An Overview of Structured Prediction Theory

Colin Graber

1 Introduction

Much of machine learning theory is devoted to the study of learning problems where the value to be predicted consists of a single number – in $\{-1, 1\}$ for the case of binary classification, and in $\mathbb{R}$ in the case of regression. However, many machine learning problems exist for which the output is more complicated. Consider, for example, the natural language processing task of part-of-speech tagging, wherein each of the words in a sentence is assigned a label from a vocabulary of part-of-speech tags:

I	proved	an	interesting	theorem	today	.
PRP	VBD	DT	JJ	NN	NN	.

There are two things to note about this task: the output space is vector-valued, and the outputs at different indices of this vector are not independent (for example, in English, you are not likely to see an adjective followed by an adverb).

Tasks with these two properties are often referred to as *structured prediction*. These sorts of learning problems are frequently found in the subfields of natural language processing (in tasks such as part-of-speech tagging, named entity recognition, and machine translation) and computer vision (in tasks such as image segmentation and object recognition).

The nature of the label space for these learning tasks prohibit us from applying approaches for simpler tasks such as binary classification without some modifications; additionally, some additional work is required to develop generalization theory for this task. This report will cover a general modeling framework for structured prediction, including a few examples of models developed for the task, before introducing recent results that derive data-dependent generalization bounds for structured prediction.

2 Modeling structured prediction

We will now provide a formal specification of this task. Inputs are provided from some input set $\mathcal{X}$, and outputs fall within an output set $\mathcal{Y} = \cup_{k=1}^{l} \prod_{i=1}^{k} \mathcal{Y}'$, where $\mathcal{Y}'$ is some set describing the labels that can be assigned to each part of the output and l is some maximum output size. Note that, because inputs may have varying sizes (e.g. English sentences tend to have different lengths), the output space consists of structures of varying size (hence the union in the definition). It is common for $\mathcal{Y}'$ to be discrete and finite, and the analysis presented later makes this assumption. In our earlier example, $\mathcal{Y}'$ consisted of the vocabulary of part-of-speech tags, while $\mathcal{Y}$ consisted of all part-of-speech tag sequences. Hypotheses take the form of scoring functions $h : \mathcal{X} \times \mathcal{Y} \to \mathcal{R}$ which take an element of the input space and an element of the output space and assign them a real-valued score. Every scoring function has a corresponding predictor function $\mathsf{h} : \mathcal{X} \to \mathcal{Y}$, where, for every $x \in \mathcal{X}$, $\mathsf{h}(x) = \text{argmax}_{y \in \mathcal{Y}} h(x, y)$.

As already mentioned, we expect the values at given indices to be correlated in some way, with the exact structure of this depending on the individual task being studied. Hence, it is important that the scoring functions we use allow us to include this knowledge so that we can better exploit it. To that end, we will represent the knowledge we have about the output structures using the language of probabilistic graphical models. Specifically, we assume that the scoring functions decompose

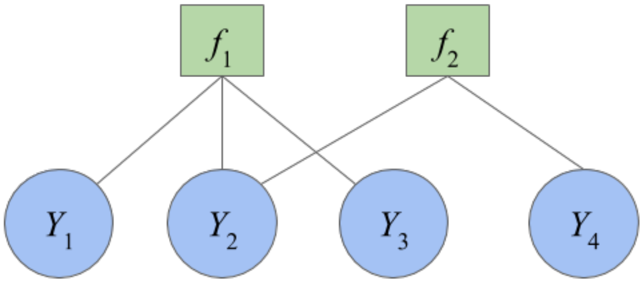

Figure 1: A sample factor graph. The scoring function encoded by this graph is $h(x, y) = h_1(x, y_1, y_2, y_3) + h_2(y_2, y_4)$.

into a sum of various components as defined by a factor graph $G = (V, F, E)$ where V is a set of k variable nodes (each representing a different component of the output), F is a set of factor nodes, and E is a set of undirected edges connecting variable and factor nodes. For each factor $f \in F$ there is a corresponding piece of the scoring function h_f which takes as arguments the input x as well as labels y_f for some subset of the output variables $\mathcal{Y}_f = \prod_{i=1}^{|\mathcal{N}(f)|} \mathcal{Y}'$, where $\mathcal{N}(f)$ represents the set of variable nodes connected to factor f. Hence, we can represent the scoring functions as follows:

$$h(x, y) = \sum_{f \in F} h_f(x, y_f)$$

It is common for the inputs to have varying size (for example, sentences contain different numbers of words); we represent this by specifying a function $G(x_i, y_i)$ that returns the appropriate factor graph for a given input. Note that, though we are borrowing elements of probabilistic graphical models to define this model, we do not require the scoring functions to be probabilistic - that is, there need not be a probabilistic interpretation of the scores output by the scoring functions, and we don't require them to enforce any independence assumptions present within the graphs.

The learning scenario is the usual statistical learning setup: we are provided with a sample of points $S = ((x_1, y_1), \ldots, (x_m, y_m))$ drawn I.I.D. from some unknown distribution $\mathcal{D}$. The goal of learning is to minimize the generalization error $R(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathsf{L}(\mathsf{h}(x), y)\right]$ where $\mathsf{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a loss function that measures how different two outputs are from each other. We require the loss to be definite; that is, $\mathsf{L}(y, y') = 0$ if and only if $y = y'$. One example of a commonly used loss in structured prediction is Hamming loss, defined as $\mathsf{L}(y, y') = \frac{1}{k}\sum_{i=1}^{k} \mathbf{1}_{\{y_i \neq y'_i\}}$.

A variety of models for structured prediction have been developed over the years. Not all of them fit neatly into this framework, but there are many that do. Here are a few examples of structured prediction models which can be seen as examples of this general framework:

- Conditional Random Fields [Lafferty et al.] were developed to be a discriminative version of earlier generative models (such as hidden Markov models) used for some structured prediction tasks. CRFs model the conditional probability of an output given an input via the following formulation:

  $$p(y_i|x_i) \propto \prod_{f \in F_i} \exp\left(w_f^T \Psi_f(x, y_f)\right)$$

  where w_f are parameter vectors and Ψ_f are "feature functions" which capture various properties of the input and provided label (e.g., for the earlier example, there could be binary indicator features for every word-label pair). Fitting this into our framework, scoring functions have a probabilistic interpretation - specifically, they represent the logarithm of a conditional distribution over labels given inputs, i.e. $\log p(y|x) \propto h(x, y)$. The individual components of the scoring functions are linear, meaning they have the form

  $$h_f(x, y_f) = w_f^T \Psi_f(x, y_f)$$

The model parameters are learned by maximizing the log-likelihood of the data:

$$\ell(w) = \sum_{i=1}^{m} \sum_{f \in F_i} w_f^T h_f(x, y_f) - \log(Z)$$

where Z is the partition function that ensures the resulting scoring functions are distributions.

- Maximum Margin Markov Networks (M3N) [Taskar et al., 2003] have the same model formulation as conditional random fields - they also model the conditional probability of an output given an input using log-linear factors, and hence the scoring functions represent the log-probabilities of the output given the input. However, the training objective for M3Ns is different - rather than maximizing the likelihood of the data, the parameters are learned using the following objective:

  $$\begin{aligned}&\text{minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i \\ &\text{s.t. } \forall i, \forall y \in \mathcal{Y} \backslash y_i : w^T[\Psi(x_i, y_i) - \Psi(x_i, y)] \geq \mathsf{L}(y_i, y) - \xi_i\end{aligned}$$

  This objective can be viewed as maximizing a loss-dependent margin between the score assigned to the correct label and the scores assigned to all other output labels.

- Structured Support Vector Machines (SSVM) [Tsochantaridis et al., 2005] are similar to the previous two models in that the scoring functions are also linear. However, SSVMs completely drop the probabilistic interpretation of the scores - this model is only concerned with finding the highest scoring output, rather than trying to model the conditional distribution of the outputs. Training is also achieved by attempting to maximize some notion of a margin. Two optimization objectives are presented in the paper as options for training: the first has the same form as that for M3Ns, and the second is the following:

  $$\begin{aligned}&\text{minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i \\ &\text{s.t. } \forall i, \forall y \in \mathcal{Y} \backslash y_i : w^T[\Psi(x_i, y_i) - \Psi(x_i, y)] \geq 1 - \frac{\xi_i}{\mathsf{L}(y_i, y)}\end{aligned}$$

  The first formulation can be described as scaling the margin by the loss, and the second can be described as scaling the slack by the loss.

Though many approaches to structured prediction have been developed over the years, until recently there has been very little analysis of generalization performance of these approaches. In the few cases where it did exist ([Taskar et al., 2003, McAllester., 2007]), specific losses or hypothesis sets were assumed. Late last year, however, [Cortes et al., 2016] developed general bounds for structured prediction using the general framework we have described here; the next section will cover some of the analysis presented in this paper.

3 Factor graph Rademacher complexity: definition and risk bound[1]

The generalization bound to be presented is defined in terms of the following quantity, which is called the empirical factor graph Rademacher complexity:

$$\widehat{\mathfrak{R}}_S^G \triangleq \frac{1}{m}\mathbb{E}_\epsilon \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{m} \sum_{f \in F_i} \max_{y \in \mathcal{Y}_f} \sqrt{|F_i|}\epsilon_{i,f} h_f(x_i, y) \right]$$

where $S = ((x_1, y_1), \ldots, (x_m, y_m))$ is a sample of points, G is the function that returns a factor graph for a given instance, and $\epsilon_{i,f}$ are Rademacher random variables. This complexity term allows

[1] The proofs presented in this section are based heavily on those in [Cortes et al., 2016]. Specifically, the generalization bound is based on the proof of Theorem 1 where a (slightly modified) version of Lemma 6 was used in place of Lemma 5.

us to base the generalization bounds for a given problem on the assumed structure that we have encoded into the form of the graph. We will now go through the proof of a generalization bound for structured prediction using this new complexity term. The bound is in terms of the following margin-based losses:

$$\widehat{R}^{add}_{S,\rho}(h) \triangleq \mathop{\mathbb{E}}_{(x,y)\sim S}\left[\Phi^*\left(\max_{y'\neq y}\mathsf{L}(y',y) - \frac{1}{\rho}[h(x,y)-h(x,y')]\right)\right]$$
$$R^{add}_{\rho}(h) \triangleq \mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}\left[\Phi^*\left(\max_{y'\neq y}\mathsf{L}(y',y) - \frac{1}{\rho}[h(x,y)-h(x,y')]\right)\right]$$

where $\Phi^*(r) = \min(M, \max(0,r))$ and $M = \max_{y,y'}\mathsf{L}(y,y')$. Upper bounds to these losses can be found that correspond to the example models described earlier. To be able to complete the proof of the bound, we will need the following two lemmas. The first provides a simple way for us to upper bound the risk using surrogate losses.

Lemma 3.1. *For any $u \in \mathbb{R}_+$, let $\Phi_u : \mathbb{R} \to \mathbb{R}$ be an upper bound on $v \to u\mathbf{1}_{v\leq 0}$. Then, the following upper bound holds for any $h \in \mathcal{H}$ and $(x,y) \in \mathcal{X}\times\mathcal{Y}$,*

$$\mathsf{L}(\mathsf{h}(x),y) \leq \max_{y'\neq y}\Phi_{\mathsf{L}(y',y)}(h(x,y)-h(x,y'))$$

Proof. See [Cortes et al., 2016] for details. □

The second lemma can be thought of as a new contraction lemma appropriate for our formulation of structured prediction. This will allow us to deal with the loss term within our generalization bound:

Lemma 3.2. *Let $\mathcal{H}$ be a hypothesis set of functions mapping $\mathcal{X}\times\mathcal{Y}$ to $\mathbb{R}$ as defined previously. Let Ψ_i, $i = 1,\ldots,m$ be functions mapping $\mathbb{R}\times\mathcal{Y}$ to $\mathcal{R}$. Assume that for all $i = 1,\ldots,m$ there exists a constant μ_i such that the following is true for any $h, h' \in \mathcal{H}$:*

$$\left|\max_{y\in\mathcal{Y}}\Psi_i(h(x_i,y),y) - \max_{y\in\mathcal{Y}}\Psi_i(h'(x_i,y),y)\right| \leq \mu_i\sqrt{\sum_{f\in F_i}\left[\max_{y\in\mathcal{Y}_f}|h_f(x_i,y)-h'_f(x_i,y)|\right]^2}$$

Then, for any sample S of m points $x_1,\ldots,x_m \in \mathcal{X}$, the following inequality holds:

$$\frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i\max_{y\in\mathcal{Y}}\Psi_i(h(x_i,y),y)\right] \leq \frac{\sqrt{2}}{m}\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sum_{f\in F_i}\max_{y\in\mathcal{Y}_f}\epsilon_{if}\mu_i h_f(x_i,y)\right]$$

where $\epsilon = (\epsilon_{if})_{i,f}$ and ϵ_{if}s are independent Rademacher random variables.

Proof. Fix a sample $S = (x_1,\ldots,x_m)$. The proof proceeds by processing one σ_i at a time and then recursing. First, we rewrite the left-hand side of equation 3.2 as follows:

$$\frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i\max_{y\in\mathcal{Y}}\Psi_i(h(x_i,y),y)\right] = \frac{1}{m}\mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_{m-1}}\left[\mathop{\mathbb{E}}_{\sigma_m}\left[\sup_{h\in\mathcal{H}}U_{m-1}(h) + \sigma_m\max_{y\in\mathcal{Y}}\Psi_m(h(x_m,y),y)\right]\right],$$

where $U_{m-1}(h) = \sum_{i=1}^{m-1}\sigma_i\max_{y\in\mathcal{Y}}\Psi_i(h(x_i,y),y)$. For the purposes of this proof, we assume that the supremum of this expression is attained for both values of σ_m; in the case that this is not true, this proof holds by instead considering ϵ-close hypotheses to these suprema. Letting h_1 and h_2 be the maximizing hypotheses for $\sigma_m = 1$ and $\sigma_m = -1$, respectively, the inner expectation can be expanded as

$$\mathop{\mathbb{E}}_{\sigma_m}\left[\sup_{h\in\mathcal{H}}U_{m-1}(h) + \sigma_m\max_{y\in\mathcal{Y}}\Psi_m(h(x_m,y),y)\right]$$

$$= \frac{1}{2}\left[U_{m-1}(h_1) + \max_{y\in\mathcal{Y}} \Psi_m(h_1(x_m,y),y)\right] + \frac{1}{2}\left[U_{m-1}(h_2) - \max_{y\in\mathcal{Y}}\Psi_m(h_2(x_m,y),y)\right]$$

Next, we apply the precondition for Ψ_m and the Khintchine-Kahane inequality to get

$$\begin{aligned}
&\frac{1}{2}\left[U_{m-1}(h_1) + \max_{y\in\mathcal{Y}} \Psi_m(h_1(x_m,y),y)\right] + \frac{1}{2}\left[U_{m-1}(h_2) - \max_{y\in\mathcal{Y}_f}\Psi_m(h_2(x_m,y),y)\right] \\
&\leq \frac{1}{2}\left[U_{m-1}(h_1) + U_{m-1}(h_2) + \mu_m\sqrt{\sum_{f\in F_m}\left[\max_{y\in\mathcal{Y}_f}|h_{1f}(x_m,y)-h_{2f}(x_m,y)|\right]^2}\right] \\
&\leq \frac{1}{2}\left[U_{m-1}(h_1) + U_{m-1}(h_2) + \mu_m\sqrt{2}\,\mathop{\mathbb{E}}_{\epsilon_m}\left[\left|\sum_{f\in F_m}\epsilon_{mf}\max_{y\in\mathcal{Y}_f}|h_{1f}(x_m,y)-h_{2f}(x_m,y)|\right|\right]\right]
\end{aligned}$$

where $\epsilon_m = (\epsilon_{mf})_{f\in F_m}$. Let $s(\epsilon_m) \in \{\pm 1\}$ denote the sign of $\sum_{f\in F_m}\epsilon_{mf}\max_{y\in\mathcal{Y}_f}|h_{1f}(x_m,y)-h_{2f}(x_m,y)|$. Then, we get the following:

$$\begin{aligned}
&\frac{1}{2}\left[U_{m-1}(h_1) + U_{m-1}(h_2) + \mu_m\sqrt{2}\,\mathop{\mathbb{E}}_{\epsilon_m}\left[\left|\sum_{f\in F_m}\epsilon_{mf}\max_{y\in\mathcal{Y}_f}|h_{1f}(x_m,y)-h_{2f}(x_m,y)|\right|\right]\right] \\
&\leq \frac{1}{2}\mathop{\mathbb{E}}_{\epsilon_m}\left[U_{m-1}(h_1) + U_{m-1}(h_2) + \mu_m\sqrt{2}s(\epsilon_m)\sum_{f\in F_m}\max_{y\in\mathcal{Y}_f}\epsilon_{mf}|h_{1f}(x_m,y)-h_{2f}(x_m,y)|\right] \\
&= \frac{1}{2}\mathop{\mathbb{E}}_{\epsilon_m}\left[U_{m-1}(h_1) + U_{m-1}(h_2) + \mu_m\sqrt{2}s(\epsilon_m)\sum_{f\in F_m}\max_{y\in\mathcal{Y}_f}\epsilon_{mf}(h_{1f}(x_m,y)-h_{2f}(x_m,y))\right] \\
&\leq \frac{1}{2}\mathop{\mathbb{E}}_{\epsilon_m}\Bigg[U_{m-1}(h_1) + \sum_{f\in F_m}\max_{y\in\mathcal{Y}_f}\epsilon_{mf}h_{1f}(x_m,y) \\
&\qquad\qquad + U_{m-1}(h_2) + \mu_m\sqrt{2}s(\epsilon_m)\sum_{f\in F_m}\max_{y\in\mathcal{Y}_f}-\epsilon_{mf}h_{2f}(x_m,y)\Bigg] \\
&\leq \frac{1}{2}\mathop{\mathbb{E}}_{\epsilon_m}\Bigg[\sup_{h\in\mathcal{H}}\left(U_{m-1}(h) + \mu_m\sqrt{2}s(\epsilon_m)\sum_{f\in F_m}\max_{y\in\mathcal{Y}_f}\epsilon_{mf}h_f(x_m,y)\right) \\
&\qquad\qquad + \sup_{h\in\mathcal{H}}\left(U_{m-1}(h) + \mu_m\sqrt{2}s(\epsilon_m)\sum_{f\in F_m}\max_{y\in\mathcal{Y}_f}-\epsilon_{mf}h_f(x_m,y)\right)\Bigg] \\
&\leq \frac{1}{2}\mathop{\mathbb{E}}_{\epsilon_m}\left[\mathop{\mathbb{E}}_{\sigma_m}\left[\sup_{h\in\mathcal{H}}U_{m-1}(h) + \mu_m\sqrt{2}\sum_{f\in F_m}\max_{y\in\mathcal{Y}_f}s(\epsilon_m)\sigma_m\epsilon_{mf}h_f(x_m,y)\right]\right] \\
&= \frac{1}{2}\mathop{\mathbb{E}}_{\epsilon_m}\left[\sup_{h\in\mathcal{H}}U_{m-1}(h) + \mu_m\sqrt{2}\sum_{f\in F_m}^{c}\max_{y\in\mathcal{Y}_f}\epsilon_{mf}h_f(x_m,y)\right]
\end{aligned}$$

Continuing the same way for every other σ_i, $i < m$, completes the proof.

□

We now have all of the tools we need to prove a generalization bound:

Theorem 3.3. *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 $h \in \mathcal{H}$:*

$$R(h) \leq R_\rho^{add}(h) \leq \widehat{R}_{S,\rho}^{add}(h) + \frac{4\sqrt{2}}{\rho}\widehat{\mathfrak{R}}_S^G(\mathcal{H}) + 3M\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

Proof. The first step is to upper bound the risk by the margin risk. Using Lemma 3.1 with $\Phi_u(v) = \Phi^*(u - \frac{v}{\rho})$ along with the monotonicity of Φ^*, we get the following:

$$\begin{aligned} R(h) &\leq \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\max_{y'\neq y}\Phi_{L(y',y)}(h(x,y) - h(x,y'))\right] \\ &\leq \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\Phi^*\left(\max_{y'\neq y}\left(\mathsf{L}(y',y) - \tfrac{1}{\rho}[h(x,y) - h(x,y')]\right)\right)\right] \\ &= R_\rho^{add}(h) \end{aligned}$$

For notational convenience, define:

$$\begin{aligned} \mathcal{H}_0 &= \left\{(x,y) \mapsto \Phi^*\left(\max_{y'\neq y}\left(\mathsf{L}(y',y) - \tfrac{1}{\rho}[h(x,y) - h(x,y')]\right)\right) : h \in \mathcal{H}\right\} \\ \mathcal{H}_1 &= \left\{(x,y) \mapsto \max_{y'\neq y}\left(\mathsf{L}(y',y) - \tfrac{1}{\rho}[h(x,y) - h(x,y')]\right) : h \in \mathcal{H}\right\} \end{aligned}$$

At this point, we apply a standard Rademacher complexity bound [Koltchinskii and Panchenko, 2002] to tell us that, for any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $h \in \mathcal{H}$:

$$R_\rho^{add}(h) \leq \widehat{R}_{S,\rho}^{add}(h) + 2\widehat{\mathfrak{R}}_S(\mathcal{H}_0) + 3M\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

where $\widehat{\mathfrak{R}}_S(\mathcal{H}_0)$ is the standard empirical Rademacher complexity of the set of functions $\mathcal{H}_0$, i.e.

$$\widehat{\mathfrak{R}}_S(\mathcal{H}_0) = \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^m \sigma_i\Phi^*\left(\max_{y'\neq y_i}\left(\mathsf{L}(y',y_i) - \tfrac{1}{\rho}[h(x_i,y_i) - h(x_i,y')]\right)\right)\right]$$

with $\sigma = (\sigma_1, \ldots, \sigma_m)$ and σ_is are Rademacher random variables. The fact that Φ^* is 1-Lipschitz allows us to apply Talagrand's contraction lemma [Ledoux and Talagrand, 1991, Mohri et al., 2012] to achieve the result $\widehat{\mathfrak{R}}_S(\mathcal{H}_0) \leq \widehat{\mathfrak{R}}_S(\mathcal{H}_1)$. The sub-additivity of the supremum along with the fact that σ_i and $-\sigma_i$ are distributed identically allows us to split the Rademacher complexity into two terms:

$$\begin{aligned} \widehat{\mathfrak{R}}_S(\mathcal{H}_1) &\leq \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^m \sigma_i \max_{y'\neq y_i}\left(\mathsf{L}(y',y_i) + \tfrac{1}{\rho}h(x_i,y')\right)\right] + \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^m \sigma_i\tfrac{1}{\rho}h(x_i,y_i)\right] \\ &\leq \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^m \sigma_i \max_{y'\in\mathcal{Y}}\left(\mathsf{L}(y',y_i) + \tfrac{1}{\rho}h(x_i,y')\right)\right] + \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^m \sigma_i \max_{y\in\mathcal{Y}}\tfrac{1}{\rho}h(x_i,y)\right] \end{aligned}$$

We bound each of these separately using Lemma 3.2. To do so, we need to prove the precondition of this Lemma for each term; for the first term, let $\Psi_i(h(x_i, y), y) = \mathsf{L}(y, y_i) + \frac{1}{\rho}h(x_i, y)$. We then have the following for any $h, h' \in \mathcal{H}$:

$$\begin{aligned}
&\left|\max_{y\in\mathcal{Y}} \left(\mathsf{L}(y, y_i) + \tfrac{1}{\rho}h(x_i, y)\right) - \max_{y\in\mathcal{Y}} \left(\mathsf{L}(y, y_i) + \tfrac{1}{\rho}h'(x_i, y)\right)\right| \\
&\qquad \leq \tfrac{1}{\rho} \max_{y\in\mathcal{Y}} |h(x_i, y) - h'(x_i, y)| \\
&\qquad \leq \tfrac{1}{\rho} \sum_{f\in F_i} \max_{y\in\mathcal{Y}_f} |h_f(x_i, y) - h'(x_i, y)| \\
&\qquad \leq \frac{\sqrt{|F_i|}}{\rho} \sqrt{\sum_{f\in F_i} \left[\max_{y\in\mathcal{Y}_f} |h_f(x_i, y) - h'_f(x_i, y)\right]^2}
\end{aligned}$$

where the last line is a consequence of the following inequality, which holds for any vector $a \in \mathcal{R}^d$:

$$\frac{1}{n}\sum_{i=1}^{d} a_i \leq \left(\frac{1}{n}\sum_{i=1}^{d} a_i^2\right)^{1/2}$$

We can prove the precondition of the Lemma for the other term in a similar manner: for any $h, h' \in \mathcal{H}$:

$$\begin{aligned}
&\left|\max_{y\in\mathcal{Y}} \tfrac{1}{\rho}h(x_i, y) - \max_{y\in\mathcal{Y}} \tfrac{1}{\rho}h'(x_i, y)\right| \\
&\qquad \leq \tfrac{1}{\rho} \max_{y\in\mathcal{Y}} |h(x_i, y) - h'(x_i, y)| \\
&\qquad \leq \tfrac{1}{\rho} \sum_{f\in F_i} \max_{y\in\mathcal{Y}_f} |h_f(x_i, y) - h'(x_i, y)| \\
&\qquad \leq \frac{\sqrt{|F_i|}}{\rho} \sqrt{\sum_{f\in F_i} \left[\max_{y\in\mathcal{Y}_f} |h_f(x_i, y) - h'_f(x_i, y)\right]^2}
\end{aligned}$$

Hence, for both terms, we can apply Lemma 3.2 with $\mu_i = \frac{\sqrt{|F_i|}}{\rho}$, giving us

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{H}_1) &\leq \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i \max_{y'\in\mathcal{Y}} \left(\mathsf{L}(y', y_i) + \tfrac{1}{\rho}h(x_i, y')\right)\right] + \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i \max_{y\in\mathcal{Y}} \tfrac{1}{\rho}h(x_i, y)\right] \\
&\leq \frac{2\sqrt{2}}{m} \mathop{\mathbb{E}}_{\epsilon} \left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \sum_{f\in F_i} \max_{y\in\mathcal{Y}_f} \epsilon_{i,f} \frac{\sqrt{|F_i|}}{\rho} h_f(x_i, y)\right] \\
&= \frac{2\sqrt{2}}{\rho} \widehat{\mathfrak{R}}_S^G(\mathcal{H})
\end{aligned}$$

Substituting this into the bound derived earlier completes the proof. □

4 Bounding factor graph Rademacher complexity

To fully instantiate the bound provided in Theorem 3.3, we need to be able to estimate the empirical factor graph Rademacher complexity for a given hypothesis set. For example, consider the set of linear functions with bounded L2-norm:

$$\mathcal{H}_2 = \left\{x \mapsto w^T\Psi(x, y) : w \in \mathcal{R}^N, \|w\|2 \leq \Lambda_2\right\}$$

where $\Psi(x,y)$ is a feature function for inputs x and y that decompose according to the factor graph, i.e. $\Psi(x,y) = \sum_{f\in F}\Psi_f(x,y_f)$. This is the hypothesis set underlying the structured prediction models mentioned earlier if we ensure that the parameters for each factor are "tied". We can bound the complexity as follows:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S^G(\mathcal{H}_2) &= \frac{1}{m}\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{\|w\|_2\le\Lambda_2} w^T\left(\sum_{i=1}^m\sum_{f\in F_i}\max_{y\in\mathcal{Y}_f}\epsilon_{i,f}\sqrt{|F_i|}\Psi_f(x_i,y)\right)\right]\\
&\le \frac{\Lambda_2}{m}\left(\mathop{\mathbb{E}}_{\epsilon}\left[\left\|\sum_{i=1}^m\sum_{f\in F_i}\max_{y\in\mathcal{Y}_f}\epsilon_{i,f}\sqrt{|F_i|}\Psi_f(x_i,y)\right\|_2\right]\right)\\
&\le \frac{\Lambda_2}{m}\left(\mathop{\mathbb{E}}_{\epsilon}\left[\left\|\sum_{i=1}^m\sum_{f\in F_i}\max_{y\in\mathcal{Y}_f}\epsilon_{i,f}\sqrt{|F_i|}\Psi_f(x_i,y)\right\|_2^2\right]\right)^{\frac{1}{2}}\\
&= \frac{\Lambda_2}{m}\left(\sum_{i=1}^m\sum_{f\in F_i}\max_{y\in\mathcal{Y}_f}|F_i|\|\Psi_f(x_i,y)\|_2^2\right)^{\frac{1}{2}}\\
&\le \frac{\Lambda_2 r_2}{m}\sqrt{\sum_{i=1}^m|F_i|^2}
\end{aligned}$$

where $r_2 = \max_{i,f,y}\|\Psi_f(x_i,y)\|_2$. To refine this bound further, we have to decide on a graph structure. For example, consider a sequence labeling problem where our graph consists of a linear chain that is the same size as the input sequence. Assuming a max sequence length of l, this bound becomes

$$\frac{\Lambda_2 r_2}{m}\sqrt{\sum_{i=1}^m|F_i|^2} \le \frac{l\Lambda_2 r_2}{\sqrt{m}}$$

As another example, consider a problem where our graph is fully pairwise, with maximum graph size of l nodes. In this case, the bound is

$$\frac{\Lambda_2 r_2}{m}\sqrt{\sum_{i=1}^m|F_i|^2} \le \frac{l^2\Lambda_2 r_2}{\sqrt{m}}$$

We can also refine r_2 based on the form of Φ we choose for the problem - for example, if we use boolean features where there is some known sparsity s where s represents the maximum number of features that are simultaneously active, then r_2 is bounded by $\sqrt{s}$.

5 Conclusion

This report presented a broad overview of structured prediction theory, including a general formulation, example models, and some generalization theory. The bounds proved are general enough to apply to a wide variety of different models. The issues presented here represent a very small subset of the problems faced when studying structured prediction; for example, for general factor graphs, finding $\text{argmax}_{y\in\mathcal{Y}}h(x,y)$ is an NP-hard problem, and therefore approximate inference is required for learning to be tractable.

References

Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, and Scott Yang. Structured prediction theory based on factor graph complexity. In *Advances in Neural Information Processing Systems*, pages 2514–2522, 2016.

Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, pages 1–50, 2002.

John Lafferty, Andrew McCallum, Fernando Pereira, et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pages 282–289.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23. Springer Science & Business Media, 1991.

David McAllester. Generalization bounds and consistency for structured labeling. In *Predicting structured data*. MIT press, 2007.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

Ben Taskar, Carlos Guestrin, and Daphne Koller. Max-margin markov networks. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, pages 25–32. MIT Press, 2003.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *Journal of machine learning research*, 6(Sep):1453–1484, 2005.