On the inconsistency of separable losses for structured prediction

Caio Corro

On the inconsistency of separable losses for structured prediction

Caio Corro
Universite Paris-Saclay, CNRS, LISN, 91400, Orsay, France
caio.corro@lisn.upsaclay.fr

Abstract

In this paper, we prove that separable negative log-likelihood losses for structured prediction are not necessarily Bayes consistent, or, in other words, minimizing these losses may not result in a model that predicts the most probable structure in the data distribution for a given input. This fact opens the question of whether these losses are well-adapted for structured prediction and, if so, why.

1 Introduction

Modern natural language processing (NLP) heavily relies on machine learning (ML), where prediction models are learned by minimizing a loss function over the training data. As such, loss functions play a central role in the design of these systems and it is important to understand their statistical properties in order to guarantee that the corresponding training objectives are well defined. Although this topic is well studied in the ML community (Lugosi and Vayatis, 2004; Lin, 2004; Zhang, 2004a,b; Bartlett et al., 2006; Gneiting and Raftery, 2007; Liu, 2007; Tewari and Bartlett, 2007; Reid and Williamson, 2010; Williamson et al., 2016; Duchi et al., 2018; Blondel et al., 2020; Nowak et al., 2022), *inter alia*, there has been less focus on the structured prediction setting apart from a few recent works (Blondel, 2019; Nowak et al., 2019, 2020).

In this paper, we emphasize the fact that, despite achievements in terms of accuracy, statistical behavior of loss functions used in practice for structured prediction in NLP are not always well understood. We illustrate this fact by proving that commonly used separable loss functions for named entity recognition (NER) and dependency parsing are not Bayes consistent, meaning that training a model with these loss functions will not necessarily result in the prediction of the most the probable output for a given input in the data distribution.

2 Bayes consistency

We denote inputs and outputs as $\boldsymbol{x} \in X$ and $\boldsymbol{y} \in Y$, respectively. We assume each $\boldsymbol{y} \in Y$ is a binary vector whose elements are indexed by a set C, *i.e.* $\boldsymbol{y} \in \{0,1\}^C$, where C is problem dependent. For example, in the k multiclass classification case, we have $C = [k]$, where we use the shorthand $[k] = \{1, 2, ..., k\}$, and Y is defined as the set of standard bases (one-hot vectors) of dimension k, meaning that $|Y| = k$. More generally, the vector $\boldsymbol{y}$ is an indicator of "selected" parts in C and, in the structured prediction case, several parts can be jointly selected. Note that it is usual to assume that the parts in C can depend on the input $\boldsymbol{x}$. Without loss of generality, we omit this detail as we will study loss functions in the pointwise setting.

A scoring model $f \in F$ is a function $f : X \to \mathbb{R}^C$ that returns scores associated with each part in C for a given input, *e.g.* the score of each class in a multiclass classification model. The actual prediction of the model is the output of maximum linear score:

$$\widehat{\boldsymbol{y}}(\boldsymbol{x}) \in \arg\max_{\boldsymbol{y} \in Y} \langle \boldsymbol{y}, f(\boldsymbol{x}) \rangle, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. We refer to computing Equation 1 as maximum *a posteriori* (MAP) inference.

A loss function compares a vector of scores with an expected output. Importantly, the 0-1 loss function is defined as follows:

$$\ell(\boldsymbol{w}, \boldsymbol{y}) = \begin{cases} 0 & \text{if } \boldsymbol{y} \in \arg\max_{\boldsymbol{y}' \in Y} \langle \boldsymbol{w}, \boldsymbol{y}' \rangle, \\ 1 & \text{otherwise,} \end{cases}$$

where $\boldsymbol{w} \in \mathbb{R}^k$ is a vector of part scores, *i.e.* $\boldsymbol{w} = f(\boldsymbol{x})$ for a given input $\boldsymbol{x}$. In order to choose a scoring function $f \in F$, it is appealing to select one that minimizes this loss over the data distribution:

$$r^* = \inf_{f \in F} r(f) = \inf_{f \in F} \mathbb{E}_{\mathbf{x},\mathbf{y}}[\ell(f(\mathbf{x}), \mathbf{y})],$$

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1491–1498

where $\mathbf{x}$ and $\mathbf{y}$ are random variables over inputs and outputs, respectively, and $\inf$ denotes the infimum. The value $r(f)$ is the Bayes risk of function f and r^* is the optimal Bayes risk. For theoretical purposes, it is often assumed that the class of functions F is rich enough (the set of all measurable mappings) to obtain the best possible risk. Then, the optimal Bayes risk is equal to:

$$r^* = \mathbb{E}_{\mathbf{x}}[1 - \max_{\boldsymbol{y} \in Y} p(\mathbf{y} = \boldsymbol{y}|\mathbf{x})],$$

or, in other words, it is the probability of making an error when the classifier predicts the most probable class for each input.

Unfortunately, in practice it is not convenient to use the 0-1 loss ℓ as it is nonconvex and has null derivatives almost everywhere. Instead, a surrogate $\widetilde{\ell}$ can be used as a loss function:

$$\widetilde{r}^* = \inf_{f \in F} \widetilde{r}(f) = \inf_{f \in F} \mathbb{E}_{\mathbf{x},\mathbf{y}}[\widetilde{\ell}(f(\mathbf{x}), \mathbf{y})],$$

where $\widetilde{r}(f)$ is the surrogate risk of function f and $\widetilde{r}^*$ is the optimal surrogate risk. An important desired property of surrogate losses is their consistency with the 0-1 loss, *i.e.* the fact that minimizing the surrrogate risk leads to a prediction model of optimal Bayes risk (Lugosi and Vayatis, 2004; Lin, 2004; Zhang, 2004a; Bartlett et al., 2006; Liu, 2007; Tewari and Bartlett, 2007).

Definition 1. *A surrogate loss $\widetilde{\ell}$ is said to be Bayes consistent*[1] *if:*

$$f^* \in \arg\min_{f \in F} \widetilde{r}(f) \implies r(f^*) = r^*.$$

Note that this property can be checked independently for each input $\boldsymbol{x}$ (called pointwise Bayes consistency) as we assume a rich enough class of functions F. In other words, we redefine the pointwise (optimal) surrogate risk as:

$$\widetilde{r}^* = \inf_{\boldsymbol{w} \in \mathbb{R}^C} \widetilde{r}(\boldsymbol{w}) = \inf_{\boldsymbol{w} \in \mathbb{R}^C} \mathbb{E}_{\mathbf{y}|\mathbf{x}=\boldsymbol{x}}[\widetilde{\ell}(\boldsymbol{w}, \mathbf{y})],$$

for any $\boldsymbol{x}$ such that $p(\mathbf{x} = \boldsymbol{x}) > 0$, and similarly for the optimal Bayes risk. The vector $\boldsymbol{w}$ should be interpreted as the model scores, *i.e.* $\boldsymbol{w} = f(\boldsymbol{x})$.

[1]This property is also referred to as Fisher consistency (Lin, 2004; Bartlett et al., 2006; Liu, 2007) and classification calibration (Williamson et al., 2016).

3 Negative log-likelihood loss

The negative log-likelihood loss (NLL), also known as the conditional random field loss (Lafferty et al., 2001), is defined as follows:

$$\widetilde{\ell}_{(nll)}(\boldsymbol{w}, \boldsymbol{y}) = -\langle \boldsymbol{w}, \boldsymbol{y} \rangle + \log \sum_{\boldsymbol{y}' \in Y} \exp \langle \boldsymbol{w}, \boldsymbol{y}' \rangle.$$

In the following, we will refer to computing the log-sum-exp term of the NLL loss as marginal inference due to its connection with marginal probabilities (Wainwright and Jordan, 2008).

Theorem 1. *Under mild conditions on the data distribution, the surrogate loss $\widetilde{\ell}_{(nll)}$ is Bayes consistent.*

Proof. The optimal pointwise surrogate Bayes risk is defined as:

$$\widetilde{r}^*_{(nll)} = \inf_{\boldsymbol{w} \in \mathbb{R}^C} \begin{array}{l} -\mathbb{E}_{\mathbf{y}|\mathbf{x}=\boldsymbol{x}}[\langle \boldsymbol{w}, \mathbf{y} \rangle] \\ +\log \sum_{\boldsymbol{y}' \in Y} \exp \langle \boldsymbol{w}, \boldsymbol{y}' \rangle. \end{array}$$

We substitute $\boldsymbol{w}(\boldsymbol{y}) = \langle \boldsymbol{w}, \boldsymbol{y} \rangle$ for all $\boldsymbol{y} \in Y$:

$$= \inf_{\substack{\forall \boldsymbol{y} \in Y: \\ \boldsymbol{w}(\boldsymbol{y}) \in \mathbb{R}}} \begin{array}{l} -\mathbb{E}_{\mathbf{y}|\mathbf{x}=\boldsymbol{x}}[\boldsymbol{w}(\mathbf{y})] \\ +\log \sum_{\boldsymbol{y}' \in Y} \exp \boldsymbol{w}(\boldsymbol{y}'). \end{array}$$

We denote $\widehat{\boldsymbol{w}}(\boldsymbol{y}), \forall \boldsymbol{y} \in Y$, an optimal solution of the minimization. By first order optimality conditions, we have:

$$\begin{aligned} &\frac{\partial}{\partial \widehat{\boldsymbol{w}}(\boldsymbol{y})} \begin{pmatrix} -\mathbb{E}_{\mathbf{y}|\mathbf{x}=\boldsymbol{x}}[\widehat{\boldsymbol{w}}(\mathbf{y})] \\ +\log \sum_{\boldsymbol{y}' \in Y} \exp(\widehat{\boldsymbol{w}}(\boldsymbol{y}')) \end{pmatrix} = 0 \\ \implies &\frac{\exp \widehat{\boldsymbol{w}}(\boldsymbol{y})}{\sum_{\boldsymbol{y}' \in Y} \exp \widehat{\boldsymbol{w}}(\boldsymbol{y}')} = p(\mathbf{y} = \boldsymbol{y}|\mathbf{x} = \boldsymbol{x}) \end{aligned} \tag{2}$$

which implies Bayes consistency under the condition that there exists a vector $\boldsymbol{w} \in \mathbb{R}^C$ such that $\forall \boldsymbol{y} \in Y : \langle \boldsymbol{w}, \boldsymbol{y} \rangle = \widehat{\boldsymbol{w}}(\boldsymbol{y})$. □

To understand why Equation 2 implies Bayes consistency, note that:

$$\frac{\exp \widehat{\boldsymbol{w}}(\boldsymbol{y})}{\sum_{\boldsymbol{y}' \in Y} \exp \widehat{\boldsymbol{w}}(\boldsymbol{y}')} \propto \exp \widehat{\boldsymbol{w}}(\boldsymbol{y}'),$$

and the exponential function is strictly increasing. This means scores of outputs $\boldsymbol{y} \in Y$ defined as $\langle \boldsymbol{y}, \widehat{\boldsymbol{w}} \rangle$ are ordered in the same way as probabilities in the data distribution $p(\mathbf{y}|\mathbf{x} = \boldsymbol{x})$. In other words, the most probable output in the data distribution will have the highest score with respect to $\widehat{\boldsymbol{w}}$.

The proof is a straightforward extension of the derivation for the multiclass classification case, see for example Blondel et al. (2020, Section 4.2). Interestingly, Equation 2 also implies that the NLL loss is strictly proper (Williamson et al., 2016), *i.e.* the Boltzmann distribution over structures in Y parameterized by minimizer $\widehat{\boldsymbol{w}}$ is equal to the data distribution $p(\mathbf{y}|\mathbf{x} = \boldsymbol{x})$. Note that Theorem 1 is not novel *per se* and a more in-depth study of NLL losses for structured prediction can be found in (Nowak et al., 2019).

One limitation of the NLL loss is that it is not (additively) separable[2] because of the log-sum-exp term. As such, this term is a bottleneck for parallel computation of the objective and doubly stochastic estimation of the training objective (Titsias, 2016). A well-known solution is to rely on independent binary classification objectives, also known as one-vs-all losses (Blondel et al., 2020, Section 6.1):

$$\widetilde{\ell}_{(one-vs-all)}(\boldsymbol{w}, \boldsymbol{y}) = -\langle \boldsymbol{w}, \boldsymbol{y} \rangle + \sum_{\boldsymbol{y}' \in Y} \log\left(1 + \exp\langle \boldsymbol{w}, \boldsymbol{y}' \rangle\right).$$

In the case of multiclass classification problems, it can be shown that this objective is Bayes consistent using similar arguments as in Theorem 1. A different approach is the one-vs-each loss function that is also Bayes consistent (Titsias, 2016).

Unfortunately, these separable surrogates cannot be applied to structured prediction problems as the set Y is often of exponential size with respect to the input length. Although tractable algorithms for marginal inference exist for many cases, there are no known algorithms to compute the one-vs-each or one-vs-all losses in an easily parallelizable fashion. As such, the NLP community often relies on token-separable losses, that is a NLL objective that decomposes as a sum of independent losses, one per token in the input sentence. Although these losses are easy to implement, we prove in the next sections that they are not Bayes consistent for two common NLP problems.

4 Named-entity recognition

Problem definition. In this Section, we focus on the flat NER problem using BIO tags (Ratinov and Roth, 2009). Without loss of generality, we assume there is a single mention label and that the input sentence $\boldsymbol{x}$ contains n words. The set of parts is defined as $C = [n] \times \{\mathrm{B}, \mathrm{I}, \mathrm{O}\}$ and Y is defined as the set of vectors $\boldsymbol{y} \in \{0, 1\}^C$ satisfying the following conditions:

1. $\forall i \in [n] : \sum_t y_{i,t} = 1$ (one tag per word);
2. $y_{1,\mathrm{I}} = 0$ (forbid inside tag for the first word);
3. $\forall i > 1 : y_{i,\mathrm{I}} = 1 \implies y_{i-1,\mathrm{B}} + y_{i-1,\mathrm{I}} = 1$ (I tag can only follow a B or I tag).

We do not include parts corresponding to transitions (this is a unigram model), otherwise it would not be possible to derive a token-separable loss.

Inference algorithms. MAP and marginal inference can be realized using the Viterbi and the forward-backward algorithms, respectively. Although the time complexity of these algorithms is $\mathcal{O}(|L|^2 n)$ where L is the set of mention labels, they can be optimized to have a $\mathcal{O}(|L|n)$ time complexity as there is no transition score. The dynamic programming algorithm is nonetheless required in order to guarantee that condition (3) is satisfied.

Separable loss. As the dynamic programming algorithm is not parallelizable over input tokens, token-separable losses are often used in practice.[3] That is, the loss is reduced to a set of n multiclass classification losses:

$$\widetilde{\ell}_{(\text{sep-bio})} = -\langle \boldsymbol{w}, \boldsymbol{y} \rangle + \sum_{i=1}^{n} \log \sum_t \exp w_{i,t},$$

where t ranges over all tags, except I if $i = 1$.

The optimal pointwise surrogate Bayes risk for the separable loss is defined as:

$$\widetilde{r}^*_{(sep-bio)} = \inf_{\boldsymbol{w} \in \mathbb{R}^C} \begin{array}{l} -\mathbb{E}_{\mathbf{y}|\mathbf{x}=\boldsymbol{x}}[\langle \boldsymbol{w}, \mathbf{y} \rangle] \\ +\sum_{i=1}^{n} \log \sum_t \exp w_{i,t}. \end{array}$$

Let $\widehat{\boldsymbol{w}}$ be an optimal solution. Then, by first order optimality conditions:

$$\frac{\partial}{\partial \widehat{w}_{i,t}} \begin{pmatrix} -\mathbb{E}_{\mathbf{y}|\mathbf{x}=\boldsymbol{x}}[\langle \boldsymbol{w}, \mathbf{y} \rangle] \\ +\sum_{i=1}^{n} \log \sum_t \exp w_{i,t} \end{pmatrix} = 0$$
$$\implies \widehat{w}_{i,t} = \log p\left(\mathrm{y}_{i,t} = 1 | \mathbf{x} = \boldsymbol{x}\right) \quad (3)$$

where $p\left(\mathrm{y}_{i,t} = 1 | \mathbf{x} = \boldsymbol{x}\right)$ denotes the marginal distribution of tag t at position i in data distribution.

Theorem 2. *The token-separable loss for NER via BIO tagging is not Bayes consistent.*

[2] A function f is additively separable if it can be written as $f = \sum_i f_i$.

[3] See for example https://github.com/huggingface/transformers/blob/v4.23.1/src/transformers/models/bert/modeling_bert.py#L1771

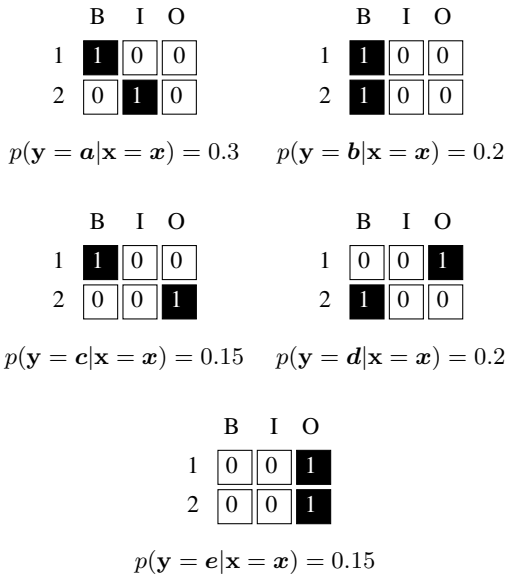

Figure 1: Example of distribution over BIO sequences for a sentence of 2 words. The set of sequences is defined as $Y = \{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}, \boldsymbol{e}\}$. The matrices represent values in elements of Y.

Proof. Let $n = 2$ and assume that the distribution $p(\mathbf{y}|\mathbf{x} = \boldsymbol{x})$ is defined as depicted in Figure 1. Then, by Equation 3 we have $\widehat{w}_{1,\mathrm{B}} = \log 0.65$, $\widehat{w}_{1,\mathrm{O}} = \log 0.35$, $\widehat{w}_{2,\mathrm{B}} = \log 0.4$, $\widehat{w}_{2,\mathrm{I}} = \log 0.3$ and $\widehat{w}_{2,\mathrm{O}} = \log 0.3$. As such:

$$\log 0.65 + \log 0.3 = \langle \widehat{\boldsymbol{w}}, \boldsymbol{a} \rangle < \langle \widehat{\boldsymbol{w}}, \boldsymbol{b} \rangle = \log 0.65 + \log 0.4,$$

but $\boldsymbol{a} \in \arg\max_{\boldsymbol{y} \in Y} p(\mathbf{y} = \boldsymbol{y}|\mathbf{x} = \boldsymbol{x})$ and $p(\mathbf{y} = \boldsymbol{a}|\mathbf{x} = \boldsymbol{x}) > p(\mathbf{y} = \boldsymbol{b}|\mathbf{x} = \boldsymbol{x})$. Therefore, the token-separable loss is not Bayes consistent for NER, *i.e.* a scoring model minimizing the surrogate risk may not lead to predicting the most probable output in the data distribution. □

Note that the inconsistency is not due to the fact that the parameterization of the model is "poor" (no transition scores). Indeed, by Equation 2, optimal scores $\widehat{\boldsymbol{w}}$ for the NLL loss satisfy the following condition for all $\boldsymbol{y} \in Y$:

$$\frac{\exp\langle \widehat{\boldsymbol{w}}, \boldsymbol{y} \rangle}{\sum_{\boldsymbol{y}' \in Y} \exp\langle \widehat{\boldsymbol{w}}, \boldsymbol{y}' \rangle} = p(\mathbf{y} = \boldsymbol{y}|\mathbf{x} = \boldsymbol{x})$$
$$\implies \langle \widehat{\boldsymbol{w}}, \boldsymbol{y} \rangle = \log p(\mathbf{y} = \boldsymbol{y}|\mathbf{x} = \boldsymbol{x})$$

The following assignment for $\widehat{\boldsymbol{w}}$ satisfies this condition: $\widehat{w}_{1,\mathrm{B}} = 0$, $\widehat{w}_{1,\mathrm{O}} = 0$, $\widehat{w}_{2,\mathrm{B}} = \log 0.2$, $\widehat{w}_{2,\mathrm{I}} = \log 0.3$, and $\widehat{w}_{2,\mathrm{O}} = \log 0.15$. That is, minimizing the NLL loss on this distribution results in a Bayes consistent classifier, as expected.

5 Syntactic dependency parsing

Problem definition. We consider a sentence of n words and, without loss of generality, restrict to the unlabeled case to simplify notations. In dependency parsing, the set of parts is defined as the set of possible bilexical dependencies between words, including a fake root at position 0 used to identify root word(s) of the sentence, *i.e.* $C = \{(h, m) \in \{0, 1, ..., n\} \times [n] | h \neq m\}$, where (h, m) denotes a dependency with the h-th word as head and the m-th word as modifier. The set Y is restricted to vectors $\boldsymbol{y} \in \{0, 1\}^C$ that can be interpreted as forming a 0-rooted spanning arborescence where words are vertices and dependencies are arcs (McDonald et al., 2005). In some cases, *e.g.* the Universal Dependency format, it is required that the fake root position has a single outgoing arc.[4]

Inference algorithms. MAP inference can be realized via the maximum spanning arborescence algorithm, which has a $\mathcal{O}(n^2)$ time complexity (Chu and Liu, 1965; Edmonds, 1967; Tarjan, 1977). The single root constraint can be taken into account using the same algorithm via the big-M trick (Fischetti and Toth, 1992, Section 2). Marginal inference can be realized via the matrix tree theorem (MTT, Koo et al., 2007; McDonald and Satta, 2007; Smith and Smith, 2007), which has $\mathcal{O}(n^3)$ time complexity.

Separable loss. The cubic-time complexity of MTT may be prohibitive in practice for training a model. Moreover, the MTT relies on a computationally unstable matrix inversion and is arguably non-trivial to implement. Hence, there has been interest in using simpler token-separable NLL loss functions (Zhang et al., 2017):

$$\widetilde{\ell}_{(sep-dep)}(\boldsymbol{w}, \boldsymbol{y}) = -\langle \boldsymbol{w}, \boldsymbol{y} \rangle + \sum_{m \in [n]} \log \sum_{h \in [n] \setminus \{m\}} \exp w_{h,m},$$

also called head selection loss. This loss is a sum of multiclass classification NLL losses, one per word in the sentence, and is therefore token-separable. As such, it can be efficiently parallelized on GPU and is trivial to implement in any ML framework.

The optimal pointwise surrogate Bayes risk for the token-separable loss is defined as:

$$\widetilde{r}^*_{(sep-dep)} = \inf_{\boldsymbol{w} \in \mathbb{R}^A} -\mathbb{E}_{\mathbf{y}|\mathbf{x}=\boldsymbol{x}}[\langle \boldsymbol{w}, \mathbf{y} \rangle] + \sum_{m \in [n]} \log \sum_{h \in [n] \setminus \{m\}} \exp w_{h,m}.$$

[4] https://universaldependencies.org/u/overview/syntax.html

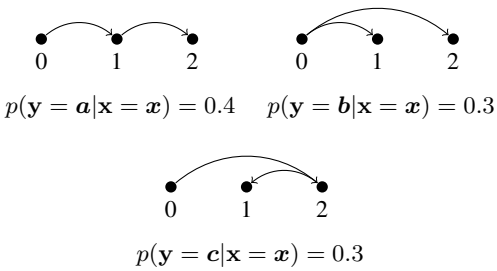

Figure 2: Example of distribution over trees for a sentence of 3 words. The set of trees is defined as $Y = \{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\}$.

Let $\widehat{\boldsymbol{w}}$ be an optimal solution. Then, by first order optimality conditions:

$$\frac{\partial}{\partial \widehat{w}_{h,m}} \left(\begin{array}{l} -\mathbb{E}_{\mathbf{y}|\mathbf{x}=\boldsymbol{x}}[\langle \widehat{\boldsymbol{w}}, \mathbf{y} \rangle] \\ + \sum\limits_{m \in [n]} \log \sum\limits_{h \in [n] \setminus \{m\}} \exp \widehat{w}_{h,m} \end{array} \right) = 0$$
$$\implies \widehat{w}_{h,m} = \log p\left(\mathrm{y}_{h,m} = 1 | \mathbf{x} = \boldsymbol{x}\right) \quad (4)$$

where $p\left(\mathrm{y}_{h,m} = 1|\mathbf{x} = \boldsymbol{x}\right)$ denotes the marginal distribution of the dependency between words at position h and m, *i.e.* the sum of the conditional probability of trees this dependency appears in.

Theorem 3. *The loss $\widetilde{\ell}_{\text{(sep-dep)}}$ is not Bayes consistent for distributions over dependency trees.*

Proof. Let $n = 2$ and assume that the distribution $p(\mathbf{y}|\mathbf{x} = \boldsymbol{x})$ is defined as depicted in Figure 2. Then, by Equation 4 we have: $\widehat{w}_{0,1} = \log 0.7$, $\widehat{w}_{0,2} = \log 0.6$, $\widehat{w}_{1,2} = \log 0.4$ and $\widehat{w}_{2,1} = \log 0.3$. As such:

$$\log 0.7 + \log 0.4 = \langle \widehat{\boldsymbol{w}}, \boldsymbol{a} \rangle$$
$$< \langle \widehat{\boldsymbol{w}}, \boldsymbol{b} \rangle = \log 0.7 + \log 0.6,$$

but $\boldsymbol{a} \in \arg\max_{\boldsymbol{y} \in Y} p(\mathbf{y} = \boldsymbol{y}|\mathbf{x} = \boldsymbol{x})$ and $p(\mathbf{y} = \boldsymbol{a}|\mathbf{x} = \boldsymbol{x}) > p(\mathbf{y} = \boldsymbol{b}|\mathbf{x} = \boldsymbol{x})$. Therefore, the token-separable loss is not Bayes consistent, *i.e.* a scoring model minimizing the surrogate risk may not lead to predicting the most probable tree in the data distribution. □

Note that using the (non-separable) NLL loss will lead to a Bayes consistent model on this distribution. Indeed, by Equation 2, optimal scores $\widehat{\boldsymbol{w}}$ satisfy the following condition for all $\boldsymbol{y} \in Y$:

$$\frac{\exp\langle \widehat{\boldsymbol{w}}, \boldsymbol{y} \rangle}{\sum_{\boldsymbol{y}' \in Y} \exp\langle \widehat{\boldsymbol{w}}, \boldsymbol{y}' \rangle} = p(\mathbf{y} = \boldsymbol{y}|\mathbf{x} = \boldsymbol{x})$$
$$\implies \langle \widehat{\boldsymbol{w}}, \boldsymbol{y} \rangle = \log p(\mathbf{y} = \boldsymbol{y}|\mathbf{x} = \boldsymbol{x})$$

The following assignment for $\widehat{\boldsymbol{w}}$ satisfies this condition: $\widehat{w}_{0,1} = 0$, $\widehat{w}_{0,2} = \log 0.3$, $\widehat{w}_{1,2} = \log 0.4$ and $\widehat{w}_{2,1} = 0$. That is, minimizing the NLL loss on this distribution results in a Bayes consistent classifier, as expected.

The single root constraint case is reported in Appendix A.

6 Conclusion

Studying statistical properties of surrogate loss functions has not been a major interest in the NLP community, although there are exceptions (Ma and Collins, 2018; Effland and Collins, 2021). We proved that token-separable losses for NER and dependency parsing are not Bayes consistent, which means that minimizing these losses will not necessarily lead to models that will predict the most probable output for a given input in the data distribution, even with infinite training data.

In the dependency parsing case, Zhang et al. (2020) experimentally observed that the structured NLL loss leads to better results than the token-separable head selection loss. As such, our analysis provides a better theoretical understanding of these experiments. However, separable losses are widely used in state-of-the-art models, which suggests that future research should study why they work in practice.

Other types of separability have also been used for constituency parsing (Corro, 2020) and semantic parsing (Pasupat et al., 2019), *inter alia*.

Limitations

Arguably, these separable loss functions perform well in practice, which questions the appropriateness of the Bayes consistency property. For example, Long and Servedio (2013) argued that assumptions usually made are too unrealistic (*e.g.* considering that F is the set of all measurable mappings) and leads to misleading theoretical knowledge when it comes to actual implementation and experiments. Maybe this is also the case of the demonstration we made in this paper. However, all in all, we hope that this work will motivate future fundamental research on ML for NLP and especially on properties of loss functions.

Acknowledgments

We thank Joseph Le Roux, François Yvon and the anonymous reviewers for their comments and suggestions.

References

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

Mathieu Blondel. 2019. Structured prediction with projection oracles. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Mathieu Blondel, André F.T. Martins, and Vlad Niculae. 2020. Learning with fenchel-young losses. *Journal of Machine Learning Research*, 21(35):1–69.

Yoeng-Jin Chu and Tseng-Hong Liu. 1965. On the shortest arborescence of a directed graph. *Scientia Sinica*.

Caio Corro. 2020. Span-based discontinuous constituency parsing: a family of exact chart-based algorithms with time complexities from O(nˆ6) down to O(nˆ3). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2753–2764, Online. Association for Computational Linguistics.

John Duchi, Khashayar Khosravi, and Feng Ruan. 2018. Multiclass classification, information, divergence and surrogate risk. *The Annals of Statistics*, 46(6B):3246 – 3275.

Jack Edmonds. 1967. Optimum branchings. *Journal of Research of the National Bureau of Standards – B. Mathematics and Mathematical Physics*.

Thomas Effland and Michael Collins. 2021. Partially supervised named entity recognition via the expected entity ratio loss. *Transactions of the Association for Computational Linguistics*, 9:1320–1335.

Matteo Fischetti and Paolo Toth. 1992. An additive bounding procedure for the asymmetric travelling salesman problem. *Mathematical Programming*, 53(1):173–197.

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

Terry Koo, Amir Globerson, Xavier Carreras, and Michael Collins. 2007. Structured prediction models via the matrix-tree theorem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 141–150, Prague, Czech Republic. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

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

Yufeng Liu. 2007. Fisher consistency of multicategory support vector machines. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 291–298, San Juan, Puerto Rico. PMLR.

Phil Long and Rocco Servedio. 2013. Consistency versus realizable h-consistency for multiclass classification. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 801–809, Atlanta, Georgia, USA. PMLR.

Gábor Lugosi and Nicolas Vayatis. 2004. On the Bayes-risk consistency of regularized boosting methods. *The Annals of Statistics*, 32(1):30 – 55.

Zhuang Ma and Michael Collins. 2018. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3698–3707, Brussels, Belgium. Association for Computational Linguistics.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 523–530, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Ryan McDonald and Giorgio Satta. 2007. On the complexity of non-projective data-driven dependency parsing. In *Proceedings of the Tenth International Conference on Parsing Technologies*, pages 121–132, Prague, Czech Republic. Association for Computational Linguistics.

Alex Nowak, Francis Bach, and Alessandro Rudi. 2019. A general theory for structured prediction with smooth convex surrogates. *arXiv preprint arXiv:1902.01958*.

Alex Nowak, Francis Bach, and Alessandro Rudi. 2020. Consistent structured prediction with max-min margin Markov networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7381–7391. PMLR.

Alex Nowak, Alessandro Rudi, and Francis Bach. 2022. On the consistency of max-margin losses. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of

Proceedings of Machine Learning Research, pages 4612–4633. PMLR.

Panupong Pasupat, Sonal Gupta, Karishma Mandyam, Rushin Shah, Mike Lewis, and Luke Zettlemoyer. 2019. Span-based hierarchical semantic parsing for task-oriented dialog. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1520–1526, Hong Kong, China. Association for Computational Linguistics.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics.

Mark D. Reid and Robert C. Williamson. 2010. Composite binary losses. *Journal of Machine Learning Research*, 11(83):2387–2422.

David A. Smith and Noah A. Smith. 2007. Probabilistic models of nonprojective dependency trees. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 132–140, Prague, Czech Republic. Association for Computational Linguistics.

Robert Endre Tarjan. 1977. Finding optimum branchings. *Networks*, 7(1):25–35.

Ambuj Tewari and Peter L. Bartlett. 2007. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(36):1007–1025.

Michalis Titsias. 2016. One-vs-each approximation to softmax for scalable estimation of probabilities. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Martin J. Wainwright and Michael Irwin Jordan. 2008. *Graphical models, exponential families, and variational inference*. Now Publishers Inc.

Robert C. Williamson, Elodie Vernet, and Mark D. Reid. 2016. Composite multiclass losses. *Journal of Machine Learning Research*, 17(222):1–52.

Tong Zhang. 2004a. Statistical analysis of some multicategory large margin classification methods. *Journal of Machine Learning Research*, 5(Oct):1225–1251.

Tong Zhang. 2004b. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56 – 85.

Xingxing Zhang, Jianpeng Cheng, and Mirella Lapata. 2017. Dependency parsing as head selection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 665–676, Valencia, Spain. Association for Computational Linguistics.

Yu Zhang, Zhenghua Li, and Min Zhang. 2020. Efficient second-order TreeCRF for neural dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3295–3305, Online. Association for Computational Linguistics.

A Inconsistency of the separable loss for single root dependency parsing

Theorem 4. *The loss $\widetilde{\ell}_{\text{(sep-dep)}}$ is not Bayes consistent for distributions over single-root dependency trees.*

Proof. Let $n = 4$ and assume that the distribution $p(\mathbf{y}|\mathbf{x} = \boldsymbol{x})$ is defined as depicted in Figure 3. Then, by Equation 4 we have, among others:

$$\widehat{w}_{0,1} = \log 0.55, \qquad \widehat{w}_{1,2} = \log 0.55,$$
$$\widehat{w}_{1,3} = \log 0.4, \qquad \widehat{w}_{2,3} = \log 0.45.$$

As such, we have:

$$\langle \widehat{\boldsymbol{w}}, \boldsymbol{a} \rangle = \log 0.55 + \log 0.55 + \log 0.40,$$
$$\langle \widehat{\boldsymbol{w}}, \boldsymbol{b} \rangle = \log 0.55 + \log 0.55 + \log 0.45,$$

which means that $\langle \widehat{\boldsymbol{w}}, \boldsymbol{a} \rangle < \langle \widehat{\boldsymbol{w}}, \boldsymbol{b} \rangle$ but $\boldsymbol{a} \in \arg\max_{\boldsymbol{y} \in Y} p(\mathbf{y} = \boldsymbol{y}|\mathbf{x} = \boldsymbol{x})$ and $p(\mathbf{y} = \boldsymbol{a}|\mathbf{x} = \boldsymbol{x}) > p(\mathbf{y} = \boldsymbol{b}|\mathbf{x} = \boldsymbol{x})$. Therefore the separable loss is not Bayes consistent for single-root dependency parsing. □

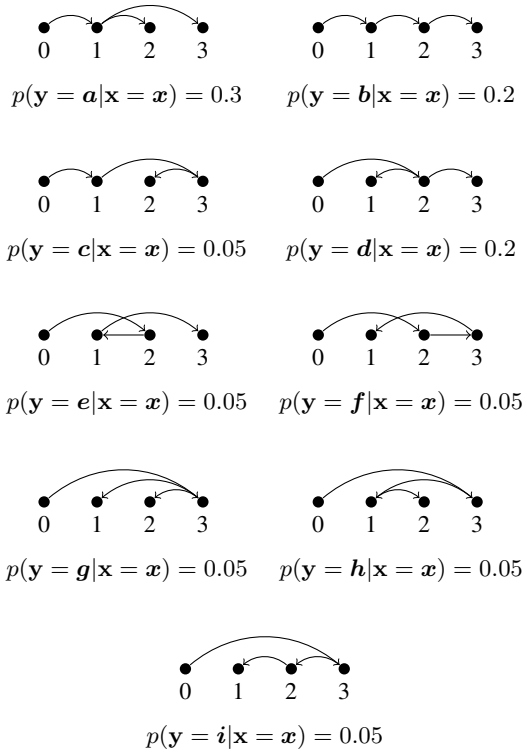

Figure 3: Example of distribution over trees for a sentence of 3 words and the single-root constraints. The set of trees is defined as $Y = \{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}, \boldsymbol{e}, \boldsymbol{f}, \boldsymbol{g}, \boldsymbol{h}, \boldsymbol{i}\}$.