GRAPH-BASED POSTERIOR REGULARIZATION FOR SEMI-SUPERVISED STRUCTURED PREDICTION

Luheng He Jennifer Gillenwater Ben Taskar
University of Pennsylvania University of Washington
GRAPH-BASED LEARNING
GRAPH-BASED LEARNING
GRAPH-BASED LEARNING
GRAPH-BASED LEARNING

Labels: verb (V), noun (N), etc.
Labels: verb (V), noun (N), etc.

they run over

blood run cold

we run out

luck run out

a run for

ninth run for

a run along
GRAPH-BASED LEARNING

Labels: verb (V), noun (N), etc.

they run over

blood run cold

we run out

luck run out

a run for

ninth run for

a run along
GRAPH-BASED LEARNING

Labels: verb (V), noun (N), etc.

they run over

blood run cold

we run out

luck run out

ninth run for

a run for

a run along
GRAPH-BASED LEARNING

Labels: verb (V), noun (N), etc.

they run over

blood run cold

we run out

luck run out

ninth run for

a run for

a run along

\[
|| \text{N} \ - \ 
\text{V} ||^2 \leq \frac{2}{2}
\]
GRAPH-BASED LEARNING

Labels: verb (V), noun (N), etc.

they run over

blood run cold

we run out

luck run out

ninth run for

a run for

a run along

0.3 || N V - V N || 2/2
GRAPH-BASED LEARNING

Labels: verb (V), noun (N), etc.

They run over

Blood run cold
We run out

Luck run out

A run for

Ninth run for

A run along

\[ w_{ab} \| q_a - q_b \|_2^2 \]
GRAPH-BASED LEARNING

Labels: verb (V), noun (N), etc.

Lap(q) = \sum_{ab} w_{ab} \|q_a - q_b\|_2^2
Labels: verb (V), noun (N), etc.

\[
\text{Lap}(q) = \sum_{a=1}^{N} \sum_{b=L+1}^{N} w_{ab} \| q_a - q_b \|_2^2
\]
STRUCTURED PREDICTION

ninth run for
ninth run for
The soldiers of the ninth run for cover
The soldiers of the ninth run for cover
The soldiers of the ninth run for cover
The soldiers of the ninth run for cover

\[ f( y_t ) \]
The soldiers of the ninth run for cover

\[ f(y_t, y_{t-1}) \]
x = The soldiers of the ninth run for cover

\[ f(y_t, y_{t-1}, x) \]
\[ x = \text{The soldiers of the ninth run for cover} \]

\[ f(y_t, y_{t-1}, x) \]

\[ p\text{-factor} \]
**Structured Prediction**

\[ x = \text{The soldiers of the ninth run for cover} \]

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

\[ p \text{-factor} \]
**STRUCTURED PREDICTION**

\[ x = \text{The soldiers of the ninth run for cover} \]

**CRF**

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

\[
\text{NLLik}(p_\theta) = - \sum_{i=1}^{\ell} \log p_\theta(y^i \mid x^i)
\]
The soldiers of the cover

\[
x = \begin{array}{c}
\text{The soldiers of the} \\
\text{cover}
\end{array}
\]

\[
\text{CRF} \quad y_1 \quad y_2 \quad y_3 \quad y_4 \quad y_5 \quad y_7 \quad y_8
\]

\[
p_\theta(y_t \mid x)
\]
WHY COMBINE?

Each type of learning incorporates different information
WHY COMBINE?

Each type of learning incorporates different information

- **graph-propagation**
- **CRF estimation**

- ninth run for
- $y_5$
- $y_7$
WHY COMBINE?

Each type of learning incorporates different information

- Graph-propagation
- CRF estimation

<table>
<thead>
<tr>
<th>Data</th>
<th>unlabeled</th>
<th>labeled</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**WHY COMBINE?**

Each type of learning incorporates different information

<table>
<thead>
<tr>
<th></th>
<th>Data</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>unlabeled</td>
<td></td>
<td>trigram</td>
</tr>
<tr>
<td>labeled</td>
<td></td>
<td>sentence</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>graph-propagation</th>
<th>CRF estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ninth run for</td>
<td></td>
<td>ninth run for</td>
</tr>
<tr>
<td></td>
<td>$y_5$</td>
<td>$y_7$</td>
</tr>
</tbody>
</table>
PRIOR WORK
Lap\((q)\) graph-propagation

ninth run for
PRIOR WORK

\[ \text{Lap}(q) \text{ graph-propagation} \quad + \quad \text{CRF estimation} \quad \text{NLik}(p_\theta) \]
PRIOR WORK

Subramanya et al. (EMNLP 2010)

\[ \text{Lap}(q) \text{ graph-propagation} + \text{CRF estimation} \quad \text{NLik}(p_\theta) \]
PRIOR WORK

Subramanya et al. (EMNLP 2010)

\[ \text{Lap}(q) \text{ graph-propagation} \quad + \quad \text{CRF estimation} \quad \text{NLik}(\theta) \]

ninth run for

ninth run for
PRIOR WORK

Subramanya et al. (EMNLP 2010)

\[ \text{Lap}(q) \text{ graph-propagation } + \text{CRF estimation } \text{NLik}(p_{\theta}) \]

This work: retains efficiency while optimizing an extendible, joint objective.
JOINT OBJECTIVE
JOINT OBJECTIVE

ninth run for

\text{Lap}(q)
JOINT OBJECTIVE

\[ \text{Lap}(q) + \text{NLik}(p_{\theta}) \]
Joint Objective

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

The soldiers of the ninth run for cover
\[ J(q, p_\theta) = \text{Lap}(q) + \text{NLik}(p_\theta) + \text{KL}(q \parallel p_\theta) \]

The soldiers of the ninth run for cover

\[ (\# \text{ tags})^8 \]

\[ \begin{array}{cccccccc}
\end{array} \]

\ldots
The soldiers of the ninth run for cover.

\[ J(q, p_\theta) = \text{Lap}(q) + \text{NLik}(p_\theta) + \text{KL}(q \parallel p_\theta) \]
The soldiers of the ninth run for cover.

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

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

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

unconstrained
OPTIMIZATION

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

\[\Delta\]

unconstrained
OPTIMIZATION

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

update:

\[ p \] update:

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

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

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

\[p \text{ update:}\]

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

Next 3 slides: Why several common techniques don’t work for updating \(q\)
OPTIMIZATION

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

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

$q$ update:

$$q_y' = q_y - \eta \frac{\partial \mathcal{J}(q, p_{\theta})}{\partial q^i_y}$$
OPTIMIZATION

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

\textbf{q} \textbf{ update:}

\[
q_y' = \text{proj}_\Delta \left( q_y - \eta \frac{\partial \mathcal{J}(q, p_\theta)}{\partial q_y} \right)
\]
OPTIMIZATION

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

\( q \) update:

\[ q^i_y' = \text{proj}_\Delta \left( q^i_y - \eta \frac{\partial \mathcal{J}(q, p_{\theta})}{\partial q^i_y} \right) \]

\( q^i \in \Delta \) of dimension (\# tags)\(^{(i's \ length)}\)
OPTIMIZATION

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

$q$ update:

$$q^i_y' = \text{proj}_\Delta\left( q^i_y - \eta \frac{\partial \mathcal{J}(q, p_\theta)}{\partial q^i_y} \right)$$

$q^i \in \Delta$ of dimension ($\#\ tags$)($i$’s length)

-Problem 1: projection is hard $q^i \not\in \Delta$
**OPTIMIZATION**

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

**q** update:

\[
q^{'i}_y = \text{proj}_\Delta \left( q^i_y - \eta \frac{\partial \mathcal{J}(q, p_\theta)}{\partial q^i_y} \right)
\]

\(q^i \in \Delta \) of dimension (\# tags)(i’s length)

- Problem 1: projection is hard
  \(q^i \not\in \Delta\)

- Problem 2: no compact form
  (\# tags)(i’s length) values
OPTIMIZATION

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

\[q\]\text { update:}

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

\[q^i \in \Delta \text{ of dimension } (\# \text{ tags})^{(i\text{'s length})}\]

-Problem 1: projection is hard
-Problem 2: no compact form
DUAL OPTIMIZATION

\[ J(q, p_\theta) \]
DUAL OPTIMIZATION

\[ \mathcal{J}(q, p_{\theta}) + \gamma \left( \sum_{y} q_{y}^{i} - 1 \right) \]
DUAL OPTIMIZATION

\[ \mathcal{J}(q, p_{\theta}) + \gamma \left( \sum_{y} q_{y}^{i} - 1 \right) \]

Posterior Regularization (PR) uses dual 
Ganchev et al. (JMLR 2010)
DUAL OPTIMIZATION

\[ J(q, p_\theta) + \gamma \left( \sum_y q_y^i - 1 \right) \]

Posterior Regularization (PR) uses dual
Ganchev et al. (JMLR 2010)

This work: \textbf{Lap}(q)
DUAL OPTIMIZATION

\[ \mathcal{J}(q, p_\theta) + \gamma \left( \sum_y q_y^i - 1 \right) \]

Posterior Regularization (PR) uses dual
Ganchev et al. (JMLR 2010)

This work: \textbf{Lap}(q) \quad \rightarrow \quad \text{Standard PR: simpler}
DUAL OPTIMIZATION

\[ \mathcal{J}(q, p_\theta) + \gamma \left( \sum_y q_y^i - 1 \right) \]

Posterior Regularization (PR) uses dual
Ganchev et al. (JMLR 2010)

This work: \text{Lap}(q) \quad \rightarrow \quad \text{Standard PR: simpler}

\[ p \text{-factors} \]

\[ y_t, y_{t-1}, x \]
DUAL OPTIMIZATION

\[ J(q, p_\theta) + \gamma \left( \sum_y q_y^i - 1 \right) \]

Posterior Regularization (PR) uses dual
Ganchev et al. (JMLR 2010)

This work: Lap(q) → Standard PR: Linear(m)
DUAL OPTIMIZATION

\[ J(q, p_\theta) + \gamma \left( \sum_y q_y^i - 1 \right) \]

Posterior Regularization (PR) uses dual
Ganchev et al. (JMLR 2010)

This work: Lap(q) → Standard PR: Linear(m)
DUAL OPTIMIZATION

\[ J(q, p_\theta) + \gamma \left( \sum_y q_y^i - 1 \right) \]

Posterior Regularization (PR) uses dual
Ganchev et al. (JMLR 2010)

This work: \textbf{Lap}(q) \quad \rightarrow \quad \text{Standard PR: Linear}(m)
DUAL OPTIMIZATION

\[ \mathcal{J}(q, p_\theta) + \gamma \left( \sum_y q_y^i - 1 \right) \]

Posterior Regularization (PR) uses dual
Ganchev et al. (JMLR 2010)

This work: \( \text{Lap}(q) \) → Standard PR: \( \text{Linear}(m) \)
DUAL OPTIMIZATION

\[ J(q, p_\theta) + \gamma \left( \sum_y q_y^i - 1 \right) \]

Posterior Regularization (PR) uses dual
Ganchev et al. (JMLR 2010)

This work: Lap(q) → Standard PR: Linear(m)

Lap(m), a quadratic function
DUAL OPTIMIZATION

\[ \mathcal{J}(q, p_\theta) + \gamma \left( \sum_y q_y^i - 1 \right) \]

Posterior Regularization (PR) uses dual
Ganchev et al. (JMLR 2010)

This work: \( \text{Lap}(q) \) \quad \text{Standard PR:} \quad \text{Linear}(m) \)

\( \text{Lap}(m) \), a quadratic function
Dual of quadratic requires:

\[
\begin{pmatrix}
1 & 2 & \cdots & N \\
1 & 2 & \cdots & N \\
\vdots & \vdots & \ddots & \vdots \\
1 & 2 & \cdots & N \\
\end{pmatrix}^{-1}
\]
Posterior Regularization (PR) uses dual
Ganchev et al. (JMLR 2010)

This work: \( \text{Lap}(q) \rightarrow \) Standard PR: \( \text{Linear}(m) \)

\[ J(q, p_\theta) + \gamma \left( \sum_y q_y^i - 1 \right) \]

Lap\((m)\), a quadratic function
Dual of quadratic requires:
\[
\begin{pmatrix}
1 & 2 & \cdots & N \\
1 & 2 & \cdots & N \\
\vdots & & \ddots & \\
1 & 2 & \cdots & N
\end{pmatrix}^{-1}
\]
EXPONENTIATED GRADIENT
EXPONENTIATED GRADIENT

\[ q_i' \propto q_i \exp \left[ -\eta \frac{\partial \mathcal{J}(q,p_\theta)}{\partial q_i} \right] \]
EXPONENTIATED GRADIENT

\[ q_y' \propto q_y \exp \left[ -\eta \frac{\partial \mathcal{J}(q,p\theta)}{\partial q_y} \right] \]

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

\[ q^i_y' \propto q^i_y \exp \left[ -\eta \frac{\partial \mathcal{I}(q,p_\theta)}{\partial q^i_y} \right] \]

\[ \exp \left[ -\eta \frac{\partial \mathcal{I}(q,p_\theta)}{\partial q^i_y} \right] = \]
EXPONENTIATED GRADIENT

\[ q^i_y' \propto q^i_y \exp \left[ -\eta \frac{\partial J(q,p_\theta)}{\partial q^i_y} \right] \]

\[ \exp \left[ -\eta \frac{\partial J(q,p_\theta)}{\partial q^i_y} \right] = \]

\[ \exp \left[ -\eta \sum_{t=1}^{T} \frac{\partial \text{Lap}(m^i_y)}{\partial m^i_{t,y_t,y_{t-1}}} \right] \]
EXPONENTIATED GRADIENT

\[ q^i_y' \propto q^i_y \exp \left[ -\eta \frac{\partial \mathcal{J}(q,p\theta)}{\partial q^i_y} \right] \]

\[
\exp \left[ -\eta \frac{\partial \mathcal{J}(q,p\theta)}{\partial q^i_y} \right] = \exp \left[ -\eta \sum_{t=1}^{T} \frac{\partial \text{Lap}(m^i_y)}{\partial m^i_{t,y_t,y_{t-1}}} \right]
\]

product of p-factors
EXPONENTIATED GRADIENT

\[ q_y^i' \propto q_y^i \exp \left[ -\eta \frac{\partial \mathcal{J}(q,p)}{\partial q_y^i} \right] \]

\[
\exp \left[ -\eta \frac{\partial \mathcal{J}(q,p)}{\partial q_y^i} \right] =
\exp \left[ -\eta \sum_{t=1}^{T} \frac{\partial \text{Lap}(m^i_y)}{\partial m^i_{t,y_t,y_{t-1}}} \right] p_{\theta}(y \mid x^i)^\eta(q_y^i)^{-\eta e}
\]

product of p-factors
EXPONENTIATED GRADIENT

\[ q^i_y' \propto q^i_y \exp \left[ -\eta \frac{\partial \mathcal{J}(q,p\theta)}{\partial q^i_y} \right] \]

\[ \exp \left[ -\eta \frac{\partial \mathcal{J}(q,p\theta)}{\partial q^i_y} \right] = \]

\[ \exp \left[ -\eta \sum_{t=1}^{T} \frac{\partial \text{Lap}(m^i_y)}{\partial m^i_{y_t, y_{t-1}}} \right] p_\theta(y \mid x^i) \eta(q^i_y)^{-\eta \epsilon} \]

product of p-factors
EXPONENTIATED GRADIENT

\( q_y^i' \propto q_y^i \exp \left[ -\eta \frac{\partial J(q,p_\theta)}{\partial q_y^i} \right] \)

\[
\exp \left[ -\eta \frac{\partial J(q,p_\theta)}{\partial q_y^i} \right] = \\
\exp \left[ -\eta \sum_{t=1}^{T} \frac{\partial \text{Lap}(m_y^i)}{\partial m_{t,y_t,y_{t-1}}^i} \right] \frac{p_\theta(y | x^i)^{\eta(q_y^i)}}{-\eta e} \]

product of p-factors
EXPONENTIATED GRADIENT

\[ q^i_y' \propto q^i_y \exp \left[ -\eta \frac{\partial \mathcal{I}(q,p_{\theta})}{\partial q^i_y} \right] \]

\[
\exp \left[ -\eta \frac{\partial \mathcal{I}(q,p_{\theta})}{\partial q^i_y} \right] = \exp \left[ -\eta \sum_{t=1}^{T} \frac{\partial \text{Lap}(m^i_{y})}{\partial m^i_{t,y_t,y_{t-1}}} \right] p_{\theta}(y \mid x^i) \eta(q^i_y)^{-\eta e} \]

product of p-factors

\[ \text{proj}_{\Delta} \]
EXPONENTIATED GRADIENT

\[ q_y^i' \propto q_y^i \exp \left[ -\eta \frac{\partial \mathcal{I}(q,p_{\theta})}{\partial q_y^i} \right] \]

\[ \exp \left[ -\eta \frac{\partial \mathcal{I}(q,p_{\theta})}{\partial q_y^i} \right] = \exp \left[ -\eta \sum_{t=1}^{T} \frac{\partial \text{Lap}(m_y^i)}{\partial m_{y_t, y_{t-1}}^i} \right] p_{\theta}(y | x^i) \eta(q_y^i)^{-\eta e} \]

product of p-factors

\[ \text{proj}_\Delta \rightarrow Z_q(x^i) \]
EXPONENTIATED GRADIENT

\[ q_y^i' \propto q_y^i \exp \left[ -\eta \frac{\partial \mathcal{J}(q, p_\theta)}{\partial q_y^i} \right] \]

\[
\exp \left[ -\eta \frac{\partial \mathcal{J}(q, p_\theta)}{\partial q_y^i} \right] = \\
\exp \left[ -\eta \sum_{t=1}^{T} \frac{\partial \text{Lap}(m_y^i)}{\partial m_{t,y_t,y_{t-1}}^i} \right] p_\theta(y | x^i) \eta(q_y^i)^{-\eta e}
\]

product of p-factors

proj_{\Delta} \rightarrow Z_q(x^i), \text{ computable via forward-backward}
\[ I(q, p_\theta) = \text{Lap}(q) + \text{NLik}(p_\theta) + \text{KL}(q \| p_\theta) \]
SUMMARY

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

\[ \theta' = \theta - \eta \frac{\partial \mathcal{I}(q, p_{\theta})}{\partial \theta} \]
\[ \mathcal{J}(q, p_\theta) = \text{Lap}(q) + \text{NLik}(p_\theta) + \text{KL}(q \parallel p_\theta) \]

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

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


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

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

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

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

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

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

Theorem:
Converges to a local optimum of \[ \mathcal{J}(q, p_{\theta}) \]
\[ J(q, p_\theta) = \text{Lap}(q) + \text{NLik}(p_\theta) + \text{KL}(q \parallel p_\theta) \]

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

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

**Theorem:**
Converges to a local optimum of \( J(q, p_\theta) \)
graph-propagation

ninth run for
ninth run for graph-propagation

POS Tagging Error

Language

EN DE ES PT DA SL SV EL IT NL Avg

GP
ninth run for

graph-propagation

POS Tagging Error

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

GP 100 labeled sentences
ninth run for

GP → CRF

POS Tagging Error

Language

EN DE ES PT DA SL SV EL IT NL Avg

GP

GP → CRF
ninth run for

Language

POS Tagging Error

GP

GP → CRF

CRF
KL

POS Tagging Error

Language

EN DE ES PT DA SL SV EL IT NL Avg

GP GP → CRF CRF J

28% avg relative gain
QUESTIONS?
QUESTIONS?

Code: https://code.google.com/p/pr-graph/