Structured Prediction with Stronger Consistency Guarantees

Anqi Mao  
Courant Institute  
New York, NY 10012  
aqmao@cims.nyu.edu

Mehryar Mohri  
Google Research & CIMS  
New York, NY 10011  
mohri@google.com

Yutao Zhong  
Courant Institute  
New York, NY 10012  
yutao@cims.nyu.edu

Abstract

We present an extensive study of surrogate losses for structured prediction supported by $\mathcal{H}$-consistency bounds. These are recently introduced guarantees that are more relevant to learning than Bayes-consistency, since they are not asymptotic and since they take into account the hypothesis set $\mathcal{H}$ used. We first show that no non-trivial $\mathcal{H}$-consistency bound can be derived for widely used surrogate structured prediction losses. We then define several new families of surrogate losses, including structured comp-sum losses and structured constrained losses, for which we prove $\mathcal{H}$-consistency bounds and thus Bayes-consistency. These loss functions readily lead to new structured prediction algorithms with stronger theoretical guarantees, based on their minimization. We describe efficient algorithms for minimizing several of these surrogate losses, including a new structured logistic loss.

1 Introduction

In most applications, the output labels of learning problems have some structure that is crucial to consider. This includes natural language processing applications, where the output may be a sentence, a sequence of parts-of-speech tags, a parse tree, or a dependency graph. It also includes image annotation, image segmentation, computer vision, video annotation, object recognition, motion estimation, computational photography, bioinformatics, and many other important applications.

Several algorithms have been designed in the past for structured prediction tasks, including Conditional Random Fields (CRFs) [Lafferty et al., 2001a, Gimpel and Smith, 2010], StructSVMs [Tsochantaridis et al., 2005a], Maximum-Margin Markov Networks (M3N) [Taskar et al., 2003a], kernel-regression-based algorithms [Cortes et al., 2005, 2007], Voted CRF and StructBoost [Cortes et al., 2016], search-based methods [Daumé III et al., 2009, Doppa et al., 2014, Lam et al., 2015, Chang et al., 2015, Ross et al., 2011] and a variety of deep learning techniques [Jurafsky and Martin, 2009, Vinyals et al., 2015a, Nadeau and Sekine, 2007, Zhang et al., 2008, Wu et al., 2016, Lucchi et al., 2013, Vinyals et al., 2015b], see Appendix A for a more comprehensive list of references and discussion.

Structured prediction tasks inherently involve a natural loss function based on substructures, which could be the Hamming loss, the $n$-gram loss, the edit-distance loss, or some other sequence similarity-based loss or task-specific structured loss. Many of the algorithms previously mentioned overlook this inherent structured loss by simply minimizing the cross-entropy loss. In contrast, the surrogate loss functions minimized by algorithms such as CRFs [Lafferty et al., 2001a, Gimpel and Smith, 2010], M3N [Taskar et al., 2003a], StructSVMs [Tsochantaridis et al., 2005a] or Voted CRF and StructBoost [Cortes et al., 2016] do take into account the natural structured loss of the task. But are these structured prediction loss functions consistent? What guarantees can we rely on when minimizing them over a restricted hypothesis set that does not include all measurable functions? Can we derive non-asymptotic guarantees?
This paper deals precisely with these theoretical problems in structured prediction.

**Previous work.** We include a detailed discussion on consistency in structured prediction in Appendix A. Here, we briefly discuss previous work by Osokin et al. [2017]. To our knowledge, this is one of the only studies proving Bayes-consistency for a family of loss functions in structured prediction (see also [Nowak et al., 2020] and other related references in Appendix A). The surrogate losses the authors proposed are the following quadratic losses (see also [Zhang, 2004]) defined for any function $h$ mapping $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$ and any loss function $\ell$ between output labels by

$$
\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \quad L^\text{quad}(h, x, y) = \sum_{y' \in \mathcal{Y}} [\ell(y', y) + h(x, y')]^2.
$$

(1)

However, the authors only consider the hypothesis set of linear scoring functions. Moreover, the feature vector in their setting only depends on the input $x$ and ignores the label $y$. In many applications such as natural language prediction, however, it is critical to allow for features that depend both on the input sequence and the output sequence, parse tree, or dependency graph. Finally, in this formulation, the structured prediction problem is cast as a regression problem. Thus, as shown below, the loss function derived is non-standard, even in the binary classification case, where $\ell = \ell_{0-1}$ is the zero-one loss and $\mathcal{Y} = \{y_1, y_2\}$. In this simple case, $L^\text{quad}(h, x, y_1)$ can be expressed as

$$
L^\text{quad}(h, x, y_1) = \sum_{y' \in \mathcal{Y}} [\ell_{0-1}(y_1, y_1) + h(x, y')]^2 = h(x, y_1)^2 + (1 + h(x, y_2))^2.
$$

(2)

This is not a typical formulation since it incorporates the magnitude of individual scores. In contrast, in standard binary classification scenario, only the difference between scores matters.

**Structure of the paper.** We present an extensive study of surrogate losses for structured prediction supported by $\mathcal{H}$-consistency bounds. These are recently introduced guarantees that are more relevant to learning than Bayes-consistency, since they are not asymptotic and since they take into account the hypothesis set $\mathcal{H}$ used. We first show that no non-trivial $\mathcal{H}$-consistency bound or even Bayes-consistency can be derived for widely used surrogate structured prediction losses (Section 3). We then define several new families of surrogate losses, including structured comp-sum losses (Section 4) and structured constrained losses (Section 5), for which we prove $\mathcal{H}$-consistency bounds and thus Bayes-consistency. These loss functions readily lead to new structured prediction algorithms with stronger theoretical guarantees, based on their minimization. We also describe efficient gradient computation algorithms for several of these surrogate losses, including a new structured logistic loss (Section 6).

2 Preliminaries

**Learning scenario.** We consider the standard structured prediction scenario with the input space $\mathcal{X}$ and output space $\mathcal{Y} = \{1, \ldots, n\}$. The output space may be discrete objects with overlapping structures, such as sequences, images, graphs, parse trees, lists, or others. We assume that the output can be decomposed into $l$ substructures. The substructures could represent words or tokens for example, or other subsequences along a sequence, resulting in the decomposition of the output space $\mathcal{Y}$ as $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_l$. Here, each $\mathcal{Y}_j$ represents the set of possible labels or classes that can be assigned to the $j$-th substructure.

**Scoring function.** Structured prediction is typically formulated via scoring functions that map $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$, which assign a score to each possible pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Let $\mathcal{H}$ be a family of such scoring functions. For any $h \in \mathcal{H}$, we denote by $h(x)$ its prediction for the input $x \in \mathcal{X}$, which is the output $y \in \mathcal{Y}$ that maximizes the score $h(x, y)$, that is, $h(x) = \arg\max_{y \in \mathcal{Y}} h(x, y)$, with a fixed deterministic strategy to break ties in selecting the label with the highest score. For simplicity, we choose the label with the highest index under the natural ordering of labels as the tie-breaking strategy. We denote by $\mathcal{H}$ all the family of all measurable scoring functions.

**Generalization error and target loss.** Given a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and a loss function $L: \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the generalization error of a hypothesis $h \in \mathcal{H}$ and the best-in-class generalization error are defined as follows:

$$
\mathcal{R}_L(h) = \mathbb{E}_{(x,y) \sim \mathcal{D}} [L(h, x, y)] \quad \text{and} \quad \mathcal{R}^*_L = \inf_{h \in \mathcal{H}} \mathcal{R}_L(h).
$$
In structured prediction, the goal is to select a hypothesis \( h \in \mathcal{H} \) with small generalization error with respect to a target loss function \( L(h, x, y) \), which can be written as \( L(h, x, y) = \ell(h(x), y) \) for some non-negative auxiliary function \( \ell(y', y) \) with any \( y', y \in Y \). \( \ell \) is assumed to be symmetric, that is, \( \ell(y', y) = \ell(y, y') \). This is a natural assumption since all instances of \( \ell \) that we are familiar with in structured prediction admit this property. A significant characteristic of structured prediction is that the target loss function can be decomposed along the substructures \( Y_k \). As an example, we may use the Hamming loss as the target loss function, which is defined as \( L_{\text{ham}}(h, x, y) = \ell_{\text{ham}}(h(x), y) \), where \( \ell_{\text{ham}}(y', y) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}_{y'_j \neq y_j} \) with any \( y'_j, y_j \in Y_j \); we can also use the zero-one loss as the target loss function, which is defined as \( L_{\text{0-1}}(h, x, y) = \ell_{\text{0-1}}(h(x), y) \), where \( \ell_{\text{0-1}}(y', y) = \mathbb{1}_{y'_j \neq y_j} \). Note that \( L_{\text{0-1}} \) can be viewed as a special case of \( L_{\text{ham}} \) when \( l = 1 \). We denote by \( \ell_{\text{max}} = \max_{y'_j \neq y_j} \ell(y', y) \) the maximum value of a target loss function. Without loss of generality, we assume that \( \ell_{\text{max}} \leq 1 \), which can be achieved by normalizing the function \( \ell \).

### Consistency guarantees and surrogate loss.

Optimizing the target loss functions in structured prediction for many choices of the hypothesis sets is NP-hard because they are not convex. One common method to address this issue is to resort to surrogate loss functions. Different surrogate loss functions readily lead to different structured prediction algorithms. A natural learning guarantee for such surrogate losses is Bayes-consistency, which guarantees that minimizing the generalization error for a surrogate loss \( L_{\text{sur}} \) over \( \mathcal{H}_{\text{all}} \) also leads to the minimization of generalization error for the target loss \( L \).

**Definition 1** (Bayes-consistency). A surrogate loss \( L_{\text{sur}} \) is Bayes-consistent in structured prediction, if for any target loss \( \ell \), hypothesis \( h_n \in \mathcal{H}_{\text{all}} \) and any distribution,

\[
(R_{L_{\text{sur}}}(h_n) - R^*_L,\mathcal{H}_{\text{all}} \xrightarrow{n \to \infty} 0) \implies \left( R_L(h_n) - R^*_L,\mathcal{H}_{\text{all}} \xrightarrow{n \to \infty} 0 \right).
\]  

(3)

Bayes-consistency is an asymptotic guarantee and does not take into account typical hypothesis sets used in structured prediction algorithms, such as linear models or neural networks. To tackle these issues, recent work by Awasthi, Mao, Mohri, and Zhong [2022a,b] propose a stronger consistency guarantee, referred to as \( \mathcal{H} \)-consistency bounds, which are bounds relating the estimation error of the target loss to the estimation error of a surrogate loss (see also [Awasthi et al., 2021a, b, Mao et al., 2023h,e,f.Zheng et al., 2023, Mao et al., 2023b,g,d, Mohri et al., 2023, Mao et al., 2023c,a, Awasthi et al., 2023a,b]):

**Definition 2** (\( \mathcal{H} \)-consistency bounds). Given a subset of the hypothesis class \( \mathcal{H} \subseteq \mathcal{H}_{\text{all}} \), a surrogate loss \( L_{\text{sur}} \) admits a \( \mathcal{H} \)-consistency bound in structured prediction, if for some non-decreasing function \( f: \mathbb{R}_+ \to \mathbb{R}_+ \), a bound of the following form holds for any target loss \( \ell \), hypothesis \( h \in \mathcal{H} \) and any distribution:

\[
R_L(h) - R^*_L,\mathcal{H} \leq f \left( R_{L_{\text{sur}}}(h) - R^*_{L_{\text{sur}},\mathcal{H}} \right).
\]  

(4)

As pointed out by Awasthi et al. [2022a,b], \( \mathcal{H} \)-consistency bounds are the state-of-the-art consistency guarantees for surrogate losses. They are much stronger and more informative than Bayes-consistency, since they account for hypothesis sets \( \mathcal{H} \) adopted and provide a quantitative, non-asymptotic relation between surrogate losses and target losses. \( \mathcal{H} \)-consistency bounds can imply Bayes-consistency when taking \( \mathcal{H} \) to be \( \mathcal{H}_{\text{all}} \). In the next sections, we will present an extensive study of surrogate losses for structured prediction supported by \( \mathcal{H} \)-consistency bounds.

### Conditional regret and minimizability gap.

We denote by \( p(x) = (p(x, 1), \ldots, p(x, n)) \) the conditional distribution of \( Y \) given \( X = x \). Then, the conditional error of a hypothesis \( h \) for a loss function \( L \), denoted by \( \mathcal{C}_L(h, x) \), can be expressed as

\[
\mathcal{C}_L(h, x) = \mathbb{E}_{y \mid x} \left[ \ell(h(x), y) \right] = \sum_{y \in Y} p(x, y) \ell(h(x), y).
\]

We further define the best-in-class conditional error and the conditional regret as \( \mathcal{C}^*_L(\mathcal{H}, x) = \inf_{h \in \mathcal{H}} \mathcal{C}_L(h, x) \) and \( \Delta \mathcal{C}_L(h, x) = \mathcal{C}_L(h, x) - \mathcal{C}^*_L(\mathcal{H}, x) \) respectively. The generalization error can then be rewritten as \( R_L(h) = \mathbb{E}_x \left[ \mathcal{C}_L(h, x) \right] \).

A key quantity appearing in \( \mathcal{H} \)-consistency bounds is the minimizability gap \( M_L(\mathcal{H}) \), which measures the difference between the best-in-class generalization error and the expected best-in-class conditional error for a loss function \( L \) and a hypothesis set \( \mathcal{H} \).

\[
M^*_L(\mathcal{H}) = \mathbb{E}_x \left[ \mathcal{C}^*_L(\mathcal{H}, x) \right] = \mathbb{E}_x \left[ \mathcal{C}^*_L(\mathcal{H}, x) \right].
\]  

This is an inherent quantity that we cannot hope to minimize or estimate. As shown by Steinwart [2007,
We denote by $H$ where $\Phi$ structured max losses which we refer to as 

Theorem 4 (The following gives a general negative result for a target loss $L$, which will be helpful for proving $\mathcal{H}$-consistency bounds in structured prediction. We denote by $H(x)$ the set of all possible predictions on a input $x$ generated by hypotheses in $\mathcal{H}$: $H(x) = \{h(x) : h \in \mathcal{H}\}$. The proof is given in Appendix B.

**Lemma 3.** The best-in-class conditional error and the conditional regret for a target loss $L$ in structured prediction can be expressed as follows:

$$
\widehat{G}^*_{L,\mathcal{H}}(x) = \min_{y' \in H(x)} \sum_{y \in Y} p(x, y) \ell(y', y)
$$

$$
\Delta \mathcal{G}_{L,\mathcal{H}}(h, x) = \sum_{y' \in Y} p(x, y) \ell(h(x), y) - \min_{y' \in H(x)} \sum_{y \in Y} p(x, y) \ell(y', y).
$$

3 Structured max losses

In this section, we examine the loss functions associated to several prominent structured prediction algorithms. We show that, while they are natural, none of them is Bayes-consistent, which implies that they cannot be supported by $\mathcal{H}$-consistency bounds.

**Theorem 3.2**, the minimizability gaps vanish $M$ Lemma 3.

guarantees.

a new family of loss functions for structured predictions that we prove to admit strong consistency.

In this section, we first analyze the Voted CRF loss function, which incorporates the auxiliary loss function $\ell$ in the CRF loss and which has been used in several previous studies. Next, we introduce a new family of loss functions for structured predictions that we prove to admit strong consistency guarantees.

4 Structured comp-sum losses

In this section, we first analyze the Voted CRF loss function, which incorporates the auxiliary loss function $\ell$ in the CRF loss and which has been used in several previous studies. Next, we introduce a new family of loss functions for structured predictions that we prove to admit strong consistency guarantees.
4.1 Voted Conditional Random Field (VCRF)

We first study a family of surrogate losses called Voted Conditional Random Field (VCRF), which corresponds to the structured prediction algorithm defined in [Cortes et al., 2016]:

$$\forall (x, y) \in X \times Y, \quad L_{VCRF}^V(h, x, y) = -\log \left[ \frac{e^{h(x, y)}}{\sum_{y' \in Y} e^{h(x, y') + v(y, y')} \Phi(v(y, y') + h(x, y) - h(x, y'))} \right].$$

This loss function has also been presented as the softmax margin [Gimpel and Smith, 2010] or the reward-augmented maximum likelihood [Norouzi et al., 2016]. It can be viewed as the softmax variant of the M3N loss. Indeed, the loss function for M3N can be written as follows:

$$L(h, x, y) = \max_{y'} \log \left( \frac{e^{h(x, y') + h(x, y) - h(x, y)}}{\sum_{y' \in Y} e^{h(x, y')}} \right).$$

(6)

If we replace the maximum function with the softmax, we obtain

$$L(h, x, y) = \log \left[ \sum_{y' \in Y} e^{\max(0, \ell(y', y) + h(x, y') - h(x, y))} \right] = \log \left[ \sum_{y' \in Y} \max(1, e^{\ell(y', y) + h(x, y') - h(x, y)}) \right].$$

(7)

Next, we show that, as with the loss function for M3N, the VCRF loss function $L_{VCRF}$ is inconsistent.

**Theorem 5 (Negative result of $L_{VCRF}$).** The Voted Conditional Random Field $L_{VCRF}$ is not Bayes-consistent.

The proof is included in Appendix D. The key observation in the proof is that the conditional error of VCRF loss function can be reduced to a specific form when the target loss function $L$ decouples, which can lead to a different Bayes classifier from that of the target loss function.

To the best of our knowledge, no prior studies in the literature have explored the consistency of the VCRF loss formulation. The most closely related discussions center around a specialized instance of which can lead to a different Bayes classifier from that of the target loss function.

Theorem 5 suggests that, even when integrating the target loss function within its formulation, the Voted Conditional Random Field cannot be consistent.

Along with Theorem 4, these results rule out consistency guarantees for commonly used surrogate loss functions in structured prediction.

4.2 Structured comp-sum loss functions

In this section, we define a family of new loss functions for structured prediction that are not only Bayes-consistent but also supported by $\mathcal{H}$-consistency bounds. These are loss functions that can be viewed as the generalization to structured prediction of loss functions defined via a composition and a sum, and that have been referred to as *comp-sum losses* in [Mao et al., 2023h]. Thus, we will refer to them as *structured comp-sum losses*. They are defined as follows:

$$\forall (x, y) \in X \times Y, \quad L_{comp}^V(h, x, y) = \sum_{y' \in Y} \ell'(y', y) \Phi_1(v(x, y')) \sum_{v \in \mathbb{R}_+} \Phi_2(v(x, y') - h(x, y'))$$

(8)

where $\ell'(y', y)$ is a non-decreasing auxiliary function and $\Phi_2: \mathbb{R} \to \mathbb{R}_+$ a non-decreasing auxiliary function. This formulation (8) can also be viewed as a weighted comp-sum loss, if we interpret $\ell'(\cdot, y)$ as a weight vector.

Specifically, we can choose $\Phi_2(v) = e^v$ and $\Phi_1(v) = \log(v)$, $\Phi_1(v) = v - 1, \Phi_1(v) = 1 - \frac{1}{v^\alpha}$, which leads to new surrogate losses for structured prediction defined in Table 1. These surrogate losses are novel strict generalization of their counterparts in the standard multi-class classification case where $\ell = \ell_{0,1}$. More precisely, when $\ell = \ell_{0,1}$, $L_{comp}^V$ coincides with the logistic loss [Verhulst, 1838, 1845; Berkson, 1944, 1951]; $L_{comp}^V$ coincides with the sum-exponential loss [Weston and Watkins, 1998; Awasthi et al., 2022b]; $L_{comp}^V$ coincides with the
We will show that these structured comp-sum losses benefit from \( \mathcal{H} \)-consistency bounds in structured prediction, when \( \mathcal{H} \) is a symmetric and complete hypothesis set. A hypothesis set \( \mathcal{H} \) is symmetric if there exists a family \( \mathcal{F} \) of real-valued functions such that \( \{ [h(x, 1), \ldots, h(x, n)]: h \in \mathcal{H} \} = \{ [f_1(x), \ldots, f_n(x)]: f_1, \ldots, f_n \in \mathcal{F} \} \) for any \( x \in \mathcal{X} \). Thus, the choice of the scoring functions does not depend on the order of the categories in \( y \). A hypothesis set \( \mathcal{H} \) is complete if it can generate scores that span \( \mathbb{R} \), that is, \( \{ h(x, y): h \in \mathcal{H} \} = \mathbb{R} \) for any \( (x, y) \in \mathcal{X} \times \mathcal{Y} \). As shown by Awasthi et al. [2022b] and Mao et al. [2023h], these assumptions are general and hold for common hypothesis sets used in practice, such as the family of linear hypotheses and that of multi-layer feed-forward neural networks, and of course that of all measurable functions.

**Theorem 6 \((H\text{-consistency bound of } L^\text{comp})\).** Assume that \( \mathcal{H} \) is symmetric and complete. Then, for any target loss \( \ell \), any hypothesis \( h \in \mathcal{H} \) and any distribution, we have

\[
R_L(h) - R^*_L \leq \Gamma(R^\text{comp}_L(h)) - M_{L, \mathcal{H}, \mathcal{S}} - M_{L, \mathcal{H}, \mathcal{G}}, \tag{9}
\]

where \( \Gamma(t) = 2 \sqrt{t} \) when \( L^\text{comp} = L^\text{comp}_\log \) or \( L^\text{comp}_\exp \); \( \Gamma(t) = 2 \sqrt{n \epsilon^2 t} \) when \( L^\text{comp} = L^\text{comp}_{\text{gce}} \); and \( \Gamma(t) = nt \) when \( L^\text{comp} = L^\text{comp}_{\text{mae}} \).

Theorem 6 represents a consolidated result for the four structured comp-sum losses, with the proofs for each being presented separately in Appendix E. The key step of the proof is to upper bound the conditional regret of the target loss (Lemma 3) by that of a surrogate loss. To achieve this, we upper bound the best-in-class conditional error by the conditional error of a carefully selected hypothesis \( \tilde{h}_\mu \in \tilde{\mathcal{H}} \). The resulting softmax \( \tilde{h}_\mu \) of this hypothesis only differs from the original softmax \( S \) corresponding to \( \mathcal{T} \) on two labels. Theorem 6 admits as special cases the \( \mathcal{H} \)-consistency bounds of Mao et al. [2023h] given for standard multi-class classification \((\ell = \ell_{0-1})\) and significantly extends them to the general structured prediction scenario.

Let us emphasize that our proof technique is novel and distinct from the approach used in [Mao et al., 2023h], which only applies to the special case where \( \ell \) is the zero-one loss and cannot be generalized to any target loss \( \ell \). In their proof, the authors choose \( \tilde{h}_\mu \) based on individual scores \( \hat{h}(x, y) \), rather than the softmax. Consequently, when \( \ell \neq \ell_{0-1} \), as is common in structured prediction, the resulting optimization problem of \( \mu \) can be very intricate and a closed-form expression of the optimization solution cannot be derived. However, our new proof method overcomes this limitation. By viewing the softmax of hypothesis as a unit and introducing a pseudo-conditional distribution \( \mathcal{T} \), we are able to solve a simple constrained optimization problem on \( \mu \) within structured prediction scenario.

By Steinwart [2007, Theorem 3.2], the minimizability gaps \( M_{L, \mathcal{H}, \mathcal{S}} \) and \( M_{L, \mathcal{H}, \mathcal{G}} \) vanish for the family of all measurable functions. Therefore, when \( \mathcal{H} = \mathcal{H}_{\text{all}} \), the \( \mathcal{H} \)-consistency bounds provided in Theorem 6 imply the Bayes-consistency of these structured comp-sum losses.

**Corollary 7.** The structured comp-sum loss \( L^\text{comp} \) is Bayes-consistent for \( L^\text{comp} = L^\text{comp}_\log, L^\text{comp}_\exp, L^\text{comp}_{\text{gce}}, \) and \( L^\text{comp}_{\text{mae}} \).

In fact, Theorem 6 provides stronger quantitative bounds than Bayes-consistency when the minimizability gaps vanish, which suggests that if the estimation error of the structured comp-sum loss \( R_{L, \text{comp}}(h) - R^*_L \) is reduced to \( \epsilon \), the estimation error of the target loss \( R_L(h) - R^*_L \) is upper bounded by \( 2\sqrt{\epsilon} \) for structured logistic loss and structured sum-exponential loss, \( 2\sqrt{n \epsilon^2} \) for structured generalized cross-entropy loss, and \( n \epsilon \) for structured mean absolute error loss.
As with the cases of structured comp-sum losses, Theorem 8 provides in fact stronger quantitative bounds than Bayes-consistency. They show that if the estimation error of the structured constrained loss \( R_{\mathcal{H}}(h) - R_{\mathcal{H}}^* \) is reduced to \( \epsilon \), then the estimation error of the target loss \( R_{\mathcal{H}}(h) - R_{\mathcal{H}}^* \) is upper bounded by \( 2\sqrt{t_{\text{max}}} \epsilon \) for \( \text{L}_{\text{hinge}} \), \( 2\sqrt{t} \epsilon \) for \( \text{L}_{\text{sq-hinge}} \), and \( \epsilon \) for \( \text{L}_{\text{cstnd}} \) and \( \text{L}_{\rho}^{\text{cstnd}} \). It is important to note that we can upper bound the minimizability gap by the approximation error, or finer terms depending on the magnitude of the parameter space as in [Mao et al., 2023b]. Furthermore, our \( \mathcal{H} \)-consistency bounds (Theorems 6 and 8) can be used to derive finite sample learning bounds for a hypothesis set \( \mathcal{H} \). These bounds depend on the Rademacher complexity of the hypothesis set and the loss function, as well as an upper bound on the minimizability gap for the surrogate loss.

### Table 2: A new family of surrogate losses for structured prediction: structured constrained losses.

<table>
<thead>
<tr>
<th>( \Phi_u(v) )</th>
<th>Name</th>
<th>Formulation (( \sum_{y \in Y} h(x, y) = 0 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( u e^{-v} )</td>
<td>Structured constrained exponential loss</td>
<td>( L_{\text{cstnd}} = \sum_{y \neq y'} (l(y', y) \max(0, 1 - h(x, y')))^2 )</td>
</tr>
<tr>
<td>( u \max{0, 1 - v}^2 )</td>
<td>Structured constrained squared-hinge loss</td>
<td>( L_{\text{cstnd}} = \sum_{y \neq y'} (l(y', y) \max(0, 1 - h(x, y')) \max(0, 1 - h(x, y')) )</td>
</tr>
<tr>
<td>( u \max{0, 1 - v} )</td>
<td>Structured constrained hinge loss</td>
<td>( L_{\text{cstnd}} = \sum_{y \neq y'} (l(y', y) \max(0, 1 - h(x, y')) )</td>
</tr>
<tr>
<td>( u \min{\max{0, 1 - \frac{v}{\rho}}, 1} )</td>
<td>Structured constrained ( \rho )-margin loss</td>
<td>( L_{\text{cstnd}} = \sum_{y \neq y'} (l(y', y) \min(0, 1 - h(x, y')) )</td>
</tr>
</tbody>
</table>

### 5 Structured constrained loss functions

In this section, we introduce another new family of surrogate losses for structured prediction that we prove to admit \( \mathcal{H} \)-consistency bounds. We will present a novel generalization of the constrained losses [Lee et al., 2004, Awasthi et al., 2022b] to structured prediction. Thus, we refer to them as structured constrained losses and define them as follows:

\[
\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \quad L_{\text{cstnd}}(h, x, y) = \sum_{y' \in \mathcal{Y}} \Phi_u(y', y)(-h(x, y')),
\]

with the constraint that \( \sum_{y' \in \mathcal{Y}} h(x, y) = 0 \) and \( \Phi_u: \mathbb{R} \to \mathbb{R}_+ \) is an upper bound on \( u \mapsto u L_{\mathcal{Y}} \) for any \( u \in \mathbb{R}_+ \). In standard constrained loss formulation, a single-variable function \( \Phi(v) \) that defines a margin-based loss is used. In (10), the single-variable function \( \Phi(v) \) is generalized to being a function of two variables \( \Phi_u(v) \), which depends on both the target loss and the scores, to accommodate the structured prediction scenario. Specifically, we can choose \( \Phi_u(v) = u e^{-v} \), \( \Phi_u(v) = u \max\{0, 1 - v\}^2 \), \( \Phi_u(v) = u \max\{0, 1 - v\} \), \( \Phi_u(v) = u \min\{\max\{0, 1 - v/\rho\}, 1\} \), which lead to new surrogate losses for structured prediction defined in Table 2. These surrogate losses are novel generalization of their corresponding counterparts [Lee et al., 2004, Awasthi et al., 2022b] in standard multi-class classification, where \( \ell = \ell_{0-1} \). As with structured comp-sum losses, we will show that these structured constrained losses benefit from \( \mathcal{H} \)-consistency bounds in structured prediction as well, for any symmetric and complete hypothesis set \( \mathcal{H} \).

**Theorem 8 (\( \mathcal{H} \)-consistency bound of \( L_{\text{cstnd}} \)).** Assume that \( \mathcal{H} \) is symmetric and complete. Then, for any target loss \( \ell \), hypothesis \( h \in \mathcal{H} \) and any distribution, we have

\[
R_{\mathcal{H}}(h) - R_{\mathcal{H}}^* \leq \Gamma\left( R_{L_{\text{cstnd}}} - R_{L_{\text{exp}}} + M_{L_{\text{cstnd}} + \mathcal{H}} - M_{L_{\mathcal{H}}} \right),
\]

where \( \Gamma(t) = 2\sqrt{t_{\text{max}}} t \) when \( L_{\text{cstnd}} = L_{\text{exp}} \), \( \Gamma(t) = 2\sqrt{t} \) when \( L_{\text{cstnd}} = L_{\text{sq-hinge}} \), and \( \Gamma(t) = t \) when \( L_{\text{cstnd}} = L_{\text{hinge}} \) or \( L_{\text{cstnd}} \).

The proof is included in Appendix F. As for Theorem 8, the key part of the proof is to upper bound the conditional regret of the target loss (Lemma 3) by that of a surrogate loss. Here too, we introduce a pseudo-conditional distribution \( q \), which can be viewed as a weighted distribution of the original one, \( p(x) \), with weights given by the target loss function. Then, we upper bound the best-in-class conditional error by the conditional error of a carefully selected hypothesis \( \overline{\mu} \).

As shown by Steinwart [2007, Theorem 3.2], for the family of all measurable functions, the minimizability gaps vanish: \( M_{L_{\text{cstnd}} + \mathcal{H}} = 0 \) and \( M_{L_{\mathcal{H}}} = 0 \). Therefore, when \( \mathcal{H} = \mathcal{H}_{\text{all}} \), the \( \mathcal{H} \)-consistency bounds provided in Theorem 6 imply the Bayes-consistency of these structured constrained losses.

**Corollary 9.** The structured constrained loss \( L_{\text{cstnd}} \) is Bayes-consistent for \( L_{\text{cstnd}} = L_{\text{exp}} \), \( L_{\text{sq-hinge}} \), \( L_{\text{hinge}} \), and \( L_{\rho}^{\text{cstnd}} \).
6 Optimization of $L_{\text{log}}^{\text{comp}}$ and $L_{\text{exp}}^{\text{comp}}$

In this section, we show that the gradient of the structured logistic loss $L_{\text{log}}^{\text{comp}}$ can be computed efficiently at any point $(x_i, y_i)$ and therefore that this loss function is both supported by $\mathcal{H}$-consistency bounds and is of practical use. We similarly show that for $L_{\text{exp}}^{\text{comp}}$ in Appendix G.2.

Fix the labeled pair $(x_i, y_i)$ and $h \in \mathcal{H}$. Observe that $L_{\text{log}}^{\text{comp}}(h, x_i, y_i)$ can be equivalently rewritten as follows:

$$L_{\text{log}}^{\text{comp}}(h, x_i, y_i) = \sum_{y' \in \mathcal{Y}} \bar{t}(y', y_i) \log \left[ \sum_{y'' \in \mathcal{Y}} e^{h(x, y'') - h(x, y')} \right]$$

$$= - \sum_{y' \in \mathcal{Y}} \bar{t}(y', y_i) h(x, y') + \left[ \sum_{y' \in \mathcal{Y}} \bar{t}(y', y_i) \right] \log \left[ \sum_{y'' \in \mathcal{Y}} e^{h(x, y'')} \right]$$

$$= - \sum_{y' \in \mathcal{Y}} \bar{t}(y', y_i) h(x, y') + \bar{t}_i \log Z_{h,i},$$

where $\bar{t}_i = \sum_{y' \in \mathcal{Y}} \bar{t}(y', y_i)$, and $Z_{h,i} = \sum_{y' \in \mathcal{Y}} e^{h(x, y')}$. Note that $\bar{t}_i$ does not depend on $h$ and can be pre-computed. Modulo normalization, this quantity is the average similarity of $y_i$ to $\mathcal{Y}$, if we interpret $\bar{t} = 1 - \ell$ as a similarity. While $|\mathcal{Y}|$ may be very large, this can be often computed straightforwardly for most loss functions $\ell$. For example, for the Hamming loss, for sequences of length $l$, we have

$$\frac{1}{|\mathcal{Y}|} \bar{t}_i = \frac{1}{l} \mathbb{E} \left[ \sum_{k=1}^{l} (1 - \mathbb{1}_{y_k = y_i}) \right] = \frac{1}{l} \sum_{k=1}^{l} \frac{1}{2} = \frac{1}{2}.$$

Thus, in this case, $\bar{t}_i$ does not depend on $i$ and is a universal constant. Similarly, $\bar{t}_i$ can be shown to be a constant for many other losses.

**Hypothesis set.** For the remaining of this section, to simplify the presentation, we will consider the hypothesis set of linear functions $\mathcal{H} = \{ x \mapsto w \cdot \Psi(x, y) : w \in \mathbb{R}^d \}$, where $\Psi$ is a feature mapping from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}^d$. Note that a number of classical structured prediction algorithms adopt the same linear hypothesis set: StructSVM [Tschantaridis et al., 2005b], Max-Margin Markov Networks (M3N) [Taskar et al., 2003b], Conditional Random Field (CRF) [Lafferty et al., 2001b], Voted Conditional Random Field (VCRF) [Cortes et al., 2016]. Our algorithms can also be incorporated into standard procedures for training neural network architectures (see [Cortes et al., 2018], Appendix B).

**Markovian features.** We will also assume Markovian features, as is common in structured prediction. Features used in practice often admit this property. Furthermore, in the absence of any such assumption, it is known that learning and inference in general are intractable. We will largely adopt here the definitions and notation from [Cortes et al., 2016] and will consider the common case where $\mathcal{Y}$ is a set of sequences of length $l$ over a finite alphabet $\Delta$ of size $r$. Other structured problems can be treated in similar ways. We will denote by $\varepsilon$ the empty string and for any sequence $y = (y_1, \ldots, y_l) \in \mathcal{Y}$, we will denote by $y_s^r = (y_s, \ldots, y_r)$ the substring of $y$ starting at index $s$ and ending at $s'$. For convenience, for $s \leq 0$, we define $y_s$ by $y_s = \varepsilon$.

We will assume that the feature vector $\Psi$ admits a Markovian property of order $q$, that is it can be decomposed as follows for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$:

$$\Psi(x, y) = \sum_{s=1}^{l} \psi(x, y_{s-q+1}, y_{s}).$$

(12)

for some position-dependent feature vector function $\psi$ defined over $\mathcal{X} \times \Delta^q \times [l]$. We note that we can write $\Psi = \sum_{k=1}^{p} \Psi_k$ with $\Psi_k = (0, \ldots, \Psi_k, \ldots, 0)$. In the following, abusing the notation, we will simply write $\Psi_k$ instead of $\Psi_k$. Each $\Psi_k$ corresponds to a Markovian feature vector based only on $k$-grams, $p$ is the largest $k$. Thus, for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have

$$\Psi(x, y) = \sum_{k=1}^{p} \Psi_k(x, y).$$

(13)
We now analyze the computation of \( \Psi(x, y) = \sum_{k=1}^{p} \Psi_k(x, y) = \sum_{k=1}^{p} \sum_{s=1}^{L} \psi_k(x, y_{s-k+1}^s) = \sum_{s=1}^{L} \sum_{k=1}^{p} \psi_k(x, y_{s-k+1}^s) = \sum_{s=1}^{L} \sum_{k=1}^{p} \tilde{\psi}(x, y_{s-p+1}^s). \) Observe then that we can write

\[
\psi(x, y_{s-p+1}^s) = \sum_{k=1}^{p} \psi_k(x, y_{s-k+1}^s) = \sum_{s=1}^{L} \psi(x, y_{s-p+1}^s).
\]

**Gradient computation.** Adopting the shorthand \( w \) for \( h \), we can rewrite the loss at \((x_i, y_i)\) as:

\[
L_{\log}^{\text{comp}}(w, x_i, y_i) = -w \cdot \left[ \sum_{y' \in \mathcal{Y}} \ell(y', y_i) \Psi(x_i, y') \right] + \ell_i \log Z_{w, i}.
\]

Thus, the gradient of \( L_{\log}^{\text{comp}} \) at any \( w \in \mathbb{R}^d \) is given by

\[
\nabla L_{\log}^{\text{comp}}(w) = -\sum_{y' \in \mathcal{Y}} \ell(y', y_i) \Psi(x_i, y') + \ell_i \sum_{y' \in \mathcal{Y}} e^{w' \Psi(x_i, y')} \Psi(x_i, y)
\]

\[
= -\sum_{y' \in \mathcal{Y}} \ell(y', y_i) \Psi(x_i, y') + \ell_i \mathbb{E}_{y' \in \mathcal{Y}} \left[ \Psi(x_i, y) \right],
\]

where \( q_w \) is defined for all \( y \in \mathcal{Y} \) by \( q_w(y) = \frac{e^{w' \Psi(x_i, y)}}{Z_w} \) with \( Z_w = \sum_{y' \in \mathcal{Y}} e^{w' \Psi(x_i, y)} \). Note that the sum defining these terms is over a number of sequences \( y \) that is exponential in \( r \) and that the computation appears to be therefore challenging. The following lemma gives the expression of the gradient of \( L_{\log}^{\text{comp}} \) and helps identify the most computationally challenging terms.

**Lemma 10.** For any \( w \in \mathbb{R}^d \), the gradient of \( L_{\log}^{\text{comp}} \) can be expressed as follows:

\[
\nabla L_{\log}^{\text{comp}}(w) = \sum_{s=1}^{L} \sum_{z \in \Delta^s} \left[ \ell_i Q_w(z, s) - L(z, s) \right] \tilde{\psi}(x_i, z, s),
\]

where \( Q_w(z, s) = \sum_{y' \in y_{s-p+1}^s \in z} q_w(y) \) and \( L(z, s) = \sum_{y' \in y_{s-p+1}^s \in z} \ell(y, y_i). \)

**Proof.** Using the decomposition of the feature vector, we can write:

\[
\sum_{y' \in \mathcal{Y}} \ell(y, y_i) \Psi(x_i, y') = \sum_{y' \in \Delta^t} \ell(y, y_i) \sum_{s=1}^{L} \sum_{z \in \Delta^s} \left[ \sum_{y' \in y_{s-p+1}^s \in z} Q_w(z, s) \right] \tilde{\psi}(x_i, z, s)
\]

\[
\mathbb{E}_{y' \in q_w} \left[ \Psi(x_i, y) \right] = \sum_{y' \in \Delta^t} q_w(y) \sum_{s=1}^{L} \sum_{z \in \Delta^s} \left[ \sum_{y' \in y_{s-p+1}^s \in z} Q_w(z, s) \right] \tilde{\psi}(x_i, z, s).
\]

This completes the proof. \( \square \)

In light of this result, the bottleneck in the gradient computation is the evaluation of \( Q_w(z, s) \) and \( L(z, s) \) for all \( s \in [l] \) and \( z \in \Delta^p \). In previous work (Cortes, Kuznetsov, Mohri, and Yang, 2016, Cortes, Kuznetsov, Mohri, Storey, and Yang, 2018), it was shown that the quantities \( Q_w(z, s) \) can be determined efficiently, all together, by running two single-source shortest-distance algorithms over the (+, ∞) semiring on an appropriate weighted finite automaton (WFA). The overall time complexity of the computation of all quantities \( Q_w(z, s) \), \( z \in \Delta^p \) and \( s \in [l] \), is then in \( O(lr^p) \), where \( r = |\Delta| \).

We now analyze the computation of \( L(z, s) \) for a fixed \( z \in \Delta^p \) and \( s \in [l] \). Note that, unlike \( Q_w(z, s) \), this term does not depend on \( w \) and can therefore be computed once and for all, before any gradient computation. The sum defining \( L(z, s) \) is over all sequences \( y \) that admit the substring \( z \) at position \( s \).

**Rational losses.** In Appendix G.1, we also give an efficient algorithm for the computation of the quantities \( L(z, s) \) in the case of Markovian losses. Here, we present an efficient algorithm for their computation in the important case of rational losses. This is a general family of loss functions based on rational kernels (Cortes, Haffner, and Mohri, 2004) that includes, in particular, n-gram losses, which can be defined for a pair of sequences \((y, y')\) as the negative inner product of the vectors of n-gram counts of \( y \) and \( y' \).
Our algorithm bears some similarity to that of Cortes et al. [2018] for the computation of the gradient of the VCRF loss function. It is however distinct because the structured prediction loss function we are considering and our definition of rational loss are both different. We will adopt a similar notation and terminology. Recall that for any sequence $y$, we denote by $y_i$ the symbol in its $i$th position and by $y'_i = y_i y_{i+1} \cdots y_j$ the substring of $y$ starting at position $i$ and ending at $j$. We denote by $E_{\mathcal{A}}$ the set of transitions of a WFA $\mathcal{A}$. Let $\mathcal{U}$ be a weighted finite-state transducer (WFST) over the $(+, \times)$ semiring over the reals, with $\Delta$ as both the input and output alphabet. Then, we define the rational loss associated to $\mathcal{U}$ for all $y, y' \in \Delta^*$ by $\mathcal{L}(y, y') = \mathcal{U}(y, y')$.

Let $\mathcal{U}$ denote a WFA over the $(+, \times)$ semiring accepting the set of all sequences of length $l$ with weight one and let $\mathcal{Y}$, denote the WFA accepting only $y_i$ with weight one. Then, by definition, the weighted transducer $\mathcal{U} \circ \mathcal{Y}$, obtained by composition maps each sequence $y$ in $\Delta^l$ to $y_i$ with weight $\mathcal{U}(y, y_i)$. The WFA $\Pi(y, y_i)$ derived from that transducer by projection on the input (that is by removing output labels) is associating to each sequence $y$ weight $\mathcal{U}(y, y_i)$. WFA $\mathcal{U}$ for $\Delta = \{a, b\}$ and $l = 3$, We use weighted determinization [Mohri, 1997] to compute an and the WFA $\mathcal{Y}$, where $y_i =$ equivalent deterministic WFA denote $\mathcal{M}$. As shown by Cortes et al. [2015][Theorem 3], $\mathcal{M}$ can be computed in polynomial time. $\mathcal{M}$ admits a unique path labeled with any sequence $y \in \Delta^l$ and the weight of that path is $\mathcal{U}(y, y_i)$. The weight of that accepting path is obtained by multiplying the weights of its transitions and that of the final state.

We now define a deterministic $p$-gram WFA $\mathcal{N}$ that accepts all sequences $y \in \Delta^l$ with each of its states $(z', s)$ encoding a $(p - 1)$-gram $z'$ read to reach it and the position $s$ in the sequence $y$ at which it is reached. The transitions of $\mathcal{N}$ are therefore defined as follows with weight one:

\[
E_{\mathcal{N}} = \left\{ \left( \left( y_{z+1}^{-1} a, 1, \left( y_{z+2}^{-1} a, s \right) \right), y \in \Delta^l, a \in \Delta, s \in [l] \right) \right\}.
\]

The initial state is $(\epsilon, 0)$ and the final states are those with the second element of the pair (the position) being $l$. Note that, by construction, $\mathcal{N}$ is deterministic. Then, the composition (or intersection) WFA $\mathcal{N} \circ \mathcal{M}$ still associates the same weight as $\mathcal{M}$ to each input string $y \in \Delta^l$. However, the states in that composition help us compute $\mathcal{L}(z, s)$. In particular, for any $z \in \Delta^p$ and $s \in [l]$, let $\mathcal{E}(z, s)$ be WFA $\mathcal{N}$ for $\Delta = \{a, b\}$, $p = 2$ the set of transitions of $\mathcal{N} \circ \mathcal{M}$ constructed by pairing the transition $(z_1^{p-1} s^{-1}, z, \omega(z, s), (z_2^p, s))$ in $\mathcal{N}$ with a transition $(q_M, z, \omega, q'_M)$ in $\mathcal{M}$. They admit the following form:

\[
\mathcal{E}(z, s) = \left\{ \left( \left( q_N, q_M \right), z_p, \omega(z, s), (q'_N, q'_M) \right) \in E_{\mathcal{N} \circ \mathcal{M}} : q_N = (z_1^{p-1} s^{-1}, 1) \right\}.
\]

The WFA $\mathcal{N} \circ \mathcal{M}$ is deterministic as a composition of two deterministic WFAs. Thus, there is a unique path labeled with a sequence $y \in \Delta^l$ in $\mathcal{N} \circ \mathcal{M}$ and $y$ admits the substring $z$ ending at position $s$ iff that path goes through a transition in $\mathcal{E}(z, s)$ when reaching position $s$. Therefore, to compute $\mathcal{L}(z, s)$, it suffices for us to compute the sum of the weights of all paths in $\mathcal{N} \circ \mathcal{M}$ going through a transition in $\mathcal{E}(z, s)$. This can be done straightforwardly using the forward-backward algorithm or two single-source shortest-distance algorithm over the $(+, \times)$ semiring [Mohri, 2002a], one from the initial state, the other one from the final states. Since $\mathcal{N} \circ \mathcal{M}$ is acyclic and admits $O(l|\Delta|^p)$ transitions, we can compute all the quantities $\mathcal{L}(z, s), s \in [l]$ and $z \in \Delta^p$, in time $O(l|\Delta|^p)$.

7 Conclusion

Our detailed study revealed shortcomings in commonly used surrogate loss functions and algorithms for structured prediction, prompting the introduction of new, strongly consistent alternatives. These findings not only enhance the theoretical and algorithmic foundations of structured prediction but also pave the way for the development of practical and effective solutions. In upcoming work, we will report an extensive empirical analysis of our algorithms. Our work provides tools and insights for future algorithm design in this domain, promising advancements in both theory and application.
References


## Contents of Appendix

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Related work</td>
<td>18</td>
</tr>
<tr>
<td>B</td>
<td>Proof of Lemma 3</td>
<td>19</td>
</tr>
<tr>
<td>C</td>
<td>Proofs for structured max losses</td>
<td>19</td>
</tr>
<tr>
<td>D</td>
<td>Proofs for Voted Conditional Random Field</td>
<td>20</td>
</tr>
<tr>
<td>E</td>
<td>Proofs for structured comp-sum losses</td>
<td>21</td>
</tr>
<tr>
<td>E.1</td>
<td>Structured logistic loss</td>
<td>21</td>
</tr>
<tr>
<td>E.2</td>
<td>Structured sum-exponential loss</td>
<td>22</td>
</tr>
<tr>
<td>E.3</td>
<td>Structured generalized cross-entropy loss</td>
<td>23</td>
</tr>
<tr>
<td>E.4</td>
<td>Structured mean absolute error loss</td>
<td>25</td>
</tr>
<tr>
<td>F</td>
<td>Proofs for structured constrained losses</td>
<td>26</td>
</tr>
<tr>
<td>F.1</td>
<td>Structured constrained exponential loss</td>
<td>26</td>
</tr>
<tr>
<td>F.2</td>
<td>Structured constrained squared-hinge loss</td>
<td>27</td>
</tr>
<tr>
<td>F.3</td>
<td>Structured constrained hinge loss</td>
<td>29</td>
</tr>
<tr>
<td>F.4</td>
<td>Structured constrained $\rho$-margin loss</td>
<td>30</td>
</tr>
<tr>
<td>G</td>
<td>Efficient gradient computation and inference</td>
<td>31</td>
</tr>
<tr>
<td>G.1</td>
<td>Efficient gradient computation for $L_{\log}^{\text{comp}}$</td>
<td>31</td>
</tr>
<tr>
<td>G.2</td>
<td>Efficient gradient computation for $L_{\exp}^{\text{comp}}$</td>
<td>32</td>
</tr>
<tr>
<td>G.3</td>
<td>Efficient Inference</td>
<td>35</td>
</tr>
</tbody>
</table>
A Related work

Structured prediction and neural networks. A variety of deep learning techniques have been used for structured prediction tasks, including a unified neural network architecture for natural language processing [Collobert et al., 2011], energy model-based structured prediction including structured prediction energy networks (SPENs) and various inference methods [Belanger and McCallum, 2016, Tu and Gimpel, 2018, Larsson et al., 2018, Tu and Gimpel, 2019, Tu et al., 2019, Graber and Schwing, 2019, Pan et al., 2020], attention networks incorporating richer structural distributions [Kim et al., 2017], tree-structured long short-term memory (LSTM) networks [Tai et al., 2015], sequence to sequence learning [Sutskever et al., 2014, Edunov et al., 2017], memory-reduced variant of best-first beam search [Meister et al., 2020], end-to-end learning for SPENs [Belanger et al., 2017], conditional generative flow [Lu and Huang, 2020], proximal methods [Wang et al., 2016], CRF using deep features [Jaderberg et al., 2014, Huang et al., 2015, Chen et al., 2014, Schwing and Urtasun, 2015, Chen et al., 2015], non-iterative feed-forward predictors [Stoyanov et al., 2011, Donke, 2013, Kunisch and Pock, 2013, Hershey et al., 2014, Li and Zemel, 2014, Belharbi et al., 2018, Zheng et al., 2015], deep value network [Gygli et al., 2017], fully convolutional networks [Long et al., 2015], constraint reasoning tool [Dragone et al., 2021, Jiang et al., 2022], multi-label box model [Patel et al., 2022], probabilistic deep networks [Zheng and Pronobis, 2019, Jang et al., 2023], autoregressive methods [Liu et al., 2022], nonlinear output transformations and embedding [Graber et al., 2018, Brodat-Motte et al., 2020], smoothing methods [Pillutla et al., 2018] and structural training [Choi et al., 2016, Ahmad et al., 2023]. Let us also mention ensemble algorithms for structured prediction algorithms [Cortes, Kuznetsov, and Mohri, 2014a,b], which can be used to combine several algorithms for this problem.

Consistency in structured prediction. Here, we discuss in detail previous work on consistency in structured prediction [Osokin et al., 2017, Ciliberto et al., 2016, Blondel, 2019, Nowak et al., 2020, 2022, Ciliberto et al., 2019, 2020, Nowak-Vila et al., 2019, Nowak et al., 2019, Cabannes et al., 2021, Cabannnes et al., 2020, Corro, 2023].

Osokin et al. [2017], Nowak et al. [2020] and Nowak et al. [2022] pointed out that the max-margin Markov networks (M3N), or more generally structural SVMs may not be Bayes-consistent. Instead, Osokin et al. [2017] proposed the first Bayes-consistent surrogate loss in the structured prediction setting, the quadratic surrogate (QS) loss. A general theory of QS was further developed in [Nowak-Vila et al., 2019, Nowak et al., 2019]. However, as pointed out in Section 1, the quadratic surrogate loss formulation casts the structured prediction problem as a regression problem and is not a typical formulation even in the binary classification case.

Nowak et al. [2020] proposed a consistent method called max-min margin Markov networks (M4N) derived from first principles for binary SVM. However, this method is restricted to SVM-type loss functions. Instead, we propose broad families of surrogate losses, which can be naturally derived from common multi-class losses including the logistic loss.

Nowak et al. [2022] addressed the inconsistency of Max-Margin loss in structured prediction by introducing the notion of Restricted-Max-Margin, where the maximization is performed over a subset of the original domain. Their method is based on an implicit embedding [Finocchiaro et al., 2019]; a general framework for structured prediction has been further developed by Ciliberto et al. [2020]. However, these methods are only applied to polyhedral-type surrogates which are not as smooth as the logistic loss. Thus, the resulting surrogate losses may not be favorable from the optimization point of view. Instead, our novel families of surrogate losses are very general and can be smooth, including a new structured logistic loss, for which we describe efficient gradient computation algorithms.

Ciliberto et al. [2016] focused on a least squares surrogate loss function and corresponding framework. In this framework, the structured prediction problem is cast as a regression problem. They derived a regularization approach to structured prediction from the least squares surrogate loss and proved the Bayes-consistency of that approach. Ciliberto et al. [2019] focused on a local structure-adapted framework for structured prediction. They proposed a novel structured prediction algorithm that adaptively leverages locality in the learning problem. Ciliberto et al. [2020] developed a general framework for structured prediction based on implicit embedding. Their methods lead to polyhedral-type surrogates losses that benefit from Bayes-consistency.

On the other hand, our work presents an extensive study of surrogate losses for structured prediction supported by β-consistency bounds. Different from the surrogate loss studied in the previous work,
the formulations of our proposed surrogate losses including structured comp-sum losses and structured constrained losses are completely novel and do not cast structured prediction problems as a regression problem. Furthermore, we prove stronger consistency guarantees that imply Bayes-consistency for these new proposed families of surrogate loss.

Other related work on structured prediction includes: projection-based losses for structured prediction [Blondel, 2019]; fast convergence rates for general structured prediction problems [Cabannes et al., 2021]; a unified framework for dealing with partial labelling [Cabannes et al., 2020]; and an analysis of the inconsistency of separable negative log-likelihood losses for structured prediction [Corro, 2023].

B Proof of Lemma 3

Lemma 3. The best-in-class conditional error and the conditional regret for a target loss \( L \) in structured prediction can be expressed as follows:

\[
\ell_{L,H}(x) = \min_{y' \in H(x)} \sum_{y \in Y} p(x,y) \ell(y', y)
\]

\[
\Delta \ell_{L,H}(h, x) = \sum_{y \neq y'} p(x,y) \ell(h(x), y) - \min_{y' \in H(x)} \sum_{y \in Y} p(x,y) \ell(y', y).
\]

Proof. By the definition, the conditional L-risk can be expressed as follows:

\[
\ell_{L}(h, x) = \sum_{y \in Y} p(x,y) \ell(h(x), y).
\] \hspace{1cm} (15)

Since \( \{ h(x) : h \in H \} = H(x) \), the best-in-class conditional error can be expressed as follows:

\[
\ell_{L,H}(x) = \min_{y' \in H(x)} \sum_{y \in Y} p(x,y) \ell(y', y),
\]

which proves the first part of the lemma. By the definition,

\[
\Delta \ell_{L,H}(h, x) = \ell_{L}(h, x) - \ell_{L,H}(x) = \sum_{y \neq y'} p(x,y) \ell(h(x), y) - \min_{y' \in H(x)} \sum_{y \in Y} p(x,y) \ell(y', y).
\]

\[\Box\]

C Proofs for structured max losses

Theorem 4 (Negative results of \( L_{\text{max}} \)). Assume that \( n > 2 \) and that \( \Phi_u(v) \) is convex and non-increasing for \( u = 1 \). Then, the max structured loss \( L_{\text{max}} \) is not Bayes-consistent.

Proof. For the structured max loss \( L_{\text{max}} \), the conditional \( L_{\text{max}} \)-risk can be expressed as follows:

\[
\ell_{L_{\text{max}}}(h, x) = \sum_{y,y'} p(x,y) \max_{y' \neq y} \Phi_1(h(x,y) - h(x,y')).
\]

Take \( \ell(y', y) = 1_{y \neq y'} \) to be the zero-one loss. Since \( \ell(y', y) = 1 \) for any \( y' \neq y' \), the conditional \( L_{\text{max}} \)-risk can be reformulated as follows:

\[
\ell_{L_{\text{max}}}(h, x) = \sum_{y \in Y} p(x,y) \max_{y' \neq y} \Phi_1(h(x,y) - h(x,y')).
\]

Consider the distribution that supports on a singleton domain \( \{ x \} \). Take \( y_1 \neq y_2 \in Y \) such that \( y_1 \neq n \) and \( y_2 \neq n \). We define the conditional distribution as \( p(x, y_1) = p(x, y_2) = \frac{1}{2} \) and \( p(x, y) = 0 \) for other \( y \in Y \). Then, by using the fact that \( \Phi_1(v) \) is convex and non-increasing, we obtain

\[
\mathcal{R}_{L_{\text{max}}}(h) = \ell_{L_{\text{max}}}(h, x) = \frac{1}{2} \max_{y' \neq y_1} \Phi_1(h(x, y_1) - h(x, y')) + \frac{1}{2} \max_{y' \neq y_2} \Phi_1(h(x, y_2) - h(x, y'))
\]

\[
= \frac{1}{2} \Phi_1(h(x, y_1) - \max_{y' \neq y_1} h(x, y')) + \frac{1}{2} \Phi_1(h(x, y_2) - \max_{y' \neq y_2} h(x, y'))
\]

\[
= \frac{1}{2} \Phi_1\left(h(x, y_1) - \frac{1}{2} \max_{y' \neq y_1} h(x, y') \right) + \frac{1}{2} \Phi_1\left(h(x, y_2) - \frac{1}{2} \max_{y' \neq y_2} h(x, y') \right)
\]

\[
\geq \frac{1}{2} \Phi_1\left(h(x, y_1) - \frac{1}{2} \max_{y' \neq y_1} h(x, y') \right) + \frac{1}{2} \Phi_1\left(h(x, y_2) - \frac{1}{2} \max_{y' \neq y_2} h(x, y') \right)
\]

\[
= \Phi_1(0) \quad (\Phi_1(v) \text{ is convex})
\]

\[
\geq \Phi_1(0) \quad (\Phi_1(v) \text{ is non-increasing})
\]

19
where the equality can be achieved by \( h^* \in \mathcal{H} \), defined as \( h^*(x, 1) = h^*(x, 2) = \ldots = h^*(x, n) \). Therefore, \( h^* \) is a Bayes classifier of the structured max loss. Note that \( h^*(x) = n \). However, by Lemma 3, in such a case, the Bayes classifier \( h^*_n \) of the target loss satisfies that

\[
h^*_n(x) = \arg\min_{y \in \mathcal{Y}} \sum_{y'} p(x, y) \ell(y', y) = \arg\min_{y \in \mathcal{Y}} (p(x, y_1) \mathbb{I}_{y' = y_1} + p(x, y_2) \mathbb{I}_{y' = y_2}) = y_1 \text{ or } y_2.
\]

Thus, we obtain \( h^* \neq h^*_n \). Therefore, \( L^{\text{max}} \) is not Bayes-consistent. \( \square \)

D Proofs for Voted Conditional Random Field

**Theorem 5 (Negative result of \( L^{\text{VCRF}} \)).** The Voted Conditional Random Field \( L^{\text{VCRF}} \) is not Bayes-consistent.

**Proof.** For the Voted Conditional Random Field \( L^{\text{VCRF}} \), the conditional \( L^{\text{VCRF}} \)-risk can be expressed as follows:

\[
\mathcal{C}_{L^{\text{VCRF}}}(h, x) = \sum_{y \in \mathcal{Y}} p(x, y) \log\left( \sum_{y' \in \mathcal{Y}} e^{\ell(h, y' - h(x, y))} \right).
\]

Consider the distribution that supports on a singleton domain \( \{x\} \). Note that \( \mathcal{R}_{L^{\text{VCRF}}} = \mathcal{C}_{L^{\text{VCRF}}} \) is convex with respect to \( h(x, y) \), \( y = 1, \ldots, n \). To find the global minimum, we will differentiate \( \mathcal{C}_{L^{\text{VCRF}}} \) with respect to \( h(x, y) \) for any \( y \in \mathcal{Y} \) and setting the derivatives to zero. Thus, we obtain for any \( y \in \mathcal{Y} \),

\[
p(x, y) = \sum_{y' \in \mathcal{Y}} p(x, y') \frac{e^{\ell(h, y' - h(x, y'))}}{\sum_{y' \in \mathcal{Y}} e^{\ell(h, y' - h(x, y'))}} = \frac{p(x, y') e^{\ell(h, y' - h(x, y'))}}{\sum_{y' \in \mathcal{Y}} e^{\ell(h, y' - h(x, y'))}}.
\]

Using the fact that \( \sum_{y' \in \mathcal{Y}} e^{\ell(h, y' - h(x, y'))} = \sum_{y' \in \mathcal{Y}} e^{\ell(h, y' - h(x, y))} - e^{\ell(h, y) - h(x, y) - h(x, y')} \), to further simplify the LHS of (16), we obtain for any \( y \in \mathcal{Y} \),

\[
p(x, y) = \sum_{y' \in \mathcal{Y}} p(x, y') \frac{e^{\ell(h, y' - h(x, y'))}}{\sum_{y' \in \mathcal{Y}} e^{\ell(h, y' - h(x, y'))}} = \frac{p(x, y') e^{\ell(h, y' - h(x, y'))}}{\sum_{y' \in \mathcal{Y}} e^{\ell(h, y' - h(x, y'))}}.
\]

Consider a target loss function \( L \) such that \( e^{\ell(h, y' - h(x, y))} = \Phi_{y, y'}^{\ell(h, y' - h(x, y))} \), that is \( \ell(h, y' - h(x, y)) = \log(\Phi_{y, y'}) + \log(\Phi_{y'}) \), where \( \Phi_{y} \) is a function mapping from \( \mathcal{Y} \) to \( \mathbb{R}_{+} \). For this special choice of the target loss function, the expression of \( \ell(h, y') \) decouple and (17) can be simplified to

\[
p(x, y) = \sum_{y' \in \mathcal{Y}} p(x, y') \frac{\Phi_{y, y'} e^{h(x, y')}}{\sum_{y' \in \mathcal{Y}} \Phi_{y, y'} e^{h(x, y')}} = \frac{\Phi_{y} e^{h(x, y)}}{\sum_{y' \in \mathcal{Y}} \Phi_{y'} e^{h(x, y')}}.
\]

Therefore, for the Bayes classifier \( h^* \) of Voted Conditional Random Field, by (18), we have

\[
\frac{\Phi_{y} e^{h^*(x, 1)}}{p(x, 1)} = \ldots = \frac{\Phi_{y} e^{h^*(x, n)}}{p(x, n)}
\]

which implies that

\[
h^*(x) = \arg\max_{y' \in \mathcal{Y}} h^*(x, y') = \arg\max_{y' \in \mathcal{Y}} e^{h^*(x, y')} = \arg\max_{y' \in \mathcal{Y}} \frac{p(x, y')}{\Phi_{y'}}.
\]

However, by Lemma 3, in such a case, the Bayes classifier \( h^*_n \) of the target loss satisfies that

\[
h^*_n(x) = \arg\min_{y' \in \mathcal{Y}} \sum_{y \in \mathcal{Y}} p(x, y) \ell(y', y) = \arg\min_{y' \in \mathcal{Y}} \sum_{y \in \mathcal{Y}} p(x, y) (\log(\Phi_{y}) + \log(\Phi_{y'})) = \arg\min_{y' \in \mathcal{Y}} \Phi_{y'}.
\]

Thus, we obtain \( h^* \neq h^*_n \) in general. Indeed, consider the case where \( p(x, y) = \frac{\Phi_{y}^{2}}{\sum_{k=1}^{n} \Phi_{y}^{2}}, y \in \mathcal{Y} \). Then, \( h^*(x) = \arg\max_{y' \in \mathcal{Y}} \sum_{y \in \mathcal{Y}} \frac{\Phi_{y}^{2}}{\sum_{k=1}^{n} \Phi_{y}^{2}} \Phi_{y'} = \arg\max_{y' \in \mathcal{Y}} \Phi_{y'} = h^*_n(x) \) when \( \{\Phi_{y}; y \in \mathcal{Y}\} \) are not equal. Therefore, \( L^{\text{VCRF}} \) is not Bayes-consistent. \( \square \)
E  Proofs for structured comp-sum losses

E.1  Structured logistic loss

Theorem 11 (H-consistency bound of $L_{\log}^\text{comp}$). Assume that $H$ is symmetric and complete. Then, for any target loss $\ell$, hypothesis $h \in H$ and any distribution, $R_{\log}(h) - R_{\log}^*(h) \leq 2 \left( R_{\log}^*(h) - R_{\log}^*(H) + M_{\log}^\text{comp} \right)^{\frac{1}{2}} - M_{\log}(h).$ (19)

Proof. For the comp-sum structured loss $L_{\log}^\text{comp}$, the conditional $L_{\log}^\text{comp}$-risk can be expressed as follows:

$$E_{\log}^\text{comp}(h, x) = - \sum_{y \in y} p(x, y) \sum_{y' \in y} \ell(y', y) \log \left( \frac{e^{h(x, y')}}{\sum_{y'' \in y'} e^{h(x, y'')}} \right)$$

$$= - \sum_{y \in y} \log \left( \sum_{y' \in y} e^{h(x, y')} \right) \sum_{y \in y} p(x, y) \ell(y', y)$$

$$= - \sum_{y \in y} \log(S(x, y')) \eta(x, y'),$$

where we denote by $\eta(x, y') = \sum_{y \in y} p(x, y) \ell(y', y) \in [0, 1]$ and $S(x, y) = \frac{e^{h(x, y)}}{\sum_{y' \in y} e^{h(x, y')}} \in [0, 1]$ with the constraint that $\sum_{y \in y} S(x, y) = 1$. Let $y_{\min} = \arg\min_{y \in y} \sum_{y \in y} p(x, y) \ell(y', y)$, where we choose the label with the highest index under the natural ordering of labels as the tie-breaking strategy. For any $h \in H$ such that $h(x) \neq y_{\min}$ and $x \in X$, by the symmetry and completeness of $H$, we can always find a family of hypotheses $\{\overline{T}_\mu : \mu \in [-S(x, y_{\min}), S(x, h(x))]\} \subset H$ such that $\overline{S}_\mu(x, \cdot) = \frac{e^{h(x, \cdot)}}{\sum_{y \in y} e^{h(x, y)}}$ take the following values:

$$\overline{S}_\mu(x, y) = \begin{cases} S(x, y) & \text{if } y \neq \{y_{\min}, h(x)\} \\ S(x, y_{\min}) + \mu & \text{if } y = h(x) \\ S(x, h(x)) - \mu & \text{if } y = y_{\min}. \end{cases}$$ (20)

Note that $\overline{S}_\mu$ satisfies the constraint:

$$\sum_{y \in y} \overline{S}_\mu(x, y) = \sum_{y \in y} S(x, y) = 1, \forall \mu \in [-S(x, y_{\min}), S(x, h(x))].$$

By (20) and using the fact that $H(x) = Y$ when $H$ is symmetric, we obtain

$$\Delta E_{\log}^\text{comp, H}(h, x) = E_{\log}^\text{comp}(h, x) - E_{\log}^\text{comp}(H(x), x)$$

$$\geq E_{\log}^\text{comp}(h, x) - \inf_{\mu \in \mathbb{R}} \left\{ \overline{\ell}(x, h(x)) \left[ -\log(S(x, y_{\min})) + \log(S(x, h(x)) - \mu) \right] \right. $$

$$\left. + \overline{\ell}(x, h(x)) \left[ -\log(S(x, h(x))) + \log(S(x, y_{\min}) + \mu) \right] \right\}.$$ 

Differentiating with respect to $\mu$ yields the optimal value $\mu^* = \frac{\overline{\ell}(x, h(x))S(x, h(x)) - \overline{\ell}(x, y_{\min})S(x, y_{\min})}{\overline{\ell}(x, y_{\min}) + \overline{\ell}(x, h(x))}$. Plugging in that value gives:

$$\Delta E_{\log}^\text{comp, H}(h, x) \geq \overline{\ell}(x, y_{\min}) \log \frac{S(x, h(x)) + S(x, y_{\min})}{S(x, y_{\min}) + \overline{\ell}(x, h(x))}$$

$$+ \overline{\ell}(x, h(x)) \log \frac{S(x, h(x)) + S(x, y_{\min})}{S(x, h(x)) + \overline{\ell}(x, h(x))}.$$ 

21
Differentiating with respect to $S$ shows that the minimum is attained for $S(x, h(x)) = S(x, y_{\text{min}})$, which gives:

$$\Delta_{L_{\log}} \leq \frac{1}{4} \Delta_{L_{\log}}(h, x)^2$$

(by Lemma 3 and $H(x) = y$)

Since the function $t \mapsto t^2$ is convex, by Jensen’s inequality, we obtain for any hypothesis $h \in \mathcal{H}$ and any distribution,

$$\frac{\left( R_L(h) - R^*_{\log} + M_{L_{\log}} \right)^2}{4} = \frac{(E_X[\Delta_{L_{\log}}(h, x)])^2}{4} \leq E_X\left[ \frac{\Delta_{L_{\log}}(h, x)^2}{4} \right] \leq E_X\left[ \Delta_{L_{\log}}(h, x) \right] = R_L(h) - R^*_{\log} + M_{L_{\log}}$$

which leads to

$$R_L(h) - R^*_{\log} \leq 2\left( R_{L_{\log}}(h) - R^*_{\log} + M_{L_{\log}} \right)^{\frac{1}{2}} - M_{L_{\log}}.$$  

\[ \square \]

\section*{E.2 Structured sum-exponential loss}

\textbf{Theorem 12 (}$\mathcal{H}$\textit{-consistency bound of }$L_{\text{comp}}$\textbf{). Assume that }$\mathcal{H}$\textit{ is symmetric and complete. Then, for any target loss }$\ell$\textit{, hypothesis }$h \in \mathcal{H}$\textit{ and any distribution,

$$R_L(h) - R^*_L \leq 2 \left( R_{L_{\text{comp}}}^L(h) - R_{L_{\text{comp}}}^* + M_{L_{\text{comp}}} \right)^{\frac{1}{2}} - M_{L_{\text{comp}}}.$$  

\textbf{Proof.} For the comp-sum structured loss $L_{\text{comp}}$, the conditional $L_{\text{comp}}$-risk can be expressed as follows:

$$\ell_{L_{\text{comp}}}^L(h, x) = \sum_{y \in Y} p(x, y) \sum_{y' \in Y} \ell(y', y) \sum_{y'' \in Y} e^{h(x, y'') - h(x, y')} = \sum_{y' \in Y} \left( \frac{1}{S(x, y')} - 1 \right) \ell(x, y'),$$

where we denote by $\tilde{\ell}(x, y') = \sum_{y \in Y} p(x, y) \ell(y', y) \in [0, 1]$ and $S(x, y) = \sum_{y'' \in Y} e^{h(x, y'')} \in [0, 1]$ with the constraint that $\sum_{y \in Y} S(x, y) = 1$. Let $y_{\text{min}} = \arg\min_{y \in Y} \sum_{y' \in Y} p(x, y) \ell(y', y)$, where we choose the label with the highest index under the natural ordering of labels as the tie-breaking strategy. For any $h \in \mathcal{H}$ such that $h(x) \neq y_{\text{min}}$ and $x \in X$, by the symmetry and completeness of $\mathcal{H}$, we can always find a family of hypotheses $\mathcal{H}_\mu = \{ \mu \in [-S(x, y_{\text{min}}), S(x, h(x))] \} \subset \mathcal{H}$ such that

$$\tilde{S}_\mu(x, y) = \begin{cases} S(x, y) & \text{if } y \notin \{y_{\text{min}}, h(x)\} \\ S(x, y_{\text{min}}) + \mu & \text{if } y = h(x) \\ S(x, h(x)) - \mu & \text{if } y = y_{\text{min}}. \end{cases}$$

(22)
Note that $\mathbb{S}_\mu$ satisfies the constraint:
\[
\sum_{y \in \mathcal{Y}} \mathbb{S}_{\mu}(x, y) = \sum_{y \in \mathcal{Y}} \mathcal{S}(x, y) = 1, \quad \forall \mu \in [-\mathcal{S}(x, y_{\text{min}}), \mathcal{S}(x, h(x))].
\]
By (22) and using the fact that $\mathcal{H}(x) = y$ when $\mathcal{H}$ is symmetric, we obtain
\[
\begin{align*}
\Delta \mathcal{C}_{\text{comp}, \mathcal{H}}(h, x) &= \mathcal{C}_{\text{comp}}(h, x) - \mathcal{C}_{\text{comp}}(\mathcal{H}, x) \\
&\geq \mathcal{C}_{\text{comp}}(h, x) - \inf_{\mu \in \mathcal{H}} \mathcal{C}_{\text{comp}}(\mathcal{H}_{\mu}, x) \\
&= \sup_{\mu \in [-\mathcal{S}(x, y_{\text{min}}), \mathcal{S}(x, h(x))] \setminus \mathcal{H}} \left\{ \mathcal{S}(x, y_{\text{min}}) \left( \frac{1}{\mathcal{S}(x, y_{\text{min}})} - \frac{1}{\mathcal{S}(x, h(x))} - \mu \right) + \mathcal{S}(x, h(x)) \left( \frac{1}{\mathcal{S}(x, y_{\text{min}})} - \frac{1}{\mathcal{S}(x, h(x))} + \mu \right) \right\} \\
&= \mathcal{S}(x, y_{\text{min}}) + \mathcal{S}(x, h(x)) \left( \frac{\sqrt{\mathcal{S}(x, y_{\text{min}})} + \sqrt{\mathcal{S}(x, h(x))}}{\mathcal{S}(x, y_{\text{min}}) + \mathcal{S}(x, h(x))} \right)^2 \\
&\geq \left( \sqrt{\mathcal{S}(x, y_{\text{min}})} - \sqrt{\mathcal{S}(x, h(x))} \right)^2 \\
&\geq \frac{\left( \sqrt{\mathcal{S}(x, h(x))} - \sqrt{\mathcal{S}(x, y_{\text{min}})} \right)^2}{\left( \sqrt{\mathcal{S}(x, h(x))} + \sqrt{\mathcal{S}(x, y_{\text{min}})} \right)^2} \\
&\geq \frac{\left( \sqrt{\mathcal{S}(x, h(x))} - \sqrt{\mathcal{S}(x, y_{\text{min}})} \right)^2}{4} (\sqrt{a} + \sqrt{b} \leq 2, \forall a, b \in [0, 1], a + b \leq 2) \\
&= \frac{1}{4} \Delta \mathcal{C}_{\text{comp}, \mathcal{H}}(h, x)^2. \\
\end{align*}
\]
Since the function $t \mapsto t^2$ is convex, by Jensen’s inequality, we obtain for any hypothesis $h \in \mathcal{H}$ and any distribution,
\[
\begin{align*}
\frac{\left( \mathcal{R}_h(h) - \mathcal{R}_{\text{comp}}^*(h) + \mathcal{M}_{\text{comp}, \mathcal{H}} \right)^2}{4} &= \frac{(\mathbb{E}[\Delta \mathcal{C}_{\text{comp}, \mathcal{H}}(h, x)])^2}{4} \\
&\leq \frac{\mathbb{E}[\Delta \mathcal{C}_{\text{comp}, \mathcal{H}}(h, x)^2]}{4} \\
&\leq \frac{\mathbb{E}[\Delta \mathcal{C}_{\text{comp}, \mathcal{H}}(h, x)]}{4} \\
&= \mathcal{R}_{\text{comp}}(h) - \mathcal{R}_{\text{comp}}^*(h) + \mathcal{M}_{\text{comp}, \mathcal{H}}, \\
\end{align*}
\]
which leads to
\[
\mathcal{R}_h(h) - \mathcal{R}_{\text{comp}}^* \leq 2 \left( \mathcal{R}_{\text{comp}}(h) - \mathcal{R}_{\text{comp}}^*(h) + \mathcal{M}_{\text{comp}, \mathcal{H}} \right)^{\frac{1}{2}} - \mathcal{M}_{\text{comp}, \mathcal{H}}.
\]

\section{E.3 Structured generalized cross-entropy loss}

\textbf{Theorem 13 ($\mathcal{H}$-consistency bound of $\mathcal{L}_{\text{comp}}^*$).} Assume that $\mathcal{H}$ is symmetric and complete. Then, for any target loss $\ell$, hypothesis $h \in \mathcal{H}$ and any distribution,
\[
\begin{align*}
\mathcal{R}_h(h) - \mathcal{R}_{\text{comp}}^* &\leq 2n^2 \left( \mathcal{R}_{\text{comp}}(h) - \mathcal{R}_{\text{comp}}^*(h) + \mathcal{M}_{\text{comp}, \mathcal{H}} \right)^{\frac{1}{2}} - \mathcal{M}_{\mathcal{H}}. \\
\end{align*}
\]
Proof. For the comp-sum structured loss \( L_{\text{comp}} \), the conditional \( L_{\text{comp}} \)-risk can be expressed as follows:

\[
\begin{align*}
\mathcal{L}_{\text{comp}}(h, x) &= \sum_{y \in Y} p(x, y) \sum_{y' \in Y} \tilde{t}(y', y) \frac{1}{\alpha} \left(1 - \left(\frac{e^{h(x, y')}}{\sum_{y'' \in Y} e^{h(x, y'')}}\right)\right) \\
&= \frac{1}{\alpha} \sum_{y' \in Y} \left(1 - \left(\frac{e^{h(x, y')}}{\sum_{y'' \in Y} e^{h(x, y'')}}\right)\right) \sum_{y \in Y} p(x, y) \tilde{t}(y', y) \\
&= \frac{1}{\alpha} \sum_{y' \in Y} (1 - \mathcal{S}(x, y')) \mathcal{S}(x, y'),
\end{align*}
\]

where we denote by \( \mathcal{S}(x, y') = \sum_{y \in Y} p(x, y) \tilde{t}(y', y) \mathcal{S}_{\text{comp}}(x, y') \) with the constraint that \( \sum_{y \in Y} \mathcal{S}(x, y) = 1 \). Let \( y_{\min} = \arg\min_{y \in Y} \mathcal{S}(x, y) \) be any strategy that satisfies the constraint. For any \( h \in \mathcal{H} \) such that \( h(x) = y_{\min} \) and \( x \in X \), by the symmetry and completeness of \( \mathcal{H} \), we can always find a family of hypotheses \( \{ h_{\mu} : \mu \in [-\mathcal{S}(x, y_{\min}), \mathcal{S}(x, h(x))] \} \subset \mathcal{H} \) such that \( \mathcal{S}_{\mu}(x, \cdot) = \frac{e^{h_{\mu}(x, \cdot)}}{\sum_{y \in Y} e^{h_{\mu}(x, y)}} \) take the following values:

\[
\mathcal{S}_{\mu}(x, y) = \begin{cases} 
\mathcal{S}(x, y) & \text{if } y \neq \{ y_{\min}, h(x) \} \\
\mathcal{S}(x, y_{\min}) + \mu & \text{if } y = h(x) \\
\mathcal{S}(x, h(x)) - \mu & \text{if } y = y_{\min}.
\end{cases}
\] (24)

Note that \( \mathcal{S}_{\mu} \) satisfies the constraint:

\[
\sum_{y \in Y} \mathcal{S}_{\mu}(x, y) = \sum_{y \in Y} \mathcal{S}(x, y) = 1, \forall \mu \in [-\mathcal{S}(x, y_{\min}), \mathcal{S}(x, h(x))].
\]

By (24) and using the fact that \( \mathcal{H}(x) = \emptyset \) when \( \mathcal{H} \) is symmetric, we obtain

\[
\begin{align*}
\Delta e_{\text{comp}, \mathcal{H}}(h, x) &= \mathcal{L}_{\text{comp}}(h, x) - \mathcal{L}_{\text{comp}}(\mathcal{H}, x) \\
&\geq \mathcal{L}_{\text{comp}}(h, x) - \inf_{\mu \in \mathcal{H}} \mathcal{L}_{\text{comp}}(h_{\mu}, x) \\
&= \frac{1}{\alpha} \sup_{\mu \in [-\mathcal{S}(x, y_{\min}), \mathcal{S}(x, h(x))]} \left\{ \mathcal{S}(x, y_{\min}) \left[ -\mathcal{S}(x, y_{\min})^\alpha + (\mathcal{S}(x, h(x)) - \mu)^\alpha \right] \\
&+ \mathcal{S}(x, h(x)) \left[ -\mathcal{S}(x, h(x))^\alpha + (\mathcal{S}(x, y_{\min}) + \mu)^\alpha \right] \right\} \\
&= \frac{1}{\alpha} \left( \mathcal{S}(x, h(x)) + (\mathcal{S}(x, y_{\min}) - \mathcal{S}(x, h(x)))^\alpha \left[ \frac{1}{\alpha \mathcal{S}(x, y_{\min})} + \frac{1}{\alpha \mathcal{S}(x, h(x))} \right] \right)^{1-\alpha} \\
&- \frac{1}{\alpha} \mathcal{S}(x, y_{\min})^\alpha - \frac{1}{\alpha} \mathcal{S}(x, h(x))^\alpha \\
&\text{(differentiating with respect to } \mu \text{ to optimize, optimum } \mu^* = \frac{\mathcal{S}(x, h(x))}{\mathcal{S}(x, y_{\min})} \right) \\
&\geq \frac{1}{\alpha \mathcal{S}(x, y_{\min})^\alpha} \left[ 2^\alpha \left( \frac{\mathcal{S}(x, y_{\min})}{\mathcal{S}(x, h(x))} + \frac{\mathcal{S}(x, h(x))}{\mathcal{S}(x, y_{\min})} \right)^{1-\alpha} - \mathcal{S}(x, y_{\min}) - \mathcal{S}(x, h(x)) \right] \\
&\geq \frac{\left( \mathcal{S}(x, h(x)) + \mathcal{S}(x, y_{\min}) \right)^2}{4\alpha} \left( \frac{a+b}{2} \right)^{1-\alpha} - \frac{a+b}{2} \geq \frac{a-b}{4}^2 \forall a, b \in [0, 1], 0 \leq a + b \leq 1 \\
&= \frac{(\sum_{y \in Y} p(x, y) \ell(h(x), y) - \sum_{y \in Y} p(x, y) \ell(y_{\min}, y))^2}{4\alpha^2} \quad \text{(by Lemma 3 and } \mathcal{H}(x) = \emptyset) \\
&= \frac{\Delta e_{\text{comp}, \mathcal{H}}(h, x)^2}{4\alpha^2}.
\end{align*}
\]
Since the function $t \mapsto \frac{t^2}{4n^\alpha}$ is convex, by Jensen’s inequality, we obtain for any hypothesis $h \in \mathcal{H}$ and any distribution,

$$\frac{(R_L(h) - R_L^* + M_L,\mathcal{H})^2}{4n^\alpha} = \frac{(E_X[\Delta e_{L,\mathcal{H}}(h,x)])^2}{4n^\alpha} \leq \frac{E_X[\Delta e_{L,\mathcal{H}}(h,x)]}{4n^\alpha} \leq \frac{E_X[\Delta e_{\text{comp},\mathcal{H}}(h,x)]}{4n^\alpha} = R_{L,\text{comp}}(h) - R_{L,\text{comp}}^* + M_{L,\text{comp},\mathcal{H}},$$

which leads to

$$R_L(h) - R_L^*_{L,\mathcal{H}} \leq 2n^\frac{\alpha}{2} \left( R_{L,\text{comp}}(h) - R_{L,\text{comp}}^* + M_{L,\text{comp},\mathcal{H}} \right) - M_{L,\mathcal{H}}. \quad \square$$

### E.4 Structured mean absolute error loss

**Theorem 14 (H-consistency bound of $L_{\text{mae}}^{\text{comp}}$).** Assume that $\mathcal{H}$ is symmetric and complete. Then, for any target loss $\ell$, hypothesis $h \in \mathcal{H}$ and any distribution,

$$R_L(h) - R_L^*_{L,\mathcal{H}} \leq n \left( R_{L,\text{mae}}^{\text{comp}}(h) - R_{L,\text{mae}}^* + M_{L,\text{mae},\mathcal{H}} \right) - M_{L,\mathcal{H}}. \quad (25)$$

**Proof.** For the comp-sum structured loss $L_{\text{mae}}^{\text{comp}}$, the conditional $L_{\text{mae}}^{\text{comp}}$-risk can be expressed as follows:

$$e_{L_{\text{mae}}^{\text{comp}}}(h,x) = \sum_{y \in \mathcal{Y}} p(x,y) \sum_{y' \in \mathcal{Y}} \ell(y',y) \left( 1 - \frac{e^{h(x,y')}}{\sum_{y'' \in \mathcal{Y}} e^{h(x,y'')}} \right)$$

$$= \sum_{y' \in \mathcal{Y}} \left( 1 - \frac{e^{h(x,y')}}{\sum_{y'' \in \mathcal{Y}} e^{h(x,y'')}} \right) \sum_{y \in \mathcal{Y}} p(x,y) \ell(y',y)$$

$$= \sum_{y' \in \mathcal{Y}} (1 - S(x,y')) \overline{\ell}(x,y'),$$

where we denote by $\overline{\ell}(x,y') = \sum_{y \in \mathcal{Y}} p(x,y) \ell(y',y) \in [0,1]$ and $S(x,y) = \frac{e^{h(x,y)}}{\sum_{y'' \in \mathcal{Y}} e^{h(x,y'')}} \in [0,1]$ with the constraint that $\sum_{y \in \mathcal{Y}} S(x,y) = 1$. Let $y_{\min} = \arg\min_{y \in \mathcal{Y}} \sum_{y' \in \mathcal{Y}} p(x,y) \ell(y',y')$, where we choose the label with the highest index under the natural ordering of labels as the tie-breaking strategy. For any $h \in \mathcal{H}$ such that $h(x) \neq y_{\min}$ and $x \in \mathcal{X}$, by the symmetry and completeness of $\mathcal{H}$, we can always find a family of hypotheses $\{\overline{\mathcal{H}}_\mu : \mu \in [-S(x,y_{\min}), S(x,h(x))]\} \subset \mathcal{H}$ such that $\overline{\mathcal{H}}_\mu(x) = \frac{e^{h(x,y)}}{\sum_{y'' \in \mathcal{Y}} e^{h(x,y'')}}$ take the following values:

$$\overline{\mathcal{H}}_\mu(x,y) = \begin{cases} S(x,y) & \text{if } y \notin \{y_{\min}, h(x)\} \\ S(x,y_{\min}) + \mu & \text{if } y = h(x) \\ S(x,h(x)) - \mu & \text{if } y = y_{\min}. \end{cases} \quad (26)$$

Note that $\overline{\mathcal{H}}_\mu$ satisfies the constraint:

$$\sum_{y \in \mathcal{Y}} \overline{\mathcal{H}}_\mu(x,y) = \sum_{y \in \mathcal{Y}} S(x,y) = 1, \forall \mu \in [-S(x,y_{\min}), S(x,h(x))].$$
By (26) and using the fact that \( H(x) = y \) when \( \mathcal{H} \) is symmetric, we obtain
\[
\Delta \mathcal{E}_{\text{cstnd}}(h, x) = \mathcal{E}_{\text{cstnd}}(h, x) - \mathcal{E}_{\text{cstnd}}(h, x)
\geq \mathcal{E}_{\text{cstnd}}(h, x) - \inf_{\mu \in \mathbb{R}} \mathcal{E}_{\text{cstnd}}(\overline{h}_{\mu}, x)
\]
\[
= \sup_{\mu \in [-S(x, y_{\min}), S(x, h(x))]} \left\{ \mathcal{I}(x, y_{\min})[-S(x, y_{\min}) + S(x, h(x)) - \mu] + \mathcal{I}(x, h(x))[S(x, h(x)) + \mu] \right\}
\]
\[
= \mathcal{I}(x, y_{\min})S(x, h(x)) - \mathcal{I}(x, h(x))S(x, h(x))
\]
(differentiating with respect to \( \mu \) to optimize, optimum \( \mu^\ast = -S(x, y_{\min}) \))
\[
\geq \frac{1}{n} (\mathcal{I}(x, y_{\min}) - \mathcal{I}(x, h(x)))
\]
(differentiating with respect to \( S \) to minimize, minimum is attained when \( S(x, h(x)) = \frac{1}{n} \))
\[
= \frac{\sum_{y \in Y} p(x, y)\ell(h(x), y) - \sum_{y \in Y} p(x, y)\ell(y_{\min}, y)}{n}
= \frac{\Delta \mathcal{E}_{\text{cstnd}}(h, x)}{n}
\]
(by Lemma 3 and \( H(x) = y \))

Therefore, we obtain for any hypothesis \( h \in \mathcal{H} \) and any distribution,
\[
\mathcal{R}_L(h) - \mathcal{R}_{\text{cstnd}}^\ast + M_{\mathcal{H}^\ast} \leq \frac{\mathbb{E}_X[\Delta \mathcal{E}_{\text{cstnd}}(h, x)]}{n}
\]
\[
= \frac{\mathbb{E}_X[\Delta \mathcal{E}_{\text{cstnd}}(h, x)]}{n}
= \mathcal{R}_{\text{cstnd}}(h) - \mathcal{R}_{\text{cstnd}}^\ast + M_{\mathcal{H}^\ast},
\]
which leads to
\[
\mathcal{R}_L(h) - \mathcal{R}_{\text{cstnd}}^\ast \leq n(\mathcal{R}_{\text{cstnd}}(h) - \mathcal{R}_{\text{cstnd}}^\ast + M_{\mathcal{H}^\ast} - M_{\mathcal{H}}).
\]

\( \square \)

F Proofs for structured constrained losses

F.1 Structured constrained exponential loss

**Theorem 15 (\( \mathcal{H} \)-consistency bound of \( L_{\text{cstnd}} \)).** Assume that \( \mathcal{H} \) is symmetric and complete. Then, for any target loss \( \ell \), hypothesis \( h \in \mathcal{H} \) and any distribution,
\[
\mathcal{R}_L(h) - \mathcal{R}_{\text{cstnd}}^\ast \leq 2\sqrt{\ell_{\max}} \left( \mathcal{R}_{L_{\text{cstnd}}}(h) - \mathcal{R}_{L_{\text{cstnd}}^\ast} + M_{L_{\text{cstnd}}} \right)^{\frac{1}{2}} - M_{\mathcal{H}}.
\]

**Proof.** Denote by \( q(x, y') = \sum_{y \in Y} p(x, y)\ell(y', y) \in [0, \ell_{\max}] \). For the constrained structured loss \( L_{\text{cstnd}} \), the conditional \( L_{\text{cstnd}} \)-risk can be expressed as follows:
\[
\mathcal{E}_{L_{\text{cstnd}}}(h, x) = \sum_{y \in Y} p(x, y) \ell(y, y')e^{h(x, y')} = \sum_{y' \in Y} e^{h(x, y')/q(x, y')},
\]
Let \( y_{\min} = \arg\max_{y' \in Y} \sum_{y \in Y} p(x, y)\ell(y', y) \), where we choose the label with the highest index under the natural ordering of labels as the tie-breaking strategy. For any \( h \in \mathcal{H} \) such that \( h(x) \neq y_{\min} \) and \( x \in X \), by the symmetry and completeness of \( \mathcal{H} \), we can always find a family of hypotheses \( \{\overline{h}_{\mu} : \mu \in \mathbb{R}\} \subset \mathcal{H} \) such that \( \overline{h}_{\mu}(x, y) \) take the following values:
\[
\overline{h}_{\mu}(x, y) = \begin{cases} 
  h(x, y) & \text{if } y \notin \{y_{\min}, h(x)\} \\
  h(x, y_{\min}) + \mu & \text{if } y = h(x) \\
  h(x, h(x)) - \mu & \text{if } y = y_{\min}.
\end{cases}
\]

(28)
Note that the hypotheses $\tilde{h}_\mu$ satisfies the constraint:
\[
\sum_{y \in Y} \tilde{h}_\mu(x, y) = \sum_{y \in Y} h(x, y) = 0, \forall \mu \in \mathbb{R}.
\]

Since $\sum_{y \in Y} h(x, y) = 0$, there must be non-negative scores. By definition of $h(x)$ as a maximizer, we must thus have $h(x, h(x)) \geq 0$. By (28) and using the fact that $H(x) = y$ when $\mathcal{H}$ is symmetric, we obtain

\[
\Delta \mathcal{E}_{\text{L-cstnd}}(h, x) = \mathcal{E}_{\text{L-cstnd}}(h, x) - \mathcal{E}_{\text{L-cstnd}}^*(\mathcal{H}, x) \geq \mathcal{E}_{\text{L-cstnd}}(h, x) - \inf_{\mu \in \mathbb{R}} \mathcal{E}_{\text{L-cstnd}}(\tilde{h}_\mu, x)
\]

\[
= \sup_{\mu \in \mathbb{R}} \left\{ q(x, y_{\min}) \left( e^{h(x,y_{\min})} - e^{h(x,h(x)) - \mu} + q(x, h(x)) \left( e^{h(x,h(x))} - e^{h(x,y_{\min}) + \mu} \right) \right\}
\]

\[
= \left( \sqrt{q(x, h(x))} e^{h(x,h(x))} - \sqrt{q(x, y_{\min})} e^{h(x,y_{\min})} \right)^2
\]

(differentiating with respect to $\mu$ to optimize, optimum $\mu^* = \frac{1}{2} \log \frac{q(x,y_{\min}) e^{h(x,h(x))}}{q(x,h(x)) e^{h(x,y_{\min})}}$

\[
\geq e^{h(x,h(x))} \left( \sqrt{q(x, y_{\min})} - \sqrt{q(x, h(x))} \right)^2 \quad \text{if } (h(x,h(x)) \geq 0)
\]

\[
\geq \left( \sqrt{q(x, y_{\min})} - \sqrt{q(x, h(x))} \right)^2 \quad \text{if } (h(x,h(x)) \geq 0)
\]

\[
= \frac{1}{4\ell_{\max}} \left( q(x, h(x)) - q(x, y_{\min}) \right)^2 \quad \text{if } (0 \leq q(x, y) \leq \ell_{\max})
\]

\[
= \frac{1}{4\ell_{\max}} \Delta \mathcal{E}_{\text{L-cstnd}}(h, x)^2.
\]

(by Lemma 3 and $H(x) = y$)

Since the function $t \mapsto \frac{t^2}{4\ell_{\max}}$ is convex, by Jensen’s inequality, we obtain for any hypothesis $h \in \mathcal{H}$ and any distribution,

\[
\frac{(R_L(h) - R^*_L, \mathcal{H} + M_L, \mathcal{H})^2}{4\ell_{\max}} = \frac{(\mathbb{E}_X [\Delta \mathcal{E}_{\text{L-cstnd}}(h, x)]^2}{4\ell_{\max}} \leq \mathbb{E}_X \left[ \frac{\Delta \mathcal{E}_{\text{L-cstnd}}(h, x)^2}{4\ell_{\max}} \right] \leq \mathbb{E}_X [\Delta \mathcal{E}_{\text{L-cstnd}}(h, x)] = R_{\text{L-cstnd}}(h) - R^*_{\text{L-cstnd}}, \mathcal{H} + M_{\text{L-cstnd}}, \mathcal{H},
\]

which leads to

\[
R_L(h) - R^*_L, \mathcal{H} \leq 2\sqrt{\ell_{\max}} \left( R_{\text{L-cstnd}}(h) - R^*_{\text{L-cstnd}}, \mathcal{H} + M_{\text{L-cstnd}}, \mathcal{H} \right) \frac{1}{2} - M_L, \mathcal{H}.
\]

\[\square\]

**F.2 Structured constrained squared-hinge loss**

**Theorem 16** ($\mathcal{H}$-consistency bound of $\mathcal{L}^\text{cstnd}_{\text{sq-hinge}}$). Assume that $\mathcal{H}$ is symmetric and complete. Then, for any target loss $\ell$, hypothesis $h \in \mathcal{H}$ and any distribution,

\[
R_L(h) - R^*_L, \mathcal{H} \leq \left( R_{\text{L-cstnd}}^{\text{sq-hinge}}(h) - R^*_{\text{L-cstnd}}^{\text{sq-hinge}}, \mathcal{H} + M_{\text{L-cstnd}}^{\text{sq-hinge}}, \mathcal{H} \right) \frac{1}{2} - M_L, \mathcal{H}.
\]
Proof. Denote by \( q(x, y') = \sum_{y \in Y} p(x, y) \ell(y', y) \in [0, \ell_{\max}] \). For the constrained structured loss \( \mathbb{L}_{\text{est}}^{\text{sq}-\text{hinge}} \), the conditional \( \mathbb{L}_{\text{est}}^{\text{sq}-\text{hinge}} \)-risk can be expressed as follows:

\[
\mathcal{E}_{\text{est}}^{\text{sq}-\text{hinge}}(h, x) = \sum_{y \in Y} p(x, y) \sum_{y' \in Y} \ell(y', y) \max\{0, 1 + h(x, y')\}^2
= \sum_{y' \in Y} \max\{0, 1 + h(x, y')\}^2 q(x, y').
\]

Let \( y_{\min} = \arg\min_{y \in Y} \sum_{y' \in Y} p(x, y) \ell(y', y) \), where we choose the label with the highest index under the natural ordering of labels as the tie-breaking strategy. For any \( h \in \mathcal{H} \) such that \( h(x) \neq y_{\min} \) and \( x \in \mathcal{X} \), by the symmetry and completeness of \( \mathcal{H} \), we can always find a family of hypotheses \( \{\overline{h}_\mu : \mu \in \mathbb{R}\} \subset \mathcal{H} \) such that \( \overline{h}_\mu(x, \cdot) \) take the following values:

\[
\overline{h}_\mu(x, y) = \begin{cases} 
  h(x, y) & \text{if } y \neq \{y_{\min}, h(x)\} \\
  h(x, y_{\min}) + \mu & \text{if } y = h(x) \\
  h(x, h(x)) - \mu & \text{if } y = y_{\min}.
\end{cases}
\] (30)

Note that the hypotheses \( \overline{h}_\mu \) satisfies the constraint:

\[
\sum_{y \in Y} \overline{h}_\mu(x, y) = \sum_{y \in Y} h(x, y) = 0, \forall \mu \in \mathbb{R}.
\]

Since \( \sum_{y \in Y} h(x, y) = 0 \), there must be non-negative scores. By definition of \( h(x) \) as a maximizer, we must thus have \( h(x, h(x)) \geq 0 \). By (30) and using the fact that \( H(x) = 1 \) when \( \mathcal{H} \) is symmetric, we obtain

\[
\begin{align*}
\Delta \mathcal{E}_{\text{est}}^{\mathcal{H}}(h, x) &= \mathcal{E}_{\text{est}}^{\mathcal{H}}(h, x) - \mathcal{E}_{\text{est}}^{\mathcal{H}}(\mathcal{H}, x) \\
&\geq \mathcal{E}_{\text{est}}^{\mathcal{H}}(h, x) - \inf_{\mu \in \mathbb{R}} \mathcal{E}_{\text{est}}^{\mathcal{H}}(\overline{h}_\mu, x) \\
&= \sup_{\mu \in \mathbb{R}} \left\{ q(x, y_{\min}) \left( \max\{0, 1 + h(x, y_{\min})\}^2 - \max\{0, 1 + h(x, h(x)) - \mu\}^2 \right) \\
&\quad + q(x, h(x)) \left( \max\{0, 1 + h(x, h(x))\}^2 - \max\{0, 1 + h(x, y_{\min}) + \mu\}^2 \right) \right\} \\
&\geq \left( 1 + h(x, h(x)) \right)^2 (q(x, y_{\min}) - q(x, h(x)))^2 \quad \text{(differentiating with respect to } \mu \text{ to optimize)} \\
&\geq \left( q(x, h(x)) - q(x, y_{\min}) \right)^2 \\
&= \left( \sum_{y \in Y} p(x, y) \ell(h(x), y) - \sum_{y \in Y} p(x, y) \ell(y_{\min}, y) \right)^2 \\
&= \Delta \mathcal{E}_{\mathcal{H}}(h, x)^2. \\
\end{align*}
\]

Since the function \( t \mapsto t^2 \) is convex, by Jensen’s inequality, we obtain for any hypothesis \( h \in \mathcal{H} \) and any distribution,

\[
(\mathcal{R}_{\mathcal{H}}(h) - \mathcal{R}^*_{\mathcal{H}, \mathcal{H}} + \mathcal{M}_{\mathcal{H}})^2 = \left( \mathbb{E}_X[\Delta \mathcal{E}_{\mathcal{H}}(h, x)] \right)^2 \\
\leq \mathbb{E}_X[\Delta \mathcal{E}_{\mathcal{H}}(h, x)^2] \\
\leq \mathbb{E}_X[\Delta \mathcal{E}_{\text{est}}^{\mathcal{H}}(h, x)] \\
= \mathcal{R}_{\text{est}}^{\mathcal{H}}(h) - \mathcal{R}^*_{\text{est}}^{\mathcal{H}, \mathcal{H}} + \mathcal{M}_{\text{est}}^{\mathcal{H}},
\]

which leads to

\[
\mathcal{R}_{\mathcal{H}}(h) - \mathcal{R}^*_{\mathcal{H}, \mathcal{H}} \leq \left( \mathcal{R}_{\text{est}}^{\text{sq}-\text{hinge}}(h) - \mathcal{R}^*_{\text{est}}^{\text{sq}-\text{hinge}, \mathcal{H}} + \mathcal{M}_{\text{est}}^{\text{sq}-\text{hinge}, \mathcal{H}} \right)^{\frac{1}{2}} - \mathcal{M}_{\mathcal{H}}.
\]
F.3 Structured constrained hinge loss

**Theorem 17 (H-consistency bound of L\textsuperscript{cstnd}.)** Assume that H is symmetric and complete. Then, for any target loss ℓ, hypothesis h ∈ H and any distribution,

\[
\mathcal{R}_L(h) - \mathcal{R}_L^* \leq \mathcal{R}_{\text{L cstnd}}(h) - \mathcal{R}_{\text{L cstnd}}^* + M_{\text{L cstnd}} - M_L.
\]

*Proof.* Denote by \( q(x, y') = \sum y \cdot p(x, y) \ell(y', y) \in [0, \ell_{\text{max}}] \). For the constrained structured loss \( \text{L cstnd} \) hinge, the conditional \( \text{L cstnd} \) hinge risk can be expressed as follows:

\[
\mathcal{E}_{\text{L cstnd}}^\text{hinge}(h, x) = \sum_{y \in Y} p(x, y) \sum_{y' \in Y} \ell(y', y) \max(0, 1 + h(x, y'))
= \sum_{y' \in Y} \max(0, 1 + h(x, y')) q(x, y').
\]

Let \( y_{\text{min}} = \arg\min_{y' \in Y} \sum_{y \in Y} p(x, y) \ell(y', y), \) where we choose the label with the highest index under the natural ordering of labels as the tie-breaking strategy. For any \( h \in H \) such that \( h(x) \neq y_{\text{min}} \) and \( x \in \mathcal{X} \), by the symmetry and completeness of \( H \), we can always find a family of hypotheses \( \{ h_\mu : \mu \in \mathbb{R} \} \subset H \) such that \( h_\mu(x, \cdot) \) take the following values:

\[
\mathcal{E}_{\text{L cstnd}}^{\text{hinge}}(h, x) = \begin{cases} 
 h(x, y) & \text{if } y \notin \{ y_{\text{min}}, h(x) \} \\
 h(x, y_{\text{min}}) + \mu & \text{if } y = h(x) \\
 h(x, h(x)) - \mu & \text{if } y = y_{\text{min}}.
\end{cases}
\]

Let that the hypotheses \( h_\mu \) satisfies the constraint:

\[
\sum_{y \in Y} h_\mu(x, y) = \sum_{y \in Y} h(x, y) = 0, \forall \mu \in \mathbb{R}.
\]

Since \( \sum_{y \in Y} h(x, y) = 0 \), there must be non-negative scores. By definition of \( h(x) \) as a maximizer, we must thus have \( h(x, h(x)) \geq 0 \). By (32) and using the fact that \( H(x) = y \) when \( H \) is symmetric, we obtain

\[
\Delta \mathcal{E}_{\text{L cstnd}}^{\text{hinge}}(h, x) \\
= \mathcal{E}_{\text{L cstnd}}^{\text{hinge}}(h, x) - \mathcal{E}_{\text{L cstnd}}^{\text{hinge}}(H, x) \\
\geq \mathcal{E}_{\text{L cstnd}}^{\text{hinge}}(h, x) - \inf_{\mu \in \mathbb{R}} \mathcal{E}_{\text{L cstnd}}^{\text{hinge}}(h_\mu, x) \\
= \sup_{\mu \in \mathbb{R}} \left\{ q(x, y_{\text{min}})(\max\{0, 1 + h(x, y_{\text{min}})\} - \max\{0, 1 + h(x, h(x)) - \mu\}) \\
+ q(x, h(x))(\max\{0, 1 + h(x, h(x))\} - \max\{0, 1 + h(x, y_{\text{min}}) + \mu\}) \right\} \\
\geq (1 + h(x, h(x)))(q(x, h(x)) - q(x, y_{\text{min}})) \quad (\text{differentiating with respect to } \mu \text{ to optimize}) \\\n\geq q(x, h(x)) - q(x, y_{\text{min}}) \quad (h(x, h(x)) \geq 0) \\
= \sum_{y \in Y} p(x, y) \ell(h(x), y) - \sum_{y \in Y} p(x, y) \ell(y_{\text{min}}, y) \\
= \Delta \mathcal{E}_{\text{L cstnd}}^{\text{hinge}}(h, x).
\]

Therefore, we obtain for any hypothesis \( h \in H \) and any distribution,

\[
\mathcal{R}_L(h) - \mathcal{R}_L^* \leq \mathcal{R}_{\text{L cstnd}}^{\text{hinge}}(h) - \mathcal{R}_{\text{L cstnd}}^* + M_{\text{L cstnd}} - M_L,
\]

which leads to

\[
\mathcal{R}_L(h) - \mathcal{R}_L^* \leq \mathcal{R}_{\text{L cstnd}}^{\text{hinge}}(h) - \mathcal{R}_{\text{L cstnd}}^* + M_{\text{L cstnd}} - M_L.
\]

\( \blacksquare \)
F.4 Structured constrained ρ-margin loss

**Theorem 18 (H-consistency bound of \(L_{\text{cstnd}}^\rho\)).** Assume that \(\mathcal{H}\) is symmetric and complete. Then, for any target loss \(\ell\), hypothesis \(h \in \mathcal{H}\) and any distribution,

\[
\mathcal{R}_L(h) - \mathcal{R}_L^*(\mathcal{H}) \leq \mathcal{R}_{L_{\text{cstnd}}}^*(h) - \mathcal{R}_{L_{\text{cstnd}}}^*(\mathcal{H}) + M_{L_{\text{cstnd}}}(\mathcal{H}) - M_{L_{\text{cstnd}}}(\mathcal{H}).
\]

**Proof.** Denote by \(q(x, y') = \sum_{y \in \mathcal{Y}} p(x, y) \ell(y', y) \in [0, \ell_{\text{max}}]\). For the constrained structured loss \(L_{\text{cstnd}}^\rho\), the conditional \(L_{\text{cstnd}}^\rho\)-risk can be expressed as follows:

\[
\mathcal{E}_{L_{\text{cstnd}}}^\rho(h, x) = \sum_{y \in \mathcal{Y}} p(x, y) \sum_{y' \in \mathcal{Y}} \ell(y', y) \max\{0, 1 + h(x, y')\} = \sum \max\{0, 1 + h(x, y')\} q(x, y').
\]

Let \(y_{\text{min}} = \arg\min_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p(x, y) \ell(y, y')\), where we choose the label with the highest index under the natural ordering of labels as the tie-breaking strategy. For any \(h \in \mathcal{H}\) such that \(h(x) \neq y_{\text{min}}\) and \(x \in \mathcal{X}\), by the symmetry and completeness of \(\mathcal{H}\), we can always find a family of hypotheses \(\{h_\mu : \mu \in \mathbb{R}\} \subset \mathcal{H}\) such that \(h_\mu(x, \cdot)\) take the following values:

\[
\mathcal{T}_\mu(x, y) = \begin{cases} 
  h(x, y) & \text{if } y \neq \{y_{\text{min}}, h(x)\} \\
  h(x, y_{\text{min}}) + \mu & \text{if } y = h(x) \\
  h(x, h(x)) - \mu & \text{if } y = y_{\text{min}}.
\end{cases}
\]

Note that the hypotheses \(\mathcal{T}_\mu\) satisfies the constraint:

\[
\sum_{y \in \mathcal{Y}} \mathcal{T}_\mu(x, y) = \sum_{y \in \mathcal{Y}} h(x, y) = 0, \forall \mu \in \mathbb{R}.
\]

Since \(\sum_{y \in \mathcal{Y}} h(x, y) = 0\), there must be non-negative scores. By definition of \(h(x)\) as a maximizer, we must thus have \(h(x, h(x)) \geq 0\). By (34) and using the fact that \(H(x) = y\) when \(\mathcal{H}\) is symmetric, we obtain

\[
\Delta \mathcal{E}_{L_{\text{cstnd}}}^\rho(h, x) = \mathcal{E}_{L_{\text{cstnd}}}^\rho(h, x) - \mathcal{E}_{L_{\text{cstnd}}}^\rho(\mathcal{H}, x) \\
\geq \mathcal{E}_{L_{\text{cstnd}}}^\rho(h, x) - \inf_{\mu \in \mathbb{R}} \mathcal{E}_{L_{\text{cstnd}}}^\rho(\mathcal{T}_\mu, x) \\
= \sup_{\mu \in \mathbb{R}} \left\{ q(x, y_{\text{min}}) \left( \min \left\{ \max \left\{ 0, 1 + \frac{h(x, y_{\text{min}})}{\rho} \right\}, 1 \right\} \right) - \min \left\{ \max \left\{ 0, 1 + \frac{h(x, h(x)) - \mu}{\rho} \right\}, 1 \right\} \right\} \\
+ q(x, h(x)) \left( \min \left\{ \max \left\{ 0, 1 + \frac{h(x, h(x))}{\rho} \right\}, 1 \right\} \right) - \min \left\{ \max \left\{ 0, 1 + \frac{h(x, y_{\text{min}}) + \mu}{\rho} \right\}, 1 \right\} \right\} \\
\geq q(x, h(x)) - q(x, y_{\text{min}}) \quad \text{(differentiating with respect to } \mu \text{ to optimize)} \\
= \sum_{y \in \mathcal{Y}} p(x, y) \ell(h(x), y) - \sum_{y \in \mathcal{Y}} p(x, y) \ell(y_{\text{min}}, y) \\
= \Delta \mathcal{E}_{L_{\text{cstnd}}}^\rho(h, x). \quad \text{(by Lemma 3 and } H(x) = y)\]

Therefore, we obtain for any hypothesis \(h \in \mathcal{H}\) and any distribution,

\[
\mathcal{R}_L(h) - \mathcal{R}_L^*(\mathcal{H}) \leq \mathbb{E}_X [\Delta \mathcal{E}_{L_{\text{cstnd}}}^\rho(h, x)] \\
\leq \mathbb{E}_X [\Delta \mathcal{E}_{L_{\text{cstnd}}}^\rho(h, x)] \\
= \mathcal{R}_{L_{\text{cstnd}}}^*(h) - \mathcal{R}_{L_{\text{cstnd}}}^*(\mathcal{H}) + M_{L_{\text{cstnd}}}(\mathcal{H}),
\]

which leads to

\[
\mathcal{R}_L(h) - \mathcal{R}_L^*(\mathcal{H}) \leq \mathcal{R}_{L_{\text{cstnd}}}^*(h) - \mathcal{R}_{L_{\text{cstnd}}}^*(\mathcal{H}) + M_{L_{\text{cstnd}}}(\mathcal{H}) - M_{L_{\text{cstnd}}}(\mathcal{H}).
\]

□
G Efficient gradient computation and inference

Here, we describe efficient algorithms for the computation of the gradients for the loss functions \( L_{\text{log}}^{\text{comp}} \) and \( L_{\text{exp}}^{\text{comp}} \). We also briefly discuss an efficient algorithm for inference.

G.1 Efficient gradient computation for \( L_{\text{log}}^{\text{comp}} \)

We first present an efficient algorithm for the computation of the quantities \( L(z, s) \) in the important case of rational losses, next in the case of Markovian losses.

Rational losses. Rational losses form a general family of loss functions based on rational kernels [Cortes et al., 2004] that includes, in particular, \( n \)-gram losses, which can be defined for a pair of sequences \((y, y')\) as the negative inner product of the vectors of \( n \)-gram counts of \( y \) and \( y' \).

Our algorithm bears some similarity to that of Cortes et al. [2018] for the computation of the gradient of the VCRF loss function. It is however distinct because the structured prediction loss function we are considering and our definition of rational loss are both different. We will adopt a similar notation and terminology. Recall that for any sequence \( y \), we denote by \( y_i \) the symbol in its \( i \)th position and by \( y_i^j = y_i y_{i+1} \ldots y_j \) the substring of \( y \) starting at position \( i \) and ending at \( j \). We denote by \( E_A \) the set of transitions of a WFA \( A \).

Let \( \mathcal{U} \) be a weighted finite-state transducer (WFST) over the \((+, \times)\) semiring over the reals, with \( \Delta \) as both the input and output alphabet. Then, we define the rational loss associated to \( \mathcal{U} \) for all \( y, y' \in \Delta^* \) by \( \ell(y, y') = \mathcal{U}(y, y') \).

Let \( \mathcal{Y} \) denote a WFA over the \((+, \times)\) semiring accepting the set of all sequences of length \( l \) with weight one and let \( \mathcal{Y}_i \) denote the WFA accepting only \( y_i \) with weight one. Then, by definition, the weighted transducer \( \mathcal{Y} \circ \mathcal{U} \circ \mathcal{Y}_i \), obtained by composition maps each sequence \( y \) in \( \Delta^l \) to \( y_i \), with weight \( \mathcal{U}(y, y_i) \). The WFA \( \Pi_l(\mathcal{Y} \circ \mathcal{U} \circ \mathcal{Y}_i) \) derived from that transducer by projection on the input (that is by removing output labels) is associating to each sequence \( y \) weight \( \mathcal{U}(y, y_i) \).

We use weighted determinization [Mohri, 1997] to compute an equivalent deterministic WFA denote \( M \). As shown by Cortes et al. [2015][Theorem 3], \( M \) can be computed in polynomial time. \( M \) admits a unique path labeled with any sequence \( y \in \Delta^l \) and the weight of that path is \( \mathcal{U}(y, y_i) \). The weight of that accepting path is obtained by multiplying the weights of its transitions and that of the final state.

We now define a deterministic \( p \)-gram WFA \( N \) that accepts all sequences \( y \in \Delta^l \) with each of its states \((z', s)\) encoding a \((p−1)\)-gram \( z' \) read to reach it and the position \( s \) in the sequence \( y \) at which it is reached. The transitions of \( N \) are therefore defined as follows with weight one:

\[
E_N = \left\{ \left((y_{s-p+1}^{s-1}, s-1), a, 1, (y_{s-p+2}^{s-1} a, s)\right) \mid y \in \Delta^l, a \in \Delta, s \in [l] \right\}.
\]

The initial state is \((\epsilon, 0)\) and the final states are those with the second element of the pair (the position) being \( l \). Note that, by construction, \( N \) is deterministic. Then, the composition (or intersection) WFA \( N \circ M \) still associates the same weight as \( M \) to each input string \( y \in \Delta^l \). However, the states in that composition help us compute \( L(z, s) \). In particular, for any \( z \in \Delta^p \) and \( s \in [l] \), let \( E(z, s) \) be the set of transitions of \( N \circ M \) constructed by pairing the transition \((z_1^{p-1}, s-1), z_p, \omega(z, s), (z_2^{p-1}, s)\) in \( N \) with a transition \((q_M, z_p, \omega, q'_M)\) in \( M \). They admit the following form:

\[
E(z, s) = \left\{ (q_M, z, s, (q_M, z_p, \omega, q'_M)) \mid q_M = \mathcal{E}_{N \circ M} : q_M = (z_1^{p-1}, s-1) \right\}.
\]

The WFA \( N \circ M \) is deterministic as a composition of two deterministic WFAs. Thus, there is a unique path labeled with a sequence \( y \in \Delta^l \) in \( N \circ M \) and \( y \) admits the substring \( z \) ending at position...
s iff that path goes through a transition in \( E(z, s) \) when reaching position \( s \). Therefore, to compute \( L(z, s) \), it suffices for us to compute the sum of the weights of all paths in \( N \circ M \) going through a transition in \( E(z, s) \). This can be done straightforwardly using the forward-backward algorithm or two single-source shortest-distance algorithm over the \((+, \times)\) semiring [Mohri, 2002a], one from the initial state, the other one from the final states. Since \( N \circ M \) is acyclic and admits \( O(l|\Delta|^p) \) transitions, we can compute all the quantities \( L(z, s), s \in [l] \) and \( z \in \Delta^p \), in time \( O(l|\Delta|^p) \).

**Markovian loss.** We consider adopting a Markovian assumption, which is commonly adopted in natural language processing [Manning and Schütze, 1999]. We will assume that \( \ell \) can be decomposed as follows for all \( y, y' \in |\Delta|^l \):

\[
\ell(y, y') = \prod_{l=1}^{t} \ell(y_{l-p+1}, y_l).
\]

To efficiently compute \( L(z, s) \), we will use a WFA representation similar to the one used by Cortes et al. [2016, 2018] and, for convenience, will adopt a similar notation. \( L(z, s) \) coincides with a flow computation in a WFA \( A \) that we now define. \( A \) has the following set of states:

\[
Q_A = \{(y_{l-p+1}, t) : y \in |\Delta|^l, t = 0, \ldots, l\},
\]

with \( l_A = (\varepsilon, 0) \) its single initial state, \( F_A = \{(y_{l-p+1}, t) : y \in |\Delta|^l\} \) its set of final states, and a transition from state \((y_{l-p+1}, t-1)\) to state \((y_{l-p+1}, b, t)\) with label \( b \) and weight \( \omega(y_{l-p+1}, b, t) = \ell(y_{l-p+1}, y_l, t) \), that is the following set of transitions:

\[
E_A = \{(y_{l-p+1}, t-1), b, \omega(y_{l-p+1}, b, t), (y_{l-p+1}, b, t) : y \in |\Delta|^l, b \in \Delta, t \in [l]\}.
\]

By construction, \( A \) is deterministic. The weight of a path in \( A \) is obtained by multiplying the weights of its constituent transitions. In view of that, \( L(z, s) \) can be seen as the sum of the weights of all paths in \( A \) going through the transition from state \((z_{l-p+1}^{-1}, s-1)\) to \((z_p^0, s)\) with label \( z_p \).

For any state \((y_{l-p+1}, t) \in Q_A \), we denote by \( \alpha((y_{l-p+1}, t)) \) the sum of the weights of all paths in \( A \) from the initial state \( l_A \) to \((y_{l-p+1}, t) \) and by \( \beta((y_{l-p+1}, t)) \) the sum of the weights of all paths from \((y_{l-p+1}, t) \) to a final state. Then, \( L(z, s) \) is given by

\[
L(z, s) = \alpha((z_{l-p+1}^{-1}, s-1)) \times \omega(z, s) \times \beta((z_p^0, s)).
\]

Since \( A \) is acyclic, \( \alpha \) and \( \beta \) can be computed for all states in linear time in the size of \( A \) using a single-source shortest-distance algorithm over the \((+, \times)\) semiring or the so-called forward-backward algorithm. Thus, since \( A \) admits \( O(l|\Delta|^p) \) transitions, we can also compute all quantities \( L(z, s), s \in [l] \) and \( z \in \Delta^p \), in time \( O(l|\Delta|^p) \).

### G.2 Efficient gradient computation for \( L_{\text{comp}}^{\text{exp}} \)

In this section, we provide a brief overview of the gradient computation for \( L_{\text{comp}}^{\text{exp}} \), which is similar to the approach used for \( L_{\log}^{\text{exp}} \).

Note that the loss \( L_{\text{comp}}^{\text{exp}} \) on a given point \((x_i, y_i)\) can be expressed as follows:

\[
L_{\text{exp}}^{\text{comp}} = \sum_{y' \in |\Delta|^l} \ell(y', y_i) \sum_{y''} e^{h(x_i, y'') - h(x_i, y')} = \sum_{y' \in |\Delta|^l} \ell(y', y_i) \sum_{y''} e^{h(x_i, y'') - h(x_i, y')} - \sum_{y' \in |\Delta|^l} \ell(y', y_i).
\]

\[
= \left[ \sum_{y' \in |\Delta|^l} e^{h(x_i, y'')} \right] \sum_{y' \in |\Delta|^l} \ell(y', y_i) e^{-h(x_i, y')} - \sum_{y' \in |\Delta|^l} \ell(y', y_i).
\]

\[
= \left[ \sum_{y' \in |\Delta|^l} e^{w \cdot \Psi(x_i, y'')} \right] \sum_{y' \in |\Delta|^l} \ell(y', y_i) e^{-w \cdot \Psi(x_i, y')} - \sum_{y' \in |\Delta|^l} \ell(y', y_i).
\]
An efficient computation of these terms is not straightforward since the summations run over an exponential number of sequences for \( y \). However, we will leverage the Markovian property of the features to design an efficient computation. This approach is similar to what we demonstrated earlier for \( L_{\text{log}} \). We start with identifying the most computationally challenging terms by rewriting the expression of the gradient of \( L_{\text{comp}} \) in the following lemma.

**Lemma 19.** For any \( w \in \mathbb{R}^d \), the gradient of \( L_{\text{comp}} \) can be expressed as follows:

\[
\nabla L_{\text{comp}}^\text{exp}(w) = \sum_{s=1}^l \sum_{z \in \Delta^*} \left[ N_w Q_w(z, s) - Z_w C_w(z, s) \right] \tilde{\psi}(x_i, z, s),
\]

where \( Q_w'(z, s) = \sum_{y/y'_{s-p+1}=z} e^{w \cdot \Psi(x, y)} \), \( C_w(z, s) = \sum_{y/y'_{s-p+1}=z} \tilde{t}(y, y_i) e^{-w \cdot \Psi(x, y)} \) and \( N_w = \sum_{y \in \Delta} \tilde{t}(y, y_i) e^{-w \cdot \Psi(x, y)} \).

**Proof.** Using the decomposition of the feature vector, we can write:

\[
\sum_{y \in \Delta^l} e^{w \cdot \Psi(x, y)} \psi(x_i, y) = \sum_{y \in \Delta^l} e^{w \cdot \Psi(x, y)} \sum_{s=1}^l \tilde{\psi}(x_i, y_{s-p+1}, s)
\]

\[
= \sum_{s=1}^l \sum_{z \in \Delta^*} \sum_{y/y'_{s-p+1}=z} e^{w \cdot \Psi(x, y)} \tilde{\psi}(x_i, z, s)
\]

\[
\sum_{y \in \Delta^l} \tilde{t}(y, y_i) e^{-w \cdot \Psi(x, y)} \psi(x_i, y) = \sum_{y \in \Delta^l} \tilde{t}(y, y_i) e^{-w \cdot \Psi(x, y)} \sum_{s=1}^l \tilde{\psi}(x_i, y_{s-p+1}, s)
\]

\[
= \sum_{s=1}^l \sum_{z \in \Delta^*} \sum_{y/y'_{s-p+1}=z} \tilde{t}(y, y_i) e^{-w \cdot \Psi(x, y)} \tilde{\psi}(x_i, z, s).
\]

This completes the proof. \( \square \)

It was shown by Cortes et al. [2016, 2018] that all of the quantities \( Q_w'(z, s) \) for \( z \in \Delta^p \) and \( s \in [l] \) and \( Z_w \) can be computed efficiently in time \( O(lr^2) \), where \( r = |\Delta| \). Thus, the remaining bottleneck in the gradient computation suggested by Lemma 19 is the evaluation of the quantities \( C_w(z, s) \) for \( z \in \Delta^p \) and \( s \in [l] \) and \( N_w \). As with the loss function \( L_{\text{comp}} \) discussed in the previous section, we will analyze the computation of these quantities first in the case of rational losses, next in that of Markovian loss.

**Rational losses.** We will adopt the same notation as in the case of the \( L_{\text{log}} \) loss with the same definition of a rational loss: \( \ell \) is a rational loss if there exists a WFST over the \((+,\times)\) semiring over the reals with \( \Delta \) as both the input and output alphabet such that for all \( y, y' \in \Delta^* \), we have \( \tilde{t}(y, y') = \ell(y, y') \).

Our algorithm is also somewhat similar to the one described for the \( L_{\text{comp}} \) loss or that of Cortes et al. [2018] for the computation of the gradient of the VCRF loss function. There are, however, several differences here too because the quantities computed and thus the automata operations required are distinct.

Exactly as in the case of \( L_{\text{log}} \) loss, we first define a deterministic WFA \( M \) over the \((+,\times)\) semiring that can be computed in polynomial time and that admits a unique path labeled with any sequence...
\(y \in \Delta_i\) whose weight is \(\ell(y, y_i)\). Next, as in [Cortes et al., 2018], we define a deterministic WFA \(A\) such that
\[
A(y) = e^{-w \cdot \Psi(x, y)} = \prod_{i=1}^l e^{-w \cdot \Psi(x, y_{l-p+1}, i)}.
\]
The set of states \(Q_A\) of \(A\) are defined as \(Q_A = \{(y_{l-p+1}, s) : y \in \Delta_i, s = 0, \ldots, l\}\), with \(l_A = (\varepsilon, 0)\) its single initial state, \(\mathcal{F}_A = \{(y_{l-p+1}, l) : y \in \Delta_i\}\) its set of final states, and with a transition from state \((y_{l-p+1}, s - 1)\) to state \((y_{l-p+2}, s, a)\) with label \(a\) and weight, that is, the following set of transitions:
\[
E_A = \{((y_{l-p+1}, s - 1), a, e^{-w \cdot \Psi(x, y_{l-p+1}, a, s)}), (y_{l-p+2}, s, a) : y \in \Delta_i, a \in \Delta, s \in [l]\}.
\]
Then, by definition of composition or intersection [Mohri, 2009], the WFA \((M \circ A)\) is deterministic and admits a unique path labeled with any given \(y \in \Delta_i\) whose weight is \((M \circ A)(y) = M(y) \cdot A(y) = \ell(y, y_i) e^{-w \cdot \Psi(x, y)}\).

Now, \(N_w\) coincides with the sum of the weights of all accepted paths in this WFA. Thus, since \((M \circ A)\) is acyclic, it can be computed in time linear in the size of \((M \circ A)\), that is the number of transitions. For any \(s \in [l]\) and \(z \in \Delta_{p^c}\), \(C_w(z, s)\) is the sum of the weights of all paths in \((M \circ A)\) labeled with a sequence \(z\) admitting \(z\) as a substring ending at position \(s\). The states in the composition \((M \circ A)\) help us compute \(C_w(z, s)\). As in the case of the \(L_{\text{comp}}^\text{log}\) loss, for any \(z \in \Delta_{p^c}\) and \(s \in [l]\), we define \(E(z, s)\) as the set of transitions of \((M \circ A)\) constructed by pairing a transition \((q_{M^c}, z, \omega_M, q_{M^c}')\) in \(M\) with the transition \((z_{l-p+1}^{-1}, s - 1), z_p, \omega(z, s), (z_{l-p+1}^{-1}, s)\) in \(A\). They admit the following form:
\[
E(z, s) = \left\{ ((q_{M^c}, q_A), z_p, \omega_M \cdot \omega(z, s), (q_M^c', q_A')) : (z_{l-p+1}^{-1}, s - 1) \right\}.
\]

The WFA \((M \circ A)\) is deterministic as a composition of two deterministic WFAs. Thus, there is a unique path labeled with a sequence \(y \in \Delta_i\) in \((M \circ A)\) and \(y\) admits the substring \(z\) ending at position \(s\) if and only if path goes through a transition in \(E(z, s)\) when reaching position \(s\). Therefore, to compute \(C_w(z, s)\), it suffices for us to compute the sum of the weights of all paths in \(C_w(z, s)\) going through a transition in \(E(z, s)\). This can be done straightforwardly using the forward-backward algorithm or two single-source shortest-distance algorithm over the \((+, \times)\) semiring [Mohri, 2002a], one from the initial state, the other one from the final states. Since \((M \circ A)\) is acyclic and admits \(O(|\Delta|)|P^c)\) transitions, we can compute all the quantities \(C_w(z, s)\), \(s \in [l]\) and \(z \in \Delta_{p^c}\), in time \(O(l|P^c)\).

**Markovian loss.** Here, we adopt the Markovian assumption and assume that \(\ell\) can be decomposed as follows for all \(y, y' \in \Delta_i\): \(\ell(y, y') = \prod_{t=1}^l \ell_t(y_{l-p+1}, y_t)\). Thus, the quantity \(C_w(z, s)\) can be written as:
\[
C_w(z, s) = \sum_{y, y'_{l-p+1} \in \Delta_{p^c}} \prod_{t=1}^l \ell_t(y_{l-p+1}, y_t) e^{-w \sum_{k=1}^l \Psi(x, y_{l-p+1}, k)}
\]
\[
= \sum_{y, y'_{l-p+1} \in \Delta_{p^c}} \prod_{t=1}^l \ell_t(y_{l-p+1}, y_t) \prod_{k=1}^l e^{-w \cdot \Psi(x, y_{l-p+1}, k)}
\]
\[
= \sum_{y, y'_{l-p+1} \in \Delta_{p^c}} \prod_{t=1}^l \ell_t(y_{l-p+1}, y_t) e^{-w \cdot \Psi(x, y_{l-p+1}, t)}.
\]

Then, we can proceed as in the Markovian loss case for the loss function \(L_{\text{comp}}^\text{log}\) except that instead of the WFA \(A\) used there, we define here a similar WFA \(A'\). The only difference is that the weight \(w(y_{l-p+1}^{-1}, b, t) = \ell_t(y_{l-p+1}^{-1}, b, y_t)\) for the WFA \(A\) is replaced with \(w'(y_{l-p+1}^{-1}, b, t) = \ell_t(y_{l-p+1}^{-1}, b, y_t) e^{-w \cdot \Psi(x, y_{l-p+1}, t)}\). With the same argument, we can compute all quantities \(C_w(z, s)\), \(s \in [l]\) and \(z \in \Delta_{p^c}\), in time \(O(l|P^c)\). The quantity \(N_w\) can also be efficiently computed in time \(O(l|P^c)\) since it is the sum of the weights of all paths in \(A'\).
G.3 Efficient Inference

We focused on the problem of efficient computation of the gradient. Inference is also a key problem in structured prediction since the label with a highest score must be determined out of an exponentially large set of possible ones. However, for the linear hypotheses considered in the previous sections, this problem can be efficiently tackled since it can be cast as a shortest-distance problem in a directed acyclic graph, as in [Cortes, Kuznetsov, Mohri, and Yang, 2016].

More generally, an efficient gradient computation, efficient inference and other related algorithms can benefit from standard weighted automata and transducer optimization algorithms such as $\epsilon$-removal [Mohri, 2000, 2002b] and determinization [Mohri, 1997, Mohri and Riley, 1997, Allauzen and Mohri, 2003, 2004] (see also the survey chapter [Mohri, 2009]).