GRAPH-BASED POSTERIOR REGULARIZATION FOR SEMI-SUPERVISED STRUCTURED PREDICTION

Luheng He             Jennifer Gillenwater†             Ben Taskar
†University of Pennsylvania     University of Washington
OVERVIEW

Structured Prediction (CRF)

Graph Propagation

A Joint Objective

\[ \mathcal{J}(q, p_{\theta}) \]

ninth run for

\( y_5 \)

\( y_7 \)
This is recognized as a face
This is recognized as a face
$X = \text{This is recognized as a face}$

$Y = \begin{array}{cccccc}
    & y_1 & y_2 & y_3 & y_4 & y_5 & y_6 \\
\end{array}$
$X = \text{This is recognized as a face}$

$Y = \{y_1, y_2, y_3, y_4, y_5, y_6\}$

$f(y_t, y_{t-1}, x)$
\[
X = \text{This is recognized as a face}
\]

\[
Y = \begin{align*}
&y_1 & y_2 & y_3 & y_4 & y_5 & y_6
\end{align*}
\]

Conditional Distribution

\[
p_\theta(y \mid x) = \frac{1}{Z_\theta(x)} \exp \left[ \sum_{t=1}^{T} \theta^\top f(y_t, y_{t-1}, x) \right]
\]

\[p\text{-factor}\]
$X = \text{This is recognized as a face}$

$Y = \{y_1, y_2, y_3, y_4, y_5, y_6\}$

Conditional Distribution

$$p_\theta(y \mid x) = \frac{1}{Z_\theta(x)} \exp \left[ \sum_{t=1}^{T} \theta^\top f(y_t, y_{t-1}, x) \right]$$

CRF objective

$$\text{NLik}(p_\theta) = - \sum_{i=1}^{\ell} \log p_\theta(y^i \mid x^i)$$
This is recognized as a face.

The painting is considered as a work of genius.

Zhu et al. (ICML 2013): Graph-based Semi-supervised Learning
This is recognized as a face.

The painting is considered as a work of genius.

Zhu et al. (ICML 2013): Graph-based Semi-supervised Learning
The painting is considered as a work of genius.

This is recognized as a face.

similar tagging

VERB

similar context
The painting is considered as a work of genius.

This is recognized as a face.

similar tagging

VERB

similar context

Zhu et al. (ICML 2013): Graph-based Semi-supervised Learning
... a run along ...

... a run for ...

... ninth run for ...

... luck run out ...

GRAPH LAPLACIAN REGULARIZER
... a run along ...

... a run for ...

... ninth run for ...

... luck run out ...

GRAPH LAPLACIAN REGULARIZER
... luck run out ...

NOUN
... a run along ...

NOUN
... a run for ...

? ... ninth run for ...

VERB

0.4

0.8

0.8

GRAPH LAPLACIAN REGULARIZER
... a run along ...

Prob (NOUN | ninth run for) = 0.6
Prob (VERB | ninth run for) = 0.4
Prob (ADV | ninth run for) = 0
Prob (DET | ninth run for) = 0
...

... luck run out ...
\[
\text{Prob}(\text{tag} \mid \text{ninth run for}) = \\
\arg \min_m 0.4 \times \| m - \text{Prob}(\text{tag} \mid \text{a run along}) \|^2_2 \\
+ 0.8 \times \| m - \text{Prob}(\text{tag} \mid \text{a run for}) \|^2_2 \\
+ 0.8 \times \| m - \text{Prob}(\text{tag} \mid \text{luck run out}) \|^2_2
\]
GRAPH LAPLACIAN REGULARIZER

\[
\min_m \text{ Lap}(m) = \sum_{a \in \text{Unlab}} \sum_{b \in \text{Neighbors}(a)} \sum_{k \in \text{Tags}} w_{ab}(m_{a,k} - m_{b,k})^2
\]

\[
\text{Prob}(\text{tag} | \text{a run along for}) = 0.8 \times \| m - \text{Prob}(\text{tag} | \text{a run for}) \|_2^2
\]

\[
m_{a,k} : \text{The proportion of time trigram } a \text{ has tag } k
\]
**COMBINING THE TWO**

<table>
<thead>
<tr>
<th>Data</th>
<th>CRF estimation</th>
<th>graph-propagation</th>
</tr>
</thead>
<tbody>
<tr>
<td>labeled</td>
<td>ninth run for</td>
<td>ninth run for</td>
</tr>
<tr>
<td>$y_5$</td>
<td>$y_7$</td>
<td></td>
</tr>
<tr>
<td>labeled + unlabeled</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>$p(tags \mid sentence; \theta)$</th>
<th>$m(tag \mid trigram)$</th>
</tr>
</thead>
</table>
PRIOR WORK

Subramanya et al. (EMNLP 2010)

graph-propagation + CRF estimation
PRIOR WORK

Subramanya et al. (EMNLP 2010)

Our work: retains efficiency while optimizing an extendible, joint objective.
HOW TO COMBINE?

introduce auxiliary variables q
HOW TO COMBINE?

introduce auxiliary variables $q$

$$p_{\theta}(y \mid x^i) = \frac{1}{Z_{\theta}(x^i)} \exp \left[ \sum_{t=1}^{T} \theta^T f(y_t, y_{t-1}, x^i) \right]$$

$$q^i_y = \frac{1}{Z_q(x^i)} \exp \left[ \sum_{t=1}^{T} r_{i,t}(y_t, y_{t-1}) \right]$$
HOW TO COMBINE?

introduce auxiliary variables q

1. Normalized

\[ p_{\theta}(y \mid x^i) = \frac{1}{Z_{\theta}(x^i)} \exp \left[ \sum_{t=1}^{T} \theta^\top f(y_t, y_{t-1}, x^i) \right] \]

\[ q_y^i = \frac{1}{Z_q(x^i)} \exp \left[ \sum_{t=1}^{T} r_{i,t}(y_t, y_{t-1}) \right] \]
HOW TO COMBINE?

introduce auxiliary variables q

1. Normalized 2. Decomposed into local factors

\[
p_\theta(y \mid x^i) = \frac{1}{Z_\theta(x^i)} \exp \left[ \sum_{t=1}^{T} \theta^\top f(y_t, y_{t-1}, x^i) \right]
\]

\[
q_y^i = \frac{1}{Z_q(x^i)} \exp \left[ \sum_{t=1}^{T} r_{i,t}(y_t, y_{t-1}) \right]
\]
JOINT OBJECTIVE

\[ J(q, p_{\theta}) = \text{Lap}(q) + \text{NLik}(p_{\theta}) + \text{KL}(q \parallel p_{\theta}) \]
JOINT OBJECTIVE

\[ \mathcal{J}(q, p_\theta) = \text{Lap}(q) + \text{NLik}(p_\theta) + \text{KL}(q \parallel p_\theta) \]

\[ \text{Lap}(q) = \sum_{a \in \text{Unlab}} \sum_{b \in \text{Neighbors}(a)} \sum_{k \in \text{Tags}} w_{ab}(m_{a,k}(q) - m_{b,k}(q))^2 \]
JOINT OBJECTIVE

\[ J(q, p_\theta) = \text{Lap}(q) + \text{NLik}(p_\theta) + \text{KL}(q \parallel p_\theta) \]

\[ \text{NLik}(p_\theta) = - \sum_{i=1}^{\ell} \log p_\theta(y^i | x^i) \]
JOINT OBJECTIVE

\[ \mathcal{J}(q, p_\theta) = \text{Lap}(q) + \text{NLik}(p_\theta) + \text{KL}(q \parallel p_\theta) \]

\[ \text{KL}(q \parallel p_\theta) = \sum_{i=1}^{n} \sum_{y} q_{y}^{i} \log \frac{q_{y}^{i}}{p_\theta(y \mid x^{i})} \]
HOW TO OPTIMIZE?

\[
\min_{q,\theta} J(q, p_\theta)
\]
HOW TO OPTIMIZE?

\[
\min_{q, \theta} J(q, p_{\theta})
\]

\[\Delta\]

unconstrained
HOW TO OPTIMIZE?

\[
\min_{q, \theta} J(q, p_\theta)
\]

\[
\Delta \quad \text{unconstrained}
\]

\[
p \text{ update:}
\theta' = \theta - \eta \frac{\partial J(q, p_\theta)}{\partial \theta}
\]
HOW TO OPTIMIZE?

\[
\min_{q, \theta} \mathcal{J}(q, p_{\theta})
\]

\[
\Delta \quad \text{unconstrained}
\]

\[p\text{ update:}
\]

\[
\theta' = \theta - \eta \frac{\partial \mathcal{J}(q, p_{\theta})}{\partial \theta}
\]

\[q\text{ update:}
\]

projection is hard \[\sum_y q_y^i = 1\]

no compact form \((\# \tags)^{i's \ length} \) values
UPDATE Q
**UPDATE Q**

$q$ can be represented by local factors $r$

\[ q^i_y = \frac{1}{Z_q(x^i)} \exp \left[ \sum_{t=1}^{T} r_{i,t}(y_t, y_{t-1}) \right] \]
UPDATE Q

\( q \) can be represented by local factors \( r \)

\[
q_y^i = \frac{1}{Z_q(x^i)} \exp \left[ \sum_{t=1}^{T} r_{i,t}(y_t, y_{t-1}) \right]
\]

doing an additive gradient update

\[
q_y^i' = q_y^i - \eta \frac{\partial \mathcal{J}(q, p_\theta)}{\partial q_y^i}
\]
**UPDATE Q**

\( \mathbf{q} \) can be represented by local factors \( \mathbf{r} \)

\[
q^i_y = \frac{1}{Z_q(x^i)} \exp \left[ \sum_{t=1}^{T} r_{i,t}(y_t, y_{t-1}) \right]
\]

doing an additive gradient update

\[
q^i_y' = q^i_y - \eta \frac{\partial \mathcal{J}(q, p_{\theta})}{\partial q^i_y}
\]

\( \mathbf{q}' \) cannot be written as product of local factors!
EXPONENTIATED GRADIENT

multiplicative gradient update:

\[ q_{y'}^i = \frac{1}{Z_{q'}(x^i)} q^i_{y} \exp \left[ -\eta \frac{\partial \mathcal{J}(q, p_{\theta})}{\partial q^i_{y}} \right] \]
EXPONENTIATED GRADIENT

multiplicative gradient update:

\[
q^i_y' = \frac{1}{Z_{q'}(x^i)} \exp \left[ \sum_{t=1}^{T} r_{i,t}(y_t, y_{t-1}) \right] \exp \left[ -\eta \frac{\partial \mathcal{J}(q, p_\theta)}{\partial q^i_y} \right]
\]

Collins et al. (JMLR 2008): Exponentiated gradient for CRFs
EXPONENTIATED GRADIENT

Collins et al. (JMLR 2008): Exponentiated gradient for CRFs

multiplicative gradient update:

\[ q_i' = \frac{1}{Z_{q'}(x^i)} \exp \left[ \sum_{t=1}^{T} r_{i,t}(y_t, y_{t-1}) \right] \exp \left[ -\eta \frac{\partial \mathcal{J}(q, p_\theta)}{\partial q_{y_i}} \right] \]

decompose into local factors
EXPONENTIATED GRADIENT

multiplicative gradient update:

\[ q^{i'}_y = \frac{1}{Z_{q'}(x^i)} \exp \left[ \sum_{t=1}^{T} r_{i,t}(y_t, y_{t-1}) \right] \exp \left[ -\eta \frac{\partial \mathcal{J}(q, p_\theta)}{\partial q^i_y} \right] \]

\[ = \frac{1}{Z_{q'}(x^i)} \exp \left[ \sum_{t=1}^{T} r'_{i,t}(y_t, y_{t-1}) \right] \]

decompose into local factors

Collins et al. (JMLR 2008): Exponentiated gradient for CRFs
Collins et al. (JMLR 2008): Exponentiated gradient for CRFs

Multiplicative gradient update:

\[
q_{y}^{i'} = \frac{1}{Z_{q'}(x^{i})} \exp \left[ \sum_{t=1}^{T} r_{i,t}(y_{t}, y_{t-1}) \right] \exp \left[ -\eta \frac{\partial \mathcal{J}(q, p_{\theta})}{\partial q_{y}^{i}} \right]
\]

Decompose into local factors

\[
= \frac{1}{Z_{q'}(x^{i})} \exp \left[ \sum_{t=1}^{T} r'_{i,t}(y_{t}, y_{t-1}) \right]
\]

Only updating \((#\text{tags})^2 \times (i's \text{ length})\) variables!
$J(q, p_\theta) = \text{Lap}(q) + \text{NLik}(p_\theta) + \text{KL}(q \parallel p_\theta)$
\[ \mathcal{J}(q, p_\theta) = \text{Lap}(q) + \text{NLik}(p_\theta) + \text{KL}(q \parallel p_\theta) \]

**M-step:** \( \theta' = \theta - \eta \frac{\partial \mathcal{J}(q, p_\theta)}{\partial \theta} \)

**E-step:** \( q_{y}^{i} = \frac{1}{Z_q(x^i)} q_{y}^{i} \exp \left[ -\eta \frac{\partial \mathcal{J}(q, p_\theta)}{\partial q_{y}^{i}} \right] \)

(update each \( r_{i,t} \) in practice)
\[ \mathcal{J}(q, p_\theta) = \text{Lap}(q) + \text{NLik}(p_\theta) + \text{KL}(q \parallel p_\theta) \]

**Theorem:** Converges to a local optimum of \( \mathcal{J}(q, p_\theta) \)

M-step: \[ \theta' = \theta - \eta \frac{\partial \mathcal{J}(q, p_\theta)}{\partial \theta} \]

E-step: \[ q_{y}' = \frac{1}{Z_q(x^i)} q_y \exp \left[ -\eta \frac{\partial \mathcal{J}(q, p_\theta)}{\partial q_y} \right] \]

(update each \( r_{i,t} \) in practice)
EXPERIMENT SETTING

10 Languages (CoNLL-X and CoNLL-2007)
100 Randomly sampled labeled sentences
Averaged over 10 sampling runs
Universal POS Tags (Petrov et al. 2011)
Second Order CRF Model \( f(y_t, y_{t-1}, y_{t-2}, x) \)
graph propagation

POS Tagging Error

Language: EN, DE, ES, PT, DA, SL, SV, EL, IT, NL, Avg

GP
ninth run for

POS Tagging Error

Language

EN  DE  ES  PT  DA  SL  SV  EL  IT  NL  Avg

GP  GP → CRF
ninth run for
POS Tagging Error

Language

GP
GP → CRF
CRF
J

28% average relative error reduction
CONCLUSION

\[ J(q, p_\theta) = \text{Lap}(q) + \text{NLik}(p_\theta) + \text{KL}(q \parallel p_\theta) \]
CONCLUSION

\[ \mathcal{J}(q, p_\theta) = \mathcal{R}(q) + \text{NLik}(p_\theta) + \text{KL}(q \parallel p_\theta) \]

any convex, differentiable regularizer
CONCLUSION

\[ J(q, p_\theta) = R(q) + \text{NLik}(p_\theta) + \text{KL}(q \parallel p_\theta) \]

any convex, differentiable regularizer

Code: [https://code.google.com/p/pr-graph/](https://code.google.com/p/pr-graph/)