Measuring Learnability in Structured Prediction using Factor Graph Complexity*

Corinna Cortes  
Google Research  
New York, NY 10011  
corinna@google.com

Vitaly Kuznetsov  
Google Research  
New York, NY 10011  
vitaly@cims.nyu.edu

Mehryar Mohri  
Courant Institute and Google  
New York, NY 10011  
mohri@cims.nyu.edu

Scott Yang  
Courant Institute  
New York, NY 10011  
yangs@cims.nyu.edu

Abstract

We present a general theoretical analysis of structured prediction. By introducing a new complexity measure that explicitly factors in the structure of the output space and the loss function, we are able to derive new learning guarantees for hypothesis sets with an arbitrary factor graph decomposition. To the best of our knowledge, these are both the most favorable and the most general guarantees for structured prediction (and multiclass classification) currently known. They are data-dependent and applicable to a broad family of losses.

1 Introduction

Structured prediction encompasses a broad family of important learning problems. These include key tasks in natural language processing, computer vision, and speech recognition. In these problems, the output space admits some structure, and the natural loss function admits a decomposition along the output substructures. The output structure and the corresponding loss function make these problems significantly different from the (unstructured) binary classification problems studied in learning theory. Moreover, it is important to properly leverage this structure as the output spaces are typically very high-dimensional. There have been relatively few comprehensive studies devoted to structured prediction theory [Bakir et al., 2007]. Existing guarantees hold primarily for decomposable losses [Taskar et al., 2003, Cortes et al., 2014, Collins, 2001] or require specific algorithmic or structural conditions [McAllester, 2007, London et al., 2016].

This paper presents a general theoretical analysis of structured prediction. We give new data-dependent learning guarantees for a broad family of loss functions and hypothesis sets based on a novel complexity measure. This measure allows us to discriminate the learnability of problems that have the same (usually very large) output dimension but different output structure. Our bounds apply more generally than the existing work above and either match or improve upon the best-known bounds in several special cases.

2 Structured prediction theory

Let \( \mathcal{X} \) denote the input space and \( \mathcal{Y} \) the output space. We assume that the output space may be decomposed into up to \( l \) structures, as for a set of sequences, images, or graphs. We denote by

---

*This paper is an extended abstract of [Cortes et al., 2016].

L: Y × Y → R_+ a loss function measuring the dissimilarity of two elements of the output space Y. A key aspect of structured prediction is that the loss function can be decomposed along the substructures Y_k. As an example, L may be the Hamming loss defined by L(y, y') = \frac{1}{m} \sum_{i=1}^{m} 1_{y_i \neq y'_i} for all y = (y_1, \ldots, y_m) and y' = (y'_1, \ldots, y'_m), with y_k, y'_k \in Y_k. We will adopt the common approach where predictions are based on a scoring function mapping X × Y to \mathbb{R}. Let \mathcal{H} be a family of scoring functions. For any h \in \mathcal{H}, we denote by h the predictor defined by h: for any x \in X, h(x) = \arg\max_{y \in Y} h(x, y).

**Factor graphs.** We will follow the standard assumption that each scoring function h \in \mathcal{H} can be decomposed as a sum. In the most general case, these sums can be written using factor graphs. A factor graph G is a tuple G = (V, F, E), where V are variable nodes, F 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\}. For any factor node f, denote by N(f) \subseteq V the set of variable nodes connected to f via an edge and define Y_f as the substructure set cross-product Y_f = \prod_{k \in N(f)} Y_k. Then, h admits the following decomposition as a sum of functions h_f: h(x, y) = \sum_{f \in F} h_f(x, y_f).

### 2.1 Complexity measure

A key ingredient of our analysis is a new data-dependent notion of complexity that extends the classical Rademacher complexity. We define the empirical factor graph Rademacher complexity \hat{\mathcal{R}}^F_m(\mathcal{H}) of a hypothesis set \mathcal{H} for a sample S = (x_1, \ldots, x_m) and factor graph G as follows:

\hat{\mathcal{R}}^F_m(\mathcal{H}) = \frac{1}{m} \mathbb{E}_\epsilon \left[ \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sum_{f \in F} \sum_{y_f \in Y_f} \sqrt{|F_f|} \epsilon_{i,f,y} h_f(x_i, y_f) \right],

where \epsilon = (\epsilon_{i,f,y}) \in [m], f \in F, y \in Y_f, with \epsilon_{i,f,y} s independent Rademacher random variables uniformly distributed over \{\pm 1\}. The factor graph Rademacher complexity of \mathcal{H} for a factor graph G is defined as the expectation: \hat{\mathcal{R}}^F_m(\mathcal{H}) = \mathbb{E}_{S \sim D^m} [\hat{\mathcal{R}}^F_m(\mathcal{H})]. The factor graph Rademacher complexity is a natural extension of the standard Rademacher complexity to vector-valued hypothesis sets (with one coordinate per factor). For binary classification, the two notions of complexity coincide. Otherwise, the factor graph complexity is bounded above by the standard one.

### 2.2 Generalization bounds

We present new margin bounds for structured prediction in terms of the following additive empirical margin loss:

\hat{\mathcal{R}}_{S,\rho}^{\text{add}}(h) = \mathbb{E}_{(x,y) \sim S} \left[ \Phi^*(\max_{y' \neq y} L(y', y) - \frac{1}{\rho} [h(x, y) - h(x, y')]) \right].

(1)

Here, \Phi^*(r) = \min(M, \max(0, r)) for all r, with M = \max_{y,y'} L(y, y'). The following is our general data-dependent margin bound for structured prediction.

**Theorem 1.** Fix \rho > 0. For any \delta > 0, with probability at least 1 - \delta over the draw of a sample S of size m, the following holds for all h \in \mathcal{H},

\[ R(h) \leq \hat{\mathcal{R}}_{\rho}^{\text{add}}(h) \leq \hat{\mathcal{R}}_{S,\rho}^{\text{add}}(h) + \frac{4\sqrt{2}}{\rho} \hat{\mathcal{R}}^F_m(\mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}. \]

Theorem 1 is the first data-dependent generalization guarantee for structured prediction with general loss functions, general hypothesis sets and arbitrary factor graphs for both multiplicative and additive margins. In the full version of this paper, [Cortes et al., 2016], we show how the bound matches or improves upon the best-known results in many of the references provided above. We also show how our theory motivates the design of novel algorithms for structure prediction.

\footnote{Factor graphs are typically used to indicate the factorization of a probabilistic model. We are not assuming these models, but they would be also captured by our general framework: h would then be \log of a probability.}
Acknowledgments

The work of M. M., S.Y. and V.K. was partly funded by NSF IIS-1117591 and CCF-1535987. S.Y. was also partially supported by the NSF GRFP under Grant No. DGE-1342536.

References


