Minimax Bounds for Structured Prediction Based on Factor Graphs

Kevin Bello
Department of Computer Science
Purdue Univeristy
West Lafayette, IN, USA
kbellome@purdue.edu

Asish Ghoshal
Department of Computer Science
Purdue Univeristy
West Lafayette, IN, USA
aghoshal@purdue.edu

Jean Honorio
Department of Computer Science
Purdue Univeristy
West Lafayette, IN, USA
jhonorio@purdue.edu

Abstract

Structured prediction can be considered as a generalization of many standard supervised learning tasks, and is usually thought as a simultaneous prediction of multiple labels. One standard approach is to maximize a score function on the space of labels, which usually decomposes as a sum of unary and pairwise potentials, each depending on one or two specific labels, respectively. For this approach, several learning and inference algorithms have been proposed over the years, ranging from exact to approximate methods while balancing the computational complexity. However, in contrast to binary and multiclass classification, results on the necessary number of samples for achieving learning are still limited, even for a specific family of predictors such as factor graphs. In this work, we provide *minimax lower bounds* for a class of *general factor-graph inference models* in the context of structured prediction. That is, we characterize the necessary sample complexity for any conceivable algorithm to achieve learning of general factor-graph predictors.

1 Introduction

Structured prediction has been continuously used over the years in multiple domains such as computer vision, natural language processing, and computational biology. Key examples of structured prediction problems include image segmentation, dependency parsing, part-of-speech tagging, named entity recognition, machine translation and protein folding. In this setting, the input x is some observation, e.g., social network, an image, a sentence. The output is a labeling y, e.g., an assignment of each individual of a social network to a cluster, or an assignment of each pixel in the image to foreground or background, or an acyclic graph as in dependency parsing. A property common to these tasks is that, in each case, the natural loss function admits a decomposition along the output substructures. Thus, a common approach to structured prediction is to exploit local features to infer the global structure. For instance, one could include a feature that encourages two individuals of a social network to be assigned to different clusters whenever there is a strong disagreement in opinions about a particular subject. Then, one can define a posterior distribution over the set of possible labelings conditioned on the input.

The output structure and corresponding loss function make these problems significantly different from the (unstructured) binary or multiclass classification problems extensively studied in learning theory. Some classical algorithms for learning the parameters of the model include conditional random fields (Lafferty et al., 2001), structured support vector machines (Taskar et al., 2003; Tsochantaridis et al., 2005; Altun and Hofmann, 2003), kernel-regression algorithm (Cortes et al., 2007), search-based structured prediction (Daumé et al., 2009). More recently, deep learning algorithms have been developed for specific tasks such as image annotation (Vinyals, Toshev, Bengio and Erhan, 2015), part-of-speech-tagging (Jurafsky and Martin, 2014; Vinyals, Kaiser, Koo, Petrov, Sutskever and Hinton, 2015), and machine translation (Zhang et al., 2008).

However, in contrast to the several algorithms developed, there have been relatively few studies devoted to the theoretical understanding of structured prediction. From the few theoretical literature, the most studied aspect has been the generalization error bounds. (Cortes et al., 2014; Collins, 2004; Taskar et al., 2004) provided learning guarantees that hold primarily for losses such as the Hamming loss. Cortes et al. (2016) presented generalization bounds for more general losses and scoring functions based on factor graphs. Similar to (Cortes et al., 2016), in this work we also study factor graph models, with the difference that we focus on lower bounds and not upper bounds. (McAllester, 2007; Honorio

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

and Jaakkola, 2016; Bello and Honorio, 2018; Ghoshal and Honorio, 2018) provided PAC-Bayesian guarantees for arbitrary losses through the analysis of randomized algorithms using count-based hypotheses.

Literature on lower bounding the sample complexity for structure prediction is scarcer even for specific classes of predictors. Information-theoretic bounds have been studied in the context of binary graphical models (Santhanam and Wainwright, 2012; Tandon et al., 2014) and Gaussian Markov random fields (Wang et al., 2010). Nevertheless, the aforementioned works apply to the modeling of the input x and not the prediction of y from x.

Our main contribution consists of characterizing the necessary sample complexity for learning factor graph models in the context of structured prediction, which to the best of our knowledge, we are the first to find such characterization. Specifically, in Theorem 1, we show that the finiteness of the PAIR-dimension (see Definition 3) is necessary for learning. We further show in Theorem 2 the connection of the PAIR-dimension to the VC-dimension (Vapnik, 2013), which will allow us to compute the PAIR-dimension from the several known results on VC-dimension.

2 Preliminaries

Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ the output space. In structured prediction, the output space usually consists of a large (e.g., exponential) set of discrete objects admitting some possibly overlapping structure. Among common structures in the literature, one finds set of sequences, graphs, images, parse trees, etc. Thus, we consider the output space $\mathcal{Y}$ to be decomposable into l substructures: $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_l$. Here, $\mathcal{Y}_i$ is the set of possible labels that can be assigned to substructure i. For example, in a webpage collective classification task (Taskar et al., 2002), each $\mathcal{Y}_i$ is a webpage label, whereas $\mathcal{Y}$ is a joint label for an entire website. In this work we assume that $\mathcal{Y}_i \in \{0, 1\}$, that is, $|\mathcal{Y}_i| = 2$ for all i. In this case, the number of possible assignments to $\mathcal{Y}$ is exponential in the number of substructures l, i.e., $|\mathcal{Y}| = 2^l$.

The Hamming loss. In order to measure the success of a prediction, we use the Hamming loss throughout this work. Specifically, for two outputs $y, y' \in \mathcal{Y}$, with $y = (y_1, \ldots, y_l)$ and $y' = (y'_1, \ldots, y'_l)$, the Hamming loss, L_H, is defined as:

$$L_H(y, y') = \sum_{i=1}^{l} \mathbf{1}\big[y_i \neq y'_i\big] .$$

The use of Hamming loss in this work is motivated for being widely used in structured prediction problems, for instance, in image segmentation one may count the number of pixels that are incorrectly assigned as foreground/background; in graphs, one may count the number of different edges between the prediction and the true label. For this reason, Globerson et al. (2015) also focused on the Hamming loss for analyzing approximate inference.

Factor graphs and scoring functions. We adopt a common approach in structured prediction where predictions are based on a scoring function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, where $\mathbb{R}_+$ denotes the set of non-negative real numbers. Let $\mathcal{F}$ be a family of scoring functions. For any $f \in \mathcal{F}$, we denote by $\mathsf{f}(x)$ the predictor induced by the scoring function f: for any $x \in \mathcal{X}$,

$$\mathsf{f}(x) \;=\; \arg\max_{y \in \mathcal{Y}} f(x, y).$$

We denote the class of induced predictors by F. Furthermore, we assume that each function $f \in \mathcal{F}$ can be decomposed as a sum, as is standard in structured prediction. We consider the most general case for such decompositions through the notion of factor graphs, motivated also in (Cortes et al., 2016). A factor graph G is a bipartite graph, and is represented as a tuple $G = (V, \Phi, E)$, where V is a set of variable nodes, Φ a set of factor nodes, and E a set of undirected edges between a variable node and a factor node. In our context, V can be identified with the set of substructure indices, that is $V = \{1, \ldots, l\}$. We further assume that G is connected.

For any factor node $\phi \in \Phi$, denote by $\mathrm{Scope}(\phi) \subseteq V$ the set of variable nodes connected to ϕ via an edge and define $\mathcal{Y}_\phi$ as the substructure set cross-product $\mathcal{Y}_\phi = \bigtimes_{i \in \mathrm{Scope}(\phi)} \mathcal{Y}_i$. Then, f decomposes as a sum of functions f_ϕ , each taking as argument an element of the input space $x \in \mathcal{X}$ and an element of $\mathcal{Y}_\phi$, $y_\phi \in \mathcal{Y}_\phi$:

$$f(x, y) = \sum_{\phi \in \Phi} f_\phi(x, y_\phi).$$

We further use $\mathcal{F}(G)$ to denote the set of scoring functions that are decomposable with respect to the graph G, and use $\mathsf{F}(G)$ to denote the set of predictors induced by $\mathcal{F}(G)$. Note also that while all $f \in \mathcal{F}(G)$ decompose with respect to same graph G, the scoring functions f_ϕ and f'_ϕ are allowed to be different for any $\phi \in \Phi$, $f, f' \in \mathcal{F}(G)$. For instance, f_ϕ can be a linear function, while f'_ϕ can be a kernel-based function. Figure 1 shows different examples of factor graphs.

Learning. We receive a training set $S = ((x_1, y_1), \ldots, (x_m, y_m))$ of m i.i.d. samples drawn according to some distribution P over $\mathcal{X} \times \mathcal{Y}$. We denote by $R_P(\mathsf{f})$ the *expected Hamming loss* and by $R_S(\mathsf{f})$ the *empirical Hamming loss* of f:

$$R_P(\mathsf{f}) = \mathop{\mathbb{E}}_{(x,y)\sim P}[L_H(\mathsf{f}(x), y)], \tag{1}$$

$$R_S(\mathsf{f}) = \frac{1}{m} \sum_{(x,y)\in S} L_H(\mathsf{f}(x), y). \tag{2}$$

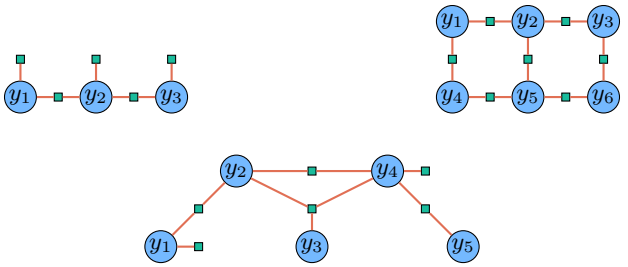

Figure 1: Examples of factor graphs. *(Left)* Tree-structured factor graph. *(Center)* Grid-structured factor graph. *(Right)* Arbitrary factor graph with decomposition: $f(x,y) = f_{\phi_1}(x,y_1)+f_{\phi_4}(x,y_4)+f_{\phi_{12}}(x,y_1,y_2)+f_{\phi_{45}}(x,y_4,y_5)+f_{\phi_{24}}(x,y_2,y_4)+f_{\phi_{234}}(x,y_2,y_3,y_4)$.

Our learning scenario consists of using the sample S to select a scoring function $f \in \mathcal{F}(G)$ with small expected Hamming loss $R_P(\mathsf{f})$.

Next, we introduce the definition of *Bayes-Hamming loss*, which in words is the minimum attainable expected Hamming loss by any predictor.

Definition 1 (Bayes-Hamming loss)**.** *For any given distribution P over $\mathcal{X} \times \mathcal{Y}$, the Bayes-Hamming loss is defined as the minimum achievable expected Hamming loss among all possible predictors $\mathsf{f} : \mathcal{X} \to \mathcal{Y}$. That is, $R^* = \min_{\mathsf{f}} R_P(\mathsf{f})$.*

Then the *Bayes-Hamming predictor*, f^*, is defined as the function that achieves the Bayes-Hamming loss, that is, $R_P(\mathsf{f}^*) = R^*$.

The following proposition shows how the Bayes-Hamming predictor makes its decision with respect to the Hamming loss.

Proposition 1. *For any given distribution P over $\mathcal{X} \times \mathcal{Y}$, the Bayes-Hamming predictor f^* is:*

$$(\mathsf{f}^*(x))_i = \begin{cases} 1 & \text{if } \eta_i(x) \geq 1/2, \\ 0 & \text{otherwise,} \end{cases}$$

where $\eta_i(x) = \mathbb{P}[y_i = 1|x]$ is the marginal probability of substructure y_i.

(See Appendix A for detailed proofs.)

We emphasize that the above definition considers the Hamming loss, L_H, as defined at the beginning of Section 2. For other types of loss functions, the Bayes predictor can have different optimal decisions.

2.1 A Review of the General Minimax Risk Framework

In this section we briefly review the minimax framework in the context of general statistical problems. The minimax framework consists of a well defined objective that aims to shed light about the optimality of algorithms and has been widely used in statistics and machine learning (Wainwright, 2019; Wasserman, 2006). The standard minimax risk considers a family of distributions $\mathcal{Q}$ over a sample space $\mathcal{Z}$, and a function $\theta : \mathcal{Q} \to \Theta$ defined on $\mathcal{Q}$, that is, a mapping $Q \mapsto \theta(Q)$. Here we call $\theta(Q)$ parameter of the distribution Q. We aim to estimate the parameter $\theta(Q)$ based on a sequence of m i.i.d. observations $Z = (z_1, \ldots, z_m)$ drawn from the (unknown) distribution Q, that is, $Z \in \mathcal{Z}^m$. To evaluate the quality of an estimator $\widehat{\theta}$, we let $\rho : \Theta \times \Theta \to \mathbb{R}_+$ denote a semi-metric on the space Θ, which we use to measure the error of an estimator $\widehat{\theta}$ with respect to the parameter $\theta(Q)$. For a distribution $Q \in \mathcal{Q}$ and for a given estimator $\widehat{\theta} : \mathcal{Z}^m \to \Theta$, we assess the quality of the estimate $\widehat{\theta}(Z)$ in terms of the (expected) risk:

$$\mathbb{E}_{Z\sim Q^m}\left[\rho(\widehat{\theta}(Z), \theta(Q))\right].$$

A common approach, first suggested by (Wald, 1939), for choosing an estimator $\widehat{\theta}$ is to select the one that minimizes the maximum risk, that is,

$$\sup_{Q\in\mathcal{Q}} \mathbb{E}_{Z\sim Q^m}\left[\rho(\widehat{\theta}(Z), \theta(Q))\right].$$

An optimal estimator for this semi-metric then gives the minimax risk, which is defined as:

$$\mathfrak{M}_m(\mathcal{Q}, \rho) := \inf_{\widehat{\theta}} \sup_{Q\in\mathcal{Q}} \mathbb{E}_{Z\sim Q^m}\left[\rho(\widehat{\theta}(Z), \theta(Q))\right],$$

where we take the supremum (worst-case) over distributions $Q \in \mathcal{Q}$, and the infimum is taken over all estimators $\widehat{\theta}$.

2.2 Minimax Risk in Structured Prediction

We now apply the framework above to our context and study a specialized notion of risk appropriate for prediction problems. In this setting, we aim to estimate a scoring function $f \in \mathcal{F}(G)$ by using samples from a distribution P. For any sample $(x, y) \sim P$, we will measure the quality of our estimation, f, by comparing the prediction $\mathsf{f}(x)$ to the structure y drawn from P through the Hamming loss. By taking expectation, we obtain the expected risk or expected Hamming loss, $R_P(\mathsf{f})$, defined in eq.(1). We then compare this risk to the best possible Hamming loss, i.e., the Bayes-Hamming loss (Definition 1). That is, we assume that at least one scoring function in $\mathcal{F}(G)$ achieves the Bayes-Hamming loss. Finally, recall that $S \in (\mathcal{X} \times \mathcal{Y})^m$ is the training set consisting of m i.i.d. samples drawn from P. Thus, we arrive to the following *minimax excess risk*:

$$\mathfrak{M}_m(\mathcal{P}) = \inf_{\mathcal{A}} \sup_{P\in\mathcal{P}} \mathbb{E}_{S\sim P^m}\left[R_P(\mathcal{A}(S)) - R_P(\mathsf{f}^*)\right], \quad (3)$$

where f^* is the induced predictor by the scoring function $f^* = \arg\min_{f\in\mathcal{F}(G)} R_P(\mathsf{f})$[1], and $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{F}(G)$

[1]Recall that f denotes the induced predictor by $f \in \mathcal{F}(G)$.

is any algorithm that returns a predictor given m training samples from P. Moreover, $\mathcal{P}$ defines a family of distributions over $\mathcal{X} \times \mathcal{Y}$.
Intuitively speaking, for a fixed distribution $P \in \mathcal{P}$, the quantity $\mathfrak{M}_m(\mathcal{P})$ represents the minimum expected excess loss achievable by any algorithm with respect to the factor graph G. Then $\mathfrak{M}_m(\mathcal{P})$ looks into the distribution that attains the worst expected excess loss.

3 Information-Theoretic Lower Bound for Structured Prediction

We are interested on finding a lower bound to the minimax risk (3) presented in Section 2.2. By doing this, we characterize the necessary number of samples to have any hope in achieving learning.

Before presenting our main result, we introduce a new type of dimension that will show up in our lower bound and will help to characterize learnability. Note that it is known that different notions of dimension of predictor classes help to characterize learnability in certain prediction problems. For example, in binary classification with the $0/1$-loss, the finiteness of the VC dimension (Vapnik, 2013) is necessary for learning (Massart et al., 2006). For multiclass classification, it was shown that the finiteness of the Natarajan dimension is necessary for learning (Daniely et al., 2015). General notions of dimensions for multiclass classification has also been study in (Ben-David et al., 1995).

For a given predictor class $\mathcal{G} \subseteq \{\mathsf{g} \mid \mathsf{g} : \mathcal{X} \to \{0,1\}^2\}$, and dataset S of m samples, we use the following shorthand notation:

$$\mathcal{G}(S) = \{(\mathsf{g}(x_1), \ldots, \mathsf{g}(x_m)) \in \{0,1\}^{m \times 2} \mid \mathsf{g} \in \mathcal{G}\}.$$

That is, $\mathcal{G}(S)$ contains all the matrices in $\{0,1\}^{m\times 2}$ that can be produced by applying all functions in $\mathcal{G}$ to the dataset S. Next we define PAIR-shattering.

Definition 2 (PAIR-shattering)**.** *A function class, $\mathcal{G}$,* PAIR*-shatters a finite set S of m samples if $\mathcal{G}(S)$ produces all possible binary matrices in $\{0,1\}^{m\times 2}$. That is, $|\mathcal{G}(S)| = 2^{2m}$.*

Definition 3 (PAIR-dimension)**.** *The* PAIR*-dimension of a function class $\mathcal{G}$, denoted* $\textsc{PairDim}(\mathcal{G})$*, is the maximal size of a set S that can be* PAIR*-shattered by $\mathcal{G}$. If $\mathcal{G}$ can shatter sets of arbitrarily large size we say that $\mathcal{G}$ has infinite* PAIR*-dimension.*

The above dimension applies to predictors with output in $\{0,1\}^2$. Next, we define the MAX-PAIR-dimension for classes of predictors $\mathcal{H} \subseteq \{\mathsf{h} \mid \mathsf{h} : \mathcal{X} \to \{0,1\}^l\}$.

Definition 4 (MAX-PAIR-dimension)**.** *For a predictor class $\mathcal{H} \subseteq \{\mathsf{h} \mid \mathsf{h} : \mathcal{X} \to \{0,1\}^l\}$, the* MAX-PAIR*-dimension of $\mathcal{H}$, denoted as* $\textsc{max-PairDim}(\mathcal{H})$*, is defined as:*

$$\textsc{max-PairDim}(\mathcal{H}) = \max_{\substack{u,v\in\{1,\ldots,l\}\\ u\neq v}} \textsc{PairDim}(\mathcal{H}_{u,v}),$$

where $\mathcal{H}_{u,v} = \{\mathsf{h}_{u,v} \mid \mathsf{h} \in \mathcal{H}, \mathsf{h}_{u,v} : \mathcal{X} \to \{0,1\}^2, \mathsf{h}_{u,v}(x) = (\mathsf{h}(x)_u, \mathsf{h}(x)_v)\}$, that is, the predictor $\mathsf{h}_{u,v}$ only takes into account the output of h at positions u and v, and becomes a mapping from $\mathcal{X}$ to $\{0,1\}^2$.

Note that Definition 4 is stated for general classes of predictors with output in $\{0,1\}^l$. However, in our context we consider predictors induced by scoring functions based on factor graphs. That is, for a predictor $\mathsf{f} : \mathcal{X} \to \{0,1\}^l$ induced by the scoring function f, we will create predictors with output in $\{0,1\}^2$ as follows. Let

$$f^{(0)}_{u,v}(x, y_u, y_v) \stackrel{\text{def}}{=} f(x, (0, \ldots, 0, y_u, 0, \ldots, 0, y_v, 0, \ldots, 0))$$

denote the scoring function $f(x, y)$ with $y_i = 0$ for all $i \in \{1, \ldots, l\} \setminus \{u, v\}$. Then, let

$$\mathsf{f}^{(0)}_{u,v}(x) = \arg\max_{y_u, y_v} f^{(0)}_{u,v}(x, y_u, y_v)$$

be the induced predictor by $f^{(0)}_{u,v}(x, y_u, y_v)$, i.e., the output of $\mathsf{f}^{(0)}_{u,v}(x)$ is in $\{0,1\}^2$.

Remark 1 (MAX-PAIR-dimension for scoring functions based on factor graphs.)**.** *For a given factor graph $G = (V, \Phi, E)$ such that $T = \{(u, v) \mid u \neq v, \{u, v\} \subseteq \text{Scope}(\phi), \phi \in \Phi\}$, $\mathcal{F}^{(0)}_{u,v}(G) = \{f^{(0)}_{u,v} \mid f \in \mathcal{F}(G)\}$, and let $\mathsf{F}^{(0)}_{u,v}(G) = \{\mathsf{f}^{(0)}_{u,v} \mid f^{(0)}_{u,v} \in \mathcal{F}^{(0)}_{u,v}(G)\}$ denote the set of predictors induced by $\mathcal{F}^{(0)}_{u,v}(G)$. Then, the* MAX-PAIR*-dimension of a class of scoring functions is given by the* MAX-PAIR*-dimension of the class of predictors it induces, i.e.,*

$$\textsc{max-PairDim}(\mathsf{F}(G)) = \max_{(u,v)\in T} \textsc{PairDim}(\mathsf{F}^{(0)}_{u,v}(G)).$$

Next, we present our main result which provides a characterization on the necessary number of samples for learning.

Theorem 1. *Let $G = (V, \Phi, E)$ be a factor graph and let $\mathcal{F}(G)$ denote a class of scoring functions where each $f \in \mathcal{F}(G)$ decomposes according to G. Let $\mathsf{F}(G)$ be the induced class of predictors by $\mathcal{F}(G)$, where $\mathsf{f} : \mathcal{X} \to \{0,1\}^l$ for each $\mathsf{f} \in \mathsf{F}(G)$, and let $d = \textsc{max-PairDim}(\mathsf{F}(G)) \geq 2$. Then, we have that for any $\gamma \in [0, 1/3]$ and any $m \geq d$:*

$$\mathfrak{M}_m(\mathcal{P}) \geq \frac{1}{81} \min\left(\frac{d-1}{\gamma m}, \sqrt{\frac{d-1}{m}}\right).$$

Proof. The proof is motivated by the work of Massart et al. (2006) for binary classifiers. As a first step it is clear that one can lower bound eq.(3) by defining the maximum over a

subset of $\mathcal{P}$. That is, we create a collection of family of distributions $\mathbb{D}_\gamma$, where $|\mathbb{D}_\gamma| = |\Phi|$. Each family distribution $\mathcal{D}_{\gamma,u,v} \in \mathbb{D}_\gamma$ is further indexed by $(u,v) \in T = \{(u,v) \mid u \neq v, \{u,v\} \subseteq \text{Scope}(\phi), \phi \in \Phi\}$. Then we have,

$$\mathfrak{M}_m(\mathcal{P}) \geq \max_{(u,v)\in T} \mathfrak{M}_m(\mathcal{D}_{\gamma,u,v}).$$

Our approach consists of first defining the families of distributions $\mathcal{D}_{\gamma,u,v} \subset \mathcal{P}$ such that its elements can be naturally indexed by the vertices of a binary hypercube. We will then relate the expected excess risk problem to an estimation of binary strings in order to apply Assouad's lemma.

Construction of $\mathcal{D}_{\gamma,u,v}$. Consider a fixed $(u,v) \in T$. We first focus on constructing a family of distributions, $\mathcal{D}_{\gamma,u,v}$, parameterized by $\gamma > 0$. Each distribution $D_{\gamma,u,v,B} \in \mathcal{D}_{\gamma,u,v}$ is further indexed by a binary matrix $B \in \{0,1\}^{(d^{(0)}_{u,v}-1)\times 2}$, where $d^{(0)}_{u,v}$ is the PAIR-dimension of $\mathsf{F}^{(0)}_{u,v}$. To construct these distributions, we will first pick the marginal distribution $D^{(x)}_{\gamma,u,v,B}$ of the feature x, and then specify the conditional distributions $D^{(y|x)}_{\gamma,u,v,B}$ of y given x, for each $B \in \{0,1\}^{(d^{(0)}_{u,v}-1)\times 2}$.

We construct $D^{(x)}_{\gamma,u,v,B}$ as follows. Since $\mathsf{F}^{(0)}_{u,v}$ is a class with PAIR-dimension $d^{(0)}_{u,v}$, there exists a set of points $\{x_1, \ldots, x_{d^{(0)}_{u,v}}\} \in \mathcal{X}$ that is shattered by $\mathsf{F}^{(0)}_{u,v}$, that is, for any binary matrix $B \in \{0,1\}^{d^{(0)}_{u,v}\times 2}$ there exists at least one function $\mathsf{f}^{(0)}_{u,v} \in \mathsf{F}^{(0)}_{u,v}$ such that $\mathsf{f}^{(0)}_{u,v}(x_i) = B_{i*}$, for all $i \in \{1, \ldots, d^{(0)}_{u,v}\}$.
We now define the marginal distribution $D^{(x)}_{\gamma,u,v,B}$ such that its support is the shattered set $\{x_1, \ldots, x_{d^{(0)}_{u,v}}\}$, i.e., $\mathbb{P}^{(x)}_{\gamma,u,v,B}[\{x_1, \ldots, x_{d^{(0)}_{u,v}}\}] = 1$. For a given parameter $p \in [0, 1/(d^{(0)}_{u,v}-1)]$, whose value is set later, we have:

$$\mathbb{P}^{(x)}_{\gamma,u,v,B}[x_i] = \begin{cases} p, & \text{if } i \in \{1, \ldots, d^{(0)}_{u,v} - 1\}, \\ 1 - (d^{(0)}_{u,v} - 1)p, & \text{otherwise.} \end{cases}$$

Next, for a fixed $B \in \{0,1\}^{(d^{(0)}_{u,v}-1)\times 2}$, the conditional distribution of y given x, $D^{(y|x)}_{\gamma,u,v,B}$, is defined as:

$$\mathbb{P}^{(y|x)}_{\gamma,u,v,B}[y|x] = \begin{cases} \frac{1-3\gamma}{4}, & \text{if } x = x_i, y_u = 1 - B_{i1}, y_v = 1 - B_{i2}, \\ & y_k = 0 \text{ for } k \in V \setminus \{u,v\}, \\ & i \in \{1, \ldots, d^{(0)}_{u,v} - 1\}, \\ \frac{1+\gamma}{4}, & \text{if } x = x_i, (y_u \neq 1 - B_{i1} \text{ or } y_v \neq 1 - B_{i2}), \\ & y_k = 0 \text{ for } k \in V \setminus \{u,v\}, \\ & i \in \{1, \ldots, d^{(0)}_{u,v} - 1\}, \\ 0, & \text{otherwise,} \end{cases}$$

here we implicitly assume that $\gamma \in (0, 1/3]$ in order to obtain a valid distribution. The above definition produces the following marginal probabilities:

$$\eta_j^{(\gamma,u,v,B)}(x) \equiv \mathbb{P}^{(y_j|x)}_{\gamma,u,v,B}[y_j = 1|x] = \begin{cases} \frac{1-\gamma}{2}, & \text{if } x = x_i \text{ for some } i \in \{1, \ldots, d^{(0)}_{u,v} - 1\}, \\ & ((j = u, B_{i1} = 0) \text{ or } (j = v, B_{i2} = 0)), \\ \frac{1+\gamma}{2}, & \text{if } x = x_i \text{ for some } i \in \{1, \ldots, d^{(0)}_{u,v} - 1\}, \\ & ((j = u, B_{i1} = 1) \text{ or } (j = v, B_{i2} = 1)), \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where we note that for each $j \in V$ and any x we have that $|2\eta_j^{(\gamma,u,v,B)}(x) - 1| \geq \gamma$. Given the above marginals, the corresponding Bayes-Hamming predictor for substructure y_j for a given input x (see Proposition 1), which we denote by $(\mathsf{f}^*_{B,u,v}(x))_j$, is given by:

$$(\mathsf{f}^*_{B,u,v}(x))_j = \begin{cases} 0, & \text{if } x = x_i \text{ for some } i \in \{1, \ldots, d^{(0)}_{u,v} - 1\}, \\ & ((j = u \text{ and } B_{i1} = 0) \text{ or } (j = v \text{ and } B_{i2} = 0)) \\ 1, & \text{if } x = x_i \text{ for some } i \in \{1, \ldots, d^{(0)}_{u,v} - 1\}, \\ & ((j = u \text{ and } B_{i1} = 1) \text{ or } (j = v \text{ and } B_{i2} = 1)) \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

That is, we have that the output of the Bayes-Hamming predictor on each x_i for $i \in \{1 \ldots d^{(0)}_{u,v} - 1\}$, for each substructure y_j for $j \in \{u,v\}$, is equal to the bit value B_{i1} or B_{i2}, and zero otherwise.

Reduction to estimation of binary strings. For any distribution $D_{\gamma,u,v,B} \in \mathcal{D}_{\gamma,u,v}$, we can further express the expected excess risk in eq.(3) as follows:

$$\begin{aligned} &R_{B,u,v}(\mathcal{A}(S)) - R_{B,u,v}(\mathsf{f}^*_{B,u,v}) = \\ &\mathop{\mathbb{E}}_{(x,y)\sim D_{\gamma,u,v,B}}\left[\sum_{j=1}^{l}(1-2y_j)((\hat{\mathsf{f}}_m(x))_j - (\mathsf{f}^*_{B,u,v}(x))_j)\right] \\ &= \sum_{j=1}^{l} \mathop{\mathbb{E}}_{x\sim D^{(x)}_{\gamma,u,v,B}} \Bigg[\\ &\qquad \mathop{\mathbb{E}}_{y_j\sim D^{(y_j|x)}_{\gamma,u,v,B}}\left[(1-2y_j)((\hat{\mathsf{f}}_m(x))_j - (\mathsf{f}^*_{B,u,v}(x))_j)\right]\Bigg] \\ &= \sum_{j=1}^{l} \mathop{\mathbb{E}}_{x\sim D^{(x)}_{\gamma,u,v,B}} \Big[\\ &\qquad \left|2\eta_j^{(\gamma,u,v,B)}(x) - 1\right| \cdot \left|(\hat{\mathsf{f}}_m(x))_j - (\mathsf{f}^*_{B,u,v}(x))_j\right| \Big] \\ &\geq \gamma \cdot \mathop{\mathbb{E}}_{x\sim D^{(x)}_{\gamma,u,v,B}}\left[\sum_{j=1}^{l}\left|(\hat{\mathsf{f}}_m(x))_j - (\mathsf{f}^*_{B,u,v}(x))_j\right|\right] \qquad (6) \\ &= \gamma \cdot \sum_{i=1}^{d^{(0)}_{u,v}}\sum_{j=1}^{l}\left|(\hat{\mathsf{f}}_m(x_i))_j - (\mathsf{f}^*_{B,u,v}(x_i))_j\right| \cdot \mathbb{P}^{(x)}_{\gamma,u,v,B}[x_i], \end{aligned}$$

$$\stackrel{\text{def}}{=} \gamma \cdot \|\hat{\mathsf{f}}_m - \mathsf{f}^*_{B,u,v}\|_{1,1}, \tag{7}$$

where $R_{B,u,v}$ denotes the expected risk and $\mathsf{f}^*_{B,u,v}$ the Bayes-Hamming predictor, both with respect to $D_{\gamma,u,v,B}$. Here $\hat{\mathsf{f}}_m$ is the output of $\mathcal{A}(S)$, with $(\hat{\mathsf{f}}_m(x))_j$ denoting the j-th substructure of the output $\hat{\mathsf{f}}_m(x)$, and $\eta_j^{(\gamma,u,v,B)}(x)$ denotes the marginal probability $\mathbb{P}_{D^{(y_j|x)}_{\gamma,u,v,B}}[y_j = 1|x]$. Equation (6) follows from our definition of $D^{(y_j|x)}_{\gamma,u,v,B}$ (see eq.(4)), and the $L_{1,1}$ matrix norm in eq.(7) is computed with respect to $D^{(x)}_{\gamma,u,v,B}$. Thus, we have that:

$$\begin{aligned}
&\mathfrak{M}_m(\mathcal{D}_{\gamma,u,v}) = \\
&\inf_{\hat{\mathsf{f}}_m} \max_{B\in\{0,1\}^{(d^{(0)}_{u,v}-1)\times 2}} \mathbb{E}_{B,u,v}\left[R_{B,u,v}(\hat{\mathsf{f}}_m) - R_{B,u,v}(\mathsf{f}^*_{B,u,v})\right] \\
&\geq \gamma \cdot \inf_{\hat{\mathsf{f}}_m} \max_{B\in\{0,1\}^{(d^{(0)}_{u,v}-1)\times 2}} \mathbb{E}_{B,u,v}\left[\|\hat{\mathsf{f}}_m - \mathsf{f}^*_{B,u,v}\|_{1,1}\right],
\end{aligned} \tag{8}$$

where $\mathbb{E}_{B,u,v}[\cdot]$ denotes the expectation with respect to $S \sim D^m_{\gamma,u,v,B}$. Equation (8) follows from eq.(7). Given any candidate estimation $\hat{\mathsf{f}}_m$, let $\widehat{B}_m \in \{0,1\}^{(d^{(0)}_{u,v}-1)\times 2}$ be defined as follows:

$$\widehat{B}_m \stackrel{\text{def}}{=} \operatorname*{arg\,min}_{B\in\{0,1\}^{(d^{(0)}_{u,v}-1)\times 2}} \|\hat{\mathsf{f}}_m - \mathsf{f}^*_{B,u,v}\|_{1,1}. \tag{9}$$

Intuitively, $\widehat{B}_m$ is the binary matrix that indexes the element of $\{\mathsf{f}^*_{B,u,v} : B \in \{0,1\}^{(d^{(0)}_{u,v}-1)\times 2}\}$ which is the closest to $\hat{\mathsf{f}}_m$ in $L_{1,1}$ norm. Then, for any B, we have

$$\begin{aligned}
&\|\mathsf{f}^*_{\widehat{B}_m,u,v} - \mathsf{f}^*_{B,u,v}\|_{1,1} \\
&\qquad \leq \|\mathsf{f}^*_{\widehat{B}_m,u,v} - \hat{\mathsf{f}}_m\|_{1,1} + \|\hat{\mathsf{f}}_m - \mathsf{f}^*_{B,u,v}\|_{1,1} \\
&\qquad \leq 2\|\hat{\mathsf{f}}_m - \mathsf{f}^*_{B,u,v}\|_{1,1},
\end{aligned}$$

where we first applied the triangle inequality, and then used eq.(9). Applying this to eq.(8), we obtain:

$$\begin{aligned}
&\mathfrak{M}_m(\mathcal{D}_{\gamma,u,v}) \\
&\quad \geq \frac{\gamma}{2} \min_{\widehat{B}_m} \max_{B\in\{0,1\}^{(d^{(0)}_{u,v}-1)\times 2}} \mathbb{E}_{B,u,v}\left[\|\mathsf{f}^*_{\widehat{B}_m,u,v} - \mathsf{f}^*_{B,u,v}\|_{1,1}\right],
\end{aligned} \tag{10}$$

here the infimum is over all estimators that take values in $\{0,1\}^{(d^{(0)}_{u,v}-1)\times 2}$ based on m samples, i.e., over $\widehat{B}_m : (\mathcal{X}\times\mathcal{Y})^m \to \{0,1\}^{(d^{(0)}_{u,v}-1)\times 2}$. We now compute $\|\mathsf{f}^*_{B,u,v} - \mathsf{f}^*_{B',u,v}\|_{1,1}$ for any two B, B'. Using eq.(5) we have:

$$\begin{aligned}
&\|\mathsf{f}^*_{B,u,v} - \mathsf{f}^*_{B',u,v}\|_{1,1} \\
&= \sum_{i=1}^{d^{(0)}_{u,v}} \sum_{j=1}^{l} \left|(\mathsf{f}^*_{B,u,v}(x_i))_j - (\mathsf{f}^*_{B',u,v}(x_i))_j\right| \cdot \mathbb{P}^{(x)}_{\gamma,u,v,B}[x_i] \\
&= p \cdot \sum_{i=1}^{d^{(0)}_{u,v}-1} \sum_{j=1}^{2} \left|B_{ij} - B'_{ij}\right| \\
&= p \cdot L_H(B, B').
\end{aligned}$$

In the last equality we abuse notation and consider the matrix $B \in \{0,1\}^{(d^{(0)}_{u,v}-1)\times 2}$ as a vector of dimension $2(d^{(0)}_{u,v}-1)$. Replacing this result into eq.(10), we get:

$$\mathfrak{M}_m(\mathcal{D}_{\gamma,u,v}) \geq \frac{p\gamma}{2} \min_{\widehat{B}_m} \max_{B\in\{0,1\}^{(d^{(0)}_{u,v}-1)\times 2}} \mathbb{E}_{B,u,v}[L_H(\widehat{B}_m, B)],$$

which is related to an estimation problem in the $\{0,1\}^{2(d^{(0)}_{u,v}-1)}$ hypercube.

Applying Assouad's lemma. In order to apply Assouad's lemma, we need an upper bound on the squared Hellinger distance $H^2(D_{\gamma,u,v,B}, D_{\gamma,u,v,B'})$ for all B, B' with $L_H(B, B') = 1$. For any two $B, B' \in \{0,1\}^{(d^{(0)}_{u,v}-1)\times 2}$ we have:

$$\begin{aligned}
&H^2(D_{\gamma,u,v,B}, D_{\gamma,u,v,B'}) \\
&= \sum_{i=1}^{d^{(0)}_{u,v}} \sum_{y\in\{0,1\}^l} \left(\sqrt{\mathbb{P}_{\gamma,u,v,B}(x_i,y)} - \sqrt{\mathbb{P}_{\gamma,u,v,B'}(x_i,y)}\right)^2 \\
&= p \sum_{i=1}^{d^{(0)}_{u,v}-1} \sum_{y\in\{0,1\}^l} \left(\sqrt{\mathbb{P}^{(y|x)}_{\gamma,u,v,B'}(y|x_i)} - \sqrt{\mathbb{P}^{(y|x)}_{\gamma,u,v,B}(y|x_i)}\right)^2.
\end{aligned}$$

In the above summation, the inner sum is zero if $B_{i*} = B'_{i*}$. Since we are interested on B and B' such that $L_H(B, B') = 1$, this implies that for only one row i from $\{1, \ldots, d^{(0)}_{u,v}-1\}$ we have $B_{i*} \neq B'_{i*}$ with exactly one bit different. Then, the Hellinger distance results in: $H^2(D_{\gamma,u,v,B}, D_{\gamma,u,v,B'}) = p\cdot(1-\gamma-\sqrt{1-2\gamma-3\gamma^2}) \leq 6p\gamma^2$. Applying Assouad's lemma we obtain:

$$\begin{aligned}
&\mathfrak{M}_m(\mathcal{D}_{\gamma,u,v}) \\
&\quad \geq \frac{p\gamma}{2} \min_{\widehat{B}_m} \max_{B\in\{0,1\}^{(d^{(0)}_{u,v}-1)\times 2}} \mathbb{E}_{B,u,v}\left[L_H(\widehat{B}_m, B)\right] \\
&\quad \geq \frac{p\gamma(d^{(0)}_{u,v}-1)}{2}(1-\sqrt{6p\gamma^2 m})
\end{aligned} \tag{11}$$

Let $p = 2/(27\gamma^2 m)$, and noting that if $\gamma \geq \sqrt{(d^{(0)}_{u,v}-1)/m}$ then the condition $p \leq 1/(d^{(0)}_{u,v}-1)$ holds. Replacing p in eq.(11) we have:

$$\mathfrak{M}_m(\mathcal{D}_{\gamma,u,v}) \geq \frac{d^{(0)}_{u,v}-1}{81\gamma m}. \tag{12}$$

If $\gamma \leq \sqrt{(d^{(0)}_{u,v}-1)/m}$, and using the same construction as above with $\widetilde{\gamma} = \sqrt{(d^{(0)}_{u,v}-1)/m}$, we see that:

$$\mathfrak{M}_m(\mathcal{D}_{\gamma,u,v}) \geq \frac{d^{(0)}_{u,v}-1}{81\widetilde{\gamma} m} = \frac{1}{81}\sqrt{\frac{d^{(0)}_{u,v}-1}{m}}. \tag{13}$$

Therefore, combining equations (12) and (13), and since the choice of (u, v) was arbitrary, we have that:

$$\begin{aligned}\mathfrak{M}_m(\mathcal{P}) &\geq \max_{(u,v)\in T} \mathfrak{M}_m(\mathcal{D}_{\gamma,u,v}) \\ &\geq \max_{(u,v)\in T} \frac{1}{81} \min\left(\frac{d_{u,v}^{(0)}-1}{\gamma m}, \sqrt{\frac{d_{u,v}^{(0)}-1}{m}}\right) \\ &= \frac{1}{81}\min\left(\frac{d-1}{\gamma m}, \sqrt{\frac{d-1}{m}}\right).\end{aligned}$$

□

4 Relation of Pair-Dimension to VC-Dimension

In this section, we show a connection of our defined PAIR-dimension to the classical VC-dimension (Vapnik, 2013).

The following theorem shows that for a function class $\mathcal{G} \subseteq \{\mathsf{g} \mid \mathsf{g} : \mathcal{X} \to \{0,1\}^2\}$, the PAIR-dimension of $\mathcal{G}$ is related to the minimum VC-dimension of a subclass of functions derived from $\mathcal{G}$.

Theorem 2. *Let $\mathcal{G} \subseteq \{\mathsf{g} \mid \mathsf{g} : \mathcal{X} \to \{0,1\}^2\}$ be a function class. Let $\mathcal{H}_{11}, \mathcal{H}_{10}, \mathcal{H}_{01}, \mathcal{H}_{00} \subseteq \{\mathsf{h} \mid \mathsf{h} : \mathcal{X} \to \{0,1\}\}$ be four function classes defined as:*

$$\begin{aligned}\mathcal{H}_{11} = \{&\mathsf{h} \mid \mathsf{h} : \mathcal{X} \to \{0,1\}, \\ &\mathsf{h}(x) = \mathsf{g}(x)_1 \mathsf{g}(x)_2, \mathsf{g} \in \mathcal{G}\}, \\ \mathcal{H}_{10} = \{&\mathsf{h} \mid \mathsf{h} : \mathcal{X} \to \{0,1\}, \\ &\mathsf{h}(x) = \mathsf{g}(x)_1 (1 - \mathsf{g}(x)_2), \mathsf{g} \in \mathcal{G}\}, \\ \mathcal{H}_{01} = \{&\mathsf{h} \mid \mathsf{h} : \mathcal{X} \to \{0,1\}, \\ &\mathsf{h}(x) = (1 - \mathsf{g}(x)_1) \mathsf{g}(x)_2, \mathsf{g} \in \mathcal{G}\}, \\ \mathcal{H}_{00} = \{&\mathsf{h} \mid \mathsf{h} : \mathcal{X} \to \{0,1\}, \\ &\mathsf{h}(x) = (1 - \mathsf{g}(x)_1)(1 - \mathsf{g}(x)_2), \mathsf{g} \in \mathcal{G}\}.\end{aligned}$$

We have that $\text{PAIRDIM}(\mathcal{G}) = \min_{i,j\in\{0,1\}} \text{VC-DIM}(\mathcal{H}_{ij})$.

Proof. Recall that for a dataset S of m samples, $\mathcal{G}(S) = \{(\mathsf{g}(x_1), \ldots, \mathsf{g}(x_m)) \in \{0,1\}^{m\times 2} \mid \mathsf{g} \in \mathcal{G}\}$. Similarly, define $\mathcal{H}_{ij}(S) = \{(\mathsf{h}(x_1), \ldots, \mathsf{h}(x_m)) \in \{0,1\}^m \mid \mathsf{h} \in \mathcal{H}_{ij}\}$ for all $i, j \in \{0,1\}$. Let $\text{PAIRDIM}(\mathcal{G}) = d$.

There exists a dataset S of d samples such that $|\mathcal{G}(S)| = 2^{2d}$. Thus for all $i, j \in \{0,1\}$ we have $|\mathcal{H}_{ij}(S)| = 2^d$, which implies that for all $i, j \in \{0,1\}$ we have $\text{VC-DIM}(\mathcal{H}_{ij}) \geq d$. Therefore,

$$d \leq \min_{i,j\in\{0,1\}} \text{VC-DIM}(\mathcal{H}_{ij}).$$

Also, for any dataset S of $d+1$ samples we have $|\mathcal{G}(S)| < 2^{2(d+1)}$. Thus there exists $i, j \in \{0,1\}$ such that $|\mathcal{H}_{ij}(S)| < 2^{d+1}$, implying that there exists $i, j \in \{0,1\}$ such that $\text{VC-DIM}(\mathcal{H}_{ij}) < d+1$. Therefore,

$$\min_{i,j\in\{0,1\}} \text{VC-DIM}(\mathcal{H}_{ij}) < d+1.$$

From the above, $\min_{i,j\in\{0,1\}} \text{VC-DIM}(\mathcal{H}_{ij}) = d$. □

5 Discussion

We consider the problem of finding the necessary number of samples for learning of scoring functions based on factor graphs in the context of structured prediction. Our work was based on the minimax framework, that is, in obtaining a lower bound to the minimax risk. We showed a lower bound that requires the MAX-PAIR-dimension to be finite in order for a function class to be learnable. We also note that in the proof of Theorem 1, our choice of setting a value of zero to many y's was for clarity purposes. In principle, one can create such distributions by fixing y's to arbitrary values in $\{0,1\}^{l-2}$, and this would result in a slightly different notion of dimension, which would take the maximum across the 2^{l-2} different values. However, our focus was on providing a clear guideline to obtain lower bounds in structured prediction, hence, we opted for simplicity. In addition, in Theorem 2 we showed the connection of the PAIR-dimension to the VC-dimension, for which there are several known results for different types of function classes.

An interesting future work is the analysis of tightness. For example, regarding tightness for linear classifiers, consider inputs $x \in \mathbb{R}^k$. We observe that our lower bound in Theorem 1 is tight with respect to k and m. Specifically, consider non-sparse linear classifiers as the scoring functions, Theorem 2 in (Cortes et al., 2016) gives $\mathcal{O}(\sqrt{k/m})$. In this case, the PAIR-dimension is equal to the VC-dimension, and the latter is equal to k. Thus, we obtain a lower bound with rate $\sqrt{k/m}$ for some γ. Similarly, consider sparse linear classifiers as the scoring functions. Then, Theorem 2 of (Cortes et al., 2016) gives $\mathcal{O}(\sqrt{\log k/m})$. In this case, the VC-dimension is $\mathcal{O}(\log k)$ (Neylon, 2006), thus, we obtain a lower bound with rate $\sqrt{\log k/m}$ for some γ. However, it remains to analyze for general functions where one possible attempt is perhaps to find an upper bound to the *factor graph Rademacher complexity* (Cortes et al., 2016) in terms of the PAIR-dimension, similar in spirit to the known result of the VC-dimension being an upper bound of the classical Rademacher complexity (see for instance, (Shalev-Shwartz and Ben-David, 2014)).

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1716609-IIS.

A Detailed Proofs

A.1 Proof of Proposition 1

Proof. Recall that $\eta_i(x) = \mathbb{P}[y_i = 1|x]$. From eq.(1) and Definition 1, the Bayes-Hamming predictor f^* minimizes the following expression (with respect to f).

$$\begin{aligned}
R_P(\mathsf{f}) &= \mathop{\mathbb{E}}_{(x,y)\sim P}[L_H(\mathsf{f}(x), y)] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim P}\left[\sum_{i=1}^{l} \mathbf{1}\big[(\mathsf{f}(x))_i \neq y_i\big]\right] \\
&= \sum_{i=1}^{l} \mathop{\mathbb{E}}_{(x,y)\sim P}\Big[\mathbf{1}\big[(\mathsf{f}(x))_i \neq y_i\big]\Big] \\
&= \sum_{i=1}^{l} \mathop{\mathbb{E}}_{x}\big[\mathbb{P}[y_i = 1|x](1 - (\mathsf{f}(x))_i) \\
&\qquad\qquad +(1 - \mathbb{P}[y_i = 1|x])(\mathsf{f}(x))_i\big] \\
&= \sum_{i=1}^{l} \mathop{\mathbb{E}}_{x}\left[\eta_i(x)(1 - (\mathsf{f}(x))_i) + (1 - \eta_i(x))(\mathsf{f}(x))_i\right].
\end{aligned}$$

In order to minimize the above expression, for any x we choose $(\mathsf{f}(x))_i = 1$ if $\eta_i(x) \geq 1/2$, and $(\mathsf{f}(x))_i = 0$ otherwise. □

References

Altun, Y. and Hofmann, T. (2003), 'Large margin methods for label sequence learning', *European Conference on Speech Communication and Technology* pp. 145–152.

Bello, K. and Honorio, J. (2018), Learning latent variable structured prediction models with gaussian perturbations, *in* 'Advances in Neural Information Processing Systems', pp. 3149–3159.

Ben-David, S., Cesa-Bianchi, N., Haussler, D. and Long, P. M. (1995), 'Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions', *Journal of Computer and System Sciences* **50**(1), 74–86.

Collins, M. (2004), Parameter estimation for statistical parsing models: Theory and practice of distribution-free methods, *in* 'New developments in parsing technology', Springer, pp. 19–55.

Cortes, C., Kuznetsov, V. and Mohri, M. (2014), Ensemble methods for structured prediction, *in* 'International Conference on Machine Learning', pp. 1134–1142.

Cortes, C., Kuznetsov, V., Mohri, M. and Yang, S. (2016), Structured prediction theory based on factor graph complexity, *in* 'Advances in Neural Information Processing Systems', pp. 2514–2522.

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

Daniely, A., Sabato, S., Ben-David, S. and Shalev-Shwartz, S. (2015), 'Multiclass learnability and the erm principle', *The Journal of Machine Learning Research* **16**(1), 2377–2404.

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

Ghoshal, A. and Honorio, J. (2018), Learning maximum-a-posteriori perturbation models for structured prediction in polynomial time, *in* 'International Conference on Machine Learning'.

Globerson, A., Roughgarden, T., Sontag, D. and Yildirim, C. (2015), How hard is inference for structured prediction?, *in* 'International Conference on Machine Learning', pp. 2181–2190.

Honorio, J. and Jaakkola, T. (2016), Structured prediction: from gaussian perturbations to linear-time principled algorithms, *in* 'Uncertainty in Artificial Intelligence'.

Jurafsky, D. and Martin, J. H. (2014), *Speech and language processing*, Vol. 3, Pearson London.

Lafferty, J., McCallum, A. and Pereira, F. C. (2001), 'Conditional random fields: Probabilistic models for segmenting and labeling sequence data'.

Massart, P., Nédélec, É. et al. (2006), 'Risk bounds for statistical learning', *The Annals of Statistics* **34**(5), 2326–2366.

McAllester, D. (2007), Generalization bounds and consistency, *in* 'Predicting Structured Data', MIT Press, pp. 247–261.

Neylon, T. (2006), Sparse Solutions for Linear Prediction Problems, PhD thesis, New York University.

Santhanam, N. P. and Wainwright, M. J. (2012), 'Information-theoretic limits of selecting binary graphical models in high dimensions', *IEEE Transactions on Information Theory* **58**(7), 4117–4134.

Shalev-Shwartz, S. and Ben-David, S. (2014), *Understanding machine learning: From theory to algorithms*, Cambridge university press.

Tandon, R., Shanmugam, K., Ravikumar, P. K. and Dimakis, A. G. (2014), On the information theoretic limits of learning ising models, *in* 'Advances in Neural Information Processing Systems', pp. 2303–2311.

Taskar, B., Abbeel, P. and Koller, D. (2002), Discriminative probabilistic models for relational data, *in* 'Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence', Morgan Kaufmann Publishers Inc., pp. 485–492.

Taskar, B., Guestrin, C. and Koller, D. (2003), 'Max-margin Markov networks', *Neural Information Processing Systems* **16**, 25–32.

Taskar, B., Guestrin, C. and Koller, D. (2004), Max-margin markov networks, *in* 'Advances in neural information processing systems', pp. 25–32.

Tsochantaridis, I., Joachims, T., Hofmann, T. and Altun, Y. (2005), 'Large margin methods for structured and interdependent output variables', *Journal of machine learning research* **6**(Sep), 1453–1484.

Vapnik, V. (2013), *The nature of statistical learning theory*, Springer science & business media.

Vinyals, O., Kaiser, Ł., Koo, T., Petrov, S., Sutskever, I. and Hinton, G. (2015), 'Grammar as a foreign language', *Advances in neural information processing systems* pp. 2773–2781.

Vinyals, O., Toshev, A., Bengio, S. and Erhan, D. (2015), 'Show and tell: A neural image caption generator', *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 3156–3164.

Wainwright, M. J. (2019), *High-dimensional statistics: A non-asymptotic viewpoint*, Cambridge University Press.

Wald, A. (1939), 'Contributions to the theory of statistical estimation and testing hypotheses', *The Annals of Mathematical Statistics* **10**(4), 299–326.

Wang, W., Wainwright, M. J. and Ramchandran, K. (2010), Information-theoretic bounds on model selection for gaussian markov random fields, *in* 'IEEE International Symposium on Information Theory', pp. 1373–1377.

Wasserman, L. (2006), *All of nonparametric statistics*, Springer Science & Business Media.

Zhang, D., Sun, L. and Li, W. (2008), A structured prediction approach for statistical machine translation, *in* 'Proceedings of the Third International Joint Conference on Natural Language Processing'.