HAL
open science

Multilabel Structured Output Learning with Random Spanning Trees of Max-Margin Markov Networks

Mario Marchand, Hongyu Su, Emilie Morvant, Johu Rousu, John Shawe-Taylor

To cite this version:

Mario Marchand, Hongyu Su, Emilie Morvant, Johu Rousu, John Shawe-Taylor. Multilabel Structured Output Learning with Random Spanning Trees of Max-Margin Markov Networks. Neural Information Processing Systems (NIPS), Dec 2014, Montréal, Canada. hal-01065586

HAL Id: hal-01065586

https://hal.science/hal-01065586v1

Submitted on 29 Oct 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Multilabel Structured Output Learning with Random Spanning Trees of Max-Margin Markov Networks

Mario Marchand
Département d'informatique et génie logiciel
Université Laval
Québec (QC), Canada
mario.marchand@ift.ulaval.ca

Hongyu Su
Helsinki Institute for Information Technology
Dept of Information and Computer Science
Aalto University, Finland
hongyu.su@aalto.fi

Emilie Morvant*
LaHC, UMR CNRS 5516
Univ. of St-Etienne, France
emilie.morvant@univ-st-etienne.fr

Juho Rousu
Helsinki Institute for Information Technology
Dept of Information and Computer Science
Aalto University, Finland
juho.rousu@aalto.fi

John Shawe-Taylor
Department of Computer Science
University College London
London, UK
j.shawe-taylor@ucl.ac.uk

Abstract

We show that the usual score function for conditional Markov networks can be written as the expectation over the scores of their spanning trees. We also show that a small random sample of these output trees can attain a significant fraction of the margin obtained by the complete graph and we provide conditions under which we can perform tractable inference. The experimental results confirm that practical learning is scalable to realistic datasets using this approach.

1 Introduction

Finding an hyperplane that minimizes the number of misclassifications is $\mathcal{NP}$-hard. But the support vector machine (SVM) substitutes the hinge for the discrete loss and, modulo a margin assumption, can nonetheless efficiently find a hyperplane with a guarantee of good generalization. This paper investigates whether the problem of inference over a complete graph in structured output prediction can be avoided in an analogous way based on a margin assumption.

We first show that the score function for the complete output graph can be expressed as the expectation over the scores of random spanning trees. A sampling result then shows that a small random sample of these output trees can attain a significant fraction of the margin obtained by the complete graph. Together with a generalization bound for the sample of trees, this shows that we can obtain good generalization using the average scores of a sample of trees in place of the complete graph. We have thus reduced the intractable inference problem to a convex optimization not dissimilar to a SVM. The key inference problem to enable learning with this ensemble now becomes finding the maximum violator for the (finite sample) average tree score. We then provide the conditions under which the inference problem is tractable. Experimental results confirm this prediction and show that

*Most of this work was carried out while E. Morvant was affiliated with IST Austria, Klosterneuburg.

practical learning is scalable to realistic datasets using this approach with the resulting classification accuracy enhanced over more naive ways of training the individual tree score functions.

The paper aims at exploring the potential ramifications of the random spanning tree observation both theoretically and practically. As such, we think that we have laid the foundations for a fruitful approach to tackle the intractability of inference in a number of scenarios. Other attractive features are that we do not require knowledge of the output graph's structure, that the optimization is convex, and that the accuracy of the optimization can be traded against computation. Our approach is firmly rooted in the maximum margin Markov network analysis [1]. Other ways to address the intractability of loopy graph inference have included using approximate MAP inference with tree-based and LP relaxations [2], semi-definite programming convex relaxations [3], special cases of graph classes for which inference is efficient [4], use of random tree score functions in heuristic combinations [5]. Our work is not based on any of these approaches, despite superficial resemblances to, *e.g.,* the trees in tree-based relaxations and the use of random trees in [5]. We believe it represents a distinct approach to a fundamental problem of learning and, as such, is worthy of further investigation.

2 Definitions and Assumptions

We consider supervised learning problems where the input space $\mathcal{X}$ is arbitrary and the output space $\mathcal{Y}$ consists of the set of all ℓ-dimensional multilabel vectors $(y_1, \ldots, y_\ell) \stackrel{\text{def}}{=} \mathbf{y}$ where each $y_i \in \{1, \ldots, r_i\}$ for some finite positive integer r_i. Each example $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ is mapped to a joint feature vector $\boldsymbol{\phi}(x, \mathbf{y})$. Given a weight vector $\mathbf{w}$ in the space of joint feature vectors, the predicted output $\mathbf{y}_{\mathbf{w}}(x)$ at input $x \in \mathcal{X}$, is given by the output $\mathbf{y}$ maximizing the *score* $F(\mathbf{w}, x, \mathbf{y})$, *i.e.*,

$$\mathbf{y}_{\mathbf{w}}(x) \stackrel{\text{def}}{=} \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}} F(\mathbf{w}, x, \mathbf{y}) \quad ; \quad \text{where} \quad F(\mathbf{w}, x, \mathbf{y}) \stackrel{\text{def}}{=} \langle \mathbf{w}, \boldsymbol{\phi}(x, \mathbf{y}) \rangle \,, \tag{1}$$

and where $\langle \cdot, \cdot \rangle$ denotes the inner product in the joint feature space. Hence, $\mathbf{y}_{\mathbf{w}}(x)$ is obtained by solving the so-called *inference* problem, which is known to be $\mathcal{NP}$-hard for many output feature maps [6, 7]. Consequently, we aim at using an output feature map for which the inference problem can be solved by a polynomial time algorithm such as dynamic programming. The *margin* $\Gamma(\mathbf{w}, x, \mathbf{y})$ achieved by predictor $\mathbf{w}$ at example $(x, \mathbf{y})$ is defined as,

$$\Gamma(\mathbf{w}, x, \mathbf{y}) \stackrel{\text{def}}{=} \min_{\mathbf{y}' \neq \mathbf{y}} \left[F(\mathbf{w}, x, \mathbf{y}) - F(\mathbf{w}, x, \mathbf{y}') \right] .$$

We consider the case where the feature map $\boldsymbol{\phi}$ is a potential function for a Markov network defined by a complete graph G with ℓ nodes and $\ell(\ell-1)/2$ undirected edges. Each node i of G represents an output variable y_i and there exists an edge (i, j) of G for each pair (y_i, y_j) of output variables. For any example $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, its joint feature vector is given by

$$\boldsymbol{\phi}(x, \mathbf{y}) = \big(\boldsymbol{\phi}_{i,j}(x, y_i, y_j)\big)_{(i,j) \in G} = \big(\boldsymbol{\varphi}(x) \otimes \boldsymbol{\psi}_{i,j}(y_i, y_j)\big)_{(i,j) \in G} \,,$$

where $\otimes$ is the Kronecker product. Hence, any predictor $\mathbf{w}$ can be written as $\mathbf{w} = (\mathbf{w}_{i,j})_{(i,j) \in G}$ where $\mathbf{w}_{i,j}$ is $\mathbf{w}$'s weight on $\boldsymbol{\phi}_{i,j}(x, y_i, y_j)$. Therefore, for any $\mathbf{w}$ and any $(x, \mathbf{y})$, we have

$$F(\mathbf{w}, x, \mathbf{y}) = \langle \mathbf{w}, \boldsymbol{\phi}(x, \mathbf{y}) \rangle = \sum_{(i,j) \in G} \langle \mathbf{w}_{i,j}, \boldsymbol{\phi}_{i,j}(x, y_i, y_j) \rangle = \sum_{(i,j) \in G} F_{i,j}(\mathbf{w}_{i,j}, x, y_i, y_j) \,,$$

where we denote by $F_{i,j}(\mathbf{w}_{i,j}, x, y_i, y_j) = \langle \mathbf{w}_{i,j}, \boldsymbol{\phi}_{i,j}(x, y_i, y_j)$ the score of labeling the edge (i, j) by (y_i, y_j) given input x.

For any vector $\mathbf{a}$, let $\|\mathbf{a}\|$ denote its L_2 norm. Throughout the paper, we make the assumption that we have a normalized joint feature space such that $\|\boldsymbol{\phi}(x, \mathbf{y})\| = 1$ for all $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ and $\|\boldsymbol{\phi}_{i,j}(x, y_i, y_j)\|$ is the same for all $(i, j) \in G$. Since the complete graph G has $\binom{\ell}{2}$ edges, it follows that $\|\boldsymbol{\phi}_{i,j}(x, y_i, y_j)\|^2 = \binom{\ell}{2}^{-1}$ for all $(i, j) \in G$.

We also have a training set $S \stackrel{\text{def}}{=} \{(x_1, \mathbf{y}_1), \ldots, (x_m, \mathbf{y}_m)\}$ where each example is generated independently according to some unknown distribution D. Mathematically, we do not assume the existence of a predictor $\mathbf{w}$ achieving some positive margin $\Gamma(\mathbf{w}, x, \mathbf{y})$ on each $(x, \mathbf{y}) \in S$. Indeed,

for some S, there might not exist any $\mathbf{w}$ where $\Gamma(\mathbf{w}, x, \mathbf{y}) > 0$ for all $(x, \mathbf{y}) \in S$. However, the generalization guarantee will be best when $\mathbf{w}$ achieves a large margin on most training points.

Given any $\gamma > 0$, and any $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, the *hinge loss* (at scale γ) incurred on $(x, \mathbf{y})$ by a unit L_2 norm predictor $\mathbf{w}$ that achieves a (possibly negative) margin $\Gamma(\mathbf{w}, x, \mathbf{y})$ is given by $\mathcal{L}^\gamma(\Gamma(\mathbf{w}, x, \mathbf{y}))$, where the so-called *hinge loss function* $\mathcal{L}^\gamma$ is defined as $\mathcal{L}^\gamma(s) \stackrel{\text{def}}{=} \max(0, 1 - s/\gamma)\ \forall s \in \mathbb{R}$. We will also make use of the *ramp loss function* $\mathcal{A}^\gamma$ defined by $\mathcal{A}^\gamma(s) \stackrel{\text{def}}{=} \min(1, \mathcal{L}^\gamma(s))\ \forall s \in \mathbb{R}$.

The proofs of all the rigorous results of this paper are provided in the supplementary material.

3 Superposition of Random Spanning Trees

Given a complete graph G of ℓ nodes (representing the Markov network), let $S(G)$ denote the set of all $\ell^{\ell-2}$ spanning trees of G. Recall that each spanning tree of G has $\ell - 1$ edges. Hence, for any edge $(i, j) \in G$, the number of trees in $S(G)$ covering that edge (i, j) is given by $\ell^{\ell-2}(\ell-1)/\binom{\ell}{2} = (2/\ell)\ell^{\ell-2}$. Therefore, for any function f of the edges of G we have

$$\sum_{T \in S(G)} \sum_{(i,j) \in T} f((i, j)) = \ell^{\ell-2} \frac{2}{\ell} \sum_{(i,j) \in G} f((i, j)) \ .$$

Given any spanning tree T of G and given any predictor $\mathbf{w}$, let $\mathbf{w}_T$ denote the projection of $\mathbf{w}$ on the edges of T. Namely, $(\mathbf{w}_T)_{i,j} = \mathbf{w}_{i,j}$ if $(i, j) \in T$, and $(\mathbf{w}_T)_{i,j} = 0$ otherwise. Let us also denote by $\boldsymbol{\phi}_T(x, \mathbf{y})$, the projection of $\boldsymbol{\phi}(x, \mathbf{y})$ on the edges of T. Namely, $(\boldsymbol{\phi}_T(x, \mathbf{y}))_{i,j} = \boldsymbol{\phi}_{i,j}(x, y_i, y_j)$ if $(i, j) \in T$, and $(\boldsymbol{\phi}_T(x, \mathbf{y}))_{i,j} = 0$ otherwise. Recall that $\|\boldsymbol{\phi}_{i,j}(x, y_i, y_j)\|^2 = \binom{\ell}{2}^{-1}\ \forall (i, j) \in G$. Thus, for all $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ and for all $T \in S(G)$, we have

$$\|\boldsymbol{\phi}_T(x, \mathbf{y})\|^2 = \sum_{(i,j) \in T} \|\boldsymbol{\phi}_{i,j}(x, y_i, y_j)\|^2 = \frac{\ell - 1}{\binom{\ell}{2}} = \frac{2}{\ell} \ .$$

We now establish how $F(\mathbf{w}, x, \mathbf{y})$ can be written as an expectation over all the spanning trees of G.

Lemma 1. *Let $\hat{\mathbf{w}}_T \stackrel{\text{def}}{=} \mathbf{w}_T / \|\mathbf{w}_T\|$, $\hat{\boldsymbol{\phi}}_T \stackrel{\text{def}}{=} \boldsymbol{\phi}_T / \|\boldsymbol{\phi}_T\|$. Let $\mathcal{U}(G)$ denote the uniform distribution on $S(G)$. Then, we have*

$$F(\mathbf{w}, x, \mathbf{y}) \ = \ \mathop{\mathbf{E}}_{T \sim \mathcal{U}(G)} a_T \langle \hat{\mathbf{w}}_T, \hat{\boldsymbol{\phi}}_T(x, \mathbf{y}) \rangle, \quad \textit{where } a_T \stackrel{\text{def}}{=} \sqrt{\frac{\ell}{2}}\, \|\mathbf{w}_T\| \ .$$

Moreover, for any $\mathbf{w}$ such that $\|\mathbf{w}\| = 1$, we have: $\mathop{\mathbf{E}}_{T \sim \mathcal{U}(G)} a_T^2 \ = \ 1$, and $\mathop{\mathbf{E}}_{T \sim \mathcal{U}(G)} a_T \ \leq \ 1$.

Let $\mathcal{T} \stackrel{\text{def}}{=} \{T_1, \ldots, T_n\}$ be a sample of n spanning trees of G where each T_i is sampled independently according to $\mathcal{U}(G)$. Given any unit L_2 norm predictor $\mathbf{w}$ on the complete graph G, our task is to investigate how the margins $\Gamma(\mathbf{w}, x, \mathbf{y})$, for each $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, will be modified if we approximate the (true) expectation over all spanning trees by an average over the sample $\mathcal{T}$.

For this task, we consider any $(x, \mathbf{y})$ and any $\mathbf{w}$ of unit L_2 norm. Let $F_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y})$ denote the estimation of $F(\mathbf{w}, x, \mathbf{y})$ on the tree sample $\mathcal{T}$,

$$F_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y}) \quad \stackrel{\text{def}}{=} \quad \frac{1}{n} \sum_{i=1}^{n} a_{T_i} \langle \hat{\mathbf{w}}_{T_i}, \hat{\boldsymbol{\phi}}_{T_i}(x, \mathbf{y}) \rangle \ ,$$

and let $\Gamma_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y})$ denote the estimation of $\Gamma(\mathbf{w}, x, \mathbf{y})$ on the tree sample $\mathcal{T}$,

$$\Gamma_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y}) \stackrel{\text{def}}{=} \min_{\mathbf{y}' \neq \mathbf{y}} \left[F_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y}) - F_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y}') \right] \ .$$

The following lemma states how $\Gamma_{\mathcal{T}}$ relates to Γ.

Lemma 2. *Consider any unit L_2 norm predictor $\mathbf{w}$ on the complete graph G that achieves a margin of $\Gamma(\mathbf{w}, x, \mathbf{y})$ for each $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, then we have*

$$\Gamma_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y}) \ \geq \ \Gamma(\mathbf{w}, x, \mathbf{y}) - 2\epsilon \quad \forall (x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y} \ ,$$

whenever we have $|F_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y}) - F(\mathbf{w}, x, \mathbf{y})| \leq \epsilon$ for all $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$.

Lemma 2 has important consequences whenever $|F_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y}) - F(\mathbf{w}, x, \mathbf{y})| \leq \epsilon$ for all $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$. Indeed, if $\mathbf{w}$ achieves a hard margin $\Gamma(\mathbf{w}, x, \mathbf{y}) \geq \gamma > 0$ for all $(x, \mathbf{y}) \in S$, then we have that $\mathbf{w}$ also achieves a hard margin of $\Gamma_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y}) \geq \gamma - 2\epsilon$ on each $(x, \mathbf{y}) \in S$ when using the tree sample $\mathcal{T}$ instead of the full graph G. More generally, if $\mathbf{w}$ achieves a ramp loss of $\mathcal{A}^{\gamma}(\Gamma(\mathbf{w}, x, \mathbf{y}))$ for each $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, then $\mathbf{w}$ achieves a ramp loss of $\mathcal{A}^{\gamma}(\Gamma_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y})) \leq \mathcal{A}^{\gamma}\left(\Gamma(\mathbf{w}, x, \mathbf{y}) - 2\epsilon\right)$ for all $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ when using the tree sample $\mathcal{T}$ instead of the full graph G. This last property follows directly from the fact that $\mathcal{A}^{\gamma}(s)$ is a non-increasing function of s.

The next lemma tells us that, apart from a slow $\ln^2(\sqrt{n})$ dependence, a sample of $n \in \Theta(\ell^2/\epsilon^2)$ spanning trees is sufficient to assure that the condition of Lemma 2 holds with high probability for all $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$. Such a fast convergence rate was made possible by using PAC-Bayesian methods which, in our case, prevented us of using the union bound over all possible $\mathbf{y} \in \mathcal{Y}$.

Lemma 3. *Consider any $\epsilon > 0$ and any unit L_2 norm predictor $\mathbf{w}$ for the complete graph G acting on a normalized joint feature space. For any $\delta \in (0, 1)$, let*

$$n \;\geq\; \frac{\ell^2}{\epsilon^2}\left(\frac{1}{16} + \frac{1}{2}\ln\frac{8\sqrt{n}}{\delta}\right)^2 . \tag{2}$$

Then with probability of at least $1 - \delta/2$ over all samples $\mathcal{T}$ generated according to $\mathcal{U}(G)^n$, we have, simultaneously for all $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, that $|F_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y}) - F(\mathbf{w}, x, \mathbf{y})| \;\leq \epsilon$.

Given a sample $\mathcal{T}$ of n spanning trees of G, we now consider an arbitrary set $\mathcal{W} \stackrel{\text{def}}{=} \{\hat{\mathbf{w}}_{T_1}, \ldots, \hat{\mathbf{w}}_{T_n}\}$ of unit L_2 norm weight vectors where each $\hat{\mathbf{w}}_{T_i}$ operates on a unit L_2 norm feature vector $\hat{\boldsymbol{\phi}}_{T_i}(x, \mathbf{y})$. For any $\mathcal{T}$ and any such set $\mathcal{W}$, we consider an arbitrary unit L_2 norm conical combination of each weight in $\mathcal{W}$ realized by a n-dimensional weight vector $\mathbf{q} \stackrel{\text{def}}{=} (q_1, \ldots, q_n)$, where $\sum_{i=1}^n q_i^2 = 1$ and each $q_i \geq 0$. Given any $(x, \mathbf{y})$ and any $\mathcal{T}$, we define the score $F_{\mathcal{T}}(\mathcal{W}, \mathbf{q}, x, \mathbf{y})$ achieved on $(x, \mathbf{y})$ by the conical combination $(\mathcal{W}, \mathbf{q})$ on $\mathcal{T}$ as

$$F_{\mathcal{T}}(\mathcal{W}, \mathbf{q}, x, \mathbf{y}) \stackrel{\text{def}}{=} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} q_i \langle \hat{\mathbf{w}}_{T_i}, \hat{\boldsymbol{\phi}}_{T_i}(x, \mathbf{y})\rangle \,, \tag{3}$$

where the $\sqrt{n}$ denominator ensures that we always have $F_{\mathcal{T}}(\mathcal{W}, \mathbf{q}, x, \mathbf{y}) \leq 1$ in view of the fact that $\sum_{i=1}^n q_i$ can be as large as $\sqrt{n}$. Note also that $F_{\mathcal{T}}(\mathcal{W}, \mathbf{q}, x, \mathbf{y})$ is the score of the feature vector obtained by the concatenation of all the weight vectors in $\mathcal{W}$ (and weighted by $\mathbf{q}$) acting on a feature vector obtained by concatenating each $\hat{\boldsymbol{\phi}}_{T_i}$ multiplied by $1/\sqrt{n}$. Hence, given $\mathcal{T}$, we define the margin $\Gamma_{\mathcal{T}}(\mathcal{W}, \mathbf{q}, x, \mathbf{y})$ achieved on $(x, \mathbf{y})$ by the conical combination $(\mathcal{W}, \mathbf{q})$ on $\mathcal{T}$ as

$$\Gamma_{\mathcal{T}}(\mathcal{W}, \mathbf{q}, x, \mathbf{y}) \;\stackrel{\text{def}}{=}\; \min_{\mathbf{y}' \neq \mathbf{y}} \left[F_{\mathcal{T}}(\mathcal{W}, \mathbf{q}, x, \mathbf{y}) - F_{\mathcal{T}}(\mathcal{W}, \mathbf{q}, x, \mathbf{y}')\right] . \tag{4}$$

For any unit L_2 norm predictor $\mathbf{w}$ that achieves a margin of $\Gamma(\mathbf{w}, x, \mathbf{y})$ for all $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, we now show that there exists, with high probability, a unit L_2 norm conical combination $(\mathcal{W}, \mathbf{q})$ on $\mathcal{T}$ achieving margins that are not much smaller than $\Gamma(\mathbf{w}, x, \mathbf{y})$.

Theorem 4. *Consider any unit L_2 norm predictor $\mathbf{w}$ for the complete graph G, acting on a normalized joint feature space, achieving a margin of $\Gamma(\mathbf{w}, x, \mathbf{y})$ for each $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$. Then for any $\epsilon > 0$, and any n satisfying Lemma 3, for any $\delta \in (0, 1]$, with probability of at least $1 - \delta$ over all samples $\mathcal{T}$ generated according to $\mathcal{U}(G)^n$, there exists a unit L_2 norm conical combination $(\mathcal{W}, \mathbf{q})$ on $\mathcal{T}$ such that, simultaneously for all $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, we have*

$$\Gamma_{\mathcal{T}}(\mathcal{W}, \mathbf{q}, x, \mathbf{y}) \;\geq\; \frac{1}{\sqrt{1+\epsilon}}\left[\Gamma(\mathbf{w}, x, \mathbf{y}) - 2\epsilon\right] .$$

From Theorem 4, and since $\mathcal{A}^{\gamma}(s)$ is a non-increasing function of s, it follows that, with probability at least $1 - \delta$ over the random draws of $\mathcal{T} \sim \mathcal{U}(G)^n$, there exists $(\mathcal{W}, \mathbf{q})$ on $\mathcal{T}$ such that, simultaneously for all $\forall (x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, for any n satisfying Lemma 3 we have

$$\mathcal{A}^{\gamma}(\Gamma_{\mathcal{T}}(\mathcal{W}, \mathbf{q}, x, \mathbf{y})) \;\leq\; \mathcal{A}^{\gamma}\left(\left[\Gamma(\mathbf{w}, x, \mathbf{y}) - 2\epsilon\right](1+\epsilon)^{-1/2}\right) .$$

Hence, instead of searching for a predictor $\mathbf{w}$ for the complete graph G that achieves a small expected ramp loss $\mathbf{E}_{(x,\mathbf{y})\sim D}\mathcal{A}^{\gamma}(\Gamma(\mathbf{w}, x, \mathbf{y})$, Theorem 4 tells us that we can settle the search for a

unit L_2 norm conical combination $(\mathcal{W}, \mathbf{q})$ on a sample $\mathcal{T}$ of randomly-generated spanning trees of G that achieves small $\mathbf{E}_{(x,\mathbf{y})\sim D}\mathcal{A}^\gamma(\Gamma_\mathcal{T}(\mathcal{W}, \mathbf{q}, x, \mathbf{y}))$. But recall that $\Gamma_\mathcal{T}(\mathcal{W}, \mathbf{q}, x, \mathbf{y}))$ is the margin of a weight vector obtained by the concatenation of all the weight vectors in $\mathcal{W}$ (weighted by $\mathbf{q}$) on a feature vector obtained by the concatenation of the n feature vectors $(1/\sqrt{n})\hat{\boldsymbol{\phi}}_{T_i}$. It thus follows that any standard risk bound for the SVM applies directly to $\mathbf{E}_{(x,\mathbf{y})\sim D}\mathcal{A}^\gamma(\Gamma_\mathcal{T}(\mathcal{W}, \mathbf{q}, x, \mathbf{y}))$. Hence, by adapting the SVM risk bound of [8], we have the following result.

Theorem 5. *Consider any sample $\mathcal{T}$ of n spanning trees of the complete graph G. For any $\gamma > 0$ and any $0 < \delta \leq 1$, with probability of at least $1 - \delta$ over the random draws of $S \sim D^m$, simultaneously for all unit L_2 norm conical combinations $(\mathcal{W}, \mathbf{q})$ on $\mathcal{T}$, we have*

$$\mathop{\mathbf{E}}_{(x,\mathbf{y})\sim D} \mathcal{A}^\gamma(\Gamma_\mathcal{T}(\mathcal{W}, \mathbf{q}, x, \mathbf{y})) \;\leq\; \frac{1}{m}\sum_{i=1}^{m} \mathcal{A}^\gamma(\Gamma_\mathcal{T}(\mathcal{W}, \mathbf{q}, x_i, \mathbf{y}_i)) + \frac{2}{\gamma\sqrt{m}} + 3\sqrt{\frac{\ln(2/\delta)}{2m}}\,.$$

Hence, according to this theorem, the conical combination $(\mathcal{W}, \mathbf{q})$ having the best generalization guarantee is the one which minimizes the sum of the first two terms on the right hand side of the inequality. Note that the theorem is still valid if we replace, in the empirical risk term, the non-convex ramp loss $\mathcal{A}^\gamma$ by the convex hinge loss $\mathcal{L}^\gamma$. This provides the theoretical basis of the proposed optimization problem for learning $(\mathcal{W}, \mathbf{q})$ on the sample $\mathcal{T}$.

4 A L_2-Norm Random Spanning Tree Approximation Approach

If we introduce the usual slack variables $\xi_k \stackrel{\text{def}}{=} \gamma \cdot \mathcal{L}^\gamma(\Gamma_\mathcal{T}(\mathcal{W}, \mathbf{q}, x_k, \mathbf{y}_k)$, Theorem 5 suggests that we should minimize $\frac{1}{\gamma}\sum_{k=1}^m \xi_k$ for some fixed margin value $\gamma > 0$. Rather than performing this task for several values of γ, we show in the supplementary material that we can, equivalently, solve the following optimization problem for several values of $C > 0$.

Definition 6. Primal L_2-norm Random Tree Approximation.

$$\begin{aligned}
\textbf{\textit{min}}_{\mathbf{w}_{T_i},\xi_k} \quad & \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{w}_{T_i}||_2^2 + C\sum_{k=1}^{m}\xi_k \\
\textbf{\textit{s.t.}} \quad & \sum_{i=1}^{n} \langle \mathbf{w}_{T_i}, \hat{\boldsymbol{\phi}}_{T_i}(x_k, \mathbf{y}_k)\rangle - \textbf{\textit{max}}_{\mathbf{y}\neq\mathbf{y}_k} \sum_{i=1}^{n} \langle \mathbf{w}_{T_i}, \hat{\boldsymbol{\phi}}_{T_i}(x_k, \mathbf{y})\rangle \geq 1 - \xi_k, \\
& \xi_k \geq 0\,, \forall\, k \in \{1, \ldots, m\},
\end{aligned}$$

where $\{\mathbf{w}_{T_i} | T_i \in \mathcal{T}\}$ are the feature weights to be learned on each tree, ξ_k is the margin slack allocated for each x_k, and C is the slack parameter that controls the amount of regularization.

This primal form has the interpretation of maximizing the joint margins from individual trees between (correct) training examples and all the other (incorrect) examples.

The key for the efficient optimization is solving the 'argmax' problem efficiently. In particular, we note that the space of all multilabels is exponential in size, thus forbidding exhaustive enumeration over it. In the following, we show how exact inference over a collection $\mathcal{T}$ of trees can be implemented in $\Theta(Kn\ell)$ time per data point, where K is the smallest number such that the average score of the K'th best multilabel for each tree of $\mathcal{T}$ is at most $F_\mathcal{T}(x, \mathbf{y}) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \langle \mathbf{w}_{T_i}, \hat{\boldsymbol{\phi}}_{T_i}(x, \mathbf{y})\rangle$. Whenever K is polynomial in the number of labels, this gives us exact polynomial-time inference over the ensemble of trees.

4.1 Fast inference over a collection of trees

It is well known that the exact solution to the inference problem

$$\hat{\mathbf{y}}_{T_i}(x) = \mathop{\text{argmax}}_{\mathbf{y}\in\mathcal{Y}}\; F_{\mathbf{w}_{T_i}}(x, \mathbf{y}) \;\stackrel{\text{def}}{=}\; \mathop{\text{argmax}}_{\mathbf{y}\in\mathcal{Y}}\; \langle \mathbf{w}_{T_i}, \hat{\boldsymbol{\phi}}_{T_i}(x, \mathbf{y})\rangle, \tag{5}$$

on an individual tree T_i can be obtained in $\Theta(\ell)$ time by dynamic programming. However, there is no guarantee that the maximizer $\hat{\mathbf{y}}_{T_i}$ of Equation (5) is also a maximizer of $F_\mathcal{T}$. In practice, $\hat{\mathbf{y}}_{T_i}$

can differ for each spanning tree $T_i \in \mathcal{T}$. Hence, instead of using only the best scoring multilabel $\hat{\mathbf{y}}_{T_i}$ from each individual $T_i \in \mathcal{T}$, we consider the set of the K highest scoring multilabels $\mathcal{Y}_{T_i,K} = \{\hat{\mathbf{y}}_{T_i,1}, \cdots, \hat{\mathbf{y}}_{T_i,K}\}$ of $F_{\mathbf{w}_{T_i}}(x, \mathbf{y})$. In the supplementary material we describe a dynamic programming to find the K highest multilabels in $\Theta(K\ell)$ time. Running this algorithm for all of the trees gives us a candidate set of $\Theta(Kn)$ multilabels $\mathcal{Y}_{\mathcal{T},K} = \mathcal{Y}_{T_1,K} \cup \cdots \cup \mathcal{Y}_{T_n,K}$. We now state a key lemma that will enable us to verify if the candidate set contains the maximizer of $F_{\mathcal{T}}$.

Lemma 7. *Let* $\mathbf{y}_K^\star = \operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}_{\mathcal{T},K}} F_{\mathcal{T}}(x, \mathbf{y})$ *be the highest scoring multilabel in* $\mathcal{Y}_{\mathcal{T},K}$*. Suppose that*

$$F_{\mathcal{T}}(x, \mathbf{y}_K^\star) \geq \frac{1}{n}\sum_{i=1}^{n} F_{\mathbf{w}_{T_i}}(x, \mathbf{y}_{T_i,K}) \stackrel{\text{def}}{=} \theta_x(K).$$

It follows that $F_{\mathcal{T}}(x, \mathbf{y}_K^\star) = \max_{\mathbf{y}\in\mathcal{Y}} F_{\mathcal{T}}(x, \mathbf{y})$.

We can use any K satisfying the lemma as the length of K-best lists, and be assured that $\mathbf{y}_K^\star$ is a maximizer of $F_{\mathcal{T}}$.

We now examine the conditions under which the highest scoring multilabel is present in our candidate set $\mathcal{Y}_{\mathcal{T},K}$ with high probability. For any $x \in \mathcal{X}$ and any predictor $\mathbf{w}$, let $\hat{\mathbf{y}} \stackrel{\text{def}}{=} \mathbf{y}_{\mathbf{w}}(x) \stackrel{\text{def}}{=} \operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{w}, x, \mathbf{y})$ be the highest scoring multilabel in $\mathcal{Y}$ for predictor $\mathbf{w}$ on the complete graph G.

For any $\mathbf{y} \in \mathcal{Y}$, let $K_T(\mathbf{y})$ be the rank of $\mathbf{y}$ in tree T and let $\rho_T(\mathbf{y}) \stackrel{\text{def}}{=} K_T(\mathbf{y})/|\mathcal{Y}|$ be the normalized rank of $\mathbf{y}$ in tree T. We then have $0 < \rho_T(\mathbf{y}) \leq 1$ and $\rho_T(\mathbf{y}') = \min_{\mathbf{y}\in\mathcal{Y}} \rho_T(\mathbf{y})$ whenever $\mathbf{y}'$ is a highest scoring multilabel in tree T. Since $\mathbf{w}$ and x are arbitrary and fixed, let us drop them momentarily from the notation and let $F(\mathbf{y}) \stackrel{\text{def}}{=} F(\mathbf{w}, x, \mathbf{y})$, and $F_T(\mathbf{y}) \stackrel{\text{def}}{=} F_{\mathbf{w}_T}(x, \mathbf{y})$. Let $\mathcal{U}(\mathcal{Y})$ denote the uniform distribution of multilabels on $\mathcal{Y}$. Then, let $\mu_T \stackrel{\text{def}}{=} \mathbf{E}_{\mathbf{y}\sim\mathcal{U}(\mathcal{Y})} F_T(\mathbf{y})$ and $\mu \stackrel{\text{def}}{=} \mathbf{E}_{T\sim\mathcal{U}(G)} \mu_T$.

Let $\mathcal{T} \sim \mathcal{U}(G)^n$ be a sample of n spanning trees of G. Since the scoring function F_T of each tree T of G is bounded in absolute value, it follows that F_T is a σ_T-sub-Gaussian random variable for some $\sigma_T > 0$. We now show that, with high probability, there exists a tree $T \in \mathcal{T}$ such that $\rho_T(\hat{\mathbf{y}})$ is decreasing exponentially rapidly with $(F(\hat{\mathbf{y}}) - \mu)/\sigma$, where $\sigma^2 \stackrel{\text{def}}{=} \mathbf{E}_{T\sim\mathcal{U}(G)} \sigma_T^2$.

Lemma 8. *Let the scoring function* F_T *of each spanning tree of* G *be a* σ_T*-sub-Gaussian random variable under the uniform distribution of labels;* i.e.*, for each* T *on* G*, there exists* $\sigma_T > 0$ *such that for any* $\lambda > 0$ *we have*

$$\mathop{\mathbf{E}}_{\mathbf{y}\sim\mathcal{U}(\mathcal{Y})} e^{\lambda(F_T(\mathbf{y}) - \mu_T)} \;\leq\; e^{\frac{\lambda^2}{2}\sigma_T^2}\,.$$

Let $\sigma^2 \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{T\sim\mathcal{U}(G)} \sigma_T^2$*, and let* $\alpha \stackrel{\text{def}}{=} \Pr_{T\sim\mathcal{U}(G)} \Big(\mu_T \leq \mu \;\wedge\; F_T(\hat{\mathbf{y}}) \geq F(\hat{\mathbf{y}}) \;\wedge\; \sigma_T^2 \leq \sigma^2\Big)$*. Then,*

$$\Pr_{\mathcal{T}\sim\mathcal{U}(G)^n} \left(\exists T \in \mathcal{T}\colon\; \rho_T(\hat{\mathbf{y}}) \leq e^{-\frac{1}{2}\frac{(F(\hat{\mathbf{y}})-\mu)^2}{\sigma^2}}\right) \geq 1 - (1-\alpha)^n\,.$$

Thus, even for very small α, when n is large enough, there exists, with high probability, a tree $T \in \mathcal{T}$ such that $\hat{\mathbf{y}}$ has a small $\rho_T(\hat{\mathbf{y}})$ whenever $[F(\hat{\mathbf{y}}) - \mu]/\sigma$ is large for G. For example, when $|\mathcal{Y}| = 2^\ell$ (the multiple binary classification case), we have with probability of at least $1 - (1-\alpha)^n$, that there exists $T \in \mathcal{T}$ such that $K_T(\hat{\mathbf{y}}) = 1$ whenever $F(\hat{\mathbf{y}}) - \mu \geq \sigma\sqrt{2\ell \ln 2}$.

4.2 Optimization

To optimize the L_2-norm RTA problem (Definition 6) we convert it to the marginalized dual form (see the supplementary material for the derivation), which gives us a polynomial-size problem (in the number of microlabels) and allows us to use kernels to tackle complex input spaces efficiently.

Definition 9. **L_2-norm RTA Marginalized Dual**

$$\max_{\boldsymbol{\mu}\in\mathcal{M}^m} \quad \frac{1}{|E_{\mathcal{T}}|}\sum_{e,k,\mathbf{u}_e} \mu(k, e, \mathbf{u}_e) - \frac{1}{2}\sum_{\substack{e,k,\mathbf{u}_e,\\ k',\mathbf{u}'_e}} \mu(k, e, \mathbf{u}_e) K_{\mathcal{T}}^e(x_k, \mathbf{u}_e; x'_k, \mathbf{u}'_e)\mu(k', e, \mathbf{u}'_e)\,,$$

where $E_{\mathcal{T}}$ *is the union of the sets of edges appearing in* $\mathcal{T}$*, and* $\boldsymbol{\mu} \in \mathcal{M}^m$ *are the marginal dual variables* $\boldsymbol{\mu} \stackrel{\text{def}}{=} (\mu(k, e, \mathbf{u}_e))_{k,e,\mathbf{u}_e}$*, with the triplet* $(k, e, \mathbf{u}_e)$ *corresponding to labeling the edge*

DATASET	MICROLABEL LOSS (%)					0/1 LOSS (%)				
	SVM	MTL	MMCRF	MAM	RTA	SVM	MTL	MMCRF	MAM	RTA
EMOTIONS	22.4	20.2	20.1	*19.5*	**18.8**	77.8	74.5	71.3	*69.6*	**66.3**
YEAST	*20.0*	20.7	21.7	20.1	**19.8**	*85.9*	88.7	93.0	86.0	**77.7**
SCENE	*9.8*	11.6	18.4	17.0	**8.8**	*47.2*	55.2	72.2	94.6	**30.2**
ENRON	6.4	6.5	6.2	**5.0**	*5.3*	99.6	99.6	92.7	*87.9*	**87.7**
CAL500	**13.7**	*13.8*	**13.7**	**13.7**	*13.8*	100.0	100.0	100.0	100.0	100.0
FINGERPRINT	**10.3**	17.3	*10.5*	*10.5*	10.7	99.0	100.0	*99.6*	*99.6*	**96.7**
NCI60	15.3	16.0	*14.6*	**14.3**	14.9	56.9	*53.0*	63.1	60.0	**52.9**
MEDICAL	*2.6*	*2.6*	**2.1**	**2.1**	**2.1**	91.8	91.8	63.8	*63.1*	**58.8**
CIRCLE10	4.7	6.3	2.6	*2.5*	**0.6**	28.9	33.2	20.3	*17.7*	**4.0**
CIRCLE50	5.7	6.2	**1.5**	*2.1*	3.8	69.8	72.3	**38.8**	*46.2*	*52.8*

Table 1: Prediction performance of each algorithm in terms of microlabel loss and 0/1 loss. The best performing algorithm is highlighted with **boldface**, the second best is in *italic*.

$e=(v,v') \in E_{\mathcal{T}}$ of the output graph by $\mathbf{u}_e=(u_v,u_{v'}) \in \mathcal{Y}_v \times \mathcal{Y}_{v'}$ for the training example x_k. Also, $\mathcal{M}^m$ is the marginal dual feasible set and

$$K^e_{\mathcal{T}}(x_k,\mathbf{u}_e;x_{k'},\mathbf{u}'_e) \stackrel{\text{def}}{=} \frac{N_{\mathcal{T}}(e)}{|E_{\mathcal{T}}|^2} K(x_k,x_{k'}) \left\langle \boldsymbol{\psi}_e(y_{kv},y_{kv'}) - \boldsymbol{\psi}_e(u_v,u_{v'}), \boldsymbol{\psi}_e(y_{k'v},y_{k'v'}) - \boldsymbol{\psi}_e(u'_v,u'_{v'}) \right\rangle$$

is the joint kernel of input features and the differences of output features of true and competing multilabels ($\mathbf{y}_k,\mathbf{u}$), projected to the edge e. Finally, $\mathcal{N}_{\mathcal{T}}(e)$ denotes the number of times e appears among the trees of the ensemble.

The master algorithm described in the supplementary material iterates over each training example until convergence. The processing of each training example x_k proceeds by finding the worst violating multilabel of the ensemble defined as

$$\bar{\mathbf{y}}_k \stackrel{\text{def}}{=} \operatorname*{argmax}_{\mathbf{y} \neq \mathbf{y}_k} F_{\mathcal{T}}(x_k,\mathbf{y}) \,, \tag{6}$$

using the K-best inference approach of the previous section, with the modification that the correct multilabel is excluded from the K-best lists. The worst violator $\bar{\mathbf{y}}_k$ is mapped to a vertex

$$\bar{\boldsymbol{\mu}}(x_k) = C \cdot ([\bar{\mathbf{y}}_e = \mathbf{u}_e])_{e,\mathbf{u}_e} \in \mathcal{M}_k$$

corresponding to the steepest feasible ascent direction (*c.f*, [9]) in the marginal dual feasible set $\mathcal{M}_k$ of example x_k, thus giving us a subgradient of the objective of Definition 9. An exact line search is used to find the saddle point between the current solution and $\bar{\boldsymbol{\mu}}$.

5 Empirical Evaluation

We compare our method RTA to Support Vector Machine (SVM) [10, 11], Multitask Feature Learning (MTL) [12], Max-Margin Conditional Random Fields (MMCRF) [9] which uses the loopy belief propagation algorithm for approximate inference on the general graph, and Maximum Average Marginal Aggregation (MAM) [5] which is a multilabel ensemble model that trains a set of random tree based learners separately and performs the final approximate inference on a union graph of the edge potential functions of the trees. We use ten multilabel datasets from [5]. Following [5], MAM is constructed with 180 tree based learners, and for MMCRF a consensus graph is created by pooling edges from 40 trees. We train RTA with up to 40 spanning trees and with K up to 32. The linear kernel is used for methods that require kernelized input. Margin slack parameters are selected from $\{100, 50, 10, 1, 0.5, 0.1, 0.01\}$. We use 5-fold cross-validation to compute the results.

Prediction performance. Table 1 shows the performance in terms of microlabel loss and 0/1 loss. The best methods are highlighted in '**boldface**' and the second best in '*italics*' (see supplementary material for full results). RTA quite often improves over MAM in 0/1 accuracy, sometimes with noticeable margin except for *Enron* and *Circle50*. The performances in microlabel accuracy are quite similar while RTA is slightly above the competition. This demonstrates the advantage of RTA that gains by optimizing on a collection of trees simultaneously rather than optimizing on individual trees as MAM. In addition, learning using approximate inference on a general graph seems less

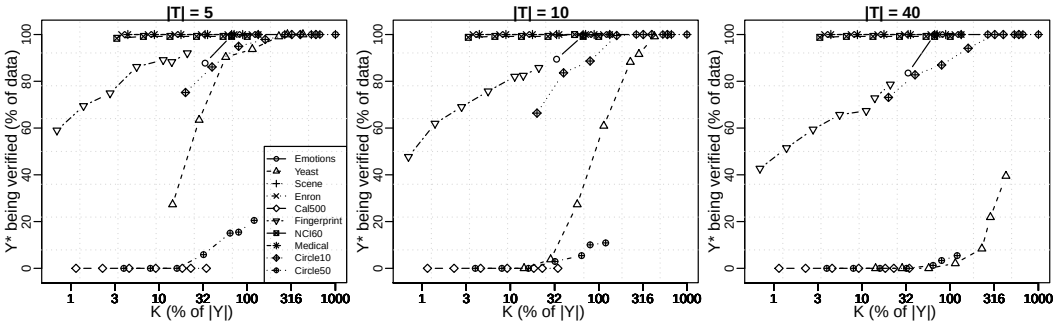

Figure 1: Percentage of examples with provably optimal $\mathbf{y}^*$ being in the K-best lists plotted as a function of K, scaled with respect to the number of microlabels in the dataset.

favorable as the tree-based methods, as MMCRF quite consistently trails to RTA and MAM in both microlabel and 0/1 error, except for *Circle50* where it outperforms other models. Finally, we notice that SVM, as a single label classifier, is very competitive against most multilabel methods for microlabel accuracy.

Exactness of inference on the collection of trees. We now study the empirical behavior of the inference (see Section 4) on the collection of trees, which, if taken as a single general graph, would call for solving an $\mathcal{NP}$-hard inference problem. We provide here empirical evidence that we can perform exact inference on most examples in most datasets in polynomial time.
We ran the K-best inference on eleven datasets where the RTA models were trained with different amounts of spanning trees $|\mathcal{T}| = \{5, 10, 40\}$ and values for $K = \{2, 4, 8, 16, 32, 40, 60\}$. For each parameter combination and for each example, we recorded whether the K-best inference was provably exact on the collection (*i.e.*, if Lemma 7 was satisfied). Figure 1 plots the percentage of examples where the inference was indeed provably exact. The values are shown as a function of K, expressed as the percentage of the number of microlabels in each dataset. Hence, 100% means $K = \ell$, which denotes low polynomial ($\Theta(n\ell^2)$) time inference in the exponential size multilabel space.

We observe, from Figure 1, on some datasets (*e.g.*, *Medical*, *NCI60*), that the inference task is very easy since exact inference can be computed for most of the examples even with K values that are below 50% of the number of microlabels. By setting $K = \ell$ (*i.e.*, 100%) we can perform exact inference for about 90% of the examples on nine datasets with five trees, and eight datasets with 40 trees. On two of the datasets (*Cal500*, *Circle50*), inference is not (in general) exact with low values of K. Allowing K to grow superlinearly on ℓ would possibly permit exact inference on these datasets. However, this is left for future studies.

Finally, we note that the difficulty of performing provably exact inference slightly increases when more spanning trees are used. We have observed that, in most cases, the optimal multilabel $\mathbf{y}^*$ is still on the K-best lists but the conditions of Lemma 7 are no longer satisfied, hence forbidding us to prove exactness of the inference. Thus, working to establish alternative proofs of exactness is a worthy future research direction.

6 Conclusion

The main theoretical result of the paper is the demonstration that if a large margin structured output predictor exists, then combining a small sample of random trees will, with high probability, generate a predictor with good generalization. The key attraction of this approach is the tractability of the inference problem for the ensemble of trees, both indicated by our theoretical analysis and supported by our empirical results. However, as a by-product, we have a significant added benefit: we do not need to know the output structure *a priori* as this is generated implicitly in the learned weights for the trees. This is used to significant advantage in our experiments that automatically leverage correlations between the multiple target outputs to give a substantive increase in accuracy. It also suggests that the approach has enormous potential for applications where the structure of the output is not known but is expected to play an important role.

References

[1] Ben Taskar, Carlos Guestrin, and Daphne Koller. Max-margin markov networks. In S. Thrun, L.K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 25–32. MIT Press, 2004.

[2] Martin J. Wainwright, Tommy S. Jaakkola, and Alan S. Willsky. MAP estimation via agreement on trees: message-passing and linear programming. *IEEE Transactions on Information Theory*, 51(11):3697–3717, 2005.

[3] Michael I. Jordan and Martin J Wainwright. Semidefinite relaxations for approximate inference on graphs with cycles. In S. Thrun, L.K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 369–376. MIT Press, 2004.

[4] Amir Globerson and Tommi S. Jaakkola. Approximate inference using planar graph decomposition. In B. Schölkopf, J.C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 473–480. MIT Press, 2007.

[5] Hongyu Su and Juho Rousu. Multilabel classification through random graph ensembles. *Machine Learning*, dx.doi.org/10.1007/s10994-014-5465-9, 2014.

[6] Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, New York, 1999.

[7] Thomas Gärtner and Shankar Vembu. On structured output training: hard cases and an efficient alternative. *Machine Learning*, 79:227–242, 2009.

[8] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[9] J. Rousu, C. Saunders, S. Szedmak, and J. Shawe-Taylor. Efficient algorithms for max-margin structured classification. *Predicting Structured Data*, pages 105–129, 2007.

[10] Kristin P. Bennett. Combining support vector and mathematical programming methods for classifications. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods—Support Vector Learning*, pages 307–326. MIT Press, Cambridge, MA, 1999.

[11] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, U.K., 2000.

[12] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[13] Yevgeny Seldin, François Laviolette, Nicolò Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58:7086–7093, 2012.

[14] Andreas Maurer. A note on the PAC Bayesian theorem. *CoRR*, cs.LG/0411099, 2004.

[15] David McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51:5–21, 2003.

[16] Juho Rousu, Craig Saunders, Sandor Szedmak, and John Shawe-Taylor. Kernel-based learning of hierarchical multilabel classification models. *Journal of Machine Learning Research*, 7:1601–1626, December 2006.

Multilabel Structured Output Learning with Random Spanning Trees of Max-Margin Markov Networks (supplementary material) by: Mario Marchand, Hongyu Su, Emilie Morvant, Juho Rousou, and John Shawe-Taylor

Lemma 1

Let $\hat{\mathbf{w}}_T \stackrel{def}{=} \mathbf{w}_T/\|\mathbf{w}_T\|$, $\hat{\boldsymbol{\phi}}_T \stackrel{def}{=} \boldsymbol{\phi}_T/\|\boldsymbol{\phi}_T\|$. *Let* $\mathcal{U}(G)$ *denote the uniform distribution on* $S(G)$. *Then, we have*

$$F(\mathbf{w}, x, \mathbf{y}) \;=\; \mathop{\mathbf{E}}_{T\sim\mathcal{U}(G)} a_T \langle \hat{\mathbf{w}}_T, \hat{\boldsymbol{\phi}}_T(x,\mathbf{y})\rangle, \text{ where } a_T \stackrel{def}{=} \sqrt{\frac{\ell}{2}}\,\|\mathbf{w}_T\|\,.$$

Moreover, for any $\mathbf{w}$ *such that* $\|\mathbf{w}\| = 1$, *we have*

$$\mathop{\mathbf{E}}_{T\sim\mathcal{U}(G)} a_T^2 \;=\; 1 \quad ; \quad \mathop{\mathbf{E}}_{T\sim\mathcal{U}(G)} a_T \;\le\; 1\,.$$

Proof.

$$\begin{aligned}
F(\mathbf{w}, x, \mathbf{y}) &= \langle \mathbf{w}, \boldsymbol{\phi}(x,\mathbf{y})\rangle \\
&= \sum_{(i,j)\in G} \langle \mathbf{w}_{i,j}, \boldsymbol{\phi}_{i,j}(x, y_i, y_j)\rangle \;=\; \frac{1}{\ell^{\ell-2}}\frac{\ell}{2}\sum_{T\in S(G)}\sum_{(i,j)\in T} \langle \mathbf{w}_{i,j}, \boldsymbol{\phi}_{i,j}(x, y_i, y_j)\rangle \\
&= \frac{\ell}{2}\mathop{\mathbf{E}}_{T\sim\mathcal{U}(G)} \langle \mathbf{w}_T, \boldsymbol{\phi}_T(x,\mathbf{y})\rangle \;=\; \frac{\ell}{2}\mathop{\mathbf{E}}_{T\sim\mathcal{U}(G)} \|\mathbf{w}_T\|\|\boldsymbol{\phi}_T(x,\mathbf{y})\|\langle \hat{\mathbf{w}}_T, \hat{\boldsymbol{\phi}}_T(x,\mathbf{y})\rangle \\
&= \mathop{\mathbf{E}}_{T\sim\mathcal{U}(G)} \sqrt{\frac{\ell}{2}}\,\|\mathbf{w}_T\|\langle \hat{\mathbf{w}}_T, \hat{\boldsymbol{\phi}}_T(x,\mathbf{y})\rangle \;=\; \mathop{\mathbf{E}}_{T\sim\mathcal{U}(G)} a_T\langle \hat{\mathbf{w}}_T, \hat{\boldsymbol{\phi}}_T(x,\mathbf{y})\rangle\,,
\end{aligned}$$

where

$$a_T \stackrel{\text{def}}{=} \sqrt{\frac{\ell}{2}}\,\|\mathbf{w}_T\|\,.$$

Now, for any $\mathbf{w}$ such that $\|\mathbf{w}\| = 1$, we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{T\sim\mathcal{U}(G)} a_T^2 &= \frac{\ell}{2}\mathop{\mathbf{E}}_{T\sim\mathcal{U}(G)} \|\mathbf{w}_T\|^2 \;=\; \frac{\ell}{2}\frac{1}{\ell^{\ell-2}}\sum_{T\in S(G)} \|\mathbf{w}_T\|^2 \;=\; \frac{\ell}{2}\frac{1}{\ell^{\ell-2}}\sum_{T\in S(G)}\sum_{(i,j)\in T} \|\mathbf{w}_{i,j}\|^2 \\
&= \sum_{(i,j)\in G} \|\mathbf{w}_{i,j}\|^2 \;=\; \|\mathbf{w}\|^2 \;=\; 1\,.
\end{aligned}$$

Since the variance of a_T must be positive, we have, for any $\mathbf{w}$ of unit L_2 norm, that

$$\mathop{\mathbf{E}}_{T\sim\mathcal{U}(G)} a_T \;\le\; 1\,.$$

□

Lemma 2

Consider any unit L_2 *norm predictor* $\mathbf{w}$ *on the complete graph* G *that achieves a margin of* $\Gamma(\mathbf{w}, x, \mathbf{y})$ *for each* $(x, \mathbf{y}) \in \mathcal{X}\times\mathcal{Y}$, *then we have*

$$\Gamma_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y}) \;\ge\; \Gamma(\mathbf{w}, x, \mathbf{y}) - 2\epsilon \quad \forall (x,\mathbf{y}) \in \mathcal{X}\times\mathcal{Y}\,,$$

whenever for all $(x, \mathbf{y}) \in \mathcal{X}\times\mathcal{Y}$, *we have*

$$|F_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y}) - F(\mathbf{w}, x, \mathbf{y})| \le \epsilon\,.$$

Proof. From the condition of the lemma, we have simultaneously for all $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ and $(x, \mathbf{y}') \in \mathcal{X} \times \mathcal{Y}$, that

$$F_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y}) \;\geq\; F(\mathbf{w}, x, \mathbf{y}) - \epsilon \quad \text{AND} \quad F_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y}') \;\leq\; F(\mathbf{w}, x, \mathbf{y}') + \epsilon\,.$$

Therefore,

$$F_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y}) - F_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y}') \;\geq\; F(\mathbf{w}, x, \mathbf{y}) - F(\mathbf{w}, x, \mathbf{y}') - 2\epsilon\,.$$

Hence, for all $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, we have

$$\Gamma_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y}) \;\geq\; \Gamma(\mathbf{w}, x, \mathbf{y}) - 2\epsilon\,.$$

□

Lemma 3

Consider any $\epsilon > 0$ and any unit L_2 norm predictor $\mathbf{w}$ for the complete graph G acting on a normalized joint feature space. For any $\delta \in (0, 1)$, let

$$n \;\geq\; \frac{\ell^2}{\epsilon^2}\left(\frac{1}{16} + \frac{1}{2}\ln\frac{8\sqrt{n}}{\delta}\right)^2\,. \tag{2}$$

Then with probability of at least $1 - \delta/2$ over all samples $\mathcal{T}$ generated according to $\mathcal{U}(G)^n$, we have, simultaneously for all $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, that

$$|F_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y}) - F(\mathbf{w}, x, \mathbf{y})| \;\leq \epsilon\,.$$

Proof. Consider an isotropic Gaussian distribution of joint feature vectors of variance σ^2, centred on $\boldsymbol{\phi}(x, \mathbf{y})$, with a density given by

$$Q_{\boldsymbol{\phi}}(\boldsymbol{\zeta}) \;\stackrel{\text{def}}{=}\; \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp - \frac{\|\boldsymbol{\zeta} - \boldsymbol{\phi}\|^2}{2\sigma^2}\,,$$

where N is the dimension of the feature vectors. When the feature space is infinite-dimensional, we can consider Q to be a Gaussian process. The end results will not depend on N.

Given the fixed $\mathbf{w}$ stated in the theorem, let us define the risk $R(Q_{\boldsymbol{\phi}}, \mathbf{w}_T)$ of $Q_{\boldsymbol{\phi}}$ on the tree T by $\mathop{\mathbf{E}}_{\boldsymbol{\zeta}\sim Q_{\boldsymbol{\phi}}} \langle \mathbf{w}_T, \boldsymbol{\zeta}\rangle$. By the linearity of $\langle\cdot,\cdot\rangle$, we have

$$R(Q_{\boldsymbol{\phi}}, \mathbf{w}_T) \;\stackrel{\text{def}}{=}\; \mathop{\mathbf{E}}_{\boldsymbol{\zeta}\sim Q_{\boldsymbol{\phi}}} \langle \mathbf{w}_T, \boldsymbol{\zeta}\rangle \;=\; \langle \mathbf{w}_T, \mathop{\mathbf{E}}_{\boldsymbol{\zeta}\sim Q_{\boldsymbol{\phi}}} \boldsymbol{\zeta}\rangle \;=\; \langle \mathbf{w}_T, \boldsymbol{\phi}\rangle\,,$$

which is independent of σ.

Gibbs' risk $R(Q_{\boldsymbol{\phi}})$ and its empirical estimate $R_{\mathcal{T}}(Q_{\boldsymbol{\phi}})$ are defined as

$$\begin{aligned} R(Q_{\boldsymbol{\phi}}) \;&\stackrel{\text{def}}{=}\; \mathop{\mathbf{E}}_{T\sim\mathcal{U}(G)} R(Q_{\boldsymbol{\phi}}, \mathbf{w}_T) \;=\; \mathop{\mathbf{E}}_{T\sim\mathcal{U}(G)} \langle \mathbf{w}_T, \boldsymbol{\phi}\rangle \\ R_{\mathcal{T}}(Q_{\boldsymbol{\phi}}) \;&\stackrel{\text{def}}{=}\; \frac{1}{n}\sum_{i=1}^{n} R(Q_{\boldsymbol{\phi}}, \mathbf{w}_{T_i}) \;=\; \frac{1}{n}\sum_{i=1}^{n} \langle \mathbf{w}_{T_i}, \boldsymbol{\phi}\rangle\,. \end{aligned}$$

Consequently, from the definitions of F and $F_{\mathcal{T}}$, we have

$$\begin{aligned} F(\mathbf{w}, x, \mathbf{y}) \;&=\; \frac{\ell}{2} R(Q_{\boldsymbol{\phi}(x,\mathbf{y})}) \\ F_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y}) \;&=\; \frac{\ell}{2} R_{\mathcal{T}}(Q_{\boldsymbol{\phi}(x,\mathbf{y})})\,. \end{aligned}$$

Recall that $\boldsymbol{\phi}$ is a normalized feature map that applies to all $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$. Therefore, if we have with probability $\geq 1 - \delta/2$ that, simultaneously for all $\boldsymbol{\phi}$ of unit L_2 norm,

$$\frac{\ell}{2}|R_{\mathcal{T}}(Q_{\boldsymbol{\phi}}) - R(Q_{\boldsymbol{\phi}})| \leq \epsilon\,, \tag{7}$$

then, with the same probability, we will have simultaneously $\forall(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, that

$$|F_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y}) - F(\mathbf{w}, x, \mathbf{y})| \leq \epsilon \,,$$

and, consequently, the lemma will be proved.

To prove that we satisfy Equation (7) with probability $\geq 1 - \delta/2$ simultaneously for all $\boldsymbol{\phi}$ of unit L_2 norm, let us adapt some elements of PAC-Bayes theory to our case. Note that we cannot use the usual PAC-Bayes bounds, such as those proposed by [13] because, in our case, the loss $\langle \mathbf{w}_T, \boldsymbol{\zeta} \rangle$ of each individual "predictor" $\boldsymbol{\zeta}$ is unbounded.

The distribution $Q_{\boldsymbol{\phi}}$ defined above constitutes the posterior distribution. For the prior P, let us use an isotropic Gaussian with variance σ^2 centered at the origin. Hence $P = Q_{\mathbf{0}}$. In that case we have

$$\mathrm{KL}(Q_{\boldsymbol{\phi}} \| P) = \frac{\|\boldsymbol{\phi}\|^2}{2\sigma^2} \;=\; \frac{1}{2\sigma^2} \,.$$

Given a tree sample $\mathcal{T}$ of n spanning trees, let

$$\Delta\mathbf{w} \;\stackrel{\text{def}}{=}\; \frac{1}{n}\sum_{k=1}^{n} \mathbf{w}_{T_k} - \mathop{\mathbf{E}}_{T\sim\mathcal{U}(G)} \mathbf{w}_T \,,$$

and consider the Gaussian quadrature

$$\begin{aligned}
\mathcal{I} \;&\stackrel{\text{def}}{=}\; \mathop{\mathbf{E}}_{\boldsymbol{\zeta}\sim P} e^{\sqrt{n}|\langle \Delta\mathbf{w}, \boldsymbol{\zeta}\rangle|} \\
&=\; e^{\frac{1}{2}n\sigma^2\|\Delta\mathbf{w}\|^2}\left(1 + \mathrm{Erf}\left[\sqrt{\frac{n}{2}}\|\Delta\mathbf{w}\|\sigma\right]\right) \\
&\leq\; 2e^{\frac{1}{2}n\sigma^2\|\Delta\mathbf{w}\|^2} \,.
\end{aligned}$$

We can then use this result for $\mathcal{I}$ to upper bound the Laplace transform $\mathcal{L}$ in the following way.

$$\begin{aligned}
\mathcal{L} \;&\stackrel{\text{def}}{=}\; \mathop{\mathbf{E}}_{\mathcal{T}\sim\mathcal{U}(G)^n} \mathop{\mathbf{E}}_{\boldsymbol{\zeta}\sim P} e^{\sqrt{n}|\langle \Delta\mathbf{w}, \boldsymbol{\zeta}\rangle|} \\
&\leq\; 2 \mathop{\mathbf{E}}_{\mathcal{T}\sim\mathcal{U}(G)^n} e^{\frac{1}{2}n\sigma^2\|\Delta\mathbf{w}\|^2} \\
&=\; 2 \mathop{\mathbf{E}}_{\mathcal{T}\sim\mathcal{U}(G)^n} e^{\frac{1}{2}n\sigma^2 \sum_{(i,j)\in G}\|(\Delta\mathbf{w})_{i,j}\|^2} \,.
\end{aligned}$$

Since

$$\mathop{\mathbf{E}}_{T\sim\mathcal{U}(G)} \mathbf{w}_T = \frac{2}{\ell}\mathbf{w} \,,$$

we can write

$$\|(\Delta\mathbf{w})_{i,j}\|^2 \;=\; \left\| \frac{1}{n}\sum_{k=1}^{n} (\mathbf{w}_{T_k})_{i,j} - \frac{2}{\ell}\mathbf{w}_{i,j} \right\|^2 \,.$$

Note that for each $(i,j) \in G$, any sample $\mathcal{T}$, and each $T_k \in \mathcal{T}$, we can write

$$(\mathbf{w}_{T_k})_{i,j} \;=\; \mathbf{w}_{i,j} Z^k_{i,j} \,.$$

where $Z^k_{i,j} = 1$ if $(i,j) \in T_k$ and $Z^k_{i,j} = 0$ if $(i,j) \notin T_k$. Hence, we have

$$\|(\Delta\mathbf{w})_{i,j}\|^2 \;=\; \|\mathbf{w}_{i,j}\|^2 \left(\frac{1}{n}\sum_{k=1}^{n} Z^k_{i,j} - \frac{2}{\ell}\right)^2 \,.$$

Hence, for $\sigma^2 \leq 4$ and $p \stackrel{\text{def}}{=} 2/\ell$, we have

$$\begin{aligned}
\mathcal{L} \;&\leq\; 2 \mathop{\mathbf{E}}_{\mathcal{T}\sim\mathcal{U}(G)^n} e^{\frac{1}{2}n\sigma^2 \sum_{(i,j)\in G}\|\mathbf{w}_{i,j}\|^2\left(\frac{1}{n}\sum_{k=1}^{n} Z^k_{i,j} - \frac{2}{\ell}\right)^2} \\
&\leq\; 2 \mathop{\mathbf{E}}_{\mathcal{T}\sim\mathcal{U}(G)^n} e^{2n \sum_{(i,j)\in G}\|\mathbf{w}_{i,j}\|^2\left(\frac{1}{n}\sum_{k=1}^{n} Z^k_{i,j} - p\right)^2} \\
&\leq\; 2 \sum_{(i,j)\in G} \|\mathbf{w}_{i,j}\|^2 \mathop{\mathbf{E}}_{\mathcal{T}\sim\mathcal{U}(G)^n} e^{2n\left(\frac{1}{n}\sum_{k=1}^{n} Z^k_{i,j} - p\right)^2} \,,
\end{aligned}$$

where the last inequality is obtained by using $\sum_{(i,j)\in G} \|\mathbf{w}_{i,j}\|^2 = 1$ and by using Jensen's inequality on the convexity of the exponential.

Now, for any $(q,p) \in [0,1]^2$, let

$$\mathrm{kl}(q\|p) \stackrel{\text{def}}{=} q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p} \,.$$

Then, by using $2(q-p)^2 \leq \mathrm{kl}(q\|p)$ (Pinsker's inequality), we have for $n \geq 8$,

$$\mathcal{L} \;\leq\; 2 \sum_{(i,j)\in G} \|\mathbf{w}_{i,j}\|^2 \mathop{\mathbf{E}}_{\mathcal{T}\sim\mathcal{U}(G)^n} e^{n\mathrm{kl}\left(\frac{1}{n}\sum_{k=1}^n Z_{i,j}^k \|p\right)} \;\leq\; 4\sqrt{n}\,,$$

where the last inequality follows from Maurer's lemma [14] applied, for any fixed $(i,j) \in G$, to the collection of n independent Bernoulli variables $Z_{i,j}^k$ of probability p.

The rest of the proof follows directly from standard PAC-Bayes theory [15, 13], which, for completeness, we briefly outline here.

Since

$$\mathop{\mathbf{E}}_{\boldsymbol{\zeta}\sim P} e^{\sqrt{n}|\langle \Delta\mathbf{w}, \boldsymbol{\zeta}\rangle|}$$

is a non negative random variable, Markov's inequality implies that with probability $> 1-\delta/2$ over the random draws of $\mathcal{T}$, we have

$$\ln \mathop{\mathbf{E}}_{\boldsymbol{\zeta}\sim P} e^{\sqrt{n}|\langle \Delta\mathbf{w}, \boldsymbol{\zeta}\rangle|} \;\leq \ln \frac{8\sqrt{n}}{\delta}\,.$$

By the change of measure inequality, we have with probability $> 1-\delta/2$ over the random draws of $\mathcal{T}$, simultaneously for all $\boldsymbol{\phi}$,

$$\sqrt{n} \mathop{\mathbf{E}}_{\boldsymbol{\zeta}\sim Q_{\boldsymbol{\phi}}} |\langle \Delta\mathbf{w}, \boldsymbol{\zeta}\rangle| \;\leq\; \mathrm{KL}\left(Q_{\boldsymbol{\phi}}\|P\right) + \ln \frac{8\sqrt{n}}{\delta}\,.$$

Hence, by using Jensen's inequality on the convex absolute value function, we have with probability $> 1-\delta/2$ over the random draws of $\mathcal{T}$, simultaneously for all $\boldsymbol{\phi}$,

$$|\langle \Delta\mathbf{w}, \boldsymbol{\phi}\rangle| \;\leq\; \frac{1}{\sqrt{n}} \left[\mathrm{KL}\left(Q_{\boldsymbol{\phi}}\|P\right) + \ln \frac{8\sqrt{n}}{\delta}\right]\,.$$

Note that we have $\mathrm{KL}(Q_{\boldsymbol{\phi}}\|P) = 1/8$ for $\sigma^2 = 4$ (which is the value we shall use). Also note that the left hand side of this equation equals to $|R_{\mathcal{T}}(Q_{\boldsymbol{\phi}}) - R(Q_{\boldsymbol{\phi}})|$. In that case, we satisfy Equation (7) with probability $1-\delta/2$ simultaneously for all $\boldsymbol{\phi}$ of unit L_2 norm whenever we satisfy

$$\frac{\ell}{2\sqrt{n}} \left[\frac{1}{8} + \ln \frac{8\sqrt{n}}{\delta}\right] \;\leq\; \epsilon\,,$$

which is the condition on n given by the theorem. □

Theorem 4

Consider any unit L_2 norm predictor $\mathbf{w}$ for the complete graph G, acting on a normalized joint feature space, achieving a margin of $\Gamma(\mathbf{w}, x, \mathbf{y})$ for each $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$. Then for any $\epsilon > 0$, and any n satisfying Lemma 3, for any $\delta \in (0,1]$, with probability of at least $1-\delta$ over all samples $\mathcal{T}$ generated according to $\mathcal{U}(G)^n$, there exists a unit L_2 norm conical combination $(\mathcal{W}, \mathbf{q})$ on $\mathcal{T}$ such that, simultaneously $\forall (x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, we have

$$\Gamma_{\mathcal{T}}(\mathcal{W}, \mathbf{q}, x, \mathbf{y}) \geq \frac{1}{\sqrt{1+\epsilon}} \left[\Gamma(\mathbf{w}, x, \mathbf{y}) - 2\epsilon\right]\,.$$

Proof. For any $\mathcal{T}$, consider a conical combination $(\mathcal{W}, \mathbf{q})$ where each $\hat{\mathbf{w}}_{T_i} \in \mathcal{W}$ is obtained by projecting $\mathbf{w}$ on T_i and normalizing to unit L_2 norm and where

$$q_i = \frac{a_{T_i}}{\sqrt{\sum_{i=1}^n a_{T_i}^2}}\,.$$

Then, from equations (3) and (4), and from the definition of $\Gamma_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y})$, we find that for all $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, we have

$$\Gamma_{\mathcal{T}}(\mathcal{W}, \mathbf{q}, x, \mathbf{y}) = \sqrt{\frac{n}{\sum_{i=1}^n a_{T_i}^2}}\Gamma_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y})\,.$$

Now, by using Hoeffding's inequality, it is straightforward to show that for any $\delta \in (0, 1]$, we have

$$\Pr_{\mathcal{T}\sim\mathcal{U}(G)^n}\left(\frac{1}{n}\sum_{i=1}^n a_{T_i}^2 \leq 1+\epsilon\right) \geq 1-\delta/2\,.$$

whenever $n \geq \frac{\ell^2}{8\epsilon}\ln\left(\frac{2}{\delta}\right)$. Since n satisfies the condition of Lemma 3, we see that it also satisfies this condition whenever $\epsilon \leq 1/2$. Hence, with probability of at least $1-\delta/2$, we have

$$\sum_{i=1}^n a_{T_i}^2 \leq n(1+\epsilon)\,.$$

Moreover Lemma 2 and Lemma 3 imply that, with probability of at least $1-\delta/2$, we have simultaneously for all $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$,

$$\Gamma_{\mathcal{T}}(\mathbf{w}, x, \mathbf{y}) \geq \Gamma(\mathbf{w}, x, \mathbf{y}) - 2\epsilon\,.$$

Hence, from the union bound, with probability of at least $1-\delta$, simultaneously $\forall (x, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, we have

$$\Gamma_{\mathcal{T}}(\mathcal{W}, \mathbf{q}, x, \mathbf{y}) \geq \frac{1}{\sqrt{1+\epsilon}}\left[\Gamma(\mathbf{w}, x, \mathbf{y}) - 2\epsilon\right]\,.$$

□

Derivation of the Primal L_2-norm Random Tree Approximation

If we introduce the usual slack variables $\xi_i \stackrel{\text{def}}{=} \gamma \cdot \mathcal{L}^\gamma(\Gamma_{\mathcal{T}}(\mathcal{W}, \mathbf{q}, x_i, \mathbf{y}_i))$, Theorem 5 suggests that we should minimize $\frac{1}{\gamma}\sum_{k=1}^m \xi_k$ for some fixed margin value $\gamma > 0$. Rather than performing this task for several values of γ, we can, equivalently, solve the following optimization problem for several values of $C > 0$.

$$\begin{aligned} \min_{\boldsymbol{\xi},\gamma,\mathbf{q},\mathcal{W}} \quad & \frac{1}{2\gamma^2} + \frac{C}{\gamma}\sum_{k=1}^m \xi_k \\ \mathbf{s.t.}: \quad & \Gamma_{\mathcal{T}}(\mathcal{W}, \mathbf{q}, x_k, \mathbf{y}_k) \geq \gamma - \xi_k,\ \xi_k \geq 0,\ \forall\, k \in \{1, \ldots, m\}\,, \\ & \sum_{i=1}^n q_i^2 = 1,\ q_i \geq 0,\ \|\mathbf{w}_{T_i}\|^2 = 1,\ \forall\, i \in \{1, \ldots, n\}\,. \end{aligned} \tag{8}$$

If we now use instead $\zeta_k \stackrel{\text{def}}{=} \xi_k/\gamma$, and $\mathbf{v}_{T_i} \stackrel{\text{def}}{=} q_i\mathbf{w}_{T_i}/\gamma$, we then have $\sum_{i=1}^n \|\mathbf{v}_{T_i}\|^2 = 1/\gamma^2$ (under the constraints of problem (8)). If $\mathcal{V} \stackrel{\text{def}}{=} \{\mathbf{v}_{T_1}, \ldots, \mathbf{v}_{T_n}\}$, optimization problem (8) is then equivalent to

$$\begin{aligned} \min_{\boldsymbol{\zeta},\mathcal{V}} \quad & \frac{1}{2}\sum_{i=1}^n \|\mathbf{v}_{T_i}\|^2 + C\sum_{k=1}^m \zeta_k \\ \mathbf{s.t.}: \quad & \Gamma_{\mathcal{T}}(\mathcal{V}, \mathbf{1}, x_k, \mathbf{y}_k) \geq 1 - \zeta_k,\ \zeta_k \geq 0,\ \forall\, k \in \{1, \ldots, m\}\,. \end{aligned} \tag{9}$$

Note that, following our definitions, we now have

$$\Gamma_{\mathcal{T}}(\mathcal{V}, \mathbf{1}, x, \mathbf{y}) \quad = \quad \frac{1}{\sqrt{n}}\sum_{i=1}^n \langle \mathbf{v}_{T_i}, \hat{\boldsymbol{\phi}}_{T_i}(x, \mathbf{y})\rangle - \max_{\mathbf{y}'\neq\mathbf{y}} \frac{1}{\sqrt{n}}\sum_{i=1}^n \langle \mathbf{v}_{T_i}, \hat{\boldsymbol{\phi}}_{T_i}(x, \mathbf{y}')\rangle\,.$$

We then obtain the optimization problem of Property 6 with the change of variables $\mathbf{w}_{T_i} \leftarrow \mathbf{v}_{T_i}/\sqrt{n}$, $\xi_k \leftarrow \zeta_k$, and $C \leftarrow C/\sqrt{n}$.

Lemma 7

Let $\mathbf{y}_K^\star = \operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}_{\mathcal{T},K}} F_{\mathcal{T}}(x,\mathbf{y})$ *be the highest scoring multilabel in* $\mathcal{Y}_{\mathcal{T},K}$*. Suppose that*

$$F_{\mathcal{T}}(x,\mathbf{y}_K^\star) \geq \frac{1}{n}\sum_{i=1}^{n} F_{\mathbf{w}_{T_i}}(x,\mathbf{y}_{T_i,K}) \stackrel{def}{=} \theta_x(K)$$

It follows that $F_{\mathcal{T}}(x,\mathbf{y}_K^\star) = \max_{\mathbf{y}\in\mathcal{Y}} F_{\mathcal{T}}(x,\mathbf{y})$*.*

Proof. Consider a multilabel $\mathbf{y}^\dagger \notin \mathcal{Y}_{\mathcal{T},K}$. It follows that for all T_i we have

$$F_{\mathbf{w}_{T_i}}(x,\mathbf{y}^\dagger) \leq F_{\mathbf{w}_{T_i}}(x,\mathbf{y}_{T_i,K}).$$

Hence,

$$F_{\mathcal{T}}(x,\mathbf{y}^\dagger) \;=\; \frac{1}{n}\sum_{i=1}^{n} F_{\mathbf{w}_{T_i}}(x,\mathbf{y}^\dagger) \;\leq\; \frac{1}{n}\sum_{i=1}^{n} F_{\mathbf{w}_{T_i}}(x,\mathbf{y}_{T_i,K}) \;\leq\; F_{\mathcal{T}}(x,\mathbf{y}_K^\star),$$

as required. □

Lemma 8

Let the scoring function F_T *of each spanning tree of* G *be a* σ_T*-sub-Gaussian random variable under the uniform distribution of labels;* i.e.*, for each* T *on* G*, there exists* $\sigma_T > 0$ *such that for any* $\lambda > 0$ *we have*

$$\mathop{\mathbf{E}}_{\mathbf{y}\sim\mathcal{U}(\mathcal{Y})} e^{\lambda(F_T(\mathbf{y})-\mu_T)} \;\leq\; e^{\frac{\lambda^2}{2}\sigma_T^2}\,.$$

Let $\sigma^2 \stackrel{def}{=} \mathop{\mathbf{E}}_{T\sim\mathcal{U}(G)} \sigma_T^2$*, and let*

$$\alpha \stackrel{def}{=} \Pr_{T\sim\mathcal{U}(G)}\Big(\mu_T \leq \mu \;\wedge\; F_T(\hat{\mathbf{y}}) \geq F(\hat{\mathbf{y}}) \;\wedge\; \sigma_T^2 \leq \sigma^2\Big)\,.$$

Then

$$\Pr_{\mathcal{T}\sim\mathcal{U}(G)^n}\left(\exists T\in\mathcal{T}\colon\; \rho_T(\hat{\mathbf{y}}) \leq e^{-\frac{1}{2}\frac{(F(\hat{\mathbf{y}})-\mu)^2}{\sigma^2}}\right) \geq 1-(1-\alpha)^n\,.$$

Proof. From the definition of $\rho(\hat{\mathbf{y}})$ and for any $\lambda > 0$, we have

$$\begin{aligned}
\rho_T(\mathbf{y}^*) &= \Pr_{\mathbf{y}\sim\mathcal{U}(\mathcal{Y})}\;(F_T(\mathbf{y}) \geq F_T(\hat{\mathbf{y}})) \\
&= \Pr_{\mathbf{y}\sim\mathcal{U}(\mathcal{Y})}\;(F_T(\mathbf{y})-\mu_T \geq F_T(\hat{\mathbf{y}})-\mu_T) \\
&= \Pr_{\mathbf{y}\sim\mathcal{U}(\mathcal{Y})}\;\Big(e^{\lambda(F_T(\mathbf{y})-\mu_T)} \geq e^{\lambda(F_T(\hat{\mathbf{y}})-\mu_T)}\Big) \\
&\leq e^{-\lambda(F_T(\hat{\mathbf{y}})-\mu_T)} \mathop{\mathbf{E}}_{\mathbf{y}\sim\mathcal{U}(\mathcal{Y})} e^{\lambda(F_T(\mathbf{y})-\mu_T)} \qquad (10)\\
&\leq e^{-\lambda(F_T(\hat{\mathbf{y}})-\mu_T)} e^{\frac{\lambda^2}{2}\sigma_T^2}\,, \qquad (11)
\end{aligned}$$

where we have used Markov's inequality for line (10) and the fact that F_T is a σ_T-sub-Gaussian variable for line (11). Hence, from this equation and from the definition of α, we have that

$$\Pr_{T\sim\mathcal{U}(G)}\Big(\rho_T(\hat{\mathbf{y}}) \leq e^{-\lambda(F_T(\hat{\mathbf{y}})-\mu_T)} e^{\frac{\lambda^2}{2}\sigma_T^2} \leq e^{-\lambda(F(\hat{\mathbf{y}})-\mu)} e^{\frac{\lambda^2}{2}\sigma^2}\Big) \;\geq\; \alpha\,.$$

Hence,

$$\Pr_{\mathcal{T}\sim\mathcal{U}(G)^n}\Big(\forall T\in\mathcal{T}\colon\; \rho_T(\hat{\mathbf{y}}) > e^{-\lambda(F(\hat{\mathbf{y}})-\mu)} e^{\frac{\lambda^2}{2}\sigma^2}\Big) \;\leq\; (1-\alpha)^n\,,$$

which is equivalent to the statement of the lemma when we choose $\lambda = [F(\hat{\mathbf{y}})-\mu]/\sigma^2$. □

The K-best Inference Algorithm

Algorithm 1 depicts the K-best inference algorithm for the ensemble of rooted spanning trees. The algorithm takes as input the collection of spanning trees $T_i \in \mathcal{T}$, the edge labeling scores

$$F_{E_{\mathcal{T}}} = \{F_{T_i,v,v'}(y_v, y_{v'})\}_{(v,v')\in E_i, y_v \in \mathcal{Y}_v, y_{v'} \in \mathcal{Y}_{v'}, T_i \in \mathcal{T}}$$

for fixed x_k and $\mathbf{w}$, the length of K-best list, and optionally (for training) also the true multilabel $\mathbf{y}_k$ for x_k.

As a rooted tree implicitly orients the edges, for convenience we denote the edges as directed $v \to pa(v)$, where $pa(v)$ denotes the parent (i.e. the adjacent node on the path towards the root) of v. By $ch(v)$ we denote the set of children of v. Moreover, we denote the subtree of T_i rooted at a node v as T_v and by $T_{v' \to v}$ the subtree consisting of $T_{v'}$ plus the edge $v' \to v$ and the node v.

The algorithm performs a dynamic programming over each tree in turn, extracting the K-best list of multilabels and their scores, and aggregates the results of the trees, retrieving the highest scoring multilabel of the ensemble, the worst violating multilabel and the threshold score of the K-best lists.

The dynamic programming is based on traversing the tree in post-order, so that children of the node are always processed before the parent. The algorithm maintains sorted K best lists of candidate labelings of the subtrees T_v and $T_{v' \to v}$, using the following data structures:

- Score matrix P_v, where element $P_v(y, r)$ records the score of the r'th best multilabel of the subtree T_v when node v is labeled as y.
- Pointer matrix C_v, where element $C_v(y, r)$ keeps track of the ranks of the child nodes $v' \in ch(v)$ in the message matrix $M_{v' \to v}$ that contributes to the score $P_v(y, r)$.
- Message matrix $M_{v \to pa(v)}$, where element $M_{v \to pa(v)}(y', r)$ records the score of r'th best multilabel of the subtree $T_{v \to pa(v)}$ when the label of $pa(v)$ is y'.
- Configuration matrix $C_{v \to pa(v)}$, where element $C_{v \to pa(v)}(y', r)$ traces the label and rank (y, r) of child v that achieves $M_{v \to pa(v)}(y', r)$.

The processing of a node v entails the following steps. First, the K-best lists of the children of the node stored in $M_{v' \to v}$ are merged in amortized $\Theta(K)$ time per child node. This is achieved by processing two child lists in tandem starting from the top of the lists and in each step picking the best pair of items to merge. This process results in the score matrix P_v and the pointer matrix C_v.

Second, the K-best lists of $T_{v \to pa(v)}$ corresponding to all possible labels y' of $pa(v)$ are formed. This is achieved by keeping the label of the head of the edge $v \to pa(v)$ fixed, and picking the best combination of labeling the tail of the edge and selecting a multilabel of T_v consistent with that label. This process results in the matrices $M_{v \to pa(v)}$ and $C_{v \to pa(v)}$. Also this step can be performed in $\Theta(K)$ time.

The iteration ends when the root v_{root} has updated its score $P_{v_{root}}$. Finally, the multilabels in form $\mathcal{Y}_{T_i,K}$ are traced using the pointers stored in C_v and $C_{v \to pa(v)}$. The time complexity for a single tree is $\Theta(K\ell)$, and repeating the process for n trees gives total time complexity of $\Theta(nK\ell)$.

Master algorithm for training the model

The master algorithm (Algorithm 2) iterates over each training example until convergence. The processing of each training example proceeds by identifying the K worst violators of each tree together with the threhold score $\theta_i = \theta_{x_i}$ (line 5), determining the worst ensemble violator from among them (line 6) and updating each tree by the worst ensemble violator (line 8). During the early stages of the algorithm, it is not essential to identify the worst violator. We therefore propose that initially $K = 2$, and the iterations continue until no violators are identified (line 7). We then increment K and continue until the condition (line 10-12) given by Lemma 7 is satisfied so that we are assured of having converged to the global optimum.

Algorithm 1 Algorithm to obtain top K multilabels on a collection of spanning trees.

$FindKBest(\mathcal{T}, F_{E_\mathcal{T}}, K, \mathbf{y}_i)$

Input: Collection of rooted spanning trees $T_i = (E_i, V_i)$,
edge labeling scores $F_{E_\mathcal{T}} = \{F_{T,v,v'}(y_v, y_{v'})\}$

Output: The best scoring multilabel $\mathbf{y}^*$, worst violator $\bar{\mathbf{y}}$, threshold θ_i

for $T_i \in \mathcal{T}$ **do**
  Initialize $P_v, C_v, M_{v\to pa(v)}, C_{v\to pa(v)}, \forall v \in V_i$
  I = nodes indices in post-order of the tree T_i
  for $j = 1 : |I|$ **do**
    $v = v_{I(j)}$
    % Collect and merge K-best lists of children
    if $ch(v) \neq \emptyset$ **then**
      $P_v(y) = P_v(y) + \underset{r_v, v' \in ch(v)}{\textbf{kmax}} \left(\sum_{v' \in ch(v)} (M_{v'\to v}(y, r_v)) \right)$
      $C_v(y) = P_v(y) + \underset{r_v, v' \in ch(v)}{\textbf{argkmax}} \left(\sum_{v' \in ch(v)} (M_{v'\to v}(y, r_v)) \right)$
    end if
    % Form the K-best list of $T_{v\to pa(v)}$
    $M_{v\to pa(v)}(y_{pa(v)}) = \underset{y,r}{\textbf{kmax}} \left(P_v(y, r) + F_{T,v\to pa(v)}(y_v, y_{pa(v)}) \right)$
    $C_{v\to pa(v)}(y_{pa(v)}) = \underset{u_v, r}{\textbf{argkmax}} \left(P_v(u_v, r) + F_{T,v\to pa(v)}(u_v, y_{pa(v)}) \right)$
  end for
  Trace back with C_v and $C_{v\to pa(v)}$ to get $\mathcal{Y}_{T_i,K}$.
end for

$$\mathcal{Y}_{\mathcal{T},K} = \bigcup_{T_i \in \mathcal{T}} \mathcal{Y}_{T_i,K}$$

$$\mathbf{y}^* = \underset{\mathbf{y} \in \mathcal{Y}_{\mathcal{T},K}}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{\substack{(v,v')= \\ e \in E_i}} F_{T_i,v,v'}(y_v, y_{v'})$$

$$\bar{\mathbf{y}} = \underset{\mathbf{y} \in \mathcal{Y}_{\mathcal{T},K} \setminus \mathbf{y}_i}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{\substack{(v,v')= \\ e \in E_i}} F_{T_i,v,v'}(y_v, y_{v'})$$

$$\theta_i = \sum_{i=1}^{n} \sum_{\substack{(v,v')= \\ e \in E_i}} F_{T_i,v,v'}(y_{T_i,K,v}, y_{T_i,k,v'})$$

Algorithm 2 Master algorithm.

Input: Training sample $\{(x_k, \mathbf{y}_k)\}_{k=1}^m$, collection of spanning trees $\mathcal{T}$, minimum violation γ_0
Output: Scoring function $F_{\mathcal{T}}$

```
K_k = 2, ∀k ∈ {1, ···, m}; w_{T_i} = 0, ∀ T_i ∈ T; converged = false
while not(converged) do
    converged = true
    for k = {1, ..., m} do
        S_T = {S_{T_i,e}(k, u_e) | S_{T_i,e}(k, u_e) = ⟨w_{T_i,e}, φ_{T_i,e}(x_k, u_e)⟩, ∀(e ∈ E_i, T_i ∈ T, u_e ∈ Y_v × Y_{v'})}
        [y*, ȳ, θ_i] = FindKBest(T, S_T, K_i, y_i)
        if F_T(x_i, ȳ) − F_T(x_i, y_i) ≥ γ_0 then
            {w_{T_i}}_{T_i∈T} = updateTrees({w_{T_i}}_{T_i∈T}, x_i, ȳ)
            converged = false
        else
            if θ_i > F_T(x_i, ȳ) then
                K_i = min(2K_i, |Y|)
                converged = false
            end if
        end if
    end for
end while
```

Derivation of the Marginal Dual

Definition 6. Primal L_2-norm Random Tree Approximation

$$\begin{aligned}
\min_{\mathbf{w}_{T_i}, \xi_k} \quad & \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{w}_{T_i}||_2^2 + C\sum_{k=1}^{m}\xi_k \\
\text{s.t.} \quad & \sum_{i=1}^{n}\langle \mathbf{w}_{T_i}, \hat{\boldsymbol{\phi}}_{T_i}(x_k, \mathbf{y}_k)\rangle - \max_{\mathbf{y}\neq\mathbf{y}_k}\sum_{i=1}^{n}\langle \mathbf{w}_{T_i}, \hat{\boldsymbol{\phi}}_{T_i}(x_k, \mathbf{y})\rangle \geq 1 - \xi_k \\
& \xi_k \geq 0, \forall\, k \in \{1, \ldots, m\},
\end{aligned}$$

where $\{\mathbf{w}_{T_i} | T_i \in \mathcal{T}\}$ are the feature weights to be learned on each tree, ξ_k is the margin slack allocated for each example x_k, and C is the slack parameter that controls the amount of regularization in the model. This primal form has the interpretation of maximizing the joint margins from individual trees between (correct) training examples and all the other (incorrect) examples.

The Lagrangian of the primal form (Definition 6) is

$$\begin{aligned}
\mathcal{L}(\mathbf{w}_{T_i}, \xi, \boldsymbol{\alpha}, \boldsymbol{\beta}) \; = & \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{w}_{T_i}||_2^2 + C\sum_{k=1}^{m}\xi_k - \sum_{k=1}^{m}\beta_k\xi_k \\
& - \sum_{k=1}^{m}\sum_{\mathbf{y}\neq\mathbf{y}_k}\alpha_{k,\mathbf{y}}\left(\sum_{i=1}^{n}\langle \mathbf{w}_{T_i}, \Delta\hat{\boldsymbol{\phi}}_{T_i}(x_k, \mathbf{y}_k)\rangle - 1 + \xi_k\right),
\end{aligned}$$

where α_k and β_k are Lagrangian multipliers that correspond to the constraints of the primal form, and $\Delta\hat{\boldsymbol{\phi}}_{T_i}(x_k, \mathbf{y}_k) = \hat{\boldsymbol{\phi}}_{T_i}(x_k, \mathbf{y}_k) - \hat{\boldsymbol{\phi}}_{T_i}(x_k, \mathbf{y})$. Note that given a training example-label pair $(x_k, \mathbf{y}_k)$ there are exponential number of $\alpha_{k,\mathbf{y}}$ one for each constraint defined by incorrect example-label pair $(x_k, \mathbf{y})$.

Setting the gradient of Lagrangian with respect to primal variables to zero, we obtain the following equalities:

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mathbf{w}_{T_i}} &= \mathbf{w}_{T_i} - \sum_{k=1}^{m}\sum_{\mathbf{y}\neq\mathbf{y}_k}\alpha_{k,\mathbf{y}}\Delta\hat{\boldsymbol{\phi}}_{T_i}(x_k, \mathbf{y}_k) = 0, \\
\frac{\partial\mathcal{L}}{\partial\xi_k} &= C - \sum_{\mathbf{y}\neq\mathbf{y}_k}\alpha_{k,\mathbf{y}} - \beta_k = 0,
\end{aligned}$$

which give the following dual optimization problem.

Definition 10. Dual L_2-norm Random Tree Approximation

$$\begin{aligned} \max_{\boldsymbol{\alpha}\geq 0} \quad & \boldsymbol{\alpha}^\intercal \mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}^\intercal \left(\sum_{i=1}^{n} K_{T_i}\right)\boldsymbol{\alpha} \\ \textbf{s.t.} \quad & \sum_{\mathbf{y}\neq\mathbf{y}_k} \alpha_{k,\mathbf{y}} \leq C, \forall\, k \in \{1,\dots,m\}, \end{aligned}$$

where $\boldsymbol{\alpha} = (\alpha_{k,\mathbf{y}})_{k,\mathbf{y}}$ is the vector of dual variables. The joint kernel

$$\begin{aligned} K_{T_i}(x_k,\mathbf{y};x_{k'},\mathbf{y}') &= \langle \hat{\boldsymbol{\phi}}_{T_i}(x_k,\mathbf{y}_k) - \hat{\boldsymbol{\phi}}_{T_i}(x_k,\mathbf{y}), \hat{\boldsymbol{\phi}}_{T_i}(x_{k'},\mathbf{y}_{k'}) - \hat{\boldsymbol{\phi}}_{T_i}(x_{k'},\mathbf{y}')\rangle \\ &= \langle \varphi(x_k),\varphi(x_{k'})\rangle_\varphi \cdot \langle \psi_{T_i}(\mathbf{y}_k) - \psi_{T_i}(\mathbf{y}), \psi_{T_i}(\mathbf{y}_{k'}) - \psi_{T_i}(\mathbf{y}')\rangle_\psi \\ &= K^\varphi(x_k,x_{k'}) \cdot \Big(K^\psi_{T_i}(\mathbf{y}_k,\mathbf{y}_{k'}) - K^\psi_{T_i}(\mathbf{y}_k,\mathbf{y}') - K^\psi_{T_i}(\mathbf{y},\mathbf{y}_{k'}) + K^\psi_{T_i}(\mathbf{y},\mathbf{y}')\Big) \\ &= K^\varphi(x_k,x_{k'}) \cdot K^{\Delta\psi}_{T_i}(\mathbf{y}_k,\mathbf{y};\mathbf{y}_{k'},\mathbf{y}') \end{aligned}$$

is composed by input kernel K^φ and output kernel $K^\psi_{T_i}$ defined by

$$\begin{aligned} K^\varphi(x_k,x_{k'}) &= \langle \varphi(x_k),\varphi(x_{k'})\rangle_\varphi \\ K^{\Delta\psi}_{T_i}(\mathbf{y}_k,\mathbf{y};\mathbf{y}_{k'},\mathbf{y}') &= K^\psi_{T_i}(\mathbf{y}_k,\mathbf{y}_{k'}) - K^\psi_{T_i}(\mathbf{y}_k,\mathbf{y}') - K^\psi_{T_i}(\mathbf{y}_{k'},\mathbf{y}) + K^\psi_{T_i}(\mathbf{y},\mathbf{y}'). \end{aligned}$$

To take advantage of the spanning tree structure in solving the problem, we further factorize the dual (Definition 10) according to the output structure [9, 16]. by defining a marginal dual variable μ as

$$\mu(k,e,\mathbf{u}_e) = \sum_{\mathbf{y}\neq\mathbf{y}_k} \mathbf{1}_{\{\psi(\mathbf{y})=\mathbf{u}_e\}}\boldsymbol{\alpha}_{k,\mathbf{y}},$$

where $e=(j,j')\in E$ is an edge in the output graph and $\mathbf{u}_e \in \mathcal{Y}\times\mathcal{Y}'$ is a possible label of edge e. As each marginal dual variable $\mu(k,e,\mathbf{u}_e)$ is the sum of a collection of dual variables $\alpha_{k,\mathbf{y}}$ that has consistent label $(u_j,u_{j'}) = \mathbf{u}_e$, we have the following

$$\sum_{\mathbf{u}_e} \mu(k,e,\mathbf{u}_e) = \sum_{\mathbf{y}\neq\mathbf{y}_k} \boldsymbol{\alpha}_{k,\mathbf{y}} \tag{12}$$

for an arbitrary edge e, independently of the structure of the trees.

The linear part of the objective (Definition 10) can be stated in term of $\boldsymbol{\mu}$ for an arbitrary collection of trees as

$$\boldsymbol{\alpha}^\intercal\mathbf{1} = \sum_{k=1}^{m}\sum_{\mathbf{y}\neq\mathbf{y}_k}\alpha_{k,\mathbf{y}} = \frac{1}{|E_\mathcal{T}|}\sum_{k=1}^{m}\sum_{e\in E_\mathcal{T}}\sum_{\mathbf{u}_e}\mu(k,e,\mathbf{u}_e) = \frac{1}{|E_\mathcal{T}|}\sum_{e,k,\mathbf{u}_e}\mu(k,e,\mathbf{u}_e)\,,$$

where edge $e=(j,j')\in E_\mathcal{T}$ appearing in the collection of trees $\mathcal{T}$.

We observe that the label kernel of tree T_i, $K^\psi_{T_i}$, decomposes on the edges of the tree as

$$K^\psi_{T_i}(\mathbf{y},\mathbf{y}') = \langle \mathbf{y},\mathbf{y}'\rangle_\psi = \sum_{e\in E_i}\langle y_e,y'_e\rangle_\psi = \sum_{e\in E_i} K^{\psi,e}(y_e,y'_e).$$

Thus, the output kernel $K^{\Delta\psi}_{T_i}$ and the joint kernel K_{T_i} also decompose

$$\begin{aligned} K^{\Delta\psi}_{T_i}(\mathbf{y}_k,\mathbf{y};\mathbf{y}_{k'},\mathbf{y}') &= \Big(K^\psi_{T_i}(\mathbf{y}_k,\mathbf{y}_{k'}) - K^\psi_{T_i}(\mathbf{y}_k,\mathbf{y}') - K^\psi_{T_i}(\mathbf{y}_{k'},\mathbf{y}) + K^\psi_{T_i}(\mathbf{y},\mathbf{y}')\Big) \\ &= \sum_{e\in E_i}\Big(K^{\psi,e}_{T_i}(y_{ke},y_{k'e}) - K^{\psi,e}_{T_i}(y_{ke},y'_e) - K^{\psi,e}_{T_i}(y_e,y_{k'e}) + K^{\psi,e}_{T_i}(y_e,y'_e)\Big) \\ &= \sum_{e\in E_i} K^{\Delta\psi,e}_{T_i}(y_{ke},y_e;y_{k'e},y'_e), \\ K_{T_i}(x_k,\mathbf{y};x_{k'},\mathbf{y}') &= K^\psi(x_k,x_{k'}) \cdot K^{\Delta\psi}_{T_i}(\mathbf{y}_k,\mathbf{y};\mathbf{y}_{k'},\mathbf{y}') \\ &= K^\psi(x_k,x_{k'}) \cdot \sum_{e\in E_i} K^{\Delta\psi,e}(y_{ke},y_e;y_{k'e},y'_e) \\ &= \sum_{e\in E_i} K^e(x_k,y_e;x_{k'},y'_e). \end{aligned}$$

The sum of joint kernels of the trees can be expressed as

$$\begin{aligned}\sum_{i=1}^{n} K_{T_i}(x_k, \mathbf{y}; x_{k'}, \mathbf{y}') &= \sum_{i=1}^{n} \sum_{e \in E_i} K^e(x_k, y_e; x_{k'}, y'_e) \\ &= \sum_{e \in E_{\mathcal{T}}} \sum_{\substack{T_i \in \mathcal{T}: \\ e \in E_i}} K^e(x_k, y_e; x_{k'}, y'_e) \\ &= \sum_{e \in E_{\mathcal{T}}} N_{\mathcal{T}}(e) K^e(x_k, y_e; x_{k'}, y'_e)\end{aligned}$$

where $N_{\mathcal{T}}(e)$ denotes the number of occurrences of edge e in the collection of trees $\mathcal{T}$.

Taking advantage of the above decomposition and of the Equation (12) the quadratic part of the objective (Definition 10) can be stated in term of $\boldsymbol{\mu}$ as

$$\begin{aligned}& -\frac{1}{2}\boldsymbol{\alpha}^{\intercal}\left(\sum_{i=1}^{n} K_{T_i}\right)\boldsymbol{\alpha} \\ =& -\frac{1}{2}\boldsymbol{\alpha}^{\intercal}\left(\sum_{e \in E_{\mathcal{T}}} N_{\mathcal{T}}(e) K^e(x_k, \mathbf{y}; x_{k'}, \mathbf{y}')\right)\boldsymbol{\alpha} \\ =& -\frac{1}{2}\sum_{k,k'=1}^{m} \sum_{e \in E_{\mathcal{T}}} N_{\mathcal{T}}(e) \sum_{\substack{\mathbf{y} \neq \mathbf{y}_k \\ \mathbf{y}' \neq \mathbf{y}_{k'}}} \alpha(k, \mathbf{y}) K^e(x_k, y_e; x_{k'}, y'_e) \alpha(k', \mathbf{y}') \\ =& -\frac{1}{2}\sum_{k,k'=1}^{m} \sum_{e \in E_{\mathcal{T}}} N_{\mathcal{T}}(e) \sum_{\mathbf{u}_e, \mathbf{u}'_e} \sum_{\substack{\mathbf{y} \neq \mathbf{y}_k : y_e = \mathbf{u}_e \\ \mathbf{y}' \neq \mathbf{y}_{k'} : y'_e = \mathbf{u}'_e}} \alpha(k, \mathbf{y}) K^e(x_k, \mathbf{u}_e; x_{k'}, \mathbf{u}'_e) \alpha(k', \mathbf{y}') \\ =& -\frac{1}{2}\sum_{k,k'=1}^{m} \sum_{e \in E_{\mathcal{T}}} \frac{N_{\mathcal{T}}(e)}{|E_{\mathcal{T}}|^2} \sum_{\mathbf{u}_e, \mathbf{u}'_e} \mu(k, e, \mathbf{u}_e) K^e(x_k, \mathbf{u}_e; x_{k'}, \mathbf{u}'_e) \mu(k', e, \mathbf{u}'_e) \\ =& -\frac{1}{2}\sum_{\substack{e, k, \mathbf{u}_e, \\ k', \mathbf{u}'_e}} \mu(k, e, \mathbf{u}_e)\, K^e_{\mathcal{T}}(x_k, \mathbf{u}_e; x_{k'}, \mathbf{u}'_e)\, \mu(k', e, \mathbf{u}'_e),\end{aligned}$$

where $E_{\mathcal{T}}$ is the union of the sets of edges appearing in $\mathcal{T}$.

We then arrive at the following definition.

Definition 9. Marginalized Dual L_2-norm Random Tree Approximation

$$\max_{\boldsymbol{\mu} \in \mathcal{M}^m} \quad \frac{1}{|E_{\mathcal{T}}|} \sum_{e, k, \mathbf{u}_e} \mu(k, e, \mathbf{u}_e) - \frac{1}{2} \sum_{\substack{e, k, \mathbf{u}_e, \\ k', \mathbf{u}'_e}} \mu(k, e, \mathbf{u}_e) K^e_{\mathcal{T}}(x_k, \mathbf{u}_e; x'_k, \mathbf{u}'_e) \mu(k', e, \mathbf{u}'_e)\,,$$

where $\mathcal{M}^m$ is marginal dual feasible set defined as (*c.f.*, [9])

$$\mathcal{M}^m = \left\{ \boldsymbol{\mu} \,|\, \mu(k, e, \mathbf{u}_e) = \sum_{\mathbf{y} \neq \mathbf{y}_k} \mathbf{1}_{\{\mathbf{y}_{ke} = \mathbf{u}_e\}} \boldsymbol{\alpha}(k, \mathbf{y})\,,\ \forall (k, e, \mathbf{u}_e) \right\}.$$

The feasible set is composed of a Cartesian product of m identical polytopes $\mathcal{M}^m = \mathcal{M} \times \cdots \times \mathcal{M}$, one for each training example. Furthermore, each $\boldsymbol{\mu} \in \mathcal{M}$ corresponds to some dual variable $\boldsymbol{\alpha}$ in the original dual feasible set $\mathcal{A} = \{\boldsymbol{\alpha} | \alpha(k, \mathbf{y}) \geq 0, \sum_{\mathbf{y} \neq \mathbf{y}_i} \alpha(k, \mathbf{y}) \leq C, \forall k\}$.

Experimental Results

Table 2 provides the standard deviation results of the prediction performance results of Table 1 for each algorithm in terms of the microlabel and 0/1 error rates. Values are obtained by five fold cross-validation.

DATASET	MICROLABEL LOSS (%)					0/1 LOSS (%)				
	SVM	MTL	MMCRF	MAM	RTA	SVM	MTL	MMCRF	MAM	RTA
EMOTIONS	1.9	1.8	0.9	1.4	0.6	3.4	3.5	3.1	4.2	1.5
YEAST	0.7	0.5	0.6	0.5	0.6	2.8	1.0	1.5	0.4	1.2
SCENE	0.3	0.5	0.3	0.1	0.3	1.4	3.6	1.2	0.9	0.6
ENRON	0.2	0.2	0.2	0.2	0.2	0.3	0.4	2.8	2.3	0.9
CAL500	0.3	0.3	0.3	0.2	0.4	0.0	0.0	0.0	0.0	0.0
FINGERPRINT	0.3	0.6	0.6	0.3	0.6	0.7	0.0	0.5	0.6	1.3
NCI60	0.7	0.6	1.3	0.9	1.6	1.3	2.0	1.4	1.2	2.2
MEDICAL	0.0	0.1	0.1	0.1	0.2	2.1	2.3	3.3	2.5	3.6
CIRCLE10	0.9	0.7	0.3	0.4	0.3	3.8	3.4	2.1	3.5	1.7
CIRCLE50	0.5	0.5	0.3	0.3	0.6	2.0	3.3	4.5	5.5	2.2

Table 2: Standard deviation of prediction performance for each algorithm in terms of microlabel loss and 0/1 loss.