JMLR: Workshop and Conference Proceedings 45:81–95, 2015 ACML 2015

Consistency of structured output learning with missing labels

Kostiantyn Antoniuk[1] ANTONKOS@CMP.FELK.CVUT.CZ
Vojtěch Franc[1] XFRANCV@CMP.FELK.CVUT.CZ
Václav Hlaváč[2] HLAVAC@FEL.CVUT.CZ
[1] *Faculty of Electrical Engineering, Czech Technical University in Prague*
[2] *Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague*

Editor: Geoffrey Holmes and Tie-Yan Liu

Abstract

In this paper we study statistical consistency of partial losses suitable for learning structured output predictors from examples containing missing labels. We provide sufficient conditions on data generating distribution which admit to prove that the expected risk of the structured predictor learned by minimizing the partial loss converges to the optimal Bayes risk defined by an associated complete loss. We define a concept of surrogate classification calibrated partial losses which are easier to optimize yet their minimization preserves the statistical consistency. We give some concrete examples of surrogate partial losses which are classification calibrated. In particular, we show that the ramp-loss which is in the core of many existing algorithms is classification calibrated.

Keywords: missing labels, statistical consistency, structured output learning

1. Introduction

This paper studies statistical consistency of risk minimization methods designed for learning structured output predictors from a set of partially annotated training examples. 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 labellings only for a subset of the local parts. In contrast, the conventional supervised methods require all local parts to be annotated. Since the missing label scenario is common in practice, several methods learning structured predictors from partial annotations have been proposed during the past few years: Vedaldi and Zisserman (2009); Lou and Hamprecht (2012); Fernandes and Brefeld (2011); Li et al. (2013); Yu et al. (2014). These algorithms implement the empirical risk minimization principle using certain surrogate loss functions that can be evaluated on partially annotated examples. Despite excellent empirical results a clear statistical justification for these methods has not been provided so far. In this paper we attempt to fix this gap by extending the concept of statistical consistent learning (e.g. Zhang (2004a); Tewari and Bartlett (2007); Gao and Zhou (2011); Ramaswamy and Agarwal (2012)) to the structured output setting with the partially annotated examples.

We assume that the target (complete) loss is additive over the local parts, that is, it is a sum of single label losses for each local part. We analyze a partial loss that can be constructed from any complete additive loss by simply neglecting those single label losses for which the labels in the training set are missing. A detailed definition of the considered setting is described in Section 2. The main contribution of this paper is sufficient conditions on data generating statistical model

which admits to prove that minimization of the partial loss yields structured predictor with expected risk converging in probability to the Bayes risk (i.e, the minimal attainable risk) defined by the associated complete loss. The statistical model is defined in Section 3 and the consistency is proved in Section 4. Since minimizing the partial loss directly is often a hard problem because of its discrete domain, in Section 5 we analyze surrogate partial losses which are easier to optimize. By adapting the general framework of Ramaswamy and Agarwal (2012) to our setting we define the concept of a surrogate classification calibrated partial loss which preserve the statistical consistency. Finally, we give some concrete examples of the classification calibrated surrogate partial losses. In particular, we show that the ramp-loss, whose variants have been previously used to construct algorithms for learning from partial annotations e.g. in Lou and Hamprecht (2012); Fernandes and Brefeld (2011); Li et al. (2013), is calibrated and hence the so far heuristic methods are statistically consistent. Conclusions are given in Section 6.

The consistency of the ramp-loss has been previously studied in McAllester and Keshet (2011). 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 in contrast 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.

2. Setting

We first describe the common fully supervised setting in section 2.1. The formulation of learning from partially annotated examples which is annalyzed in this paper is then defined in section 2.2. We use a notation adopted from Ramaswamy and Agarwal (2012) being a paper on which we built our results.

2.1. Common fully supervised setting

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 supervised setting we are given a training set $\mathcal{D}_m = \{(\boldsymbol{x}^1, \boldsymbol{y}^1), \ldots, (\boldsymbol{x}^m, \boldsymbol{y}^m)\} \in (\mathcal{X} \times \mathcal{Y}^{\mathcal{V}})^m$ 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 mapping 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 in general $\mathcal{T}$ can be different from $\mathcal{Y}$, for example in the classification with the reject option when $\mathcal{T} = \mathcal{Y} \cup \{\text{don't know}\}$.

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 $\boldsymbol{t} \in \mathcal{T}^{\mathcal{V}}$. In this paper we confine ourselves to losses which are additive over the local parts being a natural choice in many applications, i.e.

$$\ell(\boldsymbol{y}, \boldsymbol{t}) = \sum_{v \in \mathcal{V}} \ell_v(y_v, t_v) \tag{1}$$

where $\ell_v\colon \mathcal{Y} \times \mathcal{T} \to \mathbb{R}_+$, $v \in \mathcal{V}$, are single label losses. Throughout the paper 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

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

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

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

A direct minimization of the loss $\boldsymbol{\ell}$ is often a hard problem, therefore it is common to replace $\boldsymbol{\ell}$ 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

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

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 $\mathrm{pred}\colon \hat{\mathcal{T}} \to \mathcal{T}$. The ℓ-risk of the resulting decision function $\mathrm{pred}(\boldsymbol{f}(\boldsymbol{x}))$ is $R^\ell(\mathrm{pred} \circ \boldsymbol{f}; p)$. 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 $\mathrm{pred}(\hat{\boldsymbol{t}}) \in \mathrm{Argmax}_{\boldsymbol{y}\in\mathcal{Y}^\mathcal{V}} \hat{t}_{\boldsymbol{y}}$ which yields the linear structured output classifier $\boldsymbol{h}(\boldsymbol{x}) \in \mathrm{Argmax}_{\boldsymbol{y}\in\mathcal{Y}^\mathcal{V}}\langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}, \boldsymbol{y})\rangle$.

Under suitable conditions the uniform low of large numbers applies (Vapnik (1998)) and learning $\boldsymbol{f}_m$ from $\mathcal{D}_m$ by minimizing the empirical risk $R^\psi_{\mathrm{emp}}(\boldsymbol{f}) = \frac{1}{m}\sum_{i=1}^m \psi(\boldsymbol{y}^i, \boldsymbol{x}^i)$ 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 (2011)) that the consistency with respect to the ψ-risk implies the consistency with respect to the ℓ-risk provided the surrogate loss ψ is classification calibrated w.r.t the loss ℓ. We will extend this result to the setting when the training examples are partially annotated as defined in the next section.

2.2. Learning with missing labels studied in this paper

Let us consider that we are given a training set $\hat{\mathcal{D}}_m = \{(\boldsymbol{x}^1, \boldsymbol{a}^1), \ldots, (\boldsymbol{x}^m, \boldsymbol{a}^m)\} \in (\mathcal{X} \times \mathcal{A}^\mathcal{V})^m$ 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 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 $\hat{\mathcal{D}}_m$ 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 1 *For a given (complete) additive loss $\ell \colon \mathcal{Y}^{\mathcal{V}} \times \mathcal{T}^{\mathcal{V}} \to \mathbb{R}_+$ defined by (1) the associated partial loss $\ell^p \colon \mathcal{A}^{\mathcal{V}} \times \mathcal{T}^{\mathcal{V}} \to \mathbb{R}_+$ is defined as*

$$\ell^p(\boldsymbol{a}, \boldsymbol{t}) = \sum_{v \in \mathcal{V}} [\![|a_v| = 1]\!] \ell_v(a_v, t_v)^1 \,, \tag{2}$$

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.

The partial loss ℓ^p simply neglects the local losses corresponding to the missing labels. We can now learn a decision function $\boldsymbol{h} \colon \mathcal{X} \to \mathcal{T}^{\mathcal{V}}$ by minimizing the ℓ^p-risk

$$R^{\ell^p}(\boldsymbol{h}; p'') = \mathbb{E}_{p''(\boldsymbol{x}, \boldsymbol{a})} \ell^p(\boldsymbol{a}, \boldsymbol{h}(\boldsymbol{x})) = \mathbb{E}_{p(\boldsymbol{x})} \sum_{\boldsymbol{a} \in \mathcal{A}^{\mathcal{V}}} p(\boldsymbol{a} \mid \boldsymbol{x}) \ell^p(\boldsymbol{a}, \boldsymbol{h}(\boldsymbol{x})) = \mathbb{E}_{p(\boldsymbol{x})} \boldsymbol{p}_{\boldsymbol{a}}(\boldsymbol{x})^T \boldsymbol{\ell}^p_{\boldsymbol{h}(\boldsymbol{x})}$$

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

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

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. As the first contribution of this paper, we provide sufficient conditions on $p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{a})$ which guarantee that minimization of the ℓ^p-risk is equivalent to the minimization of the ℓ-risk, namely, that $R^{\ell^p}(\boldsymbol{h}_m; p'') \xrightarrow{P} R^{\ell^p}_*(p'')$ if and only if $R^{\ell}(\boldsymbol{h}_m; p') \xrightarrow{P} R^{\ell}_*(p')$ where $\boldsymbol{h}_m$ is a decision function learned from a random training set $\hat{\mathcal{D}}_m$ and $\xrightarrow{P}$ denotes convergence in probability. This claim justifies algorithms which learn $\boldsymbol{h}$ by approximately minimizing the partial loss ℓ^p by implementing the empirical risk minimization principle like those proposed e.g. in Lou and Hamprecht (2012); Fernandes and Brefeld (2011); Li et al. (2013). The conditions on the statistical model are described in Section 3 and the consistency is proved in Section 4.

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.

3. Statistical Model of Partial Annotations

In this section we describe a generative model of the partially annotated data. 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.

$$p(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{x}) = p(\boldsymbol{y} \mid \boldsymbol{x})\; p(\boldsymbol{z} \mid \boldsymbol{x})\;, \tag{3}$$

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

$$p(\boldsymbol{a} \mid \boldsymbol{x}) = \sum_{\boldsymbol{y} \in \mathcal{Y}^{\mathcal{V}}} \sum_{\boldsymbol{z} \in \mathcal{Z}^{\mathcal{V}}} p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{z} \mid \boldsymbol{x}) [\![\boldsymbol{a} = \boldsymbol{\alpha}(\boldsymbol{y}, \boldsymbol{z})]\!]\;. \tag{4}$$

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

$$p(\boldsymbol{x}, \boldsymbol{a}) = p(\boldsymbol{x})\; p(\boldsymbol{a} \mid \boldsymbol{x})\;. \tag{5}$$

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

$$p(\boldsymbol{y} \,|\, \boldsymbol{x}, \boldsymbol{a}) = \frac{p(\boldsymbol{y} \,|\, \boldsymbol{x}) c(\boldsymbol{y}, \boldsymbol{a})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}^{\mathcal{V}}} p(\boldsymbol{y}' \,|\, \boldsymbol{x}) c(\boldsymbol{y}', \boldsymbol{a})}\;, \tag{6}$$

We use the convention that $p(\boldsymbol{y} \,|\, \boldsymbol{x}, \boldsymbol{a}) = 0$ if the denominator and the numerator are zero. The distribution (6) together with (5) defines a joint distribution

$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{a}) = p(\boldsymbol{x})\; p(\boldsymbol{a} \mid \boldsymbol{x})\; p(\boldsymbol{y} \,|\, \boldsymbol{x}, \boldsymbol{a}) \tag{7}$$

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 2 *We say that 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), (6) and (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.

Example application In this paragraph 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. Consistency of partial loss

In this section we present the first main 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 3.

Theorem 3 *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 (1) and let ℓ^p be an associated partial loss defined by (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 4 *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 due to the equality following from (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

$$
h_v^*(\boldsymbol{x}) \in \operatorname*{Argmin}_{t \in \mathcal{T}} \sum_{y_v \in \mathcal{Y}} p(y_v \mid \boldsymbol{x}) \ell_v(y_v, t), \tag{8}
$$

because the marginals $p(z_v \mid \boldsymbol{x}) > 0$ thanks to Definition 2. 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 (8). ■

Note that Lemma 4 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 distributions. 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. The rigorous proof is not trivial. In the rest of the section we give a road map of the proof and defer details to the appendix.

It follows from Lemma 4 that for fixed probability model p induced by model with property A the function $H^p(\epsilon, \boldsymbol{p_{ya}})\colon \mathbb{R} \times \Delta_{|\mathcal{Y}^\mathcal{V}|\times|\mathcal{A}^\mathcal{V}|} \to \mathbb{R}^2$ defined as follows

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

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 $H(\epsilon, \boldsymbol{p_{ya}})\colon \mathbb{R} \times \Delta_{|\mathcal{Y}^\mathcal{V}|\times|\mathcal{A}^\mathcal{V}|} \to \mathbb{R}$ defined as

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

is positive as well for any $\epsilon > 0$. It is possible to show (see Lemmas 12 and 16 in Appendix) even stronger statement that for any $\epsilon > 0$, $H^p(\epsilon) \triangleq \inf_{\boldsymbol{p_{ay}}\in\mathcal{P}_{\boldsymbol{x}}} H^p(\epsilon, \boldsymbol{p_{ay}}) > 0$ and $H(\epsilon) \triangleq \inf_{\boldsymbol{p_{ay}}\in\mathcal{P}_{\boldsymbol{x}}} H(\epsilon, \boldsymbol{p_{ay}}) > 0$. 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 (1), (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

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

See Lemmas 13 and 17 in appendix for complete proof.

Functions ξ and ζ make Theorem 3 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

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

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 and we simply write $\boldsymbol{p_y}, \boldsymbol{p_a}, \boldsymbol{p_{ya}}$.
3. This proof is technically complicated, therefore it is moved to Appendix. There we provide full set of lemmas with complete proves.

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

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

given $m \to \infty$.

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

5. Surrogate partial losses

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. $\boldsymbol{f}_m \in \operatorname{Argmin}_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{m}\psi^p(\boldsymbol{a}^i, \boldsymbol{f}(\boldsymbol{x}^i))$, will converge in probability to the Bayes ψ^p-risk

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

It has been shown (e.g. Zhang (2004a); Tewari and Bartlett (2007); Gao and Zhou (2011); 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 and Theorem 3 from Ramaswamy and Agarwal (2012) to our setting.

Definition 5 *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 $\ell^p\colon \mathcal{A}^{\mathcal{V}} \times \mathcal{T} \to \mathbb{R}_+$ over $\mathcal{P} \subseteq \Delta_{|\mathcal{Y}^{\mathcal{V}}|\times|\mathcal{A}^{\mathcal{V}}|}$ if there exists a function $\operatorname{pred}\colon \hat{\mathcal{T}} \to \mathcal{T}^{\mathcal{V}}$ such that $\forall \boldsymbol{p}_{\boldsymbol{ya}} \in \mathcal{P}$:*

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

Theorem 6 *Let $\ell^p\colon \mathcal{A}^{\mathcal{V}} \times \mathcal{T} \to \mathbb{R}_+$ and $\psi^p\colon \mathcal{A}^{\mathcal{V}} \times \mathcal{T} \to \mathbb{R}_+$. 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 $\operatorname{pred}\colon \hat{\mathcal{T}} \to \mathcal{T}^{\mathcal{V}}$ 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 $\boldsymbol{f}_m\colon \mathcal{X} \to \hat{\mathcal{T}}$,*

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

Combination of Theorem 6 and Theorem 3 directly provides the following corollary:

Corollary 7 *Let $\ell^p\colon \mathcal{A}^{\mathcal{V}} \times \mathcal{T} \to \mathbb{R}_+$ be additively decomposable loss function defined by (1) and $\psi^p\colon \mathcal{A}^{\mathcal{V}} \times \mathcal{T} \to \mathbb{R}_+$. Then ψ^p is classification calibrated with repect to ℓ^p over $\mathcal{P} \subseteq \Delta_{|\mathcal{Y}^{\mathcal{V}}|\times|\mathcal{A}^{\mathcal{V}}|}$ iff there exists a function $\operatorname{pred}\colon \hat{\mathcal{T}} \to \mathcal{T}^{\mathcal{V}}$ 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 $\boldsymbol{f}_m\colon \mathcal{X} \to \hat{\mathcal{T}}$,*

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

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 7 guarantees that ℓ-risk of a decision function $\boldsymbol{h}(\boldsymbol{x}) = \operatorname{pred} \circ \boldsymbol{f}(\boldsymbol{x})$ 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.

5.1. Two examples of surrogate losses classification calibrated w.r.t partial loss

The existing algorithms for learning from missing labels, like Lou and Hamprecht (2012); Fernandes and Brefeld (2011); Li et al. (2013), are based on minimization of the ramp-loss and its mild modifications. To our best knowledge, we are the first to show that the ramp-loss is classification calibrated for learning from missing labels.

Ramp loss. The partial loss ℓ^p can be approximated by the ramp loss

$$\psi^p(\boldsymbol{a}, \hat{\boldsymbol{t}}) = \max_{\boldsymbol{t}\in\mathcal{T}^{\mathcal{V}}} \left(\ell^p(\boldsymbol{a}, \boldsymbol{t}) + \hat{\boldsymbol{t}}_t\right) - \max_{\boldsymbol{t}\in\mathcal{T}^{\mathcal{V}}} \hat{\boldsymbol{t}}_{\boldsymbol{t}}\,, \tag{9}$$

where the surrogate decision set is $\hat{\mathcal{T}} \subseteq \mathbb{R}^{|\mathcal{Y}^{\mathcal{V}}|}$. 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 $\boldsymbol{h}(\boldsymbol{x}) = \operatorname{pred}(\boldsymbol{f}(\boldsymbol{x}))$ via

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

Theorem 8 *Let ℓ^p be a partial loss (2). Then the ramp loss ψ^p constructed from ℓ^p by (9) is classification calibrated with respect to ℓ^p.*

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

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

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

$$\inf_{\hat{\boldsymbol{t}}\in\hat{\mathcal{T}}_{\text{non}}} \boldsymbol{p}_{\boldsymbol{a}}^T \boldsymbol{\psi}^p(\hat{\boldsymbol{t}}) \geq \inf_{\hat{\boldsymbol{t}}\in\hat{\mathcal{T}}_{\text{non}}} \boldsymbol{p}_{\boldsymbol{a}}^T \boldsymbol{\ell}^p_{\operatorname{pred}(\hat{\boldsymbol{t}})} > \min_{\boldsymbol{t}\in\mathcal{T}^{\mathcal{V}}} \boldsymbol{p}_{\boldsymbol{a}}^T \boldsymbol{\ell}^p_{\boldsymbol{t}}\,, \tag{10}$$

where the first inequality follows from the fact that the ramp loss $\psi^p(\boldsymbol{a}, \hat{\boldsymbol{t}})$ upper bounds the partial loss $\ell^p(\boldsymbol{a}, \operatorname{pred}(\boldsymbol{t}))$ for any $\boldsymbol{a} \in \mathcal{A}^{\mathcal{V}}, \hat{\boldsymbol{t}} \in \hat{\mathcal{T}}$ (e.g. Chuong et al. (2008)). Let $\boldsymbol{t}^* \in \operatorname{Argmin}_{\boldsymbol{t}\in\mathcal{T}^{\mathcal{V}}} \boldsymbol{p}_{\boldsymbol{a}}^T \boldsymbol{\ell}^p_{\boldsymbol{t}}$ 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 $\boldsymbol{p}_{\boldsymbol{a}}^T \boldsymbol{\ell}_{\boldsymbol{t}^*}^p = \boldsymbol{p}_{\boldsymbol{a}}^T \psi^p(\hat{\boldsymbol{t}}')$. Therefore we have $\min_{\boldsymbol{t}\in\mathcal{T}^{\mathcal{V}}} \boldsymbol{p}_{\boldsymbol{a}}^T \boldsymbol{\ell}_{\boldsymbol{t}}^p \geq \inf_{\hat{\boldsymbol{t}}\in\hat{\mathcal{T}}} \boldsymbol{p}_{\boldsymbol{a}}^T \boldsymbol{\psi}^p(\hat{\boldsymbol{t}})$ which after combining with (10) gives

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

■

Next, we give another example of classification calibrated surrogate loss which can be constructed from arbitrary single-label calibrated losses.

Additive surrogate loss. Given partial loss ℓ^p composed of label losses ℓ_v, $v \in \mathcal{V}$, according to (2) can be always approximated by the following additive surrogate loss

$$\psi^p(\boldsymbol{a},\hat{\boldsymbol{t}}) = \sum_{v\in\mathcal{V}} [\![|a_v| = 1]\!] \psi_v(a_v,\hat{\boldsymbol{t}}_v) \,, \tag{11}$$

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 $\boldsymbol{h}(\boldsymbol{x}) = \text{pred}(\boldsymbol{f}(\boldsymbol{x}))$ via $\text{pred}(\hat{\boldsymbol{t}}) = (\text{pred}_v(\hat{\boldsymbol{t}}_v) \mid v \in \mathcal{V})$ with $\text{pred}_v(\hat{\boldsymbol{t}}_v) \in \text{Argmax}_{t\in\mathcal{T}} \hat{\boldsymbol{t}}_{v,t}$, where $\hat{\boldsymbol{t}}_{v,t}$ denotes t-th component of the vector $\boldsymbol{t}_v$.

Theorem 9 *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 (11) 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 $v \in \mathcal{V}$:

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

and

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

Then we can write $\inf_{\hat{\boldsymbol{t}}\in\hat{\mathcal{T}}_{\text{non}}} \boldsymbol{p}_{\boldsymbol{a}}^T \boldsymbol{\psi}^p(\hat{\boldsymbol{t}}) = \inf_{\hat{\boldsymbol{t}}\in\hat{\mathcal{T}}_{\text{non}}} \sum_{v\in\mathcal{V}} \boldsymbol{p}_v^T \boldsymbol{\psi}_{\boldsymbol{a},v}^p(\hat{\boldsymbol{t}}_v) = \sum_{v\in\mathcal{V}} \inf_{\hat{\boldsymbol{t}}_v\in\hat{\mathcal{T}}_{\text{non}}^v} \boldsymbol{p}_{\boldsymbol{a},v}^T \boldsymbol{\psi}_v^p(\hat{\boldsymbol{t}}_v) > \sum_{v\in\mathcal{V}} \inf_{\hat{\boldsymbol{t}}_v\in\hat{\mathcal{T}}^v} \boldsymbol{p}_{\boldsymbol{a},v}^T \boldsymbol{\psi}_v^p(\hat{\boldsymbol{t}}_v) = \inf_{\hat{\boldsymbol{t}}\in\hat{\mathcal{T}}} \sum_{v\in\mathcal{V}} \boldsymbol{p}_{\boldsymbol{a},v}^T \boldsymbol{\psi}_v^p(\hat{\boldsymbol{t}}_v) = \inf_{\hat{\boldsymbol{t}}\in\hat{\mathcal{T}}} \boldsymbol{p}_{\boldsymbol{a}}^T \boldsymbol{\psi}^p(\hat{\boldsymbol{t}})$, where strict inequality follows from the fact that for every $v \in \mathcal{V}$ the inequality $\inf_{\hat{\boldsymbol{t}}_v\in\hat{\mathcal{T}}_{\text{non}}^v} \boldsymbol{p}_{\boldsymbol{a},v}^T \boldsymbol{\psi}_v^p(\hat{\boldsymbol{t}}_v) > \inf_{\hat{\boldsymbol{t}}_v\in\hat{\mathcal{T}}^v} \boldsymbol{p}_{\boldsymbol{a},v}^T \boldsymbol{\psi}_v^p(\hat{\boldsymbol{t}}_v)$ holds.

■

Theorem 9 shows that the additive surrogate loss (11) 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.

In the case of binary labels, $\mathcal{Y} = \{-1, +1\}$, a classification calibrated surrogate partial loss ψ^p can be obtained by using the hinge-loss known to be calibrated Zhang (2004a), i.e. $\psi_v(y,\hat{t}) = \max\{0, 1 - y\hat{t}\}$, $v \in \mathcal{V}$. This surrogate partial loss has been proposed in Yu et al. (2014) for learning from missing labels. To our best knowledge the additive classification calibrated losses (11) constructed for the case $|\mathcal{Y}| > 2$ has not been proposed so far.

6. Conclusions

We have defined conditions on the data generating model of partial annotations and proved that under these conditions minimization of the partial loss provides structured predictor whose expected risk converges in probability to the Bayes risk defined by an associated complete loss. Using the general framework of Ramaswamy and Agarwal (2012) we have defined the concept of a surrogate classification calibrated partial loss whose minimization preserves the statistical consistency and at the same time can be easier for optimization. We gave concrete examples of surrogate classification calibrated partial losses. Namely, we showed that the ramp-loss and the additive loss composed of any set of classification calibrated single label losses are both calibrated w.r.t the partial loss. Hence, the algorithms based on their minimization (e.g. proposed in Lou and Hamprecht (2012); Fernandes and Brefeld (2011); Li et al. (2013); Yu et al. (2014)) are statistically consistent.

Acknowledgments

The authors were supported by the Grant Agency of the Czech Republic under Project P202/12/2071, the project ERC-CZ LL1303 and EU project FP7-ICT-609763 TRADR.

Appendix A. Proofs

In this section we give detailed proofs mentioned in Section 4. We start with showing positiveness of functions $H^p(\epsilon)$ and $H(\epsilon)$. To show this first we need to show that set of all conditional distributions $p(\boldsymbol{y}, \boldsymbol{a} \mid \boldsymbol{x})$ is a compact set.

Lemma 10 *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) and (6) we see that for any $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{p}_{\boldsymbol{ya}}(\boldsymbol{x})$ is a composition of functions with vector variables $\boldsymbol{p}_{\boldsymbol{y}}(\boldsymbol{x})$ and $\boldsymbol{p}_{\boldsymbol{a}}(\boldsymbol{x})$, i.e. $\boldsymbol{p}_{\boldsymbol{ya}}(\boldsymbol{x}) = \mathcal{F}(\boldsymbol{p}_{\boldsymbol{y}}(\boldsymbol{x}), \boldsymbol{p}_{\boldsymbol{z}}(\boldsymbol{x}))$. 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 $\{\boldsymbol{p}_{\boldsymbol{y}}(\boldsymbol{x}) \in \Delta_{|\mathcal{Y}^{\mathcal{V}}|} \mid p(\boldsymbol{y} \mid \boldsymbol{x}) \geq \rho\} \times \{\boldsymbol{p}_{\boldsymbol{z}}(\boldsymbol{x}) \in \Delta_{|\mathcal{Z}^{\mathcal{V}}|} \mid p(\boldsymbol{z} \mid \boldsymbol{x}) \geq \rho\}$. Thus, $\mathcal{P}_{\boldsymbol{x}} \triangleq \{\boldsymbol{p}_{\boldsymbol{ya}}(\boldsymbol{x}) = \mathcal{F}(\boldsymbol{p}_{\boldsymbol{y}}(\boldsymbol{x}), \boldsymbol{p}_{\boldsymbol{z}}(\boldsymbol{x})) \mid p(\boldsymbol{y} \mid \boldsymbol{x}) \geq \rho, p(\boldsymbol{z} \mid \boldsymbol{x}) \geq \rho, \boldsymbol{p}_{\boldsymbol{y}}(\boldsymbol{x}) \in \Delta_{|\mathcal{Y}^{\mathcal{V}}|}, \boldsymbol{p}_{\boldsymbol{z}}(\boldsymbol{x}) \in \Delta_{|\mathcal{Z}^{\mathcal{V}}|}\}$ is a compact set. ■

Lemma 11 *Functions $\min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p}_{\boldsymbol{a}}^T \boldsymbol{\ell}_{\boldsymbol{t}'}^p$ and $\min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p}_{\boldsymbol{y}}^T \boldsymbol{\ell}_{\boldsymbol{t}'}$ are continuous functions w.r.t. $\boldsymbol{p}_{\boldsymbol{ya}} \in \Delta_{|\mathcal{Y}^{\mathcal{V}}|\times|\mathcal{A}^{\mathcal{V}}|}$.*

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 $\boldsymbol{p}_{\boldsymbol{y}}$ and $\boldsymbol{p}_{\boldsymbol{a}}$ are continuous functions of $\boldsymbol{p}_{\boldsymbol{ya}}$. Hence, both functions $\min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p}_{\boldsymbol{a}}^T \boldsymbol{\ell}_{\boldsymbol{t}'}^p$ and $\min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p}_{\boldsymbol{y}}^T \boldsymbol{\ell}_{\boldsymbol{t}'}$ 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 positives of function $H^p(\epsilon)$ for any positive ϵ.

Lemma 12 *Let $H^p(\epsilon, \boldsymbol{p}_{\boldsymbol{ya}})\colon \mathbb{R} \times \Delta_{|\mathcal{Y}^{\mathcal{V}}|\times|\mathcal{A}^{\mathcal{V}}|} \to \mathbb{R}$ be a function defined as follows*

$$\begin{aligned} \textit{minimize} \quad & \boldsymbol{p}_{\boldsymbol{a}}{}^T \boldsymbol{\ell}_{\boldsymbol{t}}^p - \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p}_{\boldsymbol{a}}{}^T \boldsymbol{\ell}_{\boldsymbol{t}'}^p \\ \textit{subject to} \quad & \boldsymbol{p}_{\boldsymbol{y}}{}^T \boldsymbol{\ell}_{\boldsymbol{t}} - \min_{\boldsymbol{t}' \in \mathcal{T}^{\mathcal{V}}} \boldsymbol{p}_{\boldsymbol{y}}{}^T \boldsymbol{\ell}_{\boldsymbol{t}'} \geq \epsilon. \end{aligned}$$

where loss functions $\ell(\boldsymbol{y},\boldsymbol{t})$ and $\ell^p(\boldsymbol{a},\boldsymbol{t})$ are defined by (1), (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 $\forall \boldsymbol{p}_{\boldsymbol{ya}}\in\mathcal{P}$ holds $H^p(\epsilon,\boldsymbol{p}_{\boldsymbol{ya}})>\delta$, i.e. $H^p(\epsilon)=\inf_{\boldsymbol{p}_{\boldsymbol{ya}}\in\mathcal{P}} H^p(\epsilon,\boldsymbol{p}_{\boldsymbol{ya}})>\delta$.

PROOF: We prove the lemma by contradiction. Assume that (16) 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 (11) 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 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. Then 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 contradiction, i.e. we have found a model $\boldsymbol{p}^*\in\mathcal{P}$ for which lemma (4) does not hold. ■

Lemma 13 *If $\forall\epsilon>0, H^p(\epsilon)\triangleq\inf\limits_{\boldsymbol{p}_{\boldsymbol{ay}}\in\mathcal{P}_{\boldsymbol{x}}}H^p(\epsilon,\boldsymbol{p}_{\boldsymbol{ay}})>0$ for the loss functions $\ell(\boldsymbol{y},\boldsymbol{t})$ and $\ell^p(\boldsymbol{a},\boldsymbol{t})$ defined by (1), (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*

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

Main idea of proof of Lemma 13 is analogical to the proof of Corollary 26 in Zhang (2004b). Thus, we provide proof only for Lemma 13 together with two auxiliary lemmas needed for its proof and proof of "flipped" version of this lemma we leave for the reader.

Lemma 14 *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*

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

PROOF: Using Jensen's inequality together with inequality $H^p(\boldsymbol{p}^T_{\boldsymbol{y}}\boldsymbol{\ell}_{\boldsymbol{t}}-\min\limits_{\boldsymbol{t}'\in\mathcal{T}^{\mathcal{V}}}\boldsymbol{p}^T_{\boldsymbol{y}}\boldsymbol{\ell}_{\boldsymbol{t}'})\leq\boldsymbol{p}^T_{\boldsymbol{a}}\boldsymbol{\ell}^p_{\boldsymbol{t}}-\min\limits_{\boldsymbol{t}'\in\mathcal{T}^{\mathcal{V}}}\boldsymbol{p}^T_{\boldsymbol{a}}\boldsymbol{\ell}^p_{\boldsymbol{t}'}$ we have $\mu(\mathbb{E}_{p(\boldsymbol{x})}\boldsymbol{p}_{\boldsymbol{y}}(\boldsymbol{x})^T\boldsymbol{\ell}_{\boldsymbol{h}(\boldsymbol{x})}-\mathbb{E}_{p(\boldsymbol{x})}\min\limits_{\boldsymbol{t}'\in\mathcal{T}^{\mathcal{V}}}\boldsymbol{p}_{\boldsymbol{y}}(\boldsymbol{x})^T\boldsymbol{\ell}_{\boldsymbol{t}'})\leq\mathbb{E}_{p(\boldsymbol{x})}\mu(\boldsymbol{p}_{\boldsymbol{y}}(\boldsymbol{x})^T\boldsymbol{\ell}_{\boldsymbol{h}(\boldsymbol{x})}-\min\limits_{\boldsymbol{t}'\in\mathcal{T}^{\mathcal{V}}}\boldsymbol{p}_{\boldsymbol{y}}(\boldsymbol{x})^T\boldsymbol{\ell}_{\boldsymbol{t}'})\leq\mathbb{E}_{p(\boldsymbol{x})}H^p(\boldsymbol{p}_{\boldsymbol{y}}(\boldsymbol{x})^T\boldsymbol{\ell}_{\boldsymbol{h}(\boldsymbol{x})}-\min\limits_{\boldsymbol{t}'\in\mathcal{T}^{\mathcal{V}}}\boldsymbol{p}_{\boldsymbol{y}}(\boldsymbol{x})^T\boldsymbol{\ell}_{\boldsymbol{t}'})\leq\mathbb{E}_{p(\boldsymbol{x})}(\boldsymbol{p}_{\boldsymbol{a}}(\boldsymbol{x})^T\boldsymbol{\ell}^p_{\boldsymbol{h}(\boldsymbol{x})}-\min\limits_{\boldsymbol{t}'\in\mathcal{T}^{\mathcal{V}}}\boldsymbol{p}_{\boldsymbol{a}}(\boldsymbol{x})^T\boldsymbol{\ell}^p_{\boldsymbol{t}'})$. ■

Lemma 15 *Let $\zeta_*(\epsilon)=\sup\limits_{a\geq0,b}\{a\epsilon+b\mid\forall z\geq0, az+b\leq H^p(z)\}$, 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 15 is a proposition 25 from Zhang (2004b) for the function $H^p(\epsilon)$, thus we omit its proof here. Now we are ready to prove Lemma 13.

PROOF: Consider $\zeta_*(\epsilon)$ in Lemma 15, 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 14.

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 pick up any $\epsilon' > 0$, and let $\delta' = \frac{\zeta_*(\epsilon')}{2}$, we know from Lemma 15 that $\delta' > 0$. This implies that $\xi(\delta) < \epsilon'$ when $\delta' < \delta$. ■

Here we just give formulation of "flipped" version of Lemma 13 and its auxiliary Lemma 16. To prove Lemma 17 we need modified Lemma 14 and 15 for the function from Lemma 16 which is straightforward to do, thus we leave it for the reader.

Lemma 16 *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*

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

where loss functions $\ell(\boldsymbol{y}, \boldsymbol{t})$ *and* $\ell^p(\boldsymbol{a}, \boldsymbol{t})$ *are defined by (1), (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 12. ■

Lemma 17 *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 (1), (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*

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

Proof of Lemma 17 is similar to proof of Lemma 13.

References

B. Do Chuong, Le Quoc, Choon Hui Teo, Olivier Chapelle, and Alex Smola. Tighter bounds for structured estimation. In *Proc. of Neural Processing Information Systems*, pages 281–288, 2008.

Eraldo R. Fernandes and Ulf Brefeld. Learning from partially annotated sequences. In *Proc. of European Conference on Machine Learning*, pages 407–422, 2011.

Wei Gao and Zhi-Hua Zhou. On the consistency of multi-label learning. *Journal of Machine Learning Research*, 2011.

Chengtao Li, Jianwen Zhang, and Zheng Chen. Structured output learning with candidate labels for local parts. In *Proc. of European Conference on Machine Learning*, 2013.

Xinghua Lou and Fred A. Hamprecht. Structured learning from partial annotations. In *Proc. of International Conference on Machine Learning*, 2012.

David A. McAllester and Joseph Keshet. Generalization bounds and consistency for latent structural probit and ramp loss. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *NIPS*, pages 2205–2212, 2011.

Harish G. Ramaswamy and Shivani Agarwal. Classification calibration dimension for general multiclass losses. In *Proc. of Neural Information Processing Systems*, pages 2087–2095, 2012.

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

Michal Uřičář, Vojtěch Franc, and Václav Hlaváč. Detector of facial landmarks learned by the structured output SVM. In Gabriela Csurka and José Braz, editors, *VISAPP '12: Proceedings of the 7th International Conference on Computer Vision Theory and Applications*, volume 1, pages 547–556, Portugal, February 2012. SciTePress — Science and Technology Publications. ISBN 978-989-8565-03-7.

Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998. ISBN 978-0-471-03003-4.

Andrea Vedaldi and Andrew Zisserman. Structured output regression for detection with partial truncation. In *Proc. of Neural Information Processing Systems*, 2009.

Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit S. Dhillon. Large-scale multi-label learning with missing labels. In *Proc. of International Conference on Machine Learning*, 2014.

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

Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004b.