CENTER FOR
MACHINE PERCEPTION

CZECH TECHNICAL
UNIVERSITY IN PRAGUE

Discriminative learning from partially annotated examples

Kostiantyn Antoniuk

Study Programme: Electrical Engineering and
Information Technology

Branch of Study: Artificial Intelligence and Biocybernetics

antonkos@fel.cvut.cz

CTU–CMP–2016–07

June 14, 2016

Available at
ftp://cmp.felk.cvut.cz/pub/cvl/articles/antoniuk/Antoniuk-TR-2016-07.pdf

Thesis Advisors: Ing. Vojtěch Franc, Ph.D. ,
prof. Ing. Václav Hlaváč, CSc.

Acknowledgements: SGS15/201/OHK3/3T/13, CAK/TE01020197, UP-Driving/688652, GACR/P103/12/G084.

Research Reports of CMP, Czech Technical University in Prague, No. 7, 2016

Published by

Center for Machine Perception, Department of Cybernetics
Faculty of Electrical Engineering, Czech Technical University
Technická 2, 166 27 Prague 6, Czech Republic
fax +420 2 2435 7385, phone +420 2 2435 7637, www: http://cmp.felk.cvut.cz

PhD THESIS

ISSN 1213-2365

Discriminative learning from partially annotated examples

Kostiantyn Antoniuk

June 14, 2016

Abstract

A number of algorithms and its applications for automatic classifiers learning from examples is ever growing. Most of existing algorithms require a training set of completely annotated examples, which are often hard to obtain. In this thesis, we tackle the problem of learning from partially annotated examples, which means that each training input comes with a set of admissible labels only one of which is correct. We contributed to two different cases of this scenario. In the first case, we studied the problem of learning the ordinal classifiers from examples with interval annotation of labels. We designed a convex learning algorithm for this case and demonstrated its advantage on real data empirically. At the same time, we made several contributions to the supervised learning of the ordinal classifiers, namely, we proposed new parametrization of the ordinal classifier, we introduced more flexible piece wise version of the ordinal classifier, and we proposed a generic cutting plane solver with convergence guarantees. In the second case, we studied the problem of learning the structured output classifiers from examples with missing annotation of a subset of labels. We have defined the concept of a surrogate classification calibrated partial loss, the minimization of which guarantees that learning is statistical consistent under fairly general conditions on the data generating process. We proved the existence of a convex classification calibrated surrogate loss for learning from partially annotated examples. We showed which existing surrogate losses are classification calibrated and which are not. Our work thus provides a missing theoretical justification for so far heuristic methods which have been successfully used in practice.

Abstrakt

Počet aplikací algoritmů pro automatické učení klasifikátorů z příkladů stále roste. Většina učících algoritmů vyžaduje trénovací množinu kompletně anotovaných příkladů, které je často težké zíkat. V této disertaci se zabýváme problémem učení z částečně anotovaných příkladů. Částečná anotace znamená, že každému trénovacímu vstup je přiřazena množina přípustných skrytých stavů, z nichž pouze jediný je spravný. V disertaci popisujeme dva případy patřící do tohoto scénaře. V prvním případě jsme zkoumali učení ordinálních klasifikátorů z příkladů anotovaných intervalem skrytých stavů. Pro tento případ jsme navrhli konvexní učící algoritmus a ověřili jeho funkčnost na reálných datech. Současně jsme přispěli k řešení problému učení ordinálních klasifikátorů z kompletně anotovaných dat, a to konkrétně návrhem nové parametrizace ordinálního klasifikátoru, flexibilnějším model pro ordinální klasifikaci a obecným optimalizačním algoritmem s garancí konvergence. V druhém případě jsme studovali problém učení strukturních klasifikátorů z příkladů s chybějící anotací u podmnožiny skrytých stavů. Definovali jsme pojem náhradní klasifikačně kalibrované částečné ztrátové funkce, jejíž minimalizace zaručuje, že učení je statisticky konzistentní za dosti obecných podmínek na proces generující data. Dokázali jsme, že existuje konvexní kalibrovaná náhradní ztrátová funkce pro učení z částečně anotovaných příkladů. Ukázali jsme, které z existujících náhradních ztrátových funkcí jsou kalibrované, a které nejsou. Naše práce tak doplňuje chybějící teoretické odůvodnění pro doposud heuristické metody úspěšně používané v praxi.

Acknowledgement

I am thankful to my two supervisors Ing. Vojtěch Franc, Ph.D. and prof. Ing. Václav Hlaváč, CSc. for fruitful discussions and brilliant ideas which led me in my research towards the fulfillment of the Ph.D. degree.

I am grateful that I was a summer intern at Google Inc. Zürich in 2014. Although I am not allowed to give details due to Non-Disclosure Agreement (NDA), it was a source of great experience.

I gratefully acknowledge that my research was supported by the Technology Agency of the Czech Republic under Project TE01020197 Center Applied Cybernetics and Automated Urban Parking and Driving under Project 688652 funded by the European Union's Horizon 2020, the Grant Agency of the CTU under the projects SGS15/201/OHK3/3T/13 and the Grant Agency of Czech Republic under project GACR/P103/12/G084.

I hereby certify that the results presented in this thesis were achieved during my own research in cooperation with my thesis advisors Ing. Vojtěch Franc, Ph.D. and prof. Ing. Václav Hlaváč, CSc.

Contents

Abbreviations

ACCPM	Analytic Center Cutting Plane Method	39
BMRM	Bundle Method for Risk Minimization	17
CCCP	Convex-Concave Procedure	17
CPA	Cutting plane algorithm	20
CRF	Conditional Random Field	16
ERM	Empirical Risk Minimization	5
II-SVOR-EXP	Interval-Insensitive SVOR-EXP	34
II-SVOR-IMC	Interval-Insensitive SVOR-IMC	34
LP	Linear Programming	18
LinCls	Linear Multi-class SVM classifier	39
LinReg	Rounded linear regressor	39
MAE	Mean Absolute Error	26
MORD	Multi-class Ordinal classifier	13
NBMRM	Non-Convex Bundle Method for Risk Minimization	17
PW-MORD	Piece-Wise Multi-class ORDinal classifier	13
QP	Quadratic Programming	37
SO-SVM	Structured Output Support Vector Machine	13
SVOR-EXP	Support Vector Ordinal Regression: Explicit Constraints on Thresholds	19
SVOR-IMC	Support Vector Ordinal Regression: Implicit Constraints on Thresholds	19
SVR	Support Vector Regression	31
VILMA	V-shaped interval insensitive loss minimization algorithm	20

Symbols

$\mathcal{X}$	Instance space
$\boldsymbol{x}$	Instance
Y	Label space
$\mathcal{Y}$	Labeling space
y	Discrete label
$\boldsymbol{y}$	Labeling
$\boldsymbol{w}$	Weight vector
Argmin	The set of all minimizers of function
argmin	Single minimizer of function
h	Classifier, a mapping from the instance space to the label space
ℓ	Original loss function
ψ, variants	Surrogate loss functiton
pred	Predictor mapping from the surrogate space to the label space
R^{ℓ}, variants	Risk induced by a loss function ℓ
$[\![A]\!]$	is the Iverson bracket. It evaluates to 1 if A holds, otherwise it is 0.
$\langle\cdot,\cdot\rangle$	Dot product
$\mathcal{D}^m_{xy}$	Dataset with fully annotated examples
$\mathcal{D}^m_{xI}$	Dataset with interval-annotated examples
$\mathcal{D}^m_{xa}$	Dataset for a structured output classifier with missing labels

Failure is an option here. If things are not failing, you are not innovating enough.

– Elon Musk

1. Introdution

In this thesis, we consider a problem of learning classifiers from partially annotated examples. This means that instead of a single label per instance, we are given a set of admissible labels only one of which is correct. Such scenario is common in practice. For instance, the problem of learning from partially annotated examples naturally arises in age recognition from facial images. Instead of acquiring a precise age for each facial image in the training set, which is often expensive or impossible, it is easier to collect age ranges that can be, for example, estimated by a human annotator. See Figure 1.1 where each subject is annotated by a range of ages instead of a precise age. Another motivating application can be image segmentation as illustrated in Figure 1.2. Obtaining a ground true label for each pixel in the image is obviously tedious and expensive, therefore very often we are provided with an incomplete labeling, meaning that some pixels in the training image are left unannotated.

To put the problem of learning from partial annotations into perspective, it is useful to list other common learning scenarios (see also Figure 1.3):

- In the **supervised** scenario each training instance is annotated with a single label.
- In the **unsupervised** scenario training instances have no label at all.
- In the **semi-supervised** scenario each training instance either has a single label or it has no label at all.
- In the **multi-instance** scenario training instances are not individually labeled but grouped into sets, which either contain at least one positive example or only negative examples.
- In the **partially annotated** scenario, i.e. the scenario analyzed in this thesis, each training instance is annotated with a set of admissible labels only one of which is correct.

There exists two standard paradigms that have been used for learning from partially annotated examples: the generative approach and the discriminative approach. The generative approach tries to model the joint probability distribution of the input observations and the labels. To this end, one has to select an appropriate class of probabilistic models. As soon as the class of the probabilistic models is chosen, the maximum likelihood method (or other

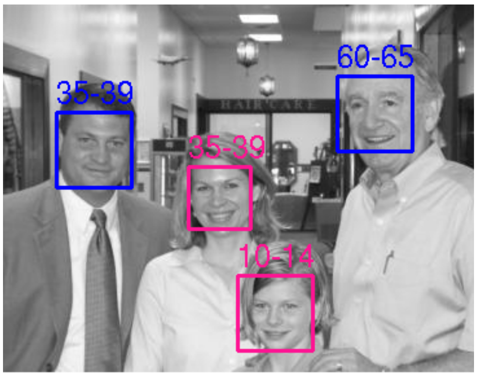


Figure 1.1. Example of facial images with partial annotation of age. Getting rough age ranges of each person is relatively easy while providing exact age is difficult.

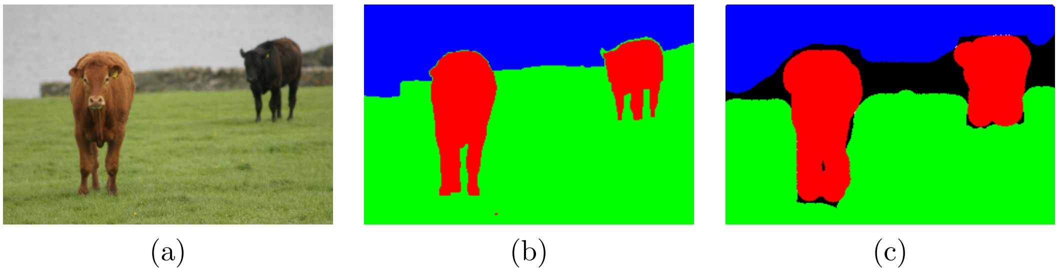
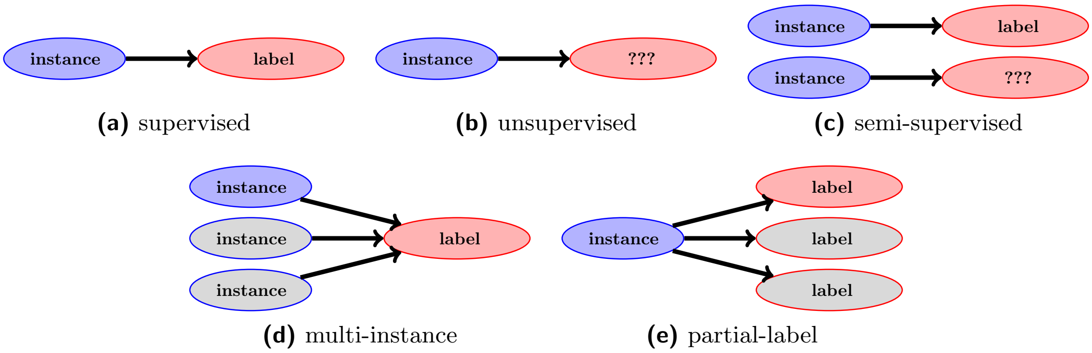


Figure 1.2. An example of a training instance when learning structured output classifier for image segmentation task. Example of an input image (a), a good(complete) labeling (b), coarse (partial) labeling (c).



(a) supervised **(b)** unsupervised **(c)** semi-supervised

(d) multi-instance **(e)** partial-label

Figure 1.3. Different learning scenarios (figure adopted from [Cour et al., 2011]).

estimation method) is used to select a single model best fitting to the training data. Finally, the required classification rule is inferred from the learned probabilistic model. On the other hand, the discriminative approach tries to learn the classification rule directly. To this end, one has to select an appropriate class of classification rules. Once the classification model is chosen, the Empirical Risk Minimization (ERM) principle (or other method) is used to select a single classification rule best fitting to the training data.

In this thesis, we follow the discriminative approach. The existing discriminative methods for learning from partially annotated examples often suffer from the following problems:

1. There is no clear connection between the target objective and the objective function actually optimized by the learning algorithm. The target objective is typically the expectation of the complete loss which evaluates the response of the classifier given the ground truth label. The objective function of the learning algorithm is typically an average of a “partial loss” computed on the partially annotated examples. The partial loss is a certain function which evaluates the response of the classifier given the partial annotation.
2. The learning problem is usually transformed into a non-convex minimization problem which is then approached by a local optimization method with no certificate of optimality.

During our work, we were mainly focused on these two problems. In short, our main contributions are the following:

- (Ad problem 1) We developed tools which allow to analyze the statistical consistency of algorithms learning the structured output classifiers from partially annotated examples. Here the partial annotation means that a subset of output labels describing the training instance is missing, e.g. like in the image segmentation examples show in Figure 1.2. We applied the proposed methodology to existing ad-hoc algorithms and we showed which of them are statistically consistent and which are not. Loosely speaking, the consistent algorithm provides a minimizer of the target objective, i.e. the expectation of the complete loss, provided the number of partially annotated training examples goes to infinity. That is, we built a missing bridge between the objective function of the consistent algorithms and the target objective.

- (Ad problem 2) We introduced a new partial loss applicable for learning the ordinal classifiers from examples with the interval annotation of the labels, e.g. like in the example shown in Figure 1.1. We establish a connection between the proposed partial loss and an associated complete target loss. We designed a convex surrogate of the partial loss which allows to convert learning into an optimization problem which can be solved efficiently and we show how to do it by cutting plane methods. As a byproduct we made several contributions to the supervised learning of the ordinal classifiers, namely, we proposed new parametrization of the ordinal classifier and we introduced more flexible piece wise version of the ordinal classifier.

In the following two sections we briefly describe the ERM based learning algorithms. We first outline the standard supervised scenario and then the scenario with the partially annotated examples. We use the two sections in order to introduce a notation which allow us to describe goals and contributions of the thesis more precisely at the very end of this chapter.

1.1. Discriminative learning from fully annotated examples

Let us briefly describe supervised learning algorithms based on the ERM principle. The supervised algorithms require a set of completely annotated training examples

$$\mathcal{D}_{xy}^m = \{(\boldsymbol{x}^1, y^1), \dots, (\boldsymbol{x}^m, y^m)\} \in (\mathcal{X} \times \mathcal{Y})^m \tag{1.1}$$

typically assumed to be drawn from independent and identically distributed (i.i.d.) random variables with some unknown distribution $p(\boldsymbol{x}, y)$. The symbol $\mathcal{X}$ denotes a set of input observations and $\mathcal{Y}$ is a set of labels to be predicted. In this thesis we assume that $\mathcal{Y}$ is finite. The goal of the supervised learning is formulated as follows. Given a loss function $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ and the training examples (1.1), the task is to learn the classifier $h\colon \mathcal{X} \to \mathcal{Y}$ whose *Bayes risk* (the target objective)

$$R^\ell(h) = \mathbb{E}_{p(\boldsymbol{x},y)} \ell(y, h(\boldsymbol{x})) \tag{1.2}$$

is as small as possible, i.e. ideally, we would like to obtain the best Bayes classifier [1]

$$h_*^\ell \in \operatorname*{Argmin}_{h\colon \mathcal{X} \to \mathcal{Y}} R^\ell(h)\,. \tag{1.3}$$

The minimization problem (1.3) cannot be solved directly due to the unknown distribution $p(\boldsymbol{x}, y)$. The ERM principle approaches the problem (1.3) by the following approximations:

[1] Strictly speaking one has to consider $\inf\limits_{h\colon \mathcal{X} \to \mathcal{Y}} R^\ell(h)$ here, however, in order to make the main message clear we assume that infimum is reachable and we can use minimum instead. Later, in Chapter 4, we describe our contribution in the strict way using infimums.

- The empirical distribution

$$s(\boldsymbol{x}, y) = \frac{1}{m} \sum_{i=1}^{m} [\![\boldsymbol{x}^i = \boldsymbol{x} \wedge y^i = y]\!] \tag{1.4}$$

  is used instead of the true but unknown distribution $p(\boldsymbol{x}, y)$.
- The set of all possible classifiers $h \colon \mathcal{X} \to \mathcal{Y}$ is restricted to some predefined set of rules $\mathcal{H}$ (the hypothesis space).

Using these approximations, the ERM amounts to solving

$$h_*^{\mathrm{emp}} \in \operatorname*{Argmin}_{h \in \mathcal{H}} R_{\mathrm{emp}}^{\ell}(h) \,, \tag{1.5}$$

where

$$R_{\mathrm{emp}}^{\ell}(h) = \mathbb{E}_{s(\boldsymbol{x},y)} \ell(y, h(\boldsymbol{x})) \tag{1.6}$$

is the empirical risk and h_*^{emp} is the learned classification rule. Under certain conditions [Vapnik, 1995], the ERM is statistically consistent learning algorithm, i.e. for the number of examples going to infinity the expected risk of the learned classifier $R^{\ell}(h_*^{\mathrm{emp}})$ converges in probability to the minimal Bayes risk $R^{\ell}(h_*^{\ell})$.

Unfortunately even simple instances of the ERM problem (1.5) are hard to solve efficiently and thus it is further simplified in the following way. The original loss function $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is replaced by a surrogate loss function $\psi \colon \mathcal{Y} \times \hat{\mathcal{T}} \to \mathbb{R}$ operating on a surrogate decision set $\hat{\mathcal{T}} \subset \mathbb{R}^Y$. With the help of the surrogate loss function ψ we learn a surrogate decision function $f \colon \mathcal{X} \to \hat{\mathcal{T}}$, which is then used to construct the decision function $h \colon \mathcal{X} \to \mathcal{Y}$ via a predefined transform $\mathrm{pred} \colon \hat{\mathcal{T}} \to \mathcal{T}$, i.e. $h = \mathrm{pred} \circ f$. As before, the set of all surrogate decision functions $f \colon \mathcal{X} \to \hat{\mathcal{T}}$ is restricted to a subset $\mathcal{F}$. With these changes, the ERM problem (1.5) is simplified to a search for the the best surrogate decision function by solving

$$f_*^{\mathrm{emp}} \in \operatorname*{Argmin}_{f \in \mathcal{F}} R_{\mathrm{emp}}^{\psi}(f) \,, \tag{1.7}$$

where

$$R_{\mathrm{emp}}^{\psi}(f) = \mathbb{E}_{s(\boldsymbol{x},y)} \psi(y, f(\boldsymbol{x})) \tag{1.8}$$

and the resulting classification rule is $h = \mathrm{pred} \circ f_*^{\mathrm{emp}}$. The surrogate loss function ψ is typically chosen to be a convex function that upper bounds the original loss ℓ. The convex surrogate loss makes the problem (1.7) convex and much easier to deal with than the original problem (1.5). Besides the convexity, however, the used surrogate loss should have a clear statistical meaning. A natural requirement is to use such surrogate losses which preserves the statistical consistency of the ERM principle.

Loosely speaking, if a given surrogate loss $\psi \colon \mathcal{Y} \times \hat{\mathcal{T}} \to \mathbb{R}$ is so called classification calibrated [Ramaswamy and Agarwal, 2012] (or statistically consistent [Zhang, 2004a,b], or Fisher consistent [Shi et al., 2015]) with respect to the original loss $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ then it holds that

$$\mathrm{pred} \circ f_*^{\psi} \in \operatorname*{Argmin}_{h \colon \mathcal{X} \to \mathcal{Y}} R^{\ell}(h) \tag{1.9}$$

for any distribution $p(\boldsymbol{x}, y)$, where

$$f_*^{\psi} \in \operatorname*{Argmin}_{f \colon \mathcal{X} \to \hat{\mathcal{T}}} R^{\psi}(f) \tag{1.10}$$

and

$$R^{\psi}(f) = \mathbb{E}_{p(\boldsymbol{x},y)}\psi(y, f(\boldsymbol{x})) \tag{1.11}$$

is the expectation of the surrogate loss. In words, using a classification calibrated surrogate guarantees that a solution of the surrogate problem (1.10) is a decision function f_*^{ψ} which defines a classification rule $h = \text{pred} \circ f_*^{\psi}$ being itself a minimizer of the original task (1.9). In practice the distribution $p(\boldsymbol{x}, y)$ is unknown and thus we do not minimize the surrogate risk $R^{\psi}(f)$ but rather its empirical estimate $R^{\psi}_{\text{emp}}(f)$. However, it can be still shown that the classification calibrated surrogate preserves the statistical consistency of the ERM.

It should be emphasized that the classification calibrated surrogate loss does not have to be an upper bound of the target loss $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ since the inclusion in (1.9) is required for the minimizers of the surrogate loss. Of course, an analysis of the minimizers is often more difficult and hence, it is common in practice to deal with surrogate losses that are upper bounds of the target loss. See Figure 1.4 for illustration.

Example. Let us consider learning of a multi-class linear classifier. A surrogate decision function is learned from a set of linear functions

$$\mathcal{F} = \left\{ f(\boldsymbol{x}) = (\langle \boldsymbol{w}_1, \boldsymbol{x}\rangle, \cdots, \langle \boldsymbol{w}_Y, \boldsymbol{x}\rangle)^T \mid (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_Y) \in \mathbb{R}^{n \times Y}, \|\boldsymbol{w}_1\|^2 + \cdots + \|\boldsymbol{w}_Y\|^2 \leq \lambda \right\}$$

with parameters whose Euclidean norm is bounded by $\lambda > 0$. Each $f \in \mathcal{F}$ maps an input $\boldsymbol{x} \in \mathbb{R}^n$ onto $\boldsymbol{t} \in \hat{\mathcal{T}} \subset \mathbb{R}^Y$. The classification rule is constructed by composing the decision function f with a transform $\text{pred}(\boldsymbol{t}) \triangleq \operatorname*{argmax}_{y \in \mathcal{Y}} \boldsymbol{t}_y$ so that the resulting classification rule reads

$$h(\boldsymbol{x}) = \text{pred} \circ f(\boldsymbol{x}) = \operatorname*{argmax}_{y \in \mathcal{Y}} f_y(\boldsymbol{x}) = \operatorname*{argmax}_{y \in \mathcal{Y}} \langle \boldsymbol{w}_y, \boldsymbol{x}\rangle \,.$$

For example, the multi-class Support Vector Machine algorithm learns the decision function f by solving the problem (1.7) with different surrogate loss functions used in practice:

1. A commonly used surrogate

$$\psi(\hat{y}, f(\boldsymbol{x})) = \max_{y \in \mathcal{Y}}(1 + f_y(\boldsymbol{x}) - f_{\hat{y}}(\boldsymbol{x}))$$

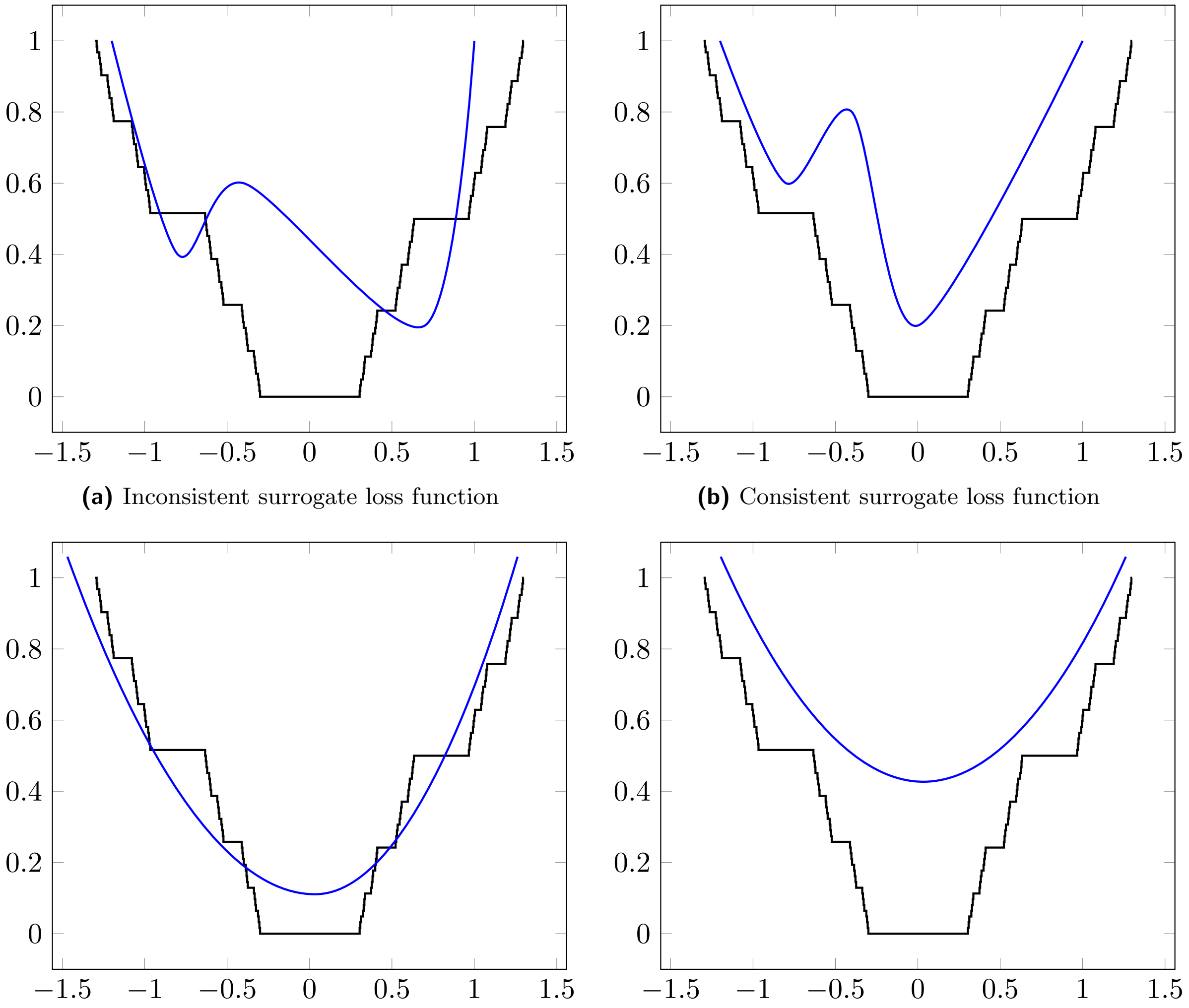
is not statistically consistent w.r.t. to the target 0/1-loss $\ell(\hat{y}, y) = [\![\hat{y} \neq y]\!]$, unless we deal only with those distributions $p(\boldsymbol{x}, y)$ such that $\exists y \in \mathcal{Y}, p(y \mid \boldsymbol{x}) > \frac{1}{2}$ ([Liu, 2007]).

2. In contrast, less commonly used surrogate loss

$$\psi(\hat{y}, f(\boldsymbol{x})) = \sum_{y \neq \hat{y}} \max(0, 1 + f_y(\boldsymbol{x}) - f_{\hat{y}}(\boldsymbol{x}))$$

is consistent with respect to the target 0/1-loss $\ell(\hat{y}, y) = [\![\hat{y} \neq y]\!]$ ([Liu, 2007]).

The statistical consistency of the ERM based fully supervised learning algorihtms have been intensively studied. Unfortunately, it is not possible to directly apply all the existing results in the case of partially annotated examples. In the next section we are going to explain main issues which raise when we apply the ERM methods to the partially annotated examples.



(a) Inconsistent surrogate loss function

(b) Consistent surrogate loss function

(c) Consistent convex surrogate loss function

(d) Consistent convex surrogate loss function that upper bounds the original loss function

Figure 1.4. The figure illustrates different cases of the surrogate loss and the target loss function. For some fixed $p(\boldsymbol{x}, y)$, $\boldsymbol{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$, we plot the value of original loss as $\ell(y, \mathrm{pred} \circ f(\boldsymbol{x}))$ and the surrogate loss as $\psi(y, f(\boldsymbol{x}))$. The x-axis corresponds to $f \in \mathcal{F}$ and the y-axis shows the value of the target loss $\ell(y, \mathrm{pred} \circ f(\boldsymbol{x}))$ shown in black and the surrogate loss $\psi(y, \mathrm{pred} \circ f(\boldsymbol{x}))$ shown in blue.

1.2. Discriminative learning from partially annotated examples

In the case of learning from partially annotated examples, we are provided with a set of admissible labels only one of each is correct. This differs from the supervised setting, where we have one to one correspondence between the input instances and labels. More precisely, we consider a set of partially annotated training examples

$$\mathcal{D}^m_{xa} = \{(\boldsymbol{x}^1, \boldsymbol{a}^1), \ldots, (\boldsymbol{x}^m, \boldsymbol{a}^m)\} \in (\mathcal{X} \times 2^{\mathcal{Y}})^m \,, \tag{1.12}$$

assumed to be drawn from i.i.d. random variables with some unknown joint distribution

$$p(\boldsymbol{x}, \boldsymbol{a}) = \sum_y p(\boldsymbol{x}, \boldsymbol{a}, y) \,.$$

Each training input $\boldsymbol{x}^i$ comes along with a set of candidate labels $\boldsymbol{a}^i \subset 2^{\mathcal{Y}}$ ($|\boldsymbol{a}| \geq 1$). A common assumption on the data generating distribution $p(\boldsymbol{x}, \boldsymbol{a}, y)$ is that the ground truth label y^i is among the known candidate labels $\boldsymbol{a}^i$, i.e. $y^i \in \boldsymbol{a}^i$.

The ultimate goal is the same as in the supervised learning, that is, for a given loss function $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ we want to learn a classifier $h \colon \mathcal{X} \to \mathcal{Y}$ whose Bayes risk (1.2) defined w.r.t

$$p(\boldsymbol{x}, y) = \sum_{\boldsymbol{a}} p(\boldsymbol{x}, \boldsymbol{a}, y)$$

is as small as possible. Although the goals are the same, the learning algorithms are not. Namely, the ERM methodology cannot be used directly because the loss function $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is undefined over the annotations (i.e. the subsets $2^{\mathcal{Y}}$) contained in the partially annotated training set $\mathcal{D}^m_{xa}$. One option to make the ERM applicable is to derive so called partial loss $\ell^P \colon \mathcal{Y} \times 2^{\mathcal{Y}} \to \mathbb{R}_+$ from a given complete (target) loss ℓ by minimizing over admissible labels:

$$\ell^P(y, \boldsymbol{a}) = \min_{\hat{y} \in \boldsymbol{a}} \ell(\hat{y}, y) \,. \tag{1.13}$$

The partial loss ℓ^P has been explicitly defined in [Cour et al., 2011] for a case when ℓ is the 0/1-loss. However implicitly, via defining a learning algorithm which in its core minimizes the partial loss, it has been used many times in various contexts. For example, it is minimized by an algorithm learning the Hidden Markov Chain based classifiers [Do and Artières, 2009], generic structured output models [Lou and Hamprecht, 2012], the multi-instance learning [Luo and Orabona, 2010] or the named entity recognizer [Fernandes and Brefeld, 2011a].

Having the partial loss, we can define the partial risk

$$R^{\ell^P}(h) = \mathbb{E}_{p(\boldsymbol{x}, \boldsymbol{a})} \ell^P(h(\boldsymbol{x}), \boldsymbol{a}) \,, \tag{1.14}$$

and search for the best (Bayes) classifier $h \colon \mathcal{X} \to \mathcal{Y}$ that minimizes the partial risk

$$h^{\ell^P}_* \in \operatorname*{Argmin}_{h \colon \mathcal{X} \to \mathcal{Y}} R^{\ell^P}(h) \,. \tag{1.15}$$

The partial risk minimization problem (1.15) can be already approached by the ERM methods. However, the central question is whether the ERM methods can provide a good approximation of the target problem (1.3). An answer to this question in the case of structured output classification (i.e. when y is a vector of labels) is one of the contributions of the thesis. Our approach is very briefly outlined in the rest of this section.

We will show that for some distributions $p(\boldsymbol{x},\boldsymbol{a},y)$ the problem (1.15) is equivalent to the target problem (1.3) in the sense that both problems share the set of solutions. In particular, the classifier $h_*^{\ell^P}$ obtained by minimizing the partial risk (1.15) is a minimizer of the target (complete) risk (1.2) as well, i.e. the inclusion

$$h_*^{\ell^P} \in \operatorname*{Argmin}_{h\colon \mathcal{X}\to\mathcal{Y}} R^\ell(h) \tag{1.16}$$

holds or equivalently

$$R^\ell(h_*^{\ell^P}) = R^\ell(h_*^\ell)\,. \tag{1.17}$$

After establishing the equivalence (1.17), one can solve the partial risk minimization problem (1.15) by the ERM methods as follows. The partial risk (1.14) can be approximated by the empirical risk

$$R^{\ell^P}_{\text{emp}}(h) = \mathbb{E}_{s(\boldsymbol{x},\boldsymbol{a})}\ell^P(h(\boldsymbol{x}),\boldsymbol{a})\,, \tag{1.18}$$

where

$$s(\boldsymbol{x},\boldsymbol{a}) = \frac{1}{m}\sum_{i=1}^m [\![\boldsymbol{x}^i = \boldsymbol{x} \wedge \boldsymbol{a}^i = \boldsymbol{a}]\!]\,. \tag{1.19}$$

As in the supervised setting, the partial empirical risk (1.18) is hard to minimize directly. Hence, the partial loss $\ell^P\colon \mathcal{Y}\times 2^{\mathcal{Y}} \to \mathbb{R}_+$ is replaced by an easier-to-minimize surrogate partial loss $\psi^P\colon 2^{\mathcal{Y}}\times\hat{\mathcal{T}} \to \mathbb{R}_+$, which operates on the surrogate decision set $\hat{\mathcal{T}} \subset \mathbb{R}^Y$. The surrogate ψ^P loss is used to learn a surrogate decision function $f\colon \mathcal{X}\to\hat{\mathcal{T}}$ such that

$$f_*^{\text{emp},\psi^p} \in \operatorname*{Argmin}_{f\in\mathcal{F}} R^{\psi^P}_{\text{emp}}(f)\,, \tag{1.20}$$

where

$$R^{\psi^P}_{\text{emp}}(f) = \mathbb{E}_{s(\boldsymbol{x},\boldsymbol{a})}\psi^P(\boldsymbol{a},f(\boldsymbol{x}))\,. \tag{1.21}$$

Finally, the resulting classification rule is constructed by composing the learned function f_*^{emp,ψ^p} and a fixed prediction function pred, i.e. $h = \text{pred}\circ f_*^{\text{emp},\psi^p}$.

Likewise in the supervised setting, in order to justify the ERM problem (1.20) we also need to study consistency of the surrogate partial loss. To this end, we introduce in this thesis a concept of a classification calibrated partial loss. Loosely speaking, if a surrogate partial loss ψ^p is classification calibrated w.r.t. the partial loss ℓ^p then for any distributions $p(\boldsymbol{x},\boldsymbol{a})$ it holds that

$$\text{pred}\circ f_*^{\psi^P} \in \operatorname*{Argmin}_{h\colon \mathcal{X}\to\mathcal{Y}} R^{\ell^P}(h)\,, \tag{1.22}$$

where $f_*^{\psi^P}$ is a minimizer of the partial surrogate risk, i.e.,

$$f_*^{\psi^P} \in \operatorname*{Argmin}_{f\colon \mathcal{X}\to\hat{\mathcal{T}}} R^{\psi^P}(f)\,, \tag{1.23}$$

$$R^{\psi^P}(f) = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{a})}\psi^P(\boldsymbol{a},f(\boldsymbol{x}))\,. \tag{1.24}$$

Consequently, using the inclusion (1.20) we can show that any minimizer of the partial surrogate risk $f_*^{\psi^P}$ is the Bayes classifier of the target (complete) risk, i.e., it holds that

$$\text{pred}\circ f_*^{\psi^P} \in \operatorname*{Argmin}_{h\colon \mathcal{X}\to\mathcal{Y}} R^\ell(h)\,. \tag{1.25}$$

In this sense the minimization of the surrogate partial risk $R^{\psi^P}(f)$ is equivalent to (or consistent with) the minimization of the target risk $R^{\ell}(h)$. Under some conditions, the equivalence is preserved even if the true risks are replaced by their empirical estimates. This allows us to show that the learning algorithms which in their core solve the problem (1.20) with calibrated surrogate partial loss are statistically consistent. Namely, we will prove that for the number of examples going to infinity the expected risk of the learned classifier $R^{\ell}(\text{pred} \circ f_*^{\text{emp},\psi^P})$ converges in probability to the minimal (Bayes) risk $R^{\ell}(h_*^{\ell})$.

1.3. Thesis goals

This thesis is centered around the ERM based algorithms learning classifiers from partially annotated examples. More precisely, we concentrated on learning algorithms which in their core solve the surrogate ERM problem (1.20). At the beginning of our work on this topic, there were many ad-hoc methods showing that algorithms implementing (1.20) give promising results, i.e. they were shown to provide a good approximations of the Bayes classifier (1.3). However, there was no firm theory which would support these empirical findings. In addition, the existing algorithms often suffer from using a non-convex surrogate partial losses making the problem (1.20) hard to optimize. And thus a further question is in which cases one can construct a good convex and, at the same time, easy-to-optimize surrogate partial loss. After recognizing the open problems, we focused our work on the following questions:

- How to design a convex surrogate of the partial loss (1.13)?
- How to solve ERM problem (1.20) efficiently?
- Under which conditions are algorithms implementing ERM problem (1.20) statistically consistent?

We have not found a complete and general answer to these questions, yet we managed to contributed to all of them. A summary of our contributions is provided in the next section.

1.4. Contributions

In this work, we investigated two different classification scenarios both falling under the umbrella of learning from partial annotations. First, learning of ordinal classifiers from interval annotations. Second, learning of structured output classifiers from examples with missing labels.

1.4.1. Learning ordinal classifier from interval annotations

We consider learning of the ordinal classifiers (i.e. classification model assuming ordered labels) from examples of inputs annotated by intervals of admissible labels.

- We propose an interval insensitive loss (IIL) function to measure discrepancy between the interval of admissible labels given in the annotation and a label predicted by the classifier. The IIL can be build from arbitrary target (complete) V-shape loss like, for example, the 0/1-loss or mean absolute error (MAE). The IIL is an instance of the generic partial loss (1.13). In contrast to existing instances of the partial loss (1.13), the IIL for ordinal classification can be approximated by tight convex surrogates as we will show.
- We show that the expectation of the IIL is a reasonable proxy of the expectation of the target complete loss. In particular, we show that the target risk R^{ℓ} is upper bounded by

a linear function of the partial risk R^{ℓ^P}. We show how the tightness of this upper bound depends on the annotations process which was used to generate the training examples.

- We show how to build tight convex surrogates of the IIL. The convex surrogates are obtained by extending surrogates known from existing supervised algorithms for ordinal regression. These surrogates are can be used as a proxy for the 0/1-loss or the MAE loss. We also propose a novel convex surrogate of a generic V-shaped interval-insensitive loss.
- We propose an efficient cutting plane solver for minimization of the ERM problem (1.20). In contrast to existing CPA solvers, it can deal with situations when the quadratic regularizer is not imposed on all model parameters which, as will be also shown, has significant influence on the final accuracy of the learned ordinal classifier.
- We have not managed to prove consistency of the IIL. Instead, we performed a thorough empirical evaluation showing that minimization of the interval insensitive loss provides a good approximation of the target Bayes classifier (1.3).

While working on this topic, we also made some progress on supervised learning of ordinal classifiers as a byproduct:

- We proved that the ordinal classifier is equivalent to a linear multi-class classifier whose class parameter vectors are collinear and with magnitude linearly increasing with the labels. We call the new representation as the Multi-class Ordinal classifier (MORD) classifier. Our equivalence proof is constructive so that we can convert any ordinal classifier to the MORD classifier and vice-versa.
- The MORD representation allows to express the space of ordinal classifiers $\mathcal{H}_{\rm ord}$ as composition of the "argmax" prediction transform $\text{pred}(\boldsymbol{t}) = \text{argmax}_{y\in\mathcal{Y}}\, t_y$ and a linear decision function $f \in \mathcal{F}$, i.e. $\mathcal{H}_{\rm ord} = \{\text{pred} \circ f \mid f \in \mathcal{F}\}$. In turn, the MORD representation can be beneficial for learning and analysis of the ordinal classifiers by using algorithms and results for well understood multi-class linear classification. For example, we show that a generic Structured Output Support Vector Machine (SO-SVM) algorithm can be applied for learning of the MORD classifier and that it delivers the same (or slightly better) results when compared to the existing learning algorithms for the ordinal classification. Moreover, the SO-SVM approach works for arbitrary loss function in contrast to existing methods which require the V-shaped losses.
- We show that the MORD representation allows introduce more complex models for ordinal classification. Namely, we propose a Piece-Wise Multi-class ORDinal classifier (PW-MORD) which subsumes the standard ordinal classifier and unrestricted multi-class classifies as special cases. We demonstrate advantages of the proposed models on standard benchmarks as well as on solving a real-life problem of estimating human age from facial images.

1.4.2. Learning structured output classifier from examples with missing labels

We concentrate on a scenario, when the object is characterized by an input observation and labelling of a set of local parts, however, a training set contains examples of inputs and labelings only for a subset of the local parts.

- We provide sufficient conditions which admit to prove that the expected risk $R^{\ell}(h)$ of the structured predictor h learned by minimizing the partial risk $R^{\ell^P}(h)$ converges in probability to the optimal Bayes risk $R^{\ell}(h^{\ell}_{*})$. The sufficient conditions restrict the target loss ℓ to be additive over the local parts while the data generating process $p(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{y})$ can be fairly generic.

- We define a concept of classification calibrated surrogate partial losses which are easier to optimize, yet their minimization preserves the statistical consistency.
- We analyze surrogate losses used by the existing algorithms implementing the ERM minimization (1.20) for learning of structured output classifiers from examples with missing labels. For example, we show that the ramp-loss and some of its modifications which have been most frequently used are classification calibrated and, in turn, the corresponding algorithms are statistically consistent. Our analysis provides a missing theoretical justification for so far heuristic methods.
- We prove the existence of a convex classification calibrated surrogate for partial learning. The proof is based on establishing a connection between learning from partially annotated examples and the recently published theory on consistency of supervised learning.

1.5. Thesis outline

Chapter 2 contains the state-of-the-art relevant to the topics studied in the thesis. In particular, we review works related to the statistical consistency of algorithms learning from fully annotated and partially annotated examples, we also review optimization algorithms which have been used to solve the ERM problem (1.20) and, finely, we review existing discriminative learning methods for the ordinal classifiers.

Chapter 3 describes our contributions to the problem of learning ordinal classifiers from fully annotated examples and examples with interval (partial) annotation of labels.

Chapter 4 describes our contributions to the problem of statistical consistency of algorithms learning structured output classifiers from examples with missing labels.

Chapter 5 contains conclusions resulting from the work done in this thesis and also a discussion of a possible future work.

We give a more detailed road map of the each individual chapter at its beginning.

2. State-of-the-art

2.1. Development of the statistical consistency of learning methods

Undisputably, the consistency of a learning method is a desirable property, i.e. a good learning method should recover the Bayes classifier at least if provided with an infinitely large training set under the condition that the class of considered classifiers contains the Bayes classifier. The design of consistent learning methods for supervised multiclass prediction received the attention in the last decade: [Zhang, 2004b,a; Bartlett et al., 2006; Hill and Doucet, 2007; Tewari and Bartlett, 2007; Liu, 2007; Santos-Rodríguez et al., 2009; Zhang et al., 2009; Ramaswamy and Agarwal, 2012]. Statistical properties of learning algorithms based on the risk minimization formulation are relatively well-understood for the supervised setting due to the aforementioned works and others. However, there are quite few works studying risk based minimization methods for the learning setting with the missing labels. Among the few exceptions belong the works of [Cour et al., 2011; Cid-Sueiro et al., 2014; Yu et al., 2014].

2.1.1. Statistical consistency of the supervised flat classifiers

[Zhang, 2004b] showed first that binary classifiers obtained by minimizing infinite-sample consistent surrogate loss for supervised learning (e.g. the hinge-loss, logistic loss, etc.) can approach Bayes classifier. [Zhang, 2004a] analysed the consistency of the hinge loss and its modifications in the context of the multiclass classification formulations such as pairwise comparison, constrained comparison and One-Versus-All methods. [Liu, 2007] considered several multiclass generalizations of the hinge-loss used in various multiclass SVMs algorithms and showed that some of them were and others were not statistically consistent. For some inconsistent losses, [Liu, 2007] showed how to modify training algorithm to make the losses behave consistently. [Tewari and Bartlett, 2007] characterized classification calibration of supervised multiclass problems in terms of geometric properties of some sets associated with the surrogate loss function. Based on these properties, they provided certain sufficient conditions for the classification calibration and examine the consistency of a few multiclass methods. [Ramaswamy and Agarwal, 2012] extended the notion of the classification calibration from 0/1-loss and/or binary classification problems to the general multiclass setting with a general loss. [Ramaswamy and Agarwal, 2012] deriveed necessary and sufficient conditions for a surrogate loss function to be classification calibrated with respect to a given target loss. They introduced the notion of so called convex classification calibration dimension of a multiclass loss matrix measuring the size of a prediction space, in which it is possible to design a convex surrogate that is calibrated with respect to the target loss. They derived lower and upper bounds of the classification calibration dimension as well. These notions can be very useful if for a given target loss one has to prove existence or non-existence of a corresponding convex calibrated surrogate loss. The consistency of multiclass losses were also considered in the development of various types of other settings, e.g. the multiclass classification with reject option [Ramaswamy et al., 2015a], hierarchical classification [Ramaswamy et al., 2015b], multiclass boosting [Zhu et al., 2009; Mukherjee and Schapire, 2013], etc.

2.1.2. Statistical consistency of the supervised ordinal classifiers

[Pedregosa et al., 2014] studied the statistical consistency of methods used for supervised ordinal classifier learning of rich family of surrogate loss functions including proportional odds and support vector ordinal regression. Authors consider the threshold and the regression based models for which they derived sufficient conditions on statistical consistency for the margin based methods and the surrogates of the V-shape loss functions. In Section 3.3, we will extend the notion of V-shaped surrogate losses for dealing with interval annotations.

2.1.3. Statistical consistency of the supervised structured output classifiers

Although there is a progress in studying the supervised multiclass setting for flat classifiers, there are only few works dedicated directly to the structured output prediction. [Shi et al., 2015] investigated the relationship between the classification calibration of multiclass losses and losses for a structured output prediction in supervised scenario. They proposed a hybrid loss for supervised multiclass and structured output problems that is a convex combination of a logarithmic loss for Conditional Random Field (CRF) and a multiclass hinge loss from the SVM methods. Their family of losses is similar to those proposed previously by [Zhang et al., 2009] for 0/1 loss. [Shi et al., 2015] provided a condition for a given loss to be statisticaly consistent for classification, which depends on a measure of dominance between labels, i.e. the gap between probabilities of the best labeling and the second best labeling. They showed that the statistical consistency is necessary also for so called parametric consistency which is needed when learning models such as the CRFs.

2.1.4. Statistical consistency of the flat classifiers learned from partially annotated examples

The literature on consistency of supervised learning methods is rich. The consistency of methods learning from partially annotated examples has been addressed very rarely so far. We are aware only of two works addressing the problem, namely [Cour et al., 2011] and [Cid-Sueiro et al., 2014]. [Cour et al., 2011] considered the multiclass learning of flat classifiers from examples with candidate set of admissible labels. They proposed a convex learning formulation based on a minimization of a certain partial loss. They also analyzed conditions under which their partial loss is asymptotically consistent against the target 0/1 loss. [Cid-Sueiro et al., 2014] proposed a generic framework which, for a given supervised target loss, allows to derive a classification calibrated surrogate suitable for learning from training examples with missing labels. Authors introduced a statistical model under which they show that consistent surrogate losses for learning with missing labels can be obtained by a linear transformation of any surrogate consistent loss for the supervised setting. Authors showed that convexity can be sometimes preserved when adapting a supervised surrogate loss to its weak consistent counterpart.

Although the setting studied in [Cour et al., 2011; Cid-Sueiro et al., 2014] is quite general, (they considered any subset of labels as candidate label set), neither of them can be used to analyze existing methods learning the structured output classifiers, i.e. the multiclass classification with exponentially large number of labels. Their convex surrogate losses designed for flat classifiers are not suitable for structured case since their evaluation requires solving computationally intractable subproblems.

Nevertheless, part of the community drops the statistical part of the problem and tries to solve the problem using ad-hoc heuristics that give very often reasonably good results in practice. For the sake of completeness, we give a brief overview of existing approaches below.

2.2. Existing methods for structured output learning from partial annotations

Most of the following works try to contribute to the practical part of the problem by improving existing heuristics for solving the non-convex ERM task (1.20) in a context of particular application. We provide a short overview of optimization methods used in the structured output learning from partially annotated examples.

2.2.1. Convex concave procedure

Many different approaches assuming that the minimization of the partial empirical risk to be good estimate have beed proposed during last decade [Chuong et al., 2008; Girshick et al., 2011; Yu and Joachims, 2009; Fernandes and Brefeld, 2011a; Zhu et al., 2010; Vedaldi and Zisserman, 2009; Wang and Mori, 2010; Luo and Orabona, 2010; Lou and Hamprecht, 2012; Sarawagi and Gupta, 2008; Yu et al., 2014]. Most of these works derived a non-convex bound on the partial empirical risk (1.14) and proved empirically that it gives good estimates in various types of applications. Non-convex optimization problems trying to solve (1.20) are reduced to a sequence of convex problems, by so called Convex-Concave Procedure (CCCP), which reduces the non-convex problem into a sequence of convex ones. Convex subproblems are often solved with the help of proximal bundle method [Kiwiel, 1990]. Most of proposed improvements for CCCP consist of:

- An adaptive increasing of the precision of supervised problem on each CCCP iteration until the required precision is reached. This reduces the number of gradient evaluations needed for Bundle Method for Risk Minimization (BMRM) on each CCCP iteration.
- Reusing "good" cutting planes across multiple CCCP iterations and avoiding computing them from the scratch.

Different kind of CCCP 's improvement was proposed by [Kumar et al., 2010], so called Self-Paced Learning. Their algorithm simultaneously selects easy examples and updates parameters on each iteration in order to escape local optima. The number of samples selected at each iteration is determined by a weight that is gradually annealed so that later iterations introduce more samples. The algorithm convergences after all samples have been considered and the objective function can not be improved further.

2.2.2. Regularized bundle methods for convex and non-convex risks

[Do and Artières, 2012] adapted the convex solver BMRM [Teo et al., 2010] for non-convex optimization problem. The main idea of the Non-Convex Bundle Method for Risk Minimization (NBMRM) is, likewise in BMRM, to build an approximation of the partial surrogate loss via the cutting plane technique iteratively. Such an approximation is not an underestimator of the objective function anymore, since the objective function is no more convex. The cutting plane approximation of the objective function may cause conflict with target function, i.e. it may lead to the overestimation of the objective function at certain points. Authors overcome the conflicts between the cutting planes similarly to classical non-convex bundle methods [Kiwiel, 1985; Gaudioso and Monaco, 1992] and prove the global convergence to cluster points which are stationary solutions (not necessarily a local minimum but may be a saddle point or even a local maximum). Their method requires fine tuning of many hyper-parameters, which makes the algorithm very sensitive for each particular application.

2.2.3. Branch and bound algorithm

[Kawahara and Washio, 2011] used branch and bound algorithm for structured output learning problems with missing labels known in global optimization theory [Horst et al., 1991]. Instead of difference of two convex functions, authors considered submodular functions, the discrete analog of convex functions. In addition, instead of solving Linear Programming (LP) problem needed for computing a lower bound in a continuous case, authors solved binary-integer linear program using state of the art optimization techniques developed for the optimization of submodular functions. Authors showed empirically the advantage of the model corresponding to the global optimum of objective function over the models corresponding to local optima. Although, the branch and bound algorithm is not really suitable for large scale problems, it shows the advantage of the global solution against the local one in the considered approach (structured output learning from partially annotated data).

2.2.4. Perceptron-like algorithms

The structured output perceptron, analogous to its “flat” counterpart, has been proposed in [Schlesinger and Hlaváč, 2002; Altun et al., 2003; Collins and Koo, 2005]. Later [Fernandes and Brefeld, 2011b] derived an extension of the loss-augmented perceptron for structured output learning that allows to deal with partialy annotated sequences. To learn from partialy annotated data [Fernandes and Brefeld, 2011b], performed a transductive step to extrapolate the partial annotations to the unlabeled part using the constrained Viterbi algorithm [Cao and Chen, 2003]. The perceptron-like algorithms have a clear advantage, namely, that they are an online type of algorithms suitable for learning from large scale data. The main disadvantage of such algorithms is an incomplete theoretical understanding like the convergence analysis (in turn it is unclear when to stop the algorithm) or a firm statistical justification [Fernandes and Brefeld, 2011b]. Nevertheless, most of the existing works show empirically that minimization of the partial loss function when learning the structured output classifiers from missing labels is a good heuristic.

2.3. Ordinal classification

First it should be mentioned that there is a plethora of works in machine learning community addressing supervised learning of ordinal classifiers. However, there is a lack of discriminative methods for learning from partially annotated examples. The existing approaches are briefly discussed below.

The ordinal classification models can be split to two groups: a regression based approaches and a threshold based approaches. The former approach involves the standard regression model the real-valued output of which is then projected on a discrete domain corresponding to the ordinal labels [Crammer and Singer, 2005]. On the other hand, the threshold based approaches provide a greater flexibility by seeking a mapping $f : \mathcal{X} \to \mathbb{R}$ along with a vector of non-decreasing thresholds which partition the real-valued prediction into an ordered set of labels. The existing learning paradigms can be split into two groups as well: the maximum likelihood based methods and the discriminative methods which are briefly outlined below.

2.3.1. Maximum likelihood methods for learning of ordinal classifier

A plug-in ordinal classifier can be constructed by substituting a probabilistic model estimated by the ML method 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 were also proposed [Chu and Ghahramani, 2005].

The maximum likelihood approach can be directly applied in the presence of incomplete annotation (e.g. the setting considered in this work when label interval is given instead of a single label) by using the Expectation-Maximization algorithms [Schlesinger, 1968; Dempster et al., 1997]. However, the maximum likelihood methods are sensitive to model mis-specification which complicates their application in modeling complex high dimensional data. In contrast, the discriminative methods reviewed below are known to be robust against the model misspecification while their extension for learning from partial annotations is not trivial.

2.3.2. Discriminative methods for supervised learning of ordinal classifiers

The existing discriminative methods learn parameters of the ordinal classifier by minimizing a 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: Explicit Constraints on Thresholds (SVOR-EXP) and the Support Vector Ordinal Regression: Implicit Constraints on Thresholds (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.

3. Learning ordinal classifiers from interval annotations

A road map of the chapter:

- **Classification models** for ordinal classification are discussed in Section 3.1. In Subsection 3.1.1 we show that the standard ordinal rule can be equivalently parametrized as a specific form of a linear multi-class classifier, which we denote as the MORD classifier. In Subsection 3.1.2 we propose a more flexible model for ordinal classification which we denoted as the PW-MORD classifier. A unified view of existing and proposed models for ordinal classification is presented in Subsection 3.1.3.
- **Supervised learning** algorithms for ordinal classification are discussed in Section 3.2. Two existing methods, the support vector ordinal machine with explicit constraints (SVOR-EXP) and the support vector ordinal machine with implicit constraints (SVOR-IMC) [Chu and Keerthi, 2005], are reviewed in Subsection 3.2.1 and Subsection 3.2.2, respectively. In Subsection 3.2.3 we show how to design an instance of the SO-SVM which can learn ordinal classifiers while minimizing a surrogate of an arbitrary V-shaped loss.
- **Learning from interval annotations** is discussed in Section 3.3. The interval-insensitive loss function and its application as an upper bound of the target (complete) loss is a subject of Subsection 3.3.1. Modifications of two existing supervised methods, the SVOR-EXP and the SVOR-IMC algorithms, in order to minimize a surrogate of the interval-insensitive loss is presented in Subsection 3.3.2 and Subsection 3.3.3, respectively. A generic surrogate of the interval-insensitive loss which can be derived from arbitrary V-shaped loss is proposed in Subsection 3.3.4. In Subsection 3.3.4 we also propose an instance of a generic ERM based algorithm, called V-shaped interval insensitive loss minimization algorithm (VILMA), and we discuss its relation to other methods.
- **Optimization algorithm** that is suitable for solving large instances of the convex problems formulated in this chapter is proposed in Subsection 3.4. The algorithm, denoted as a double-loop Cutting plane algorithm (CPA), can solve convex quadratically regularized risk minimization problems. In contrast to existing instances of the CPA, the double-loop CPA allows a subset of parameters not to be included in the quadratic regularizer which is very important when modeling the intercepts of the ordinal classification rules as will be shown experimentally.
- **Experiments** are presented in Section 3.5. The experiments provide a thorough evaluation of the supervised methods and the methods for learning from interval annotations discussed in this chapter. The evaluation is carried out on both standard UCI benchmarks as well as on a real-life problem the goal of which is the estimation of human age from facial images.

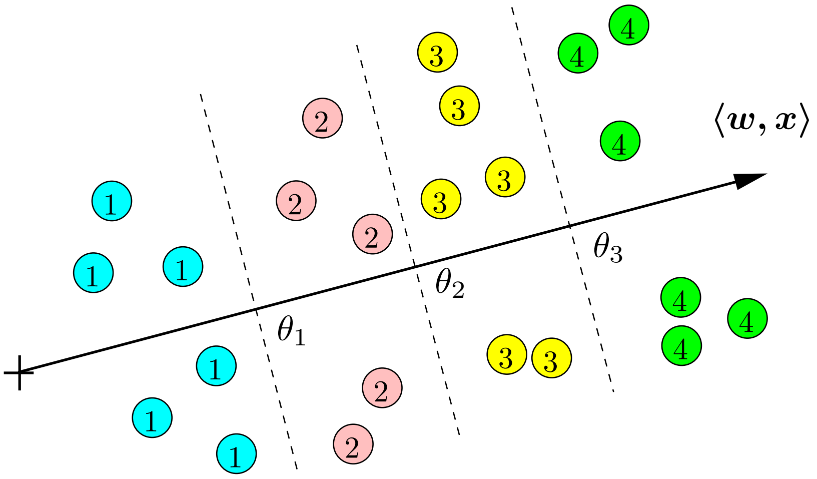


Figure 3.1. The figure vizualizes division of the 2-dimensional feature space into four classes realized by an instance of the ordinal classifier (3.1).

3.1. The model

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[1]. We consider learning of an ordinal classifier $h \colon \mathcal{X} \to \mathcal{Y}$ of the form

$$h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}) = 1 + \sum_{k=1}^{Y-1} [\![\langle \boldsymbol{x}, \boldsymbol{w} \rangle > \theta_k]\!] \,, \tag{3.1}$$

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 admissible parameters. The brackets $\langle \cdot, \cdot \rangle$ denote the dot product and the operator $[\![A]\!]$ is the Iverson bracket. It evaluates to 1 if A holds, otherwise it is 0. The classifier (3.1) splits the real line of projections $\langle \boldsymbol{x}, \boldsymbol{w} \rangle$ into Y consecutive intervals defined by thresholds $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_{Y-1}$. The observation $\boldsymbol{x}$ is assigned a label corresponding to the interval, to which the projection $\langle \boldsymbol{w}, \boldsymbol{x} \rangle$ falls to. The classifier (3.1) is a suitable 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 3.1.

We define an equivalent parametrisation of an ordinal classifier in the next section.

3.1.1. Ordinal regression as linear multi-class classification

Let us start with one-dimensional observations $x \in \mathcal{X} = \mathbb{R}$. In such case the ordinal classifier $h(x) = 1 + \sum_{k=1}^{Y-1} [\![x > \theta_k]\!]$ splits the real axis into Y intervals defined by thresholds $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_{Y-1}$. One may think of representing the ORD classifier in the form

$$h'(x) = \operatorname*{argmax}_{y \in \mathcal{Y}} f(x, y) \,, \tag{3.2}$$

where $f \colon \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ is a discriminant function. If we manage to construct the discriminant functions such that $f(x, y) > f(x, y')$, $y' \in \mathcal{Y} \setminus \{y\}$ iff $h(x) = y$ then both representations

[1] The sequence $1, \dots, Y$ is used just for a notational convenience. However, any other finite and fully ordered set can be used instead.

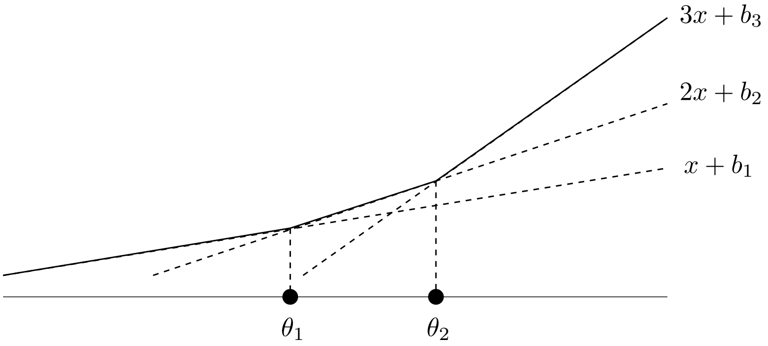


Figure 3.2. The figure illustrates the relation between the ordinal classifier $h(x) = 1 + \sum_{k=1}^{Y-1} [\![x > \theta_k]\!]$ and its alternative representation $h'(x) = \operatorname{argmax}_{y\in\mathcal{Y}}(x \cdot y + b_y)$ for the $(Y = 3)$-class problem. Note, that x and y-axes have different scale in order to save space.

will be equivalent i.e. $h'(x) = h(x)$, $x \in \mathbb{R}$. Let us consider a linear discriminant function with the slope equal to y, i.e. $f(x,y) = x \cdot y + b_y$. In such case (3.2) becomes a linear multi-class classifier. It is not difficult to see that such linear classifier also splits the real axis into intervals. Figure 3.2 shows an example of the ordinal classifier and its equivalent linear classifier $h'(x)$.

The same idea can be applied for n-dimensional observations $\boldsymbol{x} \in \mathcal{X} = \mathbb{R}^n$. The multi-class linear classifier which can represent the ordinal classifier (3.2) reads

$$h'(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{b}) = \operatorname*{argmax}_{y\in\mathcal{Y}} \Big(\langle \boldsymbol{x}, \boldsymbol{w} \rangle \cdot y + b_y \Big) , \tag{3.3}$$

where $\boldsymbol{w} \in \mathbb{R}^n$ is the parameter vector and $\boldsymbol{b} = (b_1, \ldots, b_Y) \in \mathbb{R}^Y$ is a vector of intercepts. We denote (3.3) as the Multi-class Ordinal classifier (MORD). Later in this text, we assume that the "argmax" operator returns the minimal label in the case of more than one maximizer.

A natural question is whether both representations are equivalent in the sense that any ordinal classifier can be represented by some MORD classifier and vice-versa. The following theorem gives the positive answer to the question.

Theorem 1. *The ordinal classifier (3.1) and the MORD classifier (3.3) 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 $h(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\theta}) = h'(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{b})$, $\forall \boldsymbol{x} \in \mathbb{R}^n$.*

Our proof (see Appendix A.1) is constructive in the sense that we can provide a conversion from the ordinal classifier to the MORD classifier and vice-versa.

In exotic cases, which however may appear in practice, some classes can collapse to a single point and effectively disappear. To cover all such situations, we first define the concept of non-degenerated classifier and then we give formulas for the conversions.

Definition 1 (Degenerated and non-degenerated classifier)**.** *We call class $y \in \mathcal{Y}$ non-degenerated for classifier $h'(\boldsymbol{x})$ iff $\mathcal{X}_y = \operatorname{interior}(\{\boldsymbol{x} \in \mathcal{X} : h'(\boldsymbol{x}) = y\}) \neq \emptyset$. Classifier $h'(\boldsymbol{x})$ is non-degenerated iff all classes are non-degenerated. In the opposite case, the classifier is called degenerated.*

Definition 2. *Given a MORD classifier, the class $\hat{y} \in \mathcal{Y}$ is non-degenerated iff the linear inequalities*

$$\begin{aligned} z\hat{y} + b_{\hat{y}} &> z(\hat{y} - k) + b_{\hat{y}-k},\ 1 \leq k < \hat{y}\,, \\ z\hat{y} + b_{\hat{y}} &\geq z(\hat{y} + t) + b_{\hat{y}+k},\ 1 < t \leq Y - \hat{y}\,, \end{aligned} \tag{3.4}$$

are solvable w.r.t. $z \in \mathbb{R}$.

Note that the validity of (3.4) can be verified in $\mathcal{O}(Y)$ time. The proof of Theorem 1 can be found in Appendix A.1. The proof is a constructive, i.e., it provides formulas which allow to convert the MORD classifier (3.3) to the standard ordinal classifier (3.1) and vice-versa.

Conversion formulas. Given parameters of the ordinal classifier $\boldsymbol{w} \in \mathbb{R}^n$, $\boldsymbol{\theta} \in \Theta$, the equivalent MORD classifier has parameters $\boldsymbol{w}$ and $\boldsymbol{b}$ given by

$$b_1 = 0 \qquad \text{and} \qquad b_y = -\sum_{i=1}^{y-1} \theta_i,\ y = 2, \dots, Y. \tag{3.5}$$

The conversion from the MORD classifier to the ordinal classifier is done differently for the non-generated and the degenerated classifier. Given parameters of a non-degenerated MORD classifier $\boldsymbol{w} \in \mathbb{R}^n$ and $\boldsymbol{b} \in \mathbb{R}^Y$, we can compute thresholds $\boldsymbol{\theta} \in \Theta$ of the equivalent ordinal classifier by

$$\theta_y = b_y - b_{y+1}, \qquad y = 1, \dots, Y-1\,. \tag{3.6}$$

Given parameters of a degenerated MORD classifier $\boldsymbol{w} \in \mathbb{R}^n$ and $\boldsymbol{b} \in \mathbb{R}^Y$, we compute thresholds $\boldsymbol{\theta} \in \Theta$ of the equivalent ORD classifier by

$$\theta_{y_i} = \dots = \theta_{y_{i+1}-1} = \tfrac{b_{y_i} - b_{y_{i+1}}}{(y_{i+1} - y_i)},\ i = 1, \dots, p, \tag{3.7}$$

where $y_i \in \mathcal{Y},\ i = 1, \dots, p$ is an increasing subsequence of non-degenerated classes.

Finally, let us note that the MORD classifier is represented by $n + Y$ parameters insted of $n + Y - 1$ parameters of the ordinal classifier. However, the parameters of the MORD classifier are unconstrained, which makes the MORD representation attractive for learning because no additional constraints on the intercepts $\boldsymbol{\theta} \in \Theta$ are needed.

3.1.2. Piece-wise ordinal regression classifier

The discriminative power of the ordinal classifier can be limiting in some cases. Mapping the observations into higher dimensional space via usage of kernel functions is one way to make the linear ordinal classifier more discriminative. Though the "kernalization" of the ordinal classifier is straightforward it is not suitable in all cases. For example, the kernels are prohibitive in applications, which require processing of large amounts of training examples and/or if a real-time response of the classifier is the must. Instead, we proposed to stay in the original feature space where we construct a combined classifier from a set of simpler component classifiers. In our case, the component classifiers will be the MORD classifiers, each responsible for a subset of labels.

Let $Z > 1$ be a number of cut labels $(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_Z) \in \mathcal{Y}^Z$ such that $\hat{y}_1 = 1$, $\hat{y}_Z = Y$ and $\hat{y}_z \leq \hat{y}_{z+1}$, $z \in \mathcal{Z} = \{1, \dots, Z-1\}$. The cut labels define a partitioning of $\mathcal{Y}$ into Z subsets

$\mathcal{Y}_z = \{y \in \mathcal{Y} \mid \hat{y}_z \leq y \leq \hat{y}_{z+1}\}$, $z \in \mathcal{Z}$. We will model a dependence between the observation $\boldsymbol{x}$ and a subset of labels $\mathcal{Y}_z$ by the component classifier

$$h_z(\boldsymbol{x}) = \operatorname*{argmax}_{y \in \mathcal{Y}_z} f_z(\boldsymbol{x}, y) \,, \tag{3.8}$$

where $f_z \colon \mathbb{R}^n \times \mathcal{Y}_z \to \mathbb{R}$ is a discriminant function. We define a combined classifier whose discriminant function is composed of discriminant functions of the component classifiers as follows

$$h''(\boldsymbol{x}) = \operatorname*{argmax}_{z \in \mathcal{Z}} \max_{y \in \mathcal{Y}_z} f_z(\boldsymbol{x}, y) \,. \tag{3.9}$$

We set the discriminant functions to be

$$f_z(\boldsymbol{x}, y) = \big\langle \boldsymbol{x}, \boldsymbol{w}_z(1 - \alpha(y,z)) + \boldsymbol{w}_{z+1}\alpha(y,z) \big\rangle + b_y \,, \tag{3.10}$$

where

$$\alpha(y,z) = \frac{y - \hat{y}_z}{\hat{y}_{z+1} - \hat{y}_z}$$

and $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_Z] \in \mathbb{R}^m$, $\boldsymbol{b} \in \mathbb{R}^Y$, (where $m = n \times Z$) are parameters. With these definitions, it can be claimed that:

1. the component classifiers (3.8) are the ordinal classifiers,
2. the combined classifier (3.9) is well defined because all its neighboring discriminant functions are consistent at the cut labels, i.e. $f_z(\boldsymbol{x}, \hat{y}_{z+1}) = f_{z+1}(\boldsymbol{x}, \hat{y}_{z+1})$, $z \in \mathcal{Z}$, holds.

The claim 1 is seen after substituting (3.10) into (3.8), which after some algebra yields

$$h_z(\boldsymbol{x}) = \operatorname*{argmax}_{y \in \mathcal{Y}_z} \Big(\langle \boldsymbol{x}, \boldsymbol{w}_{z+1} - \boldsymbol{w}_z \rangle \alpha(y,z) + b_y \Big) \,.$$

Since $\alpha(y,z)$ is linearly increasing with y, Theorem 1 guarantees that $h_z(\boldsymbol{x})$ is the MORD classifier equivalent to the ordinal classifier. The claim 2 follows from the fact that $\alpha(\hat{y}_{z+1}, z) = 1$ and $\alpha(\hat{y}_{z+1}, z+1) = 0$, and thus $f_z(\boldsymbol{x}, \hat{y}_{z+1}) = \langle \boldsymbol{x}, \boldsymbol{w}_{z+1} \rangle + b_{\hat{y}_{z+1}} = f_{z+1}(\boldsymbol{x}, \hat{y}_{z+1})$.

We can write explicitly the component classifier, which we call the PW-MORD, as follows

$$h''(\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{b}) = \operatorname*{argmax}_{z \in \mathcal{Z}} \operatorname*{argmax}_{y \in \mathcal{Y}_Z} \Big(\langle \boldsymbol{x}, \boldsymbol{w}_z(1 - \alpha(y,z)) + \boldsymbol{w}_{z+1}\alpha(y,z) \rangle + b_y \Big) \,. \tag{3.11}$$

Figure 3.3 visualizes the ordinal (=MORD) and the PW-MORD classifier on a toy data. It is seen that the distribution of the data cannot be well described by the ordinal classifier, while the PW-MORD composed of three ordinal classifiers provides much better model in this case.

3.1.3. Unified view of classifiers for ordinal regression

In this section, we are going to describe several instances of the classifier

$$h(\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{b}) = \operatorname*{argmax}_{y \in \mathcal{Y}} \Big(\langle \boldsymbol{x}, \sum_{z=1}^{Z} \beta(y,z) \boldsymbol{w}_z \rangle + b_y \Big) \,, \tag{3.12}$$

where $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_Z] \in \mathbb{R}^{n \times Z}$, $\boldsymbol{b} = [b_1; \ldots; b_Y] \in \mathbb{R}^Y$ are parameters and $\beta \colon \mathcal{Y} \times \{1, \ldots, Z\} \to \mathbb{R}$ are fixed numbers, that can be useful models for ordinal regression. The instances of (3.12) differ in the way how one defines β and $\mathcal{Z}$. We show below how to derive various instances of the ordinal classifier.

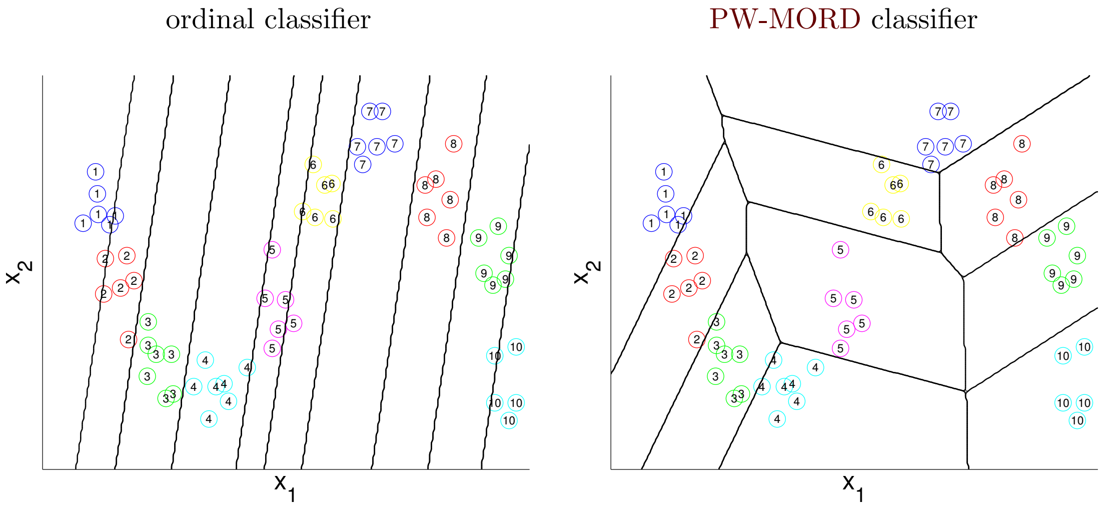


Figure 3.3. The figure shows the partitioning of 2-dimensional feature space realized by the ordinal classifier and the PW-MORD classifier with $Z = 3$ components. The cut labels for the PW-MORD classifier were set to $\{1, 4, 7, 10\}$.

1. Rounded linear-regression rule

$$h(\boldsymbol{x}, \boldsymbol{w}, b) = \max(1, \min(Y, \operatorname{round}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b))) \tag{3.13}$$

is the most simplest model for the ordinal regression obtained by clipping a rounded response of the standard linear regression rule to the interval $[1, Y]$. It is easy to show that (3.13) is an instance of (3.12) recovered after setting $Z = 1$, $\beta(1, y) = 2y$, $y \in \mathcal{Y}$, and fixing the components of the intercept vector $\boldsymbol{b}$ to $b_y = 2by - y^2$. Using the conversion formula (3.6), we can show that the rounded linear-regression rule is equivalent to the ordinal classifier with equal width of the decision intervals, namely, with $\theta_{k+1} - \theta_k = 2$, $k = 1, \ldots, Y - 2$.

2. Multi-class linear classifier

$$h(\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{b}) = \operatorname*{argmax}_{y \in \mathcal{Y}} \big(\langle \boldsymbol{w}_y, \boldsymbol{x} \rangle + b_y \big) \tag{3.14}$$

is recovered after setting $Z = Y$ and $\beta(y, z) = [\![y = z]\!]$, $y \in \mathcal{Y}$, $z \in \{1, \ldots, Z\}$. It is the most generic (and also most discriminative) form of (3.12), which completely ignores ordering of the labels.

3. The proposed MORD classifier (3.3) is recovered after setting $Z = 1$, $\boldsymbol{W} = \boldsymbol{w}_1$, and $\beta(y, 1) = y$, $y \in \mathcal{Y}$. We showed that the MORD classifier is equivalent to the standard ordinal classifier (3.1) most frequently used in the ordinal regression.

4. The proposed PW-MORD classifier (3.11) is recovered after setting $\beta(y, z)$ according to

$$\begin{array}{rcll} \beta(y,z) & = & 1 - \alpha(y,z) & \text{for} \quad z = 1, \ldots, Z-1\,, y \in \mathcal{Y}_z\,, \\ \beta(y,z) & = & \alpha(y,z-1) & \text{for} \quad z = 2, \ldots, Z\,, y \in \mathcal{Y}_z\,, \\ \beta(y,z) & = & 0 & \text{otherwise.} \end{array} \tag{3.15}$$

The PW-MORD is composed of $Z-1$ MORD classifiers each modeling a subset of labels (see Section 3.1.2). The PW-MORD is most flexible as it allows controling its complexity smoothly by a single parameter Z. It is easy to see that for $Z=2$ the PW-MORD is equivalent to the MORD (=ordinal) classifier. For $Z=Y$, it becomes the multi-class linear classifier.

To summarize, one can see PW-MORD classifier as a classifier whose discriminative power varies from MORD classifier ($Z=1$, labels are fully ordered) to multi-class linear classifier (3.14) ($Z=Y$, no order of labels) depending on the number cutting labels Z.

3.2. Supervised learning

There exist several discriminative methods for learning parameters $(\boldsymbol{w},\boldsymbol{\theta})$ of the ordinal classifier (3.1) 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 requiring a set of completely annotated training examples

$$\mathcal{D}^m_{xy} = \{(\boldsymbol{x}^1,y^1),\ldots,(\boldsymbol{x}^m,y^m)\} \in (\mathcal{X}\times\mathcal{Y})^m \tag{3.16}$$

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 the loss function $\ell\colon \mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$ and the training examples (3.16), the task is to learn the ordinal classifier $h\colon \mathcal{X}\to\mathcal{Y}$ with the *Bayes risk*

$$R^\ell(h) = \mathbb{E}_{p(\boldsymbol{x},y)}\,\ell(y,h(\boldsymbol{x})) \tag{3.17}$$

is as small as possible

$$h^\ell_* \in \operatorname*{Argmin}_{h\colon\mathcal{X}\to\mathcal{Y}} R(h)\,. \tag{3.18}$$

The loss functions most commonly used in practice for ordinal classification are the Mean Absolute Error (MAE) $\ell^{\text{MAE}}(y,y')=|y-y'|$ and the 0/1-loss $\ell^{0/1}(y,y')=[\![y\neq y']\!]$. Both MAE and 0/1-loss are instances of so called V-shaped losses.

Definition 3. *(V-shaped loss). A loss $\ell\colon\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$ is V-shaped if $\ell(y,y)=0$ and $\ell(y'',y)\geq \ell(y',y)$ 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. We constrain our analysis to the V-shaped losses.

Because the expected risk $R^\ell(h)$ 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 over a set of linear decision functions. In particular, the existing methods approximate the Bayes risk minimization (3.18) by a surrogate ERM problem

$$f^{\text{emp}}_* \in \operatorname*{Argmin}_{f\in\mathcal{F}} R^\psi_{\text{emp}}(f)\,, \tag{3.19}$$

where

$$R^\psi_{\text{emp}}(f) = \mathbb{E}_{s(\boldsymbol{x},y)}\psi(y,f(\boldsymbol{x})) \tag{3.20}$$

is the empirical risk and the resulting classification rule is $h = \text{pred} \circ f_*^{\text{emp}}$. In the case of ordinal classification, the space of decision functions is defined as

$$\mathcal{F} = \left\{ \boldsymbol{f}(\boldsymbol{x}) = (\langle \boldsymbol{w}, \boldsymbol{x} \rangle - \theta_1, \dots, \langle \boldsymbol{w}, \boldsymbol{x} \rangle - \theta_{Y-1})^T \in \mathbb{R}^{Y-1} \mid \boldsymbol{w} \in \mathbb{R}^n, \boldsymbol{\theta} \in \Theta, \Omega(\boldsymbol{w}, \boldsymbol{\theta}) \leq r \right\},$$

where $r > 0$ is a hyper-parameter and $\Omega \colon \mathbb{R}^n \times \mathbb{R}^{Y-1} \to \mathbb{R}_+$ is a convex regularization function. The form of the prediction transform pred is defined implicitly by (3.1). Because the decision function $\boldsymbol{f} \in \mathcal{F}$ is parametrized by $(\boldsymbol{w}, \boldsymbol{\theta})$ we will use a shortcut $\psi(\boldsymbol{w}, \boldsymbol{\theta}; \boldsymbol{x}, y) = \psi(y, f(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}))$ which slightly abuses the notation but should not cause a big confusion. In practice it is more convenient to solve a problem

$$(\boldsymbol{w}^*, \boldsymbol{\theta}^*) \in \underset{\boldsymbol{w} \in \mathbb{R}^n, \boldsymbol{\theta} \in \Theta}{\text{Argmin}} \left(\frac{\lambda}{2} \Omega(\boldsymbol{w}, \boldsymbol{\theta}) + \frac{1}{m} \sum_{i=1}^{m} \psi(\boldsymbol{w}, \boldsymbol{\theta}; \boldsymbol{x}^i, y^i) \right), \tag{3.21}$$

which is however equivalent to (3.19) with appropriately set regularization constant $\lambda > 0$.

In Section 3.2.1 and Section 3.2.2 we review the most polular methods, i.e. the SVOR-EXP algorithm and the SVOR-IMC algorithm [Chu and Keerthi, 2005], respectively. We will show that both algorithms are instances of (3.21) using a different surrogate loss ψ. We show that the surrogate of the SVOR-EXP is an upper bound of the 0/1-loss and that the surrogate of the SVOR-IMC is an upper bound of the MAE loss.

In Section 3.2.3 we derive an instance of the SO-SVM algorithm suitable for learning parameters of the MORD and the PW-MORD rules. In contrast to the SVOR-EXP and SVOR-IMC, the SO-SVM based algorithm uses a generic surrogate loss ψ which can approximate arbitrary target loss function. Another advantage of the SO-SVM algorithm is that it leads to an unconstrained variant of the problem (3.21). Therefore larger set of optimization solvers can be used in contrast to SVOR-EXP and SVOR-IMC which have to deal with the constraints.

A final remark is related to the regularization function Ω. A common choice in the case of ordinal classifier is either $\Omega(\boldsymbol{w}, \boldsymbol{\theta}) = \|\boldsymbol{w}\|^2 + \|\boldsymbol{\theta}\|^2$ or $\Omega(\boldsymbol{w}, \boldsymbol{\theta}) = \|\boldsymbol{w}\|^2$. The former regularizer makes the objective to be smooth, to have a unique minimizer and easier to deal with in general. However, an influence of the regularizer on the classification accuracy is unclear. In Section 3.4 we develop a generic optimization algorithm which can deal with both variants of the regularization function. In Section 3.5 we compare both variants empirically and show that the choice of the regularizer has a significant impact on the overall classifier accuracy.

3.2.1. Support vector ordinal regression: explicit constraints on thresholds

The original SVOR-EXP algorithm [Chu and Keerthi, 2005] considers only adjacent categories. In particular, it learns parameters of the ordinal classifier (3.1) from completely annotated examples $\mathcal{D}_{xy}^m$ by solving the following convex quadratic optimization problem

$$(\boldsymbol{w}^*, \boldsymbol{\theta}^*) \in \underset{\boldsymbol{w} \in \mathbb{R}^n, \boldsymbol{\theta} \in \mathbb{R}^{Y-1}}{\text{Argmin}} \left[\frac{\lambda}{2} ||\boldsymbol{w}||^2 + \sum_{j=1}^{r-1} \left(\sum_{i=1}^{n^j} \xi_i^j + \sum_{i=1}^{n^{j+1}} \xi_i^{*j+1} \right) \right] \tag{3.22}$$

subject to

$$\begin{aligned} \langle \boldsymbol{x}_i^j, \boldsymbol{w} \rangle - \theta_j \leq -1 + \xi_i^j,\ \xi_i^j \geq 0,\ \forall i = 1, \dots, n^j\,, \\ \langle \boldsymbol{x}_i^{j+1}, \boldsymbol{w} \rangle - \theta_j \geq 1 - \xi_i^{*j+1},\ \xi_i^{*j+1} \geq 0,\ \forall i = 1, \dots, n^{j+1}\,, \\ \theta_j \leq \theta_{j+1},\ \forall j \in \{1, \dots, Y-1\}\,. \end{aligned}$$

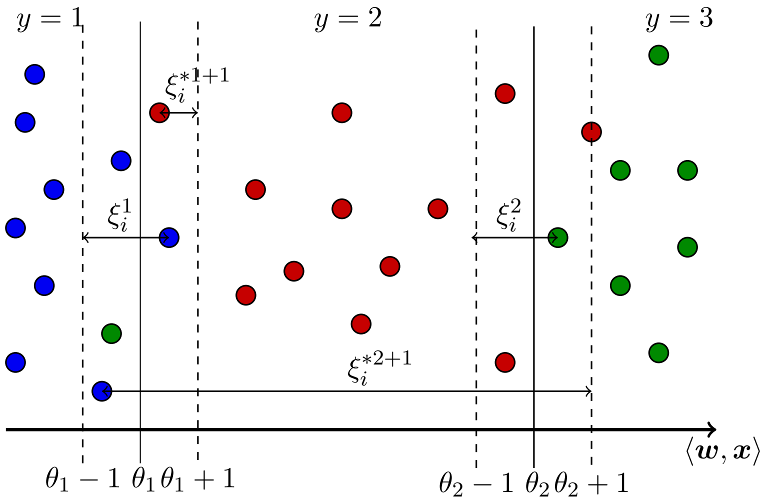


Figure 3.4. The figure explains the meaning of slack variables ξ and ξ^* for the SVOR-EXP formulation. Note, example $y+1$ can be counted twice if its projection falls into segment $[\theta_{j+1}-1, \theta_j+1]$, where $(\theta_{j+1}-1 < \theta_j+1)$. The idea of the figure taken from [Chu and Keerthi, 2005].

In this setting, the support vector formulation attempts to find the optimal mapping direction $\boldsymbol{w}$ and thresholds $\boldsymbol{w} \in \mathbb{R}^n$ and $\boldsymbol{\theta} \in \Theta = \{\boldsymbol{\theta'} \in \mathbb{R}^{Y-1} \mid \theta'_y \leq \theta'_{y+1},\ y = 1, \ldots, Y-1\}$, which define $Y-1$ parallel discriminative hyperplanes for Y ordered classes accordingly. In this formulation, each sample in the y-th category should have a function value that is less than the lower margin $\theta_y - 1$, otherwise $\langle \boldsymbol{x}_i^y, \boldsymbol{w} \rangle - (\theta_y - 1)$ is the error (denoted as ξ_i^y). Similarly, each sample from $(y+1)$-th category should have a function value that is greater than the upper margin $\theta_y + 1$, otherwise $(\theta_y + 1) - \langle \boldsymbol{x}_i^{y+1}, \boldsymbol{w} \rangle$ is the error (denoted as ξ_i^{*y}). See Figure 3.4 to get more insight to meaning of ξ_i^y and ξ_i^{*y}.

Using auxiliary variables $\theta_0 = -\infty$ and $\theta_Y = \infty$, we reformulate (3.22) as an equivalent problem in terms of ERM framework as follows

$$(\boldsymbol{w}^*, \boldsymbol{\theta}^*) \in \operatorname*{Argmin}_{\boldsymbol{w} \in \mathbb{R}^n, \boldsymbol{\theta} \in \hat{\Theta}} \left[\frac{\lambda}{2} \|\boldsymbol{w}\|^2 + \sum_{i=1}^m \psi^{\text{EXP}}(\boldsymbol{w}, \boldsymbol{\theta}; \boldsymbol{x}^i, y^i) \right], \tag{3.23}$$

where the optimized convex surrogate loss reads

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

and $\hat{\Theta} = \{\boldsymbol{\theta} \in \mathbb{R}^{Y+1} \mid \theta_0 = -\infty, \theta_Y = \infty, \theta_y \leq \theta_{y+1}, y = 1, \ldots, Y-1\}$. Note, that the surrogate $\psi^{\text{EXP}}(\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{x}, y)$ is a convex upper bound of the 0/1-loss

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

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

3.2.2. Support vector ordinal regression: implicit constraints on thresholds

Instead of considering errors only from the samples of adjacent categories in SVOR-EXP, the SVOR-IMC algorithm allows the samples in all categories to contribute errors for each threshold. That is to say, SVOR-IMC algorithm learns parameters of the ordinal classifier (3.1) from completely annotated examples $\mathcal{D}^m_{\boldsymbol{xy}}$ by solving the following quadratic optimization problem

$$(\boldsymbol{w}^*,\boldsymbol{\theta}^*) \in \operatorname*{Argmin}_{\boldsymbol{w}\in\mathbb{R}^n,\boldsymbol{\theta}\in\mathbb{R}^{Y-1}} \left[\frac{\lambda}{2}||\boldsymbol{w}||^2 + \sum_{j=1}^{r-1}\Big(\sum_{k=1}^{j}\sum_{i=1}^{n^k}\xi^j_{ki} + \sum_{k=j+1}^{r}\sum_{i=1}^{n^k}\xi^{*j}_{ki}\Big)\right] \tag{3.24}$$

subject to

$$\begin{aligned}&\langle \boldsymbol{x}^k_i,\boldsymbol{w}\rangle - \theta_j \le -1 + \xi^j_{ki},\ \xi^j_{ki}\ge 0,\ k=1,\dots,j,\ i=1,\dots,n^k\,,\\ &\langle \boldsymbol{x}^k_i,\boldsymbol{w}\rangle - \theta_j \ge 1 - \xi^{*j}_{ki},\ \xi^{*j}_{ki}\ge 0,\ \ k=j+1,\dots,r,\ i=1,\dots,n^k\,.\end{aligned}$$

The authors of [Chu and Keerthi, 2005; Li and Lin, 2006] proved that the optimal parameters are admissible, i.e. $(\boldsymbol{w}^*,\boldsymbol{\theta}^*)\in(\mathbb{R}^n,\Theta)$ holds, hence the explicit constraints $\boldsymbol{\theta}\in\Theta$ are not needed in this case. It is also shown that the sum of slack variables in (3.24) upper bounds the average of the MAE loss $\ell(y,y')=|y-y'|$ computed on the training examples. We reformulate (3.24) as an equivalent unconstrained minimization problem in terms of SO-SVM framework as follows

$$(\boldsymbol{w}^*,\boldsymbol{\theta}^*) \in \operatorname*{Argmin}_{\boldsymbol{w}\in\mathbb{R}^n,\boldsymbol{\theta}\in\hat{\Theta}} \left[\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{m}\psi^{\mathrm{IMC}}(\boldsymbol{w},\boldsymbol{\theta};\boldsymbol{x}^i,y^i)\right] \tag{3.25}$$

where the convex surrogate reads

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

As in the previous case, the problem (3.25) is an equivalent reformulation of the quadratic program defining the SVOR-IMC algorithm in [Chu and Keerthi, 2005]. It is seen that the surrogate $\psi^{\mathrm{IMC}}(\boldsymbol{w},\boldsymbol{\theta};\boldsymbol{x},y)$ is a convex upper bound of the MAE loss

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

3.2.3. Generic learning algorithm for ordinal regression

In Section 3.1.3 we showed that various models for ordinal classification can be seen as a special instances of linear classifier (3.12). In this section, we derive generic algorithm to learn (3.12) from given fully-supervised set $\mathcal{D}^m_{\boldsymbol{xy}}$ via SO-SVM framework. It is a generic and well understood framework originally developed for the structured output learning [Tsochantaridis et al., 2005].

Following [Tsochantaridis et al., 2005], we propose to approximate the empirical risk by

$$R(\boldsymbol{W},\boldsymbol{b}) = \frac{1}{m}\sum_{i=1}^{m}\max_{y\in\mathcal{Y}}\Big[\ell(y,y^i) + \langle\boldsymbol{x}^i,\sum_{z\in\mathcal{Z}}\beta(y,z)\boldsymbol{w}_z\rangle(y-y^i) + b_y - b_{y^i}\Big]\,. \tag{3.26}$$

This risk approximation uses the idea of the margin-rescaling loss functions [Tsochantaridis et al., 2005] applied to the classifier (3.12). It is easy to prove that $R(\boldsymbol{W},\boldsymbol{b})$ is a convex upper bound on the true empirical risk

$$R_{\mathrm{emp}}(\boldsymbol{W},\boldsymbol{b}) = \frac{1}{m}\sum_{i=1}^{m}\ell(y^i,h(\boldsymbol{x}^i,\boldsymbol{W},\boldsymbol{b}))$$

simply by showing that

$$\psi(\boldsymbol{w},\boldsymbol{b};\boldsymbol{x},y) = \max_{\hat{y}\in\mathcal{Y}} \Big[\ell(\hat{y},y) + \big\langle \boldsymbol{x}, \sum_{z\in\mathcal{Z}} \beta(\hat{y},z)\boldsymbol{w}_z\big\rangle(\hat{y}-y) + b_{\hat{y}} - b_y\Big] \tag{3.27}$$

is a convex upper bound on $\ell(\hat{y},y)$. We can formulate learning of the classifier (3.12) as the following convex unconstrained minimization problem

$$(\boldsymbol{W}^*,\boldsymbol{b}^*) \in \operatorname*{Argmin}_{\boldsymbol{W}\in\mathbb{R}^n,\boldsymbol{b}\in\mathbb{R}^Y} \left[\frac{\lambda}{2}\Omega(\boldsymbol{W},\boldsymbol{b}) + R(\boldsymbol{W},\boldsymbol{b})\right], \tag{3.28}$$

where $\Omega(\boldsymbol{W},\boldsymbol{b})$ is typically $\|\boldsymbol{W}\|^2$ or $\|\boldsymbol{W}\|^2 + \|\boldsymbol{b}\|^2$ and $\lambda > 0$ is a prescribed (regularization) constant used to control over-fitting.

A big effort has been put by the machine learning community into development of efficient solvers for the problem (3.28). For example, a generic cutting plane methods like the BMRM [Teo et al., 2010] or its accelerated variant [Franc and Sonneburg, 2009] can be readily applied to solve (3.28). However, the existing cutting methods require the regularizer $\Omega(\boldsymbol{W},\boldsymbol{b}) = \|\boldsymbol{W}\|^2 + \|\boldsymbol{b}\|^2$. In Section (3.4), we propose a generic solver of (3.28) able to deal with both regularizers.

Let us compare SO-SVM framework with the existing algorithms for learning the ordinal classifier. First, the SO-SVM formulation (3.28) can learn a generic rule (3.12) while the existing methods are tailored to the canonical form (3.1) only. Second, the existing algorithms consider a limited set of loss functions $\ell(y,y')$, namely MAE and 0/1-loss. The most generic approach of [Li and Lin, 2006] derives an upper bound for V-shaped losses. The third limitation of the existing algorithms is that they have to care about feasibility of thresholds $\boldsymbol{\theta} \in \Theta$ because they work directly on the parameters of the ordinal classifier (however, it does not apply to SVOR-IMC). This requires to either introduce additional constraints on the thresholds $\boldsymbol{\theta} \in \Theta$ or to impose additional constraints on the loss function, namely, that the loss must be convex [Li and Lin, 2006]. For instance, the 0/1-loss is not convex hence the learning algorithms require extra inequality constraints (like the SVOR-EXP algorithm of [Chu and Keerthi, 2005]), which may complicate the optimization. Note that in the proposed approach the problem (3.28) remains unconstrained irrespectively to the selected loss.

The generality of our framework, however, does not automatically imply that the risk approximation (3.26) is better (tighter) than those used in existing methods. We experimentally show in Section 3.5.4 that in the case of the most frequently used MAE loss, the proposed approximation (3.26) provides a slightly but consistently better test accuracy than the existing ones.

Now we are ready to formulate learning of the ordinal classifiers from partially annotated examples, namely, from interval annotations of the labels.

3.3. Learning from interval annotations

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 $p(\boldsymbol{x},y)$. In contrast to the supervised setting, the training set does 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 the 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 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

$$\mathcal{D}^m_{xI} = \{(\boldsymbol{x}^1, [y^1_l, y^1_r]), \ldots, (\boldsymbol{x}^m, [y^m_l, y^m_r])\} \in (\mathcal{X} \times \mathcal{P})^m \tag{3.29}$$

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 $p(y_l, y_r \mid \boldsymbol{x}, y)$ can not 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 process 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 $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and partially annotated examples (3.29), the task is to learn the ordinal classifier (3.1) whose Bayes risk $R^\ell(h)$ defined by (3.17) is as small as possible. Note that the objective remains the same as in the supervised setting but the information about the labels contained in the training set is reduced to intervals.

3.3.1. Learning by minimizing the 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 MORD classifier $h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}) \in \mathcal{Y}$ defined by (3.3).

Definition 4. *(Interval insensitive loss) Let $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a supervised V-shaped loss. The interval insensitive loss $\ell_I\colon \mathcal{P} \times \mathcal{Y} \to \mathbb{R}$ associated with ℓ is defined as*

$$\ell_I(y_l, y_r, y) = \min_{y' \in [y_l, y_r]} \ell(y', y) = \begin{cases} 0 & \textit{if} \quad y \in [y_l, y_r]\ , \\ \ell(y, y_l) & \textit{if} \quad y \leq y_l\ , \\ \ell(y, y_r) & \textit{if} \quad y \geq y_r\ . \end{cases} \tag{3.30}$$

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

The interval-insensitive loss (3.30) is a special case of the generic partial loss (1.13) that has been previously used in the context of different classification models like the 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 [Luo and Orabona, 2010], etc. However, as it will be shown later the ordinal classification model allows for a tight convex approximations of the partial loss in contrast to previously considered classification models which either require crude approximation or more frequently a non-convex loss function which is then hard to optimize.

Having defined the interval-insensitive loss, we can approximate minimization of the Bayes risk $R^\ell(h)$ defined in (3.17) by minimization of the expectation of the interval-insensitive loss

$$R_I^\ell(h) = \mathbb{E}_{p(\boldsymbol{x},y_l,y_r)}\, \ell_I(y_l, y_r, h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}))\,. \tag{3.31}$$

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

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

Definition 5. *(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 when $\mathcal{P}$ is not constrained to be a 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* ε. If adopted to our interval-setting, is defined as

$$\varepsilon = \max_{\boldsymbol{x},y,z\neq y} p(z \in [y_l, y_r] \mid \boldsymbol{x}, y) = \max_{\boldsymbol{x},y,z} \sum_{[y_l,y_r]\in\mathcal{P}} [\![y_l \le z \le y_r]\!]\, p(y_l, y_r \mid \boldsymbol{x}, y)\,. \tag{3.32}$$

In words, the ambiguity degree ε 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 2. *Let $p(y_l, y_r \mid \boldsymbol{x}, y)$ be a distribution describing a consistent annotation process with the ambiguity degree ε defined by (3.32). Let $R^{0/1}(h)$ be the Bayes risk (3.17) instantiated for the 0/1-loss and let $R_I^{0/1}(h)$ be the partial risk (3.31) instantiated for the interval insensitive loss associated to the 0/1-loss. Then the upper bound*

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

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

Theorem 2 is a direct application of Proposition 1 from [Cour et al., 2011].

Next we introduce a novel upper bound for the MAE loss, which is more frequently used in applications of the ordinal classifier. We again consider consistent annotation processes. We characterize the annotation process by two numbers describing the amount of uncertainty in the training data. First, we use $\alpha \in [0, 1]$ to denoted a lower bound of the portion of exactly annotated examples, that is, examples annotated by an interval having just a single label $[y_l, y_r]$, $y_l = y_r$. 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 6. *($\alpha\beta$-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*

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

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

Let us consider the extreme cases, to illustrate the meaning of the parameters α and β. If $\beta = 0$ or $\alpha = 1$ then all examples are annotated exactly. We are back in the standard supervised setting. On the other hand, if $\beta = Y - 1$ and $\alpha = 0$ then it may happen that the annotation brings no information about the hidden label because the intervals can contain all labels in $\mathcal{Y}$. With the definition of $\alpha\beta$-precise annotation, we can upper bound the Bayes risk in terms of the partial risk as follows:

Theorem 3. *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 (3.17) instantiated for the MAE-loss and let $R_I^{MAE}(h)$ be the partial risk (3.31) instantiated for the interval insensitive loss associated to the MAE-loss. Then the upper bound*

$$R^{MAE}(h) \leq R_I^{MAE}(h) + (1 - \alpha)\beta \tag{3.33}$$

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

Proof of Theorem 3 is deferred to Appendix A.2.

The bound (3.33) is obtained by the worst case analysis hence it may become trivial in some cases. For example, if all examples are annotated with wide intervals because then $\alpha = 0$ and β is large. The experimental study presented in Section 3.5 nevertheless shows that the partial risk R_I is a good proxy even in cases when the upper bound is large. This suggests that better bounds might be derived, for example, when additional information about $p(y_l, y_r \mid \boldsymbol{x}, y)$ is available.

In order to improve the performance of the resulting classifier via the bound (3.33), one needs to control the parameters α and β. A possible way, which allows to set the parameters (α, β) 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 the probability α 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 not larger than $\beta + 1$ labels. That means that approximately $m \cdot \alpha$ inputs will be annotated exactly and $m \cdot (1 - \alpha)$ inputs 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.

Above we argued that the partial risk defined as an expectation of the interval insensitive loss was a reasonable proxy of the target Bayes risk. In next section, we design algorithms learning the ordinal classifier via minimization of the quadratically regularized empirical risk used as a proxy for the expected risk. Similarly to the standard 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 translate the risk minimization to tractable convex problems.

We first show how to modify two existing supervised methods in order to learn from partially annotated examples. Namely, we extend the SVOR-EXP and SVOR-IMC algorithms. The extended interval-insensitive variants are named Interval-Insensitive SVOR-EXP (II-SVOR-EXP) (section 3.3.2) and Interval-Insensitive SVOR-IMC (II-SVOR-IMC) (section 3.3.3), 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 minimization of MAE loss.

In section 3.3.4, we show how to construct a generic convex surrogate of the interval-insensitive loss associated to an arbitrary V-shaped loss. We call a method minimizing this generic surrogate as the VILMA. We prove that the VILMA subsumes the II-SVOR-IMC (as well as the SVOR-IMC as a special case).

3.3.2. Interval insensitive support vector ordinal regression: explicit constraints on thresholds

The interval insensitive loss $\ell_I^{0/1}(y_l, y_r, y)$ derived for the target 0/1-loss reads

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

We derive its surrogate by replacing the step functions with the hinge loss which yileds

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

The surrogate $\psi_I^{\rm EXP}(\boldsymbol{w},\boldsymbol{\theta};\boldsymbol{x},y_l,y_r)$ is clearly a convex upper bound of $\ell_I^{0/1}(y_l,y_r,h(\boldsymbol{x};\boldsymbol{w},\boldsymbol{\theta}))$ as can be also seen in Figure 3.5.

We propose II-SVOR-EXP algorithm to learn parameters $(\boldsymbol{w},\boldsymbol{\theta})$ of the ordinal classifier (3.1) from partially annotated examples $\mathcal{D}_I^m$ by solving the following convex problem

$$(\boldsymbol{w}^*,\boldsymbol{\theta}^*) \in \mathop{\rm Argmin}_{\boldsymbol{w}\in\mathbb{R}^n,\boldsymbol{\theta}\in\hat{\Theta}} \left[\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^m \psi_I^{\rm EXP}(\boldsymbol{w},\boldsymbol{\theta};\boldsymbol{x}^i,y_l^i,y_r^i) \right] . \tag{3.34}$$

3.3.3. Interval insensitive support vector ordinal regression: implicit constraints on thresholds

Analogically, we derive a convex surrogate of the interval insensitive loss $\ell_I^{MAE}(y_l,y_r,y)$ associated with the MAE as follows

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

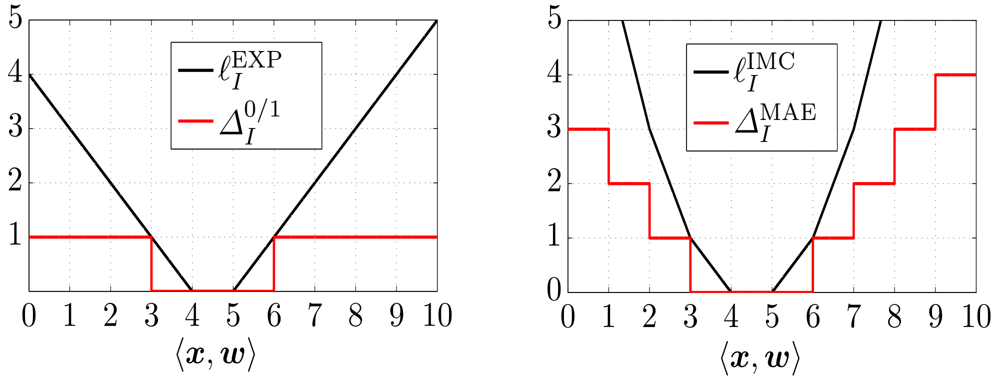


Figure 3.5. The left figure shows the interval insensitive loss $\ell_I^{0/1}(\boldsymbol{x}, y_l, y_r, h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}))$ associated with the 0/1-loss and its surrogate $\psi_I^{\text{EXP}}(\boldsymbol{w}, \boldsymbol{\theta}; \boldsymbol{x}, y_l, y_r))$. The right figure shows the interval insensitive loss $\ell_I^{\text{MAE}}(\boldsymbol{x}, y_l, y_r, h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}))$ associated with the MAE loss and its surrogate $\psi_I^{\text{IMC}}(\boldsymbol{w}, \boldsymbol{\theta}; \boldsymbol{x}, y_l, y_r))$. The losses are shown as a function of the score $\langle \boldsymbol{x}, \boldsymbol{w} \rangle$ evaluated for $\theta_1 = 1, \theta_2 = 2, \ldots, \theta_{Y-1} = Y-1$ and $y_l = 4$, $y_r = 6$. Note that for this particular setting of $\boldsymbol{\theta}$ the surrogate $\psi_I^{\text{EXP}}(\boldsymbol{w}, \boldsymbol{\theta}; \boldsymbol{x}, y_l, y_r))$ also appears to upper bound $\ell_I^{\text{MAE}}(\boldsymbol{x}, y_l, y_r, h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}))$, however, this does not hold in general.

We obtain a convex surrogate by replacing the step functions by the hinge loss

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

which is obviously an upper bound of $\ell_I^{\text{MAE}}(y_l, y_r, h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}))$ as can be also seen in Figure 3.5.

Given the partially annotated examples $\mathcal{D}_I^m$, we can learn parameters $(\boldsymbol{w}, \boldsymbol{\theta})$ of the ordinal classifier (3.1) by solving

$$(\boldsymbol{w}^*, \boldsymbol{\theta}^*) \in \operatorname*{Argmin}_{\boldsymbol{w} \in \mathbb{R}^n, \boldsymbol{\theta} \in \mathbb{R}^n} \left[\frac{\lambda}{2} \|\boldsymbol{w}\|^2 + \sum_{i=1}^{m} \psi_I^{\text{IMC}}(\boldsymbol{w}, \boldsymbol{\theta}; \boldsymbol{x}^i, y_l^i, y_r^i) \right] . \tag{3.35}$$

We denote the modified variant as the II-SVOR-IMC algorithm. Note that due to the equality $\psi_I^{\text{IMC}}(\boldsymbol{w}, \boldsymbol{\theta}; \boldsymbol{x}, y, y) = \psi^{\text{IMC}}(\boldsymbol{w}, \boldsymbol{\theta}; \boldsymbol{x}, y)$ it is clear that the proposed II-SVOR-IMC subsumes the original supervised SVOR-IMC as a special case.

3.3.4. 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. The MORD parametrization allows to adopt existing techniques for linear classification. Given a V-shaped supervised loss $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, we propose to approximate the value of the associated interval insensitive loss $\ell_I(y_l, y_r, h'(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{b}))$ by a surrogate loss $\psi_I\colon \mathbb{R}^n \times \mathbb{R}^Y \times \mathcal{X} \times \mathcal{P} \to \mathbb{R}$ defined as

$$\begin{aligned}\psi_I(\boldsymbol{w},\boldsymbol{b};\boldsymbol{x},y_l,y_r) &= \max_{y\le y_l}\Big[\ell(y,y_l)+\langle\boldsymbol{x},\boldsymbol{w}\rangle(y-y_l)+b_y-b_{y_l}\Big] \\ &\quad+\max_{y\ge y_r}\Big[\ell(y,y_r)+\langle\boldsymbol{x},\boldsymbol{w}\rangle(y-y_r)+b_y-b_{y_r}\Big]\,.\end{aligned} \tag{3.36}$$

It is seen that the function $\psi_I(\boldsymbol{w},\boldsymbol{b};\boldsymbol{x},y_l,y_r)$ is a sum of two point-wise maxima over linear functions for fixed $(\boldsymbol{x},y_l,y_r)$. Hence, it is convex in the parameters $(\boldsymbol{w},\boldsymbol{b})$. The following proposition states that the surrogate is like the previous surrogates 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*

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

holds where $h'(\boldsymbol{x};\boldsymbol{w},\boldsymbol{b})$ denotes response of the MORD classifier (3.3).

Proof is deferred to Appendix A.3.

Given partially annotated training examples $\mathcal{D}_I^m$, we can learn parameters $(\boldsymbol{w},\boldsymbol{b})$ of the MORD classifier (3.3) by solving the following unconstrained convex problem

$$(\boldsymbol{w}^*,\boldsymbol{b}^*)=\operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n,\boldsymbol{b}\in\mathbb{R}^Y}\left[\frac{\lambda}{2}\|\boldsymbol{w}\|^2+\frac{1}{m}\sum_{i=1}^m\psi_I(\boldsymbol{w},\boldsymbol{b};\boldsymbol{x}^i,y_l^i,y_r^i)\right], \tag{3.37}$$

where $\lambda\in\mathbb{R}_{++}$ is a regularization constant. A suitable value of the regularization constant is typically tuned on the validation set. In the sequel, we denote the method based on solving (3.37) as the VILMA.

As important example, let us consider the surrogate (3.36) instantiated for the MAE loss. In this case, the surrogate becomes

$$\begin{aligned}\psi_I^{\text{MAE}}(\boldsymbol{w},\boldsymbol{b};\boldsymbol{x},y_l,y_r) &= \max_{y\le y_l}\Big[y_l-y+\langle\boldsymbol{x},\boldsymbol{w}\rangle(y-y_l)+b_y-b_{y_l}\Big] \\ &\quad+\max_{y\ge y_r}\Big[y-y_r+\langle\boldsymbol{x},\boldsymbol{w}\rangle(y-y_r)+b_y-b_{y_r}\Big]\,.\end{aligned} \tag{3.38}$$

It is interesting to compare the VILMA instantiated for the MAE loss 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 (3.1) while the VILMA parameters $(\boldsymbol{w},\boldsymbol{b})$ of the MORD rule (3.3). The following proposition states that surrogates of both methods 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 (3.1) and $h'(\boldsymbol{x};\boldsymbol{w},\boldsymbol{b})$ the MORD classifier (3.3). Then the equality*

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

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

Proof is deferred to Appendix A.4.

Proposition 2 ensures that the II-SVOR-IMC algorithm and the VILMA with MAE loss both return the same classification rules although differently parametrized.

The core properties of the generic method, the VILMA, proposed in this section:

1. VILMA is applicable for an arbitrary V-shaped loss,
2. VILMA subsumes the II-SVOR-IMC algorithm optimizing the MAE loss as a special case,
3. VILMA converts learning into an unconstrained convex optimization. Note that the II-SVOR-EXP and the II-SVOR-IMC in contrast to VILMA maintain the set of linear constraints $\boldsymbol{\theta} \in \Theta = \{\boldsymbol{\theta}' \in \mathbb{R}^{Y-1} \mid \theta'_y \leq \theta'_{y+1},\ y = 1, \dots, Y-1\}$.
   In next section we will describe a solver for the optimisation problem (3.37).

3.4. Generic cutting plane solver

The proposed method VILMA translates learning into a convex optimization problem (3.37) that can be re-written as

$$(\boldsymbol{w}^*, \boldsymbol{b}^*) \in \operatorname*{Argmin}_{\boldsymbol{w}\in\mathbb{R}^n, \boldsymbol{b}\in\mathbb{R}} \left[\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \psi_I(\boldsymbol{w}, \boldsymbol{b}; \boldsymbol{x}^i, y_l^i, y_r^i) \right] . \tag{3.39}$$

Note, we dropped regularization over bias term $\boldsymbol{b}$ in formulation (3.39). The motivation for this comes from practical problems. Experiments in Section 3.5.5 show that in case of high dimensional parameter vector $\boldsymbol{w}$ and small number of classes the formulation (3.39) has an advantage over its possible alternative

$$(\boldsymbol{w}^*, \boldsymbol{b}^*) \in \operatorname*{Argmin}_{\boldsymbol{w}\in\mathbb{R}^n, \boldsymbol{b}\in\mathbb{R}} \left[\frac{\lambda}{2}(\|\boldsymbol{w}\|^2 + \|\boldsymbol{b}\|^2) + \frac{1}{m}\sum_{i=1}^{m} \psi_I(\boldsymbol{w}, \boldsymbol{b}; \boldsymbol{x}^i, y_l^i, y_r^i) \right] . \tag{3.40}$$

Therefore we concentrate our attention to problem (3.39). Of course, the task (3.39) can be reformulated as a quadratic program with $\mathcal{O}(n+m+Y)$ variables and $\mathcal{O}(Y \cdot m)$ constraints. However, generic off-the-shelf Quadratic Programming (QP) solvers are applicable only to small problems. Unlike problem (3.40), we can not plug problem (3.39) into CPA framework directly due to the regularizer that operates only on part of variables to be optimized. In this section, we derive the instance of the CPA tailored to the problem (3.39). The resulting CPA is applicable for large problems and it provides a certificate of the optimality.

More details on the CPA based solvers applied to the machine learning problems can be found for example in [Teo et al., 2010; Franc et al., 2012]. The standard CPA is suitable for solving convex tasks of the form

$$\boldsymbol{w}^* \in \operatorname*{Argmin}_{\boldsymbol{w}\in\mathbb{R}^n} F(\boldsymbol{w}) , \quad \text{where} \quad F(\boldsymbol{w}) = \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + G(\boldsymbol{w}) \tag{3.41}$$

and $G\colon \mathbb{R}^n \to \mathbb{R}$ is a convex function. In contrast to our problem (3.39), the objective of (3.41) contains a quadratic regularization imposed on all variables. It is well known that the CPA applied directly to the un-regularized problem like (3.39) exhibits a strong zig-zag behavior leading to a large number of iterations. A frequently used an ad-hoc solution is to impose an artificial regularization on $\boldsymbol{b}$, which may however significantly spoil the results as demonstrated in section 3.5. In the rest of this section, we first outline the CPA algorithm for the problem (3.41) and then show how it can be used to solve the problem (3.39).

The core idea of the CPA is to approximate the solution of the master problem (3.41) by solving a *reduced problem*

$$\boldsymbol{w}_t \in \operatorname*{Argmin}_{\boldsymbol{w}\in\mathbb{R}^n} F_t(\boldsymbol{w}) , \quad \text{where} \quad F_t(\boldsymbol{w}) = \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + G_t(\boldsymbol{w}) . \tag{3.42}$$

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 (3.41)
repeat
    $t \leftarrow t + 1$
    Compute $G(\boldsymbol{w}_{t-1})$ and $G'(\boldsymbol{w}_{t-1})$
    Update the model $G_t(\boldsymbol{w}) \leftarrow \max_{i=0,\dots,t-1} G(\boldsymbol{w}_i) + \langle G'(\boldsymbol{w}_i, \boldsymbol{w} - \boldsymbol{w}_i \rangle$
    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 $F(\boldsymbol{w}_t) - F_t(\boldsymbol{w}_t) \leq \varepsilon$;
```

The reduced problem (3.42) is obtained from (3.41) 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 of $G(\boldsymbol{w})$ reads

$$G_t(\boldsymbol{w}) = \max_{i=0,\dots,t-1} \left[G(\boldsymbol{w}_i) + \langle G'(\boldsymbol{w}_i), \boldsymbol{w} - \boldsymbol{w}_i \rangle \right] , \tag{3.43}$$

where $G'(\boldsymbol{w}) \in \mathbb{R}^n$ is a sub-gradient of G at point $\boldsymbol{w}$. Thanks to the convexity of $G(\boldsymbol{w})$, $G_t(\boldsymbol{w})$ is a piece-wise linear underestimator of $G(\boldsymbol{w})$, which is tight in the points $\boldsymbol{w}_i$, $i = 0, \dots, t-1$. In turn, the reduced problem objective $F_t(\boldsymbol{w})$ is an underestimator of $F(\boldsymbol{w})$. The cutting plane model is build iteratively by the following simple procedure. Starting from $\boldsymbol{w}_0 \in \mathbb{R}^n$, the CPA computes a new iterate $\boldsymbol{w}_t$ by solving the reduced problem (3.42). In each iteration t, the cutting-plane model (3.43) is updated by a new 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 $F(\boldsymbol{w}_t) \leq F(\boldsymbol{w}^*) + \varepsilon$ holds. The CPA is guaranteed to halt after $\mathcal{O}(\frac{1}{\lambda\varepsilon})$ iterations at most [Teo et al., 2010]. The CPA is outlined in Algorithm 1.

We can convert our problem (3.39) to (3.41) by setting

$$G(\boldsymbol{w}) = R_{\text{emp}}(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{w})) \,, \quad \text{where} \quad \boldsymbol{b}(\boldsymbol{w}) \in \operatorname*{Argmin}_{\boldsymbol{b} \in \mathbb{R}^Y} R_{\text{emp}}(\boldsymbol{w}, \boldsymbol{b}) \,. \tag{3.44}$$

It is clear that if $\boldsymbol{w}^*$ is the solution of the problem (3.41) with the function $G(\boldsymbol{w})$ defined by the equation (3.44). Consequently, $(\boldsymbol{w}^*, \boldsymbol{b}(\boldsymbol{w}^*))$ must be a solution of (3.39). Because $R_{\text{emp}}(\boldsymbol{w}, \boldsymbol{b})$ is jointly convex in $\boldsymbol{w}$ and $\boldsymbol{b}$, the function $G(\boldsymbol{w})$ in (3.44) is also convex in $\boldsymbol{w}$ (see for example [Boyd and Vandenberghe, 2004]). Hence, the application of Algorithm 1 to solve (3.39) will preserve all its convergence guarantees. To this end, we only need to provide the first-order oracle computing $G(\boldsymbol{w})$ and the sub-gradient $\nabla G(\boldsymbol{w})$ required to build the cutting plane model. For given $\boldsymbol{b}(\boldsymbol{w})$, the subgradient of $G(\boldsymbol{w})$ reads [Boyd and Vandenberghe, 2004]

$$\nabla G(\boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{x}^i (\hat{y}_l^i + \hat{y}_r^i - y_l^i - y_r^i) \tag{3.45}$$

where

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

The proposed CPA transforms solving of the problem (3.39) into a sequence of two simpler problems:

1. The reduced problem (3.42) solved in each iteration of the CPA. The problem (3.42) is a quadratic program that can be approached via its dual formulation [Teo et al., 2010] having 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 dual of (3.42) can be solved by off-the-shelf QP libraries.
2. The problem (3.44) providing $\boldsymbol{b}(\boldsymbol{w})$, which is required to compute $G(\boldsymbol{w}) = R_{\text{emp}}(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{w}))$ and the sub-gradient $G'(\boldsymbol{w})$ via equation (3.45). The problem (3.44) has only $\mathcal{O}(Y)$ (the number of labels) variables. Hence it can be approached by generic convex solvers like the Analytic Center Cutting Plane Method (ACCPM) algorithm [Gondzio et al., 1996].

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

Finally, we point out that the convex problems associated with the generic SO-SVM formulation (3.28) can be solved by the same solver. The only change is in using a different formulas for evaluating the risk and its subgradient. The same holds for the convex problems associated with the II-SVOR-EXP and the II-SVOR-IMC, in which case, however, we have to use additional constraints $\boldsymbol{\theta} \in \hat{\Theta}$ in (3.39) which propagate to the problem (3.44).

3.5. Experiments

In this section we present results of a series of experiments including:

- Section 3.5.4: Comparison of surrogate losses for supervised learning.
- Section 3.5.5: Assessment of the influence of regularization terms on the final accuracy.
- Section 3.5.6: Evaluation of the flexibility of the propose PW-MORD model.
- Section 3.5.7: Learning from the interval annotations by using the interval insensitive loss.
- Section 3.5.8: Evaluation of the tightness of the upper bound from Theorem 3.
- Section 3.5.9: Comparison of surrogate losses for learning from partial annotations.
- Section 3.5.10: Empirical verification of Proposition 2.

Before jumping on the experiments we first list the compared methods in Section 3.5.1, then we describe the used datasets in Section 3.5.2 and also the experimental protocol in Section 3.5.3.

3.5.1. Compared methods

In the experiments we evaluate all algorithms discussed in this chapter. First, the existing methods for supervised learning of the ordinal classifiers like SVOR-EXP and SVOR-IMC, as well as a simple Rounded linear regressor (LinReg) and a generic Linear Multi-class SVM classifier (LinCls). Second, we evaluate the proposed algorithms for supervised learning based on the SO-SVM algorithm and the generic algorithm VILMA for learning from partial annotations. A summary of all evaluated algorithms is presented in Table 3.1. Note that each instance of the evaluated algorithms has its unique name which encodes the learning problem, the regularization term and the loss function optimized.

Used convex solvers. All the evaluated algorithms translate learning into a certain convex minimization problem. We used BMRM solver [Teo et al., 2010] for problems where the quadratic regularization term includes all variables. We used the ACCPM solver [Antoniuk et al., 2012] for problems with no regularization term. When a part of the variables ($\boldsymbol{b}$ or $\boldsymbol{\theta}$) is

Name used in the text	Partial annotations	Classification model	Learning problem
LinReg(reg,loss)	NO	LinReg (3.13)	SO-SVM (3.28)
LinCls(reg,loss)	NO	LinCls (3.14)	SO-SVM (3.28)
SVOR-EXP(reg)	NO	ORD (3.1)	SVOR-EXP (3.23)
SVOR-IMC(reg)	NO	ORD (3.1)	SVOR-IMC (3.25)
MORD(reg,loss)	NO	MORD (3.3)	SO-SVM (3.28)
PW-MORD(reg,loss)	NO	PW-MORD (3.11)	SO-SVM (3.28)
VILMA(reg,loss)	YES	MORD (3.3)	VILMA (3.37)
II-SVOR-EXP(reg)	YES	ORD (3.1)	II-SVOR-EXP (3.34)
II-SVOR-IMC(reg)	YES	ORD (3.1)	II-SVOR-IMC (3.35)

Table 3.1. The summary of evaluated algorithms. The argument reg $\in \{\emptyset, \boldsymbol{w}, \boldsymbol{wb}, \boldsymbol{w\theta}\}$ determines whether no regularizer, $\|\boldsymbol{w}\|^2$, $\|\boldsymbol{w}\|^2+\|\boldsymbol{b}\|^2$ or $\|\boldsymbol{w}\|^2+\|\boldsymbol{\theta}\|^2$ is used, respectively. The second argument loss $\in \{\text{MAE}, 0/1\}$, applicable only for generic methods, determines which target loss is used. The column "partial annotations" indicates whether the method can deal with partial annotations. The last two columns show which classification model is learned and which optimization problem is solved by the given method, respectively.

not included in the regularized term, we use the proposed double-loop CPA (c.f. Section 3.4), which is a combination of both aforementioned methods. That is, the double-loop CPA runs BMRM in the main loop and ACCPM to solve the internal problem (3.44). In particular, we used a modified version of the BMRM from the Shogun machine learning library [Sonnenburg et al., 2010] and the Oracle Based Optimization Engine (OBOE) implementation of the Analytic Center Cutting Plane algorithm being a part of COmputational INfrastructure for Operations Research project (COIN-OR) [Gondzio et al., 1996]. We configured the used solvers to find the ε-optimal solution of the learning objective in all cases. In particular, we stopped the solver if the objective was below a factor of 1.01 of the optimal value [2].

3.5.2. Benchmark data

In our experiments, we use a subset of seven datasets[3] from UCI repository which were used in [Chu and Keerthi, 2005; Li and Lin, 2006] and two large face databases with year-precise annotation of the age of depicted subjects:

1. **UCI** data sets collection listed in Table 3.2, same as in [Chu and Keerthi, 2005; Li and Lin, 2006]. The data were produced by discretising metric regression problems into $Y = 10$ bins.
2. **MORPH** database [Ricanek and Tesafaye, 2006] is the standard benchmark for age estimation. It contains 55,134 face images with the ground true 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 most of them are of very good quality.

[2] Our implementation is available at Github: *https://github.com/K0stIa/VILMA*

[3] The link `http://www.dcc.fc.up.pt/~ltorgo/Regression/census.tar.gz` to the eight dataset "Census" was broken hence we could not include it.

3. **WILD** database is a collection of three 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 annotators. 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.

Pre-processing of MORPH and WILD database. In both MORPH and WILD data sets, we made sure that images of the same identity never appear in different parts simultaneously. The feature representation of the facial images of both MORPH and WILD data sets was computed as follows. We first localized the faces by a commercial face detector[4] 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 60×40 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.

3.5.3. Experimental protocol

Supervised setting For UCI data sets collection, we followed exactly the same evaluation protocol as in [Chu and Keerthi, 2005; Li and Lin, 2006]. Data is randomly partitioned to the training and testing part. The partitioning are repeated 20 times. The features are normalized to have zero mean and unit variance coordinate wise. The reported results are averages and standard deviations computed over the 20 partitions. The feature dimension and training and testing ratios are listed in Table 3.2. The regularization constant λ is chosen from a fixed set of values $\Lambda = \{1, 0.1, 0.01, 0.001, 0\}$ using 5-fold cross-validation estimate of the minimizing loss (MAE or 0/1) on the training split.

For both face data sets, MORPH and WILD, we used images with the year-precise age annotations. While in the MORPH data set the annotation is the biological age in the WILD

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

Dataset	number of features	number of training examples	number of test examples
Pyrimidines	27	50	24
MachineCPU	6	150	59
Boston	13	300	206
Abalone	8	1000	3177
Bank	32	3000	5192
Computer	21	4000	4192
California	8	5000	15640

Table 3.2. A subset of seven datasets from UCI repository used for benchmarking the algorithms for ordinal classification. The table shows the number of input features as well as the number of the training and the testing examples used in each random partition.

dataset it is a human estimate of the age. We constructed a sequence of training sets with the number of examples m varying from $m = 3,300$ to $m = 33,000$ in case of MORPH or from $m = 3,300$ to $m = 21,000$ in case of WILD. For each training set we learned classifiers with the regularization parameters set to $\lambda \in \{1, 0.1, 0.01, 0.001\}$ and sometimes to 0.0001, when it was needed. The classifier corresponding to λ with the smallest validation error was applied to the testing examples. This process was repeated for the three random splits on training, validation and testing part in the ratio 60/20/20. We report the averages and the standard deviations of the MAE computed on the test examples over the three splits.

Learning from partial (interval) annotations The MORPH and the WILD databases contain the year-precise annotation. We generated partial annotation from the precise one in order to have a ground-truth against which we can compare. The interval annotations were generated in a way mimicking a possible real situation as follows:

1. The number of m_P randomly selected examples were annotated precisely by taking the annotation from the databases.
2. The number of m_I randomly selected examples were annotated by intervals. The admissible annotation intervals were chosen so that they partition the set of ages and have the same width (up to the border cases). The interval width was varied from $u \in \{5, 10, 20\}$. The interval annotation was obtained by rounding the true age from the databases into 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.

Note that the used annotation process is approximately $\alpha\beta$-precise (c.f. Definition 6) with $\alpha = m_P/(m_P + m_I)$ 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_{total} is the total number of the training examples in the corresponding database.

We used the same random splits into training/validation/test examples as in the supervised setting with the only difference that instead of supervised annotation the interval annotation is used for the training set. The precise (supervised) annotation is used in the validation and the test set. We also performed experiments when the validation sets had interval annotations. However, the obtained results were almost identical hence they are not presented here.

3.5.4. Comparison of surrogate losses for supervised learning

In this experiments we consider the supervised learning of the ordinal classifier in the case when the target loss ℓ is either the 0/1-loss or the MAE loss. In particular, we compare the existing algorithms SVOR-EXP($\emptyset$), SVOR-IMC($\emptyset$) and the proposed methods MORD($\emptyset$,loss) and VILMA($\emptyset$,loss). All compared methods solve an instance of the surrogate ERM problem (3.19) with different surrogate losses ψ. The goal is to measure which surrogate ψ is a better approximation of the target loss ℓ. To this end, we evaluated the true empirical risk $R^{\ell}_{\text{emp}}(h)$ for the classifier $h = \text{pred} \circ f^{\text{emp}}_{*}$ obtained by solving the surrogate ERM problem (3.19). In this experiment we set the parameter $\lambda = 0$ in order to switch of the regularization term. Table 3.3 summarizes the empirical risk of the algorithms minimizing the surrogate of the MAE loss. The results for the algorithms minimizing the surrogate of the 0/1-loss are in Table 3.4. Note that in both tables we report the empirical risk defined by the MAE loss as well as the 0/1-loss. Based on the results we can derive the following conclusions:

- The proposed algorithm MORD($\emptyset$,MAE) achieves consistently (up to one near draw for "Computer" data) the minimal $R^{\text{MAE}}_{\text{emp}}$. That is, the SO-SVM learning the MORD classifier with the margin-rescaling loss provides the best approximation of the target MAE loss.
- The proposed algorithm VILMA($\emptyset$,MAE) and the existing algorithm SVOR-IMC($\emptyset$) yield the same $R^{\text{MAE}}_{\text{emp}}$ and $R^{0/1}_{\text{emp}}$. This observation is an experimental "check-up" of Proposition 2 which states that the surrogate losses of the two algorithms are equivalent although each is defined for different parametrizations of the ordinal classifier. In turn both methods must produce the same (up to numerical errors) classifier but differently parametrized.
- The SVOR-EXP($\emptyset$) algorithm achieves consistently the minimal $R^{0/1}_{\text{emp}}$.
- A surprising result is that the MORD($\emptyset$,MAE) and VILMA($\emptyset$,MAE) achieve consistently lower $R^{0/1}_{\text{emp}}$ than their counterparts, MORD($\emptyset$,0/1) and VILMA($\emptyset$,0/1), minimizing surrogates of the 0/1-loss. Moreover, the $R^{0/1}_{\text{emp}}$ of the MORD($\emptyset$,MAE) and VILMA($\emptyset$,MAE) is only slightly worse than that of the SVOR-EXP($\emptyset$). This suggests that the surrogates of the MAE loss are good approximations of both target losses. An explanation of this finding remains an open question.

	TrnRisk	MORD(∅,MAE)	VILMA(∅,MAE)	SVOR-IMC(∅)
Pyrimidines	MAE	**0.433 (0.093)**	0.487 (0.105)	0.482 (0.104)
	0/1	**0.343 (0.064)**	0.399 (0.070)	0.391 (0.069)
MachineCPU	MAE	**0.914 (0.052)**	0.917 (0.045)	0.920 (0.046)
	0/1	**0.602 (0.035)**	0.609 (0.024)	0.611 (0.027)
Boston	MAE	**0.812 (0.043)**	0.823 (0.045)	0.823 (0.047)
	0/1	**0.558 (0.026)**	0.575 (0.027)	0.573 (0.027)
Abalone	MAE	**1.412 (0.038)**	1.424 (0.042)	1.422 (0.041)
	0/1	**0.734 (0.015)**	0.748 (0.015)	0.748 (0.017)
Bank	MAE	**1.421 (0.021)**	1.427 (0.021)	1.429 (0.021)
	0/1	**0.700 (0.006)**	0.715 (0.009)	0.716 (0.007)
Computer	MAE	**0.632 (0.010)**	0.632 (0.009)	0.632 (0.010)
	0/1	**0.477 (0.006)**	0.481 (0.005)	0.480 (0.006)
California	MAE	**1.178 (0.013)**	1.287 (0.326)	1.182 (0.014)
	0/1	**0.692 (0.008)**	0.706 (0.028)	0.697 (0.007)

Table 3.3. Comparison of various algorithms in terms of their ability to minimize the target empirical risk on the UCI datasets. The training risk was computed w.r.t. the 0/1 loss and the MAE loss. All compared algorithms minimize various surrogates of the target MAE loss and they use no regularization.

	TrnRisk	MORD(∅,0/1)	VILMA(∅,0/1)	SVOR-EXP(∅)
Pyrimidines	MAE	0.544 (0.157)	0.506 (0.113)	**0.491 (0.125)**
	0/1	0.395 (0.083)	0.400 (0.079)	**0.329 (0.078)**
MachineCPU	MAE	1.368 (1.044)	**0.939 (0.062)**	0.972 (0.068)
	0/1	0.659 (0.084)	0.616 (0.027)	**0.594 (0.029)**
Boston	MAE	3.765 (1.465)	0.890 (0.045)	**0.869 (0.050)**
	0/1	0.901 (0.010)	0.600 (0.025)	**0.551 (0.028)**
Abalone	MAE	4.496 (0.096)	1.766 (0.923)	**1.632 (0.063)**
	0/1	0.734 (0.015)	0.777 (0.046)	**0.715 (0.016)**
Bank	MAE	3.965 (0.043)	2.108 (1.089)	**1.913 (0.051)**
	0/1	0.796 (0.005)	0.717 (0.047)	**0.690 (0.005)**
Computer	MAE	0.761 (0.015)	0.655 (0.010)	**0.653 (0.012)**
	0/1	0.551 (0.010)	0.491 (0.006)	**0.477 (0.008)**
California	MAE	4.511 (0.036)	**1.212 (0.018)**	1.233 (0.014)
	0/1	0.901 (0.002)	0.709 (0.010)	**0.681 (0.008)**

Table 3.4. Comparison of various algorithms in terms of their ability to minimize the target empirical risk on the UCI datasets. The training risk was computed w.r.t. the 0/1 loss and the target MAE loss. All compared algorithms minimize various surrogates of the 0/1-loss and they use no regularization.

	TstRisk	LinCls	LinReg	PW-MORD($\boldsymbol{wb}$, MAE)	PW-MORD($\boldsymbol{w}$, MAE)	VILMA($\boldsymbol{wb}$, MAE)	VILMA($\boldsymbol{w}$)	SVOR-IMC($\boldsymbol{w\theta}$)	SVOR-IMC($\boldsymbol{w}$, MAE)	SVOR-EXP($\boldsymbol{w\theta}$)	SVOR-EXP($\boldsymbol{w}$)
Pyrimidines	MAE	1.59 (0.25)	**1.37 (0.27)**	1.50 (0.38)	1.55 (0.35)	1.57 (0.30)	1.52 (0.29)	1.52 (0.29)	1.46 (0.29)	1.63 (0.28)	1.76 (0.28)
	0/1	0.76 (0.10)	0.76 (0.10)	**0.74 (0.09)**	0.76 (0.07)	0.79 (0.08)	0.78 (0.06)	0.79 (0.07)	0.77 (0.06)	0.80 (0.08)	0.82 (0.08)
MachineCPU	MAE	1.00 (0.15)	1.03 (0.10)	0.95 (0.12)	0.98 (0.14)	0.94 (0.11)	**0.92 (0.05)**	0.95 (0.11)	0.95 (0.11)	1.01 (0.13)	0.95 (0.10)
	0/1	0.65 (0.06)	0.70 (0.06)	0.62 (0.06)	0.64 (0.05)	0.63 (0.06)	**0.62 (0.03)**	0.63 (0.06)	0.63 (0.06)	0.65 (0.05)	0.63 (0.06)
Boston	MAE	0.94 (0.07)	0.95 (0.06)	**0.86 (0.05)**	0.89 (0.06)	0.93 (0.07)	0.92 (0.06)	0.91 (0.06)	0.91 (0.05)	0.97 (0.08)	0.93 (0.09)
	0/1	0.62 (0.03)	0.64 (0.03)	**0.58 (0.03)**	0.59 (0.03)	0.62 (0.04)	0.61 (0.03)	0.61 (0.03)	0.61 (0.03)	0.62 (0.04)	0.62 (0.04)
Abalone	MAE	1.42 (0.02)	1.51 (0.01)	**1.41 (0.02)**	**1.41 (0.02)**	1.48 (0.02)	1.48 (0.02)	1.47 (0.01)	1.48 (0.02)	1.68 (0.04)	1.47 (0.01)
	0/1	0.73 (0.01)	0.79 (0.01)	**0.73 (0.01)**	**0.73 (0.01)**	0.76 (0.01)	0.76 (0.01)	0.76 (0.01)	0.76 (0.01)	0.73 (0.01)	0.76 (0.01)
Bank	MAE	**1.45 (0.01)**	1.51 (0.01)	**1.45 (0.01)**	**1.45 (0.01)**	1.46 (0.02)	1.46 (0.01)	**1.45 (0.01)**	**1.45 (0.01)**	1.94 (0.05)	**1.45 (0.01)**
	0/1	0.70 (0.01)	0.77 (0.01)	0.70 (0.01)	0.70 (0.01)	0.72 (0.01)	0.72 (0.01)	0.72 (0.01)	0.72 (0.01)	**0.69 (0.00)**	0.72 (0.01)
Computer	MAE	0.62 (0.01)	0.72 (0.01)	**0.61 (0.01)**	**0.61 (0.01)**	0.64 (0.01)	0.64 (0.01)	0.63 (0.01)	0.63 (0.01)	0.65 (0.01)	0.63 (0.01)
	0/1	0.47 (0.00)	0.56 (0.01)	**0.47 (0.01)**	**0.47 (0.01)**	0.49 (0.01)	0.49 (0.01)	0.48 (0.01)	0.48 (0.01)	0.48 (0.00)	0.48 (0.01)
California	MAE	**1.12 (0.00)**	1.21 (0.01)	1.14 (0.00)	1.14 (0.00)	1.19 (0.01)	1.18 (0.01)	1.18 (0.01)	1.18 (0.01)	1.23 (0.01)	1.18 (0.01)
	0/1	**0.67 (0.00)**	0.71 (0.00)	0.68 (0.00)	0.68 (0.00)	0.70 (0.00)	0.70 (0.01)	0.70 (0.00)	0.70 (0.00)	0.68 (0.00)	0.70 (0.00)

Table 3.5. Comparison of various classification models in terms of the test risk measured as the MAE and the 0/1-loss.

3.5.5. Impact of the regularization term

In this experiment we compared the performance of the SVOR-IMC(reg), VILMA(reg), MORD(reg,MAE) and PW-MORD(reg,MAE) using different regularization terms. Table 3.6 shows the test accuracy of the compared methods as a function of the number of training examples obtained on the MORPH and the WILD datasets. Table 3.5 shows the test accuracy on the UCI benchmarks. The test accuracy is an estimate of the Bayes risk defined with the MAE loss.

The experiments on the MORPH and the WILD datasets show that pushing the bias terms ($\boldsymbol{b}$ or $\boldsymbol{\theta}$) towards zero by the quadratic regularizer may have a detrimental effect on the classification accuracy. This holds consistently for all methods but the PW-MORD(reg,MAE) algorithm which benefits from the bias regularizer when training from a small number of examples. This can be explained by the fact that unlike the other methods, the PW-MORD(reg,MAE) algorithm learns more flexible classification models with a higher number of parameters and hence it requires stronger regularization. The obtained bad results when regularizing the bias term of standard ordinal classifiers show that the projecting vector $\boldsymbol{w}$ and the bias term ($\boldsymbol{b}$ or $\boldsymbol{\theta}$) have different influence on the separating surface and hence they can not be treated in the same manner. However, this effect is negligable on the low-dimensional UCI benchmarks as can be seen in Table 3.5. It suggests that the choice of the regularizer becomes especially important in the case of high dimensional features.

3.5.6. Testing flexibility of the PW-MORD model

In this section, we demonstrate the benefits of the proposed PW-MORD model for ordinal classification. As shown in Section 3.1.2, the PW-MORD model subsumes the standard (i.e, the simplest) ordinal classifier as well as the unconstrained multi-class classifier as special cases. The complexity of the PW-MORD model is controlled by defining the set $\mathcal{Z}$ containing the "cut labels". Recall that the classes between the cut labels are modeled by a standard ordinal classifier. We used a different number of the cut labels and we set their position equidistantly between the minimal and the maximal label. In particular, we used the following settings:

- UCI benchmark: $\mathcal{Z} \in \big\{\{1,10\},\{1,5,10\},\{1,4,7,10\}\big\}$.
- MORPH dataset: $\mathcal{Z} \in \big\{\{0,20,40,54\},\{0,14,26,38,54\},\{0,11,22,33,44,54\}\big\}$.
- WILD dataset: $\mathcal{Z} \in \big\{\{0,25,54,79\},\{0,20,40,60,79\},\{0,16,32,48,54,79\}\big\}$.

Note that PW-MORD with Z=2 corresponds to the MORD classifier. Parameters of the PW-MORD model were learned by the PW-MORD(reg,MAE) algorithm, i.e. we used the MAE loss as the target loss. The optimal setting of $\mathcal{Z}$ was selected based on the same procedure as was used for the regularization constant, i.e. using 5-fold cross-validation in case of UCI benchmark and using validation set in case of MORPH/WILD datasets (c.f. Section 3.5.3). Results obtained on the UCI benchmark and the MORPH/WILD datasets are summarized in Table 3.5 and Table 3.6, respectively. The main empirical findings are as follows:

- **UCI benchmarks.** The PW-MORD(reg,MAE) algorithm outperformed the other methods in most cases. It is not surprising since the PW-MORD model subsumes the other models as special cases and, moreover, the surrogate minimized by PW-MORD(reg,MAE) best approximates the target MAE loss as shown in Section 3.5.4. The PW-MORD(reg,MAE) was outperformed only by the LinReg(reg,MAE) on the "Pyrimids" data and by the

MORPH

	$m = 3300$	$m = 6600$	$m = 13000$	$m = 23000$	$m = 33000$
SVOR-IMC($\boldsymbol{w}$)	5.54 ± 0.03	5.10 ± 0.02	**4.83 ± 0.01**	**4.69 ± 0.03**	4.61 ± 0.03
SVOR-IMC($\boldsymbol{w\theta}$)	5.67 ± 0.04	5.24 ± 0.02	5.04 ± 0.04	4.99 ± 0.03	5.01 ± 0.03
VILMA($\boldsymbol{w}$, MAE)	5.56 ± 0.02	5.12 ± 0.02	**4.83 ± 0.02**	**4.66 ± 0.01**	**4.55 ± 0.02**
VILMA($\boldsymbol{wb}$, MAE)	8.16 ± 0.44	8.01 ± 0.43	7.94 ± 0.42	7.79 ± 0.44	7.72 ± 0.43
MORD($\boldsymbol{w}$, MAE)	5.71 ± 0.04	5.24 ± 0.01	5.11 ± 0.04	5.06 ± 0.04	5.04 ± 0.03
MORD($\boldsymbol{wb}$, MAE)	8.04 ± 0.30	7.89 ± 0.13	7.69 ± 0.14	7.78 ± 0.40	7.70 ± 0.39
PW-MORD($\boldsymbol{w}$, MAE)	5.56 ± 0.04	5.10 ± 0.04	4.92 ± 0.05	4.74 ± 0.03	4.59 ± 0.03
PW-MORD($\boldsymbol{wb}$, MAE)	**5.52 ± 0.06**	**5.07 ± 0.05**	4.97 ± 0.06	4.96 ± 0.06	4.87 ± 0.06

WILD

	$m = 3300$	$m = 6600$	$m = 11000$	$m = 16000$	$m = 21000$
SVOR-IMC($\boldsymbol{w}$)	10.30 ± 0.11	9.51 ± 0.16	9.09 ± 0.20	8.90 ± 0.11	8.74 ± 0.10
SVOR-IMC($\boldsymbol{w\theta}$)	10.18 ± 0.12	9.54 ± 0.16	9.17 ± 0.13	9.06 ± 0.08	8.96 ± 0.15
VILMA($\boldsymbol{w}$, MAE)	10.40 ± 0.13	9.60 ± 0.13	9.14 ± 0.12	8.89 ± 0.12	8.68 ± 0.12
VILMA($\boldsymbol{wb}$, MAE)	17.27 ± 0.77	17.21 ± 0.19	16.90 ± 0.23	16.74 ± 0.21	16.71 ± 0.12
MORD($\boldsymbol{w}$, MAE)	10.68 ± 0.05	10.01 ± 0.07	9.65 ± 0.07	9.46 ± 0.09	9.39 ± 0.08
MORD($\boldsymbol{wb}$, MAE)	17.22 ± 0.85	17.13 ± 0.24	16.78 ± 0.23	16.75 ± 0.19	16.69 ± 0.10
PW-MORD($\boldsymbol{w}$, MAE)	9.25 ± 0.17	8.45 ± 0.16	**7.86 ± 0.15**	**7.54 ± 0.10**	**7.36 ± 0.12**
PW-MORD($\boldsymbol{wb}$, MAE)	**9.08 ± 0.16**	**8.41 ± 0.13**	8.32 ± 0.16	8.25 ± 0.08	8.20 ± 0.07

Table 3.6. The test MAE of the ordinal classifier learned from the precisely annotated examples by the SVOR-IMC, VILMA, MORD and PW-MORD with different regularizer settings. The results are shown for the training sets generated from the MORPH and WILD databases by randomly selecting different number of the training examples m.

LinCls(reg,MAE) on the "California" data. This result is not surprising as well because the "Pyrimids" data has very few training examples, hence the simplest regression model avoids over-fitting best. On the other hand, the "California" data are low dimensional with a high number of training examples and thus the general and most flexible classification model learned by LinCls(reg,MAE) can describe the data without over-fitting best, i.e. the ordering prior imposed by the ordinal model is not needed in this case.

- **MORPH & WILD** The PW-MORD(reg,MAE) provides consistently best results on the WILD dataset which contains more complicated photographs than those from the MORPH dataset taken under controlled conditions. In the case of less complex MORPH dataset the benefits of the PW-MORD model are not that significant. Namely, PW-MORD(reg,MAE), SVOR-IMC($\boldsymbol{w}$,MAE) and VILMA($\boldsymbol{w}$,MAE) provide similar accuracy with difference around the level of the standard deviation.

To sum up, the experiments confirm our expectation that the PW-MORD model is beneficial when the data are complex and the total ordering imposed by the standard ordinal model is partially violated.

3.5.7. Learning from interval annotations

In this section we evaluate the proposed algorithm VILMA($\boldsymbol{w}$,MAE) when used for learning from interval annotations. We preformed experiments on the age estimation datasets MORPH and WILD. We optimized MAE as the target loss since it is the standard performance measure used in this application. In the experiment we varied the number of examples with the interval annotations m_I, the number of precisely annotated examples m_p as well as the width of the annotation interval u (c.f. Section 3.5.3 describing the experimental protocol). The obtained results are summarized in Table 3.7 and Figure 3.6. The main empirical findings are as follows:

- We observe that adding the partially annotated examples improves the accuracy monotonically. This observation holds true for all tested combinations of m_I, m_P, u and both databases. This observation is of a significant 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 narrowest (5-years) intervals. In this case, the performance of the classifier learned from the partial annotations matches closely the supervised setting. In particular, the loss in accuracy resulting from using the partial annotation on the WILD database is on the level of standard deviation. Even in the most challenging case, when learning from 20-years wide intervals, the results are practically useful. For example, to get classifier with ≈ 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.

3.5.8. Tightness of the upper bound on the Bayes risk

In Section 3.3.1 we showed that the target Bayes risk R^{MAE} can be upper-bounded by a linear function of the Bayes risk R^{ℓ_I} defined by the interval-insensitive loss. Recall that the proposed algorithm VILMA($\boldsymbol{w}$,MAE) minimizes a convex surrogate of the R^{ℓ_I}. In this section we evaluate tightness of the upper bound empirically. Let us define a quantity $\gamma(\alpha,\beta) = \hat{R}^{MAE}(h^{\alpha,\beta}) - \hat{R}^{MAE}(h^*)$, where $\hat{R}^{MAE}(\cdot)$ denotes the test MAE, $h^{\alpha,\beta}$ is the classifier learned by VILMA($\boldsymbol{w}$,MAE) from partially annotated examples generated by the $\alpha\beta$-precise annotation process and h^* is the classifier learned from the precise annotations only. The quantity $\gamma(\alpha,\beta)$ thus measures the loss in test accuracy caused by using the imprecise annotation. The values of $\gamma(\alpha,\beta)$ observed on both databases are shown in Figure 3.7. 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 $\gamma(\alpha,\beta) \leq (1-\alpha)\beta$ given in Theorem 3. Although the slope of real curve $\gamma(\alpha,\beta)$, if seen as a function of $1-\alpha$, is considerably smaller than β, the tendency is approximately linear at least in the regime $1-\alpha \in [0,0.5]$.

3.5.9. Comparison of surrogate losses for learning from interval annotations

In this section we compare four different algorithms, namely, VILMA($\boldsymbol{w}$,MAE), VILMA($\boldsymbol{w}$,0/1), II-SVOR-EXP($\boldsymbol{w}$) and II-SVOR-IMC($\boldsymbol{w}$). The methods use different convex surrogates which can be evaluated on interval annotated examples. The goal is evaluate which algorithm, or which surrogate, is the best approximation of the target MAE loss. The evaluation is done on the MORPH and the WILD datasets using the same protocol as described in Section 3.5.7. The results are summarized in:

- Table 3.7 showing results of VILMA($\boldsymbol{w}$,MAE).
- Table 3.8 showing results of II-SVOR-IMC($\boldsymbol{w}$).
- Table 3.9 showing results of VILMA($\boldsymbol{w}$,0/1).
- Table 3.10 showing results of II-SVOR-EXP($\boldsymbol{w}$).

The main empirical findings are as follows:

- The proposed method VILMA($\boldsymbol{w}$,MAE) clearly outperforms VILMA($\boldsymbol{w}$,0/1), i.e., the surrogate derived for MAE, being the target loss, is indeed better than the surrogate for 0/1-loss. The margin between the both methods becomes more clear as the number of examples grows, but it is not easy to see with small number of examples. The tendency holds for all considered widths of the annotation interval $u \in \{5, 10, 20\}$.
- Unlike in the case of VILMA(reg,loss), the difference between the performance of II-SVOR-IMC($\boldsymbol{w}$) and II-SVOR-EXP($\boldsymbol{w}$) does not depend on the number of examples used in the experiment. It only grows for all considered annotation interval widths $u \in \{5, 10, 20\}$ as number of examples in data increasing.

3.5.10. Equivalence between SVOR-IMC and VILMA-MAE

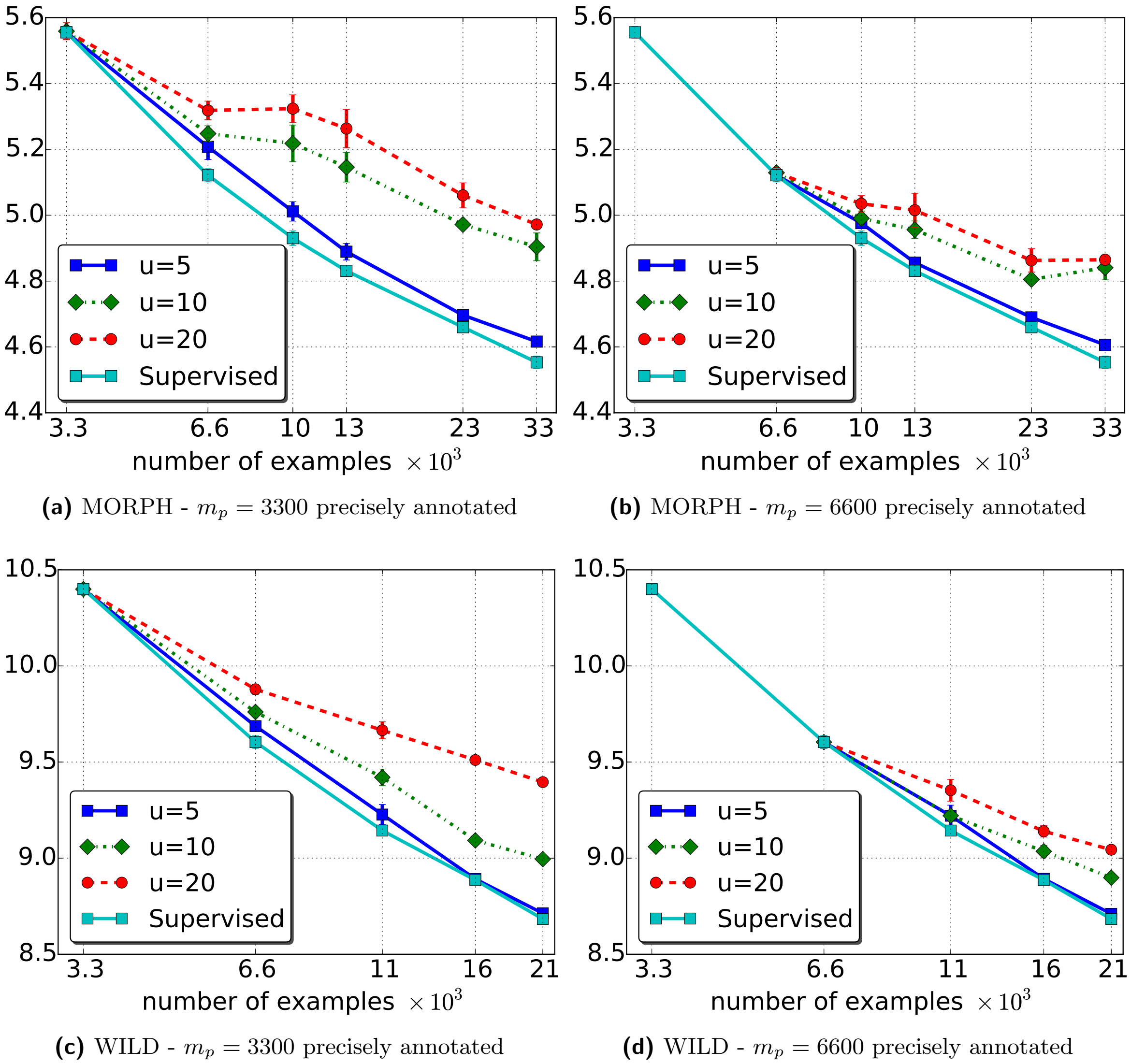
Although the VILMA(reg,MAE) and the SVOR-IMC(reg) learn different parametrizations of the ordinal classifier, the resulting rules are equivalent (up to numerical errors) as predicted by Proposition 2. The equivalence can be verified empirically as seen from comparing the results in Table 3.7 and Table 3.8. It is seen that the test accuracy of both methods differs only on the level of standard deviation.

3.6. Conclusions

We have established relationship between the classification rule used in the ordinal regression and a class of linear multi-class classifiers. The established relationship has the following benefits. First, it allows to understand various classification models better. Second, it provides a path to develop new learning algorithms for ordinal regression borrowing from well understand multi-class classification, e.g. by adopting the generic SO-SVM framework as we showed. Third, it allows to design new more flexible models for ordinal regression with higher discriminative power, e.g the proposed PW-MORD model. A functionality of the proposed methods has been successfully shown on standard benchmarks as well as on a real-life problem of estimating the human age from facial images.

We have proposed a V-shaped interval-insensitive loss 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 showed how to construct a convex surrogate of the interval-insensitive loss instantiated for an arbitrary (target) V-shaped loss. We also derived other convex surrogate losses of the interval insensitive loss by extending the existing supervised methods like the SVOR-EXP and SVOR-IMC algorithm. We derived a generic algorithm VILMA which translates learning from the interval annotations to a convex optimization problem. We have proposed a generic cutting plane solver allowing to impose a quadratic regularization on a subset of parameters which turned out to be important in practice. The experiments conducted on a real-life problem of human age estimation from facial images show that the proposed method has a practical potential. We demonstrated that a precise ordinal classifier with accuracy matching the state-of-the-art results can be obtained by learning from cheap partial annotations.

Our work is based on the interval insensitive loss and its convex surrogates, which turned out to work well empirically. We showed that under certain assumptions the expectation of the interval insensitive loss can be used to upper bound expectation of the associated target loss. However a deeper theoretical understanding is needed. For example, an open issue is whether there exists a distribution, for which the upper bound is sharp. Another interesting question is how to weaken the assumptions on the annotation process, e.g. the requirement on the consistency of the annotation. It is also unclear, which of the introduced convex surrogates is better theoretically. We believe that this issue could be resolved by analyzing statistical consistency of the surrogates which remains the open question.



(a) MORPH - $m_p = 3300$ precisely annotated

(b) MORPH - $m_p = 6600$ precisely annotated

(c) WILD - $m_p = 3300$ precisely annotated

(d) WILD - $m_p = 6600$ precisely annotated

Figure 3.6. Figures show test MAE for the ordinal classifiers learned by the VILMA($\boldsymbol{w}$, MAE) 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).

MORPH

		$m = 3300$	$m = 6600$	$m = 13000$	$m = 23000$	$m = 33000$
	Supervised	5.56 ± 0.02	5.12 ± 0.02	4.83 ± 0.02	4.66 ± 0.01	4.55 ± 0.02
m_P	u	$m_I = 0$	$m_I = 3300$	$m_I = 9700$	$m_I = 19700$	$m_I = 29700$
3300	5	5.56 ± 0.02	5.21 ± 0.04	4.89 ± 0.03	4.70 ± 0.01	4.62 ± 0.01
	10	5.56 ± 0.03	5.25 ± 0.02	5.15 ± 0.05	4.97 ± 0.01	4.90 ± 0.04
	20	5.56 ± 0.03	5.32 ± 0.03	5.26 ± 0.06	5.06 ± 0.04	4.97 ± 0.01
m_P	u	$m_I = 0$	$m_I = 0$	$m_I = 6400$	$m_I = 16400$	$m_I = 26400$
6600	5	—	5.12 ± 0.02	4.86 ± 0.02	4.69 ± 0.00	4.61 ± 0.00
	10	—	5.13 ± 0.02	4.96 ± 0.03	4.81 ± 0.01	4.84 ± 0.04
	20	—	5.13 ± 0.02	5.03 ± 0.02	4.86 ± 0.04	4.86 ± 0.01

WILD

		$m = 3300$	$m = 6600$	$m = 11000$	$m = 16000$	$m = 21000$
	Supervised	10.40 ± 0.13	9.60 ± 0.13	9.14 ± 0.12	8.89 ± 0.12	8.68 ± 0.12
m_P	u	$m_I = 0$	$m_I = 3300$	$m_I = 7700$	$m_I = 12700$	$m_I = 17700$
3300	5	10.40 ± 0.13	9.69 ± 0.12	9.23 ± 0.15	8.89 ± 0.12	8.71 ± 0.12
	10	10.40 ± 0.13	9.76 ± 0.12	9.42 ± 0.14	9.09 ± 0.12	8.99 ± 0.12
	20	10.40 ± 0.13	9.88 ± 0.13	9.67 ± 0.14	9.51 ± 0.10	9.40 ± 0.11
m_P	u	$m_I = 0$	$m_I = 0$	$m_I = 4400$	$m_I = 9400$	$m_I = 14400$
6600	5	—	9.60 ± 0.13	9.22 ± 0.16	8.89 ± 0.12	8.71 ± 0.12
	10	—	9.60 ± 0.13	9.22 ± 0.12	9.04 ± 0.13	8.90 ± 0.12
	20	—	9.60 ± 0.13	9.35 ± 0.16	9.14 ± 0.13	9.04 ± 0.12

Table 3.7. The table summarizes test MAE of the ordinal classifier learned from the training set with m examples using VILMA($\boldsymbol{w}$, MAE). 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.

MORPH

		$m = 3300$	$m = 6600$	$m = 13000$	$m = 23000$	$m = 33000$
	Supervised	5.54 ± 0.03	5.10 ± 0.02	4.83 ± 0.01	4.69 ± 0.03	4.61 ± 0.03
m_P	u	$m_I = 0$	$m_I = 3300$	$m_I = 9700$	$m_I = 19700$	$m_I = 29700$
3300	5	5.54 ± 0.03	5.20 ± 0.04	4.93 ± 0.02	4.82 ± 0.03	4.76 ± 0.03
	10	5.54 ± 0.03	5.25 ± 0.02	4.99 ± 0.02	4.95 ± 0.05	4.98 ± 0.02
	20	5.54 ± 0.03	5.34 ± 0.02	5.14 ± 0.03	5.07 ± 0.04	4.97 ± 0.01
m_P	u	$m_I = 0$	$m_I = 0$	$m_I = 6400$	$m_I = 16400$	$m_I = 26400$
6600	5	—	5.10 ± 0.02	4.88 ± 0.02	4.76 ± 0.01	4.73 ± 0.03
	10	—	5.10 ± 0.02	4.93 ± 0.02	4.81 ± 0.02	4.84 ± 0.04
	20	—	5.10 ± 0.02	5.00 ± 0.03	4.86 ± 0.03	4.86 ± 0.01

WILD

		$m = 3300$	$m = 6600$	$m = 11000$	$m = 16000$	$m = 21000$
	Supervised	10.30 ± 0.11	9.51 ± 0.16	9.09 ± 0.20	8.90 ± 0.11	8.74 ± 0.10
m_P	u	$m_I = 0$	$m_I = 3300$	$m_I = 7700$	$m_I = 12700$	$m_I = 17700$
3300	5	10.30 ± 0.11	9.57 ± 0.16	9.30 ± 0.10	9.00 ± 0.12	8.85 ± 0.13
	10	10.30 ± 0.11	9.65 ± 0.15	9.34 ± 0.17	9.10 ± 0.12	8.99 ± 0.08
	20	10.30 ± 0.11	9.87 ± 0.16	9.65 ± 0.21	9.49 ± 0.15	9.29 ± 0.11
m_P	u	$m_I = 0$	$m_I = 3300$	$m_I = 7700$	$m_I = 12700$	$m_I = 17700$
6600	5	—	9.51 ± 0.16	9.17 ± 0.15	9.05 ± 0.09	8.86 ± 0.14
	10	—	9.51 ± 0.16	9.21 ± 0.15	8.99 ± 0.11	8.89 ± 0.16
	20	—	9.51 ± 0.16	9.28 ± 0.15	9.15 ± 0.17	9.06 ± 0.16

Table 3.8. The table summarizes test MAE of the ordinal classifier learned from the training set with m examples using II-SVOR-IMC($\boldsymbol{w}$). 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.

MORPH

		$m = 3300$	$m = 6600$	$m = 13000$	$m = 23000$	$m = 33000$
	Supervised	5.72 ± 0.05	5.20 ± 0.02	5.02 ± 0.01	4.96 ± 0.04	4.92 ± 0.00
m_P	u	$m_I = 0$	$m_I = 3300$	$m_I = 9700$	$m_I = 19700$	$m_I = 29700$
3300	5	5.72 ± 0.05	5.16 ± 0.01	4.91 ± 0.02	4.79 ± 0.02	4.74 ± 0.02
	10	5.72 ± 0.05	5.23 ± 0.05	5.00 ± 0.04	4.91 ± 0.03	4.91 ± 0.04
	20	5.72 ± 0.05	5.40 ± 0.03	5.32 ± 0.04	5.22 ± 0.04	5.19 ± 0.05
m_P	u	$m_I = 0$	$m_I = 0$	$m_I = 6400$	$m_I = 16400$	$m_I = 26400$
6600	5	—	5.20 ± 0.02	4.86 ± 0.02	4.89 ± 0.03	4.72 ± 0.02
	10	—	5.20 ± 0.02	4.96 ± 0.03	4.95 ± 0.03	4.80 ± 0.03
	20	—	5.20 ± 0.02	5.03 ± 0.02	5.20 ± 0.06	5.08 ± 0.05

WILD

		$m = 3300$	$m = 6600$	$m = 11000$	$m = 16000$	$m = 21000$
	Supervised	10.31 ± 0.25	9.46 ± 0.17	9.12 ± 0.13	9.05 ± 0.16	9.00 ± 0.16
m_P	u	$m_I = 0$	$m_I = 3300$	$m_I = 7700$	$m_I = 12700$	$m_I = 17700$
3300	5	10.31 ± 0.25	9.52 ± 0.15	9.10 ± 0.11	8.93 ± 0.07	8.86 ± 0.02
	10	10.31 ± 0.25	9.52 ± 0.15	9.21 ± 0.12	9.06 ± 0.09	8.98 ± 0.02
	20	10.31 ± 0.25	9.75 ± 0.11	9.70 ± 0.15	9.48 ± 0.17	9.46 ± 0.01
m_P	u	$m_I = 0$	$m_I = 0$	$m_I = 4400$	$m_I = 9400$	$m_I = 14400$
6600	5	—	9.46 ± 0.17	9.33 ± 0.13	9.23 ± 0.11	9.15 ± 0.14
	10	—	9.46 ± 0.17	9.29 ± 0.19	9.03 ± 0.10	8.96 ± 0.11
	20	—	9.46 ± 0.17	9.43 ± 0.06	9.18 ± 0.08	9.31 ± 0.18

Table 3.9. The table summarizes test MAE of the ordinal classifier learned with help VILMA($\boldsymbol{w}$, 0/1) 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.

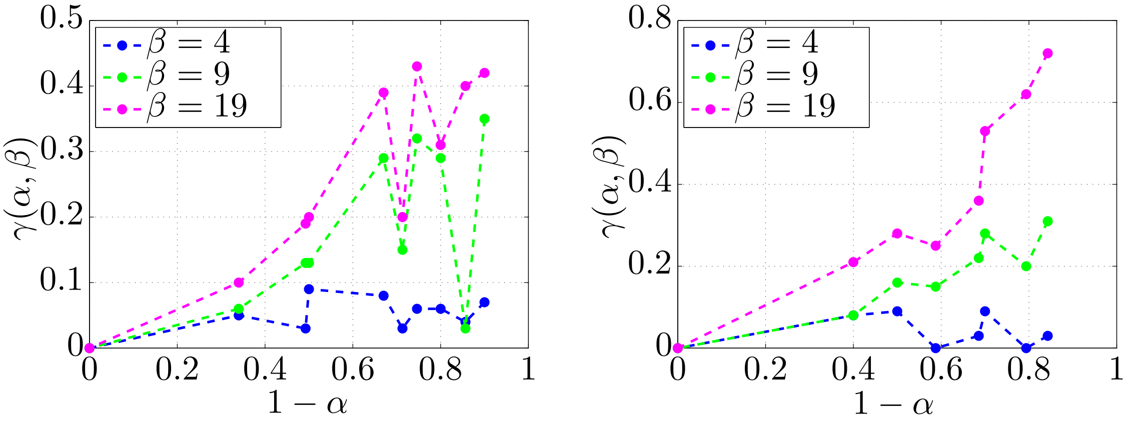
MORPH

		$m = 3300$	$m = 6600$	$m = 13000$	$m = 23000$	$m = 33000$
	Supervised	5.52 ± 0.03	5.29 ± 0.03	5.19 ± 0.02	5.17 ± 0.01	5.16 ± 0.03
m_P	u	$m_I = 0$	$m_I = 3300$	$m_I = 9700$	$m_I = 19700$	$m_I = 29700$
3300	5	5.52 ± 0.03	5.26 ± 0.07	5.01 ± 0.05	4.86 ± 0.05	4.83 ± 0.04
	10	5.52 ± 0.03	5.44 ± 0.05	5.10 ± 0.02	4.93 ± 0.02	4.85 ± 0.02
	20	5.52 ± 0.03	5.57 ± 0.03	5.59 ± 0.02	5.29 ± 0.08	5.16 ± 0.07
m_P	u	$m_I = 0$	$m_I = 0$	$m_I = 6400$	$m_I = 16400$	$m_I = 26400$
6600	5	—	5.29 ± 0.03	4.98 ± 0.03	4.83 ± 0.04	4.80 ± 0.03
	10	—	5.29 ± 0.03	5.15 ± 0.04	4.93 ± 0.03	4.83 ± 0.04
	20	—	5.29 ± 0.03	5.42 ± 0.07	5.30 ± 0.08	5.16 ± 0.07

WILD

		$m = 3300$	$m = 6600$	$m = 11000$	$m = 16000$	$m = 21000$
	Supervised	10.35 ± 0.14	9.64 ± 0.22	9.35 ± 0.14	9.28 ± 0.11	9.26 ± 0.18
m_P	u	$m_I = 0$	$m_I = 3300$	$m_I = 7700$	$m_I = 12700$	$m_I = 17700$
3300	5	10.35 ± 0.14	9.65 ± 0.17	9.26 ± 0.13	9.05 ± 0.12	8.99 ± 0.12
	10	10.35 ± 0.14	9.80 ± 0.13	9.63 ± 0.33	9.20 ± 0.08	9.11 ± 0.06
	20	10.35 ± 0.14	9.97 ± 0.07	10.11 ± 0.19	9.79 ± 0.06	9.66 ± 0.06
m_P	u	$m_I = 0$	$m_I = 3300$	$m_I = 7700$	$m_I = 12700$	$m_I = 17700$
6600	5	—	9.64 ± 0.22	9.24 ± 0.16	9.06 ± 0.13	9.01 ± 0.15
	10	—	9.64 ± 0.22	9.40 ± 0.15	9.22 ± 0.08	9.16 ± 0.06
	20	—	9.64 ± 0.22	9.59 ± 0.12	9.72 ± 0.14	9.76 ± 0.04

Table 3.10. The table summarizes test MAE of the ordinal classifier learned from the training set with m examples using II-SVOR-EXP($\boldsymbol{w}$). 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.



(a) MORPH

(b) WILD

Figure 3.7. 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 β (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.

4. Statistical consistency of structured output learning with missing labels

A roadmap of the chapter:

- **Supervised learning** of the structured output classifiers based on the ERM principle is described in Section 4.1. The main purpose of this section is to defined the class of considered structured output classifiers and the tasks of learning classifier from the completely annotated examples.
- **Minimization of the partial loss** is a subject of Section 4.2. In this section we define the task of learning from examples with missing annotation of a subset of labels and we introduce the concept of the partial loss.
- **Statistical model of partial annotations** is defined in Section 4.3. In this section we provide sufficient conditions on the data generating distribution which admit to prove the consistency of algorithms based on minimization of the partial loss.
- **Consistency of the partial loss** is stated in Section 4.4. In this section we introduce the main theorem which claims that minimization of the partial loss is equivalent to the minimization of the target (complete) loss in the sense that both problems have the same minimizers. Here by minimizers we mean sequences of classifiers converging in probability.
- **The classification calibrated surrogate of the partial loss** is defined in Section 4.5. This section combines the consistency of the partial loss stated in Section 4.4 and the results of [Ramaswamy and Agarwal, 2012] in order to show that minimizing a calibrated surrogate of the partial loss remains statistically consistency.
- **Existence of a convex surrogate of the partial loss** is stated in Section 4.6. The existence of the convex surrogate is another outcome derived from connecting the consistency of the partial loss and the existing results for supervised learning presented in [Ramaswamy and Agarwal, 2012].
- **Examples of surrogate losses** are discusses in Section 4.7. We analyze the surrogate losses which are in the core of many existing methods, namely, we analyzed the method published in: [Chuong et al., 2008; Girshick et al., 2011; Yu and Joachims, 2009; Fernandes and Brefeld, 2011a; Zhu et al., 2010; Vedaldi and Zisserman, 2009; Wang and Mori, 2010; Luo and Orabona, 2010; Lou and Hamprecht, 2012; Sarawagi and Gupta, 2008; Yu et al., 2014]. We show which surrogates are classification calibrated and which are not.
- **Relation to existing works** is discussed in Section 4.8.
- **Conclusions** are drawn in Section 4.9.

4.1. Supervised learning of structured output classifiers

Let $\mathcal{X}$ be an input space, $\mathcal{V}$ a finite set of local parts and $\mathcal{Y}$ a finite set of labels. An object is fully characterized by an input (observation) $\boldsymbol{x} \in \mathcal{X}$ and a labelling $\boldsymbol{y} = (y_v \in \mathcal{Y} \mid v \in \mathcal{V})$ of local parts $\mathcal{V}$. In the supervised setting, we are given the training set

$$\mathcal{D}^m_{\boldsymbol{xy}} = \{(\boldsymbol{x}^1, \boldsymbol{y}^1), \ldots, (\boldsymbol{x}^m, \boldsymbol{y}^m)\} \in (\mathcal{X} \times \mathcal{Y}^{\mathcal{V}})^m$$

drawn from i.i.d. random variables with distribution $p(\boldsymbol{x},\boldsymbol{y})$ defined over $\mathcal{X}\times\mathcal{Y}^{\mathcal{V}}$. We want to design a decision function $\boldsymbol{h}\colon\mathcal{X}\to\mathcal{T}^{\mathcal{V}}$, which maps an input $\boldsymbol{x}\in\mathcal{X}$ to a vector of decisions $\boldsymbol{t}=(t_v\in\mathcal{T}\mid v\in\mathcal{V})\in\mathcal{T}^{\mathcal{V}}$. We assume that the decision set $\mathcal{T}$ for each local part is finite. For example, in the most typical setting $\mathcal{T}=\mathcal{Y}$ and $\boldsymbol{h}$ is the structured output classifier predicting directly the labels. Note that $\mathcal{T}$ can be different from $\mathcal{Y}$ in general. For example, in the case of the classification with the reject option $\mathcal{T}=\mathcal{Y}\cup\{\text{don't know}\}$.

Let $\ell\colon\mathcal{Y}^{\mathcal{V}}\times\mathcal{T}^{\mathcal{V}}\to\mathbb{R}_+$ be a given loss function assigning a non-negative number to each pair of labelling $\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}}$ and a decision $\boldsymbol{t}\in\mathcal{T}^{\mathcal{V}}$. We confine ourselves to losses additive over the local parts which is a natural choice in many applications, i.e.

$$\ell(\boldsymbol{y},\boldsymbol{t})=\sum_{v\in\mathcal{V}}\ell_v(y_v,t_v)\,, \tag{4.1}$$

where $\ell_v\colon\mathcal{Y}\times\mathcal{T}\to\mathbb{R}_+$, $v\in\mathcal{V}$, are single label losses. We assume that ℓ_v are bounded and non-trivial, i.e. $\ell_v(y,t)<\infty$ and $\forall y\,\exists t$ such that $\ell_v(y,t)>0$. An example of a frequently used additive loss is the Hamming loss obtained when $\mathcal{T}=\mathcal{Y}$ and $\ell_v(y_v,t_v)=[\![y_v\neq t_v]\!]$, $v\in\mathcal{V}$. A decision function $\boldsymbol{h}$ is then evaluated by the ℓ-risk

$$R^\ell(\boldsymbol{h};p)=\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\,\ell(\boldsymbol{y},\boldsymbol{h}(\boldsymbol{x}))=\mathbb{E}_{p(\boldsymbol{x})}\sum_{\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}}}p(\boldsymbol{y}\mid\boldsymbol{x})\,\ell(\boldsymbol{y},\boldsymbol{h}(\boldsymbol{x}))=\mathbb{E}_{p(\boldsymbol{x})}\,\boldsymbol{p}_{\boldsymbol{y}}(\boldsymbol{x})^\top\boldsymbol{\ell}_{\boldsymbol{h}(\boldsymbol{x})}\,,$$

where $\boldsymbol{p}_{\boldsymbol{y}}(\boldsymbol{x})=(p(\boldsymbol{y}\mid\boldsymbol{x})\mid\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}})$ is a vector function denoting the conditional probabilities at $\boldsymbol{x}$ and $\boldsymbol{\ell}_{\boldsymbol{t}}=(\ell(\boldsymbol{y},\boldsymbol{t})\mid\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}})$ is a vector of losses for the decision $\boldsymbol{t}\in\mathcal{T}^{\mathcal{V}}$. The ultimate goal is to learn from $\mathcal{D}_m$ a decision function with the ℓ-risk close to the Bayes ℓ-risk

$$R^\ell_*(p)=\inf_{\boldsymbol{h}\colon\mathcal{X}\to\mathcal{T}^{\mathcal{V}}}R^\ell(\boldsymbol{h};p)=\mathbb{E}_{p(\boldsymbol{x})}\min_{\boldsymbol{t}\in\mathcal{T}^{\mathcal{V}}}\boldsymbol{p}_{\boldsymbol{y}}(\boldsymbol{x})^\top\boldsymbol{\ell}_{\boldsymbol{t}}\,.$$

A direct minimization of the loss $\boldsymbol{\ell}$ is often a hard problem. Therefore it is common to replace $\ell\colon\mathcal{Y}^{\mathcal{V}}\times\mathcal{T}^{\mathcal{V}}\to\mathbb{R}_+$ by a surrogate loss function $\psi\colon\mathcal{Y}^{\mathcal{T}}\times\hat{\mathcal{T}}\to\mathbb{R}_+$, which operates on a surrogate decision set $\hat{\mathcal{T}}\subseteq\mathbb{R}^d$. The goal is then to learn a function $\boldsymbol{f}\colon\mathcal{X}\to\hat{\mathcal{T}}$ minimizing the ψ-risk

$$R^\psi(\boldsymbol{f};p)=\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\,\psi(\boldsymbol{y},\boldsymbol{f}(\boldsymbol{x}))=\mathbb{E}_{p(\boldsymbol{x})}\sum_{\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}}}p(\boldsymbol{y}\mid\boldsymbol{x})\psi(\boldsymbol{y},\boldsymbol{f}(\boldsymbol{x}))=\mathbb{E}_{p(\boldsymbol{x})}\,\boldsymbol{p}_{\boldsymbol{y}}(\boldsymbol{x})^\top\boldsymbol{\psi}_{\boldsymbol{f}(\boldsymbol{x})}\,,$$

where $\boldsymbol{\psi}_{\hat{\boldsymbol{t}}}=(\psi(\boldsymbol{y},\hat{\boldsymbol{t}})\mid\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}})$ is a vector of proxy losses at the decision $\hat{\boldsymbol{t}}\in\hat{\mathcal{T}}$. The learned function $\boldsymbol{f}$ is used to construct the decision function via a transform $\text{pred}\colon\hat{\mathcal{T}}\to\mathcal{T}$. The ℓ-risk of the resulting decision function $\text{pred}(\boldsymbol{f}(\boldsymbol{x}))$ is $R^\ell(\text{pred}\circ\boldsymbol{f};p)$. For example, $\boldsymbol{f}(\boldsymbol{x})=(\langle\boldsymbol{w},\boldsymbol{\Psi}(\boldsymbol{x},\boldsymbol{y})\rangle\mid\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}})$ is a vector of scores linear in parameters $\boldsymbol{w}\in\mathbb{R}^n$ and $\text{pred}(\hat{\boldsymbol{t}})\in\text{Argmax}_{\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}}}\hat{t}_{\boldsymbol{y}}$, which yields the linear structured output classifier $\boldsymbol{h}(\boldsymbol{x})\in\text{Argmax}_{\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}}}\langle\boldsymbol{w},\boldsymbol{\Psi}(\boldsymbol{x},\boldsymbol{y})\rangle$.

Under suitable conditions the uniform law of large numbers applies (e.g. [Vapnik, 1998]) and learning $\boldsymbol{f}_m$ from $\mathcal{D}^m_{\boldsymbol{xy}}$ by minimizing the empirical risk $R^\psi_{\text{emp}}(\boldsymbol{f})=\frac{1}{m}\sum_{i=1}^m\psi(\boldsymbol{y}^i,\boldsymbol{x}^i)$ is statistically consistent, i.e. for the number of examples m going to infinity, $R^\psi(\boldsymbol{f}_m;p)$ converges in probability to the minimal (Bayes) ψ-risk

$$R^\psi_*(p)=\inf_{f\colon\mathcal{X}\to\hat{\mathcal{T}}}R^\psi(\boldsymbol{f};p)\,.$$

It has been shown (e.g. [Zhang, 2004a; Tewari and Bartlett, 2007; Gao and Zhou, 2013]) that the consistency with respect to the ψ-risk implies the consistency with respect to the ℓ-risk provided the surrogate loss ψ is so called classification calibrated w.r.t the loss ℓ. In this chaper, we will extend this result to the setting when the training examples are partially annotated as defined in the next section.

4.2. Minimization of the partial loss

Let us consider that we are given a training set

$$\mathcal{D}^m_{\boldsymbol{x}\boldsymbol{a}} = \{(\boldsymbol{x}^1, \boldsymbol{a}^1), \ldots, (\boldsymbol{x}^m, \boldsymbol{a}^m)\} \in (\mathcal{X} \times \mathcal{A}^{\mathcal{V}})^m$$

drawn from i.i.d. random variables with the distribution

$$p''(\boldsymbol{x}, \boldsymbol{a}) = \sum_{\boldsymbol{y} \in \mathcal{Y}^{\mathcal{V}}} p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{a}) \;,$$

where $\mathcal{A} = \{\mathcal{Y} \cup \{\mathcal{Y}\}\}$ denotes a set of admissible annotations of a local part and $p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{a})$ is a properly defined distribution over $\mathcal{X} \times \mathcal{Y}^{\mathcal{V}} \times \mathcal{A}^{\mathcal{V}}$. At a given part $v \in \mathcal{V}$, the label is either known $a_v \in \mathcal{Y}$ or missing $a_v = \mathcal{Y}$ meaning that all labels are possible. The partial annotation of the i-th training instance is a vector $\boldsymbol{a}^i = (a^i_v \in \mathcal{A} \mid v \in \mathcal{V})$ assigning labels to the local parts $\mathcal{V}^i_{\text{known}} = \{v \in \mathcal{V} \mid |a^i_v| = 1\}$ while the labels of the remaining local parts $\mathcal{V} \setminus \mathcal{V}^i_{\text{known}}$ are missing.

The distribution $p'(\boldsymbol{x}, \boldsymbol{y})$ over input-label space $\mathcal{X} \times \mathcal{Y}^{\mathcal{V}}$ can be obtained from $p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{a})$ by marginalization over the annotations $\mathcal{A}^{\mathcal{V}}$, i.e.,

$$p'(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\boldsymbol{a} \in \mathcal{A}^{\mathcal{V}}} p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{a}) \;.$$

Our *ultimate goal* is to learn from $\mathcal{D}^m_{\boldsymbol{x}\boldsymbol{a}}$ a decision function $\boldsymbol{h} \colon \mathcal{X} \to \mathcal{T}^{\mathcal{V}}$ with ℓ-risk $R^{\ell}(\boldsymbol{h}; p')$ close to the Bayes ℓ-risk $R^{\ell}_*(p')$. It is important to stress that the objective (i.e. the ℓ-risk) of learning from the missing labels analyzed in this paper is exactly the same as the objective in the conventional fully supervised setting, however, the annotation of the training examples is different.

In order to make learning from missing labels possible, we define a partial loss:

Definition 7. *For a given (complete) additive loss $\ell \colon \mathcal{Y}^{\mathcal{V}} \times \mathcal{T}^{\mathcal{V}} \to \mathbb{R}_+$ defined by (4.1) the associated partial loss $\ell^p \colon \mathcal{A}^{\mathcal{V}} \times \mathcal{T}^{\mathcal{V}} \to \mathbb{R}_+$ is defined as*

$$\ell^p(\boldsymbol{a}, \boldsymbol{t}) = \sum_{v \in \mathcal{V}} [\![|a_v| = 1]\!] \ell_v(a_v, t_v)^1 \;, \tag{4.2}$$

where $\ell_v \colon \mathcal{Y} \times \mathcal{T} \to \mathbb{R}_+$, $v \in \mathcal{V}$, are the same single label losses used to define the complete loss ℓ.

[1] Strictly speaking the correct formula here is $\ell^p(\boldsymbol{a}, \boldsymbol{t}) = \sum_{v \in \{v' \in \mathcal{V} \mid \, |a'_v| = 1\}} \ell_v(a_v, t_v)$. However for the sake of simplicity we slightly abuse the notation.

The partial loss ℓ^p simply neglects the local losses corresponding to the missing labels. Note that the partial loss for missing labels (4.2) is an instance of the generic partial loss (1.13). We can now learn a decision function $\boldsymbol{h}\colon \mathcal{X} \to \mathcal{T}^{\mathcal{V}}$ by minimizing the ℓ^p-risk

$$R^{\ell^p}(\boldsymbol{h}; p'') = \mathbb{E}_{p''(\boldsymbol{x},\boldsymbol{a})}\, \ell^p(\boldsymbol{a}, \boldsymbol{h}(\boldsymbol{x})) = \mathbb{E}_{p(\boldsymbol{x})} \sum_{\boldsymbol{a}\in\mathcal{A}^{\mathcal{V}}} p(\boldsymbol{a} \mid \boldsymbol{x}) \ell^p(\boldsymbol{a}, \boldsymbol{h}(\boldsymbol{x})) = \mathbb{E}_{p(\boldsymbol{x})}\, \boldsymbol{p}_{\boldsymbol{a}}(\boldsymbol{x})^\top \boldsymbol{\ell}^p_{\boldsymbol{h}(\boldsymbol{x})}\,,$$

where $\boldsymbol{p}_{\boldsymbol{a}}(\boldsymbol{x}) = (p(\boldsymbol{a} \mid \boldsymbol{x}) \mid \boldsymbol{a} \in \mathcal{A}^{\mathcal{V}})$ is a vector function denoting the conditional probabilities at $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\ell}^p_{\boldsymbol{t}} = (\ell^p(\boldsymbol{a}, \boldsymbol{t}) \mid \boldsymbol{a} \in \mathcal{A}^{\mathcal{V}})$ is a vector of partial losses for the decision $\boldsymbol{t} \in \mathcal{Y}^{\mathcal{V}}$. The Bayes ℓ^p-risk is defined as

$$R^{\ell^p}_*(p'') = \inf_{\boldsymbol{h}\colon \mathcal{X}\to\mathcal{T}^{\mathcal{V}}} R^{\ell^p}(\boldsymbol{h}; p'') = \mathbb{E}_{p(\boldsymbol{x})} \min_{\boldsymbol{t}\in\mathcal{T}^{\mathcal{V}}} \boldsymbol{p}_{\boldsymbol{a}}(\boldsymbol{x})^\top \boldsymbol{\ell}^p_{\boldsymbol{t}}\,.$$

It is clear that learning from the partial annotations is not possible without imposing constraints on the distribution $p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{a})$. For example, when $p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{a}) = p(\boldsymbol{x}, \boldsymbol{y})\, p(\boldsymbol{a})$ the annotations carry no information about the labels and hence learning is not possible. In the next section we provide sufficient conditions on $p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{a})$ which allow to prove that minimization of the ℓ^p-risk is equivalent to the minimization of the ℓ-risk.

4.3. Statistical model of partial annotations

In this section, we describe a generative model of the partially annotated data which will be used later in this chapter. The standard model $p(\boldsymbol{x}, \boldsymbol{y})$ is defined over the input-label space $\mathcal{X} \times \mathcal{Y}^{\mathcal{V}}$. We augment the standard model by additional binary random variables $\boldsymbol{z} = (z_v \in \{0, 1\} \mid v \in \mathcal{V}) \in \mathcal{Z}^{\mathcal{V}}$ assumed to be a realization of a random field distributed according to $p(\boldsymbol{z} \mid \boldsymbol{x})$. The binary variables $\boldsymbol{z} \in \mathcal{Z}^{\mathcal{V}}$ determine, which labels in $\boldsymbol{y} = (y_v \in \mathcal{Y} \mid v \in \mathcal{V})$ are annotated. Specifically, $z_v = 1$ means that the local part v is annotated, while $z_v = 0$ means that the label is missing. The annotation $\boldsymbol{a} \in \mathcal{A}^{\mathcal{V}}$ is created from $\boldsymbol{y}$ and $\boldsymbol{z}$ by copying those labels which are annotated, or formally via a vector function $\boldsymbol{\alpha}\colon \mathcal{Y}^{\mathcal{V}} \times \mathcal{Z}^{\mathcal{V}} \to \mathcal{A}^{\mathcal{V}}$ defined as $\boldsymbol{a} = (a_1, \ldots, a_{|\mathcal{V}|}) = \boldsymbol{\alpha}(\boldsymbol{y}, \boldsymbol{z}) = \big(\alpha(y_1, z_1), \cdots, \alpha(y_{|\mathcal{V}|}, z_{|\mathcal{V}|})\big)$ where

$$a_v = \alpha(y_v, z_v) = \begin{cases} y_v & \text{if} \quad z_v = 1\,, \\ \mathcal{Y} & \text{if} \quad z_v = 0\,. \end{cases}$$

We assume that the random variables $\boldsymbol{y}$ and $\boldsymbol{z}$ are conditionally independent, i.e.

$$p(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{x}) = p(\boldsymbol{y} \mid \boldsymbol{x})\, p(\boldsymbol{z} \mid \boldsymbol{x})\,, \tag{4.3}$$

which implies that for fixed $\boldsymbol{x}$ the annotation $\boldsymbol{a}$ is distributed according to

$$p(\boldsymbol{a} \mid \boldsymbol{x}) = \sum_{\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}}} \sum_{\boldsymbol{z}\in\mathcal{Z}^{\mathcal{V}}} p(\boldsymbol{y} \mid \boldsymbol{x})\, p(\boldsymbol{z} \mid \boldsymbol{x})\, [\![\boldsymbol{a} = \boldsymbol{\alpha}(\boldsymbol{y}, \boldsymbol{z})]\!]\,. \tag{4.4}$$

The model described above defines a random process generating a set of partially annotated examples according to the distribution

$$p(\boldsymbol{x}, \boldsymbol{a}) = p(\boldsymbol{x})\, p(\boldsymbol{a} \mid \boldsymbol{x})\,. \tag{4.5}$$

Let as define a function $c\colon \mathcal{Y}^{\mathcal{V}} \times \mathcal{A}^{\mathcal{V}} \to \{0,1\}$ as

$$c(\boldsymbol{y},\boldsymbol{a}) = \prod_{v\in\mathcal{V}} [\![y_v \in \boldsymbol{a}_v]\!] \,,$$

which evaluates to 1 if the labeling $\boldsymbol{y}$ is consistent with the annotation $\boldsymbol{a}$ and it is 0 otherwise. It is not difficult to show that

$$p(\boldsymbol{y}\,|\,\boldsymbol{x},\boldsymbol{a}) = \frac{p(\boldsymbol{y}\,|\,\boldsymbol{x})\, c(\boldsymbol{y},\boldsymbol{a})}{\sum_{\boldsymbol{y}'\in\mathcal{Y}^{\mathcal{V}}} p(\boldsymbol{y}'\,|\,\boldsymbol{x})\, c(\boldsymbol{y}',\boldsymbol{a})} \,, \tag{4.6}$$

We use the convention that $p(\boldsymbol{y}\,|\,\boldsymbol{x},\boldsymbol{a}) = 0$ if the denominator and the numerator are zero. The distribution (4.6) together with (4.5) defines a joint distribution

$$p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{a}) = p(\boldsymbol{x})\; p(\boldsymbol{a} \mid \boldsymbol{x})\; p(\boldsymbol{y}\,|\,\boldsymbol{x},\boldsymbol{a}) \tag{4.7}$$

describing dependency of the random variables $(\boldsymbol{x},\boldsymbol{y},\boldsymbol{a}) \in \mathcal{X} \times \mathcal{Y}^{\mathcal{V}} \times \mathcal{A}^{\mathcal{V}}$.

Definition 8. *A distribution $p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{a})$ defined over $\mathcal{X} \times \mathcal{Y}^{\mathcal{V}} \times \mathcal{A}^{\mathcal{V}}$ has a property A if there exists a triplet of properly defined distributions $p(\boldsymbol{x})$, $p(\boldsymbol{y} \mid \boldsymbol{x})$, $p(\boldsymbol{z} \mid \boldsymbol{x})$, which satisfy the following conditions:*

1. *The equations (4.4), (4.6) and (4.7) hold true simultaneously.*
2. *There exists a constant $\rho > 0$ such that $p(\boldsymbol{y} \mid \boldsymbol{x}) \geq \rho$, $\forall \boldsymbol{y} \in \mathcal{Y}^{\mathcal{V}}$ and $p(z_v = 1 \mid \boldsymbol{x}) \geq \rho$, $\forall v \in \mathcal{V}$.*

The condition 2 is required for two reasons. First, it implies that the space of probabilities with property A is a compact set which is needed to prove the consistency. Second, the nonzero marginal distributions $p(z_v = 1 \mid \boldsymbol{x}) \geq \rho$, $v \in \mathcal{V}$, guarantee that each local part has a chance to be annotated otherwise it is clear that learning from partial annotations would not be possible in general.

Example application We give an example of a prototypical application, in which the property A is guaranteed by steering the annotation process. In particular, let us consider a problem of learning structured output detector of facial landmarks (e.g. [Uřičář et al., 2012]). The facial landmarks are well discriminative features of human face like the corners of eyes or the corners of mouth. The parameters of the detector are learned from a set of training images with manually annotated landmark positions. The annotation of the training images is tedious and time consuming work. For example, in the work of [Uřičář et al., 2012] around 13,000 images had to be annotated to get desired accuracy. In the fully supervised case, the annotator is asked to mark positions of all landmarks in a given image. This corresponds to the annotation scheme $p(z_t \mid \boldsymbol{x}) = 1$, $\forall t \in \mathcal{V}$. However, we can instruct the annotator to mark only a subset of landmarks by using the following annotation scheme:

- In each even image, the annotator marks only the positions of landmarks on the left part of the face ($y_t \in \mathcal{Y}_t \mid t \in \mathcal{V}_{\text{left}}$).
- In each odd image, the annotator marks only the positions of landmarks on the right part of the face ($y_t \in \mathcal{Y}_t \mid t \in \mathcal{V}_{\text{right}}$).

Provided the annotator follows these instructions and the images are presented in a random order (which we can easily assure by randomly reshuffling the images before annotation) implies that

$$p(z_t \mid \boldsymbol{x}) = \frac{1}{2}\,, \quad t \in \mathcal{V}_{\text{left}} \cup \mathcal{V}_{\text{right}} \,.$$

This implies that with the same afford (i.e. when the annotator clicks the same amount of landmark positions) we can annotate twice as much different faces compared to the supervised framework. It is reasonable to expect that the variation in landmarks of different faces (e.g. depicting different identities) is much higher than variation between the paired landmarks of the same face. Hence the partial learning should deliver more robust landmark detector without increasing the cost of annotations.

4.4. Consistency of the partial loss

In this section, we present the principal result which justifies learning of the structured classifiers by minimization of the partial loss provided the data are generated from the statistical model defined in section 4.3.

Theorem 4. *Let $p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{a})$ be an arbitrary distribution defined over $\mathcal{X}\times\mathcal{Y}^{\mathcal{V}}\times\mathcal{A}^{\mathcal{V}}$ with property A and $p'(\boldsymbol{x},\boldsymbol{y})=\sum_{\boldsymbol{a}\in\mathcal{A}^{\mathcal{V}}} p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{a})$ and $p''(\boldsymbol{x},\boldsymbol{a})=\sum_{\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}}} p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{a})$ be the corresponding marginal distributions. Let ℓ be an additive loss (4.1) and let ℓ^p be an associated partial loss defined by (4.2). Then, for all sequences of random decision functions $\boldsymbol{h}_m\colon\mathcal{X}\to\mathcal{T}^{\mathcal{V}}$ (depending on training data generated from i.i.d variables with $p''(\boldsymbol{x},\boldsymbol{a})$), it holds*

$$R^{\ell^p}(\boldsymbol{h}_m;p'')\xrightarrow{P}R_*^{\ell^p}(p'')\Leftrightarrow R^{\ell}(\boldsymbol{h}_m;p')\xrightarrow{P}R_*^{\ell}(p')\,.$$

We start with a key lemma which shows that under proper assumptions a set of minimizers of the supervised risk is the same as the set of minimizers of the partial risk although the risk functions and their values are different.

Lemma 1. *Let $\ell\colon\mathcal{Y}^{\mathcal{V}}\times\mathcal{T}^{\mathcal{V}}\to\mathbb{R}_+$ be an additive loss function and let $\ell^p\colon\mathcal{A}^{\mathcal{V}}\times\mathcal{T}^{\mathcal{V}}\to\mathbb{R}_+$ be the associated partial loss. Let $p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{a})$ be a distribution with the property A. Then, $\boldsymbol{h}^*\colon\mathcal{X}\to\mathcal{T}^{\mathcal{V}}$ is a minimizer of $R^{\ell}(\boldsymbol{h};p')$ if and only if it is a minimizer of $R^{\ell^p}(\boldsymbol{h};p'')$, where $p'(\boldsymbol{x},\boldsymbol{y})=\sum_{\boldsymbol{a}\in\mathcal{A}^{\mathcal{V}}} p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{a})$ and $p''(\boldsymbol{x},\boldsymbol{a})=\sum_{\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}}} p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{a})$.*

PROOF: The risk $R^{\ell^p}(\boldsymbol{h};p'')$ can be rewritten as follows:

$$\begin{aligned}
R^{\ell^p}(\boldsymbol{h};p'') &= \mathbb{E}_{p(\boldsymbol{x})}\sum_{\boldsymbol{a}\in\mathcal{A}^{\mathcal{V}}} p(\boldsymbol{a}\mid\boldsymbol{x})\,\ell^p(\boldsymbol{a},\boldsymbol{h}(\boldsymbol{x}))\\
&= \mathbb{E}_{p(\boldsymbol{x})}\sum_{\boldsymbol{a}\in\mathcal{A}^{\mathcal{V}}} p(\boldsymbol{a}\mid\boldsymbol{x})\sum_{v\in\mathcal{V}}[\![|a_v|=1]\!]\,\ell_v(y_v,h_v(\boldsymbol{x}))\\
&= \mathbb{E}_{p(\boldsymbol{x})}\sum_{v\in\mathcal{V}}\sum_{a_v\in\mathcal{A}} p(a_v\mid\boldsymbol{x})\,[\![|a_v|=1]\!]\,\ell_v(y_v,h_v(\boldsymbol{x}))\\
&\overset{(*)}{=} \mathbb{E}_{p(\boldsymbol{x})}\sum_{v\in\mathcal{V}}\sum_{y_v\in\mathcal{Y}} p(z_v=1\mid\boldsymbol{x})\,p(y_v\mid\boldsymbol{x})\,\ell_v(y_v,h_v(\boldsymbol{x}))\\
&= \mathbb{E}_{p(\boldsymbol{x})}\sum_{v\in\mathcal{V}} p(z_v=1\mid\boldsymbol{x})\sum_{y_v\in\mathcal{Y}} p(y_v\mid\boldsymbol{x})\,\ell_v(y_v,h_v(\boldsymbol{x}))\,.
\end{aligned}$$

Here equality $(*)$ holds because of the equality following from (4.4):

$$p(a_v\mid\boldsymbol{x})=\sum_{y_v\in\mathcal{Y}}\sum_{z_v\in\mathcal{Z}} p(y_v\mid\boldsymbol{x})\,p(z_v\mid\boldsymbol{x})\,[\![a_v=\alpha(y_v,z_v)]\!]\,,$$

which for $a_v = \{y_v\}$ gives us the following equality

$$p(a_v \mid \boldsymbol{x}) = p(z_v = 1 \mid \boldsymbol{x})\, p(y_v \mid \boldsymbol{x})\,.$$

It is seen from the last equation that if $\boldsymbol{h}(\boldsymbol{x})^* = (h_v(\boldsymbol{x}) \mid v \in \mathcal{V})$ is a minimizer of $R^{\ell^p}(\boldsymbol{h}; p)$ then for any $\boldsymbol{x} \in \mathcal{X}$ and $v \in \mathcal{V}$ it holds that

$$h_v^*(\boldsymbol{x}) \in \operatorname*{Argmin}_{t \in \mathcal{T}} \sum_{y_v \in \mathcal{Y}} p(y_v \mid \boldsymbol{x})\, \ell_v(y_v, t), \tag{4.8}$$

because the marginals $p(z_v \mid \boldsymbol{x}) > 0$ thanks to Definition 8. Analogically, one can rewrite the risk $R^\ell(\boldsymbol{h}; p')$ as follows:

$$R^\ell(\boldsymbol{h}; p') = \mathbb{E}_{p(\boldsymbol{x})} \sum_{\boldsymbol{y} \in \mathcal{Y}^{\mathcal{V}}} p(\boldsymbol{y} \mid \boldsymbol{x})\, \ell(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x})) = \mathbb{E}_{p(\boldsymbol{x})} \sum_{v \in \mathcal{V}} \sum_{y_v \in \mathcal{Y}} p(y_v \mid \boldsymbol{x})\, \ell_v(y_v, \boldsymbol{h}_v(\boldsymbol{x}))\,,$$

showing that also any minimizer $\boldsymbol{h}(\boldsymbol{x})^* = (h_v(\boldsymbol{x}) \mid v \in \mathcal{V})$ of $R^\ell(\boldsymbol{h}; p')$ has to satisfy (4.8). ■

Note that Lemma 1 shows that the ℓ-risk and the ℓ^p-risk have the same set of minimizers. However, they do not have the same minimal value because the ℓ-risk upper bounds the ℓ^p-risk. The lemma is easy to prove and understand. However, it is not immediately applicable in practice because we cannot minimize the risks due unknown data generating distribution $p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{a})$. Instead, we resort to minimization of the empirical risk, by which we obtain approximate minimizers. The conditions under which the minimizers of the empirical risk converge are well studied [Vapnik, 1998]. It remains to show that convergence of the minimizers of the empirical partial risk to the expected partial risk implies the convergence of the same minimizers the expected true risk as stated in Theorem 4. The rigorous proof is not trivial. In the rest of the section, we give a road map of the proof and we leave the details to Appendix A.5.

It follows from Lemma 1 that for a fixed probability model p induced by a model with property A the function $H^p(\epsilon, \boldsymbol{p_{ya}})\colon \mathbb{R} \times \Delta_{|\mathcal{Y}^{\mathcal{V}}| \times |\mathcal{A}^{\mathcal{V}}|} \to \mathbb{R}^2$ whose values is defined by

$$\begin{aligned} \operatorname*{minimize}_{\boldsymbol{t} \in \mathcal{T}^{\mathcal{V}}} \quad & \boldsymbol{p_a}^\top \boldsymbol{\ell}^p_{\boldsymbol{t}} - \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_a}^\top \boldsymbol{\ell}^p_{\boldsymbol{t}'} \\ \text{subject to} \quad & \boldsymbol{p_y}^\top \boldsymbol{\ell_t} - \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_y}^\top \boldsymbol{\ell}_{\boldsymbol{t}'} \geq \epsilon \end{aligned}$$

is always positive for any $\epsilon > 0$, where $\boldsymbol{p_{ya}}(\boldsymbol{x}) = (p(\boldsymbol{y}, \boldsymbol{a} \mid \boldsymbol{x}) \mid \boldsymbol{a} \in \mathcal{A}^{\mathcal{V}}, \boldsymbol{y} \in \mathcal{Y}^{\mathcal{V}})$. Flipped function $H(\epsilon, \boldsymbol{p_{ya}})\colon \mathbb{R} \times \Delta_{|\mathcal{Y}^{\mathcal{V}}| \times |\mathcal{A}^{\mathcal{V}}|} \to \mathbb{R}$ whose values is defined by

$$\begin{aligned} \operatorname*{minimize}_{\boldsymbol{t} \in \mathcal{T}^{\mathcal{V}}} \quad & \boldsymbol{p_y}^\top \boldsymbol{\ell_t} - \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_y}^\top \boldsymbol{\ell}_{\boldsymbol{t}'} \\ \text{subject to} \quad & \boldsymbol{p_a}^\top \boldsymbol{\ell}^p_{\boldsymbol{t}} - \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_a}^\top \boldsymbol{\ell}^p_{\boldsymbol{t}'} \geq \epsilon \end{aligned}$$

is positive as well for any $\epsilon > 0$. It is possible to show (see Lemmas 4 and 8 in Appendix) even stronger statement that for any $\epsilon > 0$, $H^p(\epsilon) \triangleq \inf\limits_{\boldsymbol{p_{ay}} \in \mathcal{P}_{\boldsymbol{x}}} H^p(\epsilon, \boldsymbol{p_{ay}}) > 0$ and $H(\epsilon) \triangleq$

[2] We use $\Delta_n = \{\boldsymbol{p} \in \mathbb{R}^n \mid p_i \geq 0, \forall i \in [n], \sum_{i=1}^n p_i = 1\}$ to denote the probability simplex in $\mathbb{R}^n$. In case when $\boldsymbol{x}$ does not change, the argument $\boldsymbol{x}$ is omitted. We simply write $\boldsymbol{p_y}$, $\boldsymbol{p_a}$, $\boldsymbol{p_{ya}}$.

$\inf_{\boldsymbol{p_{ay}} \in \mathcal{P}_{\boldsymbol{x}}} H(\epsilon, \boldsymbol{p_{ay}}) > 0$. Thanks to this it is possible to show[3] that for loss functions $\ell(\boldsymbol{y}, \boldsymbol{t})$ and $\ell^p(\boldsymbol{a}, \boldsymbol{t})$ defined by (4.1), (4.2) there exist nonnegative concave functions $\xi\colon \mathbb{R} \to \mathbb{R}_+$and $\zeta\colon \mathbb{R} \to \mathbb{R}_+$, both right continuous at 0 with $\xi(0) = 0$ and $\zeta(0) = 0$, such that $\forall\ \boldsymbol{h}\colon \mathcal{X} \to \mathcal{T}^{\mathcal{V}}$ and for all distributions with property A it holds that

$$\mathbb{E}_{p(\boldsymbol{x})}\, \boldsymbol{p_y}(\boldsymbol{x})^\top \boldsymbol{\ell}_{\boldsymbol{h}(\boldsymbol{x})} - \mathbb{E}_{p(\boldsymbol{x})} \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_y}(\boldsymbol{x})^\top \boldsymbol{\ell}_{\boldsymbol{t}'} \leq$$
$$\xi\left(\mathbb{E}_{p(\boldsymbol{x})}\, \boldsymbol{p_a}(\boldsymbol{x})^\top \boldsymbol{\ell}^p_{\boldsymbol{h}(\boldsymbol{x})} - \mathbb{E}_{p(\boldsymbol{x})} \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_a}(\boldsymbol{x})^\top \boldsymbol{\ell}^p_{\boldsymbol{t}'}\right) ,$$
$$\mathbb{E}_{p(\boldsymbol{x})}\, \boldsymbol{p_a}(\boldsymbol{x})^\top \boldsymbol{\ell}^p_{\boldsymbol{h}(\boldsymbol{x})} - \mathbb{E}_{p(\boldsymbol{x})} \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_a}(\boldsymbol{x})^\top \boldsymbol{\ell}^p_{\boldsymbol{t}'} \leq$$
$$\zeta\left(\mathbb{E}_{p(\boldsymbol{x})}\, \boldsymbol{p_y}(\boldsymbol{x})^\top \boldsymbol{\ell}_{\boldsymbol{h}(\boldsymbol{x})} - \mathbb{E}_{p(\boldsymbol{x})} \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_y}(\boldsymbol{x})^\top \boldsymbol{\ell}_{\boldsymbol{t}'}\right) .$$

See Lemmas 5 and 9 in Appendix for complete proof.

Functions ξ and ζ make Theorem 4 easy to prove.

PROOF: ($\Rightarrow$) We have that for any $\epsilon > 0$ and $\boldsymbol{p_{ya}}(\boldsymbol{x}) \in \mathcal{P}_{\boldsymbol{x}}$ the inequality

$$\mathbb{P}\{\mathbb{E}_{p(\boldsymbol{x})}\, \boldsymbol{p_y}(\boldsymbol{x})^\top \boldsymbol{\ell}_{\boldsymbol{h}_m(\boldsymbol{x})} - \mathbb{E}_{p(\boldsymbol{x})} \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_y}(\boldsymbol{x})^\top \boldsymbol{\ell}_{\boldsymbol{t}'} > \epsilon\} \leq$$
$$\mathbb{P}\{\xi(\mathbb{E}_{p(\boldsymbol{x})}\, \boldsymbol{p_a}(\boldsymbol{x})^\top \boldsymbol{\ell}^p_{\boldsymbol{h}_m(\boldsymbol{x})} - \mathbb{E}_{p(\boldsymbol{x})} \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_a}(\boldsymbol{x})^\top \boldsymbol{\ell}^p_{\boldsymbol{t}'}) > \epsilon\}$$

holds. Since $\xi(x)$ is right continuous at 0, there exists $\delta > 0$ such that $\forall x\colon x - 0 \leq \delta \Rightarrow \xi(x) - \xi(0) \leq \epsilon$. Hence, if $\xi(x) > \epsilon$ then $x > \delta$, thus we obtain

$$\mathbb{P}\{\xi(\mathbb{E}_{p(\boldsymbol{x})}\, \boldsymbol{p_a}(\boldsymbol{x})^\top \boldsymbol{\ell}^p_{\boldsymbol{h}_m(\boldsymbol{x})} - \mathbb{E}_{p(\boldsymbol{x})} \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_a}(\boldsymbol{x})^\top \boldsymbol{\ell}^p_{\boldsymbol{t}'}) > \epsilon\} \leq$$
$$\mathbb{P}\{\mathbb{E}_{p(\boldsymbol{x})}\, \boldsymbol{p_a}(\boldsymbol{x})^\top \boldsymbol{\ell}^p_{\boldsymbol{h}_m(\boldsymbol{x})} - \mathbb{E}_{p(\boldsymbol{x})} \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_a}(\boldsymbol{x})^T \boldsymbol{\ell}^p_{\boldsymbol{t}'} > \delta\} \to 0 \,,$$

given $m \to \infty$.

($\Leftarrow$) implication is proved by repeating the same steps but using relation with function ζ.

■

4.5. Classification calibrated surrogates of the partial loss

In the previous section, we proved consistency of the minimization of the partial loss ℓ^p. Unfortunately, a direct minimization of the partial loss is hard due to its discrete domain. For this reason it is useful to employ a surrogate loss $\psi^p\colon \mathcal{A}^{\mathcal{V}} \times \hat{\mathcal{T}} \to \mathbb{R}_+$ and learn a function $\boldsymbol{f}\colon \mathcal{X} \to \mathbb{R}^d$ by minimizing the ψ^p-risk

$$R^{\psi^p}(\boldsymbol{f}, p'') = \mathbb{E}_{p''(\boldsymbol{x},\boldsymbol{a})} \psi^p(\boldsymbol{a}, \boldsymbol{f}(\boldsymbol{x})) \,.$$

Under suitable conditions, the ψ^p-risk of functions learned by the empirical risk minimization principle, i.e. $\boldsymbol{f}_m \in \text{Argmin}_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \psi^p(\boldsymbol{a}^i, \boldsymbol{f}(\boldsymbol{x}^i))$, will converge in probability to the Bayes ψ^p-risk

$$R^{\psi^p}_*(p'') = \inf_{\boldsymbol{f}\colon \mathcal{X} \to \hat{\mathcal{T}}} R^{\psi^p}(\boldsymbol{f}; p'') \,.$$

[3] This proof is technically complicated, therefore it is moved to Appendix. There we provide full set of lemmas with complete proofs.

It has been shown (e.g. [Zhang, 2004a; Tewari and Bartlett, 2007; Gao and Zhou, 2013; Ramaswamy and Agarwal, 2012]) that the question whether the statistically consistent estimator w.r.t ψ^p-risk implies the consistency w.r.t the ℓ^p-risk is equivalent to the question whether the surrogate loss is so called classification calibrated. Below we define a concept of a surrogate loss classification calibrated with respect to a the partial loss and the consistency theorem. These definitions are straightforward adaptations of Definition 1, Theorem 3 and Theorem 11 from [Ramaswamy and Agarwal, 2012] to our setting.

Definition 9. *A surrogate loss $\psi^p\colon \mathcal{A}^{\mathcal{V}} \times \hat{\mathcal{T}} \to \mathbb{R}_+$ is said to be classification calibrated with respect to the partial loss $\ell^p\colon \mathcal{A}^{\mathcal{V}} \times \mathcal{T}^{\mathcal{V}} \to \mathbb{R}_+$ over $\mathcal{P} \subseteq \Delta_{|\mathcal{Y}^{\mathcal{V}}|\times|\mathcal{A}^{\mathcal{V}}|}$ if there exists a function* $\mathrm{pred}\colon \hat{\mathcal{T}} \to \mathcal{T}^{\mathcal{V}}$ *such that $\forall \boldsymbol{p_{ya}} \in \mathcal{P}$:*

$$\inf_{\hat{\boldsymbol{t}}\in\hat{\mathcal{T}}\colon \mathrm{pred}(\hat{\boldsymbol{t}})\notin \mathrm{Argmin}_{\boldsymbol{t}\in\mathcal{T}^{\mathcal{V}}} \boldsymbol{p_a}^\top \boldsymbol{\ell}^p_{\boldsymbol{t}}} \boldsymbol{p_a}^\top \boldsymbol{\psi}^p(\hat{\boldsymbol{t}}) > \inf_{\hat{\boldsymbol{t}}\in\hat{\mathcal{T}}} \boldsymbol{p_a}^\top \boldsymbol{\psi}^p(\hat{\boldsymbol{t}}) \,.$$

Theorem 5. *Let $\ell^p\colon \mathcal{A}^{\mathcal{V}} \times \mathcal{T}^{\mathcal{V}} \to \mathbb{R}_+$ and $\psi^p\colon \mathcal{A}^{\mathcal{V}} \times \hat{\mathcal{T}} \to \mathbb{R}_+$. Then ψ^p is classification calibrated with respect to the partial loss ℓ^p over $\mathcal{P} \subseteq \Delta_{|\mathcal{Y}^{\mathcal{V}}|\times|\mathcal{A}^{\mathcal{V}}|}$ iff there exists a function* $\mathrm{pred}\colon \hat{\mathcal{T}} \to \mathcal{T}^{\mathcal{V}}$ *such that for all distributions $p(\boldsymbol{x},\boldsymbol{a})$ over $\mathcal{X}\times\mathcal{A}^{\mathcal{V}}$ and all sequences of random vector functions $\boldsymbol{f}_m\colon \mathcal{X} \to \hat{\mathcal{T}}$,*

$$R^{\psi^p}(\boldsymbol{f}_m; p) \xrightarrow{P} R^{\psi^p}_*(p) \qquad \textit{implies} \qquad R^{\ell^p}(\mathrm{pred}\circ \boldsymbol{f}_m; p) \xrightarrow{P} R^{\ell^p}_*(p) \,.$$

Combination of Theorem 5 and Theorem 4 directly provides the following corollary:

Corollary 6. *Let $\ell\colon \mathcal{Y}^{\mathcal{V}}\times\mathcal{T}^{\mathcal{V}} \to \mathbb{R}_+$ be additively decomposable loss function defined by (4.1) and $\psi^p\colon \mathcal{A}^{\mathcal{V}} \times \hat{\mathcal{T}} \to \mathbb{R}_+$. Then ψ^p is classification calibrated with repect to ℓ over $\mathcal{P} \subseteq \Delta_{|\mathcal{Y}^{\mathcal{V}}|\times|\mathcal{A}^{\mathcal{V}}|}$ iff there exists a function* $\mathrm{pred}\colon \hat{\mathcal{T}} \to \mathcal{T}^{\mathcal{V}}$ *such that for all distributions $p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{a})$ over $\mathcal{X}\times\mathcal{Y}^{\mathcal{V}}\times\mathcal{A}^{\mathcal{V}}$ with the property A and all sequences of random vector functions $\boldsymbol{f}_m\colon \mathcal{X} \to \hat{\mathcal{T}}$,*

$$R^{\psi^p}(\boldsymbol{f}_m; p'') \xrightarrow{P} R^{\psi^p}_*(p'') \qquad \textit{implies} \qquad R^{\ell}(\mathrm{pred}\circ \boldsymbol{f}_m; p') \xrightarrow{P} R^{\ell}_*(p') \,,$$

where $p'(\boldsymbol{x},\boldsymbol{y}) = \sum_{\boldsymbol{a}\in\mathcal{A}^{\mathcal{V}}} p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{a})$ and $p''(\boldsymbol{x},\boldsymbol{a}) = \sum_{\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}}} p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{a})$.

Corollary 6 guarantees that ℓ-risk of a decision function $\boldsymbol{h}(\boldsymbol{x}) = \mathrm{pred}\circ \boldsymbol{f}(\boldsymbol{x})$ learned by a statistically consistent algorithm minimizing the surrogate loss ψ^p, which is classification calibrated w.r.t. the partial loss ℓ^p associated to ℓ, converges in probability to the Bayes risk $R^{\ell}_*(p')$, i.e. learning algorithm minimizing ψ^p is Bayes consistent. In the next section, we give some examples of the classification calibrated surrogate partial losses.

4.6. Existence of the convex surrogate for partial loss

[Ramaswamy and Agarwal, 2012] introduced a notion of so called classification calibration dimension, which allows us to prove that for any partial loss function ℓ^p defined by (4.2) there always exists classification calibrated convex surrogate loss ψ^p. First, we give a definition of the classification calibration dimension adapted to our setting.

Definition 10. *Let $\ell^p\colon \mathcal{A}^{\mathcal{V}}\times\mathcal{T}^{\mathcal{V}} \to \mathbb{R}_+$ be the partial loss defined by (4.2). The classification calibration dimension of ℓ^p is defined as*

$$\begin{aligned}\mathrm{CCdim}(\ell^p) = \min\big\{d\in\mathbb{N} \mid\ & \textit{exists a convex set } \hat{\mathcal{T}}\subseteq\mathbb{R}^d \textit{ and a convex surrogate}\\ & \psi^P : \hat{\mathcal{T}} \to \mathbb{R}^n_+ \textit{ that is classification calibrated w.r.t. } \ell^P \textit{ over } \mathcal{P}\subseteq\Delta_{|\mathcal{Y}^{\mathcal{V}}|\times|\mathcal{A}^{\mathcal{V}}|}\big\}\end{aligned}$$

provided the above set is non-empty, and it is $\mathrm{CCdim}(\ell^p) = \infty$ *otherwise.*

In words it means that if finite, the $\mathrm{CCdim}(\ell^p)$ gives the minimal dimension of the auxiliary output space $\hat{\mathcal{T}}$ which allows to construct a convex surrogate for ℓ^p. [Ramaswamy and Agarwal, 2012] show that classification calibration dimension of any loss defined over discrete label space is always finite, see Theorem 11 in [Ramaswamy and Agarwal, 2012].

Corollary 7. *Let* $\ell^p \colon \mathcal{A}^\mathcal{V} \times \mathcal{T}^\mathcal{V} \to \mathbb{R}_+$ *be the partial additively decomposable loss defined by (4.2), then* $\mathrm{CCdim}(\ell^p) < \infty$*.*

Combination of Corollary 7 and our main result proved in Theorem 4 implies existence of a convex surrogate loss ψ^p which is classification calibrated with respect to an arbitrary additive target loss $\ell \colon \mathcal{Y}^\mathcal{V} \times \mathcal{T}^\mathcal{V} \to \mathbb{R}_+$ defined by (4.1).

Corollary 8. *There exists convex surrogate* $\psi^p \colon \mathcal{A}^\mathcal{V} \times \hat{\mathcal{T}} \to \mathbb{R}_+$ *which is classification calibrated w.r.t.* $\ell \colon \mathcal{Y}^\mathcal{V} \times \mathcal{T}^\mathcal{V} \to \mathbb{R}_+$ *defined by (4.1).*

4.7. Examples of surrogate losses

In this section, we study classification calibration of surrogate losses that have been used in existing algorithms learning structured output classifiers from partially annotated examples.

The majority of existing algorithms is based on minimization of so called ramp-loss and its mild modifications. Different modifications of the ramp-loss for structured output learning from partially annotated data were summarized in [Lou and Hamprecht, 2012] with the help of the following generic function

$$\psi^p(\boldsymbol{a}, \hat{\boldsymbol{t}}) = \left| \max_{\boldsymbol{t}\in\mathcal{U}^P} \left(\ell^p(\boldsymbol{a}, \boldsymbol{t}) + \hat{\boldsymbol{t}}_t\right) - \max_{\boldsymbol{t}\in\mathcal{U}^R} \hat{\boldsymbol{t}}_{\boldsymbol{t}} \right|_+ , \tag{4.9}$$

where $|\cdot|_+ = \max\{\cdot, 0\}$. The function $\boldsymbol{f} \colon \mathcal{X} \to \hat{\mathcal{T}}$ learned by minimizing $\psi^p(\boldsymbol{a}, \hat{\boldsymbol{t}})$ is converted to the decision function $\boldsymbol{h}(\boldsymbol{x}) = \mathrm{pred}(\boldsymbol{f}(\boldsymbol{x}))$ via

$$\mathrm{pred}(\hat{\boldsymbol{t}}) \in \operatorname*{Argmax}_{\boldsymbol{t}\in\mathcal{T}^\mathcal{V}} \hat{\boldsymbol{t}}_{\boldsymbol{t}} \,.$$

[Lou and Hamprecht, 2012] use $\mathcal{U}^P$ to denote so called “Penalty” space, elements of which contribute a positive value to the loss. Accordingly, $\mathcal{U}^R$ stands for “Reward” space whose elements make the negative contribution. Table 4.1 lists different instances of the generic function (4.9) and their appearance in the literature.

Apart from the generic form of the partial surrogate loss (4.9), one can extend loss of [Sarawagi and Gupta, 2008] for the learning with missing labels. Given partial loss ℓ^p composed of label losses ℓ_v, $v \in \mathcal{V}$, it can be approximated by the following additive surrogate loss

$$\psi^p(\boldsymbol{a}, \hat{\boldsymbol{t}}) = \sum_{v\in\mathcal{V}} [\![|a_v| = 1]\!] \, \psi_v(a_v, \hat{\boldsymbol{t}}_v) \,, \tag{4.10}$$

where $\hat{\mathcal{T}} \subseteq \mathbb{R}^{|\mathcal{T}^\mathcal{V}|}$, $\hat{\boldsymbol{t}} \in \hat{\mathcal{T}}$ is a concatenation of $|\mathcal{V}|$ vectors $\hat{\boldsymbol{t}}_v \in \hat{\mathcal{T}}_v \subseteq \mathbb{R}^{|\mathcal{T}|}$ and $\psi_v \colon \mathcal{Y} \times \hat{\mathcal{T}}_v \to \mathbb{R}_+$ are some surrogate single label losses. The function $\boldsymbol{f} \colon \mathcal{X} \to \hat{\mathcal{T}}$ is converted to the decision function $\boldsymbol{h}(\boldsymbol{x}) = \mathrm{pred}(\boldsymbol{f}(\boldsymbol{x}))$ via

$$\mathrm{pred}(\hat{\boldsymbol{t}}) = (\mathrm{pred}_v(\hat{\boldsymbol{t}}_v) \mid v \in \mathcal{V}) \quad \text{with} \quad \mathrm{pred}_v(\hat{\boldsymbol{t}}_v) \in \operatorname*{Argmax}_{t\in\mathcal{T}} \hat{\boldsymbol{t}}_{v,t} \,,$$

where $\hat{\boldsymbol{t}}_{v,t}$ denotes t-th component of the vector $\boldsymbol{t}_v$. [Yu et al., 2014] proposed an instance of the surrogate (4.10) for learning from examples with missing labels. They consider the setting with only two labels, $\mathcal{Y} = \{-1, +1\}$, and they use the hinge-loss $\psi^p(a,t) = \max\{0, 1 - a \cdot t\}$ as the single label surrogate.

Table 4.1 provides the summary of the existing surrogate losses we are aware of. For each of the listed surrogate, we either prove that it is classification calibrated or that it is not. The corresponding theorems are given in the remainder of this section.

Loss	CC	$\mathcal{U}^P$	$\mathcal{U}^R$	Appeared in the literature
Ramp	Yes	$\mathcal{T}^{\mathcal{V}}$	$\mathcal{T}^{\mathcal{V}}$	[Chuong et al., 2008; Girshick et al., 2011]
Hinge	No	$\mathcal{T}^{\mathcal{V}}$	$\boldsymbol{a}$	[Yu and Joachims, 2009; Fernandes and Brefeld, 2011a] [Zhu et al., 2010; Vedaldi and Zisserman, 2009] [Wang and Mori, 2010]
Max	Yes	$\mathcal{T}^{\mathcal{V}}/\boldsymbol{a}$	$\mathcal{T}^{\mathcal{V}}$	[Luo and Orabona, 2010]
Bridge	No	$\mathcal{T}^{\mathcal{V}}/\boldsymbol{a}$	$\boldsymbol{a}$	[Lou and Hamprecht, 2012]
S.A.L.	Yes	–	–	[Sarawagi and Gupta, 2008; Yu et al., 2014]

Table 4.1. List of surrogate loss functions, which have appeared in the literature. The naming given in the first column has been adopted from [Lou and Hamprecht, 2012] except for the S.A.L., which stands for the surrogate additive loss. The losses "Ramp", "Hinge", "Max", and "Bridge" are instances of (4.9) with particular form of the penalty space $\mathcal{U}^P$ and the reward space $\mathcal{U}^R$. The loss S.A.L. is defined in (4.10). The second column, CC, indicates whether the corresponding surrogate is classification calibrated.

Theorem 9. *Let ℓ^p be a partial loss (4.2). Then the ramp loss ψ^p constructed from ℓ^p by*

$$\psi^p(\boldsymbol{a}, \hat{\boldsymbol{t}}) = \max_{\boldsymbol{t} \in \mathcal{T}^{\mathcal{V}}} \left(\ell^p(\boldsymbol{a}, \boldsymbol{t}) + \hat{\boldsymbol{t}}_t \right) - \max_{\boldsymbol{t} \in \mathcal{T}^{\mathcal{V}}} \hat{\boldsymbol{t}}_{\boldsymbol{t}} \,, \tag{4.11}$$

is classification calibrated with respect to ℓ^p.

PROOF: Let us introduce a shortcut for a set of non-optimal decisions

$$\hat{\mathcal{T}}_{\text{non}} = \{ \hat{\boldsymbol{t}} \in \hat{\mathcal{T}} \mid \text{pred}(\hat{\boldsymbol{t}}) \notin \operatorname*{Argmin}_{\boldsymbol{t} \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p}_{\boldsymbol{a}}^T \boldsymbol{\ell}_{\boldsymbol{t}}^p \} \,.$$

Then we can write $\forall \boldsymbol{p} \in \mathcal{P}$:

$$\inf_{\hat{\boldsymbol{t}} \in \hat{\mathcal{T}}_{\text{non}}} \boldsymbol{p}_{\boldsymbol{a}}^\top \boldsymbol{\psi}^p(\hat{\boldsymbol{t}}) \geq \inf_{\hat{\boldsymbol{t}} \in \hat{\mathcal{T}}_{\text{non}}} \boldsymbol{p}_{\boldsymbol{a}}^\top \boldsymbol{\ell}^p_{\text{pred}(\hat{\boldsymbol{t}})} > \min_{\boldsymbol{t} \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p}_{\boldsymbol{a}}^\top \boldsymbol{\ell}_{\boldsymbol{t}}^p \,, \tag{4.12}$$

where the first inequality follows from the fact that the ramp loss $\psi^p(\boldsymbol{a}, \hat{\boldsymbol{t}})$ upper bounds the partial loss $\ell^p(\boldsymbol{a}, \text{pred}(\hat{\boldsymbol{t}}))$ for any $\boldsymbol{a} \in \mathcal{A}^{\mathcal{V}}$, $\hat{\boldsymbol{t}} \in \hat{\mathcal{T}}$ (e.g. [Chuong et al., 2008]). Let $\boldsymbol{t}^* \in \text{Argmin}_{\boldsymbol{t} \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p}_{\boldsymbol{a}}^\top \boldsymbol{\ell}_{\boldsymbol{t}}^p$ be an optimal decision and let us define $\hat{\boldsymbol{t}}' \in \hat{\mathcal{T}}$ such that $\hat{\boldsymbol{t}}'_{\boldsymbol{t}^*} = 0$ and $\hat{\boldsymbol{t}}'_{\boldsymbol{t}} < K$, $\forall \boldsymbol{t} \in \mathcal{T}^{\mathcal{V}} \setminus \{\boldsymbol{t}^*\}$, where $K = -\max_{\boldsymbol{a},\boldsymbol{t}} \ell^p(\boldsymbol{a}, \boldsymbol{t})$. Then, $\psi^p(\boldsymbol{a}, \hat{\boldsymbol{t}}') = \ell^p(\boldsymbol{a}, \boldsymbol{t}^*)$ for all $\boldsymbol{a} \in \mathcal{A}^{\mathcal{V}}$ and thus $\boldsymbol{p}_{\boldsymbol{a}}^\top \boldsymbol{\ell}_{\boldsymbol{t}^*}^p = \boldsymbol{p}_{\boldsymbol{a}}^\top \psi^p(\hat{\boldsymbol{t}}')$. Therefore we have $\min_{\boldsymbol{t} \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p}_{\boldsymbol{a}}^\top \boldsymbol{\ell}_{\boldsymbol{t}}^p \geq \inf_{\hat{\boldsymbol{t}} \in \hat{\mathcal{T}}} \boldsymbol{p}_{\boldsymbol{a}}^\top \boldsymbol{\psi}^p(\hat{\boldsymbol{t}})$ which after combining with (4.12) gives

$$\inf_{\hat{\boldsymbol{t}} \in \hat{\mathcal{T}}_{\text{non}}} \boldsymbol{p}_{\boldsymbol{a}}^\top \boldsymbol{\psi}^p(\hat{\boldsymbol{t}}) > \inf\nolimits_{\hat{\boldsymbol{t}} \in \hat{\mathcal{T}}} \boldsymbol{p}_{\boldsymbol{a}}^\top \boldsymbol{\psi}^p(\hat{\boldsymbol{t}}) \,.$$

■

Theorem 10. *Let ℓ^p be a partial loss (4.2). Then the max loss ψ^p constructed from ℓ^p by*

$$\psi^p(\boldsymbol{a}, \hat{\boldsymbol{t}}) = \left| \max_{\boldsymbol{t} \notin \boldsymbol{a}} \left(\ell^p(\boldsymbol{a}, \boldsymbol{t}) + \hat{\boldsymbol{t}}_t \right) - \max_{\boldsymbol{t} \in \mathcal{T}^{\mathcal{V}}} \hat{\boldsymbol{t}}_{\boldsymbol{t}} \right|_+ , \tag{4.13}$$

is classification calibrated with respect to ℓ^p if $\ell^p(\boldsymbol{a}, \boldsymbol{y}) = 0\,, \forall \boldsymbol{y} \in \boldsymbol{a}$.

PROOF: We will show how to adapt the proof of Theorem 9 to prove this theorem. In context of proof of Theorem 9, let $\boldsymbol{t}^* \in \operatorname{Argmin}_{\boldsymbol{t} \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p}_{\boldsymbol{a}}^\top \boldsymbol{\ell}_{\boldsymbol{t}}^p$ be optimal decision and let us define $\hat{\boldsymbol{t}}' \in \hat{\mathcal{T}}$ in the same way, i.e. such that $\hat{\boldsymbol{t}}'_{\boldsymbol{t}^*} = 0$ and $\hat{\boldsymbol{t}}'_{\boldsymbol{t}} < K$, $\forall \boldsymbol{t} \in \mathcal{T}^{\mathcal{V}} \setminus \{\boldsymbol{t}^*\}$, where $K = -\max_{\boldsymbol{a},\boldsymbol{t}} \ell^p(\boldsymbol{a}, \boldsymbol{t})$. What we have to show now is that $\psi^p(\boldsymbol{a}, \hat{\boldsymbol{t}}') = \ell^p(\boldsymbol{a}, \boldsymbol{t}^*)$ holds for all $\boldsymbol{a} \in \mathcal{A}^{\mathcal{V}}$. For all $\boldsymbol{a} \in \{\boldsymbol{a} \in \mathcal{A}^{\mathcal{V}} \mid \boldsymbol{t}^* \notin \boldsymbol{a}\}$ we clearly have $\psi^p(\boldsymbol{a}, \hat{\boldsymbol{t}}') = \ell^p(\boldsymbol{a}, \boldsymbol{t}^*)$. On the other hand, for all $\boldsymbol{a} \in \{\boldsymbol{a} \in \mathcal{A}^{\mathcal{V}} \mid \boldsymbol{t}^* \in \boldsymbol{a}\}$ we have $\psi^p(\boldsymbol{a}, \hat{\boldsymbol{t}}') = 0$ and $\ell^p(\boldsymbol{a}, \boldsymbol{t}^*) = 0$. Thus, proof of Theorem 9 stays valid in this case as well.

■

Theorem 11. *Let ℓ^p be a partial loss (4.2). Then the bridge loss constructed from ℓ^p by*

$$\psi^p_{\text{bridge}}(\boldsymbol{a}, \hat{\boldsymbol{t}}) = \left| \max_{\boldsymbol{t} \notin \boldsymbol{a}} \left(\ell^p(\boldsymbol{a}, \boldsymbol{t}) + \hat{\boldsymbol{t}}_t \right) - \max_{\boldsymbol{t} \in \boldsymbol{a}} \hat{\boldsymbol{t}}_{\boldsymbol{t}} \right|_+ , \tag{4.14}$$

and the hinge loss

$$\psi^p_{\text{hinge}}(\boldsymbol{a}, \hat{\boldsymbol{t}}) = \max_{\boldsymbol{t} \in \mathcal{T}^{\mathcal{V}}} \left(\ell^p(\boldsymbol{a}, \boldsymbol{t}) + \hat{\boldsymbol{t}}_t \right) - \max_{\boldsymbol{t} \in \boldsymbol{a}} \hat{\boldsymbol{t}}_{\boldsymbol{t}} \,, \tag{4.15}$$

are not classification calibrated with respect to ℓ^p.

PROOF: W.l.o.g. let us consider distribution $\boldsymbol{p}_{\boldsymbol{a}}$ so that $p(z_v = 1 \mid \boldsymbol{x}) = 1 \; \forall v \in \mathcal{V}$, i.e. our distribution generates only supervised data. Note that this assumption does not break the property A. In the supervised scenario the bridge loss (4.14) reads

$$\psi^p_{\text{bridge}}(\boldsymbol{y}, \hat{\boldsymbol{t}}) = \left| \max_{\boldsymbol{t} \notin \boldsymbol{a}} \left(\ell^p(\boldsymbol{a}, \boldsymbol{t}) + \hat{\boldsymbol{t}}_t \right) - \max_{\boldsymbol{t} \in \boldsymbol{a}} \hat{\boldsymbol{t}}_{\boldsymbol{t}} \right|_+ = \left| \max_{\boldsymbol{t} \neq \boldsymbol{y}} \left(\ell^p(\boldsymbol{y}, \boldsymbol{t}) + \hat{\boldsymbol{t}}_t \right) - \hat{\boldsymbol{t}}_{\boldsymbol{y}} \right|_+ , \tag{4.16}$$

and the hinge loss (4.15) reads

$$\psi^p_{\text{hinge}}(\boldsymbol{y}, \hat{\boldsymbol{t}}) = \max_{\boldsymbol{t} \in \mathcal{T}^{\mathcal{V}}} \left(\ell^p(\boldsymbol{y}, \boldsymbol{t}) + \hat{\boldsymbol{t}}_t \right) - \max_{\boldsymbol{t} \in \boldsymbol{y}} \hat{\boldsymbol{t}}_{\boldsymbol{t}} = \max_{\boldsymbol{t} \in \mathcal{T}^{\mathcal{V}}} \left(\ell^p(\boldsymbol{y}, \boldsymbol{t}) + \hat{\boldsymbol{t}}_t \right) - \hat{\boldsymbol{t}}_{\boldsymbol{y}} \,. \tag{4.17}$$

In both cases, it is nothing but the standard hinge-loss for the supervised multiclass classification, so called construction of [Crammer and Singer, 2002]. Using results from [Tewari and Bartlett, 2007] for a general multiclass setting or [Liu, 2007] for 0/1 multiclass loss it can be shown that the losses (4.17) and (4.16) are not classification-calibrated. ■

Theorem 12. *Let $\psi_v \colon \mathcal{Y} \times \hat{\mathcal{T}}_v \to \mathbb{R}_+$, $v \in \mathcal{V}$, be a set of single label losses classification calibrated w.r.t. to some $\ell_v \colon \mathcal{Y} \times \mathcal{T} \to \mathbb{R}_+$. Then, the loss ψ^p composed of ψ_v, $v \in \mathcal{V}$, according to (4.10) is classification calibrated w.r.t. the partial loss ℓ^p composed of ℓ_v, $v \in \mathcal{V}$.*

PROOF: Let us introduce a shortcut for a set of non-optimal decisions for $v \in \mathcal{V}$:

$$\hat{\mathcal{T}}_{\text{non}} = \{\hat{\boldsymbol{t}} \in \hat{\mathcal{T}} \mid \operatorname{pred}(\hat{\boldsymbol{t}}) \notin \operatorname*{Argmin}_{\boldsymbol{t} \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p}_{\boldsymbol{a}}^\top \boldsymbol{\ell}_{\boldsymbol{t}}^p\}$$

and

$$\hat{\mathcal{T}}^v_{\mathrm{non}} = \{\hat{\boldsymbol{t}} \in \hat{\mathcal{T}}^v \mid \mathrm{pred}_v(\hat{\boldsymbol{t}}) \notin \operatorname*{Argmin}_{t\in\mathcal{T}} \boldsymbol{p}_{\boldsymbol{a},v}^\top \boldsymbol{\ell}^p_{v,t}\} \,.$$

Then we can write

$$\inf_{\hat{\boldsymbol{t}}\in\hat{\mathcal{T}}_{\mathrm{non}}} \boldsymbol{p}_{\boldsymbol{a}}^\top \boldsymbol{\psi}^p(\hat{\boldsymbol{t}}) = \inf_{\hat{\boldsymbol{t}}\in\hat{\mathcal{T}}_{\mathrm{non}}} \sum_{v\in\mathcal{V}} \boldsymbol{p}_v^\top \boldsymbol{\psi}^p_{\boldsymbol{a},v}(\hat{\boldsymbol{t}}_v) = \sum_{v\in\mathcal{V}} \inf_{\hat{\boldsymbol{t}}_v\in\hat{\mathcal{T}}^v_{\mathrm{non}}} \boldsymbol{p}_{\boldsymbol{a},v}^\top \boldsymbol{\psi}^p_v(\hat{\boldsymbol{t}}_v) >$$

$$\sum_{v\in\mathcal{V}} \inf_{\hat{\boldsymbol{t}}_v\in\hat{\mathcal{T}}^v} \boldsymbol{p}_{\boldsymbol{a},v}^\top \boldsymbol{\psi}^p_v(\hat{\boldsymbol{t}}_v) = \inf_{\hat{\boldsymbol{t}}\in\hat{\mathcal{T}}} \sum_{v\in\mathcal{V}} \boldsymbol{p}_{\boldsymbol{a},v}^\top \boldsymbol{\psi}^p_v(\hat{\boldsymbol{t}}_v) = \inf_{\hat{\boldsymbol{t}}\in\hat{\mathcal{T}}} \boldsymbol{p}_{\boldsymbol{a}}^\top \boldsymbol{\psi}^p(\hat{\boldsymbol{t}}) \,,$$

where strict inequality follows from the fact that for every $v \in \mathcal{V}$ the inequality $\inf_{\hat{\boldsymbol{t}}_v\in\hat{\mathcal{T}}^v_{\mathrm{non}}} \boldsymbol{p}_{\boldsymbol{a},v}^\top \boldsymbol{\psi}^p_v(\hat{\boldsymbol{t}}_v) > \inf_{\hat{\boldsymbol{t}}_v\in\hat{\mathcal{T}}^v} \boldsymbol{p}_{\boldsymbol{a},v}^\top \boldsymbol{\psi}^p_v(\hat{\boldsymbol{t}}_v)$ holds.

■

Theorem 12 shows that the additive surrogate loss (4.10) preserves the property of classification calibration. This allows to convert any set of single label classification calibrated losses to the loss calibrated w.r.t. the partial loss ℓ^p. This proves that the surrogate proposed in [Yu et al., 2014] for binary labels $\mathcal{Y} = \{-1, +1\}$ is calibrated. To our best knowledge, the additive surrogate (4.10) constructed for the case $|\mathcal{Y}| > 2$ has not been proposed so far.

4.8. Relation to existing works

We have shown that under quite general assumptions on the data generating process the minimization of the partial loss yields structured output classifier whose expected risk converges in probability to the Bayes risk defined by the associated complete loss. Our result connects learning from partially annotated examples, the scenario when labels of some local parts are missing, with the supervised learning in which case all local parts are annotated. We have used the established connection to extend the results of [Ramaswamy and Agarwal, 2012], who study consistency of the supervised methods learning multiclass classifiers under generic loss function. In particular, we adopted the notation of classification calibrated surrogate loss functions to the realm of learning with missing labels. In turn, we could analyze the so far heuristic algorithms learning structured output classifiers from partially annotated examples and to show, which of them are statistically consistent and which are not. Second, the same connection allowed us to show the existence of a convex classification calibrated surrogate loss for learning from partially annotated examples.

Another work closely related to ours has been published in [Cid-Sueiro et al., 2014]. [Cid-Sueiro et al., 2014] proposed a generic framework of deriving classification calibrated surrogates for the multiclass learning of flat classifiers from examples with missing labels. The authors introduce statistical model, in which they show that consistent surrogate losses for partial learning can be obtained by a linear transformation of any conventional surrogate that is consistent in the supervised setting. The authors also show that the linear transform in some cases preserves the convexity of the original supervised surrogate. Their framework, however, is not tractable in the structured output setting because the constructed surrogate would be defined as a sum of exponential number of terms.

In this work, we do not analyze trivial extension of convex surrogate classification calibrated loss to the structured output setting, like the extension of the "one versus all loss" [Cour et al., 2011] that was mentioned (but not implemented) in [Lou and Hamprecht, 2012]. Such surrogates are not tractable in the structured output setting since their evaluation involves summation over exponentially large sets. A construction of a tractable convex surrogate losses for structured output setting thus remains an open problem. A promising direction might be to investigate the notation of so called composite proper losses introduced in [Vernet et al., 2011; Reid and Williamson, 2010]. The composite proper losses are especially interesting from practical point of view since they provide a way to design convex surrogate losses [Reid et al., 2012] that keep statistical properties. However, all these works analyze only the flat classification. It is unclear whether the extension to the structured output setting and learning from missing labels is possible.

The last work to mention is the paper of [McAllester and Keshet, 2011] who study the consistency of the ramp-loss. However, there are two major differences compared to our results. First, they analyze consistency under the PCA-Bayesian setting, which threats the parameters to be learned as random variables, while we stay in the classical frequentist statistics. Second, they consider only the standard supervised setting, when the labels to be predicted are not missing in the training set. Although they consider also latent variables these are introduced just to make the model more flexible but they do not appear in the loss function and hence the problem remains supervised in principle.

4.9. Conclusions

In this chapter, we have analyzed a partial loss which can be constructed for any (complete/supervised) additive loss function by neglecting the local parts which are not annotated in the training examples. The partial loss provides a way how to use the ERM principle for learning structured output classifiers from examples with missing labels. We have shown that under quite general assumptions on the data generating process the minimization of the partial loss yields structured output classifier whose expected risk converges in probability to the Bayes risk defined by the associated complete loss. Further, we have proposed a concept of so called classification calibrated surrogate of the partial loss. We have shown that the algorithms minimizing a classification calibrated surrogate of the partial loss are statistically consistent. We have analyzed many existing algorithms which in their core optimize a surrogate of the partial loss function. We have shown which of the used surrogates are classification calibrated and which are not. It should be mentioned that none of the existing surrogates is a convex function of the parameters of the learned decision function. We have proved that there exists a classification calibrated convex surrogate of the partial loss. Unfortunately, the existence theorem is not constructive and hence the construction of a feasible convex classification calibrated surrogate of the partial loss remains and open problem.

5. Conclusions and open questions

In this work we tried to push forward the ERM based methods for learning from partially annotated examples. We designed a convex algorithm for learning ordinal classifiers from interval annotations and demonstrated empirically its advantage on real-life data. At the same time we made several contributions to the supervised learning of the ordinal classifiers, namely, we proposed new parametrization of the ordinal classifier, we introduced more flexible piece wise version of the ordinal classifier, and we proposed a generic cutting plane solver.

In the case of learning the structured output classifiers from examples with missing labels, we have defined the concept of a surrogate classification calibrated partial loss, the minimization of which guarantees the statistical consistency under fairly general conditions on the data generating process. We showed the existence of the statistically consistent convex surrogate loss for learning from partially annotated examples. We showed which existing surrogate losses are classification calibrated and which are not. Our analysis thus provides a missing theoretical justification for so far heuristic methods which have been used in practice.

A list of open questions which can be investigated in the future is as follows:

Restrictions of consistency analysis in the structured output setting Our analysis of the statistical consistency in the structured output learning is valid only under some assumptions on the data generating process, the loss function and the hypothesis space. Namely, the loss function must be additive over the local parts which is frequent in practice, yet non-decomposable loss functions are important as well. As for the annotation process, the main restriction is that for each local part the label has to be either known exactly or completely missing. It is natural to consider an intermediate case when each local part is annotated by a candidate set of labels. Finally, our analysis requires that the Bayes classifier is contained in the hypothesis space from which we learn. The assumption, although common in theory, is often violated in practice where we often learn the linear classifiers unlikely to be Bayes optimal. Therefore providing guarantees in the case when the hypothesis space can be arbitrary is very wanted. Any extension in these directions would be interesting both from practical and theoretical point of view.

Statistically consistent convex surrogate loss for structured output learning. One of the main contribution of this work is showing the existence of a convex statistically consistent surrogate loss for learning from examples with missing labels. We were a bit disappointed to obtain this result, since we spent quite a lot of time on trying to show that such surrogate does not exist. Unfortunately, the existence proof is not constructive so a computationally feasible convex surrogate remains an open problem. An option to look in the quest for a computationally feasible convex surrogate loss is a recently proposed framework of so called composite losses [Vernet et al., 2011; Reid et al., 2012]. The composite losses allow designing convex statistical consistent surrogates for supervised learning of flat classifiers. Unfortunately, the existing results are not immediately applicable to our problem, since this framework is neither developed for the learning from partially annotated examples nor for the structured output classification. Yet the composite losses are definitely a promising direction to pursue in the future.

Statistical consistency of the V-shaped interval insensitive loss. The proposed algorithm VILMA for learning of ordinal classifiers from interval annotations is convex and works well in practice as we demonstrated. Unfortunately, the surrogate of the V-shaped interval insensitive loss minimized by the VILMA is unlikely to be consistent in general. For example, the SVOR-IMC algorithm, which is a supervised counterpart of VILMA with MAE loss, has been recently proved not to be statistically consistent [Pedregosa et al., 2014]. Still it may be interesting to investigate special cases of the data generating distribution under which the consistency can be shown.

A. Proofs

A.1. Proof of Theorem 1

Let us prove the first part of the theorem stating that for any $\boldsymbol{w} \in \mathbb{R}^n$ and admissible $\boldsymbol{\theta} \in \Theta$ there exists $\boldsymbol{b} \in \mathbb{R}^Y$ such that $h(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\theta}) = h'(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{b})$, $\forall \boldsymbol{x} \in \mathbb{R}^n$. In particular, we show that $\boldsymbol{b} \in \mathbb{R}^Y$ given by the formula (3.5) satisfies theorem.

First, suppose the ORD classifier $h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta})$ outputs $y \in \mathcal{Y}$ for some $\boldsymbol{x} \in \mathcal{X}$, i.e. $\theta_y \geq \langle \boldsymbol{w}, \boldsymbol{x} \rangle > \theta_{y-1}$ holds[1]. The MORD classifier $h'(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{b})$ outputs the same y iff the system of inequalities

$$\begin{aligned} \langle \boldsymbol{w}, \boldsymbol{x} \rangle y + b_y &> \langle \boldsymbol{w}, \boldsymbol{x} \rangle (y-k) + b_{y-k},\ 1 \leq k < y, \\ \langle \boldsymbol{w}, \boldsymbol{x} \rangle y + b_y &\geq \langle \boldsymbol{w}, \boldsymbol{x} \rangle (y+t) + b_{y+t},\ 1 \leq t \leq Y - y \end{aligned} \tag{A.1}$$

holds. The system (A.1) can be rewritten as[2]

$$\begin{aligned} \langle \boldsymbol{w}, \boldsymbol{x} \rangle k &> \sum_{i=y-k}^{y-1} \theta_i,\ 1 \leq k < y, \\ \langle \boldsymbol{w}, \boldsymbol{x} \rangle t &\leq \sum_{i=y}^{y+t-1} \theta_i,\ 1 \leq t \leq Y - y. \end{aligned} \tag{A.2}$$

The validity of (A.2) follows from

$$\begin{aligned} \langle \boldsymbol{w}, \boldsymbol{x} \rangle k > \theta_{y-1} k &\geq \sum_{i=y-k}^{y-1} \theta_i,\ 1 \leq k < y, \\ \langle \boldsymbol{w}, \boldsymbol{x} \rangle t \leq \theta_y t &\leq \sum_{i=y}^{y+t-1} \theta_i,\ 1 \leq t \leq Y - y\,, \end{aligned} \tag{A.3}$$

where the first inequality (on both lines) is induced by $\theta_y \geq \langle \boldsymbol{w}, \boldsymbol{x} \rangle > \theta_{y-1}$ and the second inequality (also on both lines) is due to $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_{Y-1}$.

Second, suppose the MORD classifier $h'(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{b})$ outputs $y \in \mathcal{Y}$ for some $\boldsymbol{x} \in \mathcal{X}$, which means that

$$\begin{aligned} \langle \boldsymbol{w}, \boldsymbol{x} \rangle y + b_y &> \langle \boldsymbol{w}, \boldsymbol{x} \rangle (y-1) + b_{y-1}, \\ \langle \boldsymbol{w}, \boldsymbol{x} \rangle y + b_y &\geq \langle \boldsymbol{w}, \boldsymbol{x} \rangle (y+1) + b_{y+1}, \end{aligned} \tag{A.4}$$

which is equivalent to

$$b_y - b_{y+1} \geq \langle \boldsymbol{w}, \boldsymbol{x} \rangle > b_{y-1} - b_y\,. \tag{A.5}$$

Finally, after combining (A.5) with (3.5) we obtain $\theta_y \geq \langle \boldsymbol{w}, \boldsymbol{x} \rangle > \theta_{y-1}$, which implies that the ordinal classifier $h(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\theta})$ outputs the same y.

Let us make the observation before proving the second part of the theorem. Let $y_1, \ldots, y_p$, denote an increasing subsequence of the non-degenerated classes of the MORD classifier

[1] The inequalities are different in the case of $y \in \{1, Y\}$. However, the analysis remains similar thus it is omited here.

[2] We use convention that a sum is zero if its upper index is less than the lower one.

$h'(\boldsymbol{x},\boldsymbol{w},\boldsymbol{b})$. For arbitrary $\boldsymbol{x}_{y_i} \in \mathcal{X}_{y_i} = \{\boldsymbol{x} \in \mathbb{R}^n \mid h'(\boldsymbol{x},\boldsymbol{w},\boldsymbol{b}) = y_i\}$, $i = 1,\dots,p$, it holds that

$$\begin{aligned}\langle \boldsymbol{w}, \boldsymbol{x}_{y_i}\rangle y_i + b_{y_i} &> \langle \boldsymbol{w}, \boldsymbol{x}_{y_{i-1}}\rangle y_{i-1} + b_{y_{i-1}},\\ \langle \boldsymbol{w}, \boldsymbol{x}_{y_i}\rangle y_i + b_{y_i} &\geq \langle \boldsymbol{w}, \boldsymbol{x}_{y_{i+1}}\rangle y_{i-1} + b_{y_{i+1}},\end{aligned} \tag{A.6}$$

It follows that

$$\tfrac{b_{y_i}-b_{y_{i+1}}}{y_{i+1}-y_i} \geq \langle \boldsymbol{w}, \boldsymbol{x}_{y_i}\rangle > \tfrac{b_{y_{i-1}}-b_{y_i}}{y_i-y_{i-1}}, \quad i = 1,\dots,p-1.$$

Thus, for any MORD classifier $h'(\boldsymbol{x},\boldsymbol{w},\boldsymbol{b})$ with non-degenerated classes $y_1,\dots,y_p$, it holds that

$$\tfrac{b_{y_{p-1}}-b_{y_p}}{y_p-y_{p-1}} > \dots > \tfrac{b_{y_{i-1}}-b_{y_i}}{y_i-y_{i-1}} > \dots > \tfrac{b_{y_1}-b_{y_2}}{y_2-y_1}. \tag{A.7}$$

We are now ready to prove the second part of the theorem stating that for any $\boldsymbol{w} \in \mathbb{R}^n$, $\boldsymbol{b} \in \mathbb{R}^Y$ and the admissible vector $\boldsymbol{\theta} \in \Theta$ computed by the formula (3.7) the equality $h(\boldsymbol{x},\boldsymbol{w},\boldsymbol{\theta}) = h'(\boldsymbol{x},\boldsymbol{w},\boldsymbol{b})$ holds $\forall \boldsymbol{x} \in \mathbb{R}^n$. It is enough to show that the ordinal classifier $h(\boldsymbol{x},\boldsymbol{w},\boldsymbol{\theta})$ outups y_i for arbitrary $\boldsymbol{x} \in \mathcal{X}$ iff the MORD classifier $h'(\boldsymbol{x},\boldsymbol{w},\boldsymbol{b})$ outputs the same output y_i.

First, suppose the MORD classifier $h'(\boldsymbol{x};\boldsymbol{w},\boldsymbol{b})$ outputs $y_i \in \mathcal{Y}$ for some $\boldsymbol{x} \in \mathcal{X}$. We want to show that the ordinal classifier $h(\boldsymbol{x};\boldsymbol{w},\boldsymbol{\theta})$ outputs the same label y_i. We shall analyse only the cases $1 < i < p$. However, the proof for $i \in \{1,p\}$ is similar and hence omitted. The equality $h'(\boldsymbol{x},\boldsymbol{w},\boldsymbol{b}) = y_i$ implies that

$$\begin{aligned}\langle \boldsymbol{w}, \boldsymbol{x}\rangle y_i + b_{y_i} &> \langle \boldsymbol{w}, \boldsymbol{x}\rangle y_{i-1} + b_{y_{i-1}},\\ \langle \boldsymbol{w}, \boldsymbol{x}\rangle y_i + b_{y_i} &\geq \langle \boldsymbol{w}, \boldsymbol{x}\rangle y_{i+1} + b_{y_{i+1}},\end{aligned} \tag{A.8}$$

which is equivalent to $\frac{b_{y_i}-b_{y_{i+1}}}{y_{i+1}-y_i} \geq \langle \boldsymbol{w}, \boldsymbol{x}\rangle > \frac{b_{y_{i-1}}-b_{y_i}}{y_i-y_{i-1}}$ and after combining with (3.7) we see that the ordinal classifier $h(\boldsymbol{x},\boldsymbol{w},\boldsymbol{\theta})$ outputs the same y_i.

Second, suppose the ordinal classifier $h(\boldsymbol{x},\boldsymbol{w},\boldsymbol{\theta})$ outputs y_i for some arbitrary $\boldsymbol{x} \in \mathcal{X}$, i.e. $\frac{b_{y_i}-b_{y_{i+1}}}{y_{i+1}-y_i} \geq \langle \boldsymbol{w}, \boldsymbol{x}\rangle > \frac{b_{y_{i-1}}-b_{y_i}}{y_i-y_{i-1}}$ holds. To show that MORD classifier $h'(\boldsymbol{x};\boldsymbol{w},\boldsymbol{\theta})$ outputs the same y_i it is enough to prove that the system

$$\langle \boldsymbol{w}, \boldsymbol{x}\rangle y_i + b_{y_i} > \langle \boldsymbol{w}, \boldsymbol{x}\rangle y_j + b_{y_j},\ \forall y_j < y_i, \tag{A.9}$$

$$\langle \boldsymbol{w}, \boldsymbol{x}\rangle y_i + b_{y_i} \geq \langle \boldsymbol{w}, \boldsymbol{x}\rangle y_j + b_{y_j},\ \forall y_j > y_i \tag{A.10}$$

holds. Indeed, from the inequality $\langle \boldsymbol{w}, \boldsymbol{x}\rangle > \frac{b_{y_{i-1}}-b_{y_i}}{y_i-y_{i-1}}$ after some algebra and applying (A.7) (after third line) we have

$$\begin{aligned}\langle \boldsymbol{w}, \boldsymbol{x}\rangle(y_i - y_j) &> (y_i - y_j)\tfrac{b_{y_{i-1}}-b_{y_i}}{y_i-y_{i-1}}\\ &= (-y_j + y_{j+1} - y_{j+1} + \dots + y_{i-1} - y_{i-1} + y_i)\tfrac{b_{y_{i-1}}-b_{y_i}}{y_i-y_{i-1}}\\ &= (y_{j+1} - y_j)\tfrac{b_{y_{i-1}}-b_{y_i}}{y_i-y_{i-1}} + \dots + (y_i - y_{i-1})\tfrac{b_{y_{i-1}}-b_{y_i}}{y_i-y_{i-1}}\\ &> (y_{j+1} - y_j)\tfrac{b_{y_j}-b_{y_{j+1}}}{y_{j+1}-y_j} + \dots + (y_i - y_{i-1})\tfrac{b_{y_{i-1}}-b_{y_i}}{y_i-y_{i-1}}\\ &= b_{y_j} - b_{y_{j+1}} + b_{y_{j+1}} - \dots - b_{y_{i-1}} + b_{y_{i-1}} - b_{y_i} = b_{y_j} - b_{y_i}\ ,\end{aligned}$$

from which the inequalities (A.9) follow for $\forall y_j < y_i$. The proof of the inequalities (A.10) is analogical. ■

A.2. Proof of Theorem 3

We will prove the bound (3.33) for each observation $\boldsymbol{x} \in \mathcal{X}$ separately, that is, we prove

$$R^{MAE}(h \mid \boldsymbol{x}) \leq R_I^{MAE}(h \mid \boldsymbol{x}) + (1-\alpha)\beta\,, \tag{A.11}$$

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 (A.11) satisfied for all $\boldsymbol{x} \in \mathcal{X}$ implies (3.33). Let us define a function, which measures a discrepancy between the MAE and the its interval insensitive counterpart:

$$\delta(h(\boldsymbol{x}), y, y_l, y_r) = |y - h(\boldsymbol{x})| - \min_{y' \in [y_l,y_r]} |y' - h(\boldsymbol{x})| = \left\{ \begin{array}{rl} |h(\boldsymbol{x}) - y| & \text{if} \quad h(\boldsymbol{x}) \in [y_l, y_r]\,, \\ y - y_l & \text{if} \quad h(\boldsymbol{x}) < y_l\,, \\ y_r - y & \text{if} \quad h(\boldsymbol{x}) > y_r\,. \end{array} \right. \tag{A.12}$$

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, 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:

$$\begin{aligned} R_I^{MAE}(h \mid \boldsymbol{x}) &= \sum_{y \in \mathcal{Y}} \sum_{[y_l,y_r] \in \mathcal{P}} p(y \mid \boldsymbol{x}) p(y_l, y_r \mid \boldsymbol{x}, y) \min_{y' \in [y_l,y_r]} |y' - h(\boldsymbol{x})| \\ &= \sum_{y \in \mathcal{Y}} p(y \mid \boldsymbol{x}) \left[\alpha |y - h(\boldsymbol{x})| + \sum_{[y_l,y_r] \notin \mathcal{P}_1} p(y_l, y_r \mid \boldsymbol{x}, y) \min_{y' \in [y_l,y_r]} |y' - h(\boldsymbol{x})| \right] \\ &= \sum_{y \in \mathcal{Y}} p(y \mid \boldsymbol{x}) \left[\alpha |y - h(\boldsymbol{x})| + \sum_{[y_l,y_r] \notin \mathcal{P}_1} p(y_l, y_r \mid \boldsymbol{x}, y) \big(|y - h(\boldsymbol{x})| - \delta(h(\boldsymbol{x}), y, y_l, y_r) \big) \right] \\ &= \sum_{y \in \mathcal{Y}} p(y \mid \boldsymbol{x}) \left[|y - h(\boldsymbol{x})| - \sum_{[y_l,y_r] \notin \mathcal{P}_1} p(y_l, y_r \mid \boldsymbol{x}, y) \delta(h(\boldsymbol{x}), y, y_l, y_r)) \right] \\ &= R^{MAE}(h \mid \boldsymbol{x}) - \sum_{y \in \mathcal{Y}} \sum_{[y_l,y_r] \notin \mathcal{P}_1} p(y \mid \boldsymbol{x}) p(y_l, y_r \mid \boldsymbol{x}, y) \delta(h(\boldsymbol{x}), y, y_l, y_r)\,. \end{aligned} \tag{A.13}$$

By (A.12), 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

$$\sum_{y \in \mathcal{Y}} \sum_{[y_l,y_r] \notin \mathcal{P}_1} p(y \mid \boldsymbol{x}) p(y_l, y_r \mid \boldsymbol{x}, y) \delta(h(\boldsymbol{x}), y, y_l, y_r) \leq (1-\alpha)\beta\,. \tag{A.14}$$

The bound (A.11) to be proved is obtained immediately by combing (A.13) and (A.14). ■

A.3. Proof of Proposifion 1

Let us first consider a triplet of labels (y, y_l, y_r) such that $y \notin [y_l, y_r]$. In this case, the left max-term $\max_{y \leq y_l} \big[\Delta(y, y_l) + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_l) + b_y - b_{y_l} \big]$ appearing in the surrogate (3.36) is an instance of the margin-rescaling loss instantiated for the supervised loss $\Delta(y, y_l)$ defined on labels $y \in [1, y_l - 1]$. The margin-rescaling loss is known to be the upper bound of the respective supervised loss [Tsochantaridis et al., 2005] and hence it upper bounds $\Delta_I(y_l, y_r, y)$. Analogically, we can see that the right max-term $\max_{y \geq y_r} \left[\Delta(y, y_r) + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_r) + b_y - b_{y_r}\right]$ of (3.36) is a margin-rescaling upper bound of the loss $\Delta(y, y_r)$ defined on labels $y \in [y_r + 1, Y]$ and hence also an 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 also 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 cannot be greater than a sum of the non-negative max-terms. ■

A.4. Proof of Proposition 2

To prove the proposition it is enough to show that following equalities hold

$$\sum_{\hat{y}=1}^{y_l-1} \max(0, 1 - \langle \boldsymbol{x}, \boldsymbol{w} \rangle + \theta_{\hat{y}}) = \max_{y \leq y_l} \Big[y_l - y + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_l) + b_y - b_{y_l} \Big] , \tag{A.15}$$

$$\sum_{\hat{y}=y_r}^{Y-1} \max(0, 1 + \langle \boldsymbol{x}, \boldsymbol{w} \rangle - \theta_{\hat{y}}) = \max_{y \geq y_r} \Big[y - y_r + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_r) + b_y - b_{y_r} \Big] . \tag{A.16}$$

We will provide the proof only for the equality (A.15). The proof for the equality (A.16) is similar and thus omitted. Let us do some algebra on (A.15):

$$\begin{aligned}
\sum_{\hat{y}=1}^{y_l-1} \max(0, 1 - \langle \boldsymbol{x}, \boldsymbol{w} \rangle + \theta_{\hat{y}}) &\overset{(a)}{=} \max_{y \leq y_l} \Big[\max(0, \sum_{\hat{y}=y}^{y_l-1} 1 - \langle \boldsymbol{x}, \boldsymbol{w} \rangle + \theta_{\hat{y}}) \Big] && \text{(A.17a)} \\
&\overset{(b)}{=} \max_{y \leq y_l} \Big[\max(0, y_l - y + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_l) + \sum_{\hat{y}=y}^{y_l-1} \theta_{\hat{y}}) \Big] && \text{(A.17b)} \\
&\overset{(c)}{=} \max_{y \leq y_l} \Big[\max(0, y_l - y + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_l) + b_y - b_{y_l}) \Big] && \text{(A.17c)} \\
&\overset{(d)}{=} \max_{y \leq y_l} \Big[y_l - y + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_l) + b_y - b_{y_l} \Big] . && \text{(A.17d)}
\end{aligned}$$

Equality (a) holds since $\theta_y \,, y = 1, \ldots, Y$ is a nondecreasing sequence. Equality (c) is possible due to conversion formulas[1] (3.5), (3.6). Since $y_l - y + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_l) + b_y - b_{y_l} \geq 0 \,, \forall y \leq y_l$ internal maximum in the equality (d) is redundant and thus can be omitted. ■

A.5. Proof of Theorem 4

In this section, we give detailed proofs mentioned in Section 4.4. We start with showing positiveness of functions $H^p(\epsilon)$ and $H(\epsilon)$. To show this, we need to show first that the set of all conditional distributions $p(\boldsymbol{y}, \boldsymbol{a} \mid \boldsymbol{x})$ is a compact set.

[1] Here, for simplicity, we provide proof in non-degenerated case, it can be adopted however for the generated case as well.

Lemma 2. *For any $\boldsymbol{x} \in \mathcal{X}$ a set $\mathcal{P}_{\boldsymbol{x}}$ containing all distributions $p(\boldsymbol{y}, \boldsymbol{a} \mid \boldsymbol{x}) = p(\boldsymbol{y} \mid \boldsymbol{x})\, p(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{x})$ induced from a distribution $p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{a})$ with the property A is a compact set.*

PROOF: Using $p(\boldsymbol{y}, \boldsymbol{a} \mid \boldsymbol{x}) = p(\boldsymbol{y} \mid \boldsymbol{x})\, p(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{x})$, (4.4) and (4.6) we see that for any $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{p_{ya}}(\boldsymbol{x})$ is a composition of functions with vector variables $\boldsymbol{p_y}(\boldsymbol{x})$ and $\boldsymbol{p_a}(\boldsymbol{x})$, i.e. $\boldsymbol{p_{ya}}(\boldsymbol{x}) = \mathcal{F}(\boldsymbol{p_y}(\boldsymbol{x}), \, \boldsymbol{p_z}(\boldsymbol{x}))$. The function $\mathcal{F}\colon \Delta_{|\mathcal{Y}^{\mathcal{V}}|} \times \Delta_{|\mathcal{Z}^{\mathcal{V}}|} \to \Delta_{|\mathcal{Y}^{\mathcal{V}}| \times |\mathcal{A}^{\mathcal{V}}|}$ is continuous on a compact set $\{\boldsymbol{p_y}(\boldsymbol{x}) \in \Delta_{|\mathcal{Y}^{\mathcal{V}}|} \mid p(\boldsymbol{y} \mid \boldsymbol{x}) \geq \rho\} \times \{\boldsymbol{p_z}(\boldsymbol{x}) \in \Delta_{|\mathcal{Z}^{\mathcal{V}}|} \mid p(\boldsymbol{z} \mid \boldsymbol{x}) \geq \rho\}$. Thus, $\mathcal{P}_{\boldsymbol{x}} \triangleq \{\boldsymbol{p_{ya}}(\boldsymbol{x}) = \mathcal{F}(\boldsymbol{p_y}(\boldsymbol{x}), \, \boldsymbol{p_z}(\boldsymbol{x})) \mid p(\boldsymbol{y} \mid \boldsymbol{x}) \geq \rho, p(\boldsymbol{z} \mid \boldsymbol{x}) \geq \rho, \boldsymbol{p_y}(\boldsymbol{x}) \in \Delta_{|\mathcal{Y}^{\mathcal{V}}|}, \boldsymbol{p_z}(\boldsymbol{x}) \in \Delta_{|\mathcal{Z}^{\mathcal{V}}|}\}$ is a compact set. ■

Lemma 3. *Functions $\min\limits_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_a}^{\top} \boldsymbol{\ell}^p_{\boldsymbol{t}'}$ and $\min\limits_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_y}^{\top} \boldsymbol{\ell}_{\boldsymbol{t}'}$ are continuous functions w.r.t. $\boldsymbol{p_{ya}} \in \Delta_{|\mathcal{Y}^{\mathcal{V}}| \times |\mathcal{A}^{\mathcal{V}}|}$.*

PROOF: Since $p(\boldsymbol{y} \mid \boldsymbol{x}) = \sum_{\boldsymbol{a}} p(\boldsymbol{y}, \boldsymbol{a} \mid \boldsymbol{x})$ and $p(\boldsymbol{a} \mid \boldsymbol{x}) = \sum_{\boldsymbol{y}} p(\boldsymbol{y}, \boldsymbol{a} \mid \boldsymbol{x})$ the functions $\boldsymbol{p_y}$ and $\boldsymbol{p_a}$ are continuous functions of $\boldsymbol{p_{ya}}$. Hence, both functions $\min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_a}^{\top} \boldsymbol{\ell}^p_{\boldsymbol{t}'}$ and $\min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_y}^{\top} \boldsymbol{\ell}_{\boldsymbol{t}'}$ are continuous since each of them is a composition of minimum over set of continuous functions. ■

Now we are going to give a proof of positive of function $H^p(\epsilon)$ for any positive ϵ.

Lemma 4. *Let $H^p(\epsilon, \boldsymbol{p_{ya}})\colon \mathbb{R} \times \Delta_{|\mathcal{Y}^{\mathcal{V}}| \times |\mathcal{A}^{\mathcal{V}}|} \to \mathbb{R}$ be a function defined as follows*

$$
\begin{aligned}
&\mathop{\textit{minimize}}_{\boldsymbol{t} \in \mathcal{T}^{\mathcal{V}}} && \boldsymbol{p_a}^{\top} \boldsymbol{\ell}^p_{\boldsymbol{t}} - \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_a}^{\top} \boldsymbol{\ell}^p_{\boldsymbol{t}'} \\
&\textit{subject to} && \boldsymbol{p_y}^{\top} \boldsymbol{\ell}_{\boldsymbol{t}} - \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_y}^{\top} \boldsymbol{\ell}_{\boldsymbol{t}'} \geq \epsilon.
\end{aligned}
$$

where loss functions $\ell(\boldsymbol{y}, \boldsymbol{t})$ and $\ell^p(\boldsymbol{a}, \boldsymbol{t})$ are defined by (4.1), (4.2). Then for any compact subset $\mathcal{P} \subseteq \Delta_{|\mathcal{Y}^{\mathcal{V}}| \times |\mathcal{A}^{\mathcal{V}}|}$ and for any $\epsilon > 0$ there exists $\delta > 0$ such that $\forall \boldsymbol{p_{ya}} \in \mathcal{P}$ holds $H^p(\epsilon, \boldsymbol{p_{ya}}) > \delta$, i.e. $H^p(\epsilon) = \inf_{\boldsymbol{p_{ya}} \in \mathcal{P}} H^p(\epsilon, \boldsymbol{p_{ya}}) > \delta$.

PROOF: We prove the lemma by contradiction. Assume that (8) does not hold, then $\exists \epsilon > 0$, and a sequence $(\boldsymbol{t}^m, \boldsymbol{p}^m_{\boldsymbol{ya}})$ with $\boldsymbol{t}^m \in \mathcal{T}^{\mathcal{V}}$ and $\boldsymbol{p}^m_{\boldsymbol{ya}} \in \mathcal{P}$ such that $\boldsymbol{p}^{m\,T}_{\boldsymbol{y}} \boldsymbol{\ell}_{\boldsymbol{t}^m} - \min\limits_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p}^{m\,T}_{\boldsymbol{y}} \boldsymbol{\ell}_{\boldsymbol{t}'} \geq \epsilon$ and $\lim\limits_{m \to \infty} \boldsymbol{p}^m_{\boldsymbol{a}} \boldsymbol{\ell}^p_{\boldsymbol{t}^m} - \min\limits_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p}^m_{\boldsymbol{a}} \boldsymbol{\ell}^p_{\boldsymbol{t}'} = 0$. Since $\mathcal{P}$ is compact, we can choose sub-sequence (which we still denoted as a whole sequence for simplicity) such that $\lim\limits_{m \to \infty} \boldsymbol{p}^m_{\boldsymbol{ya}} = \boldsymbol{p}^*_{\boldsymbol{ya}} \in \mathcal{P}$. Hence, from lemma (3) it follows that $\lim\limits_{m \to \infty} \boldsymbol{p}^m_{\boldsymbol{a}} \boldsymbol{\ell}^p_{\boldsymbol{t}^m} - \min\limits_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p}^*_{\boldsymbol{a}} \boldsymbol{\ell}^p_{\boldsymbol{t}'} = 0$ and $\lim\limits_{m \to \infty} \boldsymbol{p}^m_{\boldsymbol{y}} \boldsymbol{\ell}_{\boldsymbol{t}^m} - \min\limits_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p}^*_{\boldsymbol{y}} \boldsymbol{\ell}_{\boldsymbol{t}'} \geq \epsilon$. Sequence $(\boldsymbol{t}^m)$ consists of elements from the exponentially large but a finite set. Therefore there exists element of sequence $\boldsymbol{t}^* \in \mathcal{T}^{\mathcal{V}}$ such that the sequence contains infinite number of copies of $\boldsymbol{t}^*$. Let us choose this subsequence (which we again denoted as a whole sequence) such that $\lim\limits_{m \to \infty} \boldsymbol{t}^m = \boldsymbol{t}^*$. Note that $\lim\limits_{m \to \infty} \boldsymbol{p}^m_{\boldsymbol{ya}} = \boldsymbol{p}^*_{\boldsymbol{ya}}$ stays same. It follows that $\boldsymbol{p}^*_{\boldsymbol{a}} \boldsymbol{\ell}^p_{\boldsymbol{t}^*} - \min\limits_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p}^*_{\boldsymbol{a}} \boldsymbol{\ell}^p_{\boldsymbol{t}'} = 0$ and $\boldsymbol{p}^*_{\boldsymbol{y}} \boldsymbol{\ell}_{\boldsymbol{t}^*} - \min\limits_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p}^*_{\boldsymbol{y}} \boldsymbol{\ell}_{\boldsymbol{t}'} \geq \epsilon, \ \epsilon > 0$. We have thus obtained the contradiction, i.e. we have found a model $\boldsymbol{p}^* \in \mathcal{P}$ for which lemma (1) does not hold. ■

Lemma 5. *If $\forall \epsilon > 0, H^p(\epsilon) \triangleq \inf\limits_{\boldsymbol{p_{ay}} \in \mathcal{P}_{\boldsymbol{x}}} H^p(\epsilon, \boldsymbol{p_{ay}}) > 0$ for the loss functions $\ell(\boldsymbol{y}, \boldsymbol{t})$ and $\ell^p(\boldsymbol{a}, \boldsymbol{t})$ defined by (4.1), (4.2) then there exists a nonnegative concave function $\xi\colon \mathbb{R} \to \mathbb{R}_+$,* *right continuous at* 0 *with* $\xi(0) = 0$, *such that* $\forall\ \boldsymbol{h}\colon \mathcal{X} \to \mathcal{T}^{\mathcal{V}}$ *and for all distributions with property A it holds that*

$$\mathbb{E}_{p(\boldsymbol{x})}\boldsymbol{p_y}(\boldsymbol{x})^\top \boldsymbol{\ell}_{\boldsymbol{h}(\boldsymbol{x})} - \mathbb{E}_{p(\boldsymbol{x})} \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_y}(\boldsymbol{x})^\top \boldsymbol{\ell}_{\boldsymbol{t}'} \leq \xi\left(\mathbb{E}_{p(\boldsymbol{x})}\boldsymbol{p_a}(\boldsymbol{x})^\top \boldsymbol{\ell}^p_{\boldsymbol{h}(\boldsymbol{x})} - \mathbb{E}_{p(\boldsymbol{x})} \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_a}(\boldsymbol{x})^\top \boldsymbol{\ell}^p_{\boldsymbol{t}'}\right).$$

The main idea of Lemma 5 proof is analogical to the proof of Corollary 26 in [Zhang, 2004a]. Thus, we provide proof only for Lemma 5 together with two auxiliary lemmas needed for its proof and proof of "flipped" version of this lemma we leave to the reader.

Lemma 6. *Let* $\mu(\epsilon)\colon \mathbb{R} \to \mathbb{R}_+$ *be a convex function such that* $\mu(\epsilon) \leq H^p(\epsilon)$. *Then for any classifier* $\boldsymbol{h}(\boldsymbol{x})\colon \mathcal{X} \to \mathcal{T}$ *we have*

$$\mu(\mathbb{E}_{p(\boldsymbol{x})}\boldsymbol{p_y}(\boldsymbol{x})^\top \boldsymbol{\ell}_{\boldsymbol{h}(\boldsymbol{x})} - \mathbb{E}_{p(\boldsymbol{x})} \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_y}(\boldsymbol{x})^\top \boldsymbol{\ell}_{\boldsymbol{t}'}) \leq \mathbb{E}_{p(\boldsymbol{x})}\boldsymbol{p_a}(\boldsymbol{x})^\top \boldsymbol{\ell}^p_{\boldsymbol{h}(\boldsymbol{x})} - \mathbb{E}_{p(\boldsymbol{x})} \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_a}(\boldsymbol{x})^\top \boldsymbol{\ell}^p_{\boldsymbol{t}'}$$

PROOF: Using Jensen's inequality together with inequality

$$H^p(\boldsymbol{p_y}^\top \boldsymbol{\ell_t} - \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_y}^\top \boldsymbol{\ell}_{\boldsymbol{t}'}) \leq \boldsymbol{p_a}^\top \boldsymbol{\ell}^p_{\boldsymbol{t}} - \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_a}^\top \boldsymbol{\ell}^p_{\boldsymbol{t}'}$$

we have

$$\begin{aligned} \mu(\mathbb{E}_{p(\boldsymbol{x})}\boldsymbol{p_y}(\boldsymbol{x})^\top \boldsymbol{\ell}_{\boldsymbol{h}(\boldsymbol{x})} - \mathbb{E}_{p(\boldsymbol{x})} \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_y}(\boldsymbol{x})^\top \boldsymbol{\ell}_{\boldsymbol{t}'}) &\leq \\ \mathbb{E}_{p(\boldsymbol{x})}\mu(\boldsymbol{p_y}(\boldsymbol{x})^\top \boldsymbol{\ell}_{\boldsymbol{h}(\boldsymbol{x})} - \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_y}(\boldsymbol{x})^\top \boldsymbol{\ell}_{\boldsymbol{t}'}) &\leq \\ \mathbb{E}_{p(\boldsymbol{x})}H^p(\boldsymbol{p_y}(\boldsymbol{x})^\top \boldsymbol{\ell}_{\boldsymbol{h}(\boldsymbol{x})} - \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_y}(\boldsymbol{x})^\top \boldsymbol{\ell}_{\boldsymbol{t}'}) &\leq \\ \mathbb{E}_{p(\boldsymbol{x})}(\boldsymbol{p_a}(\boldsymbol{x})^\top \boldsymbol{\ell}^p_{\boldsymbol{h}(\boldsymbol{x})} - \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p_a}(\boldsymbol{x})^\top \boldsymbol{\ell}^p_{\boldsymbol{t}'})&. \end{aligned}$$

■

Lemma 7. *Let* $\zeta_*(\epsilon) = \sup\limits_{a \geq 0, b} \{a\epsilon + b \mid \forall z \geq 0, az + b \leq H^p(z)\}$, *then* ζ_* *is a convex function. It has the following properties:*

- $\zeta_*(\epsilon) \leq H^p(\epsilon)$,
- $\zeta_*(\epsilon)$ *is non-decreasing,*
- *for all convex functions* $\zeta(\cdot)$ *such that* $\zeta(\epsilon) \leq H^p(\epsilon)$, $\zeta(\epsilon) \leq \zeta_*(\epsilon)$.
- *Assume that* $\exists a > 0$ *and* $b \in \mathbb{R}$ *such that* $a\epsilon + b \leq H^p(\epsilon)$ *and* $\forall \epsilon > 0, H^p(\epsilon) > 0$. *Then* $\forall \epsilon > 0, \zeta_*(\epsilon) > 0$.

Lemma 7 is a proposition 25 from [Zhang, 2004a] for the function $H^p(\epsilon)$ Thus, we omit its proof here. Now we are ready to prove Lemma 5.

PROOF: Consider $\zeta_*(\epsilon)$ in Lemma 7, Let $\xi(\delta) = \sup\{\epsilon\colon \epsilon \geq 0, \zeta_*(\epsilon) \leq \delta\}$. Then $\zeta_*(\epsilon) \leq \delta$ implies $\epsilon \leq \xi(\delta)$. Therefore desired inequality comes from Lemma 6.

Given $\delta_1, \delta_2 \geq 0$: from $\zeta_*(\frac{\xi(\delta_1)+\xi(\delta_2)}{2}) \leq \frac{\delta_1+\delta_2}{2}$ we know that $\frac{\xi(\delta_1)+\xi(\delta_2)}{2} \leq \xi(\frac{\delta_1+\delta_2}{2})$. Thus, $\xi(\epsilon)$ is concave function.

We now only need to show that $\xi(\epsilon)$ is continuous at 0. From the boundedness of $\ell(\boldsymbol{y},\boldsymbol{t})$, we know that $H^p(z) = +\infty$ when $z > \max\limits_{\boldsymbol{y}\in\mathcal{Y}^\mathcal{V},\boldsymbol{t}\in\mathcal{T}^\mathcal{V}} \ell(\boldsymbol{y},\boldsymbol{t})$. Therefore $\exists a > 0$ and $b \in \mathbb{R}$ such that $a\epsilon + b \leq H^p(\epsilon)$. Now we pick up any $\epsilon' > 0$, and let $\delta' = \frac{\zeta_*(\epsilon')}{2}$. We know from Lemma 7 that $\delta' > 0$. This implies that $\xi(\delta) < \epsilon'$ when $\delta' < \delta$. ■

Here we just give formulation of "flipped" version of Lemma 5 and its auxiliary Lemma 8. To prove Lemma 9, we need modified Lemma 6 and 7 for the function from Lemma 8 which is straightforward to do, thus we leave it for the reader.

Lemma 8. *Let $H(\epsilon, \boldsymbol{p_{ya}})\colon \mathbb{R}\times\Delta_{|\mathcal{Y}^\mathcal{V}|\times|\mathcal{A}^\mathcal{V}|} \to \mathbb{R}$ be a function defined as follows*

$$\begin{aligned} &\underset{\boldsymbol{t}\in\mathcal{T}^\mathcal{V}}{\textit{minimize}} && \boldsymbol{p_y}^\top \boldsymbol{\ell_t} - \min_{\boldsymbol{t}'\in\mathcal{T}^\mathcal{V}} \boldsymbol{p_y}^\top \boldsymbol{\ell_{t'}} \\ &\textit{subject to} && \boldsymbol{p_a}^\top \boldsymbol{\ell}^p_{\boldsymbol{t}} - \min_{\boldsymbol{t}'\in\mathcal{T}^\mathcal{V}} \boldsymbol{p_a}^\top \boldsymbol{\ell}^p_{\boldsymbol{t}'} \geq \epsilon. \end{aligned}$$

where loss functions $\ell(\boldsymbol{y},\boldsymbol{t})$ and $\ell^p(\boldsymbol{a},\boldsymbol{t})$ are defined by (4.1), (4.2). Then for any compact subset $\mathcal{P} \subseteq \Delta_{|\mathcal{Y}^\mathcal{V}|\times|\mathcal{A}^\mathcal{V}|}$ and for any $\epsilon > 0$ there exists $\delta > 0$ such that $\forall \boldsymbol{p_{ya}} \in \mathcal{P}$ holds $H(\epsilon, \boldsymbol{p_{ya}}) > \delta$, i.e. $H(\epsilon) = \inf\limits_{\boldsymbol{p_{ya}}\in\mathcal{P}} H(\epsilon, \boldsymbol{p_{ya}}) > \delta$.

PROOF: The proof is analogous to the proof of Lemma 4. ■

Lemma 9. *If $\forall\epsilon > 0, H(\epsilon) \triangleq \inf\limits_{\boldsymbol{p_{ay}}\in\mathcal{P}_{\boldsymbol{x}}} H(\epsilon, \boldsymbol{p_{ay}}) > 0$ for the loss functions $\ell(\boldsymbol{y},\boldsymbol{t})$ and $\ell^p(\boldsymbol{a},\boldsymbol{t})$ defined by (4.1), (4.2) then there exists a nonnegative concave function $\zeta\colon \mathbb{R}\to\mathbb{R}_+$, right continuous at 0 with $\zeta(0) = 0$, such that $\forall\ \boldsymbol{h}\colon \mathcal{X}\to\mathcal{T}^\mathcal{V}$ and for all distributions with property A it holds that*

$$\mathbb{E}_{p(\boldsymbol{x})}\boldsymbol{p_a}(\boldsymbol{x})^\top \boldsymbol{\ell}^p_{\boldsymbol{h}(\boldsymbol{x})} - \mathbb{E}_{p(\boldsymbol{x})} \min_{\boldsymbol{t}'\in\mathcal{T}^\mathcal{V}} \boldsymbol{p_a}(\boldsymbol{x})^\top \boldsymbol{\ell}^p_{\boldsymbol{t}'} \leq$$
$$\zeta\left(\mathbb{E}_{p(\boldsymbol{x})}\boldsymbol{p_y}(\boldsymbol{x})^\top \boldsymbol{\ell}_{\boldsymbol{h}(\boldsymbol{x})} - \mathbb{E}_{p(\boldsymbol{x})} \min_{\boldsymbol{t}'\in\mathcal{T}^\mathcal{V}} \boldsymbol{p_y}(\boldsymbol{x})^\top \boldsymbol{\ell}_{\boldsymbol{t}'}\right).$$

Proof of Lemma 9 is similar to proof of Lemma 5.

Bibliography

Altun, Y., Tsochantaridis, I., and Hofmann, T. (2003). Hidden markov support vector machines. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3–10. 18

Antoniuk, K., Franc, V., and Hlaváč, V. (2012). Learning markov networks by analytic center cutting plane method. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 2250–2253. 39

Antoniuk, K., Franc, V., and Hlaváč, V. (2016). V-shaped interval insensitive loss for ordinal classification. *Machine Learning*, pages 1–23.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156. 15

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization.* Cambridge University Press, New York, NY, USA. 38

Cao, L. and Chen, C. W. (2003). A novel product coding and recurrent alternate decoding scheme for image transmission over noisy channels. *IEEE Transactions on Communications*, 51(9):1426–1431. 18

Chu, W. and Ghahramani, Z. (2005). Preference learning with gaussian processes. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 137–144. 19

Chu, W. and Keerthi, S. S. (2005). New approaches to support vector ordinal regression. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 145–152. 19, 20, 26, 27, 28, 29, 30, 40, 41

Chuong, B. D., Quoc, L., Teo, C. H., Chapelle, O., and Smola, A. (2008). Tighter bounds for structured estimation. In *Proceedings of the Neural Information Processing Systems (NIPS)*, pages 281–288. 17, 57, 67

Cid-Sueiro, J., García-García, D., and Santos-Rodríguez, R. (2014). Consistency of losses for learning from weak labels. In *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, pages 197–210. 15, 16, 69

Collins, M. and Koo, T. (2005). Discriminative reranking for natural language parsing. *Computational Linguistics*, 31(1):25–70. 18

Cour, T., Sapp, B., and Taskar, B. (2011). Learning from partial labels. *Journal of Machine Learning Research*, 12:1225–1261. 5, 10, 15, 16, 31, 32, 69

Crammer, K. and Singer, Y. (2001). Pranking with ranking. In *Proceedings of the Neural Information Processing Systems (NIPS)*, pages 641– 647. 19, 26

Crammer, K. and Singer, Y. (2002). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292. 68

Crammer, K. and Singer, Y. (2005). Online ranking by projecting. *Neural Computation*, 17(1):145–175. 18

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. 19

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). 19

Do, T.-M.-T. and Artières, T. (2009). Large margin training for hidden markov models with partially observed states. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 265–272. 10, 31

Do, T.-M.-T. and Artières, T. (2012). Regularized bundle methods for convex and non-convex risks. *Journal of Machine Learning Research*, 13(1):3539–3583. 17

Fernandes, E. R. and Brefeld, U. (2011a). Learning from partially annotated sequences. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, pages 407–422. 10, 17, 57, 67

Fernandes, E. R. and Brefeld, U. (2011b). Learning from partially annotated sequences. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, pages 407–422. 18

Franc, V. and Sonneburg, S. (2009). Optimized cutting plane algorithm for large-scale risk minimization. *Journal of Machine Learning Research*, 10:2157–2232. 30

Franc, V., Sonnenburg, S., and Werner, T. (2012). *Cutting-Plane Methods in Machine Learning*, chapter 7, pages 185–218. The MIT Press, Cambridge,USA. 37

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. 19

Gao, W. and Zhou, Z.-H. (2013). On the consistency of multi-label learning. *Artificial Intelligence*, 199-200:22–44. 59, 65

Gaudioso, M. and Monaco, M. (1992). Variants to the cutting plane approach for convex nondifferentiable optimization. *Optimization*, 25(1):65–75. 18

Girshick, R. B., Felzenszwalb, P. F., and McAllester, D. A. (2011). Object detection with grammar models. In *Proceedings of the Neural Information Processing Systems (NIPS)*, pages 442–450. 17, 57, 67

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*, 94:206–211. 39, 40

Guo, G. and Mu, G. (2010). Human age estimation: What is the influence across race and gender? In *Proceeding of the Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 71–78.

Hill, S. I. and Doucet, A. (2007). A framework for kernel-based multi-category classification. *Journal of Artificial Intelligence Research*, 30:525–564. 15

Horst, R., Phong, T., Thoai, N., and Vries, J. (1991). On solving a d.c. programming problem by a sequence of linear programs. *Journal of Global Optimization*, 1(2):183–203. 18

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. 41

Kawahara, Y. and Washio, T. (2011). Prismatic algorithm for discrete d.c. programming problem. In *Proceedings of the Neural Information Processing Systems (NIPS)*, pages 2106–2114. 18

Kiwiel, K. C. (1985). *Methods of descent for nondifferentiable optimization*. Lecture notes in mathematics. Springer-Verlag, Berlin, New York. 18

Kiwiel, K. C. (1990). Proximity control in bundle methods for convex nondifferentiable minimization. *Mathematical Programming*, 46:105–122. 17

Kumar, M. P., Packer, B., and Koller, D. (2010). Self-paced learning for latent variable models. In *Proceedings of the Neural Information Processing Systems (NIPS)*, pages 1189–1197. 17

Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. (2009). Attribute and simile classifiers for face verification. In *Proceeding of the International Conference on Computer Vision (ICCV)*, pages 365–372. 41

Li, C., Zhang, J., and Chen, Z. (2013). Structured output learning with candidate labels for local parts. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, pages 336–352.

Li, L. and Lin, H.-T. (2006). Ordinal regression by extended binary classification. In *Proceedings of the Neural Information Processing Systems (NIPS)*, pages 865–872. 19, 26, 29, 30, 40, 41

Liu, Y. (2007). Fisher consistency of multicategory support vector machines. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTAS)*, pages 291–298. 8, 15, 68

Lou, X. and Hamprecht, F. A. (2012). Structured learning from partial annotations. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1519–1526. 10, 17, 31, 57, 66, 67, 70

Luo, J. and Orabona, F. (2010). Learning from candidate labeling sets. In *Proceedings of the Neural Information Processing Systems (NIPS)*, pages 1504–1512. 10, 17, 31, 57, 67

McAllester, D. A. and Keshet, J. (2011). Generalization bounds and consistency for latent structural probit and ramp loss. In *Proceedings of the Neural Information Processing Systems (NIPS)*, pages 2205–2212. 70

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statical Society*, 42(2):109–142. 19, 21

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. 41

Mukherjee, I. and Schapire, R. E. (2013). A theory of multiclass boosting. *Journal of Machine Learning Research*, 14(1):437–497. 16

Pedregosa, F., Bach, F. R., and Gramfort, A. (2014). On the consistency of ordinal regression methods. *CoRR*, abs/1408.2327. 16, 72

Ramaswamy, H. G. and Agarwal, S. (2012). Classification calibration dimension for general multiclass losses. In *Proceedings of the Neural Information Processing Systems (NIPS)*, pages 2087–2095. 7, 15, 57, 65, 66, 69

Ramaswamy, H. G., Tewari, A., and Agarwal, S. (2015a). Consistent algorithms for multiclass classification with a reject option. *CoRR*, abs/1505.04137. 16

Ramaswamy, H. G., Tewari, A., and Agarwal, S. (2015b). Convex calibrated surrogates for hierarchical classification. In *Proceedings of the International Conference on Machine Learning, (ICML)*, pages 1852–1860. 16

Reid, M. D. and Williamson, R. C. (2010). Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422. 70

Reid, M. D., Williamson, R. C., and Sun, P. (2012). The convexity and design of composite multiclass losses. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 687–694. 70, 71

Rennie, J. D. and Srebro, N. (2005). Loss functions for preference levels: Regression with discrete ordered labels. In *Proceedings of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling*, pages 180–186. 19

Ricanek, K. and Tesafaye, T. (2006). Morph: A longitudial image database of normal adult age-progression. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition (FGR)*, pages 341–345. 40

Santos-Rodríguez, R., Guerrero-Curieses, A., Alaiz-Rodríguez, R., and Cid-Sueiro, J. (2009). Cost-sensitive learning based on bregman divergences. *Machine Learning Journal*, 76(2):271–285. 15

Sarawagi, S. and Gupta, R. (2008). Accurate max-margin training for structured output spaces. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 888–895. 17, 57, 66, 67

Schlesinger, M. (1968). A connection between learning and self-learning in the pattern recognition (in Russian). *Kibernetika*, 2:81–88. 19

Schlesinger, M. and Hlaváč, V. (2002). *Ten Lectures on Statistical and Structural Pattern Recognition*. Computational Imaging and Vision. Kluwer Academic Publishers, Dordrecht, The Netherlands. 18

Shashua, A. and Levin, A. (2002). Ranking with large margin principle: Two approaches. In *Proceedings of the Neural Information Processing Systems (NIPS)*, pages 961–968. 19, 26

Shi, Q., Reid, M. D., Caetano, T. S., van den Hengel, A., and Wang, Z. (2015). A hybrid loss for multiclass and structured prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 37(1):2–12. 7, 16

Sonnenburg, S. and Franc, V. (2010). Coffin: A computational framework for linear svms. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 999–1006. 41

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. *Journal of Machine Learning Research*, 11:1799–1802. 40

Teo, C. H., Vishwanthan, S., Smola, A. J., and Le, Q. V. (2010). Bundle methods for regularized risk minimization. *Journal of Machine Learning Research*, 11:311–365. 17, 30, 37, 38, 39

Tewari, A. and Bartlett, P. L. (2007). On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025. 15, 59, 65, 68

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. 29, 76

Uřičář, M., Franc, V., and Hlaváč, V. (2012). Detector of facial landmarks learned by the structured output SVM. In *Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 547–556. 41, 61

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA. 7

Vapnik, V. N. (1998). *Statistical learning theory*. Adaptive and Learning Systems. Wiley, New York, New York, USA. 31, 58, 63

Vedaldi, A. and Zisserman, A. (2009). Structured output regression for detection with partial truncation. In *Proceedings of the Neural Information Processing Systems (NIPS)*, pages 1928–1936. 17, 57, 67

Vernet, E., Williamson, R. C., and Reid, M. D. (2011). Composite multiclass losses. In *Proceedings of the Neural Information Processing Systems (NIPS)*, pages 1224–1232. 70, 71

Wang, Y. and Mori, G. (2010). A discriminative latent model of object classes and attributes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 155–168. 17, 57, 67

Yu, C.-N. J. and Joachims, T. (2009). Learning structural svms with latent variables. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1169–1176. 17, 57, 67

Yu, H.-F., Jain, P., Kar, P., and Dhillon, I. S. (2014). Large-scale multi-label learning with missing labels. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 593–601. 15, 17, 57, 67, 69

Zhang, T. (2004a). Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251. 7, 15, 59, 65, 78

Zhang, T. (2004b). Statistical behaviour and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 31(1):56–134. 7, 15

Zhang, Z., Jordan, M. I., Li, W., and Yeung, D. (2009). Coherence functions for multicategory margin-based classification methods. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTAS)*, pages 647–654. 15, 16

Zhu, J., Zou, H., Rosset, S., and Hastie, T. (2009). Multi-class adaboost. *Statistics and Its Interface*, 2:349–360. 16

Zhu, L., Chen, Y., Yuille, A. L., and Freeman, W. T. (2010). Latent hierarchical structural learning for object detection. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, pages 1062–1069. 17, 57, 67

A. Author's publications

A.1. Publications related to the thesis

Impacted journal papers excerpted by ISI

Antoniuk, K., Franc, V., and Hlaváč, V. (2016). V-shaped interval insensitive loss for ordinal classification. *Machine Learning*, 103(2):261–283. [45%].

Conference papers excerpted by ISI

Antoniuk, K., Franc, V., and Hlaváč, V. (2012). Learning markov networks by analytic center cutting plane method. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 2250–2253. [45%].

Antoniuk, K., Franc, V., and Hlavac, V. (2013). Mord: Multi-class classifier for ordinal regression. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, pages 96–111. [45%].

Antoniuk, K., Franc, V., and Hlavac, V. (2014). Interval insensitive loss for ordinal classification. In *Proceedings of the Asian Conference on Machine Learning (ACML)*, pages 189–204. [45%].

Antoniuk, K., Franc, V., and Hlaváč, V. (2015). Consistency of structured output learning with missing labels. In *Proceedings of the Asian Conference on Machine Learning (ACML)*, pages 81 – 95. [55%].

Other conference papers

Antoniuk, K. (2013). Statistical formulation of structured output learning from partially annotated examples. In *Proceedings of the International Student Conference on Electrical Engineering*, pages 1–5. [100%].

A.2. Other publications

Impacted journal papers excerpted by ISI

Schlesinger, M. I. and Antoniuk, K. V. (2011). Diffusion algorithms and structural recognition optimization problems. *Cybernetics and Systems Analysis*, 47(2):175–192.

Impacted journal papers not excerpted by ISI

Schlesinger, M. I., Antoniuk, K. V., and Vodolazskii, E. V. (2011). Optimal labelling problems, their relaxation and equivalent transformations. *Control Systems and Computers*, (2):55–70.

B. Citations of author's work

Holešovský, O. (2015). Face descriptor learned by convolutional neural networks. Master's thesis, Czech Technical University in Prague.

Horaud, R. (2013). Humanoids with auditory and visual abilities in populated spaces (humavips). Technical report, INRIA Grenoble Rhône-Alpes 655.

Hritsenko, V. and Schlesinger, M. I. (2014). Formal models, problems and algorithms of visual thinking. In *Proceedings of the Actual Problems of Telecommunication, Computer Science and Engineers Teaching (TCSET).*

Neil, H. (2014). *Efficient Bayesian active learning and matrix modelling.* PhD thesis, University of Cambridge.

Savchynskyy, B., Kappes, J. H., Swoboda, P., and Schnörr, C. (2013). Global map-optimality by shrinking the combinatorial search area with convex relaxation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1950–1958.

Schlesinger, M., Vodolazskiy, E., and Lopatka, N. (2011). Proceedings of the energy minimization methods in computer vision and pattern recognition (emmcvpr). In *Stop condition for subgradient minimization in dual relaxed (max,+) Problem*, pages 118–131.

Shekhovtsov, A., Swoboda, P., and Savchynskyy, B. (2015). Maximum persistency via iterative relaxed inference with graphical models. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, pages 521–529.

Uřičář, M., Timofte, R., Rothe, R., Matas, J., and Van Gool, L. (2016). Structured output SVM prediction of apparent age, gender and smile from deep features. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.*

Werner, T. (2011). Zero-temperature limit of a convergent algorithm to minimize the bethe free energy. *CoRR*, abs/1112.5298.

Werner, T. (2013). Marginal consistency: unifying convergent message passing and constraint propagation. Technical Report CTU-CMP-2013-26, Czech Technical University in Prague.

Werner, T. (2015). Marginal consistency: Upper-bounding partition functions over commutative semirings. *IEEE Transactions on pattern analysis and machine intelligence (PAMI)*, 37(7):1455–1468.