Moment Distributionally Robust Tree Structured Prediction

Yeshu Li **Danyal Saeed** **Xinhua Zhang** **Brian D. Ziebart**
Department of Computer Science
University of Illinois at Chicago
{yli299, dsaeed3, zhangx, bziebart}@uic.edu

Kevin Gimpel
Toyota Technological Institute at Chicago
kgimpel@ttic.edu

Abstract

Structured prediction of tree-shaped objects is heavily studied under the name of syntactic dependency parsing. Current practice based on maximum likelihood or margin is either agnostic to or inconsistent with the evaluation loss. Risk minimization alleviates the discrepancy between training and test objectives but typically induces a non-convex problem. These approaches adopt explicit regularization to combat overfitting without probabilistic interpretation. We propose a moment-based distributionally robust optimization approach for tree structured prediction, where the worst-case expected loss over a set of distributions within bounded moment divergence from the empirical distribution is minimized. We develop efficient algorithms for arborescences and other variants of trees. We derive Fisher consistency, convergence rates and generalization bounds for our proposed method. We evaluate its empirical effectiveness on dependency parsing benchmarks.

1 Introduction

Structured prediction is an important learning setting for joint prediction of interdependent variables. The output space typically consists of an exponential number of structured objects whose inherent relations can be exploited to develop efficient learning algorithms and capture key properties of data [Ciliberto et al., 2019]. Trees are widely used structures that offer expressiveness and simplicity. We distinguish between two different tree structured prediction tasks in the literature. The first task is a structure learning problem in graphical models [Bradley and Guestrin, 2010], aimed at constructing trees underlying a predictive model from training data. The optimal tree is found easily with greedy algorithms for generative models [Chow and Liu, 1968], while it is NP-hard for the discriminative max-margin setting [Meshi et al., 2013]. The second task requires prediction itself to be a tree-shaped object (e.g., an incidence vector). Dependency parsing is a crucial application of this problem that has inspired a flurry of work in natural language processing. The first-order spanning tree prediction assuming factorization over arcs can be done in $\mathcal{O}(n^2)$ [Stanojević and Cohen, 2021], whereas exact inference is NP-hard for certain (non-projective) higher-order trees (e.g., considering siblings) [McDonald and Satta, 2007]. We study the latter in this work.

A common evaluation criterion in dependency parsing is the attachment score, namely, the score we would like to maximize on test data. It is cost-sensitive to allow partially correct prediction. Ideally, the training objective should be aligned with the test objective. An early attempt to directly mimic test conditions leads to a non-convex piece-wise constant objective [Och, 2003]. Risk minimization in appropriate parametric form has a non-convex smooth objective, solvable with gradient descent,

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

but still losing global convergence and generalization guarantees. Maximum likelihood approaches formulate a convex smooth problem minimizing a logistic loss, consistent with conditional probability estimates but oblivious to test losses. Maximum margin methods have convex objectives able to implicitly incorporate custom losses by scaling margins, but are known to be inconsistent with test losses generally [Nowak-Vila et al., 2021]. Unfortunately, none of these approaches yield a Bayes optimal estimator for test losses with global convergence and finite-sample generalization guarantees.

Consistent structured prediction methods include Ciliberto et al. [2016], Blondel [2019], Nowak-Vila et al. [2020], the latter two of which are based on Fenchel-Young losses [Blondel et al., 2020]. However, none of them have addressed the tree structured prediction problem explicitly. For instance, Blondel [2019] calls for Euclidean or Kullback-Leibler projection oracles, which do not exist in an efficient sense from what we know for arborescence (directed tree) polytopes. In addition, the Frank-Wolfe type algorithm adopted by Nowak-Vila et al. [2020] requires a max-min oracle and converges in a rate of $\mathcal{O}(\frac{1}{\epsilon})$. Furthermore, all of the above methods belong to empirical risk minimization (ERM) that requires explicit regularization to combat overfitting, which can be quite vulnerable in high-dimensional settings (e.g., scarce data).

To address the above issues, we propose an estimator from first principles in distributionally robust optimization (DRO). It minimizes the worst-case risk over an ambiguity set of distributions within bounded moment divergence from the empirical distribution. We seek probabilistic prediction by assuming non-deterministic groundtruth labels, which, together with the ambiguity set, models uncertainty about the unknown true distribution. We interpret the primal problem as a dual-norm-regularized surrogate loss minimization problem. Note that prior art applying moment-based DRO to tree-structured graphical models [Fathony et al., 2018b] and bipartite matching [Fathony et al., 2018a] adopts a special case of our ambiguity set in which the empirical feature moments are matched exactly and regularization has to be imposed manually. This moment-based DRO also allows us to derive generalization bounds regarding true worst-case risks. When the ambiguity radius is zero, the DRO estimator is shown to be consistent. We develop two practical algorithms, one based on game theory and the other based on marginal probabilities of tree parts. We further propose efficient Euclidean projection oracles onto the arborescence polytope with linearly convergent guarantees. We conduct experiments on three common dependency parsing datasets, suggesting that our method is particularly effective with little training data.

Contributions. Our contributions are summarized as follows. (1) We propose a distributionally robust tree structured prediction method and show its equivalence to regularized surrogate minimization. (2) We derive its generalization bounds and consistency. (3) We propose efficient algorithms based on projection oracles for arborescence polytopes. (4) We perform empirical study on real-world datasets.

Paper structure. We begin with problem setup and existing work in Section 2. We present our method with theoretical analysis in Section 3. Section 4 proposes efficient projection oracles. Section 5 discusses extensions beyond first-order directed trees. Experimental results of comparing our method with a competitive baseline are given in Section 6. We conclude the paper in Section 7.

2 Background and Related Works

2.1 Tree Structured Prediction

Consider a weighted directed multi-graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where each arc $(i, j, l) \in \mathcal{E}$ from node i to j has a label l. By designating a root node $r \in \mathcal{V}$, we say that $\mathcal{A} \subseteq \mathcal{E}$ is an r-arborescence of $\mathcal{G}$ if $(\mathcal{V}, \mathcal{A})$ is a directed spanning tree rooted at r. For any $v \in \mathcal{V}$, denote by $\delta^-(v) := \{(i, j, l) \in \mathcal{E} : j = v\}$ the set of its incoming arcs, and $\delta^+(v) := \{(i, j, l) \in \mathcal{E} : i = v\}$ the set of its outgoing arcs.

Let $\mathcal{X}$ be the input space and $\mathcal{Y} \triangleq \bigcup_{\boldsymbol{x} \in \mathcal{X}} \mathcal{Y}(\boldsymbol{x})$ be the output space where $\mathcal{Y}(\boldsymbol{x})$ represents the set of r-arborescences of a graph $\mathcal{G}(\boldsymbol{x})$ formed by $\boldsymbol{x}$. Dependence on $\boldsymbol{x}$ is suppressed when context is clear. Let $\mathcal{R} \subseteq 2^{\mathcal{E}}$ be a set of parts with $\mathcal{E} \subseteq \mathcal{R}$. Each part $s \in \mathcal{R}$ is a subset of arcs. It is convenient to represent $\boldsymbol{y} \in \mathcal{Y}$ as a binary vector with $y_s = 1$ iff part s appears in $\boldsymbol{y}$. Let $w_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) \triangleq \sum_{s \in \mathcal{R}} w_{\boldsymbol{\theta}}(\boldsymbol{x}, y_s)$ be a score function decomposing over parts, parameterized by $\boldsymbol{\theta}$. Let $\{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\}_{i=1}^m$ be a set of m training examples drawn i.i.d. from a distribution $\mathbb{P} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, where each $\boldsymbol{y}^{(i)}$ is an r-arborescence. The goal of tree structured prediction is to learn a function $h : \mathcal{X} \to \mathcal{Y}$ from training data. Assume that the evaluation criterion is a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$.

We introduce existing methods in the setting of (graph-based, non-projective, syntactic) dependency parsing where $\boldsymbol{x}$ is a sequence of tokens and $\mathcal{G}(\boldsymbol{x})$ encodes dependencies among tokens.

2.2 Maximum Likelihood

A probabilistic modeling approach based on exponential family distributions maximizes the conditional log-likelihood of the training data:

$$\min_{\boldsymbol{\theta}} -\sum_{i=1}^{m} \log p_{\boldsymbol{\theta}}(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)}) := -\sum_{i=1}^{m} \log \left[\exp\left(w_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\right)/Z(\boldsymbol{x}^{(i)})\right],$$

where $Z(\boldsymbol{x}) \triangleq \sum_{\boldsymbol{y}\in\mathcal{Y}(\boldsymbol{x})} \exp\left(w_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y})\right)$. This problem is convex for log-linear models, but intractable for general $\mathcal{R}$ [Koller and Friedman, 2009]. The first-order arc-factored model ($\mathcal{R} = \mathcal{E}$) is equivalent to a loop-free factor graph, rendering it tractable via the matrix-tree theorem [Kirchhoff, 1847, William, 1984, Koo et al., 2007, McDonald and Satta, 2007, Smith and Smith, 2007]. Neural parsers either leverage the same theorem to compute the partition function [Ma and Hovy, 2017] or consider the parent node distribution independently for each node by local normalization [Dozat and Manning, 2017, Zhang et al., 2017]. Higher-order models require approximate algorithms such as loopy belief propagation [Murphy et al., 1999] and Markov chain Monte Carlo [Brooks, 1998]. This approach does not incorporate task-specific losses. In fact, with maximum a posteriori (MAP) decoding, it is not consistent with any specific loss in general [Nowak-Vila et al., 2019].

2.3 Maximum Margin

An alternative approach based on maximum margin Markov networks [Taskar et al., 2003] or structured support vector machines [Tsochantaridis et al., 2005] optimizes a hinge-type surrogate:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{m} -w_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) + \max_{\boldsymbol{y}} \ell(\boldsymbol{y}^{(i)}, \boldsymbol{y}) + w_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}, \boldsymbol{y}),$$

which inspires a rich line of work based on MAP inference with manual features [Taskar et al., 2004, McDonald et al., 2005, McDonald and Pereira, 2006, Martins et al., 2009, 2010, 2015, Zhang et al., 2014] or deep learning [Kiperwasser and Goldberg, 2016, Wang and Chang, 2016]. Approximate MAP inference is required for models beyond first-order. A smooth variant called softmax-margin [Gimpel and Smith, 2010] incorporates the task-specific loss ℓ but still implicitly minimizes it. Margin-based objectives are known to be consistent only under very restrictive conditions [Liu, 2007, Nowak-Vila et al., 2021] (i.e., data with majority label, loss being a distance).

2.4 Minimum Risk

Empirical risk minimization suggests directly optimizing the expected target loss on training data:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{m} \sum_{\boldsymbol{y}} p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}^{(i)}) \ell(\boldsymbol{y}^{(i)}, \boldsymbol{y}),$$

which is commonly non-convex due to normalization of $p_{\boldsymbol{\theta}}$. There are a few parsers optimizing this objective via back-propagation [Stoyanov and Eisner, 2012], k-best lists [Smith and Eisner, 2006], semirings [Li and Eisner, 2009, Zmigrod et al., 2021] and other differentiable approximations [Gormley et al., 2015, Mensch and Blondel, 2018]. Local optima found by these algorithms do not satisfy the premise of Fisher consistency and make it difficult to quantify generalization errors.

2.5 Distributionally Robust Optimization

Distributionally robust optimization has attracted emerging interests in improving machine learning models due to its connections to robustness, regularization and generalization. It proposes to minimize a risk with respect to the worst-case distribution chosen by an adversary in some uncertainty set:

$$\min_{\boldsymbol{\theta}} \max_{\mathbb{Q}\in\mathcal{B}} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{Y}, h_{\boldsymbol{\theta}}(\boldsymbol{X}))],$$

where $\mathcal{B}$ is an ambiguity set that can be defined by discrepancies [Shafieezadeh-Abadeh et al., 2019, Duchi and Namkoong, 2019], moments [Delage and Ye, 2010, Farnia and Tse, 2016], shapes [Popescu, 2005, Hanasusanto et al., 2015] and kernels [Shang et al., 2017, Staib and Jegelka, 2019]. A thorough review can be found in Rahimian and Mehrotra [2019]. We focus on moment-matching discriminative approaches while a similar generative method is proposed in Ganapathi et al. [2008].

3 Method

We introduce the formulation, followed by practical algorithms for learning and inference. Afterwards, we present the theoretical guarantees. We defer all proofs to Appendix A.

3.1 Formulation

We assume that the evaluation criterion is the Hamming loss $\ell(\boldsymbol{y}, \boldsymbol{y}') := \sum_i \mathbb{1}(y_i \neq y'_i)$ with $\mathbb{1}(\cdot)$ being the 0-1 indicator function, but the results in this paper generalize to losses with affine decomposition [Ramaswamy et al., 2013] easily.

Let $\mathbb{P}^{\text{true}}$ be the true distribution and $\mathbb{P}^{\text{emp}}$ be the empirical distribution. Our approach builds upon a probabilistic predictor that non-parametrically minimizes the expected loss with regard to the most adverse distribution in an uncertainty set where the distributions are ε away from the empirical distribution in terms of feature moment difference:

$$\min_{\mathbb{P}} \max_{\mathbb{Q} \in \mathcal{B}(\mathbb{P}^{\text{emp}})} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{X},\check{\boldsymbol{Y}}}, \mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}), \tag{1}$$

where $\mathcal{B}(\mathbb{P}^{\text{emp}}) := \{\mathbb{Q} : \mathbb{Q}_{\boldsymbol{X}} = \mathbb{P}^{\text{emp}}_{\boldsymbol{X}} \wedge \|\mathbb{E}_{\mathbb{P}^{\text{emp}}} \boldsymbol{\phi}(\cdot) - \mathbb{E}_{\mathbb{Q}} \boldsymbol{\phi}(\cdot)\| \leq \varepsilon\}$ with $\varepsilon \geq 0$ and $\boldsymbol{\phi} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ is a joint feature mapping decomposable over parts: $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) \triangleq \sum_s \boldsymbol{\phi}(\boldsymbol{x}, y_s)$. In Farnia and Tse [2016], cross-moments are adopted: $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) := \boldsymbol{\phi}_{\boldsymbol{X}}(\boldsymbol{x}) \otimes \boldsymbol{\phi}_{\boldsymbol{Y}}(\boldsymbol{y})$ where $\otimes$ is the tensor product.

By Fenchel duality [Altun and Smola, 2006] and strong duality [Von Neumann and Morgenstern, 1947], we show that Eq. (1) is analogous to dual-norm-regularized surrogate loss minimization:

Proposition 1. *The distributionally robust tree structured prediction problem based on moment divergence in Eq.* (1) *can be rewritten as*

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}^{emp}_{\boldsymbol{X},\boldsymbol{Y}}} \underbrace{\min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}, \mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^{\mathsf{T}}(\boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \boldsymbol{\phi}(\boldsymbol{X}, \boldsymbol{Y})) + \varepsilon \|\boldsymbol{\theta}\|_*}_{\ell_{adv}(\boldsymbol{\theta}, (\boldsymbol{X}, \boldsymbol{Y}))}, \tag{2}$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ is the vector of Lagrangian multipliers and $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

3.2 Constraint Generation Solution

From a game-theoretic rationale [Topsøe, 1979, Grünwald and Dawid, 2004], Eq. (1) is considered as an adversary-constrained zero-sum game. A prediction player chooses a set of stochastic strategies (conditional distributions over arborescences) in order to minimize the expected payoff whereas an adversarial player chooses constrained strategies to maximize it. The payoff for a pair of pure strategies is the incurred loss, $\ell(\hat{\boldsymbol{y}}, \check{\boldsymbol{y}})$. The constrained game is transformed to a set of unconstrained ones in Eq. (2) whose payoffs are parameterized by $\boldsymbol{\theta}$: payoff$(\hat{\boldsymbol{y}}, \check{\boldsymbol{y}}) \triangleq \ell(\hat{\boldsymbol{y}}, \check{\boldsymbol{y}}) + \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{\phi}(\boldsymbol{x}, \check{\boldsymbol{y}})$. Note that the games in Eq. (1) are jointly constrained for all $\boldsymbol{x}$'s in the support of $\mathbb{P}^{\text{emp}}_{\boldsymbol{X}}$ while the ones in Eq. (2) are conditionally independent given $\boldsymbol{x}$. The unconstrained game can be solved by a linear program [Von Neumann and Morgenstern, 1947]. However, there are $\mathcal{O}(n^n)$ spanning trees in a complete graph, thus making explicit construction of the full payoff matrix impractical.

We adopt a constraint generation algorithm named double oracle [McMahan et al., 2003], shown in Appendix B. It builds a payoff sub-matrix starting from small initial sets of strategies. In each iteration, each player takes their turn based on the game payoff sub-matrix by finding the best response among all possible strategies to the opponent's optimal mixture strategies. The response is added to a player's strategy set if it improves the value of the game, with the sub-matrix updated. The algorithm terminates and converges to a Nash equilibrium of the original game when the strategy sets no longer grow. The size of the final sub-matrix is usually small in practice but there are no known theoretical guarantees, thus no way to analyze the convergence behavior. Finding the best response requires an oracle, equivalent to finding the minimum weight arborescence. The objective in Eq. (2) is a convex function of $\boldsymbol{\theta}$, so we can optimize it with sub-gradients based on solutions of the inner zero-sum games. Although lacking convergence guarantees, this algorithm is flexible with custom losses and provides a game-theoretic perspective to a typical DRO problem.

3.3 Marginal Distribution Formulation

The r-arborescence polytope is defined as the convex hull of all vectors representing r-arborescences: $\mathcal{A}_{\text{arb}}(\boldsymbol{x}) := \text{Conv}(\{\boldsymbol{y} \in \mathbb{R}^{|\mathcal{R}|} : \boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})\})$. Note that each $\boldsymbol{p} \in \mathcal{A}_{\text{arb}}$ is a convex combination of all r-arborescences: $\boldsymbol{p} \triangleq \sum_{\boldsymbol{y}} \text{Prob}(\boldsymbol{y})\boldsymbol{y}$, where p_s denotes the marginal probability of part s. Here we adopt the squared ℓ_2 norm as the dual norm and an ambiguity radius of $\varepsilon = \lambda/2$. By substituting the marginal probability vectors and switching min-max optimization orders, we simplify Eq. (2) into

$$\max_{\boldsymbol{q}^{(i)} \in \mathcal{A}_{\text{arb}}} \min_{\boldsymbol{\theta}} \frac{1}{m} \sum_{i=1}^{m} \min_{\boldsymbol{p} \in \mathcal{A}_{\text{arb}}} (\boldsymbol{q}^{(i)} - \boldsymbol{p}_{\text{emp}}^{(i)})^\intercal \boldsymbol{\Phi}^{(i)} \boldsymbol{\theta} - \langle \boldsymbol{p}, \boldsymbol{q}^{(i)} \rangle + \frac{\mu}{2} \|\boldsymbol{p}\|_2^2 - \frac{\mu}{2} \|\boldsymbol{q}^{(i)}\|_2^2 + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2, \quad (3)$$

where $\boldsymbol{\Phi}^{(i)} \in \mathbb{R}^{|\mathcal{R}| \times d}$ denotes the feature matrix of the i-th training data, $\mu \in \mathbb{R}_{\geq 0}$ is a smoothing parameter to induce strong convexity. We push the maximization over $\boldsymbol{q}$ to the outermost level because of its large computational cost. If $\mu = 0$, the solution to Eq. (3) is also optimal to Eq. (2) by strong duality but the problem becomes non-smooth. Therefore we expect $\boldsymbol{\theta}^*$ obtained with a very small positive μ to be a good approximation of $\boldsymbol{\theta}^*$ obtained with $\mu = 0$.

To optimize it, with fixed $\boldsymbol{q}$, due to strong convexity, the unconstrained minimization over $\boldsymbol{\theta}$ yields $\boldsymbol{\theta}^* = -\frac{1}{m\lambda} \sum_{i=1}^{m} (\boldsymbol{\Phi}^{(i)})^\intercal (\boldsymbol{q}^{(i)} - \boldsymbol{p}_{\text{emp}}^{(i)})$. In contrast, the constrained minimization over $\boldsymbol{p}$ admits no closed-form solution but can be cast as Euclidean projection onto $\mathcal{A}_{\text{arb}}$ instead, independently for each $i \in [m]$: $\boldsymbol{p}^* = \min_{\boldsymbol{p} \in \mathcal{A}_{\text{arb}}} \|\boldsymbol{p} - \frac{1}{\mu} \boldsymbol{q}^{(i)}\|_2^2 \triangleq \text{Proj}_{\mathcal{A}_{\text{arb}}}(\frac{1}{\mu} \boldsymbol{q}^{(i)})$. Given $\boldsymbol{\theta}^*$ and $\boldsymbol{p}^*$, the outermost maximization can be solved by a projected quasi-Newton algorithm [Schmidt et al., 2009] that also requires the projection oracle $\text{Proj}_{\mathcal{A}_{\text{arb}}}(\cdot)$, elaborated in Section 4.

3.4 Inference

We propose two algorithms to make inference with given $\boldsymbol{\theta}^*$.

Weight construction. Construct the part weights as $\boldsymbol{\Phi}\boldsymbol{\theta}^* \in \mathbb{R}^{|\mathcal{R}|}$ and find the maximum weight arborescence: $\boldsymbol{y}^* \in \arg\max_{\boldsymbol{y}} \boldsymbol{y}\boldsymbol{\Phi}\boldsymbol{\theta}^*$ by the Gabow-Tarjan (GT) algorithm [Gabow et al., 1986, Zmigrod et al., 2020] or approximate methods for higher-order trees.

Minimum Bayes risk decoding. The optimal probabilistic prediction $\mathbb{P}^*$ or $\boldsymbol{p}^*$ can be obtained from Eq. (2) or Eq. (3). The marginal probabilities enable minimum Bayes risk decoding: $\boldsymbol{y}^* \in \arg\min_{\boldsymbol{y}} \mathbb{E}_{\mathbb{P}^*_{\hat{\boldsymbol{Y}}|\boldsymbol{x}}} \ell(\boldsymbol{y}, \hat{\boldsymbol{Y}}) \triangleq \arg\max_{\boldsymbol{y}} \sum_{s:y_s=1} \boldsymbol{p}_s^*$, a maximum weight arborescence problem.

3.5 Statistical Properties

Basic generalization bounds of DRO methods derived from measure concentration are not appropriate for an ambiguity set defined by low-order moments in this paper since it fails to converge [Shafieezadeh-Abadeh et al., 2019]. We take an alternate approach following Farnia and Tse [2016] to obtain excess out-of-sample risk bounds by assuming boundedness on features and losses.

Theorem 2. *Given m samples, a non-negative loss $\ell(\cdot,\cdot)$ such that $|\ell(\cdot,\cdot)| \leq K$, a feature function $\boldsymbol{\phi}(\cdot,\cdot)$ such that $\|\boldsymbol{\phi}(\cdot,\cdot)\| \leq B$, a positive ambiguity level $\varepsilon > 0$, then, for any $\rho \in (0, 1]$, with a probability at least $1 - \rho$, the following excess true worst-case risk bound holds:*

$$\max_{\mathbb{Q} \in \mathcal{B}(\mathbb{P}^{\textit{true}})} R_{\mathbb{Q}}^{L}(\boldsymbol{\theta}^*_{\textit{emp}}) - \max_{\mathbb{Q} \in \mathcal{B}(\mathbb{P}^{\textit{true}})} R_{\mathbb{Q}}^{L}(\boldsymbol{\theta}^*_{\textit{true}}) \leq \frac{4KB}{\varepsilon\sqrt{m}} \left(1 + \frac{3}{2}\sqrt{\frac{\ln(4/\rho)}{2}}\right),$$

where $\boldsymbol{\theta}^*_{\textit{emp}}$ and $\boldsymbol{\theta}^*_{\textit{true}}$ are the optimal parameters learned in Eq. (2) under $\mathbb{P}^{\textit{emp}}$ and $\mathbb{P}^{\textit{true}}$ respectively. The original risk of $\boldsymbol{\theta}$ under $\mathbb{Q}$ is $R_{\mathbb{Q}}^{L}(\boldsymbol{\theta}) := \mathbb{E}_{\mathbb{Q}_{\boldsymbol{X},\boldsymbol{Y}}, \mathbb{P}^{\boldsymbol{\theta}}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \boldsymbol{Y})$ with Bayes prediction $\mathbb{P}^{\boldsymbol{\theta}}_{\boldsymbol{Y}|\boldsymbol{x}} \in \arg\min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{x}} \mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{x}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^\intercal \boldsymbol{\phi}(\boldsymbol{x}, \check{\boldsymbol{Y}})$.

Theorem 2 presents a bound based on uniform convergence and Rademacher complexities [Bartlett and Mendelson, 2002], which improves the results in Asif et al. [2015], who merely show that the worst-case risk upper bounds the risk under any distribution in the ambiguity set.

The dual problem in Eq. (2) suggests an adversarial surrogate loss $\ell_{\text{adv}}(\boldsymbol{\theta}, (\boldsymbol{x}, \boldsymbol{y}))$ in a ERM form. The special case of $\varepsilon = 0$ in our DRO estimator has a similar form to the max-min surrogate loss in Nowak-Vila et al. [2020] except that we assume probabilistic prediction. A conclusion of its Fisher consistency can thus be drawn based on Fathony et al. [2018a], Nowak-Vila et al. [2020].

Corollary 3. *When $\varepsilon = 0$, ℓ_{adv} is Fisher consistent with respect to ℓ. Namely, $\mathbb{P}^{\boldsymbol{\theta}^*_{true}}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}$ is the probabilistic prediction made by the Bayes optimal decision rule, where $\boldsymbol{\theta}^*_{true}$ is defined in Theorem 2.*

If $\varepsilon > 0$, the decoded prediction for each $\boldsymbol{x}$ will not belong to the convex hull of true conditional distributions, thus not a minimizer of ℓ. On the other hand, if ε is chosen as $m^{-\alpha}$ for $0 < \alpha < 1/2$, ℓ_{adv} will be universally consistent according to the comparison inequality in Nowak-Vila et al. [2020].

4 Projection onto Arborescence Polytopes

The Euclidean projection onto an r-arborescence polytope is a quadratic programming problem[1]:

$$\min_{\boldsymbol{x} \in \mathcal{A}_{\text{arb}}} f(\boldsymbol{x}) := \|\boldsymbol{x} - \boldsymbol{w}\|_2^2.$$

We focus on first-order models and discuss the extensions to other classes of trees in Section 5.

4.1 Frank-Wolfe Algorithm

The Frank-Wolfe (FW) method [Frank et al., 1956] is an iterative first-order algorithm that enforces constraints by optimizing a linear objective over the feasible set at each iteration t:

$$\boldsymbol{s}^t \in \arg\min_{\boldsymbol{s} \in \mathcal{A}_{\text{arb}}} \boldsymbol{s}^\intercal \nabla f(\boldsymbol{x}^t), \tag{4}$$

which is a minimum weight arborescence problem with weights $\nabla f(\boldsymbol{x}^t)$ in our case. The solution is updated and stays feasible: $\boldsymbol{x}^{t+1} \leftarrow \boldsymbol{x}^t + \gamma_t(\boldsymbol{s}^t - \boldsymbol{x}^t)$, where γ_t is a step size typically set to $\frac{2}{t+2}$. FW style algorithms are known to have a convergence rate of $\mathcal{O}(\frac{1}{\epsilon})$ [Jaggi, 2013].

4.2 Martin's Polytope

A compact representation of $\mathcal{A}_{\text{arb}}$ with a polynomial number of linear constraints is attractive to lead to efficient algorithms. To the best of our knowledge, there is no existing projection method exploiting special structures of this polytope. An extended formulation of the arborescence polytope [Friesen, 2019, Martin, 1991] follows a lift-and-project approach. It relates each element to existence of k-arboresences of the underlying undirected graph for all $k \in \mathcal{V}$. We extend it to multi-graphs:

$$\mathcal{A}_{\text{marb}} := \{\boldsymbol{z}^r : \exists \boldsymbol{z}^k \geq \boldsymbol{0} \sum_{a \in \delta^-(j)} z_a^k = \mathbb{1}(j \neq k)\ \forall k, j \in \mathcal{V} \wedge \sum_{a \in \mathcal{E}'_{ij}} z_a^k = \sum_{a \in \mathcal{E}_{ij}} z_a^r\ \forall k \neq r, i, j \in \mathcal{V} \wedge \boldsymbol{z}^r \geq \boldsymbol{0}\},$$

where $\boldsymbol{z}^r \in \mathbb{R}^{|\mathcal{E}|}$ is associated with the original arcs $\mathcal{E}$, $\boldsymbol{z}^k \in \mathbb{R}^{|\mathcal{E}'|}$ for $k \neq r$ is associated with a simple directed graph $(\mathcal{V}, \mathcal{E}')$ formed by removing directions and splitting each edge $\{i, j\}$ into two directed ones, $\mathcal{E}_{ij} := \{a \in \mathcal{E} : \bar{a} = \{i, j\}\}$ is the set of arcs connecting i and j with $\bar{a} \triangleq \overline{(i, j, l)} := \{i, j\}$ denoting the underlying undirected edge. We show exact correspondence between $\mathcal{A}_{\text{marb}}$ and $\mathcal{A}_{\text{arb}}$ based on a similar argument for simple graphs [Friesen, 2019]:

Proposition 4. *Let $\mathcal{G}$ be a multi-graph. $\mathcal{A}_{marb} \triangleq \mathcal{A}_{arb}$.*

[1] This is a well-defined convex optimization problem, different from that in differentiable structured prediction methods [Peng et al., 2018, Mihaylova et al., 2020] which elicit gradients with respect to inputs.

To solve $\min_{\boldsymbol{x}\in\mathcal{A}_{\text{marb}}}\|\boldsymbol{x}-\boldsymbol{w}\|_2^2$, we propose to adopt the alternating direction method of multipliers (ADMM) and rewrite it into the following separable form:

$$\begin{aligned}
&\min_{\boldsymbol{u}} g(\boldsymbol{u}) := \sum_{k\in\mathcal{V}} \frac{1}{|\mathcal{V}|}\|\boldsymbol{u}_k-\boldsymbol{w}\|_2^2 + I_{\mathcal{U}_k}(\boldsymbol{u}_k)\\
\text{s.t.}\quad &\mathcal{U}_k := \{\boldsymbol{x}\in\mathbb{R}^{|\mathcal{E}|} : \exists \boldsymbol{z}\in\mathbb{R}^{|\mathcal{E}'|}_{\geq 0} \sum_{a\in\delta^-(j)} z_a = \mathbb{1}(j\neq k)\wedge \sum_{a\in\mathcal{E}'_{ij}} z_a = \sum_{a\in\mathcal{E}_{ij}} x_a \;\forall i,j\in\mathcal{V}\}\\
&\boldsymbol{u}_r = \boldsymbol{u}_k \quad \forall k\in\mathcal{V}\setminus r,\quad \mathcal{U}_r := \{\boldsymbol{x}\in\mathbb{R}^{|\mathcal{E}|}_{\geq 0} : \sum_{a\in\delta^-(j)} x_a = \mathbb{1}(j\neq r)\;\forall j\in\mathcal{V}\},
\end{aligned}$$

where $I_{\mathcal{U}}(\cdot)$ is the characteristic function with $I_{\mathcal{U}}(\boldsymbol{x})=0$ if $\boldsymbol{x}\in\mathcal{U}$ and ∞ otherwise.

Let $\boldsymbol{\lambda}'_k$ be the dual variables and $\boldsymbol{\lambda}_k := \frac{1}{\rho_k}\boldsymbol{\lambda}'_k$. The scaled augmented Lagrangian function is $L_\rho(\boldsymbol{u},\boldsymbol{\lambda}) = g(\boldsymbol{u}) + \sum_{k\neq r}\frac{\rho_k}{2}\|\boldsymbol{u}_r-\boldsymbol{u}_k+\boldsymbol{\lambda}_k\|_2^2 - \frac{\rho_k}{2}\|\boldsymbol{\lambda}_k\|_2^2$.

The ADMM algorithm updates the parameters as follows:

$$\begin{aligned}
\boldsymbol{u}_k^{t+1} &:= \arg\min_{\boldsymbol{u}_k\in\mathcal{U}_k} L_\rho((\boldsymbol{u}_r^t,\boldsymbol{u}_k^t),\boldsymbol{\lambda}^t) \triangleq \text{Proj}_{\mathcal{U}_k}(\frac{2\boldsymbol{w}+\rho_k|\mathcal{V}|(\boldsymbol{u}_r^t+\boldsymbol{\lambda}_k^t)}{2+\rho_k|\mathcal{V}|}) \quad \forall k\neq r\\
\boldsymbol{u}_r^{t+1} &:= \arg\min_{\boldsymbol{u}_r\in\mathcal{U}_r} L_\rho((\boldsymbol{u}_r^t,\boldsymbol{u}_k^{t+1}),\boldsymbol{\lambda}^t) \triangleq \text{Proj}_{\mathcal{U}_r}(\frac{2\boldsymbol{w}+|\mathcal{V}|\sum_{k\neq r}\rho_k(\boldsymbol{u}_k^{t+1}-\boldsymbol{\lambda}_k^t)}{2+|\mathcal{V}|\sum_{k\neq r}\rho_k})\\
\boldsymbol{\lambda}_k^{t+1} &:= \boldsymbol{\lambda}_k^t + (\boldsymbol{u}_r^{t+1}-\boldsymbol{u}_k^{t+1}) \quad \forall k\neq r.
\end{aligned}$$

This decomposes the original projection problem into simpler projection problems. Projection onto $\mathcal{U}_k$ for $k=r$ decomposes over $j\in\mathcal{V}$ into $|\mathcal{V}|$ projections onto simplex, solvable as fast as $\mathcal{O}(n)$ in the worst case [Condat, 2016]. For $k\neq r$, computation of $\boldsymbol{u}_k^{t+1}$ can be done in parallel. The Lagrange dual problem of $\text{Proj}_{\mathcal{U}_k}(\cdot)$ can be written as

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^{|\mathcal{V}|}} \sum_{\{i,j\}\in\bar{\mathcal{E}}} h_{ij}(\boldsymbol{\alpha}) - \sum_{j\neq k}\alpha_j \quad \text{s.t. } h_{ij}(\boldsymbol{\alpha}) = \begin{cases} w_{ij}^2/n_{ij} & \text{if } \alpha_{ij} > 2w_{ij}/n_{ij},\\ -n_{ij}\alpha_{ij}^2/4+\alpha_{ij}w_{ij} & \text{if } \alpha_{ij}\leq 2w_{ij}/n_{ij},\end{cases}$$

where $w_{ij} := \sum_{a\in\mathcal{E}_{ij}} w_a$, $n_{ij} := |\mathcal{E}_{ij}|$, $\alpha_{ij} := \min(\alpha_i,\alpha_j)$ and $\alpha_k := +\infty$. Strong duality holds by linear constraint qualification. Primal solutions are recovered by $x_a^* = w_a - \min(\alpha_{\bar{a}}^*/2, w_{\bar{a}}/n_{\bar{a}})$.

Convergence. The dual objective of $\text{Proj}_{\mathcal{U}_k}(\cdot)$ is strongly concave on $\{\boldsymbol{\alpha}\in\mathbb{R}^{|\mathcal{V}|} : \forall i\exists j\ \{i,j\}\in\overline{\mathcal{E}}\wedge\alpha_i\leq\alpha_j\wedge\alpha_i\leq 2w_{ij}/n_{ij}\}$, with a unique global maximizer. This implies fast convergence in practice given good initialization. The negative Lagrange dual function has restricted strong convexity with $\nu = \min_{ij}(n_{ij}/2)$, near the optimum, suggesting linear convergence [Zhang and Cheng, 2015]. Alternatively, exact solutions can be found by enumerating rankings (with duplicates) of $\boldsymbol{\alpha}$ in $\mathcal{O}(|\mathcal{V}|^{|\mathcal{V}|})$. In this manner, the ADMM algorithm with a strongly convex objective has a linear convergence rate $\mathcal{O}(\log\frac{1}{\epsilon})$ with either exact [Deng and Yin, 2016] or linearly convergent approximate solution [Hager and Zhang, 2020] of $\text{Proj}_{\mathcal{U}_k}(\cdot)$. Using Nesterov's accelerated gradient algorithm [Nesterov, 2003] to optimize Eq. (3) leads to iteration complexity $\mathcal{O}(C\log\frac{1}{\epsilon})$ with constant C dependent on Lipschitz constants of gradients and μ.

5 Extensions

5.1 Undirected Spanning Trees

An straight-forward way of extending to undirected spanning trees is to split $\{i,j\}$ into two arcs (i,j), (j,i) and make the feature mapping direction-invariant, i.e., $\boldsymbol{\phi}(\boldsymbol{x},y_s)=\boldsymbol{\phi}(\boldsymbol{x},y_{s'})$ for s and s' having the same underlying undirected graph. We post-process the prediction by removing directions.

Alternatively, we seek projection oracles for undirected graphs. Projection via FW is done by using any minimum spanning tree algorithm in Eq. (4). For ADMM, the formulation in Martin [1991] is originally for undirected trees: $\mathcal{A}_{\text{mund}} := \{\boldsymbol{x} : \exists\boldsymbol{z}\geq\boldsymbol{0}\sum_{a\in\delta^-(j)} z_a^k = \mathbb{1}(j\neq k)\wedge z_{ij}^k + z_{ji}^k = x_{\{i,j\}}\forall k,i,j\in\mathcal{V}\}$. ADMM is easily adapted to this case with $\sum_{a\in\mathcal{E}_{ij}} x_a$ replaced by $x_{\{i,j\}}$.

5.2 Dependency Trees

The spanning tree structure in dependency parsing is a special one where the outdegree of root is restricted to be one. We can use the GT algorithm for inference with either the same training objective or an aligned objective where a dependency tree polytope is considered: $\mathcal{A}_{\text{dep}}(\boldsymbol{x}) := \text{Conv}(\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}) : |\delta^+(r)| = 1\})$. A straightforward extension of $\mathcal{A}_{\text{marb}}$ to characterizing dependency trees is $\mathcal{A}_{\text{mdep}} := \{\boldsymbol{z}^r : \boldsymbol{z}^r \in \mathcal{A}_{\text{marb}} \wedge \sum_{a \in \delta^+(r)} z^r_a = 1\}$, equivalent to $\mathcal{A}_{\text{dep}}$ by the following proposition:

Proposition 5. *Let $\mathcal{G}$ be a multi-graph. $\mathcal{A}_{mdep} \triangleq \mathcal{A}_{dep}$.*

FW methods leverage the GT algorithm in Eq. (4). As for ADMM, the dual problem of projection onto $\mathcal{U}'_r := \{\boldsymbol{x} : \boldsymbol{x} \in \mathcal{U}_r \wedge \sum_{a \in \delta^+(r)} x_a = 1\}$ becomes

$$\max_{\boldsymbol{\alpha}, \beta} \sum_{a \in \mathcal{E}} h_a(\boldsymbol{\alpha}, \beta) - \sum_{j \neq r} \alpha_j - \beta \quad \text{s.t. } h_a(\boldsymbol{\alpha}, \beta) = \begin{cases} w_a^2 & \gamma_a > 2w_a, \\ w_a \gamma_a - \gamma_a^2/4 & \gamma_a \leq 2w_a, \end{cases}$$

where $\gamma_{(i,j,l)} := \alpha_j + \mathbb{1}(i = r)\beta$. This can be solved in $\mathcal{O}(|\mathcal{E}| \log |\mathcal{E}|)$ [Zhang et al., 2010].

5.3 Higher-order Polytope

Compact higher-order polytope descriptions exist for undirected spanning trees but are still unknown for arborescences with even one monomial [Friesen, 2019]. FW requires a linear oracle that is NP-hard to solve exactly in higher-order settings [McDonald and Pereira, 2006].

Instead, we can approximate it with a local polytope where the marginal probabilities of each part s is required to be locally consistent with that of each arc a. For simplicity, we consider only features for the all-true assignments, i.e., all arcs exist in part s. The resulting polytope can be written as $\mathcal{A}_{\text{mloc}} := \{\boldsymbol{x} : \boldsymbol{x}_{\mathcal{E}} \in \mathcal{A}_{\text{marb}} \wedge \forall s \in \mathcal{R}, a \in s \quad p_s \leq p_a\}$, which suggests an ADMM algorithm with additional constraint sets for each part: $\mathcal{U}_s := \{\boldsymbol{x} \in \mathbb{R}^{|\mathcal{R}|}_{\geq 0} : x_s \leq x_a \quad \forall a \in s\}$, the projection onto which can be done in $\mathcal{O}(|s| \log |s|)$. See Appendix D for details.

6 Experiments

We evaluate our proposed method on dependency parsing tasks and compare its ability to *BiAF* [Dozat and Manning, 2017], arguably the state-of-the-art neural dependency parser. We implement our methods in Python and C[2]. We leverage the implementations in SuPar[3] [Zhang et al., 2020] for the baseline. All experiments are conducted on a computer with an Intel Core i7 CPU (2.7 GHz) and an NVIDIA Tesla P100 GPU (16 GB).

We adopt three public datasets, the English Penn Treebank (PTB v3.0) [Marcus et al., 1993], the Penn Chinese Treebank (CTB v5.1) [Xue et al., 2002] and the Universal Dependencies (UD v2.3) [Nivre et al., 2016]. See Appendix C for data-processing details.

Representation learning is not the focus of this paper. We follow Levy et al. [2020] and compare our method with the last biaffine classification layer in *BiAF* on top of pretrained features preceding this layer (backbone's output). The pretrained embeddings produced by complicated non-linear models make Fisher consistency's assumption of optimizing over all measurable functions less violated. To featurize the data, for each dataset, we train a *BiAF* network with the whole training set to obtain a pretrained model. Note that this may create unfair advantages for the baseline because the last layer was optimized together with the backbone network in an end-to-end manner during pretraining. Moreover, pretraining uses a standard ERM objective with the cross-entropy loss and local normalization over head nodes. The pretrained features are thus more adequate for the ERM objective than for our DRO objective. To make use of the features as inputs in our method, we take the outer product of the embedding vectors for two nodes as the arc feature vector. Our method and the biaffine layer therefore share the same number of parameters (501×501, including bias terms). We focus on predicting the unlabeled dependency tree while relying on pretrained models for relation label prediction. The evaluation criteria are the labeled/unlabeled attachment scores (LAS/UAS) and

[2] Our code is publicly available at `https://github.com/DanielLeee/drtreesp`.

[3] `https://github.com/yzhangcs/parser`

Table 1: Comparison of mean UAS and execution time under different training set sizes. Time refers to the CPU time taken to finish one gradient descent step. Statistically significant differences compared to *BiAF* are marked with † (paired t-test, $p < 0.05$). The best UAS are highlighted in bold.

		PTB				CTB				UD Dutch				UD Turkish (low resource)			
Method	Time (s)	m = 10	50	100	1000	m = 10	50	100	1000	m = 10	50	100	1000	m = 10	50	100	1000
BiAF	0.34	93.48	**96.87**	**96.95**	**97.16**	88.45	90.89	91.15	**91.70**	90.86	93.80	94.15	94.98	17.64	26.59	30.75	42.82
Marginal	0.28	94.51†	96.81†	96.92	97.12	89.19†	91.03†	91.27	91.67	**92.41**†	94.22†	94.50†	**95.15**†	24.85†	**32.83**†	**33.75**†	**43.18**
Stochastic	2.72	**94.62**†	96.81	96.93	97.14	**89.27**†	91.03†	**91.27**	91.66	92.40†	94.23†	94.47	95.14†	**25.06**†	31.35†	33.62†	41.20†
Game	7.25	94.51†	96.86	96.92	97.08†	89.22†	**91.06**†	91.22	91.57†	92.32†	**94.34**†	**94.59**†	95.01	19.85	23.18†	27.12†	36.30†

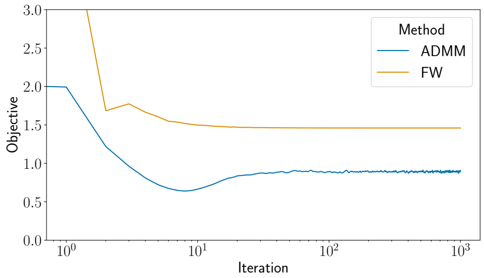

Figure 1: Convergence of ADMM and FW for random points with 95% confidence intervals.

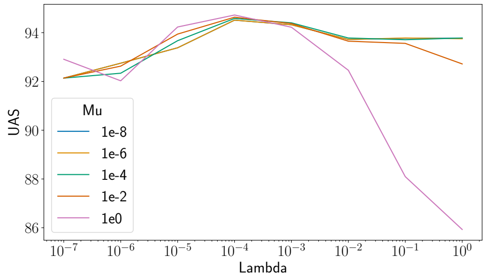

Figure 2: The best UAS with the Marginal algorithm as μ and λ vary in logarithmic scales.

labeled/unlabeled complete matches (LCM/UCM). The attachment score can be transformed to the Hamming loss with linear mapping: $\mathrm{AS}(\boldsymbol{y}, \boldsymbol{y}') \triangleq |\mathcal{V}| - 1 - \ell(\boldsymbol{y}, \boldsymbol{y}')/2$.

Full batch learning is adopted for *Marginal* (Eq. (3)). Mini-batch training is adopted for *Game*, the game-theoretic algorithm, and *Stochastic*, which solves the inner min-max problem in Eq. (2) using Eq. (3) with fixed $\boldsymbol{\theta}$. All models are trained with the training set only. The optimal hyperparameters and parameters are chosen based on the validation set. See Appendix C for detailed parameter values.

To showcase the ability of DRO methods tackling scarce data, in each run, we randomly draw $m \in \{10, 50, 100, 1000\}$ samples without replacement from the training set and keep the original validation and test sets. All the models are trained on the same set of sampled data. The process is repeated 5 times for each m. The main UAS results on the PTB, CTB and UD Dutch Lassy Small datasets are reported in Table 1 with complete results provided in Appendix C. Our methods consistently deliver higher UAS than *BiAF* especially with a small amount of data[4]. With little training data, DRO approaches minimize the worst-case risk to avoid overfitting. With more training data available, our method is still comparable to *BiAF* which is not significantly better than our methods by statistical tests. This illustrates the advantages of replacing conditional log-likelihood with our Fisher consistent surrogate loss without changing the number of model parameters. Moreover, we study a low-resource setting with the UD Turkish dataset in which only the sampled data is used for pretraining without BERT embeddings. The binary cross-entropy loss (single normalization) is adopted during pretraining in this setting to avoid pretrained features biased towards the multi-class cross-entropy loss (local normalization) adopted by *BiAF*. We observe consistently competitive performance of our methods in the low-resource setting in Table 1 as well.

We report computational time of one gradient descent step in the second column of Table 1, averaged across 10 runs. For fair comparisons, all the models are run with CPU only, with a batch size of 200. All the methods achieve their optimal validation set performance in 150-300 steps. *BiAF* and *Marginal* are the fastest because the most time-consuming step of computing dot products of features and parameters is only performed once whereas the other two methods perform it multiple times. However, since *Marginal* is unable to leverage stochastic gradients, its execution time grows linearly in the full batch size. Henceforth, there is a trade-off between *Marginal* and *Stochastic*/*Game* for computational efficiency. The extra cost compared to *BiAF* with cross entropy is expected because distributional robustness against a set of adversarial distributions is guaranteed.

[4] The UAS is high with 10 training samples possibly because (1) the backbone sub-network and linear layer were trained together with the whole training set; (2) BERT embeddings yield data representation that is easily linearly separable; (3) 10 samples result in as many as $10 \times 20 \times 20$ balanced head-selection instances for *BiAF*.

We compare ADMM and FW by performing for 100 times projection of random points in $[-5, 5]^{75}$ on a graph with 5 nodes and 3 parallel arcs between each (i, j). We subtract the integral part of the observed minimum values in each run for better illustration. As shown in Figure 1, ADMM usually finds a better solution in the arborescence polytope than FW does within 1000 iterations[5]. That being said, the per-iteration cost of ADMM is about $8n$ times higher than that of FW due to consensus optimization of n subproblems. In practice, the solution computed with FW usually leads to an approximately good sub-derivative to optimize the DRO objective. We have verified that the solutions suggested by ADMM satisfy the polytope constraints for graphs of up to 10 nodes.

We conduct sensitivity analysis by varying μ and λ on UD Dutch with 100 training samples. Figure 2 implies that moderate smoothing is beneficial to generalization. The ambiguity radius should be judiciously chosen because a small λ causes overfitting while a large λ leads to conservative models.

7 Discussion and Conclusion

We proposed a distributionally robust and consistent tree structured prediction method. We showed its equivalence to regularized surrogate loss minimization. We put forward a provably convergent algorithm based on efficient projection oracles for arborescence polytopes. Our proposed method enjoys Fisher consistency and robustness against noise in conditional distributions in terms of feature moments. Theoretical and empirical results validate its effectiveness.

We assume that an expressive feature mapping is given such that a sufficiently good linear discriminant rule can be learned. The class-sensitive form $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$ is general but consumes more memory than the decomposable form $\boldsymbol{\phi}_{\boldsymbol{X}}(\boldsymbol{x}) \otimes \boldsymbol{\phi}_{\boldsymbol{Y}}(\boldsymbol{y})$. The ADMM projection algorithm is efficient theoretically with high per-iteration costs in practice. We expect this work to be a principled way of learning to predict tree-structured objects. Future directions include a more efficient implementation and general structured prediction with DRO. Potential negative societal impacts of our work include using its prediction without verification to guide human-centered design in policy-making.

Representation learning. Our method can be easily adapted to a representation learning framework with automatic differentiation. Although this may lead to a non-convex problem without the theoretical guarantees derived in this paper, it is highly desired in practice if feature mappings are optimized as well. We discuss a possible approach as follows. Modern neural networks for supervised learning typically have a linear layer in the end without activation. Assume the penultimate layer outputs $\boldsymbol{\Phi}(\boldsymbol{x})$ for input $\boldsymbol{x}$, the last layer with parameters $\boldsymbol{\theta}$ will typically output $\boldsymbol{\psi}(\boldsymbol{x}) := \boldsymbol{\Phi}(\boldsymbol{x})\boldsymbol{\theta} \in \mathbb{R}^k$, sometimes called logits, with $k = n^2$ labels for all arcs when parsing a sentence of n tokens. Note that $\boldsymbol{\theta}$ in our formulation naturally serves as the parameters of this linear layer. Moreover, knowing $\boldsymbol{\psi}(\boldsymbol{x})$ is sufficient for us to solve the inner minimax problem in Eq. (2) to get $\mathbb{P}^*_{\hat{\boldsymbol{Y}}|\boldsymbol{x}}$ and $\mathbb{Q}^*_{\check{\boldsymbol{Y}}|\boldsymbol{x}}$. In this way, our DRO method can be considered a loss layer without learnable parameters, which backpropagates the sub-derivative of the objective with respect to $\boldsymbol{\psi}(\boldsymbol{x})$:

$$\frac{\partial}{\partial \boldsymbol{\psi}(\boldsymbol{x})} \ell_{\text{adv}} \in \frac{1}{B} \sum_{i=1}^{B} (\boldsymbol{q}^{(i)*} - \boldsymbol{p}^{(i)*}_{\text{emp}}),$$

where B is the batch size. The sub-derivative of the regularization term with respect to $\boldsymbol{\theta}$ should be added to the linear layer. Now we are able to take advantage of automatic differentiation and focus on solving the inner adversarial problem given $\boldsymbol{\psi}(\boldsymbol{x})$ and $\boldsymbol{y}$. Since the computational bottleneck lies in computing $\boldsymbol{\psi}(\boldsymbol{x})$ and backward passes, the overhead of computing the adversarial loss may be dominated and not significant compared to the cross-entropy loss. We leave investigations on its effective applications to future work.

Acknowledgments and Disclosure of Funding

This material is based upon work supported by the National Science Foundation under Grant Nos. 1652530, 1910146, and 1934915.

[5] One explanation is that FW relies on first-order approximations while there are exponential number of facets in the arborescence polytope.

References

Yasemin Altun and Alex Smola. Unifying divergence minimization and statistical inference via convex duality. In *International Conference on Computational Learning Theory*, pages 139–153. Springer, 2006.

Kaiser Asif, Wei Xing, Sima Behpour, and Brian D Ziebart. Adversarial cost-sensitive classification. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 92–101, 2015.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Mathieu Blondel. Structured prediction with projection oracles. *Advances in Neural Information Processing Systems*, 32:12145–12156, 2019.

Mathieu Blondel, André FT Martins, and Vlad Niculae. Learning with Fenchel-Young losses. *J. Mach. Learn. Res.*, 21(35):1–69, 2020.

Stephen Boyd, Neal Parikh, and Eric Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

Joseph K Bradley and Carlos Guestrin. Learning tree conditional random fields. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 127–134, 2010.

Stephen Brooks. Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society: Series D (the Statistician)*, 47(1):69–100, 1998.

Danqi Chen and Christopher D Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750, 2014.

C.K. Chow and Cong Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. *Advances in neural information processing systems*, 29:4412–4420, 2016.

Carlo Ciliberto, Francis Bach, and Alessandro Rudi. Localized structured prediction. *Advances in Neural Information Processing Systems*, 32, 2019.

Laurent Condat. Fast projection onto the simplex and the l_1 ball. *Mathematical Programming*, 158 (1-2):575, 2016.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019. URL `http://arxiv.org/abs/1911.02116`.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online, November 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.findings-emnlp.58`.

Marie-Catherine De Marneffe and Christopher D Manning. The Stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8, 2008.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

Wei Deng and Wotao Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 66(3):889–916, 2016.

Timothy Dozat and Christopher D. Manning. Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=Hk95PK9le`.

John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *The Journal of Machine Learning Research*, 20(1):2450–2504, 2019.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A Smith. Transition-based dependency parsing with stack long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 334–343, 2015.

Farzan Farnia and David Tse. A minimax approach to supervised learning. *Advances in Neural Information Processing Systems*, 29:4240–4248, 2016.

Rizal Fathony, Sima Behpour, Xinhua Zhang, and Brian Ziebart. Efficient and consistent adversarial bipartite matching. In *International Conference on Machine Learning*, pages 1457–1466. PMLR, 2018a.

Rizal Fathony, Ashkan Rezaei, Mohammad Ali Bashiri, Xinhua Zhang, and Brian D Ziebart. Distributionally robust graphical models. In *Advances in Neural Information Processing Systems*, pages 8354–8365, 2018b.

Marguerite Frank, Philip Wolfe, et al. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Mirjam Friesen. *Extended formulations for higher order polytopes in combinatorial optimization*. PhD thesis, Otto von Guericke University Magdeburg, 2019.

Harold N Gabow, Zvi Galil, Thomas Spencer, and Robert E Tarjan. Efficient algorithms for finding minimum spanning trees in undirected and directed graphs. *Combinatorica*, 6(2):109–122, 1986.

Varun Ganapathi, David Vickrey, John Duchi, and Daphne Koller. Constrained approximate maximum entropy learning of Markov random fields. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 196–203, 2008.

Kevin Gimpel and Noah A Smith. Softmax-margin crfs: Training log-linear models with cost functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 733–736, 2010.

Matthew R Gormley, Mark Dredze, and Jason Eisner. Approximation-aware dependency parsing by belief propagation. *Transactions of the Association for Computational Linguistics*, 3:489–501, 2015.

Peter D Grünwald and A Philip Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *the Annals of Statistics*, 32(4):1367–1433, 2004.

William W Hager and Hongchao Zhang. Convergence rates for an inexact admm applied to separable convex optimization. *Computational Optimization and Applications*, 77(3):729–754, 2020.

Grani A Hanasusanto, Vladimir Roitch, Daniel Kuhn, and Wolfram Wiesemann. A distributionally robust perspective on uncertainty quantification and chance constrained programming. *Mathematical Programming*, 151(1):35–62, 2015.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435. PMLR, 2013.

Eliyahu Kiperwasser and Yoav Goldberg. Simple and accurate dependency parsing using bidirectional lstm feature representations. *Transactions of the Association for Computational Linguistics*, 4: 313–327, 2016.

Gustav Kirchhoff. Ueber die auflösung der gleichungen, auf welche man bei der untersuchung der linearen vertheilung galvanischer ströme geführt wird. *Annalen der Physik*, 148(12):497–508, 1847.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

Terry Koo, Amir Globerson, Xavier Carreras Pérez, and Michael Collins. Structured prediction models via the matrix-tree theorem. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 141–150, 2007.

Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

Zhifei Li and Jason Eisner. First-and second-order expectation semirings with applications to minimum-risk training on translation forests. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 40–51, 2009.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL `http://arxiv.org/abs/1907.11692`.

Yufeng Liu. Fisher consistency of multicategory support vector machines. In *Artificial Intelligence and Statistics*, pages 291–298. PMLR, 2007.

Xuezhe Ma and Eduard Hovy. Neural probabilistic model for non-projective MST parsing. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 59–69, 2017.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. 1993.

R Kipp Martin. Using separation algorithms to generate mixed integer model reformulations. *Operations Research Letters*, 10(3):119–128, 1991.

André Filipe Torres Martins. *The geometry of constrained structured prediction: applications to inference and learning of natural language syntax*. PhD thesis, Carnegie Mellon University, 2012.

André FT Martins, Noah A Smith, and Eric Xing. Concise integer linear programming formulations for dependency parsing. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 342–350, 2009.

André FT Martins, Noah A Smith, Eric Xing, Pedro Aguiar, and Mario Figueiredo. Turbo parsers: Dependency parsing by approximate variational inference. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 34–44, 2010.

André FT Martins, Mário AT Figueiredo, Pedro MQ Aguiar, Noah A Smith, and Eric P Xing. Ad3: Alternating directions dual decomposition for map inference in graphical models. *The Journal of Machine Learning Research*, 16(1):495–545, 2015.

Ryan McDonald and Fernando Pereira. Online learning of approximate dependency parsing algorithms. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

Ryan McDonald and Giorgio Satta. On the complexity of non-projective data-driven dependency parsing. In *Proceedings of the Tenth International Conference on Parsing Technologies*, pages 121–132, 2007.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajic. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 523–530, 2005.

H Brendan McMahan, Geoffrey J Gordon, and Avrim Blum. Planning in the presence of cost functions controlled by an adversary. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 536–543, 2003.

Arthur Mensch and Mathieu Blondel. Differentiable dynamic programming for structured prediction and attention. In *International Conference on Machine Learning*, pages 3462–3471. PMLR, 2018.

Ofer Meshi, Elad Eban, Gal Elidan, and Amir Globerson. Learning max-margin tree predictors. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 411–420, 2013.

Tsvetomila Mihaylova, Vlad Niculae, and André FT Martins. Understanding the mechanics of spigot: Surrogate gradients for latent structure learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2186–2202, 2020.

Kevin P Murphy, Yair Weiss, and Michael I Jordan. Loopy belief propagation for approximate inference: an empirical study. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 467–475, 1999.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, 2016.

Alex Nowak-Vila, Francis Bach, and Alessandro Rudi. A general theory for structured prediction with smooth convex surrogates. *arXiv preprint arXiv:1902.01958*, 2019.

Alex Nowak-Vila, Francis Bach, and Alessandro Rudi. Consistent structured prediction with max-min margin Markov networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Alex Nowak-Vila, Alessandro Rudi, and Francis Bach. Max-margin is dead, long live max-margin! *arXiv preprint arXiv:2105.15069*, 2021.

Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st annual meeting of the Association for Computational Linguistics*, pages 160–167, 2003.

Hao Peng, Sam Thomson, and Noah A Smith. Backpropagating through structured argmax using a spigot. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1863–1873, 2018.

Ioana Popescu. A semidefinite programming approach to optimal-moment bounds for convex classes of distributions. *Mathematics of Operations Research*, 30(3):632–657, 2005.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

Harish G Ramaswamy, Shivani Agarwal, and Ambuj Tewari. Convex calibrated surrogates for low-rank loss matrices with applications to subset ranking losses. In *Advances in Neural Information Processing Systems*, pages 1475–1483, 2013.

Mark Schmidt, Ewout Berg, Michael Friedlander, and Kevin Murphy. Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm. In *Artificial Intelligence and Statistics*, pages 456–463. PMLR, 2009.

Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

Chao Shang, Xiaolin Huang, and Fengqi You. Data-driven robust optimization based on kernel learning. *Computers & Chemical Engineering*, 106:464–479, 2017.

David A Smith and Jason Eisner. Minimum risk annealing for training log-linear models. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 787–794, 2006.

David A Smith and Noah A Smith. Probabilistic models of nonprojective dependency trees. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 132–140, 2007.

Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems*, 32:9134–9144, 2019.

Miloš Stanojević and Shay B Cohen. A root of a problem: Optimizing single-root dependency parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10540–10557, 2021.

Veselin Stoyanov and Jason Eisner. Minimum-risk training of approximate CRF-based NLP systems. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 120–130, 2012.

Ben Taskar, Carlos Guestrin, and Daphne Koller. Max-margin Markov networks. *Advances in Neural Information Processing Systems*, 16, 2003.

Ben Taskar, Dan Klein, Mike Collins, Daphne Koller, and Christopher D Manning. Max-margin parsing. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 1–8, 2004.

Flemming Topsøe. Information-theoretical optimization techniques. *Kybernetika*, 15(1):8–27, 1979.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 252–259, 2003.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, Yasemin Altun, and Yoram Singer. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6(9), 2005.

John Von Neumann and Oskar Morgenstern. Theory of games and economic behavior, 2nd rev. 1947.

Wenhui Wang and Baobao Chang. Graph-based dependency parsing with bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2306–2315, 2016.

T William. Tutte. graph theory. *Encyclopedia of Mathematics and its Applications*, 21, 1984.

Richard T Wong. Integer programming formulations of the traveling salesman problem. In *Proceedings of the IEEE international Conference of Circuits and Computers*, pages 149–152. IEEE Press Piscataway NJ, 1980.

Zheng Xu, Gavin Taylor, Hao Li, Mário AT Figueiredo, Xiaoming Yuan, and Tom Goldstein. Adaptive consensus ADMM for distributed optimization. In *International Conference on Machine Learning*, pages 3841–3850. PMLR, 2017.

Nianwen Xue, Fu-Dong Chiou, and Martha Palmer. Building a large-scale annotated Chinese corpus. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

Hui Zhang and Lizhi Cheng. Restricted strong convexity and its applications to convergence analysis of gradient-type methods in convex optimization. *Optimization Letters*, 9(5):961–979, 2015.

Xingxing Zhang, Jianpeng Cheng, and Mirella Lapata. Dependency parsing as head selection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 665–676, 2017.

Xinhua Zhang, Ankan Saha, and SVN Vishwanathan. Regularized risk minimization by Nesterov's accelerated gradient methods: Algorithmic extensions and empirical studies. *arXiv preprint arXiv:1011.0472*, 2010.

Yu Zhang, Zhenghua Li, and Min Zhang. Efficient second-order TreeCRF for neural dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3295–3305, 2020.

Yuan Zhang, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Greed is good if randomized: New inference for dependency parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1013–1024, 2014.

Ran Zmigrod, Tim Vieira, and Ryan Cotterell. Please mind the root: Decoding arborescences for dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4809–4819, 2020.

Ran Zmigrod, Tim Vieira, and Ryan Cotterell. Efficient computation of expectations under spanning tree distributions. *Transactions of the Association for Computational Linguistics*, 9:675–690, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Technical Proofs

Proposition 1. *The distributionally robust tree structured prediction problem based on moment divergence in Eq.* (1) *can be rewritten as*

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}^{emp}_{\boldsymbol{X},\boldsymbol{Y}}} \underbrace{\min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}},\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^\intercal(\boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \boldsymbol{\phi}(\boldsymbol{X}, \boldsymbol{Y})) + \varepsilon\|\boldsymbol{\theta}\|_*}_{\ell_{adv}(\boldsymbol{\theta},(\boldsymbol{X},\boldsymbol{Y}))},$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ is the vector of Lagrangian multipliers and $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

Proof. Recall the primal problem

$$\min_{\mathbb{P}} \max_{\mathbb{Q}\in\mathcal{B}(\mathbb{P}^{\text{emp}})} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{X},\check{\boldsymbol{Y}}}\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}),$$

where $\mathcal{B}(\mathbb{P}^{\text{emp}}) := \{\mathbb{Q} : \mathbb{Q}_{\boldsymbol{X}} = \mathbb{P}^{\text{emp}}_{\boldsymbol{X}} \wedge \|\mathbb{E}_{\mathbb{P}^{\text{emp}}}\boldsymbol{\phi}(\cdot) - \mathbb{E}_{\mathbb{Q}}\boldsymbol{\phi}(\cdot)\| \le \varepsilon\}$ with $\varepsilon \ge 0$.

Note the feature function $\boldsymbol{\phi}(\cdot)$ is fixed and given. Since $\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}} \in \Delta$ and $\mathbb{Q}_{\boldsymbol{X},\check{\boldsymbol{Y}}} \in \Delta \cap \mathcal{B}(\mathbb{P}^{\text{emp}})$ where Δ is the probability simplex with dimension omitted, the constraint sets are convex. The objective function is convex in $\mathbb{P}$ and concave in $\mathbb{Q}$ because it is affine in both. Therefore strong duality holds:

$$\max_{\mathbb{Q}\in\mathcal{B}(\mathbb{P}^{\text{emp}})} \min_{\mathbb{P}} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{X},\check{\boldsymbol{Y}}}\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}).$$

Let $\mathcal{C} := \{\boldsymbol{u} : \|\boldsymbol{u} - \mathbb{E}_{\mathbb{P}^{\text{emp}}}\boldsymbol{\phi}(\cdot)\| \le \varepsilon\}$. Rewrite the problem with this constraint:

$$\begin{aligned} &\sup_{\mathbb{Q},\boldsymbol{u}} \min_{\mathbb{P}} \mathbb{E}_{\mathbb{P}^{\text{emp}}_{\boldsymbol{X}}\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) - I_{\mathcal{C}}(\boldsymbol{u}) \\ \text{s.t.} \quad &\boldsymbol{u} = \mathbb{E}_{\mathbb{P}^{\text{emp}}_{\boldsymbol{X}}\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}} \boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}), \end{aligned}$$

where $I_{\mathcal{C}}(\cdot)$ is the indicator function with $I_{\mathcal{C}}(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \in \mathcal{C}$ and $+\infty$ otherwise. The simplex constraints are omitted.

The dual problem by relaxing the equality constraint is

$$\sup_{\mathbb{Q},\boldsymbol{u}} \min_{\boldsymbol{\theta}} \min_{\mathbb{P}} \mathbb{E}_{\mathbb{P}^{\text{emp}}_{\boldsymbol{X}}\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) - I_{\mathcal{C}}(\boldsymbol{u}) + \boldsymbol{\theta}^\intercal \mathbb{E}_{\mathbb{P}^{\text{emp}}_{\boldsymbol{X}}\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}} \boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \boldsymbol{\theta}^\intercal \boldsymbol{u},$$

where $\boldsymbol{\theta}$ is the vector of Lagrange multipliers.

Given $\boldsymbol{X} = \boldsymbol{x}$, optimization of $\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{x}}$ and $\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{x}}$ can be done independently. Again by strong duality, we can rearrange the terms:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}^{\text{emp}}_{\boldsymbol{X}}} \min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^\intercal \boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}) + \sup_{\boldsymbol{u}} -I_{\mathcal{C}}(\boldsymbol{u}) - \boldsymbol{\theta}^\intercal \boldsymbol{u}.$$

The associated dual norm $\|\cdot\|_*$ of the norm $\|\cdot\|$ is defined as

$$\|\boldsymbol{z}\|_* := \sup\{\boldsymbol{z}^\intercal \boldsymbol{x} : \|\boldsymbol{x}\| \le 1\},$$

based on which we are able to simplify the optimization over $\boldsymbol{u}$ as

$$\sup_{\boldsymbol{u}} -I_{\mathcal{C}}(\boldsymbol{u}) - \boldsymbol{\theta}^\intercal \boldsymbol{u} = \sup_{\boldsymbol{u}\in\mathcal{C}} -\boldsymbol{\theta}^\intercal \boldsymbol{u} = \sup_{\boldsymbol{e}:\|\boldsymbol{e}\|\le 1} -\boldsymbol{\theta}^\intercal(\mathbb{E}_{\mathbb{P}^{\text{emp}}}\boldsymbol{\phi}(\cdot) - \varepsilon\boldsymbol{e}) = -\boldsymbol{\theta}^\intercal \mathbb{E}_{\mathbb{P}^{\text{emp}}}\boldsymbol{\phi}(\cdot) + \varepsilon\|\boldsymbol{\theta}\|_*.$$

Plugging it back to the dual problem, we have

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}^{\text{emp}}_{\boldsymbol{X},\boldsymbol{Y}}} \min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^\intercal(\boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \boldsymbol{\phi}(\boldsymbol{X}, \boldsymbol{Y})) + \varepsilon\|\boldsymbol{\theta}\|_*.$$

□

Theorem 2. *Given m samples, a non-negative loss $\ell(\cdot,\cdot)$ such that $|\ell(\cdot,\cdot)| \le K$, a feature function $\boldsymbol{\phi}(\cdot,\cdot)$ such that $\|\boldsymbol{\phi}(\cdot,\cdot)\| \le B$, a positive ambiguity level $\varepsilon > 0$, then, for any $\rho \in (0, 1]$, with a probability at least $1 - \rho$, the following excess true worst-case risk bound holds:*

$$\max_{\mathbb{Q}\in\mathcal{B}(\mathbb{P}^{true})} R^L_{\mathbb{Q}}(\boldsymbol{\theta}^*_{emp}) - \max_{\mathbb{Q}\in\mathcal{B}(\mathbb{P}^{true})} R^L_{\mathbb{Q}}(\boldsymbol{\theta}^*_{true}) \le \frac{4KB}{\varepsilon\sqrt{m}} \left(1 + \frac{3}{2}\sqrt{\frac{\ln(4/\rho)}{2}}\right),$$

where $\boldsymbol{\theta}^*_{emp}$ and $\boldsymbol{\theta}^*_{true}$ are the optimal parameters learned in Eq. (2) *under $\mathbb{P}^{emp}$ and $\mathbb{P}^{true}$ respectively. The original risk of $\boldsymbol{\theta}$ under $\mathbb{Q}$ is $R^L_{\mathbb{Q}}(\boldsymbol{\theta}) := \mathbb{E}_{\mathbb{Q}_{\boldsymbol{X},\boldsymbol{Y}},\mathbb{P}^{\boldsymbol{\theta}}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \boldsymbol{Y})$ with Bayes prediction $\mathbb{P}^{\boldsymbol{\theta}}_{\boldsymbol{Y}|\boldsymbol{x}} \in$ $\arg\min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{x}}\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{x}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^{\intercal}\boldsymbol{\phi}(\boldsymbol{x}, \check{\boldsymbol{Y}})$.*

Proof. Define the adversarial surrogate risk of $\boldsymbol{\theta}$ with respect to $\tilde{\mathbb{P}}$ as
$R^S_{\tilde{\mathbb{P}}}(\boldsymbol{\theta}) := \mathbb{E}_{\tilde{\mathbb{P}}_{\boldsymbol{X},\boldsymbol{Y}}} \ell_{\text{adv}}(\boldsymbol{\theta}, (\boldsymbol{X}, \boldsymbol{Y})) := \mathbb{E}_{\tilde{\mathbb{P}}_{\boldsymbol{X},\boldsymbol{Y}}} \min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^{\intercal}(\boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \boldsymbol{\phi}(\boldsymbol{X}, \boldsymbol{Y})) + \varepsilon\|\boldsymbol{\theta}\|_*$.

Let $\boldsymbol{\theta}^*_{\text{true}} \in \arg\min_{\boldsymbol{\theta}} R^S_{\mathbb{P}^{\text{true}}}(\boldsymbol{\theta})$ and $\boldsymbol{\theta}^*_{\text{emp}} \in \arg\min_{\boldsymbol{\theta}} R^S_{\mathbb{P}^{\text{emp}}}(\boldsymbol{\theta})$ be the optimal parameters learned with $\mathbb{P}^{\text{true}}_{\mathbf{X},\boldsymbol{Y}}$ and $\mathbb{P}^{\text{emp}}_{\mathbf{X},\boldsymbol{Y}}$ respectively.

Given $\boldsymbol{x}$, define the decoded prediction by $\boldsymbol{\theta}$ as

$$\mathbb{P}^{\boldsymbol{\theta}}_{\boldsymbol{Y}|\boldsymbol{x}} \in \arg\min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{x}}\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{x}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^{\intercal}\boldsymbol{\phi}(\boldsymbol{x}, \check{\boldsymbol{Y}}).$$

Let the original risk of loss ℓ under some distribution $\mathbb{Q}$ be

$$R^L_{\mathbb{Q}}(\boldsymbol{\theta}) := \mathbb{E}_{\mathbb{Q}_{\boldsymbol{X},\boldsymbol{Y}},\mathbb{P}^{\boldsymbol{\theta}}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \boldsymbol{Y}).$$

According to Proposition 1, for any fixed $\mathbb{P}$, we have similarly

$$\max_{\mathbb{Q}\in\mathcal{B}(\mathbb{P}^{\text{emp}})} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{X},\check{\boldsymbol{Y}}}\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) \triangleq \min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}^{\text{emp}}_{\boldsymbol{X},\boldsymbol{Y}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^{\intercal}(\boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \boldsymbol{\phi}(\boldsymbol{X}, \boldsymbol{Y})) + \varepsilon\|\boldsymbol{\theta}\|_*.$$

We start by looking at the worst-case risk of $\boldsymbol{\theta}^*_{\text{true}}$ and $\boldsymbol{\theta}^*_{\text{emp}}$.

$$\begin{aligned}
&\max_{\mathbb{Q}\in\mathcal{B}(\mathbb{P}^{\text{true}})} R^L_{\mathbb{Q}}(\boldsymbol{\theta}^*_{\text{emp}}) \\
=&\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}^{\text{true}}_{\boldsymbol{X},\boldsymbol{Y}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}^{\boldsymbol{\theta}^*_{\text{emp}}}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^{\intercal}(\boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \boldsymbol{\phi}(\boldsymbol{X}, \boldsymbol{Y})) + \varepsilon\|\boldsymbol{\theta}\|_* \\
\leq&\mathbb{E}_{\mathbb{P}^{\text{true}}_{\boldsymbol{X},\boldsymbol{Y}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}^{\boldsymbol{\theta}^*_{\text{emp}}}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^*_{\text{emp}} \cdot (\boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \boldsymbol{\phi}(\boldsymbol{X}, \boldsymbol{Y})) + \varepsilon\|\boldsymbol{\theta}^*_{\text{emp}}\|_*,
\end{aligned}$$

where the last inequality holds because $\boldsymbol{\theta}^*_{\text{emp}}$ is not necessarily a minimizer. Similarly for $\boldsymbol{\theta}^*_{\text{true}}$,

$$\max_{\mathbb{Q}\in\mathcal{B}(\mathbb{P}^{\text{true}})} R^L_{\mathbb{Q}}(\boldsymbol{\theta}^*_{\text{true}}) \leq \mathbb{E}_{\mathbb{P}^{\text{true}}_{\boldsymbol{X},\boldsymbol{Y}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}^{\boldsymbol{\theta}^*_{\text{true}}}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^*_{\text{true}} \cdot (\boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \boldsymbol{\phi}(\boldsymbol{X}, \boldsymbol{Y})) + \varepsilon\|\boldsymbol{\theta}^*_{\text{true}}\|_*.$$

On the other hand,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}^{\text{true}}_{\boldsymbol{X},\boldsymbol{Y}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}^{\boldsymbol{\theta}^*_{\text{true}}}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^*_{\text{true}} \cdot (\boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \boldsymbol{\phi}(\boldsymbol{X}, \boldsymbol{Y})) + \varepsilon\|\boldsymbol{\theta}^*_{\text{true}}\|_* \\
=&\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}^{\text{true}}_{\boldsymbol{X},\boldsymbol{Y}}} \min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^{\intercal}(\boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \boldsymbol{\phi}(\boldsymbol{X}, \boldsymbol{Y})) + \varepsilon\|\boldsymbol{\theta}\|_* \\
=&\min_{\mathbb{P}} \min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}^{\text{true}}_{\boldsymbol{X},\boldsymbol{Y}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^{\intercal}(\boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \boldsymbol{\phi}(\boldsymbol{X}, \boldsymbol{Y})) + \varepsilon\|\boldsymbol{\theta}\|_* \\
\leq&\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}^{\text{true}}_{\boldsymbol{X},\boldsymbol{Y}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}^{\boldsymbol{\theta}^*_{\text{true}}}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^{\intercal}(\boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \boldsymbol{\phi}(\boldsymbol{X}, \boldsymbol{Y})) + \varepsilon\|\boldsymbol{\theta}\|_* \\
=&\max_{\mathbb{Q}\in\mathcal{B}(\mathbb{P}^{\text{true}})} R^L_{\mathbb{Q}}(\boldsymbol{\theta}^*_{\text{true}}),
\end{aligned}$$

where the first equality holds according to the definition of $\boldsymbol{\theta}^*_{\text{true}}$. The above two inequalities imply the equality:

$$\max_{\mathbb{Q}\in\mathcal{B}(\mathbb{P}^{\text{true}})} R^L_{\mathbb{Q}}(\boldsymbol{\theta}^*_{\text{true}}) = \mathbb{E}_{\mathbb{P}^{\text{true}}_{\boldsymbol{X},\boldsymbol{Y}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}^{\boldsymbol{\theta}^*_{\text{true}}}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^*_{\text{true}} \cdot (\boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \boldsymbol{\phi}(\boldsymbol{X}, \boldsymbol{Y})) + \varepsilon\|\boldsymbol{\theta}^*_{\text{true}}\|_*.$$

Therefore,

$$\begin{aligned}
&\max_{\mathbb{Q}\in\mathcal{B}(\mathbb{P}^{\text{true}})} R^L_{\mathbb{Q}}(\boldsymbol{\theta}^*_{\text{emp}}) - \max_{\mathbb{Q}\in\mathcal{B}(\mathbb{P}^{\text{true}})} R^L_{\mathbb{Q}}(\boldsymbol{\theta}^*_{\text{true}}) \\
\leq&\mathbb{E}_{\mathbb{P}^{\text{true}}_{\boldsymbol{X},\boldsymbol{Y}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}^{\boldsymbol{\theta}^*_{\text{emp}}}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^*_{\text{emp}} \cdot (\boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \boldsymbol{\phi}(\boldsymbol{X}, \boldsymbol{Y})) + \varepsilon\|\boldsymbol{\theta}^*_{\text{emp}}\|_* \\
&- (\mathbb{E}_{\mathbb{P}^{\text{true}}_{\boldsymbol{X},\boldsymbol{Y}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}^{\boldsymbol{\theta}^*_{\text{true}}}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^*_{\text{true}} \cdot (\boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \boldsymbol{\phi}(\boldsymbol{X}, \boldsymbol{Y})) + \varepsilon\|\boldsymbol{\theta}^*_{\text{true}}\|_*). \qquad (5)
\end{aligned}$$

The main idea is thus to use uniform convergence bounds. Firstly, by substituting $\mathbb{Q} = \mathbb{P}^{\text{true}}$, note that

$$\min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^{\intercal}(\boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \boldsymbol{\phi}(\boldsymbol{X}, \boldsymbol{Y})) \geq \min_{\mathbb{P}} \mathbb{E}_{\mathbb{P}^{\text{true}}_{\boldsymbol{Y}|\boldsymbol{X}}\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \boldsymbol{Y}) \geq 0.$$

We can get an upper bound of the norm of any optimal solution $\boldsymbol{\theta}^*_{\text{true}}$ or $\boldsymbol{\theta}^*_{\text{emp}}$ as follows:

$$0 + \varepsilon \|\boldsymbol{\theta}^*_{\text{true}}\|_* \leq R^S_{\mathbb{P}^{\text{true}}}(\boldsymbol{\theta}^*_{\text{true}}) \leq R^S_{\mathbb{P}^{\text{true}}}(\mathbf{0}) \leq \mathbb{E}_{\mathbb{P}^{\text{true}}_{\boldsymbol{X},\boldsymbol{Y}}} \min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) \leq K \implies \|\boldsymbol{\theta}^*_{\text{true}}\|_* \leq \frac{K}{\varepsilon}.$$

Let $\psi(\boldsymbol{X}, \boldsymbol{Y}) := \boldsymbol{\theta}^{\intercal}\boldsymbol{\phi}(\boldsymbol{X}, \boldsymbol{Y})$ and $\boldsymbol{\psi}_{\boldsymbol{x}} := (\psi(\boldsymbol{x}, \boldsymbol{y}))_{\boldsymbol{y} \in \mathcal{Y}}$. Define

$$\begin{aligned} f(\boldsymbol{\theta}, \tilde{\mathbb{P}}) &:= \mathbb{E}_{\tilde{\mathbb{P}}_{\boldsymbol{X},\boldsymbol{Y}}} \min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^{\intercal}(\boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \boldsymbol{\phi}(\boldsymbol{X}, \boldsymbol{Y})) \\ &\triangleq \mathbb{E}_{\tilde{\mathbb{P}}_{\boldsymbol{X},\boldsymbol{Y}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}^{\boldsymbol{\theta}}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \boldsymbol{\theta}^{\intercal}(\boldsymbol{\phi}(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \boldsymbol{\phi}(\boldsymbol{X}, \boldsymbol{Y})) \\ &\triangleq \mathbb{E}_{\tilde{\mathbb{P}}_{\boldsymbol{X},\boldsymbol{Y}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}\mathbb{P}^{\boldsymbol{\theta}}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}}) + \psi(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \psi(\boldsymbol{X}, \boldsymbol{Y}) \\ &\triangleq g(\boldsymbol{\psi}, \tilde{\mathbb{P}}). \end{aligned}$$

Let $\boldsymbol{q}_{\boldsymbol{x}} \in \Delta$ be the probability vector of $\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{x}}$ and $\boldsymbol{e}_{\boldsymbol{y}}$ be the standard basis vector with $\boldsymbol{y}$-th entry equal to 1. We have that for any $(\boldsymbol{x}, \boldsymbol{y})$,

$$\frac{\partial}{\partial \boldsymbol{\psi}_{\boldsymbol{x}}} g(\boldsymbol{\psi}, \delta_{(\boldsymbol{x},\boldsymbol{y})}) \subseteq \text{Conv}(\{\boldsymbol{q}_{\boldsymbol{x}} - \boldsymbol{e}_{\boldsymbol{y}} : \boldsymbol{q}_{\boldsymbol{x}} \in \Delta\}) \implies \|\frac{\partial}{\partial \boldsymbol{\psi}_{\boldsymbol{x}}} g(\boldsymbol{\psi}, \delta_{(\boldsymbol{x},\boldsymbol{y})})\|_1 \leq \max_{\boldsymbol{q}_{\boldsymbol{x}} \in \Delta} \|\boldsymbol{q}_{\boldsymbol{x}} - \boldsymbol{e}_{\boldsymbol{y}}\|_1 \leq 2,$$

where $\delta_{(\boldsymbol{x},\boldsymbol{y})}$ is the Dirac point measure. $g(\cdot, \tilde{\mathbb{P}})$ is therefore 2-Lipschitz with respect to the ℓ_1 norm. As per the assumption, $\|\boldsymbol{\phi}(\cdot, \cdot)\| \leq B$. This further implies that

$$f(\boldsymbol{\theta}_1, \delta_{(\boldsymbol{x}_1,\boldsymbol{y}_1)}) - f(\boldsymbol{\theta}_2, \delta_{(\boldsymbol{x}_2,\boldsymbol{y}_2)}) \leq \frac{4KB}{\varepsilon} \quad \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2 \quad \text{s.t.} \quad \|\boldsymbol{\theta}_i\|_* \leq \frac{K}{\varepsilon} \quad \forall i = 1, 2.$$

We then follow the proof of Theorem 3 in Farnia and Tse [2016]. According to Theorem 26.12 in Shalev-Shwartz and Ben-David [2014], by uniform convergence, for any $\rho \in (0, 2]$, with a probability at least $1 - \frac{\rho}{2}$,

$$f(\boldsymbol{\theta}^*_{\text{emp}}, \mathbb{P}^{\text{true}}) - f(\boldsymbol{\theta}^*_{\text{emp}}, \mathbb{P}^{\text{emp}}) \leq \frac{4KB}{\varepsilon\sqrt{m}} \left(1 + \sqrt{\frac{\ln(4/\rho)}{2}}\right).$$

According to the definition of $\boldsymbol{\theta}^*_{\text{true}}$, the following inequality holds:

$$f(\boldsymbol{\theta}^*_{\text{emp}}, \mathbb{P}^{\text{emp}}) + \varepsilon\|\boldsymbol{\theta}^*_{\text{emp}}\|_* - f(\boldsymbol{\theta}^*_{\text{true}}, \mathbb{P}^{\text{emp}}) - \varepsilon\|\boldsymbol{\theta}^*_{\text{true}}\|_* \leq 0.$$

Since $\boldsymbol{\theta}^*_{\text{true}}$ do not depend on samples, according to the Hoeffding's inequality, with a probability $1 - \rho/2$,

$$f(\boldsymbol{\theta}^*_{\text{true}}, \mathbb{P}^{\text{emp}}) - f(\boldsymbol{\theta}^*_{\text{true}}, \mathbb{P}^{\text{true}}) \leq \frac{2KB}{\varepsilon\sqrt{m}} \sqrt{\frac{\ln(4/\rho)}{2}}.$$

Applying the union bound to the above three inequations, with a probability $1 - \rho$, we have

$$f(\boldsymbol{\theta}^*_{\text{emp}}, \mathbb{P}^{\text{true}}) + \varepsilon\|\boldsymbol{\theta}^*_{\text{emp}}\|_* - f(\boldsymbol{\theta}^*_{\text{true}}, \mathbb{P}^{\text{true}}) - \varepsilon\|\boldsymbol{\theta}^*_{\text{true}}\|_* \leq \frac{4KB}{\varepsilon\sqrt{m}} \left(1 + \frac{3}{2}\sqrt{\frac{\ln(4/\rho)}{2}}\right).$$

As stated by Inequation (5), we conclude with the following excess risk bound:

$$\max_{\mathbb{Q} \in \mathcal{B}(\mathbb{P}^{\text{true}})} R^L_{\mathbb{Q}}(\boldsymbol{\theta}^*_{\text{emp}}) - \max_{\mathbb{Q} \in \mathcal{B}(\mathbb{P}^{\text{true}})} R^L_{\mathbb{Q}}(\boldsymbol{\theta}^*_{\text{true}}) \leq \frac{4KB}{\varepsilon\sqrt{m}} \left(1 + \frac{3}{2}\sqrt{\frac{\ln(4/\rho)}{2}}\right).$$

□

Corollary 3. *When $\varepsilon = 0$, ℓ_{adv} is Fisher consistent with respect to ℓ. Namely,*

$$\mathbb{P}^{\boldsymbol{\theta}^*_{true}}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}} \in \arg\min_{\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \mathbb{E}_{\mathbb{P}^{true}_{\boldsymbol{X},\boldsymbol{Y}},\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}} \ell(\hat{\boldsymbol{Y}}, \boldsymbol{Y}),$$

where $\boldsymbol{\theta}^*_{true}$ is learned with ℓ_{adv} and $\mathbb{P}^{true}$ as in Theorem 2.

Proof. Our formulation differs from Nowak-Vila et al. [2020] in the fact that we allow probabilistic prediction to be ground truth. By defining $y^*(\mu)$ as the gold standard probabilistic prediction and $\mathcal{Y}$ as the set of all possible probabilistic predictions in Proposition C.2 in Nowak-Vila et al. [2020], we have

$$\mathbb{P}^{\boldsymbol{\theta}^*_{\text{true}}}_{\hat{\boldsymbol{Y}}|\boldsymbol{x}} \in \text{Conv}(\arg\min_{\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{x}}} \mathbb{E}_{\mathbb{P}^{\text{true}}_{\boldsymbol{Y}|\boldsymbol{x}},\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{x}}} \ell(\hat{\boldsymbol{Y}}, \boldsymbol{Y})).$$

Therefore,

$$\mathbb{P}^{\boldsymbol{\theta}^*_{\text{true}}}_{\hat{\boldsymbol{Y}}|\boldsymbol{x}} \in \arg\min_{\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{x}}} \mathbb{E}_{\mathbb{P}^{\text{true}}_{\boldsymbol{Y}|\boldsymbol{x}},\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{x}}} \ell(\hat{\boldsymbol{Y}}, \boldsymbol{Y}).$$

□

Proposition 4. *Let $\mathcal{G}$ be a multi-graph. $\mathcal{A}_{marb} \triangleq \mathcal{A}_{arb}$.*

Proof. We follow the proof of Friesen [2019] for simple graphs. Recall the definition of $\mathcal{A}_{\text{marb}}$:

$$\begin{aligned}
&\mathcal{A}_{\text{marb}} := \{\boldsymbol{z}^r : \exists \boldsymbol{z} \geq \boldsymbol{0} \\
&\sum_{a \in \delta^-(j)} z^k_a = \mathbb{1}(j \neq k)\ \forall k, j \in \mathcal{V} \wedge \qquad (6) \\
&\sum_{a \in \mathcal{E}'_{ij}} z^k_a = \sum_{a \in \mathcal{E}_{ij}} z^r_a \quad \forall k \neq r, i, j \in \mathcal{V}\}. \qquad (7)
\end{aligned}$$

On one hand, given a legal r-arborescence with characteristic vector $\boldsymbol{z}^r$, Eq. (6) and Eq. (7) hold by the definition of arborescences. The equality also holds for a convex combination of the characteristic vectors of r-arborescences.

On the other hand, given $\boldsymbol{z} \in \mathcal{A}_{\text{marb}}$. Consider Edmond's definition of r-arborescence polytope based on rank constraints:

$$\begin{aligned}
&\sum_{a \in S} x_a \leq |S| - 1 \quad \forall S \subset \mathcal{V} \text{ with } S \neq \emptyset \qquad (8) \\
&\sum_{a \in \delta^-(j)} x_a = \mathbb{1}(j \neq r)\ \forall j \in \mathcal{V} \qquad (9) \\
&\boldsymbol{x} \geq \boldsymbol{0}.
\end{aligned}$$

We have Eq. (6) directly implies Eq. (9). According to Eq. (7),

$$\sum_{a \in S} z^r_a = \sum_{a \in S} z^u_a \quad \forall S \subseteq \mathcal{V} \wedge u \in \mathcal{V}.$$

Therefore,

$$\sum_{a \in S} z^r_a = \sum_{a \in S} z^u_a \leq \sum_{j \in S} \sum_{a \in \delta^-(j)} z^u_a = |S| - 1 \quad \forall S \subseteq \mathcal{V} \wedge u \in S,$$

which is exactly Eq. (8). □

Proposition 5. *Let $\mathcal{G}$ be a multi-graph. $\mathcal{A}_{mdep} \triangleq \mathcal{A}_{dep}$.*

Proof. Recall the definition of $\mathcal{A}_{\text{mdep}}$:

$$\begin{aligned}
\mathcal{A}_{\text{mdep}} := \{&\boldsymbol{z}^r : \boldsymbol{z}^r \in \mathcal{A}_{\text{marb}} \wedge \\
&\sum_{a \in \delta^+(r)} z^r_a = 1\}. \qquad (10)
\end{aligned}$$

Algorithm 1 Double Oracle Game Solver

```
Input: Lagrange multipliers $\boldsymbol{\theta}$; feature function $\boldsymbol{\phi}(\cdot,\cdot)$; initial set of trees $\{\boldsymbol{y}_{\text{initial}}\}$
Output: A sparse Nash equilibrium $(\hat{\mathcal{T}}, \check{\mathcal{T}}, \mathbb{P}, \mathbb{Q})$
Initialize $\hat{\mathcal{T}} \leftarrow \check{\mathcal{T}} \leftarrow \{\boldsymbol{y}_{\text{initial}}\}$
repeat
    $(\mathbb{P}, \hat{v}_{\text{Nash}}) \leftarrow \text{SolveZeroSumGame}_{\hat{\mathcal{T}}}(\ell, \boldsymbol{\theta}^{\intercal}\boldsymbol{\phi}, \hat{\mathcal{T}}, \check{\mathcal{T}})$
    $(\check{\boldsymbol{y}}_{\text{BR}}, \check{v}_{\text{BR}}) \leftarrow \text{FindBestResponse}(\ell, \boldsymbol{\theta}^{\intercal}\boldsymbol{\phi}, \mathbb{P}, \hat{\mathcal{T}})$
    if $\hat{v}_{\text{Nash}} \neq \check{v}_{\text{BR}}$ then
        $\check{\mathcal{T}} \leftarrow \check{\mathcal{T}} \cup \{\check{\boldsymbol{y}}_{\text{BR}}\}$
    end if
    $(\mathbb{Q}, \check{v}_{\text{Nash}}) \leftarrow \text{SolveZeroSumGame}_{\check{\mathcal{T}}}(\ell, \boldsymbol{\theta}^{\intercal}\boldsymbol{\phi}, \hat{\mathcal{T}}, \check{\mathcal{T}})$
    $(\hat{\boldsymbol{y}}_{\text{BR}}, \hat{v}_{\text{BR}}) \leftarrow \text{FindBestResponse}(\ell, \boldsymbol{\theta}^{\intercal}\boldsymbol{\phi}, \mathbb{Q}, \check{\mathcal{T}})$
    if $\check{v}_{\text{Nash}} \neq \hat{v}_{\text{BR}}$ then
        $\hat{\mathcal{T}} \leftarrow \hat{\mathcal{T}} \cup \{\hat{\boldsymbol{y}}_{\text{BR}}\}$
    end if
until $\hat{v}_{\text{Nash}} = \check{v}_{\text{BR}} = \check{v}_{\text{Nash}} = \hat{v}_{\text{BR}}$
return $(\hat{\mathcal{T}}, \check{\mathcal{T}}, \mathbb{P}, \mathbb{Q})$
```

On one hand, given a legal dependency tree $\boldsymbol{z}^r \in \mathcal{A}_{\text{dep}}$, it satisfies Eq. (6) and Eq. (7) by Proposition 4. It also satisfies Eq. (10) by the definition of $\mathcal{A}_{\text{dep}}$.

On the other hand, given $\boldsymbol{z}^r \in \mathcal{A}_{\text{mdep}}$, firstly, $\boldsymbol{z}^r$ must be in $\mathcal{A}_{\text{arb}}$ by Proposition 4, which implies that we can write it as a convex combination of k r-arborescences vectors: $\boldsymbol{z}^r \triangleq \alpha_1 \boldsymbol{t}^1 + \alpha_2 \boldsymbol{t}^2 + \cdots + \alpha_k \boldsymbol{t}^k$. All of them are legal r-arborescences, so $\sum_{a \in \delta^+(r)} t_a^i \geq 1$ for all $i \in [k]$. Now if $\sum_{a \in \delta^+(r)} t_a^i > 1$ for some i, we would have a contradiction, $\sum_{a \in \delta^+(r)} z_a^r > 1$. □

B Algorithm Details

The pseudo-code of the constraint generation algorithm proposed in Section 3.2 is illustrated in Algorithm 1.

C More on Experiments

We adopt three public datasets, the English Penn Treebank (PTB v3.0) [Marcus et al., 1993], the Penn Chinese Treebank (CTB v5.1) [Xue et al., 2002], the Dutch Lassy Small Treebank and the Turkish Treebank in Universal Dependencies (UD v2.3) [Nivre et al., 2016]. We follow conventions in Chen and Manning [2014], Dyer et al. [2015] to prepare our data. We make standard train/validation/test splits. We use Stanford Dependencies (SD v3.3.0) [De Marneffe and Manning, 2008] to convert dependencies in PTB and CTB. The predicted POS tags with Stanford POS tagger [Toutanova et al., 2003] are adopted for PTB whereas gold POS tags are adopted for CTB and UD. Punctuation is excluded during evaluation[6].

The pretrained models are trained with the suggested hyperparameters in SuPar. The pretrained models achieve 97.25%, 91.91% and 94.78% UAS on PTB, CTB and UD Dutch respectively, where RoBERTa [Liu et al., 2019], ELECTRA [Cui et al., 2020] and XLM-RoBERTa [Conneau et al., 2019] are adopted as encoders. No BERT embeddings are adopted for the UD Turkish dataset.

For our ADMM algorithm, we adopt the adaptive scheme of varying penalty parameters ($\tau_{\text{incr}} = \tau_{\text{decr}} = 1.1$, $\mu = 1$) in Boyd et al. [2011] and the stopping criterion ($\epsilon_{\text{tol}} = 10^{-2}$) for consensus ADMM in Xu et al. [2017]. In FW, the learning rate is set to $\frac{2}{t+2}$. The smoothness weight μ and ambiguity radius $\lambda = 2\varepsilon$ are tuned using a logarithmic scale on $[10^{-7}, 1]$. The batch size for the game-theoretic algorithm is 10. The batch size for *Stochastic* is 200. The error tolerance in *Game* is set to 10^{-2}. In stochastic gradient training, we use Adam with $lr = 10^{-2}$, $\beta_1 = 0.9$,

[6] A token is a punctuation if its gold POS tag is space, semi-colon, comma or period for English and PU for Chinese.

$\beta_2 = 0.999$, $\epsilon = 10^{-8}$. In our experiments, for efficiency, we again adopt the FW algorithm for the outer maximization in *Marginal*.

Complete main experimental results including all the metrics are shown in Table 2.

D Extension Details

For the dependency tree polytope, recall that the dual problem of projection onto $\mathcal{U}'_r := \{\boldsymbol{x} : \boldsymbol{x} \in \mathcal{U}_r \wedge \sum_{a\in\delta^+(r)} x_a = 1\}$ is

$$\max_{\boldsymbol{\alpha},\beta} \sum_{a\in\mathcal{E}} h_a(\boldsymbol{\alpha},\beta) - \sum_{j\neq r} \alpha_j - \beta \quad \text{s.t. } h_a(\boldsymbol{\alpha},\beta) = \begin{cases} w_a^2 & \gamma_a > 2w_a, \\ w_a\gamma_a - \gamma_a^2/4 & \gamma_a \leq 2w_a, \end{cases}$$

where $\gamma_{(i,j,l)} := \alpha_j + \mathbb{1}(i = r)\beta$. Following Zhang et al. [2010] similarly, we sort $2w_{(i,j,l)}$ for each j and compute the optimal α_j^* with $\beta = 0$. Let the sorted w's be $(w_1^{(j)}, \ldots, w_n^{(j)})$ for each j. We blend create a set $\{w_x^{(j)} - \alpha_j^*\}$ for all j and x. Let the sorted sequence be $-\infty = t_1 < t_2 < \cdots < t_{n_t} = \infty$. The derivative with respect to β is piecewise-linear in each interval $[t_k, t_{k+1}]$. Since the objective is concave in β, we can iterate over all the intervals or find the optimal β^* with binary search.

For higher-order tree local polytopes, the central problem is the projection onto

$$\mathcal{U}_s := \{\boldsymbol{x} \in \mathbb{R}_{\geq 0}^{|\mathcal{R}|} : x_s \leq x_a \quad \forall a \in s\}.$$

The only variables of interest are x_a and x_s, given x_s, the optimal x_a is simply $x_a^* = \max(w_a, x_s)$. We can sort $(w_a, w_s)_{a\in s}$ and enumerate the range x_s takes over this set.

E Wong's Arborescence Polytope

We introduce another extended formulation of the arborescence polytope based on a multi-commodity flow representation [Wong, 1980, Martins, 2012, Friesen, 2019] as follows, which may be of independent interest:

$$\sum_{a\in\delta^-(j)} x_a = \mathbb{1}(j \neq r) \quad \forall j \in \mathcal{V} \tag{11}$$

$$\sum_{a\in\delta^-(j)} f_a^k - \sum_{a\in\delta^+(j)} f_a^k = \mathbb{1}(j = k) - \mathbb{1}(j = r) \quad \forall k \in \mathcal{V}\setminus\{r\}, j \in \mathcal{V} \tag{12}$$

$$0 \leq f_a^k \leq x_a \quad \forall a \in \mathcal{E}, k \in \mathcal{V}\setminus\{r\}. \tag{13}$$

Thus we have the arborescence polytope:

$$\mathcal{A}_{\text{mc}} = \{\boldsymbol{x} \in \mathbb{R}^{|\mathcal{E}|} | \exists \boldsymbol{f} : (\boldsymbol{x}, \boldsymbol{f}) \text{ satisfy equations } (11) - (13)\}.$$

According to Martins [2012], Friesen [2019], $\mathcal{A}_{\text{mc}} \triangleq \mathcal{A}_{\text{arb}}$ instead of an outer polytope of $\mathcal{A}_{\text{arb}}$.

We are interested in the following quadratic programming problem with linear inequality constraints:

$$\min_{\boldsymbol{x}\in\mathcal{A}_{\text{mc}}} \|\boldsymbol{x} - \boldsymbol{w}\|_2^2.$$

We can reformulate the problem as

$$\begin{aligned} &\min_{\boldsymbol{x},\boldsymbol{u}} g(\boldsymbol{x},\boldsymbol{u}) := \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{w}\|_2^2 + \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{w}\|_2^2 + I_{\mathcal{X}}(\boldsymbol{x}) + I_{\mathcal{U}}(\boldsymbol{u}) \\ \text{s.t.}\quad & \boldsymbol{x} = \boldsymbol{u} \\ & \mathcal{X} := \{\boldsymbol{x} : \sum_{a\in\delta^-(j)} x_a = \mathbb{1}(j \neq r) \forall j \in \mathcal{V} \wedge x_a \geq 0 \forall a \in \mathcal{E}\} \\ & \mathcal{U} := \{\boldsymbol{u} : \exists \boldsymbol{f} \sum_{a\in\delta^-(j)} f_a^k - \sum_{a\in\delta^+(j)} f_a^k = \mathbb{1}(j = k) - \mathbb{1}(j = r) \quad \forall k \in \mathcal{V}\setminus\{r\}, j \in \mathcal{V} \\ & 0 \leq f_a^k \leq u_a \quad \forall k \in \mathcal{V}\setminus\{r\}, a \in \mathcal{E}\}. \end{aligned}$$

Table 2: Comparison of mean UAS, LAS, UCM and LCM under different training set sizes. Statistically significant differences compared to BiAF are marked with † (paired t-test, $p < 0.05$). We highlight in bold the best results among the four methods.

Dataset	# train	Metric	BiAF	Marginal	Stochastic	Game
PTB	10	UAS	93.48 ± 2.30	94.51 ± 1.71†	**94.62 ± 1.60**†	94.51 ± 1.75†
		LAS	92.02 ± 2.26	93.04 ± 1.69†	**93.14 ± 1.58**†	93.04 ± 1.73†
		UCM	47.17 ± 10.28	52.30 ± 8.71†	**52.62 ± 8.18**†	52.50 ± 8.60†
		LCM	39.73 ± 7.96	43.63 ± 6.71†	**43.97 ± 6.39**†	43.86 ± 6.58†
	50	UAS	**96.87 ± 0.06**	96.81 ± 0.05†	96.81 ± 0.05	96.86 ± 0.05
		LAS	**95.34 ± 0.06**	95.28 ± 0.05†	95.28 ± 0.05	95.33 ± 0.05
		UCM	67.65 ± 0.81	67.38 ± 0.62	67.18 ± 0.79	**67.73 ± 0.64**
		LCM	**55.46 ± 0.59**	54.93 ± 0.56†	54.79 ± 0.59†	55.17 ± 0.49
	100	UAS	**96.95 ± 0.05**	96.92 ± 0.06	96.93 ± 0.05	96.92 ± 0.03
		LAS	**95.42 ± 0.05**	95.39 ± 0.06	95.40 ± 0.04	95.39 ± 0.02
		UCM	**68.79 ± 0.42**	68.27 ± 0.72	68.36 ± 0.41	68.29 ± 0.34
		LCM	**56.21 ± 0.14**	55.68 ± 0.56	55.67 ± 0.45	55.66 ± 0.33
	1000	UAS	**97.16 ± 0.02**	97.12 ± 0.03	97.14 ± 0.02	97.08 ± 0.03†
		LAS	**95.63 ± 0.03**	95.59 ± 0.02	95.60 ± 0.02	95.55 ± 0.03†
		UCM	**70.99 ± 0.23**	70.59 ± 0.49	70.61 ± 0.32	69.94 ± 0.34†
		LCM	**57.57 ± 0.09**	57.18 ± 0.28†	57.24 ± 0.28†	56.80 ± 0.23†
CTB	10	UAS	88.45 ± 0.67	89.19 ± 0.38†	**89.27 ± 0.33**†	89.22 ± 0.39†
		LAS	84.79 ± 0.62	85.50 ± 0.35†	**85.58 ± 0.30**†	85.53 ± 0.36†
		UCM	35.21 ± 1.67	36.83 ± 1.20	**37.14 ± 0.94**†	36.95 ± 1.23†
		LCM	25.86 ± 0.87	26.82 ± 0.62	**26.95 ± 0.59**†	26.95 ± 0.63†
	50	UAS	90.89 ± 0.10	91.03 ± 0.05†	91.03 ± 0.05†	**91.06 ± 0.05**†
		LAS	87.08 ± 0.10	87.20 ± 0.05†	87.20 ± 0.05†	**87.23 ± 0.06**†
		UCM	42.54 ± 0.24	42.92 ± 0.24†	42.86 ± 0.12†	**42.99 ± 0.30**
		LCM	29.70 ± 0.23	29.69 ± 0.36	29.72 ± 0.38	**29.79 ± 0.23**
	100	UAS	91.15 ± 0.16	91.27 ± 0.08	**91.27 ± 0.10**	91.22 ± 0.05
		LAS	87.32 ± 0.14	87.42 ± 0.06	**87.42 ± 0.08**	87.37 ± 0.05
		UCM	43.41 ± 0.35	**43.91 ± 0.27**†	43.86 ± 0.43†	43.81 ± 0.22
		LCM	30.02 ± 0.22	**30.27 ± 0.25**	30.23 ± 0.28	30.26 ± 0.26
	1000	UAS	**91.70 ± 0.04**	91.67 ± 0.03	91.66 ± 0.03	91.57 ± 0.03†
		LAS	**87.84 ± 0.04**	87.80 ± 0.03	87.79 ± 0.03	87.70 ± 0.03†
		UCM	**45.80 ± 0.27**	45.43 ± 0.11†	45.41 ± 0.12†	45.36 ± 0.27†
		LCM	31.14 ± 0.19	31.11 ± 0.18	31.08 ± 0.17	**31.20 ± 0.11**
UD Dutch	10	UAS	90.86 ± 1.23	**92.41 ± 0.94**†	92.40 ± 0.91†	92.32 ± 1.03†
		LAS	86.54 ± 1.26	**88.10 ± 0.95**†	88.08 ± 0.91†	87.99 ± 1.00†
		UCM	64.11 ± 2.18	**67.26 ± 2.16**†	67.21 ± 1.91†	67.26 ± 1.97†
		LCM	48.33 ± 1.88	50.32 ± 1.75†	**50.48 ± 1.45**†	50.46 ± 1.30†
	50	UAS	93.80 ± 0.43	94.22 ± 0.26†	94.23 ± 0.18†	**94.34 ± 0.24**†
		LAS	89.36 ± 0.33	89.79 ± 0.21†	89.79 ± 0.12†	**89.89 ± 0.18**†
		UCM	70.57 ± 1.52	72.42 ± 0.90†	72.05 ± 0.99	**72.60 ± 1.39**
		LCM	52.40 ± 0.61	53.47 ± 0.62†	53.40 ± 0.59	**53.58 ± 0.76**
	100	UAS	94.15 ± 0.18	94.50 ± 0.18†	94.47 ± 0.13	**94.59 ± 0.12**†
		LAS	89.69 ± 0.18	90.04 ± 0.15†	90.01 ± 0.12	**90.12 ± 0.10**†
		UCM	71.71 ± 0.92	73.24 ± 0.88†	73.01 ± 0.99	**73.63 ± 0.75**†
		LCM	53.01 ± 0.81	53.79 ± 0.40	53.70 ± 0.55	**54.13 ± 0.44**†
	1000	UAS	94.98 ± 0.07	**95.15 ± 0.10**†	95.14 ± 0.11†	95.01 ± 0.05
		LAS	90.44 ± 0.06	**90.59 ± 0.08**†	90.59 ± 0.08†	90.44 ± 0.06
		UCM	74.73 ± 0.33	**75.87 ± 0.63**†	75.64 ± 0.57†	75.41 ± 0.56
		LCM	54.59 ± 0.13	**55.21 ± 0.17**†	55.16 ± 0.21†	54.70 ± 0.22
UD Turkish	10	UAS	17.64 ± 2.45	24.85 ± 2.35†	**25.06 ± 0.58**†	19.85 ± 0.46
		LAS	4.86 ± 2.74	5.33 ± 2.97	**5.40 ± 2.85**	5.02 ± 3.04
		UCM	7.69 ± 1.72	9.03 ± 1.33	7.88 ± 2.27	**10.03 ± 0.54**
		LCM	1.46 ± 1.03	1.50 ± 1.07	1.50 ± 1.07	**1.74 ± 1.38**
	50	UAS	26.59 ± 2.37	**32.83 ± 1.50**†	31.35 ± 1.10†	23.18 ± 2.03†
		LAS	10.14 ± 0.57	10.73 ± 0.86	**10.74 ± 0.54**	10.10 ± 0.69
		UCM	10.03 ± 1.31	10.63 ± 0.50	**10.81 ± 0.50**	10.34 ± 0.36
		LCM	3.24 ± 0.31	3.26 ± 0.24	3.38 ± 0.27	**3.43 ± 0.27**
	100	UAS	30.75 ± 1.13	**33.75 ± 0.86**†	33.62 ± 1.49†	27.12 ± 1.25†
		LAS	10.84 ± 0.80	11.48 ± 0.75	**11.69 ± 0.67**†	10.48 ± 0.70†
		UCM	**11.61 ± 1.22**	11.30 ± 0.29	11.34 ± 0.26	11.08 ± 0.44
		LCM	3.53 ± 0.60	**3.61 ± 0.31**	3.57 ± 0.23	3.55 ± 0.23
	1000	UAS	42.82 ± 1.82	**43.18 ± 1.73**	41.20 ± 2.17†	36.30 ± 2.79†
		LAS	**18.44 ± 1.00**	18.24 ± 1.62	18.13 ± 1.13	16.38 ± 1.20†
		UCM	**15.86 ± 0.40**	15.18 ± 0.81	13.78 ± 0.30†	13.52 ± 0.43†
		LCM	**4.49 ± 0.47**	4.37 ± 0.46	4.31 ± 0.41†	4.29 ± 0.38†

The scaled augmented Lagrangian function is

$$\begin{aligned}
L_\rho(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{y}) &= g(\boldsymbol{x}, \boldsymbol{u}) + \boldsymbol{\lambda}'^\intercal(\boldsymbol{x} - \boldsymbol{u}) + \frac{\rho}{2}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 \\
&= g(\boldsymbol{x}, \boldsymbol{u}) + \frac{\rho}{2}\|\boldsymbol{x} - \boldsymbol{u} + \frac{1}{\rho}\boldsymbol{\lambda}'\|_2^2 - \frac{1}{2\rho}\|\boldsymbol{\lambda}'\|_2^2 \\
&= g(\boldsymbol{x}, \boldsymbol{u}) + \frac{\rho}{2}\|\boldsymbol{x} - \boldsymbol{u} + \boldsymbol{\lambda}\|_2^2 - \frac{\rho}{2}\|\boldsymbol{\lambda}\|_2^2, \\
\text{where} \qquad \boldsymbol{\lambda} &:= \frac{1}{\rho}\boldsymbol{\lambda}'.
\end{aligned}$$

The ADMM algorithm updates the parameters as follows:

$$\begin{aligned}
\boldsymbol{x}^{t+1} &:= \arg\min_{\boldsymbol{x}} L_\rho(\boldsymbol{x}, \boldsymbol{u}^t, \boldsymbol{\lambda}^t) \\
&= \arg\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{w}\|_2^2 + I_{\mathcal{X}}(\boldsymbol{x}) + \frac{\rho}{2}\|\boldsymbol{x} - \boldsymbol{u}^t + \boldsymbol{\lambda}^t\|_2^2 \\
&= \arg\min_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x} - \frac{1}{\rho+1}(\boldsymbol{w} + \rho\boldsymbol{u}^t - \rho\boldsymbol{\lambda}^t)\|_2^2, \\
&\triangleq \text{Proj}_{\mathcal{X}}(\frac{1}{\rho+1}(\boldsymbol{w} + \rho\boldsymbol{u}^t - \rho\boldsymbol{\lambda}^t)) \\
\boldsymbol{u}^{t+1} &:= \arg\min_{\boldsymbol{u}} L_\rho(\boldsymbol{x}^{t+1}, \boldsymbol{u}, \boldsymbol{\lambda}^t) \\
&= \arg\min_{\boldsymbol{u}} \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{w}\|_2^2 + I_{\mathcal{F}}(\boldsymbol{u}) + \frac{\rho}{2}\|\boldsymbol{x}^{t+1} - \boldsymbol{u} + \boldsymbol{\lambda}^t\|_2^2 \\
&= \arg\min_{\boldsymbol{u}\in\mathcal{U}} \|\boldsymbol{u} - \frac{1}{\rho+1}(\boldsymbol{w} + \rho\boldsymbol{x}^{t+1} + \rho\boldsymbol{\lambda}^t)\|_2^2, \\
&\triangleq \text{Proj}_{\mathcal{U}}(\frac{1}{\rho+1}(\boldsymbol{w} + \rho\boldsymbol{x}^{t+1} + \rho\boldsymbol{\lambda}^t)) \\
\boldsymbol{\lambda}^{t+1} &:= \boldsymbol{\lambda}^t + (\boldsymbol{x}^{t+1} - \boldsymbol{u}^{t+1}).
\end{aligned}$$

Projection onto $\mathcal{X}$ is decomposable over each $j \in \mathcal{V}$. And for each j, the optimal value of the group can be computed in $\mathcal{O}(n)$ in almost closed form via Section 5.5.1 in Zhang et al. [2010] or other simplex projection algorithms in $\mathcal{O}(n \log n)$.

Projection onto $\mathcal{U}$ is a minimum quadratic capacity expansion cost problem for fixed multi-commodity flows:

$$\min_{\boldsymbol{u}\in\mathcal{U}} \|\boldsymbol{u} - \boldsymbol{w}\|_2^2.$$

A partially relaxed problem is

$$\begin{aligned}
&\max_{\boldsymbol{\beta}} \min_{\boldsymbol{u}, \boldsymbol{f}} \|\boldsymbol{u} - \boldsymbol{w}\|_2^2 + \sum_{a,k} \beta_a^k (f_a^k - u_a) \\
\text{s.t.} \quad &\sum_{a\in\delta^-(j)} f_a^k - \sum_{a\in\delta^+(j)} f_a^k = \mathbb{I}(j = k) - \mathbb{I}(j = r) \quad \forall k \in \mathcal{V} \setminus \{r\}, j \in \mathcal{V} \\
&f_a^k \geq 0, \beta_a^k \geq 0 \quad \forall k \in \mathcal{V} \setminus \{r\}, a \in \mathcal{E}.
\end{aligned}$$

Given $\boldsymbol{\beta}$, the sub-problem for $\boldsymbol{u}$ is

$$\min_{\boldsymbol{u}} \sum_a u_a^2 - 2u_a w_a - \sum_k \beta_a^k u_a,$$

with an analytical solution

$$\boldsymbol{u}^* = \boldsymbol{w} + \frac{1}{2}\boldsymbol{\beta}^k.$$

Given $\boldsymbol{\beta}$, the sub-problem for $\boldsymbol{f}$ is

$$\begin{aligned}
&\min_{\boldsymbol{f}} \sum_{a,k} \beta_a^k f_a^k \\
\text{s.t.} \quad &\sum_{a \in \delta^-(j)} f_a^k - \sum_{a \in \delta^+(j)} f_a^k = \mathbb{I}(j = k) - \mathbb{I}(j = r) \forall k \in \mathcal{V} \setminus \{r\}, j \in \mathcal{V} \\
&f_a^k \geq 0 \quad \forall k \in \mathcal{V} \setminus \{r\}, a \in \mathcal{E},
\end{aligned}$$

which is a minimum-cost multi-commodity flow problem.

With $\boldsymbol{u}^*$ and $\boldsymbol{f}^*$, we can optimize $\boldsymbol{\beta}$ with sub-gradient ascent.

Alternatively, another partially relaxed problem is

$$\begin{aligned}
&\max_{\boldsymbol{\beta}} \min_{\boldsymbol{u},\boldsymbol{f}} \|\boldsymbol{u} - \boldsymbol{w}\|_2^2 + \sum_{a,k} f_a^k (\beta_{h(a)}^k - \beta_{t(a)}^k) + \sum_k \beta_r^k - \beta_k^k \\
\text{s.t.} \quad &0 \leq f_a^k \leq u_a, \beta_a^k \geq 0 \quad \forall k \in \mathcal{V} \setminus \{r\}, a \in \mathcal{E},
\end{aligned}$$

where $h(a)$ and $t(a)$ are the head and tail of arc a respectively.

Given $\boldsymbol{\beta}$, the inner minimization problem is decomposed over a:

$$\begin{aligned}
&\min_{\boldsymbol{u},\boldsymbol{f}} u_a^2 - 2u_a w_a + \sum_k f_a^k (\beta_{h(a)}^k - \beta_{t(a)}^k) \\
\text{s.t.} \quad &0 \leq f_a^k \leq u_a \quad \forall k \in \mathcal{V} \setminus \{r\},
\end{aligned}$$

which is a convex continuous knapsack problem for each a.

The above optimization requires sub-gradient methods, which are usually slower than FW ($\mathcal{O}(\frac{1}{\epsilon^2})$).