - L^*(p))^2}{k-1} \leq A(\alpha, p) - A^*$.
- *Logistic AT*: $\gamma(\theta) = \frac{\theta^2}{2} \implies \frac{(L(\alpha,p) - L^*(p))^2}{2(k-1)} \leq A(\alpha, p) - A^*$.
- *Exponential AT*: $\gamma(\theta) = 1 - \sqrt{1-\theta^2} \implies$
  $(k-1)(1 - \sqrt{1 - \frac{(L(\alpha,p) - L^*(p))^2}{k-1}}) \leq A(\alpha, p) - A^*$.
- *Squared AT*: $\gamma(\theta) = \theta^2 \implies \frac{(L(\alpha,p) - L^*(p))^2}{k-1} \leq A(\alpha, p) - A^*$.

For $k = 2$, these results generalize the known excess risk bounds for binary surrogates. For $k > 2$, the normalizing factor $\frac{1}{k-1}$ is not surprising, since contrary to the 0-1 loss, the absolute error is not bounded by 1 but by $k - 1$ instead. While similar excess risk bounds are known for multiclass classification (Zhang, 2004b; Ávila Pires et al., 2013), to the best of our knowledge this is the first time that such bounds have been developed for the AT surrogate ($k > 2$).

3.2 Cumulative Link (CL)

We now focus on the CL loss function defined in Eq. (11). This is a maximum likelihood estimator, and so the function $f : \mathcal{X} \to \mathcal{S}$ that verifies

$$\sigma(f(x)_i) = P(Y \leq i | X = x) \quad ,$$

maximizes the likelihood. Hence, assuming that the inverse of the link function σ exists, a minimizer of the surrogate loss function (given by the negative log-likelihood) $\alpha^* \in \mathcal{S}$ and A^* are given by

$$\alpha_i^*(p) = \sigma^{-1}(u_i(p)), \quad A^*(p) = \sum_{i=1}^{k} p_i \log(p_i) \quad . \tag{21}$$

This immediately leads to a characterization of consistency based on the link function σ:

Theorem 7 *Suppose σ is an invertible function. Then the CL surrogate is consistent if and only if the inverse link function verifies*

$$\left(\sigma^{-1}(t)\right)(2t - 1) > 0 \text{ for } t \neq \frac{1}{2} \quad . \tag{22}$$

Proof ($\Longleftarrow$) Suppose σ^{-1} does not verify Eq. (22), i.e., there exists a $\xi \neq 1/2$ such that $\sigma^{-1}(\xi)(2\xi - 1) \leq 0$. We consider a probability distribution where $u_1(p) = \xi$ for all $p \in \Delta^k$. In that case, by Eq. (21) $\alpha_1^*(p) = \xi$ and so this has a sign opposite to the Bayes decision function. By Eq. (16) this implies that $L(\alpha^*) > L^*$, contradiction since CL is consistent by assumption.

($\Longrightarrow$) Let $0 < i < k$. For $u_i(p) \neq 1/2$, $\alpha_i^*(p)$ agrees in sign with $2u_i(p) - 1$ and so by Lemma 3 has minimal risk. If $u_i(p) = 1/2$, then in light of Eq. (15) the risk is the same no matter the value of $\alpha_i^*(p)$. We have proved that $\alpha^*(p)$ has the same risk as a Bayes decision function, hence the CL model is consistent. This completes the proof. ■

The previous theorem captures the notion that the inverse of the link function should agree in sign with $2t - 1$. When the link function is the sigmoid function, i.e., $\sigma(t) = 1/(1 + e^{-t})$ this surrogate is convex and its inverse link function is given by the logit function, which verifies the assumptions of the theorem and hence is consistent. Its optimal decision function is given by

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

which coincides with the logistic AT surrogate. Despite the similarities between both surrogates, we have not been able to derive excess risk bounds for this surrogate since the

separability properties of the AT are not met in this case. We finish our treatment of the CL surrogate by stating the convexity of the logistic CL loss, which despite being a fundamental property of the loss, has not been proven before to the best of our knowledge.

Lemma 8 *The logistic CL surrogate is a convex function of in its second argument in the domain of definition.*

Proof We recall that the logistic CL surrogate is defined as

$$\psi_{\mathrm{CL}}(y,\alpha) := \begin{cases} -\log(\sigma(\alpha_1)) & \text{if } y = 1 \\ -\log(\sigma(\alpha_y) - \sigma(\alpha_{y-1})) & \text{if } 1 < y < k \\ -\log(1 - \sigma(\alpha_{k-1})) & \text{if } y = k \quad , \end{cases}$$

where σ is the sigmoid function. $\psi_{\mathrm{CL}}(1,\alpha)$ and $\psi_{\mathrm{CL}}(k,\alpha)$ are convex because they are log-sum-exp functions. It is thus sufficient to prove that $\psi_{\mathrm{CL}}(i,\cdot)$ 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 $a > b$ is the domain of definition.

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-concavity of the function $f(a,b) = \exp(a) - \exp(b)$ can be settled by proving the positive-definiteness of the matrix $Q = \nabla f(a,b)\nabla f(a,b)^T - f(a,b)\nabla^2 f(a,b)$ for all (a,b) in the domain $\{b > a\}$ (Boyd and Vandenberghe, 2004). In our case,

$$Q = \begin{pmatrix} \exp(a+b) & -\exp(a+b) \\ -\exp(a+b) & \exp(a+b) \end{pmatrix} = \exp(a+b) \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \quad ,$$

which is a positive semidefinite matrix with eigenvalues $2\exp(a+b)$ and 0. This proves that Q is positive semidefinite and thus $\psi_{\mathrm{CL}}(i,\cdot)$ is a convex function. ■

3.3 Least absolute deviation

We will now prove consistency of the least absolute deviation (LAD) surrogate. Consistency of this surrogate was already proven for the case $k = 3$ by Ramaswamy and Agarwal (2012). For completeness, we provide here an alternative proof for an arbitrary number of classes.

Theorem 9 *The least absolute deviation surrogate is consistent.*

Proof Recall that for $y \in \mathcal{Y}, \alpha \in \mathcal{S}$, the LAD surrogate is given by

$$\psi_{\mathrm{LAD}}(y,\alpha) = \left| y + \alpha_1 - \frac{3}{2} \right| \quad .$$

The pointwise surrogate risk is then given by

$$A(\alpha, p) = \sum_{i=1}^{k} p_i \psi_{\text{LAD}}(y, \alpha) = \mathbb{E}_{Y \sim p}\left[\left|Y + \alpha_1 - \frac{3}{2}\right|\right] \quad ,$$

where $Y \sim p$ means that Y is distributed according to a multinomial distribution with parameter $p \in \Delta^k$. By the optimality conditions of the median, a value that minimizes this conditional risk is given by

$$\alpha_1^*(p) \in \text{Median}_{Y \sim p}\left(\frac{3}{2} - Y\right) \quad ,$$

where Med is the median, that is, $\alpha_1^*(p)$ is any value that verifies

$$P\left(\frac{3}{2} - Y \leq \alpha_1^*(p)\right) \geq \frac{1}{2} \text{ and } P\left(\frac{3}{2} - Y \geq \alpha_1^*(p)\right) \geq \frac{1}{2} \quad .$$

We will now prove that LAD is consistent by showing that $L(\alpha^*(p), p) = L(\underline{\alpha}(p), p)$, where $\underline{\alpha}$ is the Bayes decision function described in Lemma 3. Let $r^* = \text{pred}(\underline{\alpha}(p))$ and I denote the set $I = \{i : \alpha_i^*(p)(2u_i(p) - 1) < 0\}$. Suppose this set is non-empty and let $i \in I$. We distinguish the cases $\alpha_i^*(p) > 0$ and $\alpha_i^*(p) < 0$:

- $\alpha_i^*(p) < 0$. By Eq. (13), α_i^* and α_1^* are related by $\alpha_i^* = i - 1 + \alpha_1^*$. Then it is verified that

$$\begin{aligned} P\left(\frac{3}{2} - Y \geq \alpha_1^*(p)\right) &= P\left(\frac{3}{2} - Y \geq \alpha_i^* - i + 1\right) = P\left(\frac{1}{2} + i - \alpha_i^* \geq Y\right) \\ &\geq P\left(\frac{1}{2} + i \geq Y\right) = u_i(p) \quad . \end{aligned}$$

  By assumption, $\alpha_i^*(p)(2u_i(p) - 1) < 0$, which implies $u_i(p) > 1/2$. Hence, by the above we have that $P\left(\frac{3}{2} - Y \geq \alpha_1^*(p)\right) > 1/2$. At the same time, by the definition of median, $P\left(\frac{3}{2} - Y \geq \alpha_1^*(p)\right) \leq 1/2$, contradiction.

- $\alpha_i^*(p) > 0$. Using the same reasoning as before, it is verified that

$$\begin{aligned} P\left(\frac{3}{2} - Y \leq \alpha_1^*(p)\right) &= P\left(\frac{3}{2} - Y \leq \alpha_i^* - i + 1\right) = P\left(\frac{1}{2} + i - \alpha_i^* \leq Y\right) \\ &\geq P\left(\frac{1}{2} + i \leq Y\right) = 1 - u_i(p) \quad . \end{aligned}$$

  By assumption $u_i(p) < 1/2 \implies P\left(\frac{3}{2} - Y \leq \alpha_1^*(p)\right) > 1/2$. At the same time, by the definition of median, $P\left(\frac{3}{2} - Y \geq \alpha_1^*(p)\right) \leq 1/2$, contradiction.

Supposing I not empty has lead to contradictions in both cases, hence $I = \emptyset$. By Eq. (16), $L(\alpha^*(p), p) = L(\underline{\alpha}(p), p)$, which concludes the proof. ■

4. Surrogates of the 0-1 loss

Perhaps surprisingly, one popular model for ordinal regression is not a surrogate of the absolute error but of the 0-1 loss. In this section we focus on the 0-1 loss and we provide a characterization of consistency for the immediate threshold loss function.

4.1 Immediate Thresholds

For the immediate threshold, the conditional risk can be expressed as

$$A(\alpha, p) = \sum_{i=1}^{k} p_j \psi_{\mathrm{IT}(j,\alpha)} = \sum_{i=1}^{k-1} p_i \varphi(-\alpha_i) + p_{i+1}\varphi(\alpha_i)$$

As pointed out by Chu and Keerthi (2005), and contrary to what happened for AT surrogate, the constraints can not be ignored in general when computing A^*. Results that rely on this property such as the excess error bound of Theorem 6 will not translate directly for the IT loss. However, we will still be able to characterize the functions φ that lead to a consistent surrogate, in a result analogous to Theorem 5 for the AT surrogate.

Theorem 10 *Let φ be convex. Then the immediate threshold surrogate is Fisher consistent if and only if φ is differentiable at 0 and $\varphi'(0) < 0$.*

Proof As for the AT surrogate, this can be seen as a particular case of Theorem 11 with ℓ the 0-1 loss. We will postpone the proof until Section 5. ■

5. Extension to other admissible loss functions

In this section we will show that the AT and IT surrogates can be seen as particular instances of a family of loss functions for which we will be able to provide a characterization of consistency.

The admissibility criterion that we require on the loss function is that this is of the form $\ell(i, \alpha) = g\left(\left|i - \mathrm{pred}(\alpha)\right|\right)$, where g is an increasing function. Intuitively, this condition implies that labels further away from the true label are penalized more than those closer by. This criterion is general enough to contain all losses considered before such as the absolute error and (albeit in a degenerate sense) 0-1 loss. It also contains other loss functions that we have not yet considered such as the squared error ($g(t) = t^2$). A very similar condition is the V-shape property of (Li and Lin, 2007). This property captures the notion that the loss should not decrease as the predicted value moves away from the true value by imposing that ℓ verifies $\ell(y, \alpha) \leq \ell(y, \alpha')$ for $\left|y - \mathrm{pred}(\alpha)\right| \leq \left|y - \mathrm{pred}(\alpha')\right|$. The only difference between the two conditions is that our admissibility criterion adds a symmetric condition, i.e., ℓ verifies that the loss of predicting a when the true label is b is the same as the loss of predicting b when the true label is a, which is not necessarily true for V-shaped loss functions. We conjecture that the results in this section are valid for general V-shaped loss functions, although for simplicity we have only proven results for symmetric V-shaped loss functions. For the rest of this section, we will consider that ℓ is a loss function that verifies the admissibility criterion.

Let $c_i = g(i) - g(i-1)$ and note that with this notation g can always be written as a sum of c_i with the formula $g(i) = \sum_{j=1}^{i} c_j$ and that $c_i \geq 0$ by the admissibility property. Using this notation, and following the same development as in Eq. (8), any admissible loss function can be written as a sum of c_i as

$$\ell(y, \alpha) = g\left(\sum_{i=1}^{y-1} [\![\alpha_i \geq 0]\!] + \sum_{i=y}^{k-1} [\![\alpha_i < 0]\!]\right) = \sum_{i=1}^{y-1} c_{y-i} [\![\alpha_i \geq 0]\!] + \sum_{i=y}^{k-1} c_{i-y+1} [\![\alpha_i < 0]\!] \quad . \tag{23}$$

In light of this, it seems natural to define a surrogate for this general loss function by replacing the 0-1 loss with a surrogate as the hinge or logistic that we will denote by φ. This defines a new surrogate that we will denote *generalized all threshold (GAT)*:

$$\psi_{\text{GAT}}(y, \alpha) := \sum_{i=1}^{y-1} \varphi(-\alpha_i) c_{y-i} + \sum_{i=y}^{k-1} \varphi(\alpha_i) c_{i-y+1} \quad .$$

In the special case of the absolute error, c_i is identically equal to 1 and we recover AT loss of Eq. (10). Likewise, for the zero-one loss, c_i will be one for $i \in \{y-1, y\}$ and zero otherwise, recovering the IT loss of Eq. (12). We will now present the main result of this section, which has Theorems 5 and 10 as particular cases.

Theorem 11 *Let φ be convex. Then the GAT surrogate is consistent if and only if φ is differentiable at 0 and $\varphi'(0) < 0$.*

Before presenting the proof this theorem, we will need some auxiliary results. Unlike for the absolute error, in this case we will not be able to derive a closed form of the optimal decision function. However, we will be able to derive a formula for the excess risk in terms of the functions $u, v : \Delta^k \to \mathbb{R}^{k-1}$, defined as

$$u_i(p) = \sum_{j=1}^{i} p_j c_{i-j+1} \quad v_i(p) = \sum_{j=i+1}^{k} p_j c_{j-i} \quad .$$

Note that we have overloaded the function u defined in Section 3. This is not a coincidence, as when ℓ is the absolute error both definitions coincide. Using this notation, the surrogate risk can be conveniently written as

$$\begin{aligned} A(\alpha, p) &= \sum_{i=1}^{k} p_i \psi_{\text{GAT}}(i, \alpha) = \sum_{i=1}^{k} p_i \left(\sum_{j=1}^{i-1} \varphi(-\alpha_j) c_{i-j} + \sum_{j=i}^{k-1} \varphi(\alpha_j) c_{j-i+1}\right) \\ &= \sum_{i=1}^{k-1} v_i(p) \varphi(-\alpha_i) + u_i(p) \varphi(\alpha_i) \quad , \end{aligned}$$

and we have the following formulas for the excess risk:

Lemma 12 *Let $p \in \Delta^k$, $\alpha \in \mathcal{S}$, $r = pred(\alpha)$ and r^* be the label predicted by any Bayes decision function at p. Then, it is verified that*

$$L(\alpha, p) - L^*(p) = \begin{cases} \sum_{i=r}^{r^*} (v_i(p) - u_i(p)) & \text{if } r < r^* \\ \sum_{i=r^*}^{r-1} (u_i(p) - v_i(p)) & \text{if } r > r^* \end{cases} .$$

Proof The risk can be expressed in terms of u_i and v_i as

$$\begin{aligned} L(\alpha) &= \sum_{i=1}^{k} p_i g(|r-i|) = \sum_{i=1}^{r-1} p_i g(r-i) + \sum_{i=r+1}^{k} p_i g(i-r) \\ &= \sum_{i=1}^{r-1} p_i \sum_{j=1}^{r-i} c_j + \sum_{i=r+1}^{k} p_i \sum_{j=1}^{i-r} c_j \\ &= \sum_{i=1}^{r-1} u_i + \sum_{i=r}^{k-1} v_i \quad , \end{aligned} \tag{24}$$

hence for $r < r^*$,

$$0 \leq L(\alpha) - L^* = \sum_{i=1}^{r-1} u_i + \sum_{i=r}^{k-1} v_i - \left(\sum_{i=1}^{r^*-1} u_i + \sum_{i=r^*}^{k-1} v_i \right) = \sum_{i=r}^{r^*-1} (v_i - u_i) \quad ,$$

and similarly for $r > r^*$

$$0 \leq L(\alpha) - L^* = \sum_{i=1}^{r-1} u_i + \sum_{i=r}^{k-1} v_i - \left(\sum_{i=1}^{r^*-1} u_i + \sum_{i=r^*}^{k-1} v_i \right) = \sum_{i=r^*}^{r-1} (u_i - v_i) \quad ,$$

■

Proof [Proof of Theorem 11] This proof loosely follows the steps by Bartlett et al. (2003, Theorem 6), with special care to ensure that the optimal value of the surrogate risk lies within $\mathcal{S}$ and adapted to consider multiple classes. We denote by α^* the value in $\mathcal{S}$ that minimizes $A(\cdot)$, by r the prediction at α^* and by r^* the prediction of any Bayes decision function. For simplicity, we consider p to be fixed and write u_i, v_i to denote $u_i(p), v_i(p)$ respectively.

($\Longrightarrow$) We first prove that consistency implies φ is differentiable at 0 and $\varphi'(0) < 0$. Since φ is convex, we can find subgradients $g_1 \geq g_2$ such that, for all $\beta \in \mathbb{R}$

$$\begin{aligned} \varphi(\beta) &\geq g_1 \beta + \varphi(0) \\ \varphi(\beta) &\geq g_2 \beta + \varphi(0) \quad . \end{aligned}$$

Then we have for all i

$$\begin{aligned} v_i\varphi(-\beta) + u_i\varphi(\beta) &\geq v_i(g_1\beta + \varphi(0)) + u_i(-g_2\beta + \varphi(0)) \\ &= (v_i g_1 - u_i g_2)\beta + (v_i + u_i)\varphi(0) \quad . \end{aligned} \tag{25}$$

For $0 < \varepsilon < 1/2$, we will consider the following vector of conditional probabilities

$$p = \left(0, \cdots, 0, \frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon\right) \quad ,$$

from where u_i and v_i take the following simple form

$$u_i = \begin{cases} p_{k-1}c_1 & \text{if } i = k-1 \\ 0 & \text{otherwise} \end{cases} , \quad v_i = \begin{cases} p_k c_1 & \text{if } i = k-1 \\ p_{k-1}c_{k-r-1} + p_k c_{k-r} & \text{otherwise} \end{cases}$$

hence by Eq. (24) consistency implies $r = k$ and so we must have $\alpha^*_{k-1} < 0$.

Let now $\tilde{\alpha} \in \mathcal{S}$ be a vector that equals α^* in all except the last component, which is zero (i.e., $\tilde{\alpha}_{k-1} = 0$). We will now prove that if $g_1 > g_2$ then $A(\tilde{\alpha}, p) < A(\alpha^*, p)$ leading to a contradiction. Given that $g_1 > g_2$, it is always possible to choose $\varepsilon > 0$ such that it verifies

$$1 < \frac{v_{k-1}}{u_{k-1}} = \frac{1 + 2\varepsilon}{1 - 2\varepsilon} < \frac{g_2}{g_1} \quad ,$$

in which case $v_{k-1}g_1 - u_{k-1}g_2 < 0$ and so $(v_{k-1}g_1 - u_{k-1}g_2)\alpha^*_{k-1} > 0$. Replacing in Eq. (25) yields

$$\begin{aligned} A(\alpha^*, p) &= \sum_{i=1}^{k-1} v_i(p)\varphi(-\alpha^*_i) + u_i\varphi(\alpha^*_i) \\ &= \sum_{i=1}^{k-2} \left\{v_i(p)\varphi(-\alpha^*_i) + u_i\varphi(\alpha^*_i)\right\} + v_{k-1}\varphi(-\alpha_{k-1}) + u_{k-1}\varphi(\alpha_{k-1}) \\ &> \sum_{i=1}^{k-2} \left\{v_i(p)\varphi(-\alpha^*_i) + u_i\varphi(\alpha^*_i)\right\} + v_{k-1}\varphi(0) + u_{k-1}\varphi(0) \\ &= A(\tilde{\alpha}, p) \quad , \end{aligned}$$

contradiction, so we can say that if the GAT loss is consistent, then φ is differentiable at 0. To see that we must also have $\varphi'(0) < 0$, notice that from Eq. (25) we have

$$A_i(\beta) \geq (v_i - u_i)\varphi'(0)\beta + A_i(0) \quad .$$

But for any $v_i > u_i$ and $\beta < 0$, if $\varphi'(0) \geq 0$, then this expression is greater than $A_i(0)$. Hence, if GAT is consistent then $\varphi'(0) < 0$, which concludes one of the implications of the proof.

($\Longleftarrow$) We now prove that if φ is differentiable at 0 and $\varphi'(0) < 0$, then GAT is consistent.

The first order optimality conditions states that there exists $\lambda_i \geq 0$ such that the optimal value of $A(\alpha, p)$ subject to $\alpha \in \mathcal{S}$ is the minimizer of the following unconstrained function:

$$G(\alpha) = A(\alpha, p) + \sum_{i=1}^{k-1} \lambda_i(\alpha_i - \alpha_{i+1}) \quad .$$

We show that assuming GAT is not consistent (i.e., $L(\alpha^*) > L^*$) leads to a contradiction and hence GAT must be consistent.

We start by computing the partial derivative of G at zero:

$$\frac{\partial G}{\partial \alpha_i}\Big|_{\alpha_i=0} = (u_i - v_i)\varphi'(0) - \lambda_{i-1} + \lambda_i \quad ,$$

were for convenience $\lambda_0 = 0$. Note that at $i = r$, α^* verifies $\alpha^*_{r-1} < 0 \leq \alpha^*_r$ by definition of prediction function. Hence, by complementary slackness $\lambda_{r-1} = 0$. Suppose $r < r^*$. Then, the addition of all partial derivatives between r and r^* yields

$$\sum_{i=r}^{r^*} \frac{\partial G}{\partial \alpha_i}\Big|_{\alpha_i=0} = \left(\sum_{i=r}^{r^*} u_i - v_i\right) \varphi'(0) + \lambda_{r^*} \quad ,$$

which by Lemma 12 is strictly positive. However, by the definition of prediction function, $\alpha^*_{i \geq r} \geq 0$ and so by convexity $\partial G / \partial \alpha_{i \geq r} \leq 0$ at $\alpha_i = 0$, contradiction.

Likewise, suppose $r > r^*$. Then, the addition of all partial derivatives between r^* and r yields

$$\sum_{i=r^*}^{r} \frac{\partial G}{\partial \alpha_i}\Big|_{\alpha_i=0} = \left(\sum_{i=r^*}^{r} u_i - v_i\right) \varphi'(0) - \lambda_{r^*-1} \quad ,$$

which by Lemma 12 is strictly negative. However, by the definition of prediction function, $\alpha^*_{i<r} < 0$ and so by convexity $\partial G / \partial \alpha_{i \geq r} \geq 0$ at $\alpha_i = 0$, contradiction. ■

6. Threshold-based decision functions and parametric consistency

In this section we revisit the assumption that the optimal decision function can be estimated independently at each point $x \in \mathcal{X}$. This is implicitly assumed on most consistency studies, however in practice it is often the case to have models that enforce inter-observational constraints (e.g. smoothness). In the case of ordinal regression it is often the case that the decision functions are of the form

$$f(x) = (\theta_1 - g(x), \theta_2 - g(X), \ldots, \theta_{k-1} - g(X)) \quad , \tag{26}$$

where $(\theta_1, \ldots, \theta_{k-1})$ is an increasing vector (i.e., its components form an increasing sequence) known as the *vector of thresholds* (hence the appearance of the name thresholds in many models) and g is a measurable function. We will call decision functions of this form *threshold-based* decision functions. All the examined models with the exception of the least

absolute deviation are commonly constrained to this family of decision functions (Chu and Keerthi, 2005; Rennie and Srebro, 2005; Lin and Li, 2006; Shashua and Levin, 2003).

The main issue with such decision functions is that since the vector of thresholds is estimated from the data, it is no longer true that the optimal decision function can be estimated independently at each point. This implies that the pointwise characterization of Fisher consistency described in Lemma 2 does no longer hold when restricted to this family and hence the consistency proofs of past sections break down.

Let $\mathcal{F}$ be the set of functions of the form of Eq. (26). We will now apply the notion of *$\mathcal{F}$-consistency* or *parametric consistency* of (Shi et al., 2015) to the threshold-based setting. This is merely the notion of Fisher consistency where the decision functions are restricted to a family of interest:

Definition 13 *($\mathcal{F}$-**Consistency**) Given a surrogate loss function $\psi : \mathcal{Y} \times \mathcal{S} \to \mathbb{R}$, we will say that the surrogate loss function ψ is $\mathcal{F}$-consistent with respect to the loss $\ell : \mathcal{Y} \times \mathcal{S} \to \mathcal{Y}$ if for every probability distribution over $X \times Y$ it is verified that every minimizer of the ψ-risk reaches the optimal risk in $\mathcal{F}$, that is,*

$$f^* \in \arg\min_{f \in \mathcal{F}} \mathcal{A}(f) \implies \mathcal{L}(f^*) = \inf_{f \in \mathcal{F}} \mathcal{L}(f) \quad .$$

We will show that by imposing additional constraints on the probability distribution P we will be able to derive $\mathcal{F}$-consistency for particular surrogates in the following theorem.

Theorem 14 *Let R denote the odds-ratio at $x \in \mathcal{X}$, that is,*

$$R_i(x) = \frac{u_i(\eta(x))}{1 - u_i(\eta(x))} \cdot \frac{1 - u_{i+1}(\eta(x))}{u_{i+1}(\eta(x))} \quad ,$$

where $\eta(x)$ is the vector of conditional probabilities defined in Section 1.2. Then, if the odds-ratio is independent of x for all $0 < i < k - 1$, that is, if

$$R_i(x_1) = R_i(x_2) \quad \forall x_1, x_2 \in \mathcal{X}$$

then the logistic all threshold and the logistic CL are $\mathcal{F}$-consistent.

Proof It will be sufficient to prove that under the constraints on P, the optimal decision function for the unconstrained problem belongs to $\mathcal{F}$. In Section 3 we derived the optimal decision function for the logistic all threshold and the logistic CL. Hence, we can write

$$\alpha_i^*(\eta(x)) - \alpha_{i+1}^*(\eta(x)) = \log\left(\frac{u_i(\eta(x))}{1 - u_i(\eta(x))}\right) - \log\left(\frac{u_{i+1}(\eta(x))}{1 - u_{i+1}(\eta(x))}\right) = \log\left(\frac{R_i(x)}{R_{i+1}(x)}\right)$$

Since the condition that α is a threshold-based decision if and only if $\alpha_i - \alpha_{i+1}$ does not depend on $x \in \mathcal{X}$, the result follows by the monotonicity of log. ■

7. Experiments: A novel surrogate for the squared error

While the focus of this work is a theoretical investigation of consistency, we have also conducted experiments to study a novel surrogate suggested by the results of the last section. There, we constructed a surrogate that is consistent with any loss function that verifies an admissible criterion. One example of such loss function is the squared error,

$$\ell(y, \alpha) = (y - \text{pred}(\alpha))^2 \quad .$$

One of the byproducts of Theorem 11 is that this surrogate is consistent with respect to the squared error:

$$\psi(y, \alpha) = \sum_{i=1}^{y-1} \varphi(-\alpha_i)(2(y-i)-1) + \sum_{i=y}^{k-1} \varphi(\alpha_i)(2(i-y)+1) \quad .$$

To the best of our knowledge, this is a novel surrogate. We compare the cross-validation error of this surrogate on different datasets against the least squares surrogate:

$$\psi_{\text{LS}}(y, \beta) = (y - \beta)^2 \quad ,$$

where $\beta \in \mathbb{R}$ and prediction is given by rounding to the closest integer. In both cases, we consider linear decision functions, i.e.

$$\alpha = (\theta_1 - \langle w, x \rangle, \ldots, \theta_{k-1} - \langle w, x \rangle) \quad \text{and} \quad \beta = \langle w, x \rangle \quad .$$

In each case, the optimal values of w, θ where found by minimizing the empirical surrogate risk. For the training sample $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, $x_i \in \mathbb{R}^p$, it yielding the following optimization problems for GAT and least squares, respectively:

$$\underset{\theta \in \mathcal{S}, w \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{y_i - 1} \varphi(\langle w, x_i \rangle - \theta_j)(2(y-j)-1) + \sum_{j=y_i}^{k-1} \varphi(\theta_j - \langle w, x_i \rangle)(2(j-y)+1) \right\}$$

$$\underset{w \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} (y_i - \langle w, x_i \rangle)^2$$

The different datasets that we will consider are described in (Chu and Keerthi, 2005) and can be download from the authors website[4]. We display results for the 9 datasets of SET I using the version of the datasets with 5 bins, although similar results were observed when using the datasets with 10 bins. Given the small dimensionality of the datasets (between 6 and 60) and the comparatively high number of samples (between 185 and 4000), we did not consider the use of regularization.

Performance is computed as the squared error on left out data, averaged over 20 folds. We report this performance in Figure 7, where it can be seen that the GAT surrogate outperforms LS on 7 out of 9 datasets, showing that GAT yields a highly competitive surrogate in practice.

4. `http://www.gatsby.ucl.ac.uk/~chuwei/ordinalregression.html`.

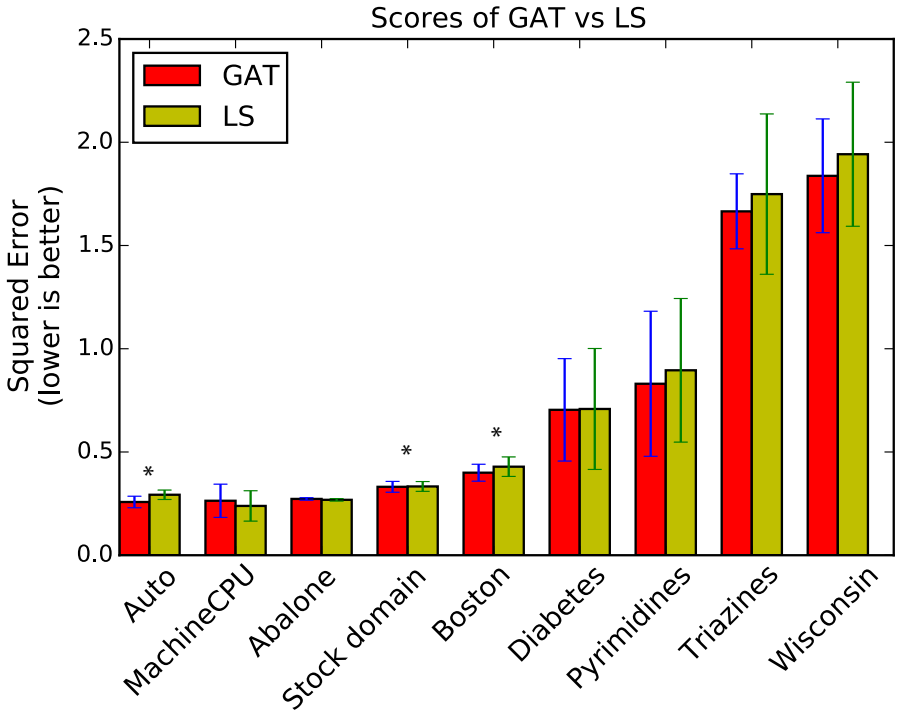

Figure 1: Scores of the generalized all threshold (GAT) and least squares (LS) surrogate on 6 different datasets. The scores are computed as the squared error between the prediction and the true labels on left out data, averaged over 20 cross-validation folds. On 7 out of 9 datasets all the GAT surrogate outperforms the least squares estimator, showing that this surrogate yields a highly competitive model. Datasets for which a Wilcoxon signed-rank test rejected the null hypothesis that the means are equal with p-value < 0.01 are highlighted by a $*$ symbol over the bars.

8. Conclusions

In this paper we have characterized the consistency for a rich family of surrogate loss functions used for ordinal regression. Our aim is to bridge the gap between the consistency properties known for classification and ranking and those known for ordinal regression.

We have first described a wide family of ordinal regression methods under the same framework. The surrogates of the absolute error that we have considered are the all threshold (AT), cumulative link (CL), and least absolute deviation (LAD), while the surrogate for the 0-1 loss is the immediate threshold (IT).

For all the surrogates considered, we have characterized its consistency. For AT and IT, consistency is characterized by the derivative of a real-valued convex function at zero (Theorems 5 and 10 respectively). For CL, consistency is characterized by a simple condition on its link function (Theorem 7) and for LAD we have extended the proof of Ramaswamy and Agarwal (2012) to an arbitrary number of classes (Theorem 5). Furthermore, we have proven that AT verifies a decomposability property and using this property we have provided excess risk bounds that generalize those of Bartlett et al. (2003) for binary classification(Theorem 6).

The derivation we have given when introducing IT and AT are identical except for the underlying loss function. This suggest that both can be seen as special cases of a more general family of surrogates. In Section 5 we have constructed such surrogate and characterized its consistency with respect to a general loss function that verifies an admissibility condition. Again, the characterization only relies on the derivative at zero of a convex real-valued function. We named this surrogate generalized all threshold (GAT).

In Section 6 we have turned back to examine one of the assumptions described in the introduction and that is common to the vast majority of consistency studies, i.e., that the optimal decision function can be estimated independently at every sample. However, in the setting of ordinal regression it is common for decision functions to have a particular structure known as threshold-based decision functions and which violates this assumption. Following (Shi et al., 2015), we are able to prove a restricted notion of consistency known as $\mathcal{F}$-consistency or parametric consistency on two surrogates by enforcing constraints on the probability distribution P. We believe this restricted notion of consistency to be important in practice and we look forward to see consistency studies extended to consider different types of decision functions, such as smooth functions, polynomial functions, etc.

Finally, in Section 7 we provide an empirical comparison for the GAT surrogate. The underlying loss function that we consider in this case is the squared error, in which case GAT yields a novel surrogate. We compare this surrogate against the least squares surrogate in terms of cross-validation error. Our results show that GAT outperforms least squares on 7 out of 9 datasets, showing the pertinence of such surrogate on real-world datasets.

9. Acknowledgments

We would like to thank our colleague Guillaume Obozinski for fruitful discussions. FP acknowledges financial support from INRIA, Parietal project-team under grants IRMGroup ANR-10-BLAN-0126-02 and BrainPedia ANR-10-JCJC 1408-01 and the "Chaire Économie

des Nouvelles Données", under the auspices of Institut Louis Bachelier, Havas-Media and Université Paris-Dauphine.

References

Shivani Agarwal. Generalization bounds for some ordinal regression algorithms. *ALT '08 Proceedings of the 19th international conference on Algorithmic Learning Theory*, 2008.

Alan Agresti. *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons, 2010.

Cande V. Ananth and David G. Kleinbaum. Regression models for ordinal responses: a review of methods and applications. *International journal of epidemiology*, 26(6):1323–1333, 1997.

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

Bernardo Ávila Pires, Csaba Szepesvari, and Mohammad Ghavamzadeh. Cost-sensitive multiclass classification risk bounds. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1391–1399, 2013.

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.

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.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

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

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, University of Pennsylvania, November 2005.

Clément Calauzenes, Nicolas Usunier, Patrick Gallinari, et al. On the (non-) existence of convex, calibrated surrogate losses for ranking. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

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

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.

Koby Crammer and Yoram Singer. Pranking with ranking. In *Advances in Neural Information Processing Systems (NIPS)*, 2001.

Koby Crammer and Yoram Singer. Online ranking by projecting. *Neural Computation*, 17 (1):145–175, 2005.

Orla M. Doyle, John Ashburner, F.O. Zelaya, Stephen C.R. Williams, Mitul A. Mehta, and Andre F. Marquand. Multivariate decoding of brain images using ordinal regression. *NeuroImage*, 81:347–357, 2013.

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.

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.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

William H. Greene. *Econometric analysis*. Prentice-Hall, Inc, 1997.

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.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Support vector learning for ordinal regression. *IET Conference Proceedings*, pages 97–102(5), January 1999.

Donald E Knuth. Two notes on notation. *American Mathematical Monthly*, pages 403–422, 1992.

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

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.

Ling Li and Hsuan-tien Lin. Ordinal regression by extended binary classification. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2007.

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.

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

Tie-Yan Liu. *Learning to rank for information retrieval*. Springer Science & Business Media, 2011.

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

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

Harish G Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. *arXiv preprint arXiv:1408.2764*, 2014.

Mark D Reid and Robert C Williamson. Composite Binary Losses. 11:2387–2422, 2010.

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.

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

Qinfeng Shi, Mark Reid, Tiberio Caetano, Anton Van den Hengel, and Zhenhua Wang. A hybrid loss for multiclass and structured prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):2–12, 2015.

Ingo Steinwart. Support vector machines are universally consistent. *Journal of Complexity*, 18(3):768–791, September 2002.

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

Ambuj Tewari and Peter L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32:56–85, 2004a.

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