Moment Distributionally Robust Tree Structured Prediction

Yeshu Li [1] Danyal Saeed [1] Xinhua Zhang [1] Brian D. Ziebart [1]
Kevin Gimpel [2]

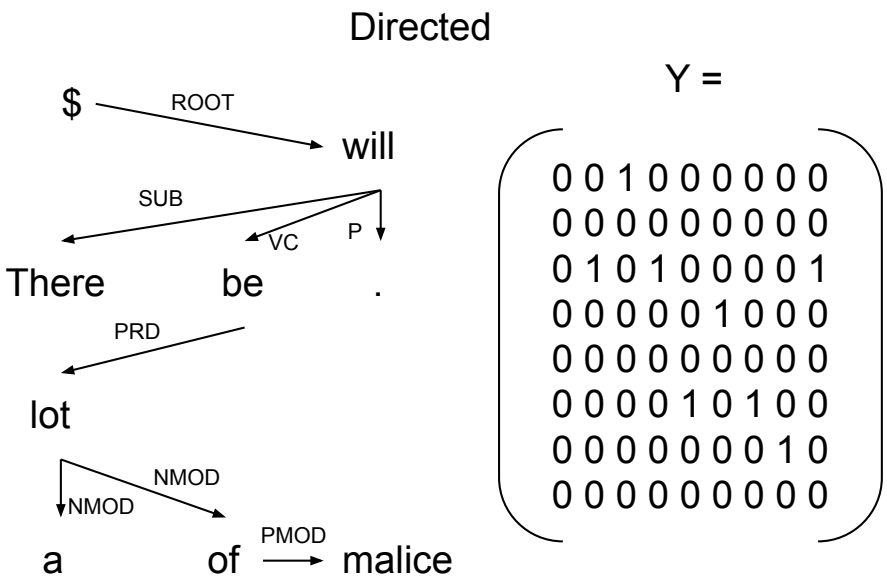
[1]Department of Computer Science
University of Illinois at Chicago

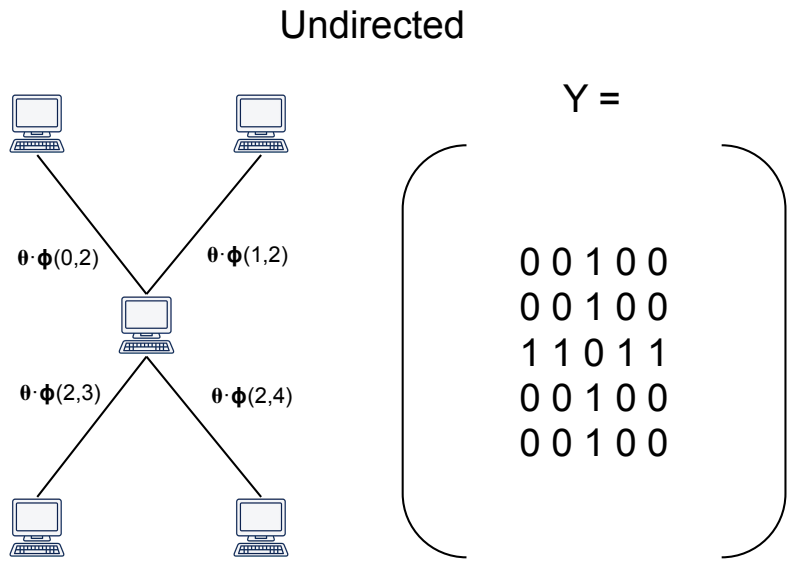
[2]Toyota Technological Institute at Chicago

Nov. 2022

Tree Structured Prediction

- Task: Given training data $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \ldots, (\boldsymbol{x}_m, \boldsymbol{y}_m)\}$, learn a mapping $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y}$ is the space of tree-shaped objects (discrete, exponential but structured)
- Dependency Parsing: A large body of literature focuses on predicting dependency tree structures in natural language processing





Statistical Learning: Test vs Train

- The ultimate goal is to perform well on test data, in terms of a performance metric L, typically piecewise-constant and discontinuous, usually intractable to optimize on training data

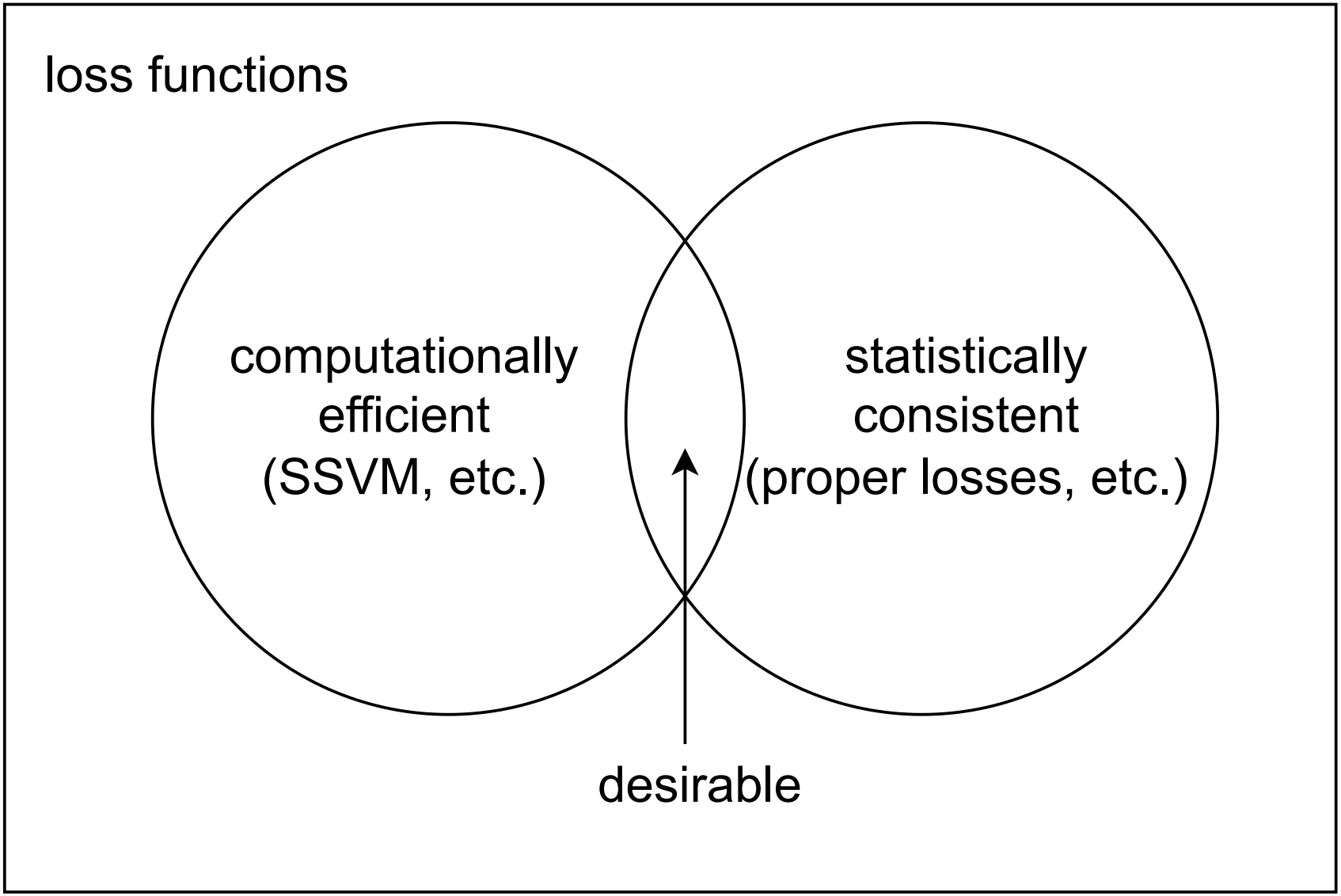
$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n_{\text{train}}} L(f(\boldsymbol{x}_i), \boldsymbol{y}_i)$$

- Resort to surrogate loss and mappings for easier optimization

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n_{\text{train}}} S(g_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \boldsymbol{y}_i)$$

- Might lead to discrepancy between training and testing objectives

Design of Losses



Fisher Consistency

Definition (Fisher consistency)

Given a space $\mathcal{H}$ of all measurable functions $\mathcal{X} \mapsto \mathcal{P}(\mathcal{Y})$, a space $\mathcal{G}$ of all measurable functions $\mathcal{X} \mapsto \mathcal{T}$, and a surrogate loss function $S : \mathcal{T} \times \mathcal{P}(\mathcal{Y}) \mapsto \mathbb{R}_+$, we say that the surrogate loss S is Fisher consistent with respect to the loss $L : \mathcal{P}(\mathcal{Y}) \times \mathcal{P}(\mathcal{Y}) \mapsto \mathbb{R}_+$ if there exists a function $d : \mathcal{T} \mapsto \mathcal{P}(\mathcal{Y})$ such that for every probability distribution $\mathbb{P}$ on $\mathcal{X} \times \mathcal{Y}$, every minimizer g^* of the surrogate risk reaches Bayes optimal risk:

$$R^S_{\mathbb{P}}(g^*) = \min_{g \in \mathcal{G}} R^S_{\mathbb{P}}(g) \implies R^L_{\mathbb{P}}(d \circ g^*) = \min_{h \in \mathcal{H}} R^L_{\mathbb{P}}(h).$$

Existing Methods

- Empirical risk minimization (K-best lists (Smith and Eisner, 2006), automatic differentiation (Stoyanov and Eisner, 2012))

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{m} \sum_{\boldsymbol{y} \in \mathcal{Y}} L(\boldsymbol{y}_i, \boldsymbol{y}) \Pr_{\boldsymbol{\theta}}(\boldsymbol{y} | \boldsymbol{x}_i) = \sum_{i=1}^{m} \sum_{\boldsymbol{y} \in \mathcal{Y}} L(\boldsymbol{y}_i, \boldsymbol{y}) \frac{\exp(\text{score}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{y}))}{\sum_{\boldsymbol{y}'} \exp(\text{score}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{y}'))}$$

- Conditional log-likelihood (deep learning (Dozat and Manning, 2017), matrix tree theorem (Koo et al., 2007))

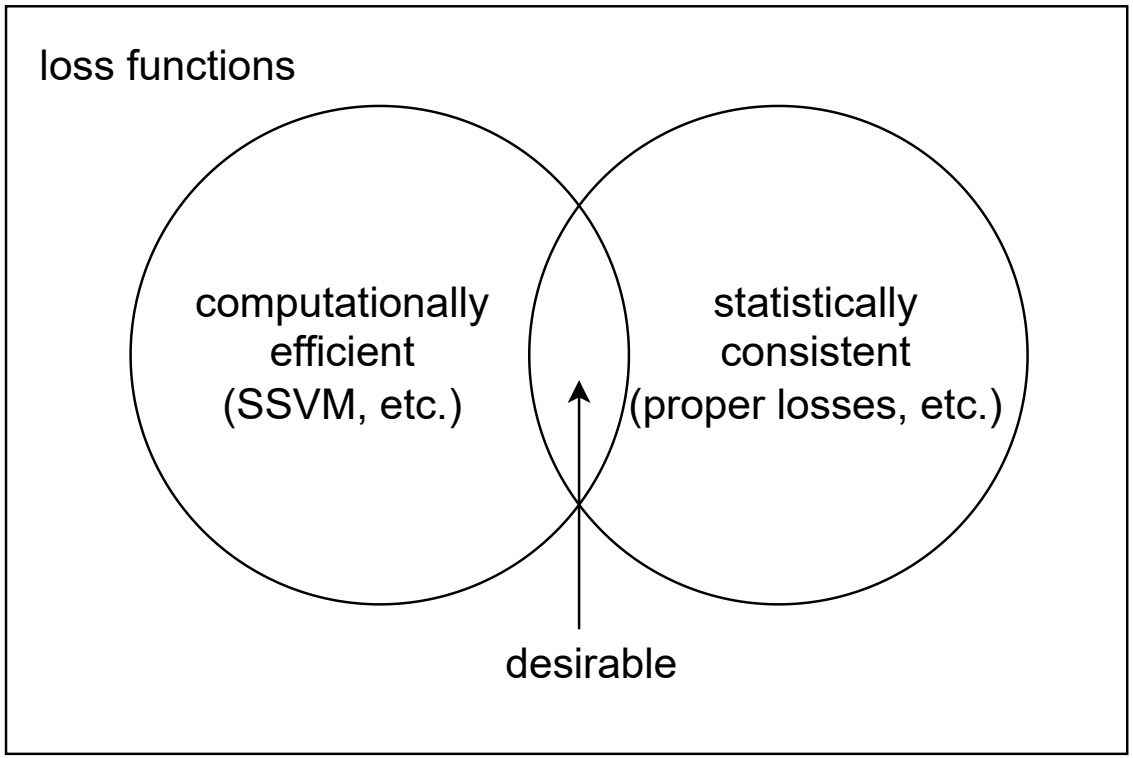
$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{m} -\text{score}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{y}_i) + \log \sum_{\boldsymbol{y}} \exp(\text{score}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{y}))$$

- Max-margin (Structured SVM (Taskar et al., 2004))

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{m} -\text{score}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{y}_i) + \max_{\boldsymbol{y}} (\text{score}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{y}) + L(\boldsymbol{y}_i, \boldsymbol{y}))$$

Motivation

- Existing methods are either **NOT** Fisher consistent or **NOT** convex
- Existing methods rely on instinctive regularization to combat overfitting, without good probabilistic interpretation

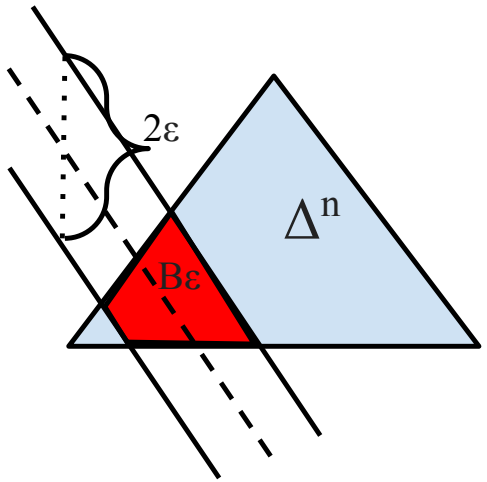


Proposed Method (Primal)

Distributionally robust optimization based on moment divergence

$$\inf_{\hat{\mathbb{P}}} \sup_{\check{\mathbb{P}} \in B_\varepsilon(\tilde{\mathbb{P}})} \mathbb{E}_{\tilde{\mathbb{P}}_{\boldsymbol{X}}, \hat{\mathbb{P}}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}, \check{\mathbb{P}}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}} L(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}})$$

- $B_\varepsilon(\tilde{\mathbb{P}}) := \{\check{\mathbb{P}} : \|\mathbb{E}_{\check{\mathbb{P}}_{\boldsymbol{X},\check{\boldsymbol{Y}}}} \phi(\boldsymbol{X}, \check{\boldsymbol{Y}}) - \mathbb{E}_{\tilde{\mathbb{P}}_{\boldsymbol{X},\boldsymbol{Y}}} \phi(\boldsymbol{X}, \boldsymbol{Y})\| \leq \varepsilon, \check{\mathbb{P}}_{\boldsymbol{X}} = \tilde{\mathbb{P}}_{\boldsymbol{X}}\}$ is a convex compact ambiguity set quantifying the uncertainty about the underlying true distribution
- $\phi(\boldsymbol{x}, \boldsymbol{y}) := \sum_{f \in \mathcal{F}} \phi(\boldsymbol{x}, \boldsymbol{y}_f)$ is a feature mapping decomposable over factors f



Proposed Method (Dual)

Proposition (dual norm regularization)

The distributionally robust tree structured prediction problem based on moment divergence can be rewritten as

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

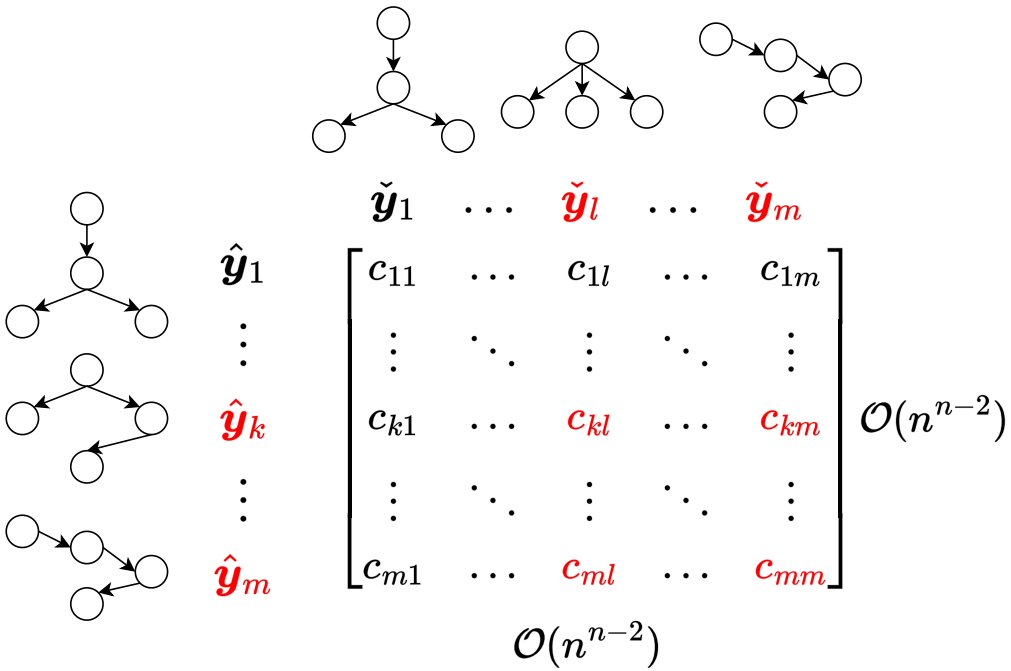
where $\boldsymbol{\theta} \in \mathbb{R}^d$ is the vector of Lagrangian multipliers and $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

Algorithm: Game Theory

- The inner minimax problems are independent conditioned on $\boldsymbol{X}$

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}^{\text{emp}}_{\boldsymbol{X},\boldsymbol{Y}}} \min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{X}},\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{X}}} \text{payoff}(\hat{\boldsymbol{Y}}, \check{\boldsymbol{Y}})$$

- Use constraint generation to find a Nash equilibrium $\mathbb{P}^*_{\hat{\boldsymbol{Y}}|\boldsymbol{X}}$, $\mathbb{Q}^*_{\check{\boldsymbol{Y}}|\boldsymbol{X}}$
- No convergence guarantees

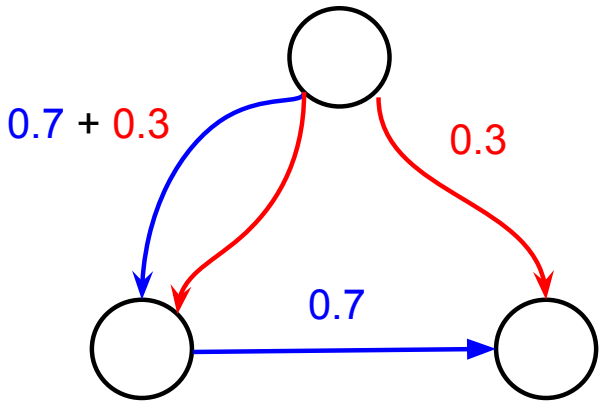


Algorithm: Marginal Distribution

- Assuming additive L, rearrange the order of optimization variables

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

- Converges to the global optimum
- Requires an oracle for projection onto the arborescence (directed spanning tree) polytopes

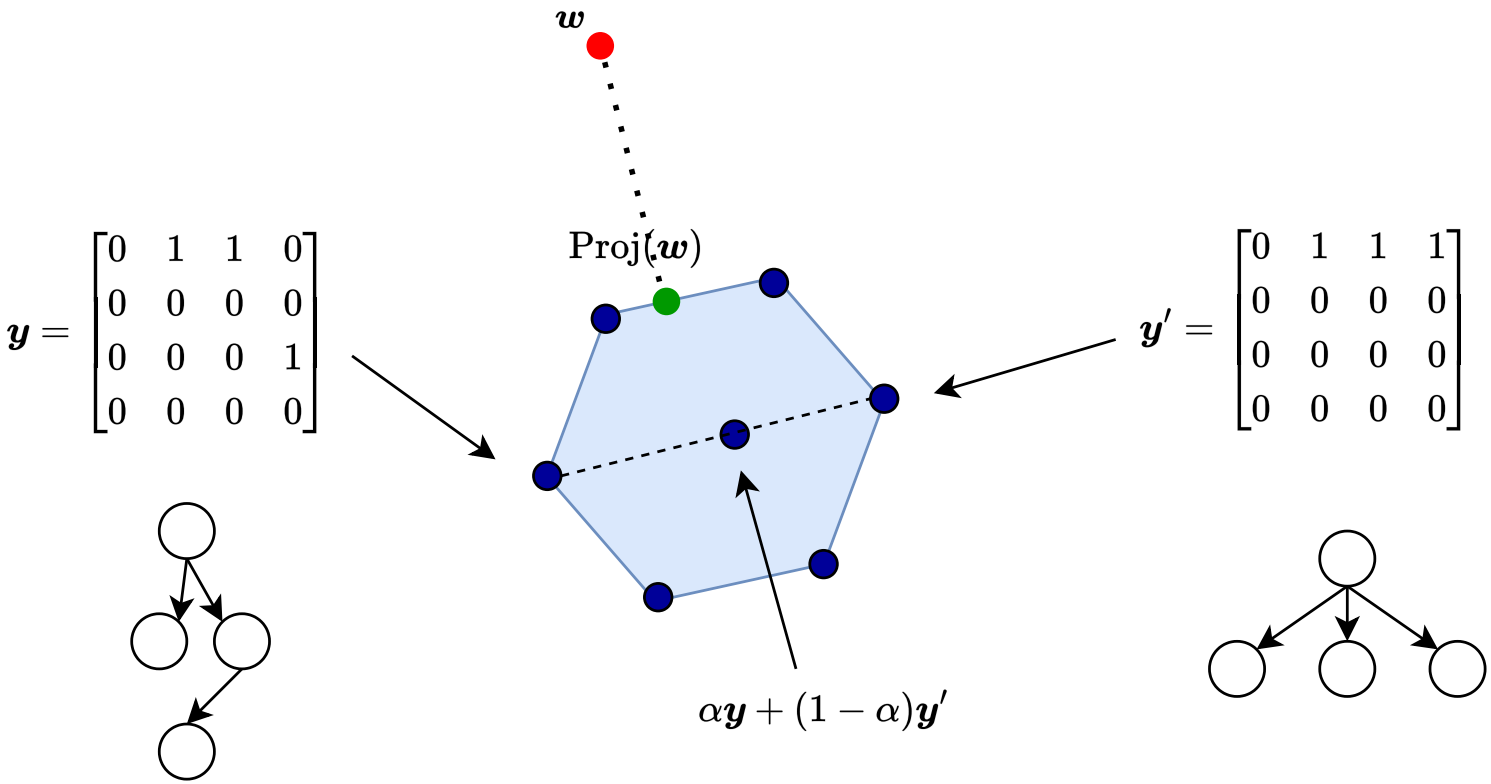


$$\begin{pmatrix} 0.0 & 1.0 & 0.3 \\ 0.0 & 0.0 & 0.7 \\ 0.0 & 0.0 & 0.0 \end{pmatrix}$$

Projection on the Arborescence Polytope

Constrained quadratic programming

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



Projection on the Arborescence Polytope

Proposed solutions

- Frank-Wolfe (Frank and Wolfe, 1956)
  - Based on minimum weight directed spanning trees
  - Low per-iteration cost
  - Sub-linear convergence rate $\mathcal{O}(\frac{1}{\epsilon})$
- Alternating Direction Method of Multipliers (Boyd et al., 2011)
  - Based on a compact representation of the first-order arborescence polytope (Friesen, 2019)
  - Higher per-iteration cost
  - Linear convergence rate $\mathcal{O}(\log \frac{1}{\epsilon})$

Excess True Risk Bound

Theorem

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:

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

where $\boldsymbol{\theta}^*_{\text{emp}}$ and $\boldsymbol{\theta}^*_{\text{true}}$ are the optimal parameters learned under $\mathbb{P}^{\text{emp}}$ and $\mathbb{P}^{\text{true}}$ 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 $\mathbb{P}^{\boldsymbol{\theta}}_{\boldsymbol{Y}|\boldsymbol{X}} \in \arg\min_{\mathbb{P}}\max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\boldsymbol{Y}}|\boldsymbol{x}}\mathbb{P}_{\hat{\boldsymbol{Y}}|\boldsymbol{x}}}\ell(\hat{\boldsymbol{Y}},\check{\boldsymbol{Y}}) + \boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{\phi}(\boldsymbol{x},\check{\boldsymbol{Y}})$.

Fisher Consistency

Corollary of Proposition C.2 in Nowak et al. (2020)

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

Results

- Datasets: PTB, CTB, UD
- Baseline: Deep Biaffine Attention (Dozat and Manning, 2017)
- Marginal: use marginal probabilities with full gradient
- Stochastic: use marginal probabilities with mini-batch
- Game: constraint generation (double oracle)

Table 1: Comparison of mean unlabeled attachment score (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			
Method	Time (s)	m = 10	50	100	1000	m = 10	50	100	1000	m = 10	50	100	1000
BiAF (baseline)	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
Marginal (ours)	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**†
Stochastic (ours)	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†
Game (ours)	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

More Results

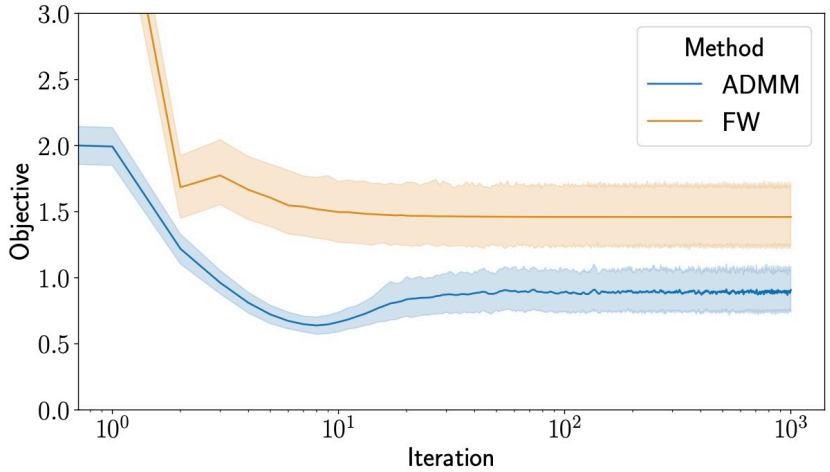


Figure 1: Convergence of Alternating Direction Method of Multipliers (ADMM) and Frank-Wolfe (FW) for random points with 95% confidence intervals.

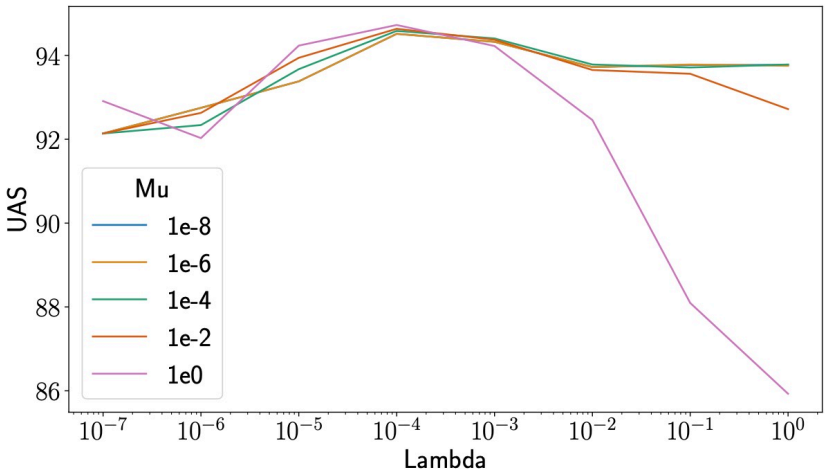
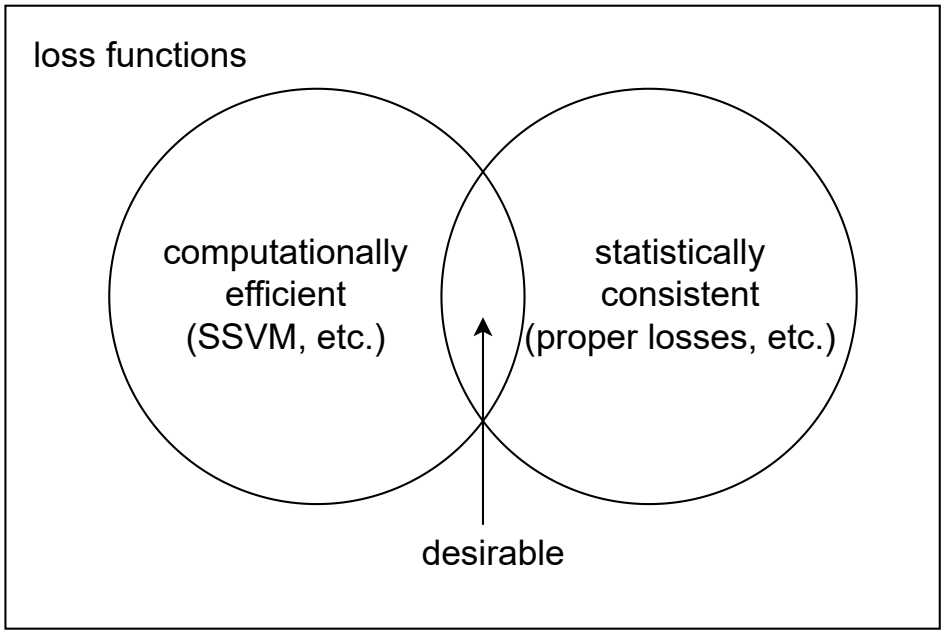


Figure 2: The best unlabeled attachment score (UAS) with the Marginal algorithm as μ and λ vary in logarithmic scales.

Conclusion

- Tree structured prediction from first principles in DRO
  - Dependency, directed, undirected, higher-order trees
- Generalization bounds and Fisher consistency
- Efficient projection oracles on arborescence polytopes
- Code available at `https://github.com/DanielLeee/drtreesp`



Future Work

- General structured prediction tasks
- Other ambiguity sets in distributionally robust optimization
- End-to-end representation learning (code available)

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

References

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122.

Dozat, T. and Manning, C. D. (2017). Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

Frank, M. and Wolfe, P. (1956). An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110.

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

Koo, T., Globerson, A., Carreras Pérez, X., and Collins, M. (2007). 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.

Nowak, A., Bach, F., and Rudi, A. (2020). Consistent structured prediction with max-min margin markov networks. In *International Conference on Machine Learning*, pages 7381–7391. PMLR.

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

Stoyanov, V. and Eisner, J. (2012). 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.

Taskar, B., Klein, D., Collins, M., Koller, D., and Manning, C. D. (2004). Max-margin parsing. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 1–8.

Thank you!

Q & A