Towards Sharper Generalization Bounds for Structured Prediction

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper, we investigate the generalization performance of structured prediction learning and obtain state-of-the-art generalization bounds. Our analysis is based on factor graph decomposition of structured prediction algorithms, and we present novel margin guarantees from three different perspectives: Lipschitz continuous, smoothness, and the space capacity condition. In the Lipschitz continuous scenario, we improve the square-root dependency on the cardinality of the label set of existing bounds to a logarithmic dependency. In the smoothness scenario, we provide generalization bounds that are not only a logarithmic dependency on the cardinality of the label set but a faster convergence rate of order $\mathcal{O}(\frac{1}{n})$ w.r.t. the sample size n. In the space capacity scenario, generalization bounds not related to the cardinality of the label set and with faster convergence rates than $\mathcal{O}(\frac{1}{\sqrt{n}})$ are presented. In each scenario, applications are provided to suggest that these conditions are easy to be satisfied.

1 Introduction

Structured prediction [16, 10, 15] covers a wide range of machine learning problems, such as computer vision, natural language processing, and computational biology. Several examples of structured prediction problems include part-of-speech tagging, dependency parsing, named entity recognition and machine translation in natural language understanding, image segmentation and objection recognition in compute vision, protein folding in computational biology. An important property of structured prediction is that the output space admits some structure, such as strings, graphs, trees or sequences [16, 10, 11]. Meanwhile, another property common to above tasks is that the natural loss function admits a decomposition along the output substructures. The complex output structure and corresponding loss function make structured prediction different from the widely studied binary or multi-class classfication problem.

The design of the structured prediction algorithms have been thrived for many years, including conditional random fields [27], structured support vector machines [47], kernel-regression algorithm [17], search-based structured prediction [18], maximum-margin markov network [25], image segmentation [33], part-of-speech-tagging [24], and machine translation [50]. Compared with the prosperity and development of structured prediction algorithm designing, the theoretical analysis of structured prediction appears to be not sufficiently well-documented [11], especifically the sharper generalization bound studying. However, it is known that the theoretical study of structured prediction algorithms is definitely significant [16, 10, 11, 1].

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Several theoretical studies of structured prediction consider the specific algorithm and the simple loss such as Hamming loss [25, 14, 12, 20]. Subsequently, more work considers the analysis on general loss. These works can be roughly cast into three categories: PAC-bayesian approach, factor graph decomposition approach, and implicit embedding approach. [35, 23, 4] provide PAC-Bayesian guarantees for arbitrary losses through the analysis of randomized algorithms using count-based hypotheses. [10, 11, 40, 39, 6, 9, 45, 5] use implicit loss embedding to construct a connection between discrete output of structured prediction and regression, to investigate the generalization performance of structured prediction. [16, 3] present generalization bounds for general loss based on factor graph decomposition. There has a special structured prediction problem: multi-class classification [30, 29, 32, 42, 38, 49, 36, 37, 34], whose scoring function can be seen as factor graphs with factor node size and substructure size equal to one.

Despite these works provide generalization bounds from different perspectives, there still are some problems urgently to be solved in structured prediction.

[I.] Existing generalization bounds mostly have a slow convergence rate w.r.t. the sample size n. In PAC-bayesian approach, [35, 23, 4, 20] provide the generalization bounds of order $\mathcal{O}(\frac{1}{\sqrt{n}})$. In implicit embedding approach, [10, 11, 39, 9, 45, 5] provide the convergence rate of order $\mathcal{O}(\frac{1}{n^{1/4}})$, [40] of order $\mathcal{O}(\frac{1}{\sqrt{n}})$. Also, in the factor graph decomposition approach, [16] presents the generalization upper bounds of order $\mathcal{O}(\frac{1}{\sqrt{n}})$. Therefore, there raises a question: can the convergence rate of structured prediction achieves faster order $\mathcal{O}(\frac{1}{n})$?

[II.] In structured prediction, the number of possible labels is potentially infinite. Thus it is important to obtain an upper bound that has a lower order dependency on the cardinality of possible labels. The factor graph decomposition approach can provide the explicit dependency on the properties of the factor graph and help us to explicitly show the dependency on the number of possible labels, which shed new light on the role played by the graph in generalization [16]. Therefore, our analysis is based on the factor graph decomposition [16, 3]. However, the upper bounds in [16] show a square-root dependency on the cardinality of possible labels. If we consider the extremely large number of possible labels, their results may be divergent and can not explain the good performance of structured prediction algorithms in practice. Thus, can the upper bound presents a lower order dependency on the cardinality of possible labels, or even, no dependency?

[III.] Although some work in multi-class classification [32, 42, 49], a special structured prediction problem, show the faster convergence rate of order $\mathcal{O}(\frac{1}{n})$, it is unknown in more difficult and complex structured prediction problem. Additionally, we naturally wonder to know whether the generalization bounds of structured prediction can combine **[I]** with **[II]** and show faster convergence rate simultaneously?

This paper intends to answer the three interesting questions. We consider the general loss, and present novel margin-based theoretical analysis for structured prediction from three different perspectives: Lipschitz continuous, smoothness, and space capacity-dependent perspective. We first assume the loss function used in structured prediction is Lipschitz continuous and try to answer question **[II]**. Under this condition, we improve the generalization bounds to a logarithmic dependency on the cardinality of the label set compared with [16] whose bounds are square-root dependency, but a slow convergence rate of order $\mathcal{O}(\frac{1}{\sqrt{n}})$ w.r.t. the sample size n. Then, we assume the loss is smooth and intend to answer questions **[I]** and **[III]**. Under this assumption, we obtain sharper bounds that not only have faster convergence rate of order $\mathcal{O}(\frac{1}{n})$ but also have the logarithmic dependency on the cardinality of the label set. Furthermore, we consider the space capacity-dependent assumptions: logarithmic covering number and polynomial covering number assumptions, which are commonly used in learning theory [28, 51, 43]. This scenario attempts to answer questions **[I]**, **[II]** and **[III]**. Under this condition, we show that the bound can present no dependency on the cardinality of the label set, simultaneously with faster convergence rates than $\frac{1}{\sqrt{n}}$. In the above three perspectives, we all provide applications to suggest that these conditions are easy to be satisfied in practice.

This paper is organized as follows. In Section 2, we introduce the notations and definitions relevant to our discussion. In Section 3, we present main results, deriving a series of new learning guarantees for structured prediction. Section 4 concludes this paper.

2 Preliminaries

Notations. Let P be a probability measure defined on a sample space $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} \in \mathbb{R}^d$ being the input space and $\mathcal{Y}$ being the output space. A distinguishing property of strutured prediction is that the natural loss function admits a decomposition along the output substructures [16]. We assume that the output space $\mathcal{Y}$ can be decomposed into l substructures: $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_l$, where $\mathcal{Y}_i$ is the set of possible labels that can be assigned to substructure i. Take the simple webpage collective classification task for example [46], each $\mathcal{Y}_i$ is a webpage label and $\mathcal{Y}$ is a joint label for an entire website. If we assume each $\mathcal{Y}_i \in \{0, 1\}$, the number of possible labels to $\mathcal{Y}$ is exponential in the number of substructures l, i.e., $|\mathcal{Y}| = 2^l$.

Factor graphs and scoring function. In structured prediction, we aim to learn a scoring function $f = \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Let $\mathcal{F}$ be a family of scoring function. For each $f \in \mathcal{F}$, we define the predictor $\hat{f}(x)$ as: $\hat{f}(x) = \arg\max_{y \in \mathcal{Y}} f(x, y)$. Moreover, we assume that the scoring function f can be decomposed as a sum [47, 25, 27, 16], which is standard in structured prediction and can be formulated via the notation of factor graph [16]. A factor graph G is a bipartite graph, and is represented as a tuple $G = (V, H, E)$, where $V = \{1, ..., l\}$ is the set of variables nodes, H is a set of factor nodes, and E is a set of undirected edges between a variable node and a factor node. Let $\mathcal{N}(h)$ be the set of variable nodes connected to the factor h by an edge and $\mathcal{Y}_h$ be the substructure set cross-product $\mathcal{Y}_h = \prod_{k \in \mathcal{N}(h)} \mathcal{Y}_k$. Based on above notations, the scoring function $f(x, y)$ can be decomposed as a sum of functions f_h [16], with an input $x \in \mathcal{X}$ and an input $y_h \in \mathcal{Y}_h$ as arguments:

$$f(x, y) = \sum_{h \in H} f_h(x, y_h).$$

In this paper, we consider the more generally setting: for each example (x_i, y_i), the corresponding factor graph may be different, that is, $G(x_i, y_i) = G_i = ([l_i], H_i, E_i)$. A special case of this setting is that, for example, when the size l_i of each example is allowed to vary and where the number of possible labels $|\mathcal{Y}|$ is potentially infinite. Figure 1 shows different examples of factor graphs.

Learning. In order to measure the success of a prediction, we use a loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ to measure the dissimilarity of two elements of the output space $\mathcal{Y}$, such as the Hamming loss defined as: $L(y, y') = \frac{1}{l} \sum_{k=1}^{l} \mathbb{I}_{y_k \neq y'_k}$ for two outputs $y, y' \in \mathcal{Y}$, with $y = (y_1, ..., y_l)$ and $y = (y'_1, ..., y'_l)$, or edit-distance loss in natural language and speech processing applications, or other losses [16, 8]. Given a training set $S = \{(x_1, y_1), ..., (x_n, y_n)\}$ of size n being independently drawn from the underlying distribution P, our goal is to learn a hypothesis $f \in \mathcal{F}$ with good generalization performance from the sample S by optimizing the loss L. The performance is measured by expected risk, defined as

$$R(f) := \mathbb{E}_{(x,y) \sim P}[L(\hat{f}(x), y)].$$

However, the above mentioned loss function $f \to L(\hat{f}(x), y)$ is typically the $0 - 1$ loss of f, which is hard to handle in learning machine. Therefore, one usually consider the surrogate loss [10, 11, 16].

Margin and loss function class. We now introduce the definition of standard margin and the surrogate loss. The margin of any scoring function f for an example (x, y) is defined as:

$$\rho_{f(x,y)} = f(x, y) - \max_{y' \neq y} f(x, y').$$

We consider the scoring function f of the form $f(x, y) = \langle w, \Psi(x, y) \rangle$, where Ψ is a feature mapping from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}^N$ such that $\Psi(x, y) = \sum_{h \in H} \Psi_h(x, y_h)$ due to the decomposition of f [16]. And we define the following scoring function space:

$$\mathcal{F} = \{x \mapsto \langle w, \Psi(x, y) \rangle : w \in \mathbb{R}^N, \|w\|_p \leq \Lambda_p\}, \tag{1}$$

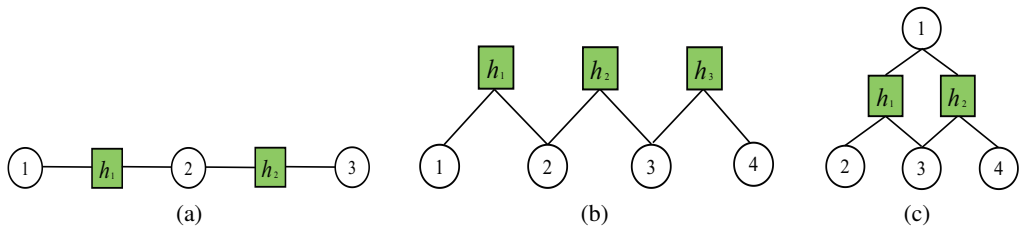

Figure 1: Example of factor graphs. (a) represents an example of pairwise Markov network decomposition: $f(x,y) = f_{h_1}(x,y_1,y_2) + f_{h_2}(x,y_2,y_3)$; (b) represents an example of factor graph that having a sequence structure: $f(x,y) = f_{h_1}(x,y_1,y_2)+f_{h_2}(x,y_2,y_3)+f_{h_3}(x,y_3,y_4)$; (c) represents a tree-structured factor graph.

where $\|w\|_p = (\sum_{i=1}^N |w_i|^p)^{\frac{1}{p}}$. Then, the corresponding loss function class is defined as:

$$\mathcal{L}_\rho = \{\ell_\rho(x,y,f) := \ell(\rho_f(x,y)) : f \in \mathcal{F}\},$$

where ℓ_ρ is the surogate loss of L and is bounded by M, that is $0 \leq L(\hat{f}(x),y) = L(\hat{f}(x),y)_{\rho_{f(x,y)}\leq 0} \leq \ell(\rho_f(x,y)) \leq M$. Additionally, we define the empirical risk of any scoring function f as

$$\hat{R}(\ell(\rho_f)) := \frac{1}{n}\sum_{i=1}^n \ell_\rho(f,x_i,y_i),$$

and the expected risk as

$$R(\ell(\rho_f)) := \mathbb{E}_{(x,y)\sim P}\left[\ell_\rho(f,x,y)\right].$$

Clearly, we have $R(f) \leq R(\ell(\rho_f))$.

3 Main Results

In this section, we provide sharper generalization bounds for structured prediction from three perspectives: Lipschitz continuous, smoothness, and space capacity. We finally obtain sharper generalization bounds of order $\mathcal{O}(\frac{1}{n})$ w.r.t. the sample size n and a logarithmic dependency, or even no dependency, on the cardinality of the label set. In each case, we provide applications to suggest that these assumptions are easy to be satisfied.

3.1 Lipschitz Continuous Dependent Case

Assumption 1 *Assume that ℓ is μ-Lipschitz continuous, that is*

$$|\ell(t) - \ell(s)| \leq \mu|t-s|.$$

Assumption 1 is a pretty mild assumption. The hinge loss $\ell(\rho_f) = (1-\rho_f)_+$ and the margin loss $\ell(\rho_f) = 1_{\rho_f\leq 0} + (1-\rho_f\rho^{-1})1_{0<\rho_f\leq\rho}$ satisfy Assumption 1 [29]. Moreover, the loss of truncated margin functions in [22] satisfies Assumption 1. The additive and the multiplicative margin loss provided by [16], which covers many existing structured prediction algorithms, satisfying Assumption 1 (see Corollary 1 for details).

Theorem 1 *Under Assumption 1, for any $\delta > 0$, with probability $1-\delta$ over the sample S of size n, the following holds for all $f \in \mathcal{F}$,*

$$R(f) \leq R(\ell(\rho_f)) \leq \hat{R}(\ell(\rho_f)) + \mathcal{O}\left(\frac{\mu s \ln n}{\sqrt{n}}\left(\log^{\frac{1}{2}}(n^2) + \log^{\frac{1}{2}}(sk)\right) + \sqrt{\frac{\log 1/\delta}{n}}\right),$$

where $s = \max_{i\in[n]}|H_i|$, $k = \max_{i\in[n]}|H_i|d_i$, and where $d_i = \max_{h\in|H_i|}|\mathcal{Y}_h|$.

Remark 1 **[Sketch of Theorem 1.]** Since our task is to prove bounds with better rates, that is, which decrease fast with $n \to \infty$, thus for brevity, we leave the explicit results of all Theorems in Appendix. In Theorem 1, we focus on terms in the numerator. From Theorem 1, one can see that our generalization bound has a linear dependency on the maximum factor size of a given sample set, that is $s = \max_{i\in[n]} |H_i|$. The terms of order $\mathcal{O}(\log(\cdot))$ is small enough and typically can be ignored. On terms $\log^{\frac{1}{2}}(n^2)$ and $\log^{\frac{1}{2}}(sk)$, if we consider the case $n^2 \geq sk$, our result suggests that the influence in generalization error of sample size n is larger than the cardinality of the label set. It implies that if the sample size n is large enough, hyper-parameter n will play a dominated role in generalization performance of structured prediction. While for cardinality of the label size, our bound has a logarithmic dependency, that is $\log(d_i)$ (Since the factor size s is typically far smaller than label size d_i [16], we just consider the label size d_i in term $\log^{\frac{1}{2}}(sk)$), and is faster than the results in [16]. When the function class $\mathcal{F}$ is assigned with $p = 1$ or $p = 2$, the generalization bounds in [16] are $R(\ell^{add}(\rho_f)) \leq \hat{R}(\ell^{add}(\rho_f)) + \mathcal{O}(\frac{s\sqrt{d_i}}{\rho\sqrt{n}} + \sqrt{\frac{\log 1/\delta}{n}})$ and $R(\ell^{mult}(\rho_f)) \leq \hat{R}(\ell^{mult}(\rho_f)) + \mathcal{O}(\frac{Ms\sqrt{d_i}}{\rho\sqrt{n}} + \sqrt{\frac{\log 1/\delta}{n}})$ (ℓ^{add} and ℓ^{mult} are defined in Application 1). Thus the generalization bounds in [16] have a linear dependency on s but a square-root dependency on d_i, that is $\sqrt{d_i}$. The term $\sqrt{d_i}$ may be extremely large since it is the largest cardinality of the label set of certain factor in H_i. If we consider the extremely large number of possible labels, the bounds in [16] may be divergent. Thus, compared with their bounds, our generalization bound has a significant improvement. The complete proof of Theorem 1 is provided in Appendix A. Besides, the Lipschitz continuous constant μ is typically small, we present two applications for example in the following.

Application 1. [16] provides two margin loss, the additive and the multiplicative margin losses, that can be used to guarantee many existing structured prediction algorithms, defined as:

$$\ell^{add}(\rho_f(x,y)) = \Phi^*\left(\max_{y'\neq y} L(y',y) - \frac{1}{\rho}\left[f(x,y) - f(x,y')\right]\right),$$
$$\ell^{mult}(\rho_f(x,y)) = \Phi^*\left(\max_{y'\neq y} L(y',y)\left(1 - \frac{1}{\rho}\left[f(x,y) - f(x,y')\right]\right)\right),$$

where $\Phi^*(r) = \min\left(\max_{y,y'} L(y,y'), \max(0,r)\right)$ for any $y, y' \in \mathcal{Y}$. For the additive and the multiplicative margin losses, we show that they are $\frac{1}{\rho}$ and $\frac{M}{\rho}$ Lipschitz continuous w.r.t. the margin ρ_f respectively in Appendix B, thus we have Corollary 1.

Corollary 1 *Fix $\rho > 0$. For any $\delta > 0$, with probability $1 - \delta$ over the sample S of size n, the following holds for all $f \in \mathcal{F}$,*

$$R(f) \leq R(\ell^{add}(\rho_f)) \leq \hat{R}(\ell^{add}(\rho_f)) + \mathcal{O}\left(\frac{s\ln n}{\rho\sqrt{n}}\left(\log^{\frac{1}{2}}(n^2) + \log^{\frac{1}{2}}(sk)\right) + \sqrt{\frac{\log 1/\delta}{n}}\right),$$
$$R(f) \leq R(\ell^{mult}(\rho_f)) \leq \hat{R}(\ell^{mult}(\rho_f)) + \mathcal{O}\left(\frac{Ms\ln n}{\rho\sqrt{n}}\left(\log^{\frac{1}{2}}(n^2) + \log^{\frac{1}{2}}(sk)\right) + \sqrt{\frac{\log 1/\delta}{n}}\right),$$

where $s = \max_{i\in[n]} |H_i|$, $k = \max_{i\in[n]} |H_i| d_i$, and where $d_i = \max_{h\in|H_i|} |\mathcal{Y}_h|$.

Application 2. In many structured prediction applications, such as natural language processing and computer vision, people may wish to take advantage of very rich features. However, using a rich family of hypotheses can lead to overfitting. In this application, we further consider to derive learning guarantees for ensembles of structured prediction rules that explicitly account for the differing complexities between families, called Voted Risk Minimization (VRM) [13, 16]. Assume that we are given p families $F_1, ..., F_p$ of functions mapping from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$. Define the ensemble family $\mathcal{G} = conv(\cup_{i=1}^p F_i)$, that is the family of functions g of the form $g = \sum_{t=1}^T \alpha_t f_t$, where $\alpha = (\alpha_1, ..., \alpha_T)$ is in the simplex Δ and where, for each $t \in [1, T]$, f_t is in F_{i_t} for some $i_t \in [1, p]$.

Consider the following loss function:

$$\ell_\tau^{add}(\rho_g(x,y)) = \Phi^*\left(\max_{y'\neq y} L(y',y) + \tau - \frac{1}{\rho}\left[g(x,y) - g(x,y')\right]\right),$$
$$\ell_\tau^{mult}(\rho_g(x,y)) = \Phi^*\left(\max_{y'\neq y} L(y',y)\left(1 + \tau - \frac{1}{\rho}\left[g(x,y) - g(x,y')\right]\right)\right),$$

where τ can be seen as a margin term that acts in conjuction with ρ. We assume these families $F_1, ..., F_p$ have differing complexities. This setting is the same as [16], and under this setting we have Corollary 2 for VRM.

Corollary 2 *Fix $\rho > 0$. For any $\delta > 0$, with probability $1-\delta$ over the sample S of size n, the following holds for all $g \in \mathcal{G}$,*

$$R(g) \leq \hat{R}(\ell_1^{add}(\rho_g)) + \mathcal{O}\left(\frac{s\ln n}{\rho\sqrt{n}}\left(\log^{\frac{1}{2}}(n^2) + \log^{\frac{1}{2}}(sk)\right) + \sqrt{C(n,p,\rho,|\mathcal{Y}|,\delta)}\right),$$
$$R(g) \leq \hat{R}(\ell_1^{mult}(\rho_g)) + \mathcal{O}\left(\frac{Ms\ln n}{\rho\sqrt{n}}\left(\log^{\frac{1}{2}}(n^2) + \log^{\frac{1}{2}}(sk)\right) + \sqrt{C(n,p,\rho,|\mathcal{Y}|,\delta)}\right),$$

where $C(n,p,\rho,|\mathcal{Y}|,\delta) = \lceil \frac{1}{\rho^2}\log(\frac{|\mathcal{Y}|^2\rho^2 n}{4\log p})\rceil \frac{\log p}{n} + \frac{\log 2/\delta}{n}$, $s = \max_{i\in[n]}|H_i|$, $k = \max_{i\in[n]}|H_i|d_i$, and where $d_i = \max_{h\in|H_i|}|\mathcal{Y}_h|$.

Remark 2 **[Sketch of Applications 1 and 2.]** Corollary 1 shows that, under the same assumption, our generalization bounds improve rsults in [16] from a square-root dependency $\sqrt{d_i}$ to a logarithmic dependency $\log(d_i)$. For VRM, [16] shows that $R(g) \leq \hat{R}(\ell_1^{add}(\rho_g)) + \mathcal{O}(\frac{1}{\rho\sqrt{n}}\sum_{t=1}^T \alpha_t \Re^G(F_{k_t}) + \sqrt{C(n,p,\rho,|\mathcal{Y}|,\delta)})$ and $R(g) \leq \hat{R}(\ell_1^{mult}(\rho_g)) + \mathcal{O}(\frac{M}{\rho\sqrt{n}}\sum_{t=1}^T \alpha_t \Re^G(F_{k_t}) + \sqrt{C(n,p,\rho,|\mathcal{Y}|,\delta)})$. Thus, the generalization bounds of VRM in [16] are dominated by term $\sum_{t=1}^T \alpha_t \Re^G(F_{k_t})$, where $\Re^G(F_{k_t})$ is the factor graph Rademacher complexity of fucntion calss F_{k_t} and is defined in their paper. The explicit dependency of their bounds on parameter vector α reveals that learning even with highly complex hypothesis sets could be possible so long as the complexity term. However, Corollary 2 shows the explicit bounds rather than $\Re^G(F_{k_t})$ term and suggests that the generalization bounds of VRM do not dependent on the parameter vector α, which suggests that extremely large number of complex families can be used. Therefore, it is feasible to use rich families in conjunction with highly complex families in structured prediction. Furthermore, in Theorem 2 of [16], the term $\Re^G(F_{k_t})$ is of order $\frac{\sqrt{\max_{i\in[n]}|H_i|^2 d_i}}{\sqrt{n}}$ where $d_i = \max_{h\in|H_i|}|\mathcal{Y}_h|$, which presents a square-root dependency on label size, that is $\sqrt{d_i}$. Compared with their results, our bounds are tighter. Therefore, the generalization bounds in Corollary 2 provide further insights into the learning guarantees of VRM. The complete proof of Corollary 1 and Corollary 2 is provided in Appendix B and C, respectively.

3.2 Smooth Dependent Case

Assumption 2 *Assume that ℓ is β-smooth, that is*

$$|\nabla\ell(s) - \nabla\ell(t)| \leq \beta|s-t|.$$

Assumption 2 is a pretty mild condition. Both the square hinge loss $\ell(\rho_f) = (1-\rho_f)_+^2$ and the square margin loss $\ell(\rho_f) = (1_{\rho_f\leq 0} + (1-\rho_f\rho^{-1})1_{0<\rho_f\leq\rho})^2$ satisfy Assumption 2 [32].

Theorem 2 *Under Assumption 2, $\forall v > \max(1, \frac{\sqrt{2}}{2M})$, for any $\delta > 0$, with probability $1-\delta$ over the sample S of size n, the following holds for all $f \in \mathcal{F}$,*

$$R(f) \leq R(\ell(\rho_f)) \leq \max\left\{\frac{v}{v-1}\hat{R}(\ell(\rho_f)), \hat{R}(\ell(\rho_f)) + \mathcal{O}\left(\frac{\beta s^2\log^4 n}{n}\left(\log(n^2) + \log(sk)\right) + \frac{\log(\frac{1}{\delta})}{n}\right)\right\},$$

where $s = \max_{i\in[n]}|H_i|$, $k = \max_{i\in[n]}|H_i|d_i$, and where $d_i = \max_{h\in|H_i|}|\mathcal{Y}_h|$.

Remark 3 [**Sketch of Theorem 2.**] Theorem 2 suggests that the generalization bound of structured prediction is of order $\mathcal{O}(\frac{s^2}{n})$ under smoothness assumption when hide the logarithmic term. Since the sample size n is typically far larger than s, this bound is obviously sharper than bound in Theorem 1. Compared with [16] whose bounds are of order $\frac{\sqrt{\max_{i\in[n]}|H_i|^2 d_i}}{\sqrt{n}}$ where $d_i = \max_{h\in|H_i|}|\mathcal{Y}_h|$, our generalization bound in Theorem 2 is of order $\frac{\max_{i\in[n]}|H_i|^2}{n}$ when we hide the logarithmic term. And if we consider the case $d_i \geq \max_{i\in[n]}|H_i|^2$ which is very possible since the cardinality of the label set d_i can be far larger than factor size $|H_i|$, our bound is sharper than bounds in [16] whether in terms of sample size n or the order of the numerator. For cardinality of the label set, our bound has a logarithmic dependency, that is $\log(d_i)$. Overall, Theorem 2 obtains sharper bounds not only with faster convergence rate of order $\mathcal{O}(\frac{1}{n})$ but with the logarithmic dependency on the cardinality of the label set. Moreover, compared with PAC-bayesian approach [35, 23, 4, 20] that provide the generalization bounds of order $\mathcal{O}(\frac{1}{\sqrt{n}})$ and the implicit embedding approach [10, 11, 39, 9, 45, 5] that mostly provide the convergence rate of order $\mathcal{O}(\frac{1}{n^{1/4}})$ w.r.t. the sample size n, the generalization bound in Theorem 2 is clearly sharper. The complete proof of Theorem 2 is provided in Appendix D. Besides, the smooth constant β is also typically small, and we provide Application 3 for example in the following.

Application 3. Consider the square hinge loss

$$\ell_{sh}(\rho_f(x,y)) = \left(1 - \left(f(x,y) - \max_{y'\neq y} f(x,y')\right)\right)_+^2,$$

which is $\beta = 2$ smooth w.r.t. the margin ρ_f, thus the generalization bound of ℓ_{sh} for structured prediction is immediate.

3.3 Capacity Dependent Case

Definition 1 (Covering Number [52, 26]) *Let $\mathcal{F}$ be class of real-valued fucntions, defined over a space $\mathcal{Z}$ and $S := \{\mathbf{z}_1, ..., \mathbf{z}_n\} \in \mathcal{Z}^n$ of cardinality n. For any $\epsilon > 0$, the empirical ℓ_∞-norm covering number $\mathcal{N}_\infty(\epsilon, \mathcal{F}, S)$ w.r.t S is defined as the minimal number m of a collection of vectors $\mathbf{v}^1, ..., \mathbf{v}^m \in \mathbb{R}^n$ such that ($\mathbf{v}_i^j$ is the i-th component of the vector $\mathbf{v}^j$)*

$$\sup_{f\in\mathcal{F}} \min_{j=1,...,m} \max_{i=1,...,n} |f(\mathbf{z}_i) - \mathbf{v}_i^j| \leq \epsilon.$$

In this case, we call $\{\mathbf{v}^1, ..., \mathbf{v}^m\}$ an (ϵ, ℓ_∞)-cover of $\mathcal{F}$ w.r.t. S. Furthermore, the following covering number is introduced:

$$\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_\infty) := \sup_n \sup_S \mathcal{N}_\infty(\epsilon, \mathcal{F}, S).$$

To present Theorem 3, we use the property $\Psi(x,y) = \sum_{h\in H}\Psi_h(x, y_h)$ to decompose $\Psi(x,y)$ and introduce the following space (different from $\mathcal{F}$ defined in (1)):

$$\mathcal{F}_h := \{\Psi_h(x, y_h) \mapsto \langle w, \Psi_h(x, y_h)\rangle : w \in \mathbb{R}^N, \|w\|_p \leq \Lambda_p\}. \tag{2}$$

Assumptions 3 and 4 are assumptions w.r.t. the covering number bounds (space capacity) on $\mathcal{F}_h$.

Assumption 3 (polynomial covering number) *We assume that the function class defined in (2) satisfies that*

$$\log \mathcal{N}(\epsilon, \mathcal{F}_h, \|\cdot\|_\infty) \leq \frac{\gamma_p}{\epsilon^p} \quad \epsilon > 0,$$

where $0 < p < 2$ and γ_p is a positive constant dependent on p.

Classes that fulfill Assumption 3 are known as satisfying the uniform entropy condition [48]. The popular RKHSs of Gaussian, polynomial and finite rank kernels satisfy this assumption [21]. For example, the function class $\mathcal{F}_h$ specified by Gaussian kernel and $0 < p \leq 1$ satisfy Assumption 3 [19, 36].

Assumption 4 (logarithmic covering number) *We assume that the function class defined in (2) satisfies that*

$$\log \mathcal{N}(\epsilon, \mathcal{F}_h, \|\cdot\|_\infty) \leq D \log^p(\frac{\gamma}{\epsilon}) \quad \epsilon > 0,$$

where p, D and γ are three positive numbers.

Many popular function classes satisfy the above assumption when the hypothesis $\mathcal{F}_h$ is bounded: Any function space with finite VC-dimension [48, 41], including linear functions and univariate polynomials of degree k (for which $d = k + 1$, and $p = 1$) as special cases; Any RKHS based on a kernel with rank D when $p = 1$ [7]; Any unit ball $\mathcal{B}_2 \subset \mathbb{R}^D$ with fixed $\epsilon \in (0, 1)$ [44].

Theorem 3 *With different space capacity assumptions, we have the following different results:*

1) Under Assumptions 1 and 3, $\forall v > \max(1, \frac{\sqrt{2}}{2M})$, for any $\delta > 0$, with probability $1 - \delta$ over the sample S, we have

$$R(f) \leq R(\ell(\rho_f)) \leq \max\left\{\frac{v}{v-1}\hat{R}(\ell(\rho_f)), \hat{R}(\ell(\rho_f)) + \mathcal{O}\left(\frac{\gamma_p \mu^p s^p}{n^{\frac{2}{p+2}}} + \frac{\log(\frac{1}{\delta})}{n}\right)\right\},$$

for any $f \in \mathcal{F}$, where $0 < p < 2$, $s = \max_{i\in[n]} |H_i|$, and γ_p is a positive constant dependent on p.

2) Under Assumptions 1 and 4, $\forall v > \max(1, \frac{\sqrt{2}}{2M})$, for any $\delta > 0$, with probability $1 - \delta$ over the sample S, we have

$$R(f) \leq R(\ell(\rho_f)) \leq \max\left\{\frac{v}{v-1}\hat{R}(\ell(\rho_f)), \hat{R}(\ell(\rho_f)) + \mathcal{O}\left(\frac{D \log^p(\mu s \gamma \sqrt{n})}{n} + \frac{\log(\frac{1}{\delta})}{n}\right)\right\},$$

for any $f \in \mathcal{F}$, where $p > 0$ and $s = \max_{i\in[n]} |H_i|$, and where p, D and γ are three positive numbers.

Remark 4 **[Sketch of Theorem 3.]** Theorem 3 suggests that, when the function space $\mathcal{F}_h$ satisfies the specific space capacity assumption, the generalization bound of structured prediction can have no dependency on the cardinality of the label set, just represents the dependency on the factor size: s^p or $\log^p(s)$. It also implies that the factor node size s can presents a lower order $\mathcal{O}(s^p)$ (when $0 < p < 1$) or $\mathcal{O}(\log^p(s))$ than linear dependency $\mathcal{O}(s)$ in influencing the generalization performance of structured prediction. From Theorem 3, one can also see that, under the polynomial covering number assumption, this bound is sharper than results in Theorem 1 w.r.t. the sample size n since $0 < p < 2$. Under the logarithmic covering number assumption, the bound presents faster order $\mathcal{O}(\frac{1}{n})$ w.r.t. the sample size n. Overall, Theorem 3 obtains generalization bounds that have no dependency on the cardinality of the label set and have faster convergence rates than $\mathcal{O}(\frac{1}{\sqrt{n}})$ simultaneously. Besides, as we showed in Corollary 2, the Lipschitz continuous constant μ is typically small. The complete proof of Theorem 3 is provided in Appendix E. An example satisfying Assumption 4 is given in the following application.

Application 4. We assume that any $\|\Psi_h(x, y_h)\|$ is bounded by B which is a standard assumption in structured prediction, for example, [16] assumes $\max_{i,h,y_h} \|\Psi_h(x, y_h)\|_2 \leq r_2$ where r_2 is a constant. Thus, there holds that $|\langle w_1, \Psi_h(x, y_h)\rangle - \langle w_2, \Psi_h(x, y_h)\rangle| \leq B\|w_1 - w_2\|$ for any w_1, w_2 and sample $(x, y) \in \mathcal{X} \times \mathcal{Y}$, which implies that $\langle w, \Psi_h(x, y_h)\rangle$ is B-Lipschitz w.r.t. w. Assume that $w \in \mathcal{W}$, where $\mathcal{W} \subset \mathbb{R}^N$ is an unit ball. According to Lemma 1.1.8 in [44], any unit ball has the logarithmic covering number bounds. Therefore, combined with the Lipschitz property, we have

$$\log \mathcal{N}(\epsilon, \mathcal{F}_h, \|\cdot\|_\infty) \leq N \log\left(\frac{3B}{\epsilon}\right), \tag{3}$$

which satisfies Assumption 4. Substituting (3) into Theorem 3, the generalization bound of structured prediction in this setting is immediate.

Remark 5 [**Sketch of Application 4.**] To compare our bounds with the extensively studied multi-class classfication task where $|H_i| = 1$, $d_i = c$, and where c is the number of classes of multiclass classfication task, we consider our bounds in the multi-class classfication case, which is $\mathcal{O}(\frac{N\log(\mu\sqrt{n})}{n})$. To the best of our knowledge, this is the first bound not dependent on the label size c in high probability in multi-class classfication. In related work of multiclass classfication [30, 29, 32, 42, 38, 49, 36, 37, 34, 16, 31], [16] shows a convergence rate of order $\mathcal{O}(\frac{c}{\sqrt{n}})$ under 2-norm regularization condition; [30] shows a convergence rate of order $\mathcal{O}(\frac{\log^2(nc)}{\sqrt{n}})$ under ℓ_∞-norm Lipschitz continuous condition; [29] shows a convergence rate of order $\mathcal{O}(\sqrt{\frac{c}{n}})$ under ℓ_2-norm Lipschitz continuous condition; [32] shows a convergence rate of order $\mathcal{O}(\frac{\sqrt{c}\log^3(n)}{n})$ under ℓ_2-norm Lipschitz continuous, smoothness and low noise conditions; [31] shows a convergence rate of order $\mathcal{O}(\frac{\sqrt{c}}{n})$ under ℓ_2-norm Lipschitz continuous and decay of singular value conditions; [49] shows a convergence rate of order $\mathcal{O}(\frac{\log^3(nc)}{n})$ under ℓ_∞-norm Lipschitz continuous and strong convexity conditions (Due to length limit, we just compare our result with the recent work). However, our result shows that if the loss is μ-Lipschitz continuous w.r.t the margin function, the generalization bound of multi-class classfication can show a fast convergence rate of order $\mathcal{O}(\frac{N\log(\sqrt{n})}{n})$. Note that in this bound, we do not assume strong curvature conditions: strong convexity or smoothness. If we consider the case $N\log(\sqrt{n}) \leq \log^3(nc)$, our bound is also sharper than the state-of-the-art generalization bound in multi-class classfication, under much milder assumptions.

Excess Risk Bounds. Define $\hat{f}^* = \arg\min_{f\in\mathcal{F}} \hat{R}(\ell(\rho_f))$ and $f^* = \arg\min_{f\in\mathcal{F}} R(\ell(\rho_f))$, the excess risk of structured prediction is $R(\ell(\rho_{\hat{f}^*})) - R(\ell(\rho_{f^*}))$, which demonstrates the performance of the empirical risk minimizer obtained by samples on the population level and is also an important term for understanding the performance of structured prediction [2]. On excess risk bounds of structured prediction, we have the following Corollary 3.

Corollary 3 *With different space capacity assumptions, we have the following different results:*

1) Under Assumptions 1 and 3, for any $\delta > 0$, with probability $1-\delta$ over the sample S of size n, we have

$$R(f) \leq R(\ell(\rho_{\hat{f}^*})) \leq R(\ell(\rho_{f^*})) + \mathcal{O}\left(\frac{\gamma_p \mu^p s^p}{n^{\frac{2}{p+2}}} + \frac{\log(\frac{1}{\delta})}{n}\right).$$

for any $f \in \mathcal{F}$, where $0 < p < 2$, $s = \max_{i\in[n]} |H_i|$, and γ_p is a positive constant dependent on p.

2) Under Assumptions 1 and 4, for any $\delta > 0$, with probability $1-\delta$ over the sample S of size n, we have

$$R(f) \leq R(\ell(\rho_{\hat{f}^*})) \leq R(\ell(\rho_{f^*})) + \mathcal{O}\left(\frac{\log^p(\mu s \gamma\sqrt{n})}{n} + \frac{\log(\frac{1}{\delta})}{n}\right).$$

for any $f \in \mathcal{F}$, where $p > 0$ and $s = \max_{i\in[n]} |H_i|$, and where p, D and γ are three positive numbers.

Remark 6 [**Sketch of Corollary 3.**] The excess risk bounds in Corollary 3 have the same order convergence rates as results in Theorem 3. And the meaning of these results have been discussed in Remark 4 and Remark 5. We provide the complete proof of Corollary 3 in Appendix F.

4 Conclusion

In this paper, we towards sharper generalization bounds for structured prediction. The analysis is based on the factor graph decomposition approach, which can help shed new light on the role played by the graph in generalization. We present the state-of-the-art generalization bounds from different perspectives. Overall, these bounds presented have answered the three questions posed in Section 1. We believe our theoretical findings can provide deep insights into the learning guarantees of structured prediction. Additionally, we are also concerned about whether the convergence rate of structured prediction can reach super-linear order? In future work, we will investigate this problem, and design new algorithms based on our theoretical analysis.

References

[1] G. BakIr, T. Hofmann, B. Schölkopf, A. J. Smola, and B. Taskar. *Predicting structured data*. MIT press, 2007.

[2] P. L. Bartlett, O. Bousquet, S. Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[3] K. Bello, A. Ghoshal, and J. Honorio. Minimax bounds for structured prediction based on factor graphs. In *International Conference on Artificial Intelligence and Statistics*, pages 213–222. PMLR, 2020.

[4] K. Bello and J. Honorio. Learning latent variable structured prediction models with gaussian perturbations. *arXiv preprint arXiv:1805.09213*, 2018.

[5] L. Brogat-Motte, A. Rudi, C. Brouard, J. Rousu, and F. d'Alché Buc. Learning output embeddings in structured prediction. *arXiv preprint arXiv:2007.14703*, 2020.

[6] V. Cabannes, A. Rudi, and F. Bach. Fast rates in structured prediction. *arXiv preprint arXiv:2102.00760*, 2021.

[7] B. Carl and H. Triebel. Inequalities between eigenvalues, entropy numbers, and related quantities of compact operators in banach spaces. *Mathematische Annalen*, 251(2):129–133, 1980.

[8] H. Choi. Efficient structured surrogate loss and regularization in structured prediction. *arXiv preprint arXiv:1809.05550*, 2018.

[9] C. Ciliberto, F. Bach, and A. Rudi. Localized structured prediction. *arXiv preprint arXiv:1806.02402*, 2018.

[10] C. Ciliberto, L. Rosasco, and A. Rudi. A consistent regularization approach for structured prediction. *Advances in neural information processing systems*, 29:4412–4420, 2016.

[11] C. Ciliberto, L. Rosasco, and A. Rudi. A general framework for consistent structured prediction with implicit loss embeddings. *Journal of Machine Learning Research*, 21(98):1–67, 2020.

[12] M. Collins. Parameter estimation for statistical parsing models: Theory and practice of distribution-free methods. In *New developments in parsing technology*, pages 19–55. Springer, 2004.

[13] C. Cortes, P. Goyal, V. Kuznetsov, and M. Mohri. Kernel extraction via voted risk minimization. In *Feature Extraction: Modern Questions and Challenges*, pages 72–89. PMLR, 2015.

[14] C. Cortes, V. Kuznetsov, and M. Mohri. Ensemble methods for structured prediction. In *International Conference on Machine Learning*, pages 1134–1142. PMLR, 2014.

[15] C. Cortes, V. Kuznetsov, M. Mohri, D. Storcheus, and S. Yang. Efficient gradient computation for structured output learning with rational and tropical losses. *Advances in Neural Information Processing Systems*, 31:6810–6821, 2018.

[16] C. Cortes, V. Kuznetsov, M. Mohri, and S. Yang. Structured prediction theory based on factor graph complexity. *Advances in Neural Information Processing Systems*, 29:2514–2522, 2016.

[17] C. Cortes, M. Mohri, and J. Weston. A general regression framework for learning string-to-string mappings. *Predicting Structured Data*, 2(4), 2007.

[18] H. Daumé, J. Langford, and D. Marcu. Search-based structured prediction. *Machine learning*, 75(3):297–325, 2009.

[19] M. Farooq and I. Steinwart. Learning rates for kernel-based expectile regression. *Machine Learning*, 108(2):203–227, 2019.

[20] A. Ghoshal and J. Honorio. Learning maximum-a-posteriori perturbation models for structured prediction in polynomial time. In *International Conference on Machine Learning*, pages 1754–1762. PMLR, 2018.

[21] C. Gu. *Smoothing spline ANOVA models*, volume 297. Springer Science & Business Media, 2013.

[22] Y. Guermeur. Combinatorial and structural results for gamma-psi-dimensions. *arXiv preprint arXiv:1809.07310*, 2018.

[23] J. Honorio and T. Jaakkola. Structured prediction: from gaussian perturbations to linear-time principled algorithms. *arXiv preprint arXiv:1508.00945*, 2015.

[24] D. Jurafsky and C. Manning. Natural language processing. *Instructor*, 212(998):3482, 2012.

[25] B. T. C. G. D. Koller. Max-margin markov networks. *Advances in neural information processing systems*, 16:25, 2004.

[26] A. N. Kolmogorov and V. M. Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959.

[27] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[28] Y. Lei, L. Ding, and Y. Bi. Local rademacher complexity bounds based on covering numbers. *Neurocomputing*, 218:320–330, 2016.

[29] Y. Lei, Ü. Dogan, A. Binder, and M. Kloft. Multi-class svms: From tighter data-dependent generalization bounds to novel algorithms. *arXiv preprint arXiv:1506.04359*, 2015.

[30] Y. Lei, Ü. Dogan, D.-X. Zhou, and M. Kloft. Data-dependent generalization bounds for multi-class classification. *IEEE Transactions on Information Theory*, 65(5):2995–3021, 2019.

[31] J. Li, Y. Liu, and W. Wang. Learning vector-valued functions with local rademacher complexity. *CoRR*, 2019.

[32] J. Li, Y. Liu, R. Yin, H. Zhang, L. Ding, and W. Wang. Multi-class learning: From theory to algorithm. In *Advances in Neural Information Processing Systems*, pages 1586–1595, 2018.

[33] A. Lucchi, Y. Li, and P. Fua. Learning for structured prediction using approximate subgradient descent with working sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1987–1994, 2013.

[34] A. Maurer. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer, 2016.

[35] D. McAllester. Generalization bounds and consistency. *Predicting structured data*, pages 247–261, 2007.

[36] K. Musayeva. *Generalization Performance of Margin Multi-category Classifiers*. PhD thesis, Université de Lorraine, 2019.

[37] K. Musayeva. Sample complexity result for multi-category classifiers of bounded variation. *arXiv preprint arXiv:2003.09176*, 2020.

[38] K. Musayeva, F. Lauer, and Y. Guermeur. A sharper bound on the rademacher complexity of margin multi-category classifiers. In *ESANN*, 2018.

[39] A. Nowak-Vila, F. Bach, and A. Rudi. A general theory for structured prediction with smooth convex surrogates. *arXiv preprint arXiv:1902.01958*, 2019.

[40] A. Osokin, F. Bach, and S. Lacoste-Julien. On structured prediction theory with calibrated convex surrogate losses. *arXiv preprint arXiv:1703.02403*, 2017.

[41] A. Pakes and D. Pollard. Simulation and the asymptotics of optimization estimators. *Econometrica: Journal of the Econometric Society*, pages 1027–1057, 1989.

[42] H. W. Reeve and A. Kaban. Optimistic bounds for multi-output prediction. *arXiv preprint arXiv:2002.09769*, 2020.

[43] W. Rejchel. On ranking and generalization bounds. *Journal of Machine Learning Research*, 13(5), 2012.

[44] P. Rigollet and J.-C. Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 813:814, 2015.

[45] A. Rudi, C. Ciliberto, G. M. Marconi, and L. Rosasco. Manifold structured prediction. *arXiv preprint arXiv:1806.09908*, 2018.

[46] B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. *arXiv preprint arXiv:1301.0604*, 2012.

[47] I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun, and Y. Singer. Large margin methods for structured and interdependent output variables. *Journal of machine learning research*, 6(9), 2005.

[48] J. Wellner et al. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.

[49] L. Wu, A. Ledent, Y. Lei, and M. Kloft. Fine-grained generalization analysis of vector-valued learning. In *AAAI Conference on Artificial Intelligence*, 2021.

[50] D. Zhang, L. Sun, and W. Li. A structured prediction approach for statistical machine translation. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*, 2008.

[51] T. Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.

[52] D.-X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] This work does not present any foreseeable societal consequence.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A] We did not ran experiments.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A] We did not ran experiments.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] We did not ran experiments.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] We did not ran experiments.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A] We did not use existing assets.
   (b) Did you mention the license of the assets? [N/A] We did not use existing assets.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] We did not use existing assets.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We did not use existing assets.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We did not use existing assets.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] Not applicable.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] Not applicable.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] Not applicable.