THE UNIVERSITY OF ILLINOIS AT CHICAGO

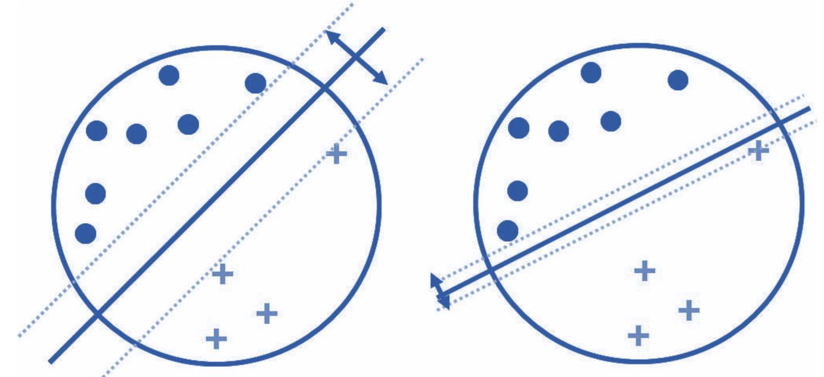
Distributionally Robust Graphical Models

Rizal Fathony, Ashkan Rezaei, Mohammad Ali Bashiri, Xinhua Zhang, Brian D. Ziebart
{rfatho2,arezae4,mbashi4,zhangx,bziebart}@uic.edu

UIC COMPUTER SCIENCE

Structured Prediction with Label Interactions

Structured SVM (Tsochantaridis et. al., 2005)



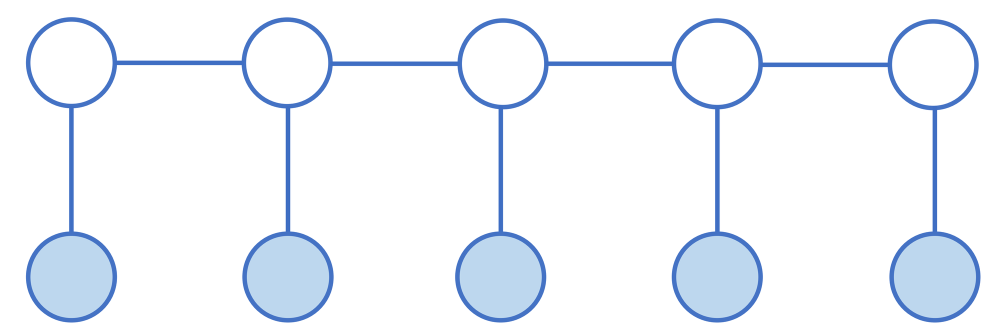
$$\text{hinge}_\theta(\check{\mathbf{y}}) = \max_{\mathbf{y}} \left\{ \text{loss}(\mathbf{y}, \check{\mathbf{y}}) + \theta \cdot (\Phi(\mathbf{x}, \check{\mathbf{y}}) - \Phi(\mathbf{x}, \mathbf{y})) \right\}$$

$$\Phi(\mathbf{x}, \mathbf{y}) = \sum_c \phi(\mathbf{x}, \mathbf{y}_c)$$

✔ Align with the loss/performance metrics

✘ No Fisher consistency guarantee
Based on Crammer & Singer's Multiclass SVM

Conditional Random Fields (Lafferty et. al., 2001)



$$P_\theta(\mathbf{y}|\mathbf{x}) = \frac{e^{\theta \cdot \Phi(\mathbf{x}, \mathbf{y})}}{\sum_{\mathbf{y}' \in \mathcal{Y}} e^{\theta \cdot \Phi(\mathbf{x}, \mathbf{y}')}}$$

$$\Phi(\mathbf{x}, \mathbf{y}) = \sum_c \phi(\mathbf{x}, \mathbf{y}_c)$$

✔ Fisher Consistent
Produce Bayes optimal prediction in ideal case.

✘ No easy mechanism to incorporate customized loss/performance metrics

Adversarial Graphical Models (AGM)

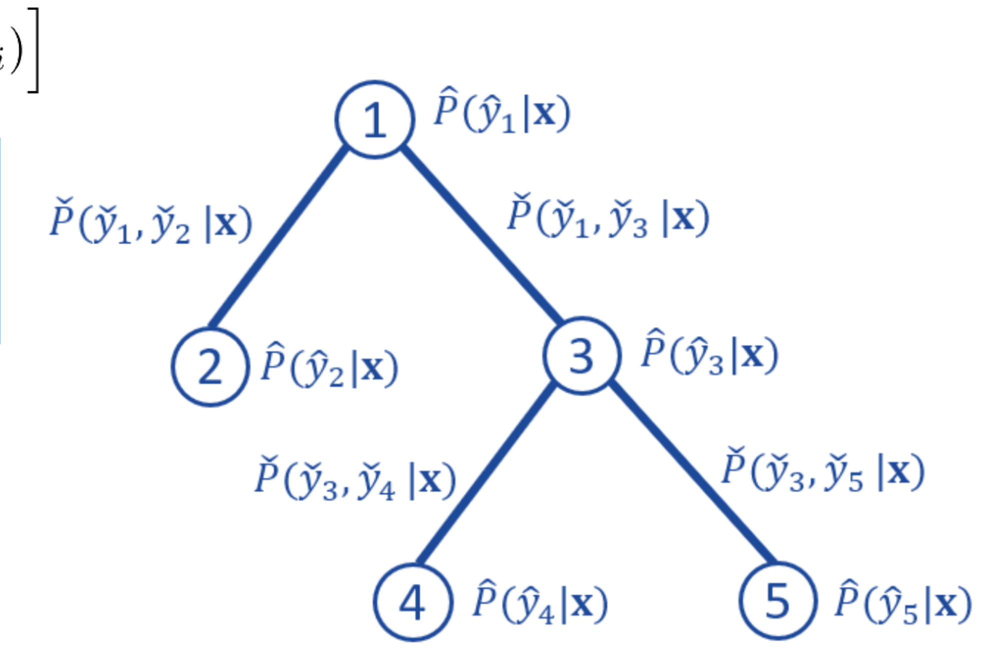
- A distributionally robust approach
- Seek a predictor that robustly minimize a loss metric against the worst-case conditional distribution that match the statistics of the training data

$$\min_{\hat{P}(\hat{\mathbf{y}}|\mathbf{x})} \max_{\check{P}(\check{\mathbf{y}}|\mathbf{x})} \mathbb{E}_{\mathbf{X}\sim\tilde{P}; \hat{\mathbf{Y}}|\mathbf{X}\sim\hat{P}; \check{\mathbf{Y}}|\mathbf{X}\sim\check{P}} \left[\text{loss}(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) \right] \text{ subject to: } \mathbb{E}_{\mathbf{X}\sim\tilde{P}; \check{\mathbf{Y}}|\mathbf{X}\sim\check{P}} \left[\Phi(\mathbf{X}, \check{\mathbf{Y}}) \right] = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\tilde{P}} \left[\Phi(\mathbf{X}, \mathbf{Y}) \right]$$

- Focus: pairwise graphical models: interactions between label = edges in graphs
- Feature function $\Phi(\mathbf{X}, \mathbf{Y})$: additively decomposed over nodes and edges
- Loss metric: additively decomposed over each y_i variables, $\text{loss}(\hat{y}, \check{y}) = \sum_{i=1}^{n} \text{loss}(\hat{y}_i, \check{y}_i)$
- Dual problem: can be written in terms of node and edge marginal distributions:

$$\min_{\theta_e, \theta_v} \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\tilde{P}} \max_{\check{P}(\check{\mathbf{y}}|\mathbf{x})} \min_{\hat{P}(\hat{\mathbf{y}}|\mathbf{x})} \Big[\sum_i^n \sum_{\hat{y}_i, \check{y}_i} \hat{P}(\hat{y}_i|\mathbf{x}) \check{P}(\check{y}_i|\mathbf{x}) \text{loss}(\hat{y}_i, \check{y}_i) + \sum_{(i,j)\in E} \sum_{\check{y}_i, \check{y}_j} \check{P}(\check{y}_i, \check{y}_j|\mathbf{x}) \left[\theta_e \cdot \phi(\mathbf{x}, \check{y}_i, \check{y}_j) \right] - \sum_{(i,j)\in E} \theta_e \cdot \phi(\mathbf{x}, y_i, y_j) + \sum_i^n \sum_{\check{y}_i} \check{P}(\check{y}_i|\mathbf{x}) \left[\theta_v \cdot \phi(\mathbf{x}, \check{y}_i) \right] - \sum_i^n \theta_v \cdot \phi(\mathbf{x}, y_i) \Big]$$

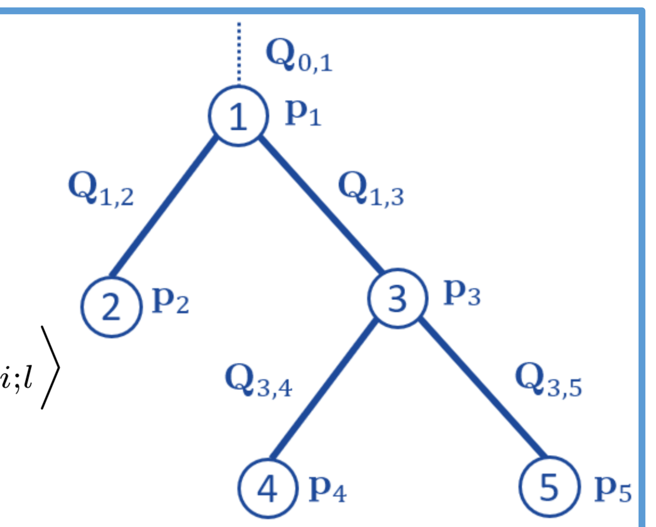
Similar to CRF and SSVM: General Graphical Models: Intractable	Focus: Graphs with low tree-width, e.g., chain, tree. Tractable optimization



✔ Align with the loss/performance metrics

✔ Fisher Consistent

AGM | Optimization



- Focus: tree-structured graphical models
- Matrix & vector notations:

$$\min_{\theta_e,\theta_v} \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\tilde{P}} \max_{\mathbf{Q}\in\Delta} \min_{\mathbf{p}\in\Delta} \sum_i^n \Big[\mathbf{p}_i \mathbf{L}_i (\mathbf{Q}_{pt(i);i}^{\mathrm{T}} \mathbf{1}) + \Big\langle \mathbf{Q}_{pt(i);i} - \mathbf{Z}_{pt(i);i}, \textstyle\sum_l \theta_e^{(l)} \mathbf{W}_{pt(i);i;l} \Big\rangle + (\mathbf{Q}_{pt(i);i}^{\mathrm{T}} \mathbf{1} - \mathbf{z}_i)^{\mathrm{T}} (\textstyle\sum_l \theta_v^{(l)} \mathbf{w}_{i;l}) \Big]$$

$$\text{subject to: } \mathbf{Q}_{pt(pt(i));pt(i)}^{\mathrm{T}} \mathbf{1} = \mathbf{Q}_{pt(i);i} \mathbf{1}, \ \forall i \in \{1, \ldots, n\}$$

Optimization Technique:

- Stochastic (sub)-gradient descent (outer opt. for θ_e and θ_v)
- Dual decomposition (inner $\mathbf{Q}$ optimization)
- Discrete optimal transport solver (recovering $\mathbf{Q}$)
- Closed-form solution (inner $\mathbf{p}$ optimization)

Runtime (for a single subgradient update):

- Depends on the loss metric
- Additive zero-one loss metric: $O(nlk \log k + nk^2)$
  k: # classes, n: # nodes,
  l: # iterations in dual decomposition
- Competitive with CRF $[O(nk^2)]$ and SSVM $[O(nk^2)]$

General Graphical Structure with Low Treewidth:

- Create a junction tree representation, then run the same optimization technique.
- Runtime: $O\left(nlwk^{(w+1)} \log k + nk^{2(w+1)}\right)$, where: n: # cliques, w: treewidth of the graph

Experiments

1. Facial Emotion Intensity Prediction (Chain Structure, Labels with Ordinal Category)

- Predict emotion intensity of each picture in a video
- Each node: 3 class classification: *neutral = 1< increasing = 2 < apex = 3*
- Ordinal loss metrics: zero-one loss, absolute loss, and squared loss
- Weighted and unweighted. Weights reflect the focus of prediction.
- Results: Overall, AGM has advantages over SSVM & CRF in terms of the average loss and number of "indistinguishably best" performance.

Table 1: The average loss metrics for the emotion intensity prediction. Bold numbers indicate the best or not significantly worse than the best results (Wilcoxon signed-rank test with $\alpha = 0.05$).

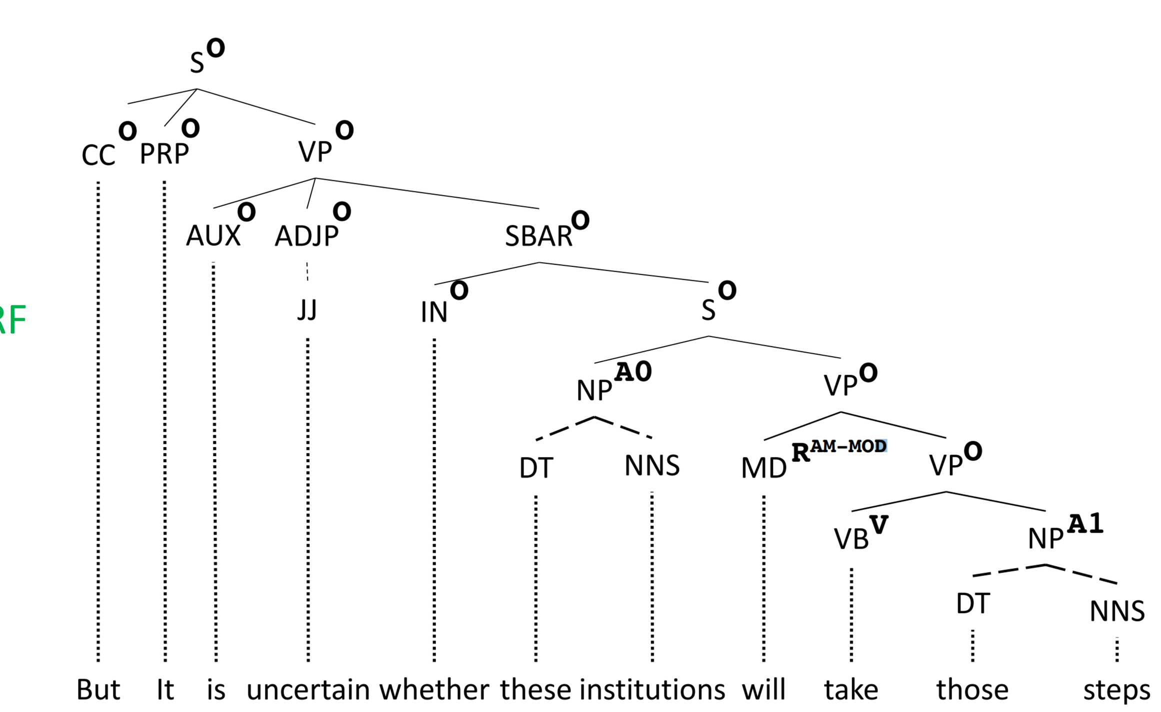
Loss metrics	AGM	CRF	SSVM
zero-one, unweighted	0.34	**0.32**	0.37
absolute, unweighted	**0.33**	0.34	0.40
squared, unweighted	**0.38**	**0.38**	0.40
zero-one, weighted	**0.28**	0.32	0.29
absolute, weighted	**0.29**	0.36	**0.29**
squared, weighted	0.36	0.40	**0.33**
average	0.33	0.35	0.35
# bold	4	2	2

2. Semantic Role Labeling (Tree Structure)

- Predict label of each node given known parse tree.
- Cost-sensitive loss metric is used reflect the importance of each label
- CoNLL 2005 dataset
- Result: AGM: competitive with SSVM & better than CRF
- Incorporating loss metric in learning is important

Table 2: The average loss metrics for the semantic role labeling task.

Loss metrics	AGM	CRF	SSVM
cost-sensitive loss	0.14	0.19	0.14



Acknowledgement. This research was supported in part by National Science Foundation under Grant No. 1652530, and by the Future of Life Institute (futureoflife.org) FLI-RFP-AI1 program