**Noname manuscript No.**
(will be inserted by the editor)

# V-shaped Interval Insensitive Loss for Ordinal Classification

**Kostiantyn Antoniuk · Vojtěch Franc · Václav Hlaváč**

Received: date / Accepted: date

**Abstract** We address a problem of learning ordinal classifiers from partially annotated examples. We introduce a V-shaped interval-insensitive loss function to measure discrepancy between predictions of an ordinal classifier and a partial annotation provided in the form of intervals of candidate labels. We show that under reasonable assumptions on the annotation process the Bayes risk of the ordinal classifier can be bounded by the expectation of an associated interval-insensitive loss. The bounds justify learning the ordinal classifier from partially annotated examples via minimization of an empirical estimate of the interval-insensitive loss. We propose several convex surrogates of the interval-insensitive loss which are used to formulate convex learning problems. We described a variant of the cutting plane method which can solve large instances of the learning problems. Experiments on a real-life application of human age estimation show that the ordinal classifier learned from cheap par-

Kostiantyn Antoniuk
Center for Machine Perception, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Technická 2, 166 27 Prague 6 Czech Republic
Tel.: +420-22435-5729
E-mail: antonkos@cmp.felk.cvut.cz

Vojtěch Franc
Center for Machine Perception, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Technická 2, 166 27 Prague 6 Czech Republic
Tel.: +420-22435-7665
E-mail: xfrancv@cmp.felk.cvut.cz

Václav Hlaváč
Center for Machine Perception, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Technická 2, 166 27 Prague 6 Czech Republic
Tel.: +420-22435-7465
Fax:+420-22435-7385
E-mail: hlavac@fel.cvut.cz

tially annotated examples can achieve accuracy matching the results of the so-far used supervised methods which require expensive precisely annotated examples.

**Keywords** ordinal classification, partially annotated examples, risk minimization

## 1 Introduction

The ordinal classification model (also ordinal regression) is used in problems where the set of labels is fully ordered, for example, the label can be an age category (0-9,10-19,. . .,90-99) or a respondent answer to certain question (from strongly agree to strongly disagree). The ordinal classifiers are routinely used in social sciences, epidemiology, information retrieval or computer vision.

Recently, many supervised algorithms have been proposed for discriminative learning of the ordinal classifiers. The discriminative methods learn parameters of an ordinal classifier by minimizing a regularized convex proxy of the empirical risk. A Perceptron-like on-line algorithm PRank has been proposed in [Crammer and Singer, 2001]. A large-margin principle has been applied for learning ordinal classifiers in [Shashua and Levin, 2002]. The paper [Chu and Keerthi, 2005] proposed Support Vector Ordinal Regression algorithm with explicit constraints (SVOR-EXP) and the SVOR algorithm with implicit constraints (SVOR-IMC). Unlike [Shashua and Levin, 2002], the SVOR-EXP and SVOR-IMC guarantee the learned ordinal classifier to be statistically plausible. The same approach have been proposed independently by [Rennie and Srebro, 2005] who introduce so called immediate-threshold loss and all-thresholds loss functions. Minimization of a quadratically regularized immediate-threshold loss and the all-threshold loss are equivalent to the SVOR-EXP and the SVOR-IMC formulation, respectively. A generic framework proposed in [Li and Lin, 2006], of which the SVOR-EXP and SVOR-IMC are special instances, allows to convert learning of the ordinal classifier into learning of two-class SVM classifier with weighted examples.

Estimating parameters of a probabilistic model by the Maximum Likelihood (ML) method is another paradigm that can be used to learn ordinal classifiers. A plug-in ordinal classifier can be then constructed by substituting the estimated model to the optimal decision rule derived for a particular loss function (see e.g. [Debczynski et al., 2008] for a list of losses and corresponding decision functions suitable for ordinal classification). Parametric probability distributions suitable for modeling the ordinal labels have been proposed in [McCullagh, 1980, Fu and Simpson, 2002, Rennie and Srebro, 2005]. Besides the parametric methods, the non-parametric probabilistic approaches like the Gaussian processes have been also applied [Chu and Ghahramani, 2005].

Properties of the discriminative and the ML based methods are complementary to each other. The ML approach can be directly applied in the presence of incomplete annotation (e.g. when label interval is given instead of a single label as considered in this paper) by using the Expectation-Maximization

algorithms [Dempster et al., 1997]. However, the ML methods are sensitive to model mis-specification which complicates their application in modeling complex high-dimensional data. In contrast, the discriminative methods are known to be robust against the model mis-specification while their extension for learning from partial annotations is not trivial. To our best knowledge, the existing discriminative approaches for ordinal classification assume the precisely annotation only, that is, each training instance is annotated by exactly one label.

In this paper, we consider learning of the ordinal classifiers from partially annotated examples. We assume that each training input is annotated by an interval of candidate labels rather than a single label. This setting is common in practice. For example, let us assume a computer vision problem of learning an ordinal classifier predicting age from a facial image (e.g. [Ramanathan et al., 2009, Chang et al., 2011]). In this case, examples of face images are typically downloaded from the Internet and the age of depicted people is estimated by a human annotator. Providing a reliable year-exact age just from a face image is difficult if not possible. It is it is more natural and easier for humans to provide an interval of ages. The interval annotation can be also obtained in an automated way e.g. by the method of [Kotlowski et al., 2008] removing inconsistencies in the data.

To deal with the interval annotations, we propose an interval-insensitive loss function which extends an arbitrary (supervised) V-shaped loss to the interval setting. The interval-insensitive loss measures a discrepancy between the interval of candidate labels given in the annotation and a label predicted by the classifier. Our interval-insensitive loss can be seen as the ordinal regression counterpart of the $\epsilon$-insensitive loss used in the Support Vector Regression [Vapnik, 1998]. We prove that under reasonable assumptions on the annotation process, the Bayes risk of the ordinal classifier can be bounded by the expectation of the interval-insensitive loss. The bounds justify learning the ordinal classifier via minimization of an empirical estimate of the interval-insensitive loss. The tightness of the bound depends on two intuitive parameters characterizing the annotations process. Moreover, we show how to control the parameters in practice by properly designing the annotation process. We propose a convex surrogate of an arbitrary V-shaped interval-insensitive loss which is used to formulate a convex learning problem. We also show how to modify the existing supervised methods, the SVOR-EXP and the SVOR-IMC algorithms, in order to minimize a convex surrogate of the interval-insensitive loss associated with the 0/1-loss and the Mean Absolute Error (MAE) loss. We design a variant of the cutting plane algorithm which can solve large instances of the learning problems efficiently.

Discriminative learning from partially annotated examples has been recently studied in the context of a generic multi-class classifiers [Cour et al., 2011], the Hidden Markov Chain based classifiers [Do and Artières, 2009], generic structured output models [Lou and Hamprecht, 2012], the multi-instance learning [Jie and Orabona, 2010] etc. All these methods translate learning to minimization of a partial loss evaluating discrepancy between the classifier pre-

dictions and partial annotations. The partial loss is defined as minimal value of a supervised loss (defined on a pair of labels, e.g. 0/1-loss) over all candidate labels consistent with the partial annotation. Our interval-insensitive loss can be seen as an application of such type of partial losses in the context of the ordinal classification. In particular, we analyze the partial annotation in the form of intervals of the candidate labels and the Mean Absolute Error which is the most typical target loss in the ordinal classification. The bounds of the Bayes risk via the expectation of the partial loss have been studied in [Cour et al., 2011] but only for the 0/1-loss which is much less suitable for ordinal classification. It worth mentioning that the ordinal classification model allows for a tight convex approximations of the partial loss in contrast to previously considered classification models which often require hard to optimize non-convex surrogates [Do and Artières, 2009, Lou and Hamprecht, 2012, Jie and Orabona, 2010].

The paper is organized as follows. Formulation of the learning problem and its solution via minimization of the interval insensitive loss is presented in section 2. Algorithms approximating minimization of the interval-insensitive loss by convex optimization problems are proposed in section 3. A cutting plane based method solving the convex programs is described in section 4. Section 5 presents experimental and section 6 concludes the paper.

## 2 Learning ordinal classifier from weakly annotated examples

### 2.1 Learning from completely annotated examples

Let $\mathcal{X} \subset \mathbb{R}^n$ be a space of input observations and $\mathcal{Y} = \{1, \dots, Y\}$ a set of hidden labels endowed with a natural order. We consider learning of an ordinal classifier $h\colon \mathcal{X} \to \mathcal{Y}$ of the form

$$h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}) = 1 + \sum_{k=1}^{Y-1} [\![ \langle \boldsymbol{x}, \boldsymbol{w} \rangle > \theta_k ]\!] \tag{1}$$

where $\boldsymbol{w} \in \mathbb{R}^n$ and $\boldsymbol{\theta} \in \Theta = \{\boldsymbol{\theta}' \in \mathbb{R}^{Y-1} \mid \theta'_y \leq \theta'_{y+1},\ y = 1, \dots, Y-1\}$ are admisible parameters. The operator $[\![A]\!]$ evaluates to 1 if $A$ holds, otherwise it is 0. The classifier (1) splits the real line of projections $\langle \boldsymbol{x}, \boldsymbol{w} \rangle$ into $Y$ consecutive intervals defined by thresholds $\theta_1 \leq \theta_2 \leq \dots \leq \theta_{Y-1}$. The observation $\boldsymbol{x}$ is assigned a label correspoding to the interval to which the projection $\langle \boldsymbol{w}, \boldsymbol{x} \rangle$ falls to. The classifier (1) is a sutiable model if the label can be thought of as a rough measurement of a continuous random variable $\xi(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{w} \rangle + \text{noise}$ [McCullagh, 1980]. An example of the ordinal classifier applied to a toy 2D problem is depicted in Figure 1.

There exist several discriminative methods for learning parameters $(\boldsymbol{w}, \boldsymbol{\theta})$ of the classifier (1) from examples, e.g. [Crammer and Singer, 2001, Shashua and Levin, 2002, Chu and Keerthi, 2005, Li and Lin, 2006]. To our best knowledge, all the existing methods are fully supervised algorithms which require a set of completely

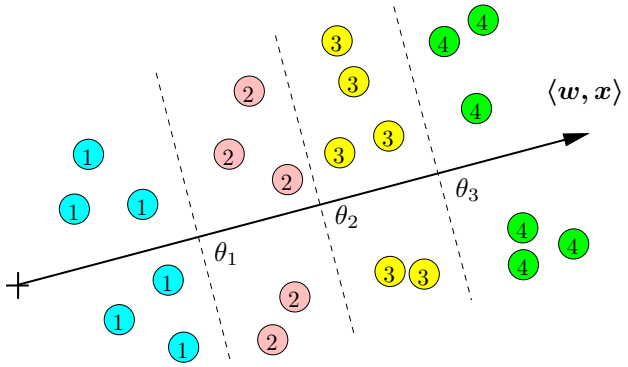



Fig. 1: The figure vizualizes division of the 2-dimensional feature space into four classes realized by an instance of the ordinal classifier (1).

annotated training examples

$$\{(\boldsymbol{x}^1, y^1), \ldots, (\boldsymbol{x}^m, y^m)\} \in (\mathcal{X} \times \mathcal{Y})^m \tag{2}$$

typically assumed to be drawn from i.i.d. random variables with some unknown distribution $p(\boldsymbol{x}, y)$. The goal of the supervised learning algorithm is formulated as follows. Given a loss function $\Delta\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and the training examples (2), the task is to learn the ordinal classifier $h$ whose *Bayes risk*

$$R(h) = \mathbb{E}_{(\boldsymbol{x},y)\sim p(\boldsymbol{x},y)} \Delta(y, h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta})) \tag{3}$$

is as small as possible. The loss functions most commonly used in practice are the Mean Absolute Error (MAE) $\Delta(y, y') = |y - y'|$ and the 0/1-loss $\Delta(y, y') = [\![y \neq y']\!]$. The MAE and 0/1-loss are examples of so called V-shaped losses.

**Definition 1** (V-shaped loss). A loss $\Delta\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is V-shaped if $\Delta(y, y) = 0$ and $\Delta(y'', y) \geq \Delta(y', y)$ holds for all triplets $(y, y', y'') \in \mathcal{Y}^3$ such that $|y'' - y'| \geq |y' - y|$.

That is, the value of a V-shaped loss grows monotonically with the distance between the predicted and the true label. In this paper we constrain our analysis to the V-shaped losses.

Because the expected risk $R(h)$ is not accessible directly due to the unknown distribution $p(\boldsymbol{x}, y)$, the discriminative methods like [Shashua and Levin, 2002, Chu and Keerthi, 2005, Li and Lin, 2006] minimize a convex surrogate of the empirical risk augmented by a quadratic regularizer. We follow the same framework but with novel surrogate loss functions suitable for learning from partially annotated examples.

### 2.2 Learning from partially annotated examples

Analogically to the supervised setting we assume that the observation $\boldsymbol{x} \in \mathcal{X}$ and the corresponding hidden label $y \in \mathcal{Y}$ are generated from some unknown distribution $p(\boldsymbol{x}, y)$. However, in contrast to the supervised setting the training set do not contain a single label for each instance. Instead, we assume that an annotator provided with the observation $\boldsymbol{x}$, and possibly with the label $y$, returns a partial annotation in the form of an interval of candidate labels $[y_l, y_r] \in \mathcal{P}$. The symbol $\mathcal{P} = \{[y_l, y_r] \in \mathcal{Y}^2 \mid y_l \leq y_r\}$ denotes a set of all possible partial annotations. The partial annotation $[y_l, y_r]$ means that the true label $y$ is from the interval $[y_l, y_r] = \{y \in \mathcal{Y} \mid y_l \leq y \leq y_r\}$. We shell assume that the annotator can be modeled by a stochastic process determined by a distribution $p(y_l, y_r \mid \boldsymbol{x}, y)$. That is, we are given a set of partially annotated examples

$$\{(\boldsymbol{x}^1, [y_l^1, y_r^1]), \ldots, (\boldsymbol{x}^m, [y_l^m, y_r^m])\} \in (\mathcal{X} \times \mathcal{P})^m \tag{4}$$

assumed to be generated from i.i.d. random variables with the distribution

$$p(\boldsymbol{x}, y_l, y_r) = \sum_{y \in \mathcal{Y}} p(y_l, y_r \mid \boldsymbol{x}, y)\, p(\boldsymbol{x}, y)$$

defined over $\mathcal{X} \times \mathcal{P}$. The learning algorithms described below do not require the knowledge of $p(\boldsymbol{x}, y)$ and $p(y_l, y_r \mid \boldsymbol{x}, y)$. However, it is clear that the annotation process given by $p(y_l, y_r \mid \boldsymbol{x}, y)$ cannot be arbitrary in order to make learning possible. For example, in the case when $p(y_l, y_r \mid \boldsymbol{x}, y) = p(y_l, y_r)$ the annotation would carry no information about the true label. Therefore we will later assume that the annotation is consistent in the sense that $y \notin [y_l, y_r]$ implies $p(y_l, y_r \mid \boldsymbol{x}, y) = 0$. The consistency of the annotation proces is a standard assumption used e.g. in [Cour et al., 2011].

The goal of learning from the partially annotated examples is formulated as follows. Given a (supervised) loss function $\Delta\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and partially annotated examples (4), the task is to learn the ordinal classifier (1) whose Bayes risk $R(h)$ defined by (3) is as small as possible. That is, the objective remains the same as in the supervised setting but the information about the labels containted in the training set is reduced to intervals.

### 2.3 Interval insensitive loss

We define an interval-insensitive loss function in order to measure discrepancy between the interval annotation $[y_l, y_r] \in \mathcal{P}$ and the predictions made by the classifier $h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}) \in \mathcal{Y}$.

**Definition 2** (Interval insensitive loss) Let $\Delta\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a supervised V-shaped loss. The interval insensitive loss $\Delta_I\colon \mathcal{P} \times \mathcal{Y} \to \mathbb{R}$ associated with $\Delta$ is defined as

$$\Delta_I(y_l, y_r, y) = \min_{y' \in [y_l, y_r]} \Delta(y', y) = \begin{cases} 0 \text{ if } y \in [y_l, y_r]\ , \\ \Delta(y, y_l) \text{ if } y \leq y_l\ , \\ \Delta(y, y_r) \text{ if } y \geq y_r\ . \end{cases} \tag{5}$$

The interval-insensitive loss $\Delta_I(y_l, y_r, y)$ does not penalize predictions which are in the interval $[y_l, y_r]$. Otherwise the penalty is either $\Delta(y, y_l)$ or $\Delta(y, y_r)$ depending which border of the interval $[y_l, y_r]$ is closer to the prediction $y$. In the special case of the Mean Absolute Error (MAE) $\Delta(y, y') = |y - y'|$, one can think of the associated interval-insensitive loss $\Delta_I(y_l, y_r, y)$ as the discrete counterpart of the $\epsilon$-insensitive loss used in the Support Vector Regression [Vapnik, 1998].

Having defined the interval-insensitive loss, we can approximate minimization of the Bayes risk $R(h)$ defined in (3) by minimization of the expectation of the interval-insensitive loss

$$R_I(h) = \mathbb{E}_{(\boldsymbol{x}, y_l, y_r) \sim p(\boldsymbol{x}, y_l, y_r)} \Delta_I(y_l, y_r, h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta})) \,. \tag{6}$$

In the sequel we denote $R_I(h)$ as the *partial risk*. The question is how well the partial risk $R_I(h)$ approximates the Bayes risk $R(h)$ being the target quantity to be minimized. In the rest of this section we first analyze this question for the 0/1-loss addopting results of [Cour et al., 2011] and then we present a novel bound for the MAE loss. In particular, we show that in both case the Bayes risk $R(h)$ can be upper bounded by a linear function of the partial risk $R_I(h)$.

In the sequel we will assume that the annotated process governed by the distribution $p(y_l, y_r \mid \boldsymbol{x}, y)$ is consistent in the following sense:

**Definition 3** (Consistent annotation process) Let $p(y_l, y_r \mid \boldsymbol{x}, y)$ be a properly defined distribution over $\mathcal{P}$ for any $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$. The annotation process governed by $p(y_l, y_r \mid \boldsymbol{x}, y)$ is consistent if any $y \in \mathcal{Y}$, $[y_l, y_r] \in \mathcal{P}$ such that $y \notin [y_l, y_r]$ implies $p(y_l, y_r \mid \boldsymbol{x}, y) = 0$.

The consistent annotation process guarantees that the true label is always contained among the candidate labels in the annotation.

We first apply the excess bound for the 0/1-loss function which has been studied in [Cour et al., 2011] for a generic partial annotations, i.e. when $\mathcal{P}$ is not necessarily the set of label intervals. The tightness of the resulting bound depends on the annotation process $p(y_l, y_r \mid \boldsymbol{x}, y)$ characterized by so called *ambiguity degree* $\varepsilon$ which, if addopted to our interval-setting, is defined as

$$\varepsilon = \max_{\boldsymbol{x}, y, z \neq y} p(z \in [y_l, y_r] \mid \boldsymbol{x}, y) = \max_{\boldsymbol{x}, y, z} \sum_{[y_l, y_r] \in \mathcal{P}} [\![ y_l \leq z \leq y_r ]\!] \, p(y_l, y_r \mid \boldsymbol{x}, y) \,. \tag{7}$$

In words, the ambiguity degree $\varepsilon$ is the maximum probability of an extra label $z$ co-occurring with the true label $y$ in the annotation interval $[y_l, y_r]$, over all labels and observations.

**Theorem 1** *Let $p(y_l, y_r \mid \boldsymbol{x}, y)$ be a distribution describing a consistent annotation process with the ambiguity degree $\varepsilon$ defined by (7). Let $R^{0/1}(h)$ be the Bayes risk (3) instantiated for the 0/1-loss and let $R_I^{0/1}(h)$ be the partial risk (6) instantiated for the interval insensitive loss associated to the 0/1-loss. Then the upper bound*

$$R^{0/1}(h) \leq \frac{1}{1 - \varepsilon} R_I^{0/1}(h)$$

*holds true for any* $h \in \mathcal{X} \to \mathcal{Y}$.

Theorem 1 is a direct application of Proposition 1 from [Cour et al., 2011].

Next we will introduce a novel upper bound for the MAE loss used in a majority of applications of the ordinal classifier. We again consider consistent annotation processes. We will characterize the annotation process by two numbers describing an amount of uncertainty in the training data. First, we use $\alpha \in [0, 1]$ to denoted the lower bound of the portion of exactly annotated examples, that is, examples annotated by an interval $[y_l, y_r]$, $y_l = y_r$, having just a single label. Second, we use $\beta \in \{0, \ldots, Y-1\}$ to denote the maximal uncertainty in annotation, that is, $\beta+1$ is the maximal width of the annotation interval which can appear in the training data with non-zero probability.

**Definition 4** ($\alpha\beta$-precise annotation process) Let $p(y_l, y_r \mid \boldsymbol{x}, y)$ be a properly defined distribution over $\mathcal{P}$ for any $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$. The annotation process governed by $p(y_l, y_r \mid \boldsymbol{x}, y)$ is $\alpha\beta$-precise if

$$\alpha \leq p(y, y \mid \boldsymbol{x}, y) \quad \text{and} \quad \beta \geq \max_{[y_l, y_r] \in \mathcal{P}} [\![ p(y_l, y_r \mid \boldsymbol{x}, y) > 0 ]\!] \, (y_r - y_l)$$

hold for any $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$.

To illustrate the meaning of the parameters $\alpha$ and $\beta$ let us consider the extreme cases. If $\alpha = 1$ of $\beta = 0$ then all examples are annotated exactly, that is, we are back in the standard supervised setting. If $\beta = Y - 1$ then in the worst case the annotation brings no information about the hidden label because the intervals contain the whole $\mathcal{Y}$. With the definition of $\alpha\beta$-precise annotation we can upper bound the Bayes risk in terms of the partial risk as follows:

**Theorem 2** *Let* $p(y_l, y_r \mid \boldsymbol{x}, y)$ *be a distribution describing a consistent* $\alpha\beta$*-precise annotation process. Let* $R^{MAE}(h)$ *be the Bayes risk (3) instantiated for the MAE-loss and let* $R_I^{MAE}(h)$ *be the partial risk (6) instantiated for the interval insensitive loss associated to the MAE-loss. Then the upper bound*

$$R^{MAE}(h) \leq R_I^{MAE}(h) + (1 - \alpha)\beta \tag{8}$$

*holds true for any* $h \in \mathcal{X} \to \mathcal{Y}$.

Proof of Theorem 2 is deferred to the appendix.

The bound (8) is obtained by the worst case analysis hence it may become trivial in some cases, for example, if all examples are annotated with large intervals because then $\alpha = 0$ and $\beta$ is large. The experimental study presented in section 5 nevertheless shows that the partial risk $R_I$ is a good proxy even for cases when the bound upper bound is big. This suggests that better bounds might be derive, for example, when additional information of $p(y_l, y_r \mid \boldsymbol{x}, y)$ is available.

In order to improve the performance of the resulting classifier via the bound (8) one needs to control the parameters $\alpha$ and $\beta$. A possible way which allows to set the parameters $(\alpha, \beta)$ exactly is to control the annotation process. For example, given a set of unannotated randomly drawn input samples $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\} \in \mathcal{X}^m$ we can proceed as follows:

1. We generate a vector of binary variables $\boldsymbol{\pi} \in \{0,1\}^m$ according to Bernoulli distribution with probability $\alpha$ that the variable is 1.
2. We instruct the annotator to provide just a single label for each input example with index from $\{i \in \{1,\ldots,m\} \mid \pi_i = 1\}$ while the remaining inputs (with $\pi_i = 0$) can be annotated by intervals but not larger than $\beta+1$ labels. That means that approximately $m \cdot \alpha$ inputs will be annotated exactly and $m \cdot (1-\alpha)$ with intervals.

This simple procedure ensures that the annotation process is $\alpha\beta$-precise though the distribution $p(y_l, y_r \mid \boldsymbol{x}, y)$ itself is unknown and depends on the annotator.

## 3 Algorithms

In the previous section we argue that the partial risk defined as an expectation of the interval insensitive loss is a reasonable proxy of the target Bayes risk. In this section we design algorithms learning the ordinal classifier via minimization of the regularized of the empirical risk which is known to be a reasonable proxy of the expected risk. Similarly to the supervised case we cannot minimize the empirical risk directly due to a discrete domain of the interval insensitive loss. For this reason we derive several convex surrogates which allow to translates the risk minimization to tractable convex problems.

We first show how to modify two state-of-the-art supervised methods in order to learn from partially annotated examples. Namely, we extend the Support Vector Ordinal Regression algorithm with explicit constraints (SVOR-EXP) and the Support Vector Ordinal Regression algorithm with implicit constraints (SVOR-IMC). The extended interval-insensitive variants are named II-SVOR-EXP and II-SVOR-IMC, respectively. The II-SVOR-EXP is a method minimizing a convex surrogate of the interval-insensitive loss associated to the 0/1-loss while the II-SVOR-IMC is designed for the MAE loss.

We also show how to construct a convex surrogate of the interval-insensitive loss associated to an arbitrary V-shaped loss. The generic surrogate is used to define the V-shaped interval insensitive loss minimization algorithm (VILMA). We also prove that the VILMA subsumes the II-SVOR-IMC (and the SVOR-IMC) as a special case.

### 3.1 Interval insensitive SVOR-EXP algorithm

The original SVOR-EXP algorithm [Chu and Keerthi, 2005] learns parameters of the ordinal classifier (1) from completely annotated examples $\{(\boldsymbol{x}^1, y^1), \ldots, (\boldsymbol{x}^m, y^m)\} \in (\mathbb{R}^n \times \mathcal{Y})^m$ by solving the following convex problem

$$(\boldsymbol{w}^*, \boldsymbol{\theta}^*) = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n, \boldsymbol{\theta}\in\Theta} \left[ \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{m} \ell^{\mathrm{EXP}}(\boldsymbol{x}^i, y^i, \boldsymbol{w}, \boldsymbol{\theta}) \right] \tag{9}$$

where

$$\ell^{\mathrm{EXP}}(\boldsymbol{x}, y, \boldsymbol{w}, \boldsymbol{\theta}) = \max(0, 1 - \langle \boldsymbol{x}, \boldsymbol{w} \rangle + \theta_{y^i}) + \max(0, 1 + \langle \boldsymbol{x}, \boldsymbol{w} \rangle - \theta_y)$$

and $\theta_0 = -\infty, \theta_Y = \infty$ are auxiliary constants used for notational convenience. In the original paper [Chu and Keerthi, 2005], the SVOR-EXP algorithm is formulated as an equivalent quadratic program which can be easily obtained from (9). We rather use the formulation (9) because it shows the optimized surrogate loss in its explicit form. The surrogate $\ell^{\mathrm{EXP}}(\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{x}, y)$ is a convex upper bound of the 0/1-loss

$$\Delta^{0/1}(y, h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta})) = [\![ y \neq h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}) ]\!] = [\![ \langle \boldsymbol{x}, \boldsymbol{w} \rangle < \theta_{y-1} ]\!] + [\![ \langle \boldsymbol{x}, \boldsymbol{w} \rangle \geq \theta_y ]\!] \,,$$

obtained by replacing the step functions $[\![ t \leq 0 ]\!]$ with the hinge-loss $\max(0, 1 - t)$.

We apply the same idea to approximate the interval insensitive loss $\Delta_I^{0/1}(y_l, y_r, y)$ associated with the 0/1-loss. According to the definition (5) we have

$$\begin{aligned} \Delta_I^{0/1}(y_l, y_r, h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta})) &= \min_{y' \in [y_l, y_r]} [\![ y' \neq h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}) ]\!] \\ &= [\![ \langle \boldsymbol{x}, \boldsymbol{w} \rangle < \theta_{y_l - 1} ]\!] + [\![ \langle \boldsymbol{x}, \boldsymbol{w} \rangle \geq \theta_{y_r} ]\!] \,. \end{aligned}$$

By replacing the step functions with the hinge-losses we get the surrogate

$$\ell_I^{\mathrm{EXP}}(\boldsymbol{x}, y_l, y_r, \boldsymbol{w}, \boldsymbol{\theta}) = \max(0, 1 - \langle \boldsymbol{x}, \boldsymbol{w} \rangle + \theta_{y_l - 1}) + \max(0, 1 + \langle \boldsymbol{x}, \boldsymbol{w} \rangle - \theta_{y_r})$$

which is a convex upper bound of $\Delta_I^{0/1}(y_l, y_r, h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}))$. For visualization see Figure 2.

We can modify the SVOR-EXP algorithm for learning from partially annotated examples by replacing the loss $\ell^{\mathrm{EXP}}(\boldsymbol{x}, y, \boldsymbol{w}, \boldsymbol{\theta})$ in the definition of (9) by its interval insensitive counterpart $\ell_I^{\mathrm{EXP}}(\boldsymbol{x}, y_l, y_r, \boldsymbol{w}, \boldsymbol{\theta})$. We denote the modified variant as the II-SVOR-EXP algorithm.

### 3.2 Interval insensitive SVOR-IMC algorithm

The original SVOR-IMC algorithm [Chu and Keerthi, 2005] learns parameters of the ordinal classifier (1) from completely annotated examples $\{(\boldsymbol{x}^1, y^1), \ldots, (\boldsymbol{x}^m, y^m)\} \in (\mathbb{R}^n \times \mathcal{Y})^m$ by solving the following convex optimization problem

$$(\boldsymbol{w}^*, \boldsymbol{\theta}^*) = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^n, \boldsymbol{\theta} \in \Theta} \left[ \frac{\lambda}{2} \|\boldsymbol{w}\|^2 + \sum_{i=1}^{m} \ell^{\mathrm{IMC}}(\boldsymbol{x}^i, y^i, \boldsymbol{w}, \boldsymbol{\theta}) \right] \tag{10}$$

where

$$\ell^{\mathrm{IMC}}(\boldsymbol{x}, y, \boldsymbol{w}, \boldsymbol{\theta}) = \sum_{y'=1}^{y-1} \max(0, 1 - \langle \boldsymbol{x}, \boldsymbol{w} \rangle + \theta_{y'-1}) + \sum_{y'=y}^{Y-1} \max(0, 1 + \langle \boldsymbol{x}, \boldsymbol{w} \rangle - \theta_{y'})$$

and using the convention $\sum_{i=m}^{n} a_i = 0$ if $m > n$. As in the previous case, the problem (10) is an equivalent reformulation of the quadratic program defining

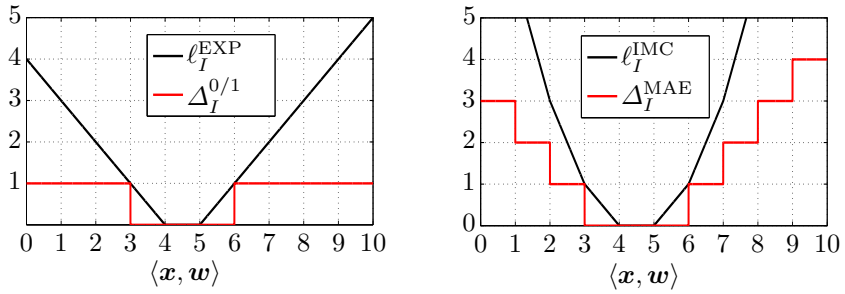



Fig. 2: The left figure shows the interval insensitive loss $\Delta_I^{0/1}(\boldsymbol{x}, y_l, y_r, h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}))$ associated with the 0/1-loss and its surrogate $\ell_I^{\mathrm{EXP}}(\boldsymbol{x}, y_l, y_r, \boldsymbol{w}, \boldsymbol{\theta}))$. The right figure shows the interval insensitive loss $\Delta_I^{\mathrm{MAE}}(\boldsymbol{x}, y_l, y_r, h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}))$ associated with the MAE and its surrogate $\ell_I^{\mathrm{IMC}}(\boldsymbol{x}, y_l, y_r, \boldsymbol{w}, \boldsymbol{\theta}))$. The value of the losses is shown as a function of the dot product $\langle \boldsymbol{x}, \boldsymbol{w} \rangle$ for $\theta_1 = 1, \theta_2 = 2, \ldots, \theta_{Y-1} = Y - 1$ and $y_l = 4$, $y_r = 6$.

the SVOR-IMC algorithm in [Chu and Keerthi, 2005]. It is seen that the surrogate $\ell^{\mathrm{IMC}}(\boldsymbol{x}, y, \boldsymbol{w}, \boldsymbol{\theta})$ is a convex upper bound of the MAE loss which can be written as

$$\Delta^{\mathrm{MAE}}(y, h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta})) = |y - h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta})| = \sum_{y'=1}^{y-1} [\![ \langle \boldsymbol{x}, \boldsymbol{w} \rangle < \theta_y ]\!] + \sum_{y'=y}^{Y-1} [\![ \langle \boldsymbol{x}, \boldsymbol{w} \rangle \geq \theta_y ]\!].$$

The surrogate is obtained by replacing the step functions in the sums with the hinge loss. Analogically, we can derive a convex surrogate of the interval insensitive loss associated with the MAE. By definition (5), the interval insensitive loss associated with MAE reads

$$\begin{aligned} \Delta_I^{\mathrm{MAE}}(y_l, y_r, h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta})) &= \min_{y' \in [y_l, y_r]} |y' - h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta})| \\ &= \sum_{y'=1}^{y_l - 1} [\![ \langle \boldsymbol{x}, \boldsymbol{w} \rangle < \theta_y ]\!] + \sum_{y'=y_r}^{Y-1} [\![ \langle \boldsymbol{x}, \boldsymbol{w} \rangle \geq \theta_y ]\!] \,. \end{aligned}$$

Replacing the step functions by the hinge loss we obtain a convex surrogate

$$\ell_I^{\mathrm{IMC}}(\boldsymbol{x}, y_l, y_r, \boldsymbol{w}, \boldsymbol{\theta}) = \sum_{y'=1}^{y_l - 1} \max(0, 1 - \langle \boldsymbol{x}, \boldsymbol{w} \rangle + \theta_{y'-1}) + \sum_{y'=y_r}^{Y-1} \max(0, 1 + \langle \boldsymbol{x}, \boldsymbol{w} \rangle - \theta_y),$$

which is obviously an upper bound of $\Delta_I^{\mathrm{MAE}}(y_l, y_r, h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}))$. See Figure 2 for visualization.

Given the partially annotated examples $\{(\boldsymbol{x}^1, [y_l^1, y_r^1]), \ldots, (\boldsymbol{x}^m, [y_l^m, y_r^m])\} \in (\mathcal{X} \times \mathcal{P})^m$, we can learn parameters of the ordinal classifier (1) by solving (10)

with $\ell_I^{\mathrm{IMC}}(\boldsymbol{x}, y_l, y_r, \boldsymbol{w}, \boldsymbol{\theta})$ substituted for $\ell^{\mathrm{IMC}}(\boldsymbol{x}, y, \boldsymbol{w}, \boldsymbol{\theta})$. We denote the variant optimizing $\ell_I^{\mathrm{IMC}}(\boldsymbol{x}, y_l, y_r, \boldsymbol{w}, \boldsymbol{\theta})$ as the II-SVOR-IMC algrithm. Due to the equality $\ell_I^{\mathrm{IMC}}(\boldsymbol{x}, y, y, \boldsymbol{w}, \boldsymbol{\theta}) = \ell^{\mathrm{IMC}}(\boldsymbol{x}, y, \boldsymbol{w}, \boldsymbol{\theta})$ it is clear that the proposed II-SVOR-IMC subsumes the original SVOR-IMC as a special case.

### 3.3 VILMA: V-shaped interval insensitive loss minimization algorithm

In this section we propose a generic method for learning the ordinal classifiers with arbitrary interval insensitive V-shaped loss. We start by introducing an equivalent parametrization of the ordinal classifier (1) originally proposed in [Antoniuk et al., 2013]. The ordinal classifier (1) can be re-parametrized as a multi-class linear classifier, in the sequel denoted as multi-class ordinal (MORD) classifier, which reads

$$h'(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{b}) = \operatorname*{argmax}_{y \in \mathcal{Y}} \Big( \langle \boldsymbol{x}, \boldsymbol{w} \rangle \cdot y + b_y \Big) \tag{11}$$

where $\boldsymbol{w} \in \mathbb{R}^n$ and $\boldsymbol{b} = (b_1, \ldots, b_Y) \in \mathbb{R}^Y$ are parameters. Note that the MORD classifier has $n + Y$ parameters and any pair $(\boldsymbol{w}, \boldsymbol{b}) \in (\mathbb{R}^n \times \mathbb{R}^Y)$ is admissible. In contrast, the original ordinal classifier (1) has $n+Y-1$, however, the admissible parameters must satisfy $(\boldsymbol{w}, \boldsymbol{\theta}) \in (\mathbb{R}^n \times \Theta)$. The following theorem states that both parametrizations are equivalent.

**Theorem 3** *The ordinal classifier (1) and the MORD classifier (11) are equivalent in the following sense. For any $\boldsymbol{w} \in \mathbb{R}^n$ and admissible $\boldsymbol{\theta} \in \Theta$ there exists $\boldsymbol{b} \in \mathbb{R}^Y$ such that $h(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\theta}) = h'(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{b})$, $\forall \boldsymbol{x} \in \mathbb{R}^n$. For any $\boldsymbol{w} \in \mathbb{R}^n$ and $\boldsymbol{b} \in \mathbb{R}^n$, there exists admissible $\boldsymbol{\theta} \in \Theta$ such that $h(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\theta}) = h'(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{b})$, $\forall \boldsymbol{x} \in \mathbb{R}^n$.*

Proof of Theorem 3 is given in [Antoniuk et al., 2013] as well as conversion formulas between the two parametrizations.

The MORD parametrization allows to adopt techniques known for linear classifiers. Namely, we can replace the interval insensitive loss by a convex surrogate similar to the margin-rescaling loss known from the structured output leaning [Tsochantaridis et al., 2005]. Given a V-shaped supervised loss $\Delta\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, we propose to approximate the value of the associated interval insensitive loss $\Delta_I(y_l, y_r, h'(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{b}))$ by a surrogate loss $\ell_I\colon \mathcal{X} \times \mathcal{P} \times \mathbb{R}^n \times \mathbb{R}^Y \to \mathbb{R}$

$$\begin{aligned} \ell_I(\boldsymbol{x}, y_l, y_r, \boldsymbol{w}, \boldsymbol{b}) = &\max_{y \le y_l} \Big[ \Delta(y, y_l) + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_l) + b_y - b_{y_l} \Big] \\ &+ \max_{y \ge y_r} \Big[ \Delta(y, y_r) + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_r) + b_y - b_{y_r} \Big] . \end{aligned} \tag{12}$$

It is seen that for fixed $(\boldsymbol{x}, y_l, y_r)$ the function $\ell_I(\boldsymbol{x}, y_l, y_r, \boldsymbol{w}, \boldsymbol{b})$ is a sum of two point-wise maxima over linear functions hence it is convex in the parameters $(\boldsymbol{w}, \boldsymbol{b})$. The following proposition states that the surrogate is also an upper bound of the interval insensitive loss.

**Proposition 1** *For any* $\boldsymbol{x} \in \mathbb{R}^n$, $[y_l, y_r] \in \mathcal{P}$, $\boldsymbol{w} \in \mathbb{R}^n$ *and* $\boldsymbol{b} \in \mathbb{R}^Y$ *the inequality*

$$\Delta_I(y_l, y_r, h'(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{b})) \leq \ell_I(\boldsymbol{x}, y_l, y_r, \boldsymbol{w}, \boldsymbol{b})$$

*holds where* $h'(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{b})$ *denotes response of the MORD classifier (11).*

Proof of Proposition 1 is deferred to the appendix.

As an example let us consider the surrogate (12) instantiated for the MAE, then

$$\begin{aligned}\ell_I^{\mathrm{MAE}}(\boldsymbol{x}, y_l, y_r, \boldsymbol{w}, \boldsymbol{b}) = &\max_{y \leq y_l} \left[ y_l - y + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_l) + b_y - b_{y_l} \right] \\ &+ \max_{y \geq y_r} \left[ y - y_r + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_r) + b_y - b_{y_r} \right].\end{aligned} \tag{13}$$

Given partially annotated training examples $\{(\boldsymbol{x}^1, [y_l^1, y_r^1]), \ldots, (\boldsymbol{x}^m, [y_l^m, y_r^m])\} \in (\mathcal{X} \times \mathcal{P})^m$, we can learn parameters $(\boldsymbol{w}, \boldsymbol{b})$ of the MORD classifier (11) by solving the following unconstrained convex problem

$$(\boldsymbol{w}^*, \boldsymbol{b}^*) = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^n, \boldsymbol{b} \in \mathbb{R}^Y} \left[ \frac{\lambda}{2} \|\boldsymbol{w}\|^2 + \frac{1}{m} \sum_{i=1}^{m} \ell_I(\boldsymbol{x}^i, y_l^i, y_r^i, \boldsymbol{w}, \boldsymbol{b}) \right] \tag{14}$$

where $\lambda \in \mathbb{R}_{++}$ is a regularization constant. In the sequal we denote the method based on solving (14) as the V-shape Interval insensitive Loss Minimization Algorithm (VILMA).

It is interesting to compare the VILMA instantiated for the MAE with the II-SVOR-IMC algorithm which optimizes a different surrogate of the same loss. Note that the II-SVOR-IMC learns the parameters $(\boldsymbol{w}, \boldsymbol{\theta})$ of the ordinal classifier (1) while the VILMA parameters $(\boldsymbol{w}, \boldsymbol{b})$ of the MORD rule (11). The following proposition states that the surrogate losses of the algorithms are equivalent.

**Proposition 2** *Let* $\boldsymbol{w} \in \mathbb{R}^n, \boldsymbol{\theta} \in \Theta, \boldsymbol{b} \in \mathbb{R}^Y$ *be a triplet of vectors such that* $h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}) = h'(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{b})$ *holds for all* $\boldsymbol{x} \in \mathcal{X}$ *where* $h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta})$ *denotes the ordinal classifier (1) and* $h'(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{b})$ *the MORD classifier (11). Then the equality*

$$\ell_I^{\mathrm{IMC}}(\boldsymbol{x}, y_l, y_r, \boldsymbol{w}, \boldsymbol{\theta}) = \ell_I^{\mathrm{MAE}}(\boldsymbol{x}, y_l, y_r, \boldsymbol{w}, \boldsymbol{b})$$

*holds true for any* $\boldsymbol{x} \in \mathcal{X}$ *and* $[y_l, y_r] \in \mathcal{P}$.

Proof is given in the appendix.

The corollary of Proposition 2 is that the II-SVOR-IMC and the VILAM with MAE loss return the same classification rules although differently parametrized. To sum up, the VILMA has the following properties:

- It is applicable for arbitrary V-shaped loss.
- It subsumes the II-SVOR-IMC and the original SVOR-IMC as special cases.
- It converts learning into an unconstrained convex optimization in contrast to the II-SVOR-EXP and the II-SVOR-IMC which maintain the constraints $\boldsymbol{\theta} \in \Theta$.

## 4 Double-loop Cutting Plane Solver

The proposed method VILMA translates learning into a convex optimization problem (14) that can be re-written as

$$(\boldsymbol{w}^*, \boldsymbol{b}^*) = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n, \boldsymbol{b}\in\mathbb{R}^Y} \left[\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + R_{\text{emp}}(\boldsymbol{w}, \boldsymbol{b})\right] \tag{15}$$

where $R_{\text{emp}}(\boldsymbol{w}, \boldsymbol{b}) = \frac{1}{m}\sum_{i=1}^{m} \ell_I(\boldsymbol{x}^i, y^i, \boldsymbol{w}, \boldsymbol{b})$ is non-differentiable convex function w.r.t. variables $\boldsymbol{w}$ and $\boldsymbol{b}$. Thanks to the form $\ell_I(\boldsymbol{x}^i, y^i, \boldsymbol{w}, \boldsymbol{b})$ defined in (12) the task (15) can be reformulated as a quadratic program with $n + m + Y$ variables and $Y \cdot m$ constraints. The size of the QP rules out the off-the-shelf optimization methods. The common strategy in machine learning is to solve the problem approximately, for example, by the stochastic gradient methods. The stochastic methods are applicable on large problems but they require fine tuning of free parameters of the solver. Another method frequently used for the convex optimization is the cutting plane algorithm (CPA) [Teo et al., 2010, Franc et al., 2012]. The CPA provides a certificate of the optimality and has no free parameters to be tuned. The CPA solves efficiently

$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} F(\boldsymbol{w}) \quad \text{where} \quad F(\boldsymbol{w}) = \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + G(\boldsymbol{w}) \tag{16}$$

and $G\colon \mathbb{R}^n \to \mathbb{R}$ is a convex function. In contrast to (15), the objective of (16) contains a quadratic regularization of all variables. It is well known that the CPA applied directly to the un-regularized problem like (15) exhibits a strong zig-zag behavior leading to a large number of iterations. An ad-hod solution would be to impose an additional regularization on $\boldsymbol{b}$ which, however, can significantly spoil the results as demonstrated in section 5.2. In the rest of this section we first outline the CPA algorithm for the problem (16) and then show how it can be used to solve the problem (15).

The problem (16) can be approximated by its *reduced problem*

$$\boldsymbol{w}_t \in \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} F_t(\boldsymbol{w}) \quad \text{where} \quad F_t(\boldsymbol{w}) = \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + G_t(\boldsymbol{w})\,. \tag{17}$$

The reduced problem (17) is obtained from (16) by substituting a cutting-plane model $G_t(\boldsymbol{w})$ for the convex function $G(\boldsymbol{w})$ while the regularizer remains unchanged. The cutting plane model reads

$$G_t(\boldsymbol{w}) = \max_{i=0,\ldots,t-1} \left[G(\boldsymbol{w}_i) + \langle G'(\boldsymbol{w}_i), \boldsymbol{w} - \boldsymbol{w}_i\rangle\right] \tag{18}$$

where $G'(\boldsymbol{w}) \in \mathbb{R}^n$ is a sub-gradient of $G$ at point $\boldsymbol{w}$. Thanks to the convexity of $G(\boldsymbol{w})$, the objective $F_t(\boldsymbol{w})$ of the reduced problem is a piece-wise linear underestimator of $F(\boldsymbol{w})$. The CPA is outlined in Algorithm 1. Starting from $\boldsymbol{w}_0 \in \mathbb{R}^n$, the CPA computes a new iterate $\boldsymbol{w}_t$ by solving the reduced problem (17). In each iteration $t$, the cutting-plane model (18) is updated by a new

**Algorithm 1:** Cutting Plane Algorithm

**Input**: $\varepsilon > 0$, $\boldsymbol{w}_0 \in \mathbb{R}^n$, $t \leftarrow 0$
**Output**: vector $\boldsymbol{w}_t$ being $\varepsilon$-precise solution of (16)
**repeat**
  $t \leftarrow t + 1$
  Compute $G(\boldsymbol{w}_{t-1})$ and $G'(\boldsymbol{w}_{t-1})$
  Update the model $G_t(\boldsymbol{w}) \leftarrow \max_{i=0,\ldots,t-1} G(\boldsymbol{w}_i) + \langle G'(\boldsymbol{w}_i, \boldsymbol{w} - \boldsymbol{w}_i \rangle$
  Solve the reduced problem $\boldsymbol{w}_t \leftarrow \operatorname{argmin}_{\boldsymbol{w}} F_t(\boldsymbol{w})$ where
  $F_t(\boldsymbol{w}) = \lambda \Omega(\boldsymbol{w}) + R_t(\boldsymbol{w})$
**until** $F(\boldsymbol{w}_t) - F_t(\boldsymbol{w}_t) \leq \varepsilon$;

cutting plane computed at the intermediate solution $\boldsymbol{w}_t$ leading to a progressively tighter approximation of $F(\boldsymbol{w})$. The CPA halts if the gap between $F(\boldsymbol{w}_t)$ and $F_t(\boldsymbol{w}_t)$ gets below a prescribed $\varepsilon > 0$, meaning that $F(\boldsymbol{w}_t) \leq F(\boldsymbol{w}^*) + \varepsilon$. The CPA halts after $\mathcal{O}(\frac{1}{\lambda\varepsilon})$ iterations at most [Teo et al., 2010].

We can convert (15) to (16) by setting

$$G(\boldsymbol{w}) = R_{\text{emp}}(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{w})) \quad \text{where} \quad \boldsymbol{b}(\boldsymbol{w}) = \operatorname*{argmin}_{\boldsymbol{b} \in \mathbb{R}^Y} R_{\text{emp}}(\boldsymbol{w}, \boldsymbol{b}) \,. \tag{19}$$

It is clear that if $\boldsymbol{w}^*$ is a solution of (16) with $G(\boldsymbol{w})$ defined by (19) then $(\boldsymbol{w}^*, \boldsymbol{b}(\boldsymbol{w}^*))$ is a solution of (15). Because $R_{\text{emp}}(\boldsymbol{w}, \boldsymbol{b})$ is jointly convex in $\boldsymbol{w}$ and $\boldsymbol{b}$ the function $G$ in (19) is convex in $\boldsymbol{w}$ (see e.g. [Boyd and Vandenberghe, 2004]). Hence we can apply the CPA to solve (18) preserving all its convergence guarantees. To this end, we have to provide a first-order oracle computing $G(\boldsymbol{w})$ and the sub-gradient $G'(\boldsymbol{w})$. Knowning $\boldsymbol{b}(\boldsymbol{w})$ the subgradient can be computed as

$$G'(\boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{x}^i (\hat{y}_l^i + \hat{y}_r^i - y_l^i - y_r^i) \tag{20}$$

where

$$\hat{y}_l^i = \operatorname*{argmax}_{y \leq y_l^i} \left[ \Delta(y, y_l^i) + \langle \boldsymbol{w}, \boldsymbol{x}^i \rangle y - b_y(\boldsymbol{w}) \right] ,$$
$$\hat{y}_r^i = \operatorname*{argmax}_{y \geq y_r^i} \left[ \Delta(y, y_r^i) + \langle \boldsymbol{w}, \boldsymbol{x}^i \rangle y - b_y(\boldsymbol{w}) \right] .$$

To sum up, the CPA transforms solving (15) to a sequence of two simpler problems:

1. The reduced problem (17) which is a quadratic program that can be solved efficiently via its dual formulation [Teo et al., 2010]. The dual QP has only $t$ variables where $t$ is the number of iterations of the CPA. Since the CPA rarely needs more than a few hundred iterations the off-the-shelf QP solvers can be used.
2. The problem (19) providing $\boldsymbol{b}(\boldsymbol{w})$ needed to compute $G(\boldsymbol{w}) = R_{\text{emp}}(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{w}))$ and the sub-gradient $G'(\boldsymbol{w})$ via (20). The problem (19) has only $Y$ (the number of labels) variables. Hence it can be approached by off-the-shelf convex solver like the Analytic Center Cutting Plane algorithm [Gondzio et al., 1996].

Because we use another cutting plane method in the inner loop to implement the first-order oracle, we call the proposed solver as the *double-loop CPA*.

Finally, we point out that the convex problems associated with the II-SVOR-EXP and the II-SVOR-IMC can be solved by a similar method. The only change is in using additional constraints $\boldsymbol{\theta} \in \Theta$ in (15) which propagate to the problem (19).

## 5 Experiments

We evaluate the proposed methods on a real-life computer vision problem of estimating age of a person from his/her facial image. The age estimation is a prototypical problem calling for the ordinal classification and for learning from the interval annotations. The set of labels corresponds to individual ages which form an ordered set. Training examples of the facial images are cheap, for example, they can be downloaded from the Internet. Obtaining the ground truth age for the facial images is often very complicated for obvious reasons. The typical solution used in practice is to endow the collected images with age estimated manually by a human annotator. Annotating large image databases with year-precise estimate of the age is not only tedious but often results in inconsistent estimates. On the other hand, providing the interval annotations is more natural for humans and clearly such annotation will be more consistent.

The experiments have two parts. First, in section 5.2, we present results on the precisely annotated examples. By conducting these experiments we i) set a baseline for the later experiments on the partially annotated examples, ii) numerically verify that the VILMA subsumes the SVOR-IMC algorithm as a special case and iii) justify usage of the proposed double-loop CPA. Second, in section 5.3 we thoroughly analyze the performance of the VILMA on the partially annotated examples. We emphasize that all tested algorithms are designed to optimize the MAE loss which is the standard accuracy measure for the age estimation systems.

### 5.1 Databases and implementation details

We use two large face databases with year-precise annotation of the age:

1. MORPH database [Ricanek and Tesafaye, 2006] is the standard benchmark for age estimation. It contains 55,134 face images with exact age annotation ranging from 16 to 77 years. Because the age category 70+ is severely under-represented (only 9 examples in total) we removed faces with age higher than 70. The database contains frontal police mugshots taken under controlled conditions. The images have resolution 200×240 pixels and are mostly of very good quality.
2. WILD database is a collection of there public databases: Labeled Faces in the Wild [Huang et al., 2007], PubFig [Kumar et al., 2009] and PAL [Minear and Park, 2004]. The images are annotated by several independent persons. We selected a

subset of near-frontal images (yaw angle in $[-30°, 30°]$) containing 34,259 faces in total with the age from 1 to 80 years. The WILD database contains challenging "in-the-wild" images exhibiting a large variation in the resolution, illumination changes, race and background clutter.

The faces were split randomly three times into training, validation and testing part in the ratio 60/20/20. We made sure that images of the same identity never appear in different parts simultaneously.

*Preprocessing .* The feature representation of the facial images were computed as follow. We first localized the faces by a commercial face detector [1] and consequently applied a Deformable Part Model based detector [Uřičář et al., 2012] to find facial landmarks like the corners of eyes, mouth and tip of the nose. The found landmarks were used to transform the input face by an affine transform into its canonical pose. Finally, the canonical face of size $60 \times 40$ pixels was described by multi-scale LBP descriptor [Sonnenburg and Franc, 2010] resulting in $n = 159,488$-dimensional binary sparse vector serving as an input of the ordinal classifier.

*Implementation of the solver.* We implemented the double-loop CPA and the standard CPA in C++ by modifying the code from the Shogun machine learning library [Sonnenburg et al., 2010]. To solve internal problem (19) we used the Oracle Based Optimization Engine (OBOE) implementation of the Analytic Center Cutting Plane algorithm implemented as part of COmputational INfrastructure for Operations Research project (COIN-OR) [Gondzio et al., 1996].

### 5.2 Supervised setting

The purpose of experiments in this section is three fold. First, to present the results for the standard supervised setting which is later used as a baseline. Second, to numerically verify Proposition 2 which states that the VILMA subsumes the SVOR-IMC as a special case if instantiated for the MAE loss. Third, to show that imposing an extra quadratic regularization on the biases $\boldsymbol{b}$ of the MORD rule (11) severly harms the results which justifies using the proposed double-loop CPA.

We used images with the year-precise age annotations from the MORPH database. We constructed a sequence of training sets with the number of examples $m$ varying from $m = 3,300$ to $m = 33,000$ (the total number of training example in the MORPH). For each training set we learned the ordinal classifier with the regularization parameters set to $\lambda \in \{1, 0.1, 0.01, 0.001\}$. The classifier corresponding to $\lambda$ with the smallest validation error was applied on the testing examples. This process was applied for three random splits. We report the averages and the standard deviations of the MAE computed

[1] Courtesy of Eydea Recognition Ltd, www.eyedea.cz

on the test examples over the three splits. The same evaluation procedure was applied for the three algorithms: i) the proposed method VILMA, ii) the standard SVOR-IMC and iii) the VILMA-REG which learns by solving (14) but using the regularization $\frac{\lambda}{2}(\|\boldsymbol{w}|^2 + \|\boldsymbol{b}\|^2)$ instead of $\frac{\lambda}{2}\|\boldsymbol{w}\|^2$. We used the double-loop CPA for the VILMA and the SVOR-IMC and the standard CPA for the VILMA-REG. Table 1 summarizes the results.

We observe that the prediction error steeply decreases with adding new precisely annotated examples. The MAE for the largest training set is $4.55 \pm 0.02$ which closely matches the state-of-the-art methods like [Guo and Mu, 2010] reporting MAE 4.45 on the same database. The next section shows that similar results can be obtained with cheaper partially annotated examples.

Although the VILMA and the SVOR-IMC learn different parametrizations of the ordinal classifier the resulting rules are equivalent up a numerical error as predicted by Proposition 2. We did the same experiment applying the VILMA and the II-SVOR-IMC on the partially annotated examples as described in the next section. The results of both methods were the same up a numerical error. Therefore in the next section we include the results only for the VILMA.

The test MAE of the classifier learned by the VILMA-REG is almost doubled compared to the classifier learned by VILMA via the double-loop CPA. This shows that pushing the biases $\boldsymbol{b}$ towards zero by the quadratic regularizer which is necessary in the standard CPA has a detrimental effect on the accuracy.

| | $m = 3300$ | $m = 6600$ | $m = 13000$ | $m = 23000$ | $m = 33000$ |
|---|---|---|---|---|---|
| VILMA | $5.56 \pm 0.02$ | $5.12 \pm 0.02$ | $4.83 \pm 0.02$ | $4.66 \pm 0.01$ | $4.55 \pm 0.02$ |
| SVOR-IMC | $5.56 \pm 0.03$ | $5.14 \pm 0.02$ | $4.83 \pm 0.01$ | $4.68 \pm 0.03$ | $4.54 \pm 0.01$ |
| VILMA-REG | $9.57 \pm 0.03$ | $9.21 \pm 0.06$ | $9.07 \pm 0.05$ | $9.04 \pm 0.05$ | $9.06 \pm 0.02$ |

Table 1: The test MAE of the ordinal classifier learned from the precisely annotated examples by the VILMA, the standard SVOR-IMC and the VILMA-REG using the $\frac{\lambda}{2}(\|\boldsymbol{w}\|^2 + \|\boldsymbol{b}\|)$ regularizer. The results are shown for the training sets generated from the MORPH database by randomly selecting different number of examples $m$.

### 5.3 Learning from partially annotated examples

The goal is to evaluate the VILMA algorithm when it is applied to learning from partially annotated examples. The MORPH and WILD contain the year-precise annotation of the age which is necessary to make the comparison with supervised methods. We generated the partial annotation in a way which simulates a practical setting:

– $m_P$ randomly selected examples were annotated precisely.

– $m_I$ randomly selected examples were annotated by intervals. The admissible intervals were chosen to partition the set of ages and to have the same width (up to border cases). The interval width was varied from $u \in \{5, 10, 20\}$. The interval annotation was obtained by rounding the true age to the admissible intervals. For example, in case of $(u = 5)$-years wide intervals the true ages $y \in \{1, 2, \ldots, 5\}$ were transformed to the interval annotation $[1, 5]$, the ages $y \in \{6, 7, \ldots, 10\}$ to $[6, 10]$ and so on.

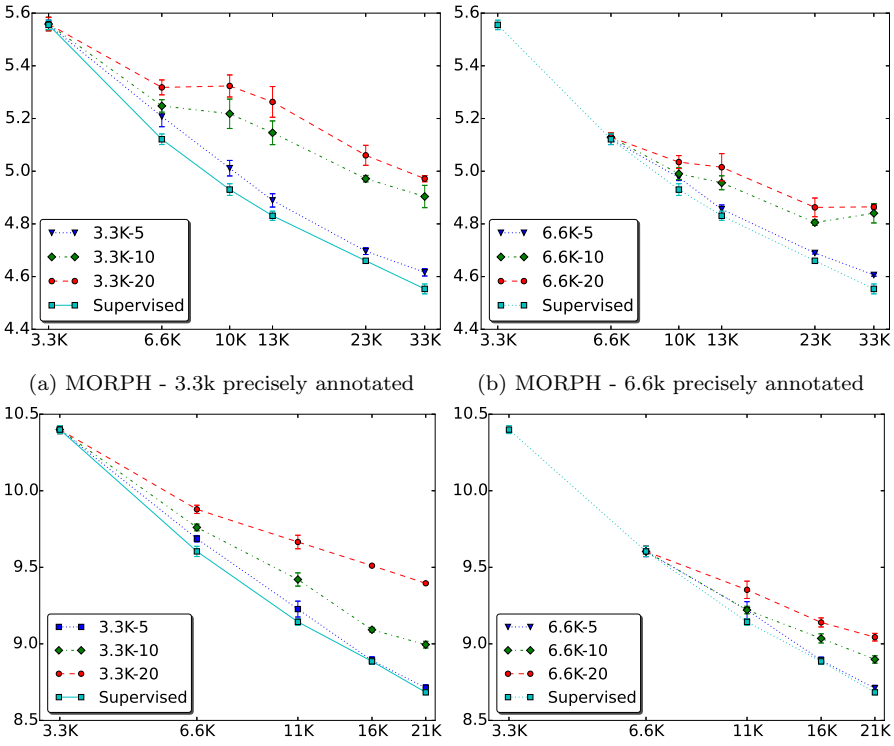
The described annotation process is approximately $\alpha\beta$-precise with $\alpha = \frac{m_P}{m_P+m_I}$ and $\beta = u - 1 \in \{4, 9, 19\}$. We varied $m_P \in \{3300, 6600\}$ and $m_I$ from 0 to $m_{\text{total}} - m_P$ where $m_{\text{total}}$ is the total number of the training examples .

For each training set we run the VILMA with the regularization constant $\lambda$ set to $\{1, 0.1, 0.01, 0.001\}$ and selected the best value according to the MAE error computed on the validation examples. The best model was then evaluated on the test part. This process was repeated for the three random splits. The reported errors are the averages and the standard deviations of the MAE computed on the test examples. The results are summarized in Figure 3 and Table 2.

We observe that adding the partially annotated examples monotonically improves the accuracy. This observation holds true for all tested combinations of $m_I$, $m_P$, $u$ and both databases. This observation is of great practical importance. It suggests that adding cheap partially annotated examples only improves and never worsens the accuracy of the ordinal classifier.

It is seen that the improvement caused by adding the partially annotated examples can be substantial. Not surprisingly the best results are obtained for the annotation with the 5-years wide intervals. In this case, the performance of the classifier learned from partial annotations closely matches the supervised setting. In particular, the loss in accuracy observed for the WILD database is on the level of the standard deviation. Even in the most challenging case, the 20-years wide intervals, the results are practically useful. For example, to get classifier with $\approx 9$ MAE on the WILD database one can either learn from $\approx 12,000$ precisely annotated examples or instead from $6,600$ precisely annotated plus $14,400$ partially annotated with 20-years wide intervals.

Let $\gamma(\alpha, \beta) = \hat{R}^{MAE}(h^{\alpha,\beta}) - \hat{R}^{MAE}(h^*)$ be the loss in the test accuracy caused by learning from the partially annotated examples generated by $\alpha\beta$-precise annotation process instead of learning by the supervised algorithm. The values of $\gamma(\alpha, \beta)$ are shown in Figure 4. We see that the loss in accuracy grows proportionally with the interval width $u = 1 + \beta$ and with the portion of partially annotated examples $1 - \alpha$. This observation complies with the theoretical upper bound $\gamma(\alpha, \beta) \leq (1 - \alpha)\beta$ following from Theorem 2. Although the slope of the real curve $\gamma(\alpha, \beta)$, if seen as a function of $1 - \alpha$, is much smaller than $\beta$, the tendency is approximately linear at least in the regime $1 - \alpha \in [0, 0.5]$.




(a) MORPH - 3.3k precisely annotated

(b) MORPH - 6.6k precisely annotated

(c) WILD - 3.3k precisely annotated

(d) WILD - 6.6k precisely annotated

Fig. 3: The figures show test MAE for the ordinal classifiers learned by the VILMA from different training sets. The x-axis corresponds to the total number of examples in the training set. In the case of partial annotation, x-axis corresponds $m_P + m_I$ where $m_P$ is the number of partial and $m_I$ the number of precisely annotated examples, respectively. The figures (a)(c) show results for $m_P = 3300$ and figures (b)(d) for $m_P = 6600$, respectively. In the supervised case, the x-axis is just the number of precisely annotated examples. Each figure shows one curve for the supervised setting plus three curves corresponding to the partial setting with different width $u \in \{5, 10, 20\}$ of the annotation intervals. The results for MORPH database are in figures (a)(b) and the results for WILD in (c)(d).

## 6 Conclusions

We have proposed a V-shape interval-insensitive loss function suitable for risk minimization based learning of ordinal classifiers from partially annotated examples. We proved that under reasonable assumption on the annotation process the Bayes risk of the ordinal classifier can be bounded by the expectation of the associated interval-insensitive loss. We proposed a convex surrogate

MORPH

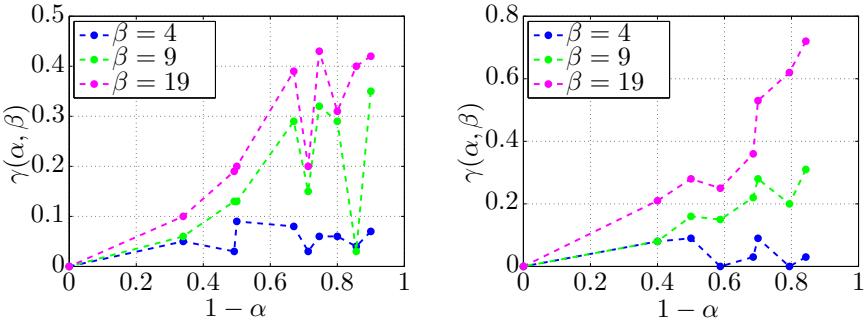
| | $m = 3300$ | $m = 6600$ | $m = 13000$ | $m = 23000$ | $m = 33000$ |
|---|---|---|---|---|---|
| Supervised | $5.56 \pm 0.02$ | $5.12 \pm 0.02$ | $4.83 \pm 0.02$ | $4.66 \pm 0.01$ | $4.55 \pm 0.02$ |
| 3.3K-5 | $5.56 \pm 0.02$ | $5.21 \pm 0.04$ | $4.89 \pm 0.03$ | $4.70 \pm 0.01$ | $4.62 \pm 0.01$ |
| 3.3K-10 | $5.56 \pm 0.03$ | $5.25 \pm 0.02$ | $5.15 \pm 0.05$ | $4.97 \pm 0.01$ | $4.90 \pm 0.04$ |
| 3.3K-20 | $5.56 \pm 0.03$ | $5.32 \pm 0.03$ | $5.26 \pm 0.06$ | $5.06 \pm 0.04$ | $4.97 \pm 0.01$ |
| 6.6K-5 | — | $5.12 \pm 0.02$ | $4.86 \pm 0.02$ | $4.69 \pm 0.00$ | $4.61 \pm 0.00$ |
| 6.6K-10 | — | $5.13 \pm 0.02$ | $4.96 \pm 0.03$ | $4.81 \pm 0.01$ | $4.84 \pm 0.04$ |
| 6.6K-20 | — | $5.13 \pm 0.02$ | $5.03 \pm 0.02$ | $4.86 \pm 0.04$ | $4.86 \pm 0.01$ |

WILD

| | $m = 3300$ | $m = 6600$ | $m = 11000$ | $m = 16000$ | $m = 21000$ |
|---|---|---|---|---|---|
| Supervised | $10.40 \pm 0.03$ | $9.60 \pm 0.03$ | $9.14 \pm 0.02$ | $8.89 \pm 0.02$ | $8.68 \pm 0.02$ |
| 3.3K-5 | $10.40 \pm 0.03$ | $9.69 \pm 0.02$ | $9.23 \pm 0.05$ | $8.89 \pm 0.02$ | $8.71 \pm 0.02$ |
| 3.3K-10 | $10.40 \pm 0.03$ | $9.76 \pm 0.02$ | $9.42 \pm 0.04$ | $9.09 \pm 0.02$ | $8.99 \pm 0.02$ |
| 3.3K-20 | $10.40 \pm 0.03$ | $9.88 \pm 0.03$ | $9.67 \pm 0.04$ | $9.51 \pm 0.00$ | $9.40 \pm 0.01$ |
| 6.6K-5 | — | $9.60 \pm 0.03$ | $9.22 \pm 0.06$ | $8.89 \pm 0.02$ | $8.71 \pm 0.02$ |
| 6.6K-10 | — | $9.60 \pm 0.03$ | $9.22 \pm 0.02$ | $9.04 \pm 0.03$ | $8.90 \pm 0.02$ |
| 6.6K-20 | — | $9.60 \pm 0.03$ | $9.35 \pm 0.06$ | $9.14 \pm 0.03$ | $9.04 \pm 0.02$ |

Table 2: The summarizes test MAE of the ordinal classifier learned from the training set with $m$ examples. The upper row shows results of the supervised setting when all $m$ examples are precisely annotated. The bottom rows show results of learning from $m_p$ precisely annotated examples and $m_I = m - m_P$ examples annotated by intervals of width $u$. For example, the row 3.3K-5 contains results for $m_P = 3300$, $u = 5$ and $m$ shown in corresponding column.

of the interval-insensitive loss associated to an arbitrary supervised V-shaped loss. We derived a generic V-shaped Interval insensitive Loss Minimization Algorithm (VILMA) which translates learning to a convex optimization problem. We also derived other convex surrogate losses by extending the existing state-of-the-art SVOR-EXP and SVOR-IMC algorithm. We showed that VILMA subsumes the SVOR-IMC as a special case. We have proposed a cutting plane method which can solve large instances of the convex learning problems. The experiments conducted on a real-life problem of human age estimation show that the proposed method has strong practical potential. The results show that the ordinal classifier with accuracy closely matching the state-of-the-art results can be obtained by learning from cheap partial annotations in contrast to so far used supervised methods which require expensive precisely annotated examples.




(a) MORPH (b) WILD

Fig. 4: The figures show $\gamma(\alpha,\beta) = \hat{R}^{MAE}(h^{\alpha,\beta}) - \hat{R}^{MAE}(h^*)$ which is the loss in accuracy caused by training from partially annotated examples generated by $\alpha\beta$-precise annotation process relatively to the supervised case. The value of $\gamma(\alpha,\beta)$ is shown for different $\beta$ (note that $u = \beta + 1$ is the interval width) as a function of the portion of the partially annotated examples $1-\alpha$. The figure (a) and (b) contains the results obtained on the MORPH and the WILD database, respectively.

## Appendix

Proof of Theorem 2

We will prove the bound (8) for each observation $\boldsymbol{x} \in \mathcal{X}$ separately, that is, we prove

$$R^{MAE}(h \mid \boldsymbol{x}) \leq R_I^{MAE}(h \mid \boldsymbol{x}) + (1-\alpha)\beta\,, \tag{21}$$

where $R^{MAE}(h \mid \boldsymbol{x}) = \mathbb{E}_{y\sim p(y|\boldsymbol{x})}|y - h(\boldsymbol{x})|$ and

$$R_I^{MAE}(h \mid \boldsymbol{x}) = \mathbb{E}_{[y_l,y_r]\sim p(y_l,y_r|\boldsymbol{x})} \min_{y'\in[y_l,y_r]} |y' - h(\boldsymbol{x})|\,.$$

It is clear that (21) satisfied for all $\boldsymbol{x} \in \mathcal{X}$ implies (8). Let us define a function which measures a discrepancy between the MAE and the its interval insensitive counterpart:

$$\delta(h(\boldsymbol{x}), y, y_l, y_r) = |y - h(\boldsymbol{x})| - \min_{y'\in[y_l,y_r]} |y' - h(\boldsymbol{x})| = \begin{cases} |h(\boldsymbol{x}) - y| \text{ if } h(\boldsymbol{x}) \in [y_l, y_r]\,, \\ y - y_l \text{ if } h(\boldsymbol{x}) < y_l\,, \\ y_r - y \text{ if } h(\boldsymbol{x}) > y_r\,. \end{cases} \tag{22}$$

Let us denote a set of intervals of unit length as $\mathcal{P}_1 = \{[y_l, y_r] \in \mathcal{P} | y_l = y_r\}$. Recall also that due to the assumption that $p(y_l, y_r \mid \boldsymbol{x}, y)$ is consistent and $\alpha\beta$-precise then we have $p(y, y \mid \boldsymbol{x}, y) = \alpha$ and $\sum_{[y_l,y_r]\in\mathcal{P}_1} p(y_l, y_r \mid \boldsymbol{x}, y) = (1-\alpha)$.

With these definitions we can write the following chain of equations:

$$
\begin{aligned}
R_I^{MAE}(h \mid \boldsymbol{x}) &= \sum_{y\in\mathcal{Y}} \sum_{[y_l,y_r]\in\mathcal{P}} p(y \mid \boldsymbol{x}) p(y_l, y_r \mid \boldsymbol{x}, y) \min_{y'\in[y_l,y_r]} |y' - h(\boldsymbol{x})| \\
&= \sum_{y\in\mathcal{Y}} p(y \mid \boldsymbol{x}) \Bigg[ \alpha|y - h(\boldsymbol{x})| + \sum_{[y_l,y_r]\notin\mathcal{P}_1} p(y_l, y_r \mid \boldsymbol{x}, y) \min_{y'\in[y_l,y_r]} |y' - h(\boldsymbol{x})| \Bigg] \\
&= \sum_{y\in\mathcal{Y}} p(y \mid \boldsymbol{x}) \Bigg[ \alpha|y - h(\boldsymbol{x})| + \sum_{[y_l,y_r]\notin\mathcal{P}_1} p(y_l, y_r \mid \boldsymbol{x}, y) \big( |y - h(\boldsymbol{x})| - \delta(h(\boldsymbol{x}), y, y_l, y_r) \big) \Bigg] \\
&= \sum_{y\in\mathcal{Y}} p(y \mid \boldsymbol{x}) \Bigg[ |y - h(\boldsymbol{x})| - \sum_{[y_l,y_r]\notin\mathcal{P}_1} p(y_l, y_r \mid \boldsymbol{x}, y) \delta(h(\boldsymbol{x}), y, y_l, y_r) \big) \Bigg] \\
&= R^{MAE}(h \mid \boldsymbol{x}) - \sum_{y\in\mathcal{Y}} \sum_{[y_l,y_r]\notin\mathcal{P}_1} p(y \mid \boldsymbol{x}) p(y_l, y_r \mid \boldsymbol{x}, y) \delta(h(\boldsymbol{x}), y, y_l, y_r) \,.
\end{aligned}
\tag{23}
$$

By (22) we have that $\delta(h(\boldsymbol{x}), y, y_l, y_r) \leq \beta$ for all $\boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}, [y_l, y_r] \in \mathcal{P}$ and hence

$$
\sum_{y\in\mathcal{Y}} \sum_{[y_l,y_r]\notin\mathcal{P}_1} p(y \mid \boldsymbol{x}) p(y_l, y_r \mid \boldsymbol{x}, y) \delta(h(\boldsymbol{x}), y, y_l, y_r) \leq (1-\alpha)\beta \,. \tag{24}
$$

The bound (21) to be proved is obtained immediately by combing (23) and (24).

Proof of Proposifion 1

Let us first consider triplet of labels $(y, y_l, y_r)$ such that $y \notin [y_l, y_r]$. In this case, the left max-term $\max_{y\leq y_l} \left[\Delta(y, y_l) + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_l) + b_y - b_{y_l}\right]$ is an instance of margin-rescaling loss instantiatred for the suppervised loss $\Delta(y, y_l)$ defined on labels $y \in [1, y_l - 1]$. The margin-rescaling loss is known to be an upper bound of the respective suppervised loss (for proof see [Tsochantaridis et al., 2005]) and, in turn, it is also an upper bound of $\Delta_I(y_l, y_r, y)$. Analogically, we can see that the right max-term $\max_{y\geq y_r} \left[\Delta(y, y_r) + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_r) + b_y - b_{y_r}\right]$ is margin-rescaling upper bound of the loss $\Delta(y, y_r)$ on labels $y \in [y_r + 1, Y]$ and, in turn, also upper bound of $\Delta_I(y_l, y_r, y)$. The V-shaped loss $\Delta(y, y')$ is non-negative by definition and hence both max-terms are non-negative and their sum upper bounds the value of $\Delta_I(y_l, y_r, y)$ for $y \notin [y_l, y_r]$. In the case when $y \in [y_l, y_r]$ the value of $\Delta_I(y_l, y_r, y)$ is defined to be zero and hence it is also upper bounded by the sum of the non-negative max-terms.

Proof of Proposition 2

First let us prove $\sum_{\hat{y}=1}^{y_l-1} \max(0, 1 - \langle \boldsymbol{x}, \boldsymbol{w} \rangle + \theta_{\hat{y}}) = \max_{y\leq y_l} \Big[ y_l - y + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_l) + b_y - b_{y_l} \Big]$. Since $\theta_y, y = 1, \ldots, Y$ is nondecreasing sequence, sum $\sum_{\hat{y}=1}^{y_l-1} \max(0, 1 -$

$\langle \boldsymbol{x}, \boldsymbol{w} \rangle + \theta_{\hat{y}})$ equals to $\max_{y \leq y_l} \Big[ \max(0, \sum_{\hat{y}=y}^{y_l-1} 1 - \langle \boldsymbol{x}, \boldsymbol{w} \rangle + \theta_{\hat{y}}) \Big]$. Simplifying it we get $\max_{y \leq y_l} \Big[ \max(0, y_l - y + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_l) + \sum_{\hat{y}=y}^{y_l-1} \theta_{\hat{y}}) \Big]$ or taking into account conversion formulas between MORD and ORD non-degenerated classifiers [Antoniuk et al., 2013] $b_1 = 0$, $b_y = -\sum_{\hat{y}=1}^{y-1} \theta_{\hat{y}}$, we can conclude that this sum is basically nothing else than $\max_{y \leq y_l} \Big[ \max(0, y_l - y + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_l) + b_y - b_{y_l}) \Big]$. Since $y_l - y + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_l) + b_y - b_{y_l} \geq 0 \ , \forall y = 1, \ldots, Y$ we can simply omit internal maximum and write $\max_{y \leq y_l} \Big[ y_l - y + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_l) + b_y - b_{y_l} \Big]$ instead. To summarise, we just have shown that $\sum_{\hat{y}=1}^{y_l-1} \max(0, 1 - \langle \boldsymbol{x}, \boldsymbol{w} \rangle + \theta_{\hat{y}}) = \max_{y \leq y_l} \Big[ y_l - y + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_l) + b_y - b_{y_l} \Big]$. Following same login one can conclude that $\sum_{\hat{y}=y_r}^{Y-1} \max(0, 1 + \langle \boldsymbol{x}, \boldsymbol{w} \rangle - \theta_{\hat{y}}) = \max_{y \geq y_r} \Big[ y - y_r + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_r) + b_y - b_{y_r} \Big]$. An analogical technique can be used for degenerated ordinal classifiers.

## References


[Antoniuk et al., 2013] Antoniuk, K., Franc, V., and Hlavac, V. (2013). Mord: Multi-class classifier for ordinal regression. In *ECML/PKDD (3)*, pages 96–111.

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, New York, NY, USA.

[Chang et al., 2011] Chang, K., Chen, C., and Hung, Y. (2011). Ordinal hyperplane ranker with cost sensitivities for age estimation. In *Computer Vision and Pattern Recognition*.

[Chu and Ghahramani, 2005] Chu, W. and Ghahramani, Z. (2005). Preference learning with gaussian processes. In *Proc. of the International Conference on Machine Learning*.

[Chu and Keerthi, 2005] Chu, W. and Keerthi, S. S. (2005). New approaches to support vector ordinal regression. In *In Proceedings of the 22nd international conference on Machine Learning (ICML)*, pages 145–152.

[Cour et al., 2011] Cour, T., Sapp, B., and Taskar, B. (2011). Learning from partial labels. *Journal of Machine Learning Research*, 12:1225–1261.

[Crammer and Singer, 2001] Crammer, K. and Singer, Y. (2001). Pranking with ranking. In *Advances in Neural Information Processing Systems (NIPS)*, pages 641–647. MIT Press.

[Debczynski et al., 2008] Debczynski, K., Kotlowski, W., and Slowinski, R. (2008). Ordinal classification with decision rules. In *Mining Complex Data, Lecture Notes in Computer Science*, volume 4944, pages 169–181.

[Dempster et al., 1997] Dempster, A., Laird, N., and Rubin, D. (1997). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 1(39).

[Do and Artières, 2009] Do, T.-M.-T. and Artières, T. (2009). Large margin training for hidden markov models with partially observed states. In *In Proceedings of the international conference on Machine Learning (ICML)*.

[Franc et al., 2012] Franc, V., Sonnenburg, S., and Werner, T. (2012). *Cutting-Plane Methods in Machine Learning*, chapter 7, pages 185–218. The MIT Press, Cambridge,USA.

[Fu and Simpson, 2002] Fu, L. and Simpson, D. G. (2002). Conditional risk models for ordinal response data: simultaneous logistic regression analysis and generalized score test. *Journal of Statistical Planning and Inference*, pages 201–217.

[Gondzio et al., 1996] Gondzio, J., du Merle, O., Sarkissian, R., and Vial, J.-P. (1996). Accpm - a library for convex optimization based on an analytic center cutting plane method. *European Journal of Operational Research*.

[Guo and Mu, 2010] Guo, G. and Mu, G. (2010). Human age estimation: What is the influence across race and gender? In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

[Huang et al., 2007] Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

[Jie and Orabona, 2010] Jie, L. and Orabona, F. (2010). Learning from candidate labeling sets. In *Advances in Neural Information Processing Systems (NIPS)*.

[Kotlowski et al., 2008] Kotlowski, W., Dembczynski, K., Greco, S., and Slowinski, R. (2008). Stochastic dominance-based rough set model for ordinal classification. *Journal of Information Sciences*, 178(21):4019–4037.

[Kumar et al., 2009] Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. (2009). Attribute and Simile Classifiers for Face Verification. In *IEEE International Conference on Computer Vision (ICCV)*.

[Li and Lin, 2006] Li, L. and Lin, H.-T. (2006). Ordinal regression by extended binary classification. In *Advances in Neural Information Processing Systems (NIPS)*.

[Lou and Hamprecht, 2012] Lou, X. and Hamprecht, F. A. (2012). Structured learning from partial annotations. In *In Proceedings of the international conference on Machine Learning (ICML)*.

[McCullagh, 1980] McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statical Society*, 42(2):109–142.

[Minear and Park, 2004] Minear, M. and Park, D. (2004). A lifespan database of adult facial stimuli. *Behavior research methods, instruments, & computers: a journal of the Psychonomic Society*, 36:630–633.

[Ramanathan et al., 2009] Ramanathan, N., Chellappa, R., and Biswas, S. (2009). Computational methods for modeling facial aging: Asurvey. *Journal of Visual Languages and Computing*.

[Rennie and Srebro, 2005] Rennie, J. D. and Srebro, N. (2005). Loss functions for preference levels: Regression with discrete ordered labels. In *Proc. of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling*.

[Ricanek and Tesafaye, 2006] Ricanek, K. and Tesafaye, T. (2006). Morph: A longitudial image database of normal adult age-progression. In *In proc. of Automated Face and Gesture Recognition*.

[Shashua and Levin, 2002] Shashua, A. and Levin, A. (2002). Ranking with large margin principle: Two approaches. In *In Proceedings of Advances in Neural Information Processing Systems (NIPS)*.

[Sonnenburg and Franc, 2010] Sonnenburg, S. and Franc, V. (2010). Coffin: A computational framework for linear svms. In *Proceedings of the 27th Annual International Conference on Machine Learning (ICML 2010)*, Madison, USA. Omnipress.

[Sonnenburg et al., 2010] Sonnenburg, S., Rätsch, G., Henschel, S., Widmer, C., Behr, J., Zien, A., Bona, F. d., Binder, A., Gehl, C., and Franc, V. (2010). The shogun machine learning toolbox. *J. Mach. Learn. Res.*, 11:1799–1802.

[Teo et al., 2010] Teo, C. H., Vishwanthan, S., Smola, A. J., and Le, Q. V. (2010). Bundle methods for regularized risk minimization. *J. Mach. Learn. Res.*, 11:311–365.

[Tsochantaridis et al., 2005] Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y., and Singer, Y. (2005). Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484.

[Uřičář et al., 2012] Uřičář, M., Franc, V., and Hlaváč, V. (2012). Detector of facial landmarks learned by the structured output SVM. In Csurka, G. and Braz, J., editors, *VISAPP '12: Proceedings of the 7th International Conference on Computer Vision Theory and Applications*, volume 1, pages 547–556, Porto, Portugal. SciTePress - Science and Technology Publications.

[Vapnik, 1998] Vapnik, V. N. (1998). *Statistical learning theory*. Adaptive and Learning Systems. Wiley, New York, New York, USA.