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

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 $\mathcal{C} := \{\boldsymbol{u} : \|\boldsymbol{u} - \mathbb{E}_{\mathbb{P}^{\text{emp}}}\boldsymbol{\phi}(\cdot)\| \leq \varepsilon\}$. 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

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

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

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

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 $|\ell(\cdot,\cdot)| \leq K$, 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:*

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

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 $\boldsymbol{\theta}^*_{\text{emp}}$.

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

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

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

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}\|_* \\
\le&\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,

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

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 $\mathcal{A}_{\text{marb}}$:

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

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:

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

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 $\mathcal{A}_{\text{mdep}}$:

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

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:

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

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:

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

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 ($\mathcal{O}(\frac{1}{\epsilon^2})$).