Dynamic Tree Block Coordinate Ascent

Daniel Tarlow[1], Dhruv Batra[2]
Pushmeet Kohli[3], Vladimir Kolmogorov[4]

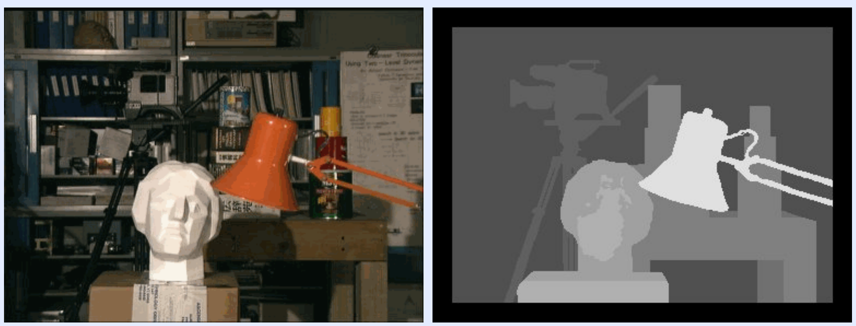
1: University of Toronto
2: TTI Chicago
3: Microsoft Research Cambridge
4: University College London

International Conference on Machine Learning (ICML), 2011

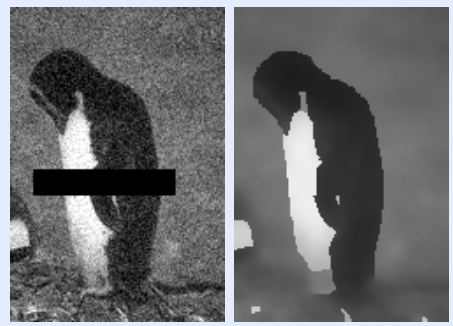
MAP in Large Discrete Models

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

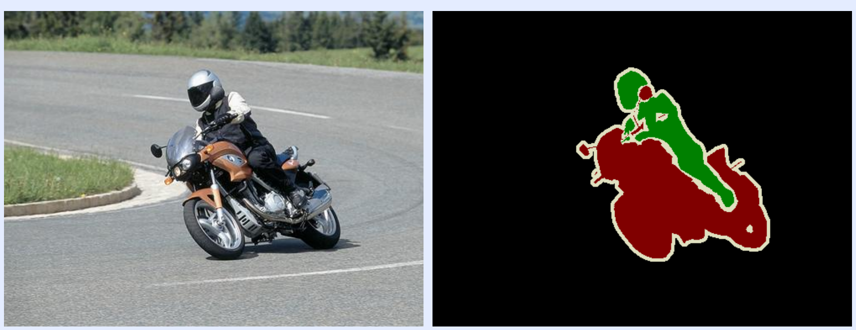
$$\min_{\boldsymbol{x} \in X} E(\boldsymbol{x}) = \min_{\boldsymbol{x} \in X} \sum_{i \in \mathcal{V}} \theta_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \theta_{ij}(x_i, x_j)$$



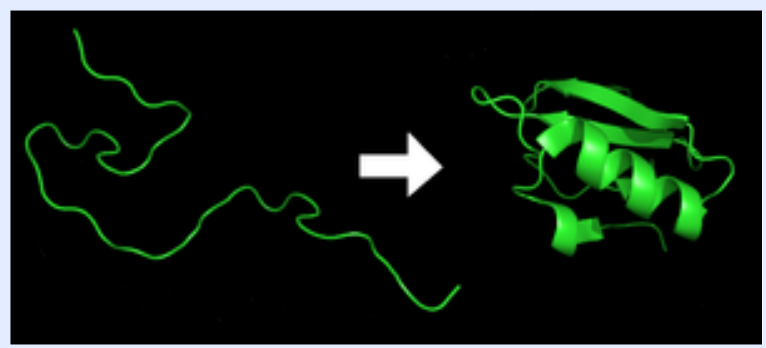
Stereo



Inpainting



Object Class Labeling



Protein Design / Side Chain Prediction

Primal and Dual

Primal   Dual

$$\min_{\mathbf{x}} \sum_{A \in \mathcal{V} \cup \mathcal{E}} \theta_A(x_A) \quad \geq \quad \sum_{A \in \mathcal{V} \cup \mathcal{E}} \min_{x_A} \tilde{\theta}_A(x_A) = \sum_{A \in \mathcal{V} \cup \mathcal{E}} h_A^*$$

- 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:

$$\sum_{A \in \mathcal{V} \cup \mathcal{E}} \theta_A(x_A) = \sum_{A \in \mathcal{V} \cup \mathcal{E}} \tilde{\theta}_A(x_A) \qquad \forall \{x_A\}$$

LP-based message passing: find reparameterization to maximize dual

Standard Linear Program-based Message Passing

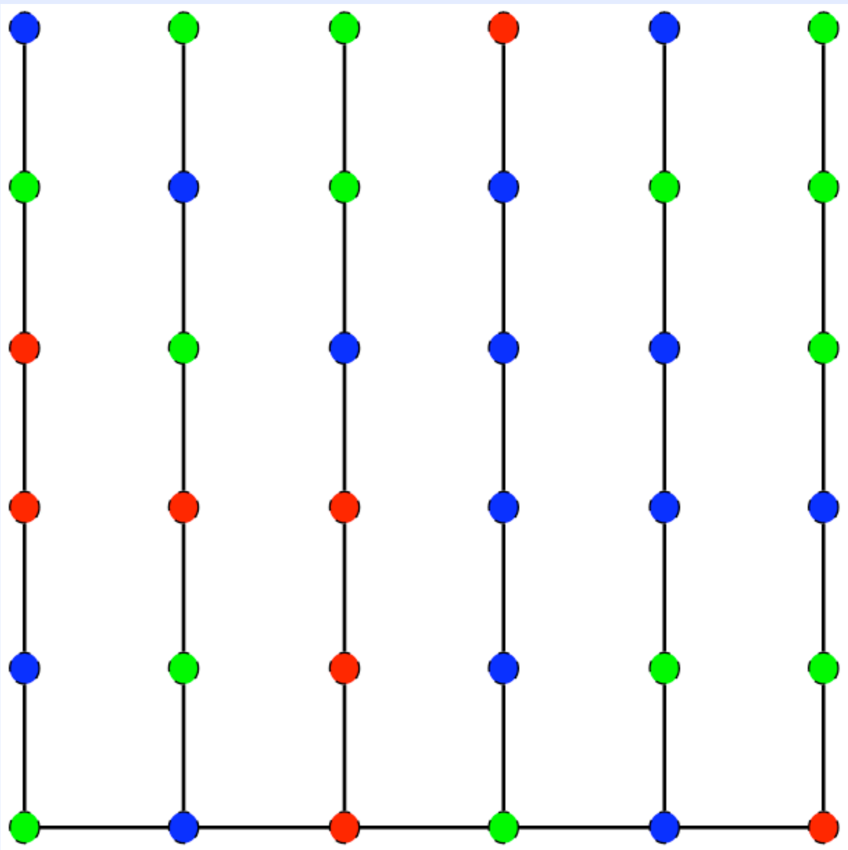
- Max Product Linear Programming (MPLP)
  - Update edges **in fixed order**
- Sequential Tree-Reweighted Max Product (TRW-S)
  - Sequentially iterate over variables **in fixed order**
- Tree Block Coordinate Ascent (TBCA) [Sontag & Jaakkola, 2009]
  - Update trees **in fixed order**

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

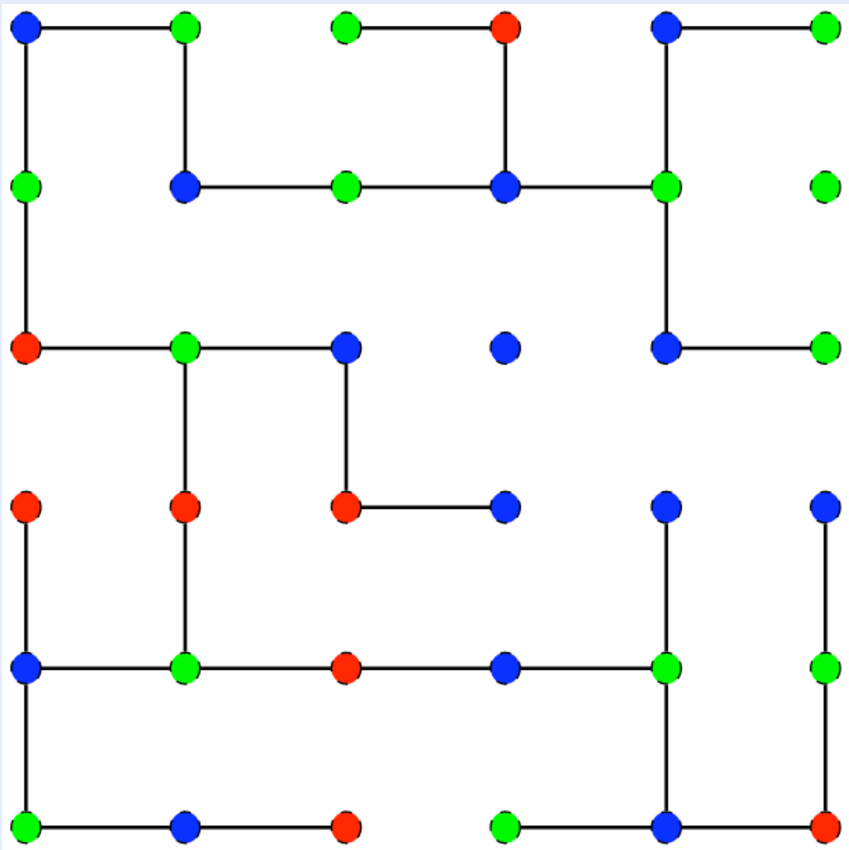
Can we do better by being ***energy aware***?

Example

TBCA with Static Schedule:
630 messages needed



TBCA with Dynamic Schedule:
276 messages needed



Benefit of Energy Awareness

Static settings

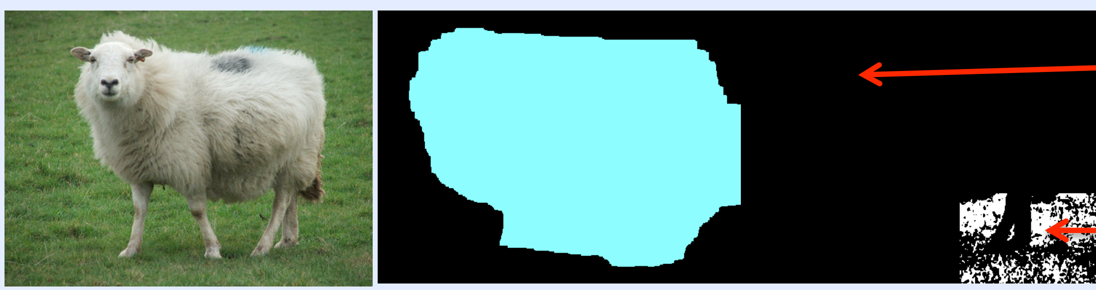
- Not all graph regions are equally difficult
- Repeating computation on easy parts is wasteful

Easy region

Harder region

Dynamic settings (e.g., learning, search)

- Small region of graph changes.
- Computation on unchanged part is wasteful



Unchanged

Changed

Image Previous Optimum Change Mask

References and Related Work

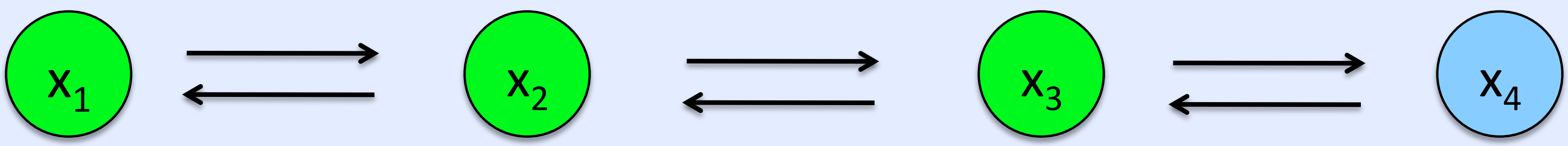
- [Elidan et al., 2006], [Sutton & McCallum, 2007]
  - Residual Belief Propagation. Pass most different messages first.
- [Chandrasekaran et al., 2007]
  - Works only on continuous variables. Very different formulation.
- [Batra et al., 2011]
  - Local Primal Dual Gap for Tightening LP relaxations.
- [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)

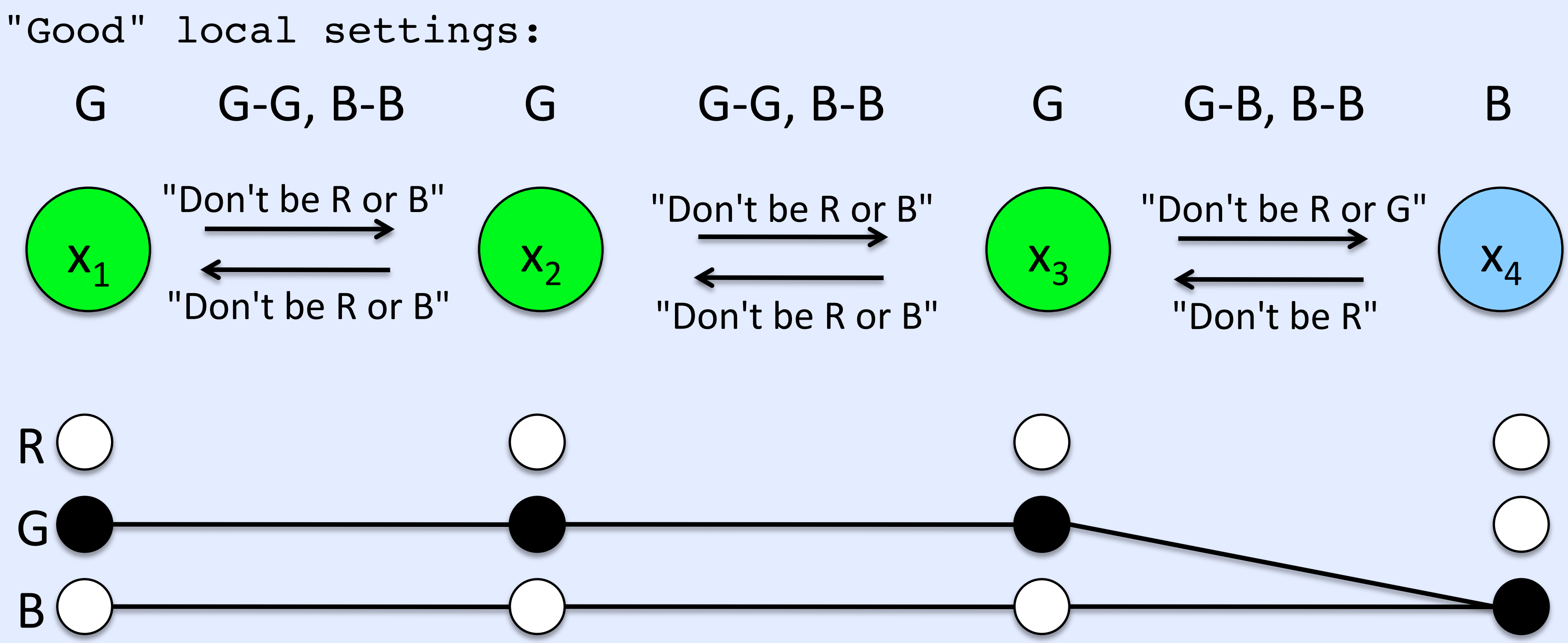
"Good" local settings: (can assume "good" has cost 0, otherwise cost 1)

G G-G, B-B G G-G, B-B G G-B, B-B B



Visualization of reparameterized energy $\tilde{\theta}$

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

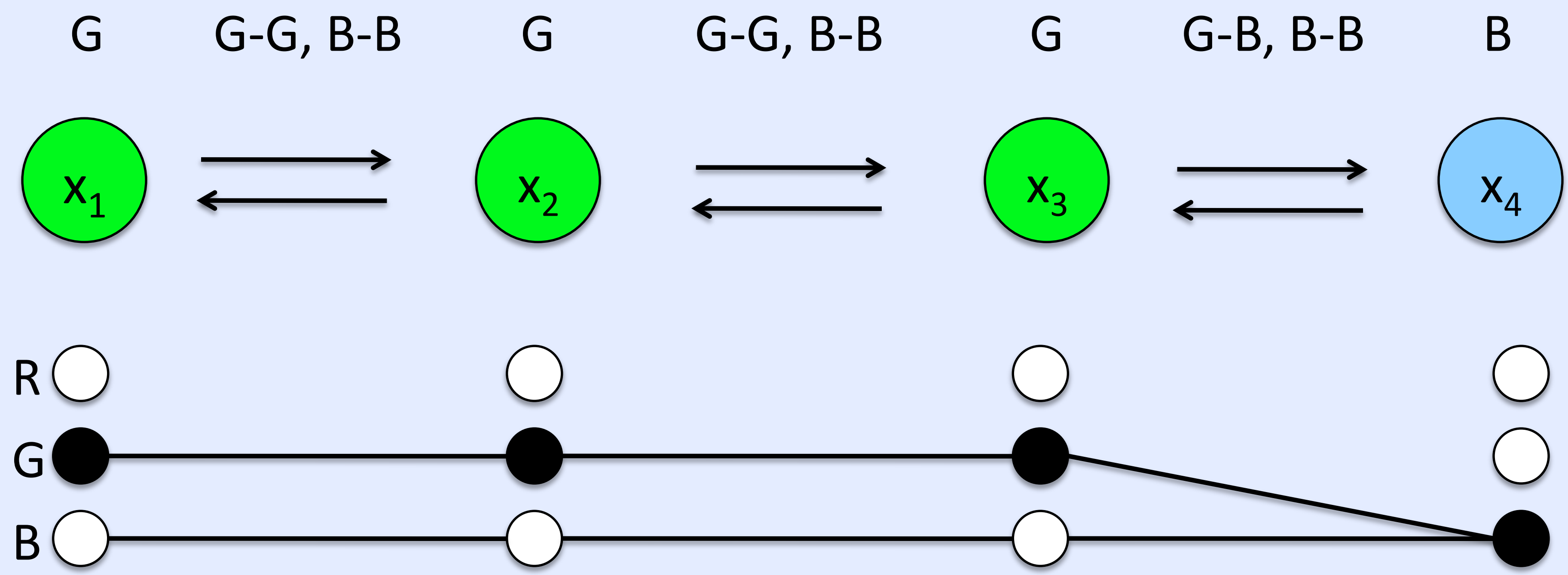


Hypothetical messages that e.g. residual max-product would send.

Visualization of reparameterized energy $\tilde{\theta}$

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

`"Good" local settings:`



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

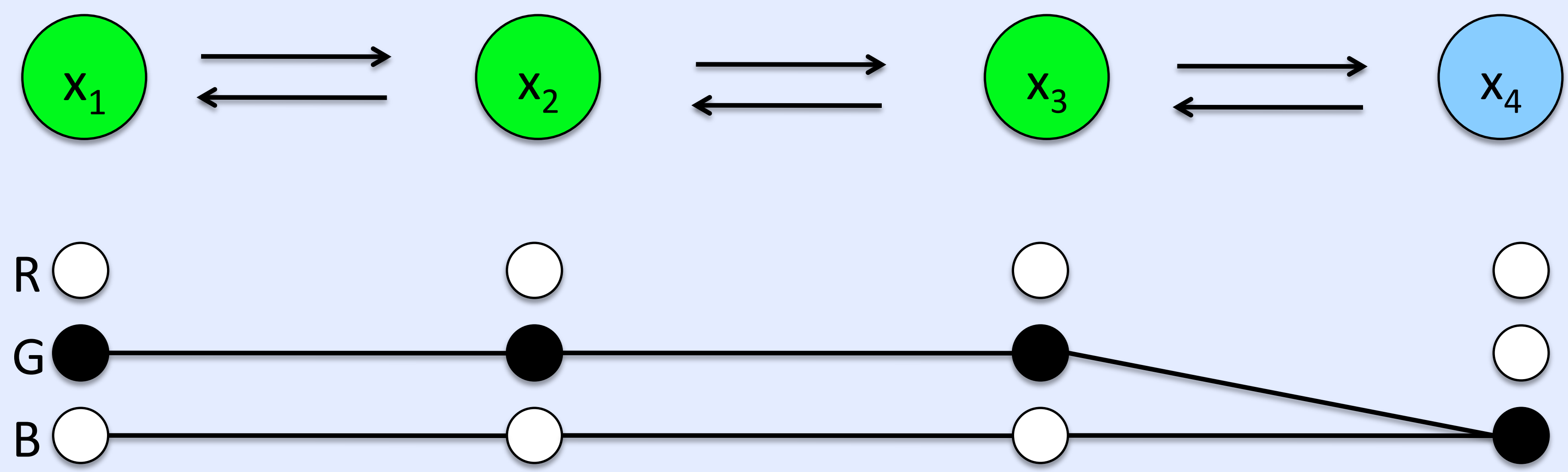
Our scores (see later slides) are 0, so we wouldn't send any messages here.

Visualization of reparameterized energy $\tilde{\theta}$

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

`"Good" local settings:`

B G-G, B-B **B** G-G, B-B **B** G-B, B-B B



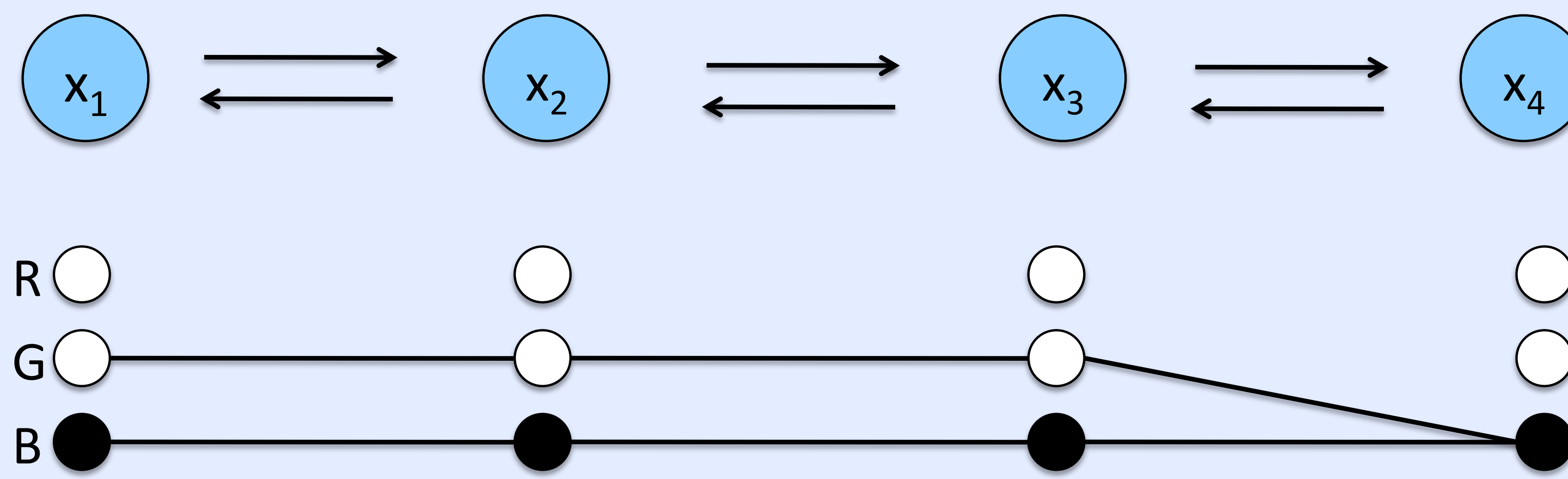
Change unary potentials (e.g., during learning or search)

Visualization of reparameterized energy $\tilde{\theta}$

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

`"Good" local settings:`

B G-G, B-B **B** G-G, B-B **B** G-B, B-B B

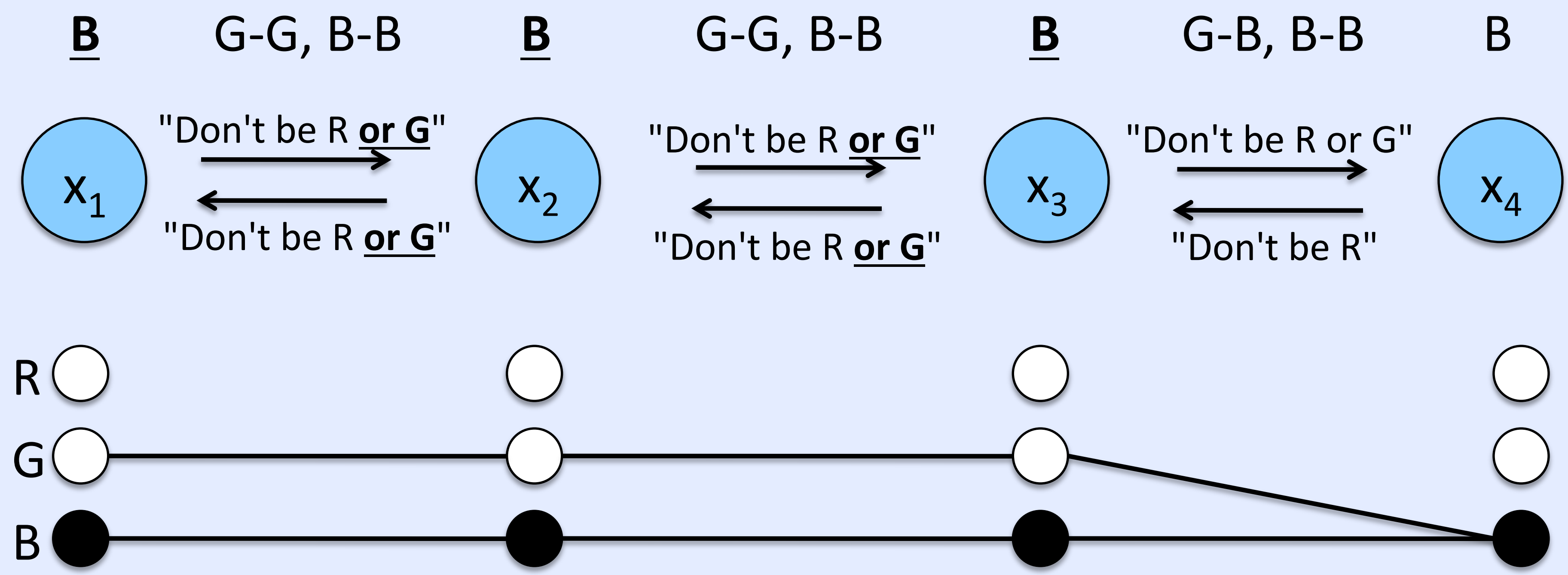


Locally, best assignment for some variables change.

Visualization of reparameterized energy $\tilde{\theta}$

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

```
"Good" local settings:
```

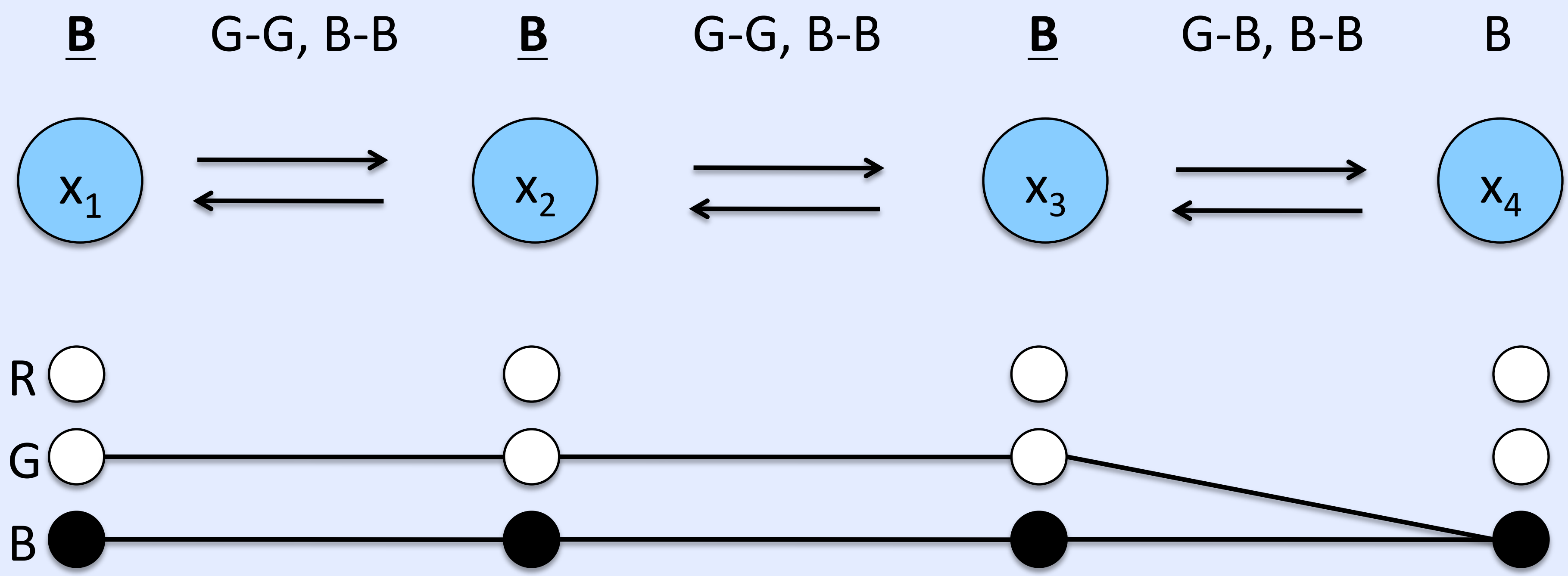


Hypothetical messages that e.g. residual max-product would send.

Visualization of reparameterized energy $\tilde{\theta}$

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

`"Good" local settings:`

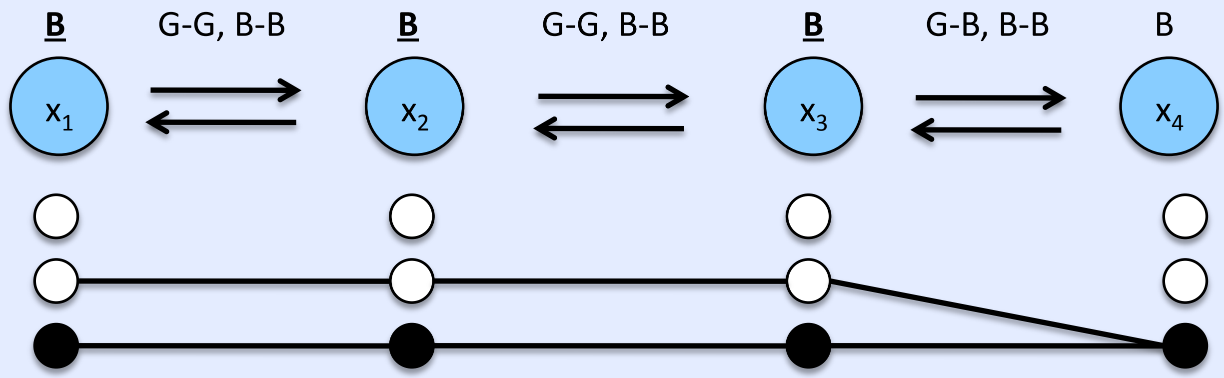


But we don't need to send any messages. We are at the global optimum.

Our scores (see later slides) are 0, so we wouldn't send any messages here.

Visualization of reparameterized energy $\tilde{\theta}$

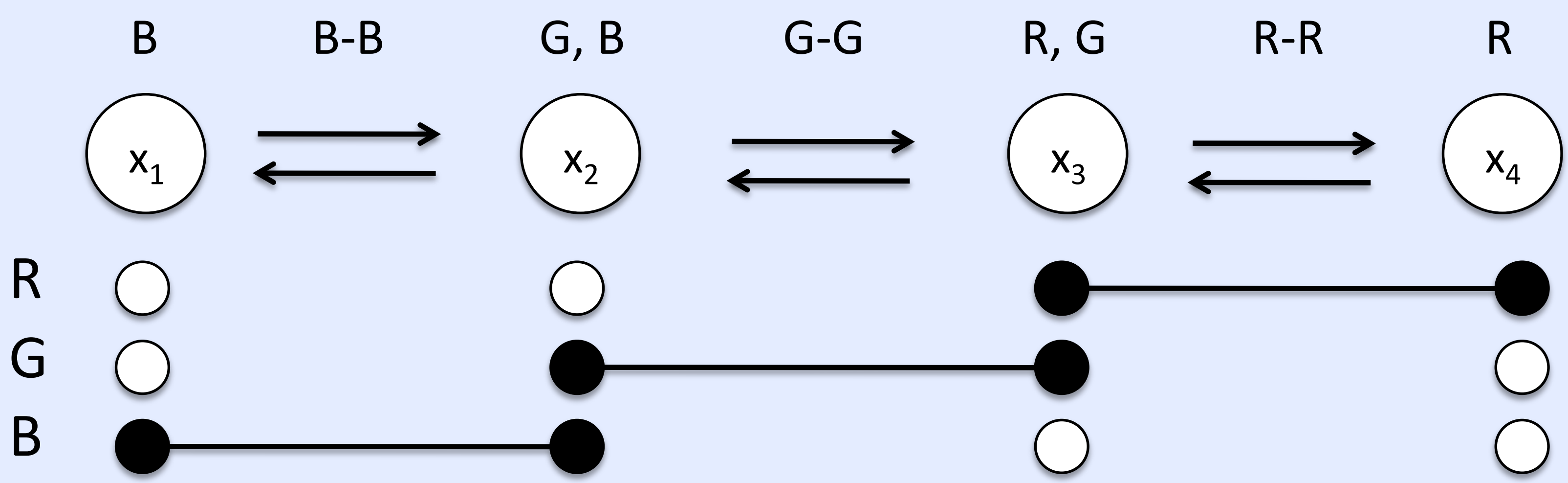
"Good" local settings:



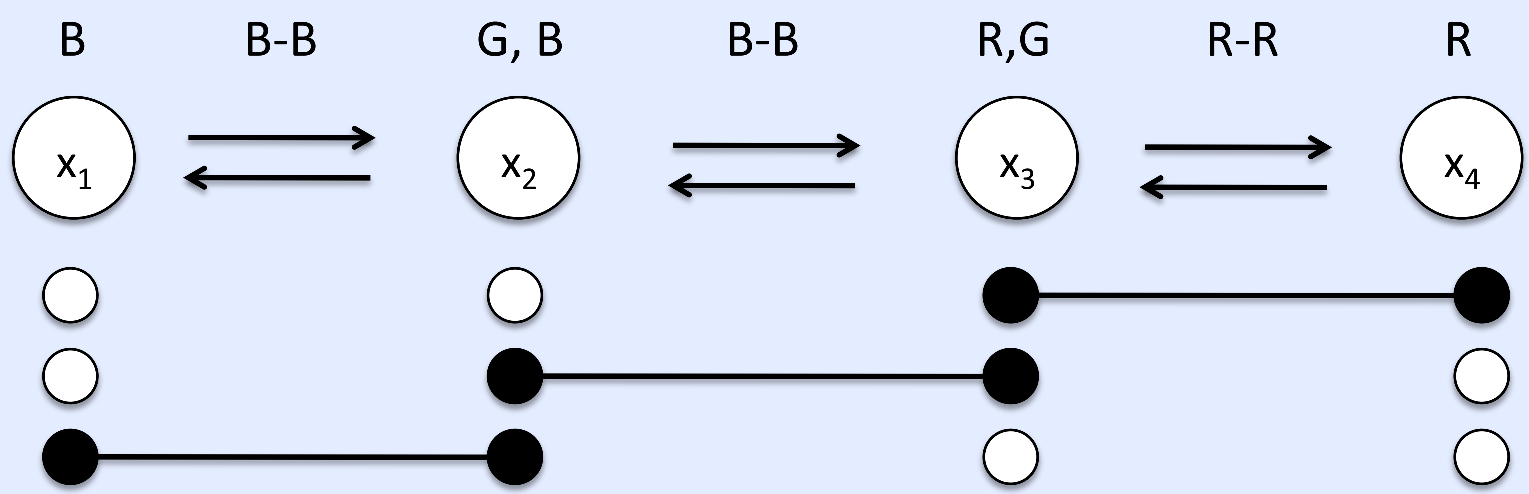
Possible fix: look at how much sending messages on edge would improve dual.

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

"Good" local settings:

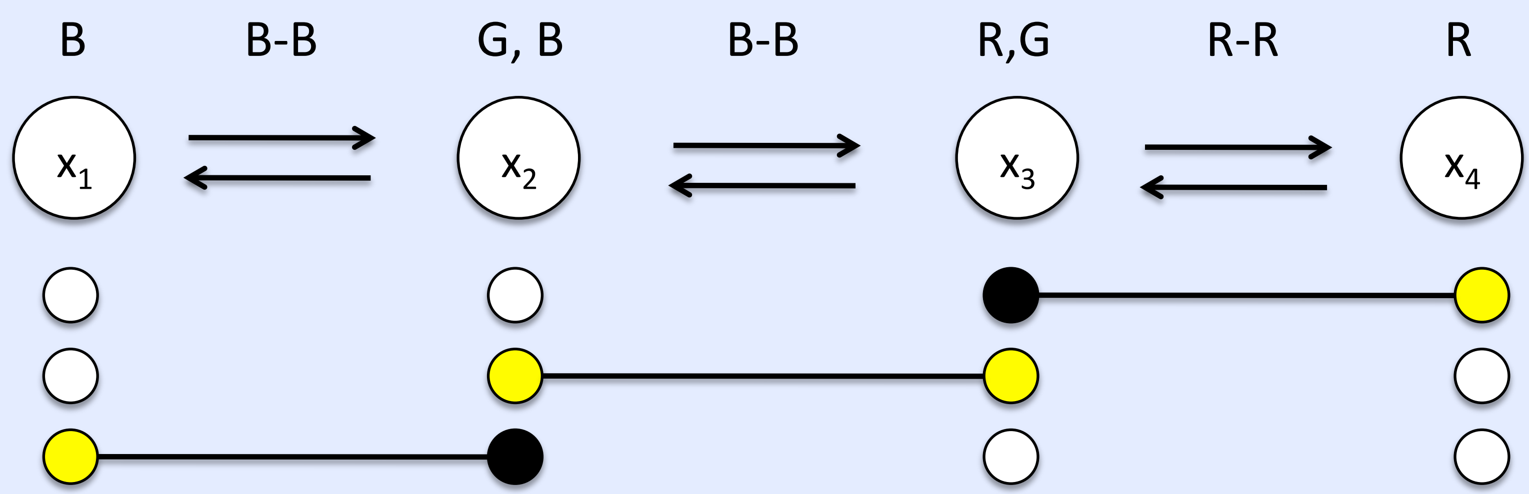


Key Slide



Locally, everything looks optimal

Key Slide

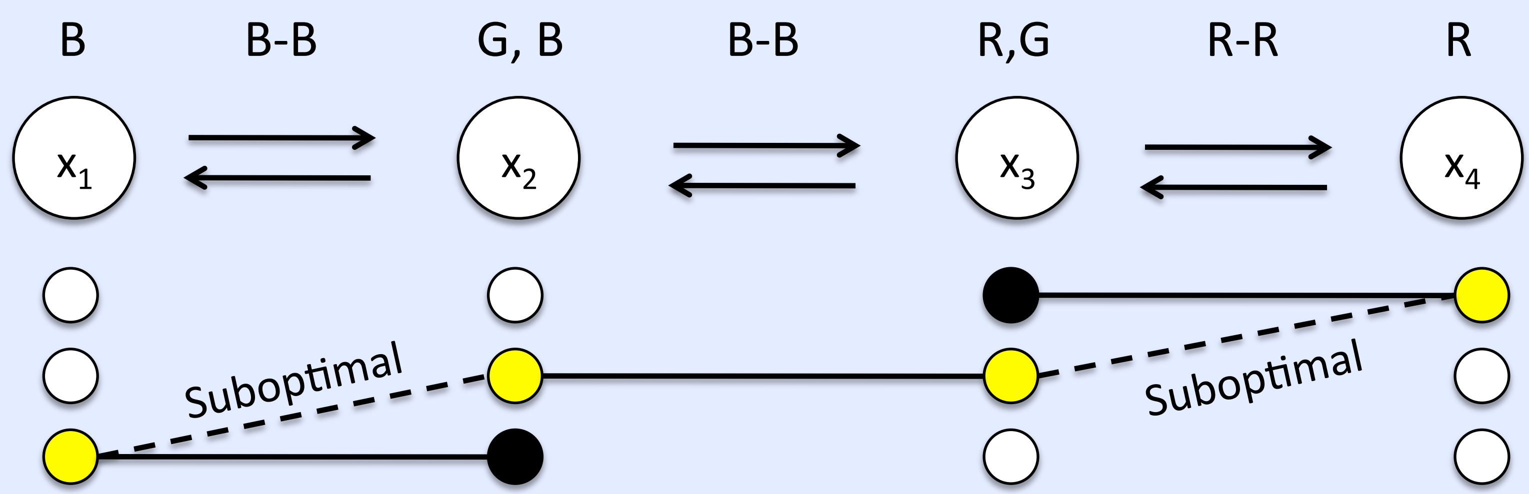


Try assigning a value to each variable

Key Slide

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 $\mathbf{x}^p$ and dual variables (messages) defining $\tilde{\theta}$, primal-dual gap is

Primal-dual gap $$\underbrace{\sum_{A \in \mathcal{V} \cup \mathcal{E}} \theta_A(x_A^p)}_{\text{primal}} - \underbrace{\sum_{A \in \mathcal{V} \cup \mathcal{E}} \min_{x_A} \tilde{\theta}_A(x_A)}_{\text{dual}}$$

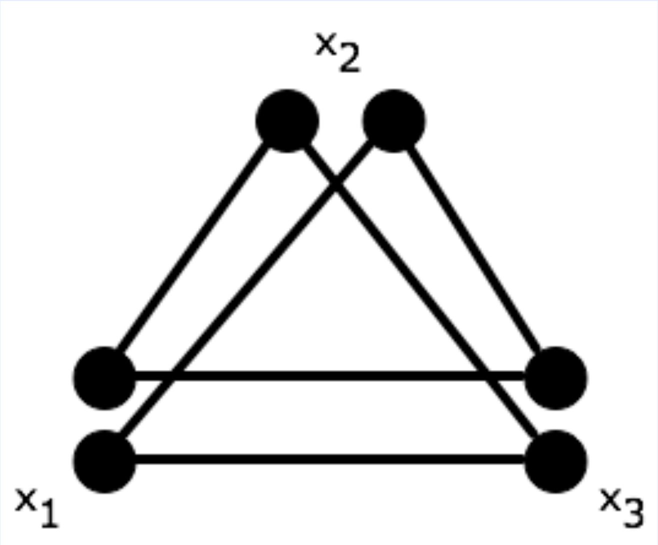
$$= \sum_{A \in \mathcal{V} \cup \mathcal{E}} \left(\tilde{\theta}_A(x_A^p) - \min_{x_A} \tilde{\theta}_A(x_A) \right) = \sum_{A \in \mathcal{V} \cup \mathcal{E}} \mathrm{LPDG}\,(A)$$

Primal cost of node/edge

Dual bound at node/edge

e: "local disagreement" measure: $e_A = \mathrm{LPDG}\,(A)$

Shortcoming of LPDG Score: Loose Relaxations



LPDG > 0,
but dual optimal

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

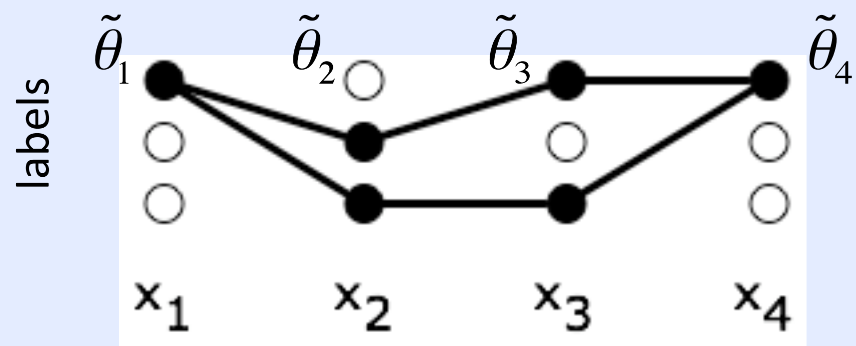
Weak Tree Agreement (WTA) [Kolmogorov 2006]

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

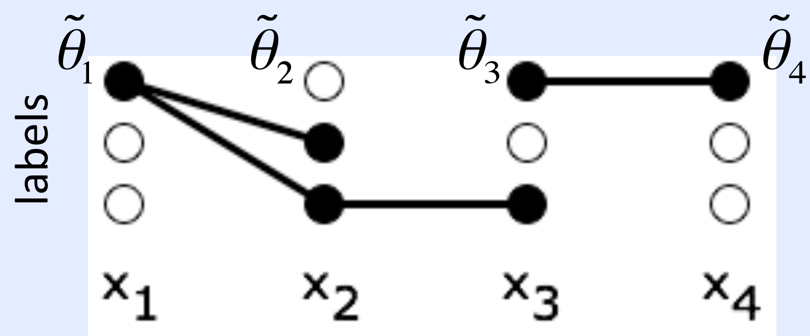
$$\tilde{\theta}_i(x_i) = h_i^* \qquad \forall x_i \in D_i$$

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

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

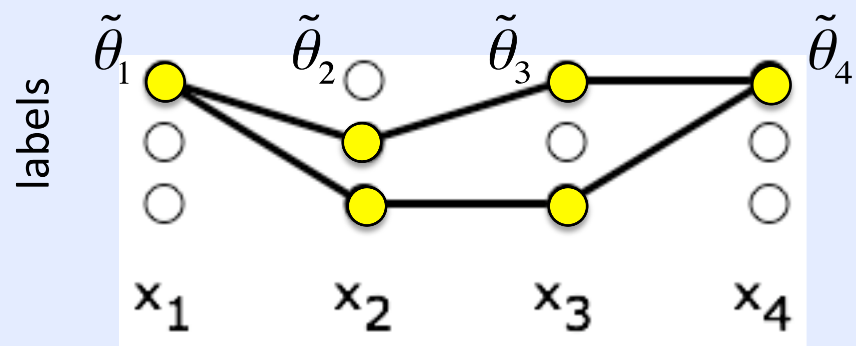
Weak Tree Agreement (WTA) [Kolmogorov 2006]

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

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

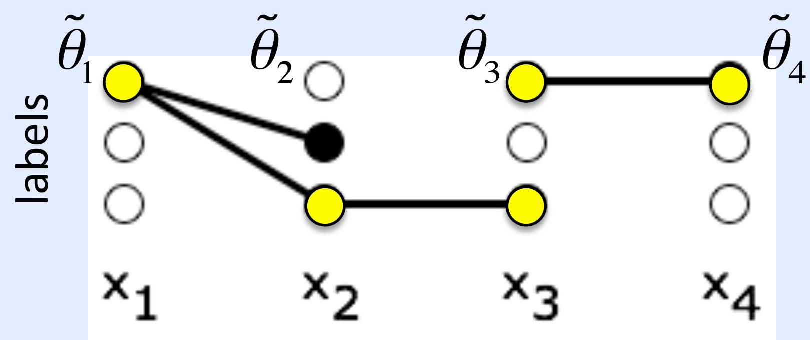
$$\min_{x_j \in D_j} \tilde{\theta}_{ij}(x_i, x_j) = h_{ij}^* \qquad \forall x_i \in D_i, (i,j) \in \mathcal{E}$$

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}

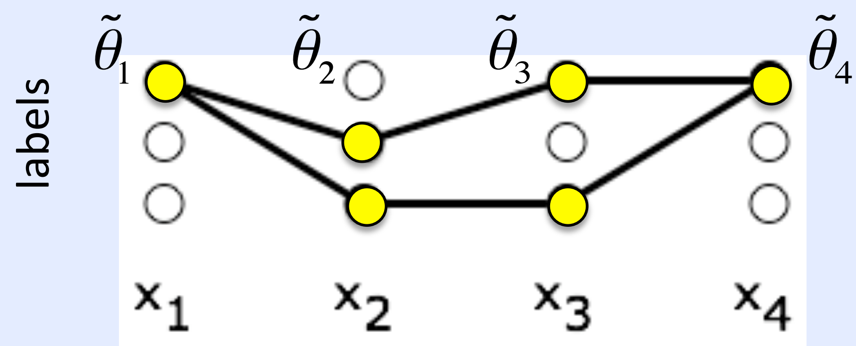
Weak Tree Agreement (WTA) [Kolmogorov 2006]

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

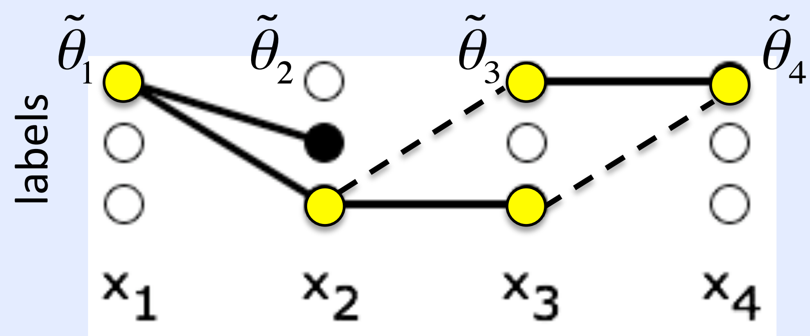
$$\tilde{\theta}_i(x_i) = h_i^* \qquad \forall x_i \in D_i$$

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

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

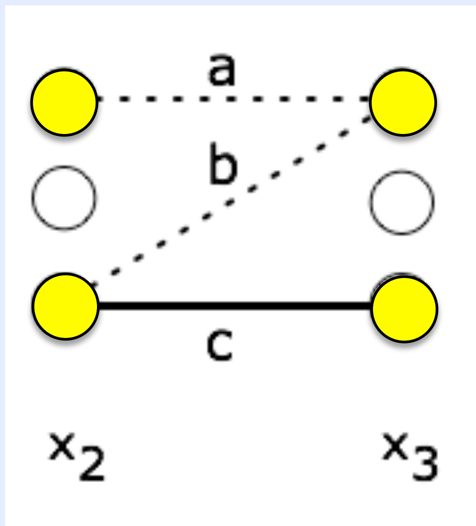
D_1={0} D_2={2} D_2={0,2} D_3={0}

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}

$$e_{23} = \max\Big(\min(a, high),\ \min(\mathrm{b}, c)\Big) - c$$

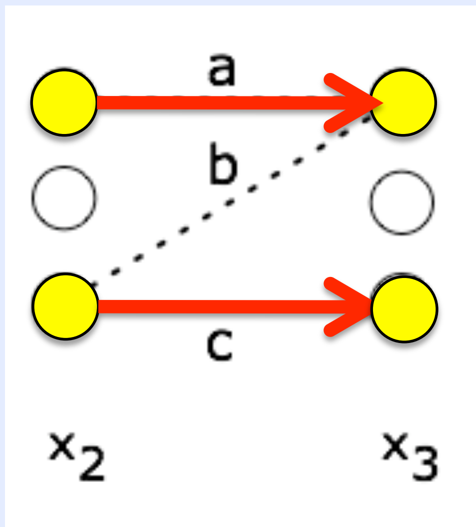
Filled circle means $\tilde{\theta}_i(x_i) = h_i^*$, black edge means $\tilde{\theta}_{ij}(x_i, x_j) = h_{ij}^*$

WTA Score

e: “local disagreement” measure

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

Costs:
solid – low
dotted – medium
else – high



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

$$e_{23} = \max\left(\boxed{\min(a, high)}, \boxed{\min(\mathrm{b}, c)}\right) - c$$

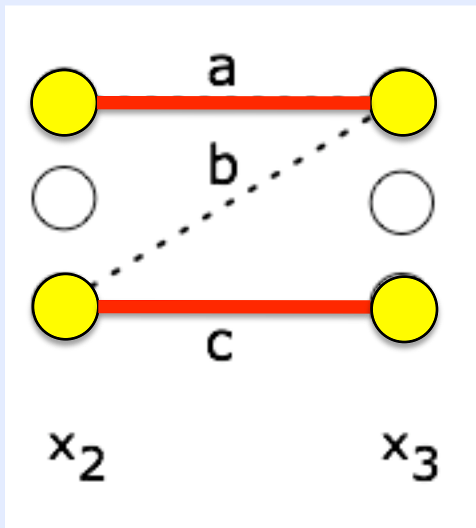
Filled circle means $\tilde{\theta}_i(x_i) = h_i^*$, black edge means $\tilde{\theta}_{ij}(x_i, x_j) = h_{ij}^*$

WTA Score

e: “local disagreement” measure

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

Costs:
solid – low
dotted – medium
else – high



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

$$e_{23} = \boxed{\max(a,\ c)} - c = a - c$$

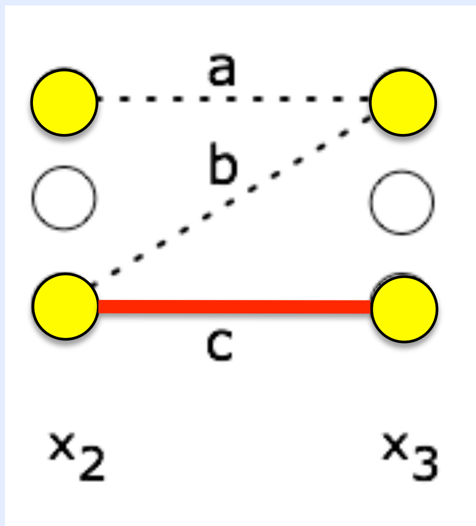
Filled circle means $\tilde{\theta}_i(x_i) = h_i^*$, black edge means $\tilde{\theta}_{ij}(x_i, x_j) = h_{ij}^*$

WTA Score

e: "local disagreement" measure

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

Costs:
solid – low
dotted – medium
else – high



$D_2 = \{0,2\}$

$D_3 = \{0,2\}$

$$e_{23} = \max(a,\ c) - \boxed{c} = a - c$$

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

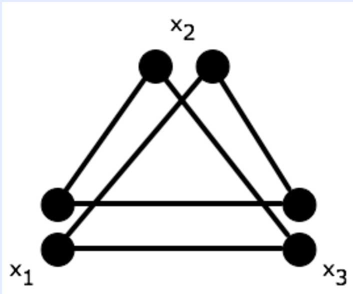
R=1: LPDG score. R>1: WTA score.

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.



LPDG > 0
WTA = 0

- 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

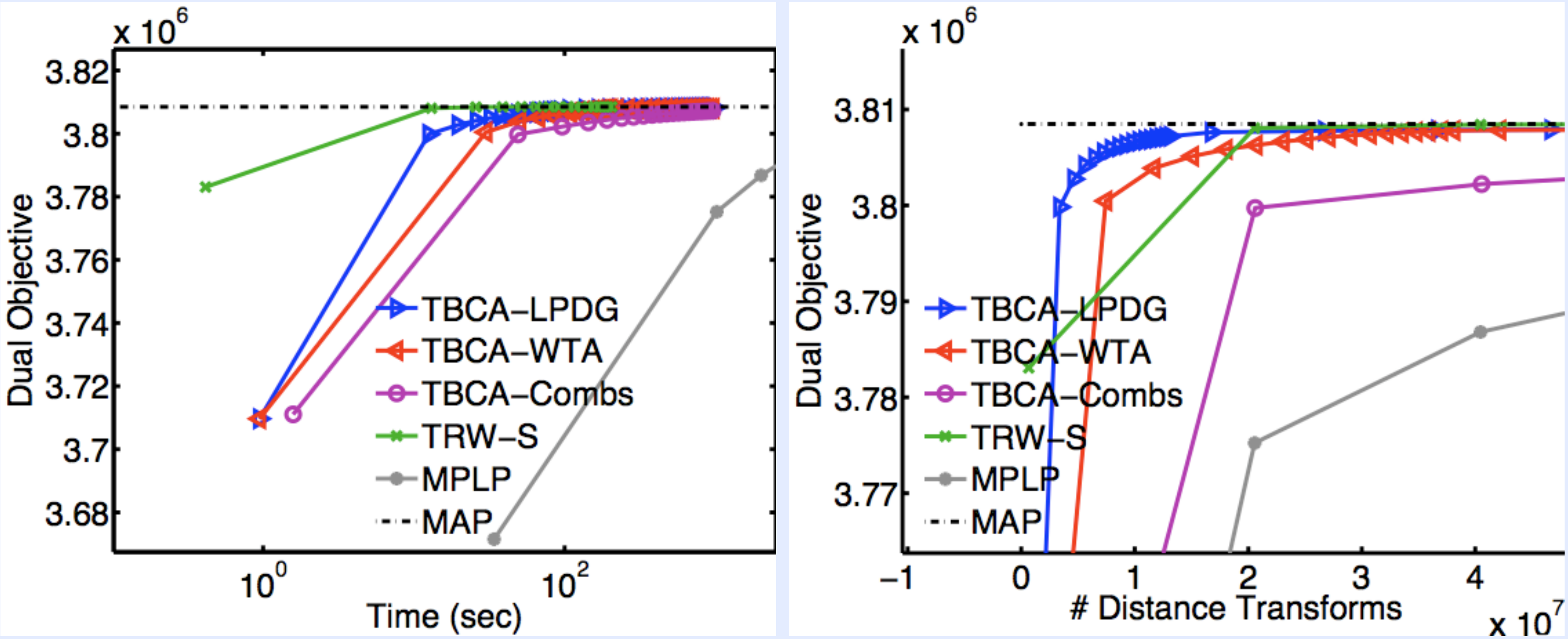
Protein Design:

- Static problem
- Correlation between measure and dual improvement
- Dynamic search application

Algorithms

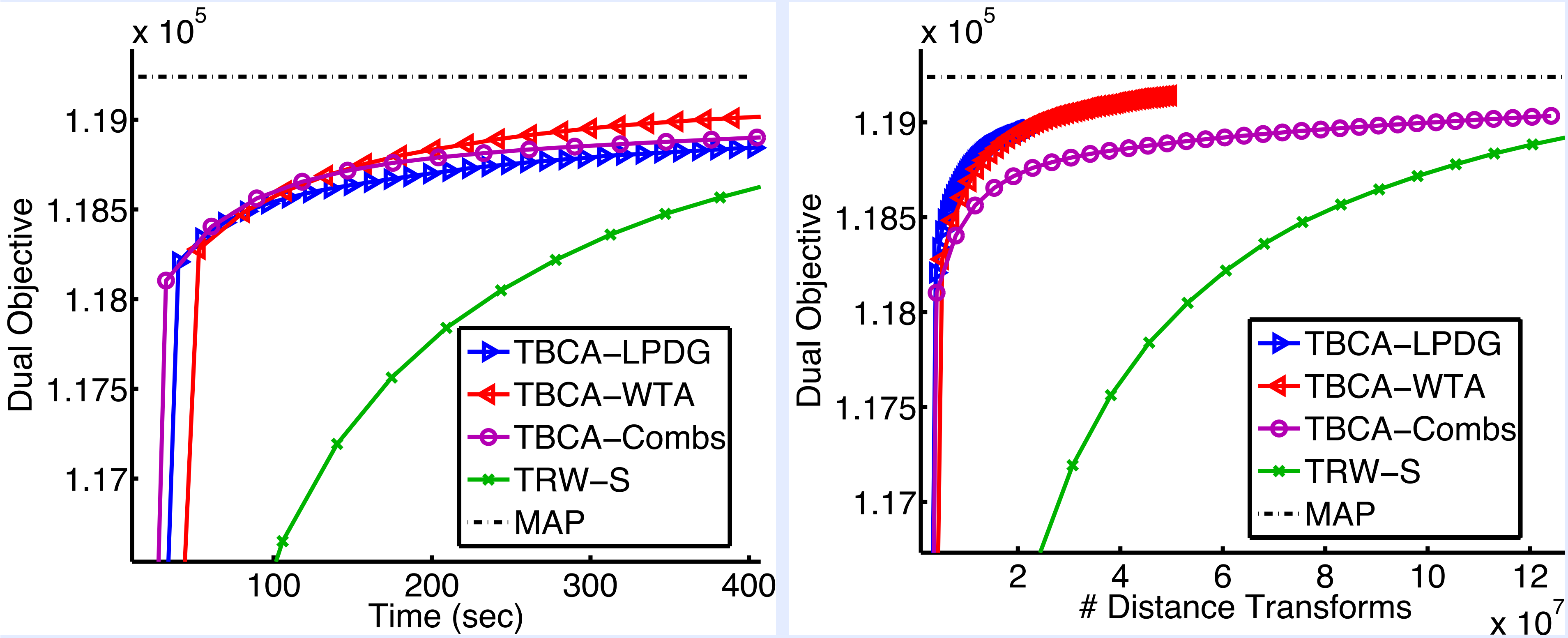
- TBCA: Static Schedule, LPDG Schedule, WTA Schedule
- MPLP [Sontag and Globerson implementation]
- TRW-S [Kolmogorov Implementation]

Experiments: Stereo



383x434 pixels, 16 labels. Potts potentials.

Experiments: Image Segmentation

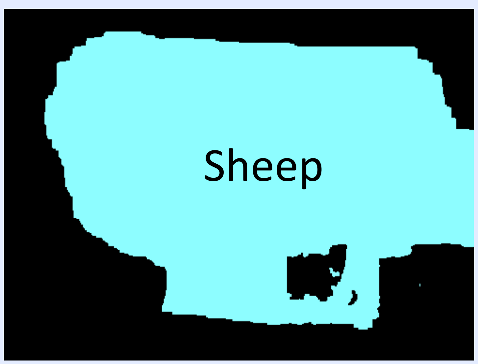


375x500 pixels, 21 labels. General potentials based on label co-occurence.

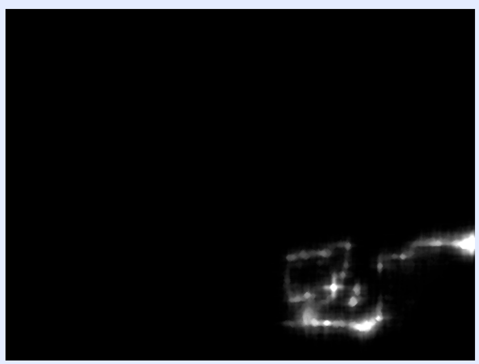
Experiments: Dynamic Image Segmentation

Previous Opt

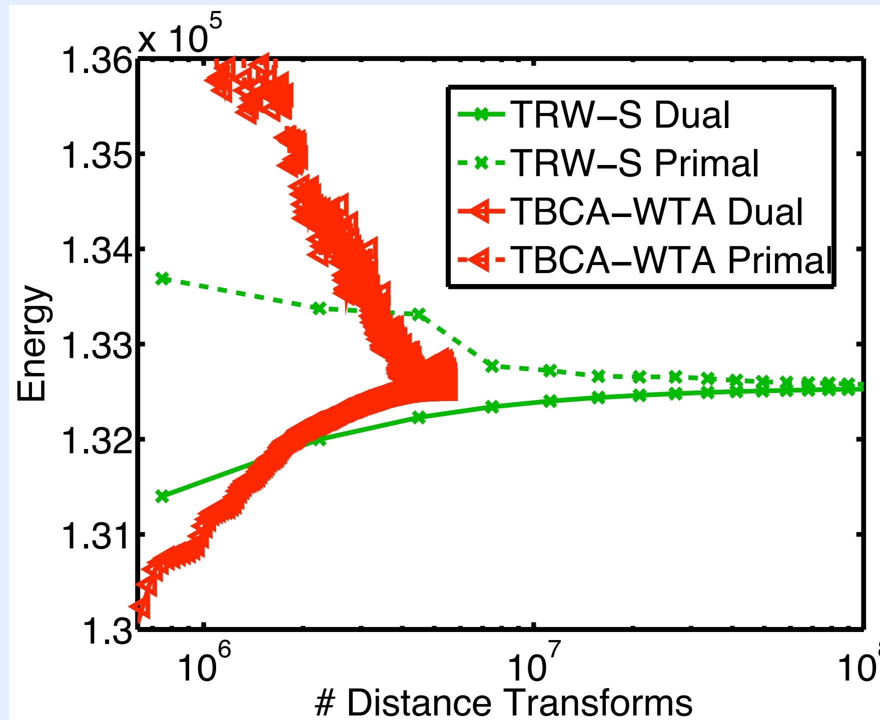
Modify White Unaries



New Opt



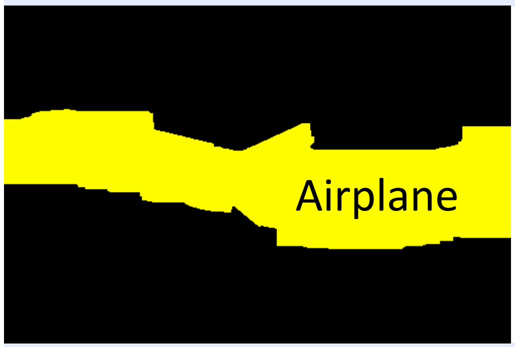
Heatmap of Messages



Warm-started DTBCA vs Warm-started TRW-S

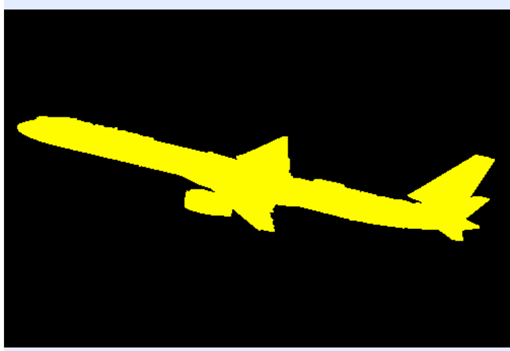
375x500 pixels, 21 labels. Potts potentials.

Experiments: Dynamic Image Segmentation

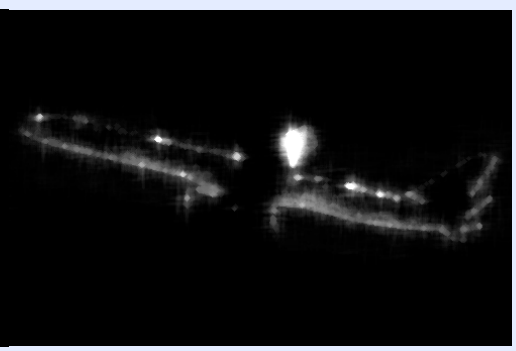


Previous Opt

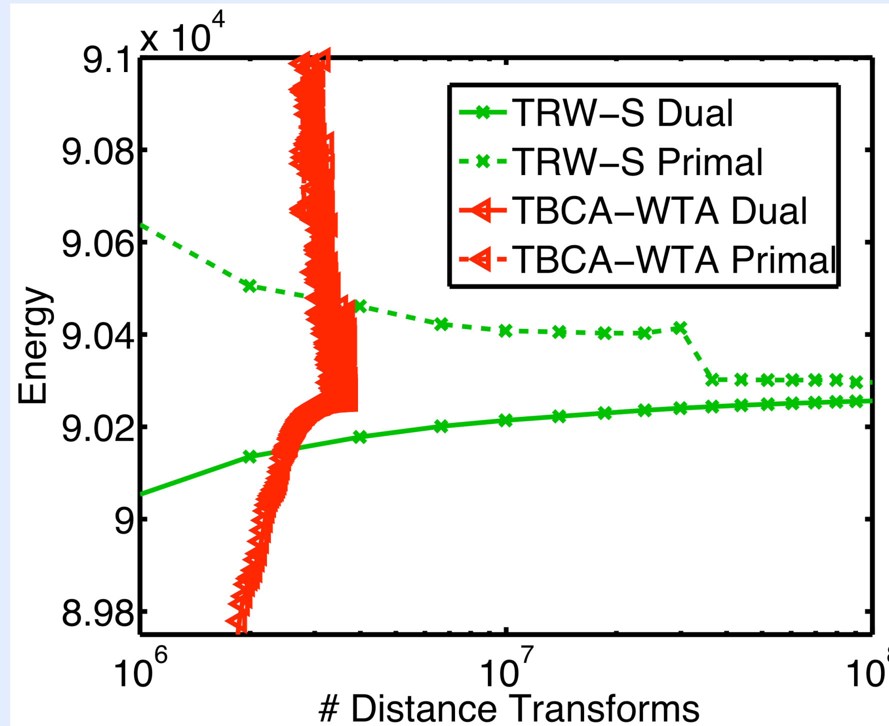
Modify White Unaries



New Opt



Heatmap of Messages

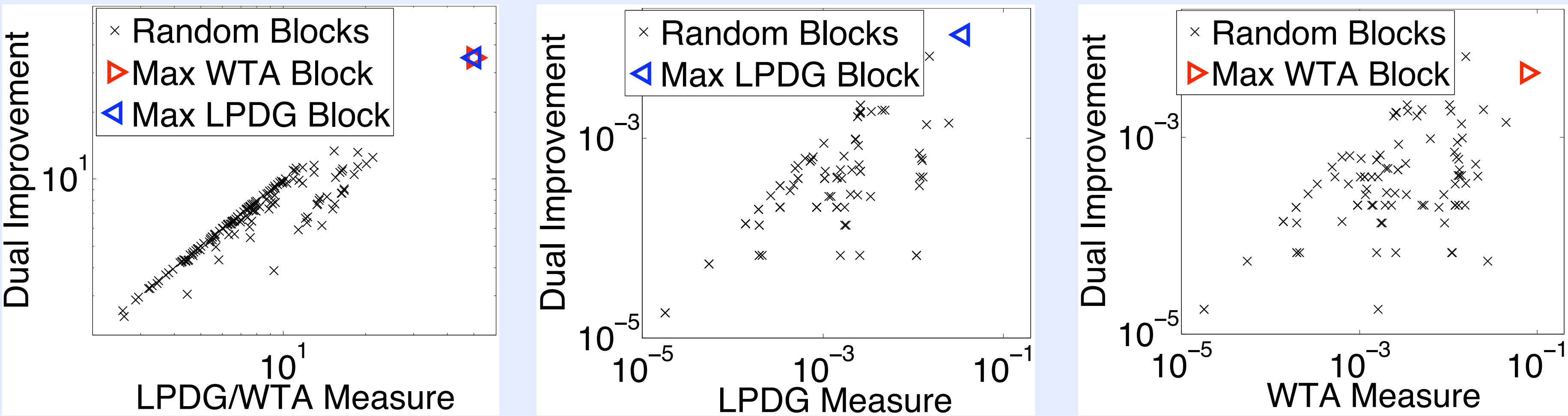


Warm-started DTBCA vs Warm-started TRW-S

375x500 pixels, 21 labels. Potts potentials.

Experiments: Protein Design

Dual Improvement vs. Measure on Forest



<u>Other protein experiments: (see paper)</u>

- 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, ...

Thank You!

Unused slides

Schlesinger's Linear Program (LP)

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

$\Big\downarrow$ exact

Real-valued

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

Marginal polytope

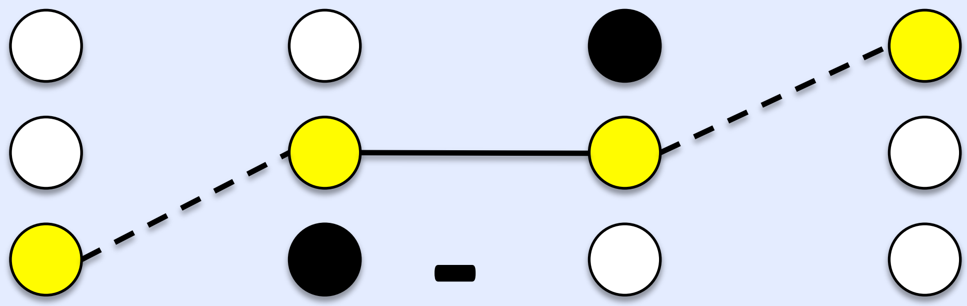
$\Big\downarrow$ approx

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

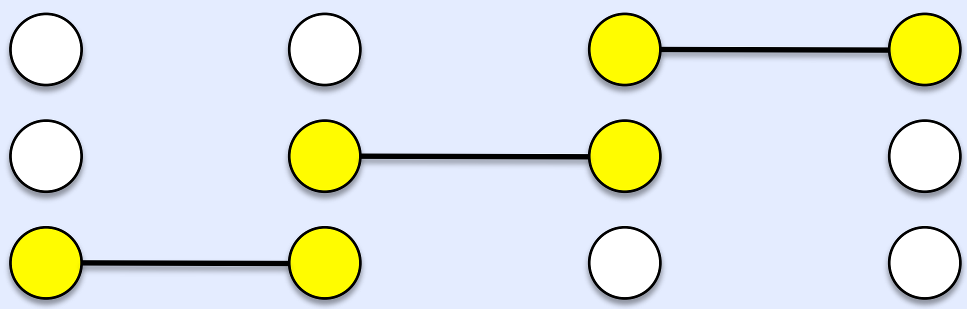
LOCAL polytope

(see next slide)

Primal



Dual



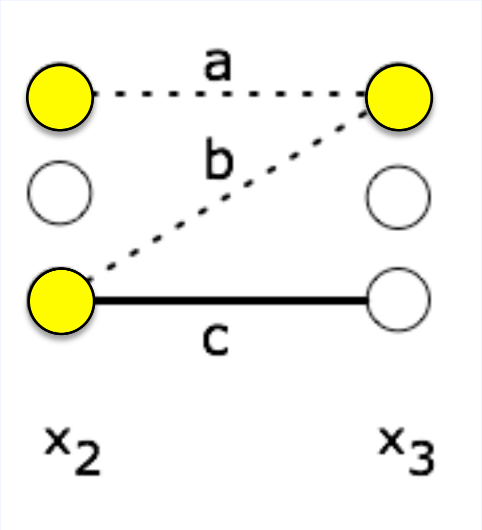
Algorithm 1 Dynamic Tree-Block Coordinate Ascent

$\hat{\mathcal{V}} \leftarrow \mathcal{V}$ {Dirty nodes}
$\hat{\mathcal{E}} \leftarrow \mathcal{E}$ {Dirty edges (see Sec. 5.1 for details)}
$R \leftarrow$ RUN-LPDG ? $1 : R_{WTA}$ {History size}
for $t = 1 : t_{\max}$ **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
  $T \leftarrow$ KRUSKAL-FOREST$(\boldsymbol{w})$
  $\tilde{\theta} \leftarrow$ REPARAMETERIZE-FOREST$(T, \tilde{\theta})$
end for

WTA Score

e: "local disagreement" measure

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



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

$$e_{23} = \max(a, b) - c \qquad e_{32} = \min(a, b) - c$$

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