Dynamic Tree Block Coordinate Ascent

Daniel Tarlow\textsuperscript{1}, Dhruv Batra\textsuperscript{2}
Pushmeet Kohli\textsuperscript{3}, Vladimir Kolmogorov\textsuperscript{4}

1: University of Toronto  3: Microsoft Research Cambridge
2: TTI Chicago  4: University College London

International Conference on Machine Learning (ICML), 2011
MAP in Large Discrete Models

- Many important problems can be expressed as a discrete Random Field (MRF, CRF)
- MAP inference is a fundamental problem

\[
\min_{x \in X} E(x) = \min_{x \in X} \sum_{i \in V} \theta_i(x_i) + \sum_{(i,j) \in E} \theta_{ij}(x_i, x_j)
\]
**Primal and Dual**

<table>
<thead>
<tr>
<th>Primal</th>
<th>Dual</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ \min_x \sum_{A \in \mathcal{V} \cup \mathcal{E}} \theta_A(x_A) \geq \sum_{A \in \mathcal{V} \cup \mathcal{E}} \min_{x_A} \tilde{\theta}<em>A(x_A) = \sum</em>{A \in \mathcal{V} \cup \mathcal{E}} h^*_A ]</td>
<td></td>
</tr>
</tbody>
</table>

- Dual is a lower bound: less constrained version of primal
- \( \tilde{\theta} \) is a *reparameterization*, determined by messages
- \( h^*_A \) is *height* of unary or pairwise potential

- Definition of reparameterization:
  \[ \sum_{A \in \mathcal{V} \cup \mathcal{E}} \theta_A(x_A) = \sum_{A \in \mathcal{V} \cup \mathcal{E}} \tilde{\theta}_A(x_A) \quad \forall \{x_A\} \]

  LP-based message passing: find reparameterization to maximize dual
Standard Linear Program-based Message Passing

• Max Product Linear Programming (MPLP)
  – Update edges in fixed order

• Sequential Tree-Reweighted Max Product (TRW-S)
  – Sequentially iterate over variables in fixed order

• Tree Block Coordinate Ascent (TBCA) [Sontag & Jaakkola, 2009]
  – Update trees in fixed order

**Key:** these are all *energy oblivious*

Can we do better by being *energy aware*?
Example

TBCA with Static Schedule: 630 messages needed

TBCA with Dynamic Schedule: 276 messages needed
Benefit of Energy Awareness

**Static settings**
- Not all graph regions are equally difficult
- Repeating computation on easy parts is wasteful

**Dynamic settings (e.g., learning, search)**
- Small region of graph changes.
- Computation on unchanged part is wasteful
References and Related Work

• [Elidan et al., 2006], [Sutton & McCallum, 2007]
  – Residual Belief Propagation. Pass most different messages first.

• [Chandrasekaran et al., 2007]
  – Works only on continuous variables. Very different formulation.

• [Batra et al., 2011]
  – Local Primal Dual Gap for Tightening LP relaxations.

• [Kolmogorov, 2006]
  – Weak Tree Agreement in relation to TRW-S.

• [Sontag et al., 2009]
  – Tree Block Coordinate Descent.
Visualization of reparameterized energy $\tilde{\theta}$

States for each variable: red (R), green (G), or blue (B)

"Good" local settings: (can assume "good" has cost 0, otherwise cost 1)

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>G-G, B-B</td>
<td>G</td>
<td>G-G, B-B</td>
<td>G</td>
</tr>
</tbody>
</table>

$\begin{array}{c}
X_1 \\
X_2 \\
X_3 \\
X_4
\end{array}$

$\begin{array}{c}
R \\
G \\
B
\end{array}$

VisualizaJon of reparameterized energy
States for each variable: red (R), green (G), or blue (B)

"Good" local settings:

<table>
<thead>
<tr>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>G-G, B-B</td>
<td>G</td>
<td>G-G, B-B</td>
</tr>
<tr>
<td>&quot;Don't be R or B&quot;</td>
<td>&quot;Don't be R or B&quot;</td>
<td>&quot;Don't be R or B&quot;</td>
<td>&quot;Don't be R or G&quot;</td>
</tr>
</tbody>
</table>

Holographic messages that e.g. residual max-product would send.
States for each variable: red (R), green (G), or blue (B)

"Good" local settings:

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>G-G, B-B</td>
<td>G</td>
<td>G-G, B-B</td>
<td>G</td>
</tr>
</tbody>
</table>

But we don't need to send any messages. We are at the global optimum.

Our scores (see later slides) are 0, so we wouldn't send any messages here.
Visualization of reparameterized energy $\tilde{\theta}$

States for each variable: red (R), green (G), or blue (B)

"Good" local settings:

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>G-G, B-B</td>
<td>G-G, B-B</td>
<td>B</td>
</tr>
<tr>
<td>G</td>
<td>G-B, B-B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
</tbody>
</table>

Change unary potentials (e.g., during learning or search)
States for each variable: red (R), green (G), or blue (B)

"Good" local settings:


Locally, best assignment for some variables change.
Visualization of reparameterized energy $\tilde{\theta}$

States for each variable: red (R), green (G), or blue (B)

"Good" local settings:

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>G-G, B-B</td>
<td>B</td>
<td>G-G, B-B</td>
</tr>
<tr>
<td>&quot;Don't be R or G&quot;</td>
<td>&quot;Don't be R or G&quot;</td>
<td>&quot;Don't be R or G&quot;</td>
<td>&quot;Don't be R or G&quot;</td>
</tr>
</tbody>
</table>

Hypothetical messages that e.g. residual max-product would send.
"Good" local settings:

But we don't need to send any messages. We are at the global optimum.

Our scores (see later slides) are 0, so we wouldn't send any messages here.
VisualizaJon of reparameterized energy $\tilde{\theta}$

"Good" local settings:

$B$, $G\text{-}G, B\text{-}B$

$B$

$x_1$

$\leftrightarrow$

$x_2$

$\leftrightarrow$

$x_3$

$\leftrightarrow$

$x_4$

Possible fix: look at how much sending messages on edge would improve dual.

- Would work in above case, but incorrectly ignores e.g. the subgraph below:

"Good" local settings:

$B$

$B\text{-}B$

$G, B$

$G\text{-}G$

$R, G$

$R\text{-}R$

$R$

$x_1$

$\leftrightarrow$

$x_2$

$\leftrightarrow$

$x_3$

$\leftrightarrow$

$x_4$

$R$

$G$

$B$
Locally, everything looks optimal
Try assigning a value to each variable
Our main contribution

Use primal (and dual) information to choose regions on which to pass messages

Try assigning a value to each variable
Our Formulation

• Measure primal-dual local agreement at edges and variables
  – Local Primal Dual Gap (LPDG).
  – Weak Tree Agreement (WTA).

• Choose forest with maximum disagreement
  – Kruskal's algorithm, possibly terminated early

• Apply TBCA update on maximal trees

**Important! Minimize overhead.**
Use quantities that are already computed during inference, and carefully cache computations
Local Primal-Dual Gap (LPDG) Score

• Difference between primal and dual objectives
  – Given primal assignment $x^p$ and dual variables (messages) defining $\tilde{\Theta}$, primal-dual gap is

$$\text{Primal-dual gap} = \sum_{A \in \mathcal{V} \cup \mathcal{E}} \theta_A(x^p_A) - \sum_{A \in \mathcal{V} \cup \mathcal{E}} \min_{x_A} \tilde{\Theta}_A(x_A)$$

$$= \sum_{A \in \mathcal{V} \cup \mathcal{E}} \left( \tilde{\Theta}_A(x^p_A) - \min_{x_A} \tilde{\Theta}_A(x_A) \right) = \sum_{A \in \mathcal{V} \cup \mathcal{E}} \text{LPDG} (A)$$

- Primal cost of node/edge
- Dual bound at node/edge

e: “local disagreement” measure: $e_A = \text{LPDG} (A)$
Shortcoming of LPDG Score: Loose Relaxations

\[
\text{Filled circle means } \tilde{\theta}_i(x_i) = h_i^*, \text{ black edge means } \tilde{\theta}_{ij}(x_i, x_j) = h_{ij}^*
\]

LPDG > 0, but dual optimal
Weak Tree Agreement (WTA) [Kolmogorov 2006]

Reparameterized potentials $\tilde{\theta}$ are said to satisfy WTA if there exist non-empty subsets $D_i \subseteq X_i$ for each node $i$ such that

$$\tilde{\theta}_i(x_i) = h_i^*$$
$$\forall x_i \in D_i$$

$$\min_{x_j \in D_j} \tilde{\theta}_{ij}(x_i, x_j) = h_{ij}^*$$
$$\forall x_i \in D_i, (i, j) \in E$$

Filled circle means $\tilde{\theta}_i(x_i) = h_i^*$
Black edge means $\tilde{\theta}_{ij}(x_i, x_j) = h_{ij}^*$

At Weak Tree Agreement

Not at Weak Tree Agreement
**Weak Tree Agreement (WTA)** [Kolmogorov 2006]

Reparameterized potentials $\tilde{\theta}$ are said to satisfy WTA if there exist non-empty subsets $D_i \subseteq X_i$ for each node $i$ such that

$$\tilde{\theta}_i(x_i) = h_i^*$$

$$\min_{x_j \in D_j} \tilde{\theta}_{ij}(x_i, x_j) = h_{ij}^*$$

$\forall x_i \in D_i$  $\forall x_i \in D_i, (i, j) \in E$

Filled circle means $\tilde{\theta}_i(x_i) = h_i^*$  Black edge means $\tilde{\theta}_{ij}(x_i,x_j) = h_{ij}^*$

At Weak Tree Agreement

$D_1=\{0\}$  $D_2=\{0,2\}$  $D_2=\{0,2\}$  $D_3=\{0\}$

Not at Weak Tree Agreement

$D_1=\{0\}$  $D_2=\{2\}$  $D_2=\{0,2\}$  $D_3=\{0\}$
Weak Tree Agreement (WTA) [Kolmogorov 2006]

Reparameterized potentials $\tilde{\theta}$ are said to satisfy WTA if there exist non-empty subsets $D_i \subseteq X_i$ for each node $i$ such that

\[
\tilde{\theta}_i(x_i) = h_i^* \\
\min_{x_j \in D_j} \tilde{\theta}_{ij}(x_i, x_j) = h_{ij}^* \\
\forall x_i \in D_i \\
\forall x_i \in D_i, (i, j) \in E
\]

Filled circle means $\tilde{\theta}_i(x_i) = h_i^*$  
Black edge means $\tilde{\theta}_{ij}(x_i, x_j) = h_{ij}^*$

Filled circle means $\tilde{\theta}_i(x_i) = h_i^*$  
Black edge means $\tilde{\theta}_{ij}(x_i, x_j) = h_{ij}^*$

At Weak Tree Agreement

\[
D_1 = \{0\} \quad D_2 = \{2\} \quad D_2 = \{0, 2\} \quad D_3 = \{0\}
\]

Not at Weak Tree Agreement
WTA Score

e: “local disagreement” measure

\[
e_{ij} = \max_{x_i \in D_i, x_j \in D_j} \min \tilde{\theta}_{ij}(x_i, x_j) - \min_{x_i, x_j} \tilde{\theta}_{ij}(x_i, x_j)
\]

Costs:
solid – low
dotted – medium
else – high

\[
D_2 = \{0, 2\}
D_3 = \{0, 2\}
\]

\[
e_{23} = \max(\min(a, high), \min(b, c)) - c
\]

Filled circle means \(\tilde{\theta}_i(x_i) = h_i^*\), black edge means \(\tilde{\theta}_{ij}(x_i, x_j) = h_{ij}^*\)
WTA Score

e: “local disagreement” measure

\[ e_{ij} = \max_{x_i \in D_i, x_j \in D_j} \min_{x_i, x_j} \tilde{\theta}_{ij}(x_i, x_j) - \min_{x_i, x_j} \tilde{\theta}_{ij}(x_i, x_j) \]

Costs:
- solid – low
- dotted – medium
- else – high

\[ D_2 = \{0, 2\} \]
\[ D_3 = \{0, 2\} \]

\[ e_{23} = \max \left( \min(a, \text{high}), \min(b, c) \right) - c \]

Filled circle means \( \tilde{\theta}_i(x_i) = h_i^* \), black edge means \( \tilde{\theta}_{ij}(x_i, x_j) = h_{ij}^* \)
**WTA Score**

\[ e_{ij} = \max_{x_i \in D_i} \min_{x_j \in D_j} \tilde{\theta}_{ij}(x_i, x_j) - \min_{x_i, x_j} \tilde{\theta}_{ij}(x_i, x_j) \]

 Costs:
- solid – low
- dotted – medium
- else – high

\[ D_2 = \{0, 2\} \]
\[ D_3 = \{0, 2\} \]

\[ e_{23} = \max(a, c) - c = a - c \]

Filled circle means \( \tilde{\theta}_i(x_i) = h_i^* \), black edge means \( \tilde{\theta}_{ij}(x_i, x_j) = h_{ij}^* \)
WTA Score

e: “local disagreement” measure

\[ e_{ij} = \max_{x_i \in D_i} \min_{x_j \in D_j} \tilde{\theta}_{ij}(x_i, x_j) - \min_{x_i, x_j} \tilde{\theta}_{ij}(x_i, x_j) \]

Costs:
solid – low
dotted – medium
else – high

D_2 = \{0, 2\}
D_3 = \{0, 2\}

\[ e_{23} = \max(a, c) - c = a - c \]

Filled circle means \( \tilde{\theta}_i(x_i) = h_i^* \), black edge means \( \tilde{\theta}_{ij}(x_i, x_j) = h_{ij}^* \)
WTA Score

e: “local disagreement” measure: node measure

\[ e_i = \max_{x_i \in D_i} \tilde{\theta}_i(x_i) - \min_{x_i} \tilde{\theta}_i(x_i) \]
Single Formulation of LPDG and WTA

- Set a max history size parameter R.
- Store most recent R labelings of variable $i$ in label set $D_i$

  $R=1$: LPDG score. 
  $R>1$: WTA score.

Combine scores into undirected edge score:

$$w_{ij} = \max(e_{ij}, e_{ji}) + e_i + e_j$$
Properties of LPDG/WTA Scores

- LPDG measure gives upper bound on possible dual improvement from passing messages on forest.

- LPDG may overestimate "usefulness" of an edge e.g., on non-tight relaxations.

- WTA measure addresses overestimate problem: is zero shortly after normal message passing would converge.

- Both only change when messages are passed on nearby region of graph.
Experiments

Computer Vision:
• Stereo
• Image Segmentation
• Dynamic Image Segmentation

Protein Design:
• Static problem
• Correlation between measure and dual improvement
• Dynamic search application

Algorithms
• TBCA: Static Schedule, LPDG Schedule, WTA Schedule
• MPLP  [Sontag and Globerson implementation]
• TRW-S  [Kolmogorov Implementation]
Experiments: Stereo

383x434 pixels, 16 labels. Potts potentials.
Experiments: Image Segmentation

375x500 pixels, 21 labels. General potentials based on label co-occurrence.
Experiments: Dynamic Image Segmentation

Sheep

Previous Opt
Modify White Unaries

Sheep

New Opt
Heatmap of Messages

Warm-started DTBCA vs Warm-started TRW-S

375x500 pixels, 21 labels. Potts potentials.
Experiments: Dynamic Image Segmentation

375x500 pixels, 21 labels. Potts potentials.
Experiments: Protein Design

Dual Improvement vs. Measure on Forest

Other protein experiments: (see paper)
- DTBCA vs. static "stars" on small protein
  DTBCA converges to optimum in .39s vs TBCA in .86s
- Simulating node expansion in A* search on larger protein
  Similar dual for DTBCA in 5s as Warm-started TRW-S in 50s.

Protein Design from Yanover et al.
Discussion

• Energy oblivious schedules can be wasteful.

• For LP-based message passing, primal information is useful for scheduling.
  – We give two low-overhead ways of including it

• Biggest win comes from dynamic applications
  – Exciting future dynamic applications: search, learning, ...
Discussion

• Energy oblivious schedules can be wasteful.

• For LP-based message passing, primal information is useful for scheduling.
  – We give two low-overhead ways of including it

• Biggest win comes from dynamic applications
  – Exciting future dynamic applications: search, learning, ...
Unused slides
Schlesinger's Linear Program (LP)

\[
\min_{x \in X} \sum_{i \in V} \theta_i(x_i) + \sum_{ij \in E} \theta_{ij}(x_i, x_j)
\]

\[
\min_{\mu \in \mathcal{M}(G)} \sum_{i \in V} \mu_i(x_i) \theta_i(x_i) + \sum_{ij \in E} \mu_{ij}(x_i, x_j) \theta_{ij}(x_i, x_j)
\]

\[
\min_{\mu \in \mathcal{L}(G)} \sum_{i \in V} \mu_i(x_i) \theta_i(x_i) + \sum_{ij \in E} \mu_{ij}(x_i, x_j) \theta_{ij}(x_i, x_j)
\]

- **Real-valued**
- **Marginal polytope**
- **LOCAL polytope** (see next slide)
**Algorithm 1 Dynamic Tree-Block Coordinate Ascent**

\[ \hat{\mathcal{V}} \leftarrow \mathcal{V} \quad \{\text{Dirty nodes}\} \]
\[ \hat{\mathcal{E}} \leftarrow \mathcal{E} \quad \{\text{Dirty edges (see Sec. 5.1 for details)}\} \]
\[ R \leftarrow \text{RUN-LPDG} \quad ? \quad 1 : R_{WTA} \quad \{\text{History size}\} \]

for \( t = 1 : t_{\text{max}} \) do

for \( i \in \hat{\mathcal{V}} \) do {Node scores}

\[ x_i^p \leftarrow \arg \min_{x_i} \tilde{\theta}_i(x_i) \]
ADD-TO-HISTORY(\( x_i^p, D_i, R \))
\[ e_i \leftarrow \max_{x_i \in D_i} \tilde{\theta}_i(x_i) - \min_{x_i} \tilde{\theta}_i(x_i) \]
end for

for \( (i, j) \in \hat{\mathcal{E}} \) do {Directed edge scores}

\[ h_{ij} \leftarrow \min_{x_i, x_j} \tilde{\theta}_{ij}(x_i, x_j) \]
\[ e_{ij} \leftarrow \max_{x_i \in D_i} \min_{x_j \in D_j} \tilde{\theta}_{ij}(x_i, x_j) - h_{ij} \]
\[ e_{ji} \leftarrow \max_{x_j \in D_j} \min_{x_i \in D_i} \tilde{\theta}_{ij}(x_i, x_j) - h_{ij} \]
end for

for \( (i, j) \in \mathcal{E} \) do {Undirected edge scores}

\[ w_{ij} \leftarrow \max(e_{ij}, e_{ji}) + e_i + e_j \]
end for

\[ T \leftarrow \text{KRUSKAL-Forest}(w) \]
\[ \tilde{\theta} \leftarrow \text{REPARAMETERIZE-Forest}(T, \tilde{\theta}) \]
end for
WTA Score

\[ e_{ij} = \max_{x_i \in D_i} \min_{x_j \in D_j} \tilde{\theta}_{ij}(x_i, x_j) - \min_{x_i, x_j} \tilde{\theta}_{ij}(x_i, x_j) \]

Filled circle means \( \tilde{\theta}(x_i) = h_i^* \), black edge means \( \tilde{\theta}(x_i, x_j) = h_{ij}^* \)

\[ D_2 = \{0, 2\} \]
\[ D_3 = \{0\} \]

\[ e_{23} = \max(a, b) - c \quad e_{32} = \min(a, b) - c \]