Structured Prediction Theory and Algorithms

Joint work with Corinna Cortes (Google Research)
Vitaly Kuznetsov (Google Research)
Scott Yang (Courant Institute)

MEHRYAR MOHRI MOHRI@
COURANT INSTITUTE & GOOGLE RESEARCH

Structured Prediction

- Structured output:

$$\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_l.$$

- Loss function: $\mathsf{L} \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ decomposable.
  - Example: Hamming loss.

$$\mathsf{L}(y, y') = \frac{1}{l} \sum_{k=1}^{l} 1_{y_k \neq y'_k}.$$

  - Example: edit-distance loss.

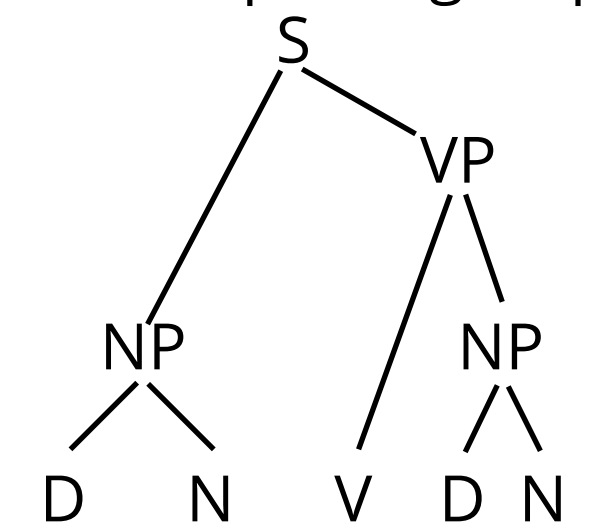
$$\mathsf{L}(y, y') = \frac{1}{l} \mathsf{d}_{\text{edit}}(y_1 \cdots y_l, y'_1 \cdots y'_l).$$

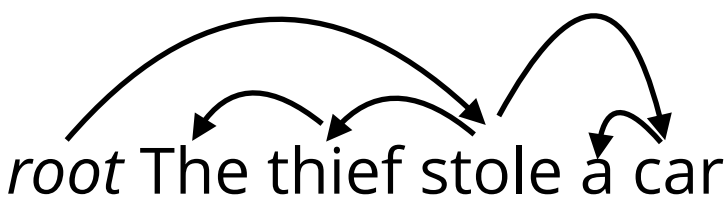
Examples

- Pronunciation modeling.
- Part-of-speech tagging.
- Named-entity recognition.
- Context-free parsing.
- Dependency parsing.
- Machine translation.
- Image segmentation.

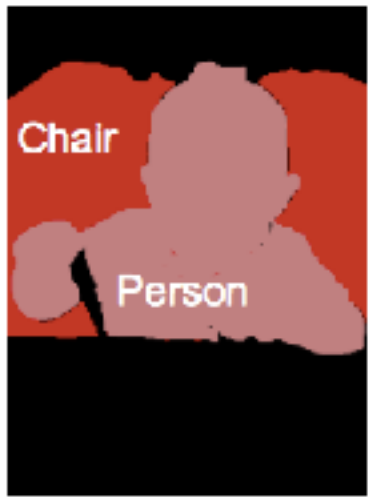
Examples: NLP Tasks

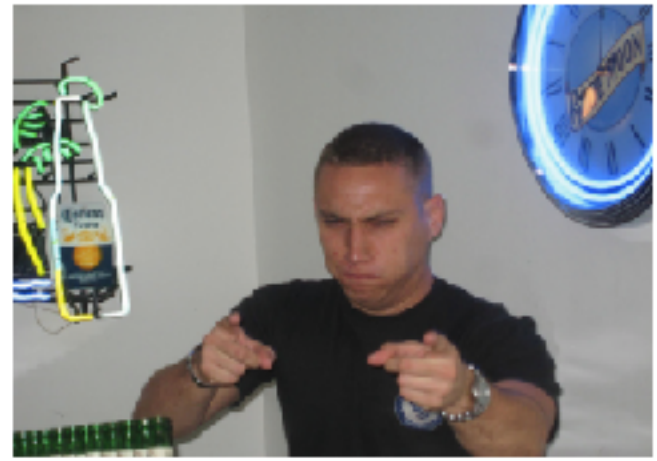
- Pronunciation: I have formulated a
  ay hh ae v f ow r m y ax l ey t ih d ax
- POS tagging: The thief stole a car
  D N V D N
- Context-free parsing/Dependency parsing:

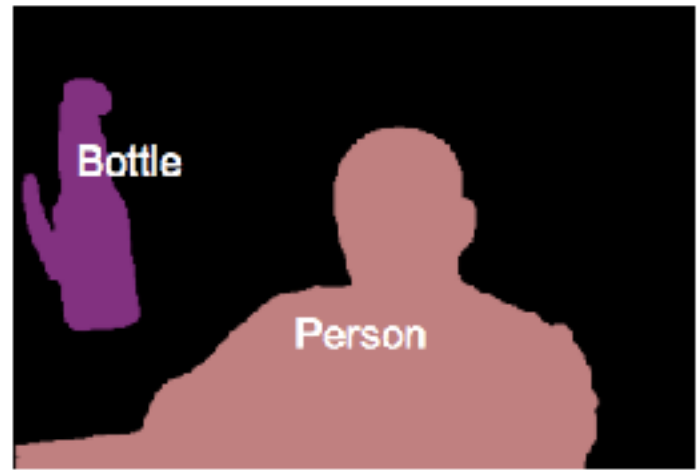




Examples: Image Segmentation







Predictors

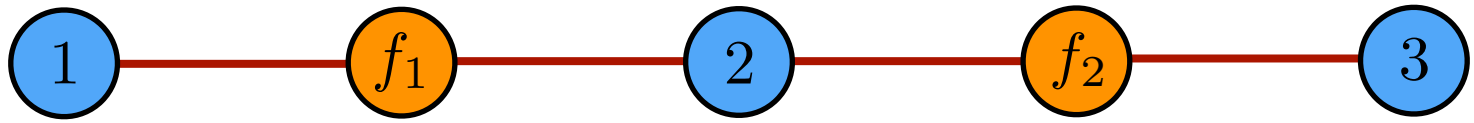
- Family of scoring functions $\mathcal{H}$ mapping from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$.
- For any $h \in \mathcal{H}$, prediction based on highest score:

$$\forall x \in \mathcal{X},\ \mathsf{h}(x) = \operatorname*{argmax}_{y \in \mathcal{Y}} h(x, y).$$

- Decomposition as a sum modeled by factor graphs.

Factor Graph Examples

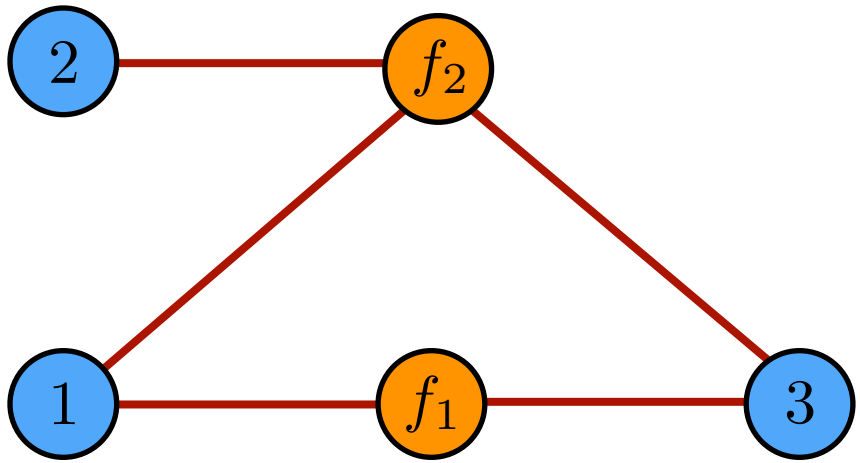
- Pairwise Markov network decomposition:



$$h(x, y) = h_{f_1}(x, y_1, y_2) + h_{f_2}(x, y_2, y_3).$$

- Other decomposition:

$$h(x, y) = h_{f_1}(x, y_1, y_3) + h_{f_2}(x, y_1, y_2, y_3).$$



Factor Graphs

- $G = (V, F, E)$: factor graph.
- $\mathcal{N}(f)$: neighborhood of f.
- $\mathcal{Y}_f = \prod_{k \in \mathcal{N}(f)} \mathcal{Y}_k$: substructure set cross-product at f.
- Decomposition:

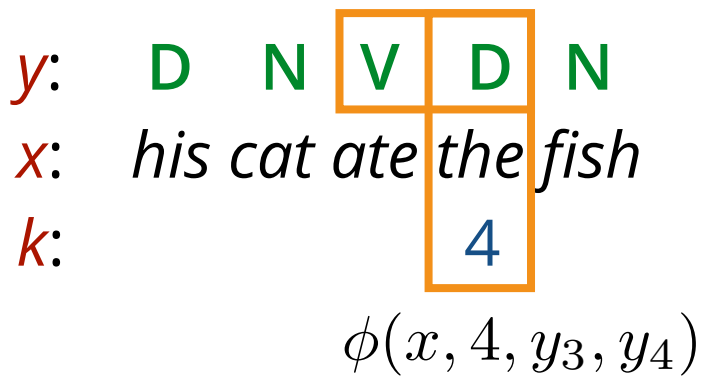
$$h(x, y) = \sum_{f \in F} h_f(x, y_f).$$

- More generally, example-dependent factor graph,

$$G_i = G(x_i, y_i) = (V_i, F_i, E_i).$$

Linear Hypotheses

- Feature decomposition → Hypothesis decomposition.
  - Example: bigram decomposition.



$$\mathbf{\Phi}(x, y) = \sum_{s=1}^{l} \phi(x, s, y_{s-1}, y_s).$$

$$h(x, y) = \mathbf{w} \cdot \mathbf{\Phi}(x, y) = \sum_{s=1}^{l} \underbrace{\mathbf{w} \cdot \phi(x, s, y_{s-1}, y_s)}_{h_s(x, y_{s-1}, y_s)}.$$

Structured Prediction Problem

- Training data: sample drawn i.i.d. from $\mathcal{X} \times \mathcal{Y}$ according to some distribution $\mathcal{D}$,

$$S = ((x_1, y_1), \ldots, (x_m, y_m)) \in \mathcal{X} \times \mathcal{Y}.$$

- Problem: find hypothesis $h \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ in $\mathcal{H}$ with small expected loss:

$$R(h) = \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}} [\mathsf{L}(\mathsf{h}(x), y)].$$

  - learning guarantees?
  - role of factor graph?
  - better algorithms?

This Talk

- Theory.
- Voted risk minimization (VRM).
- Algorithms.
- Experiments.

Theory

Previous Work

- Standard multi-class learning bounds:
  - number of classes is exponential!
- Structured prediction bounds:
  - covering number bounds: Hamming loss, linear hypotheses (Taskar et al., 2003).
  - PAC-Bayesian bounds (randomized algorithms) (David McAllester, 2007).

→ can we derive learning guarantees for general hypothesis sets and general loss functions?

Factor Graph Complexity

- Empirical factor graph complexity for hypothesis set $\mathcal{H}$ and sample $S = (x_1, \ldots, x_m)$:

$$\widehat{\mathfrak{R}}_S^G(\mathcal{H}) = \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}} \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{m} \sum_{f \in F_i} \sum_{y \in \mathcal{Y}_f} \sqrt{|F_i|}\, \epsilon_{i,f,y}\, h_f(x_i, y) \right]$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \underbrace{\left[\begin{array}{c} \vdots \\ \epsilon_{i,f,y} \\ \vdots \end{array} \right] \cdot \left[\begin{array}{c} \vdots \\ \sqrt{|F_i|} h_f(x_i, y) \\ \vdots \end{array} \right]}_{\text{correlation with random noise}} \right].$$

- Factor graph complexity:

$$\mathfrak{R}_m^G(\mathcal{H}) = \mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} \left[\widehat{\mathfrak{R}}_S^G(\mathcal{H}) \right].$$

Margin

- Definition: the margin of h at a labeled point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is

$$\rho_h(x, y, y') = \min_{y' \neq y} h(x, y) - h(x, y').$$

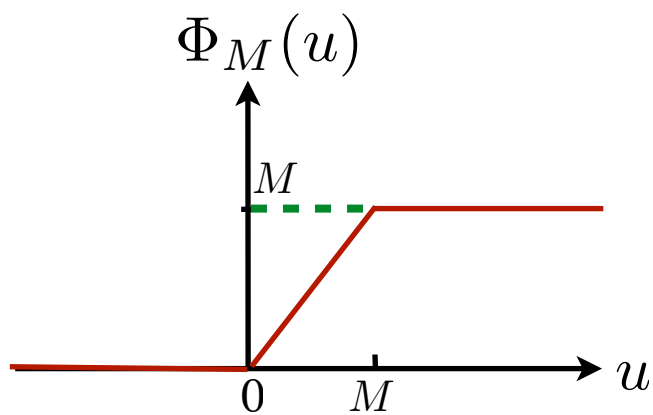
  - error when $\rho_h(x, y, y') \leq 0$.
  - small margin interpreted as low confidence.

Empirical Margin Losses

- For any $\rho > 0$,

$$\widehat{R}^{\text{add}}_{S,\rho}(h) = \mathop{\mathbb{E}}_{(x,y)\sim S}\left[\Phi_M\left(\max_{y'\neq y} \mathsf{L}(y',y) - \frac{h(x,y)-h(x,y')}{\rho}\right)\right]$$

$$\widehat{R}^{\text{mult}}_{S,\rho}(h) = \mathop{\mathbb{E}}_{(x,y)\sim S}\left[\Phi_M\left(\max_{y'\neq y} \mathsf{L}(y',y)\left(1 - \frac{h(x,y)-h(x,y')}{\rho}\right)\right)\right],$$



Generalization Bounds

- Theorem: for any $\delta > 0$, with probability at least $1 - \delta$, each of the following holds for all $h \in \mathcal{H}$:

$$R(h) \leq \widehat{R}^{\text{add}}_{S,\rho}(h) + \frac{4\sqrt{2}}{\rho}\mathfrak{R}^{G}_{m}(\mathcal{H}) + M\sqrt{\frac{\log \frac{1}{\delta}}{2m}},$$

$$R(h) \leq \widehat{R}^{\text{mult}}_{S,\rho}(h) + \frac{4\sqrt{2}M}{\rho}\mathfrak{R}^{G}_{m}(\mathcal{H}) + M\sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

  - tightest margin bounds for structured prediction.
  - data-dependent.
  - improve upon bound of (Taskar et al., 2003) by log terms (in the special case they study).

Linear Hypotheses

- Hypothesis set used by most convex structured prediction algorithms (StructSVM, M3N, CRF):

$$\mathcal{H}_p = \left\{ x \mapsto \mathbf{w} \cdot \mathbf{\Psi}(x, y) \colon \mathbf{w} \in \mathbb{R}^N, \|\mathbf{w}\|_p \leq \Lambda_p \right\},$$

with $p \geq 1$ and $\mathbf{\Psi}(x, y) = \sum_{f \in F} \mathbf{\Psi}_f(x, y_f)$.

Complexity Bounds

- Bounds on factor graph complexity of linear hypothesis sets:

$$\widehat{\mathfrak{R}}_S^G(\mathcal{H}_1) \leq \frac{\Lambda_1 r_\infty \sqrt{s \log(2N)}}{m}$$

$$\widehat{\mathfrak{R}}_S^G(\mathcal{H}_2) \leq \frac{\Lambda_2 r_2 \sqrt{\sum_{i=1}^m \sum_{f \in F_i} \sum_{y \in \mathcal{Y}_f} |F_i|}}{m}$$

with $r_q = \max_{i,f,y} \|\Psi_f(x_i, y)\|_q$

$$s = \max_{j \in [1,N]} \sum_{i=1}^m \sum_{f \in F_i} \sum_{y \in \mathcal{Y}_f} |F_i| 1_{\Psi_{f,j}(x_i,y) \neq 0}.$$

Key Term

- Sparsity parameter:

$$s \leq \sum_{i=1}^{m} \sum_{f \in F_i} \sum_{y \in \mathcal{Y}_f} |F_i| \leq \sum_{i=1}^{m} |F_i|^2 d_i \leq m \max_i |F_i|^2 d_i,$$

where $d_i = \max_{f \in F_i} |\mathcal{Y}_f|$.

- factor graph complexity in $O\left(\sqrt{\log(N) \max_i |F_i|^2 d_i / m}\right)$ for hypothesis set $\mathcal{H}_1$.
- key term: average factor graph size.

NLP Applications

- Features:
  - $\Psi_{f,j}$ is often a binary function, non-zero for a single pair $(x, y) \in \mathcal{X} \times \mathcal{Y}_f$.
  - example: presence of n-gram (indexed by j) at position f of the output with input sentence x_i.
  - complexity term only in $O\big(\max_i |F_i| \sqrt{\log(N)/m}\big)$.

Theory Takeaways

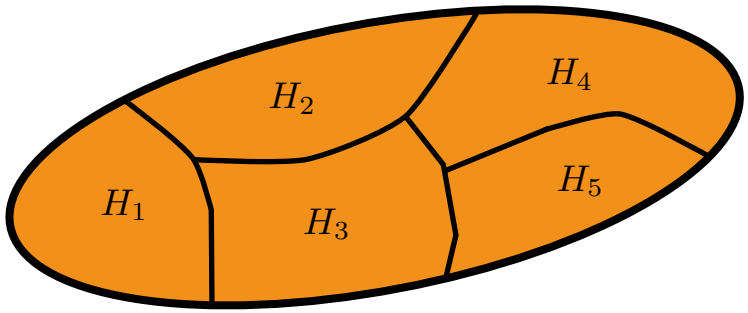
- Key generalization terms:
  - average size of factor graphs.
  - empirical margin loss.
- But, is learning with very complex hypothesis sets (factor graph complexity) possible?
  - richer families needed for difficult NLP tasks.
  - but generalization bound indicates risk of overfitting.

→ Voted Risk Minimization (VRM) theory.

Voted Risk Minimization

Decomposition of *H*

- Decomposition in terms of sub-families.

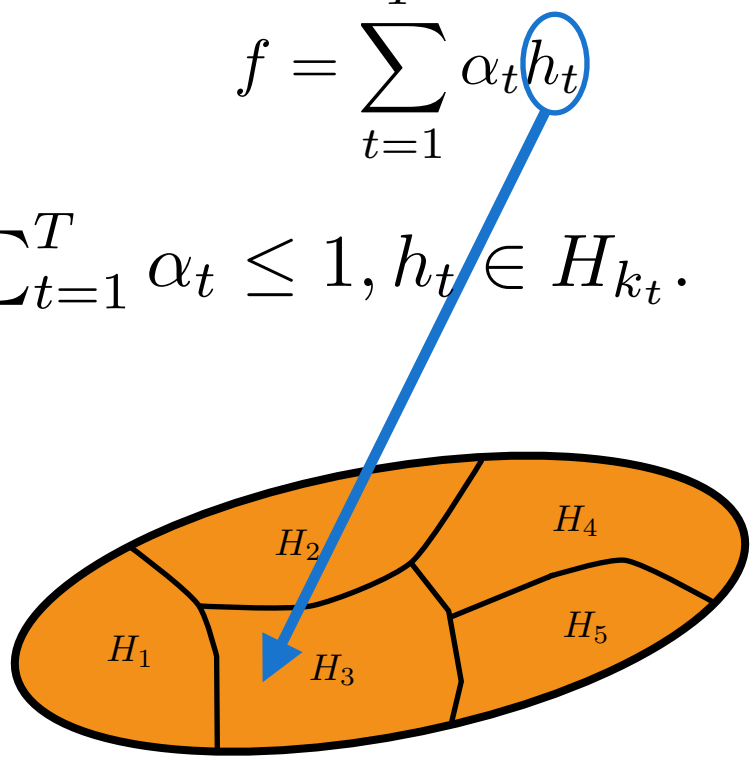


Ensemble Family

- Non-negative linear ensembles $\mathcal{F} = \mathrm{conv}(\cup_{k=1}^{p} H_k)$:

$$f = \sum_{t=1}^{T} \alpha_t h_t$$

with $\alpha_t \geq 0, \sum_{t=1}^{T} \alpha_t \leq 1, h_t \in H_{k_t}$.



Ideas

(Cortes, MM, and Syed, 2014)

- Use hypotheses drawn from H_ks with larger ks but allocate more weight to hypotheses drawn from smaller ks.
  - how can we determine quantitatively the amounts of mixture weights apportioned to different families?
  - can we provide learning guarantees guiding these choices?

Learning Guarantee

- Theorem: Fix $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $f = \sum_{t=1}^{T} \alpha_t h_t \in \mathcal{F}$:

$$R(f) - \widehat{R}^{\text{add}}_{S,\rho,1}(f) \leq \frac{4\sqrt{2}}{\rho} \sum_{t=1}^{T} \alpha_t \mathfrak{R}^G_m(H_{k_t}) + \widetilde{O}\left(M\sqrt{\frac{\log p}{\rho^2 m}}\right)$$

$$R(f) - \widehat{R}^{\text{mult}}_{S,\rho,1}(f) \leq \frac{4\sqrt{2}M}{\rho} \sum_{t=1}^{T} \alpha_t \mathfrak{R}^G_m(H_{k_t}) + \widetilde{O}\left(M\sqrt{\frac{\log p}{\rho^2 m}}\right).$$

Consequences

- Complexity term with explicit dependency on mixture weights.
  - quantitative guide for controlling weights assigned to more complex sub-families.
  - bound can be used to directly define an ensemble algorithm.

Algorithms

Surrogate Loss Framework

- Lemma: assume that $u\,1_{v\leq 0} \leq \Phi_u(v)$ for any $u \in \mathbb{R}_+$ and $v \in \mathbb{R}$. Then, for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\mathsf{L}(\mathsf{h}(x), y) \leq \max_{y' \neq y} \Phi_{\mathsf{L}(y', y)}(h(x, y) - h(x, y')).$$

- Proof: if $\mathsf{h}(x) = y$, then $\mathsf{L}(\mathsf{h}(x), y) = 0$ and result is trivial. Otherwise, $\mathsf{h}(x) \neq y$ and

$$\begin{aligned}
\mathsf{L}(\mathsf{h}(x), y) &= \mathsf{L}(\mathsf{h}(x), y) 1_{h(x,y) - \max_{y' \neq y} h(x,y') \leq 0} \\
&\leq \Phi_{\mathsf{L}(\mathsf{h}(x), y)}(h(x, y) - \max_{y' \neq y} h(x, y')) \quad (\Phi_u(v) \text{ upper bound on } u1_{v\leq 0}) \\
&= \Phi_{\mathsf{L}(\mathsf{h}(x), y)}(h(x, y) - h(x, \mathsf{h}(x))) \\
&\leq \max_{y' \neq y} \Phi_{\mathsf{L}(y', y)}(h(x, y) - h(x, y')). \quad (\mathsf{h}(x) \neq y)
\end{aligned}$$

Application

- Convex surrogate losses:
  - $\Phi_u(v) = \max(0, u(1-v))$: StructSVM (Tsochantaridis et al., 2005).
  - $\Phi_u(v) = \max(0, u-v)$: M3N (Taskar et al., 2003).
  - $\Phi_u(v) = \log(1+e^{u-v})$: CRF (Lafferty et al., 2003).
  - $\Phi_u(v) = ue^{-v}$: StructBoost (Cortes et al., 2016).

Voted Cond. Random Field

- Hypothesis set:
  - linear functions: $h\colon (x, y) \mapsto \mathbf{w} \cdot \mathbf{\Phi}(x, y)$.
  - complex feature vector $\mathbf{\Phi}$.
  - decomposition in blocks: $\mathbf{\Phi} = \begin{bmatrix} \Phi_1 \\ \vdots \\ \Phi_p \end{bmatrix}$.
- Upper bound:

$$\max_{y' \neq y} \log(1 + e^{\mathsf{L}(y,y') - \mathbf{w} \cdot (\mathbf{\Psi}(x,y) - \mathbf{\Psi}(x,y'))})$$

$$\leq \log \left(\sum_{y' \in \mathcal{Y}} e^{\mathsf{L}(y,y') - \mathbf{w} \cdot (\mathbf{\Psi}(x,y) - \mathbf{\Psi}(x,y'))} \right).$$

Voted Cond. Random Field

- Optimization problem (VCRF):

$$\min_{\mathbf{w}} \frac{1}{m} \sum_{i=1}^{m} \log \left(\sum_{y \in \mathcal{Y}} e^{\mathsf{L}(y, y_i) - \mathbf{w} \cdot (\mathbf{\Psi}(x_i, y_i) - \mathbf{\Psi}(x_i, y))} \right) + \sum_{k=1}^{p} (\lambda r_k + \beta) \|\mathbf{w}_k\|_1,$$

with $r_k = r_\infty |F(k)| \sqrt{\log N}$.

- solution via stochastic gradient descent (SGD).
- efficient gradient computation for Markovian features.
- relationship with L1-CRF.
- other regularization, e.g., L2-VCRF.

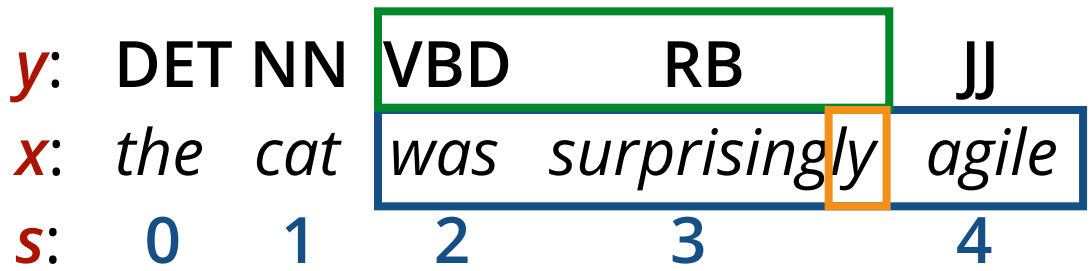
Experiments

Preliminary Experiments

- Part-of-speech tagging.
- Multiple data sets.

Dataset	Full name	Sentences	Tokens	Unique tokens	Labels
Basque	Basque UD Treebank	8993	121443	26679	16
Chinese	Chinese Treebank 6.0	28295	782901	47570	37
Dutch	UD Dutch Treebank	13735	200654	29123	16
English	UD English Web Treebank	16622	254830	23016	17
Finnish	Finnish UD Treebank	13581	181018	53104	12
Finnish-FTB	UD_Finnish-FTB	18792	160127	46756	15
Hindi	UD Hindi Treebank	16647	351704	19232	16
Tamil	UD Tamil Treebank	600	9581	3583	14
Turkish	METU-Sabanci Turkish Treebank	5635	67803	19125	32
Twitter	Tweebank	929	12318	4479	25

Features - Example



$$h_1(x) = 1_{x_2=\text{'was'}, x_3=\text{'surprisingly'}, x_4=\text{'agile'}}(x)$$

$$h_2(y) = 1_{y_2=\text{'VBD'}, y_3=\text{'RB'}}(y)$$

$$h_3(x) = 1_{\text{suff}(x_3,2)=\text{'ly'}}(x).$$

Features

- Feature families:
  - definition: for each choice of the window sizes (k_1, k_2, k_3), sum of products of indicators over positions along the sequence.
  - complexity:

$$r(H_{k_1,k_2,k_3}) \leq \sqrt{\frac{2(k_1 \log |V| + k_2 \log |\Delta| + k_3 \log |\Sigma|}{m}}.$$

Experiments

- Parameters λ and β determined via cross-validation.
- Comparison with L1-CRF.
- Two sets of results:
  - original data sets.
  - artificial noise added: tokens corresponding to features that commonly appear in the dataset (at least five times), POS labels flipped with some probability (20% noise).

Experimental Results

	VCRF error (%)		**CRF error(%)**	
Dataset	Token	Sentence	Token	Sentence
Basque	**7.26 ± 0.13**	**57.67 ± 0.82**	7.68 ± 0.20	59.78 ± 1.39
Chinese	**7.38 ± 0.15**	**67.73 ± 0.46**	7.67 ± 0.12	68.88 ± 0.49
Dutch	5.97 ± 0.08	49.27 ± 0.71	6.01 ± 0.92	49.48 ± 1.02
English	5.51 ± 0.04	44.40 ± 1.30	5.51 ± 0.06	44.32 ± 1.31
Finnish	**7.48 ± 0.05**	**55.96 ± 0.64**	7.86 ± 0.13	57.17 ± 1.36
Finnish-FTB	**9.79 ± 0.22**	**51.23 ± 1.21**	10.55 ± 0.22	52.98 ± 0.75
Hindi	**4.84 ± 0.10**	**51.69 ± 1.07**	4.93 ± 0.08	53.18 ± 0.75
Tamil	**19.82 ± 0.69**	**89.83 ± 2.13**	22.50 ± 1.57	92.00 ± 1.54
Turkish	**11.28 ± 0.40**	**59.63 ± 1.55**	11.69 ± 0.37	61.15 ± 1.01
Twitter	**17.98 ± 1.25**	**75.57 ± 1.25**	19.81 ± 1.09	76.96 ± 1.37

Average No. of Features

Dataset	VCRF	CRF	Ratio
Basque	7028	94712653	0.00007
Chinese	219736	552918817	0.00040
Dutch	2646231	2646231	1.00000
English	4378177	357011992	0.01226
Finnish	32316	89333413	0.00036
Finnish-FTB	53337	5735210	0.00930
Hindi	108800	448714379	0.00024
Tamil	1583	668545	0.00237
Turkish	498796	3314941	0.15047
Twitter	18371	26660216	0.00069

Experimental Results

	VCRF error (%)		**CRF error(%)**	
Dataset	Token	Sentence	Token	Sentence
Basque	**9.13 ± 0.18**	**67.43 ± 0.93**	9.42 ± 0.31	68.61 ± 1.08
Chinese	**96.43 ± 0.33**	**100.00 ± 0.01**	96.81 ± 0.43	100.00 ± 0.01
Dutch	**8.16 ± 0.52**	**62.15 ± 1.77**	8.57 ± 0.30	63.55 ± 0.87
English	**8.79 ± 0.23**	**61.27 ± 1.21**	9.20 ± 0.11	63.60 ± 1.18
Finnish	**9.38 ± 0.27**	**64.96 ± 0.89**	9.62 ± 0.18	65.91 ± 0.93
Finnish-FTB	**11.39 ± 0.29**	**72.56 ± 1.30**	11.76 ± 0.25	73.63 ± 1.19
Hindi	**6.63 ± 0.51**	**63.84 ± 2.86**	7.85 ± 0.33	71.93 ± 1.20
Tamil	20.77 ± 0.70	93.00 ± 1.35	21.36 ± 0.86	93.50 ± 1.78
Turkish	14.28 ± 0.46	69.72 ± 1.51	14.31 ± 0.53	69.62 ± 2.04
Twitter	90.92 ± 1.67	100.00 ± 0.00	92.27 ± 0.71	100.00 ± 0.00

Conclusion

- Structured prediction theory:
  - tightest margin guarantees for structured prediction.
  - general loss functions, data-dependent.
  - key notion of factor graph complexity.
  - guarantees for complex hypothesis sets (VRM theory).
  - VCRF and StructBoost algorithms.
  - favorable preliminary experiments.
  - additionally, tightest margin bounds for standard classification.