Tightening LP Relaxations for MAP using Message-Passing

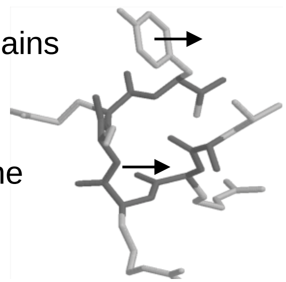
David Sontag

Joint work with Talya Meltzer, Amir Globerson, Tommi Jaakkola, and Yair Weiss

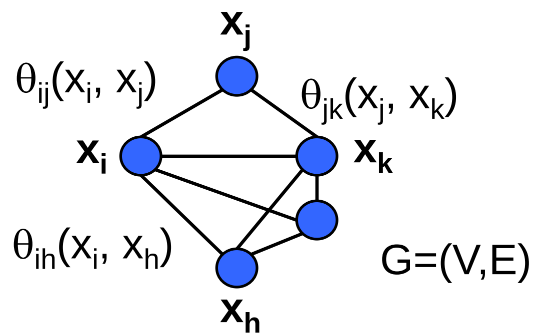
Massachusetts Institute of Technology

MAP in Undirected Graphical Models

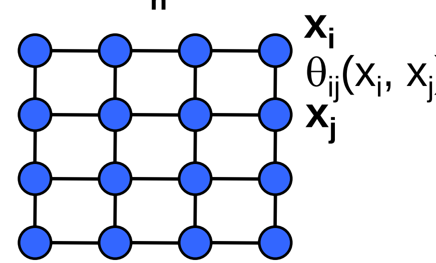
Real-world problems:



Protein design



Stereo vision



$$\Pr(x;\theta) \propto \exp\left(\sum_{(i,j)\in E} \theta_{ij}(x_i, x_j)\right)$$

Find most likely assignment:

$$x_{\text{map}} = \arg\max_x \sum_{(i,j)\in E} \theta_{ij}(x_i, x_j)$$

How to solve MAP?

- MAP is NP-hard
- Are real-world MAP problems really so hard?
- We give an algorithm which:
  - Improves approximation, using more computation
  - Problem-specific
  - If we *do* find best assignment, we know it
- Solves real-world problems, *exactly*

MAP as a linear program

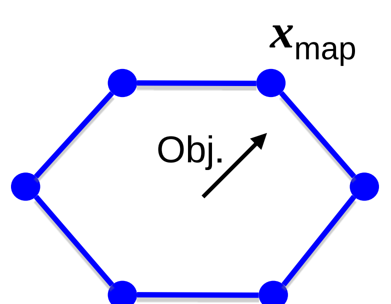
We can formulate the MAP problem as a linear program

$$\max_{\boldsymbol{x}} \sum_{(i,j)\in E} \theta_{ij}(x_i,x_j) = \max_{\boldsymbol{\mu}\in\mathcal{M}(G)} \sum_{(i,j)\in E} \sum_{x_i,x_j} \mu_{ij}(x_i,x_j)\theta_{ij}(x_i,x_j)$$

where the variables μ_{ij} are defined over edges.

The *marginal polytope* constrains the μ_{ij} to be marginals of *some* distribution:

$$\mathcal{M}(G) = \{\boldsymbol{\mu} \mid \exists \Pr(\boldsymbol{x};\boldsymbol{\theta}) \text{ s.t. } \mu_{ij}(x_i,x_j) = \Pr(x_i,x_j;\boldsymbol{\theta})\}$$

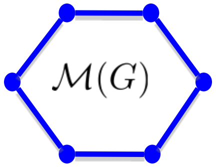


Very many constraints!

Vertices correspond to assignments

Relaxing the MAP LP

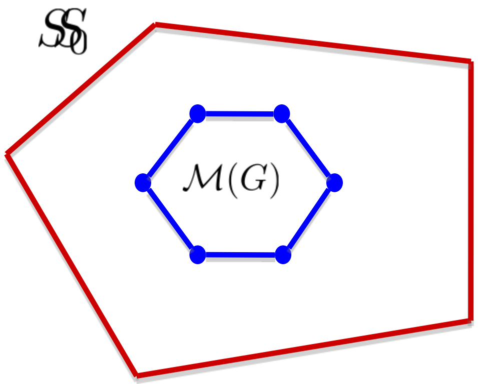
$$\max_{\boldsymbol{x}} \sum_{(i,j)\in E} \theta_{ij}(x_i, x_j) \quad = \quad \max_{\boldsymbol{\mu}\in\mathcal{M}(G)} \sum_{(i,j)\in E} \sum_{x_i, x_j} \mu_{ij}(x_i, x_j)\theta_{ij}(x_i, x_j)$$

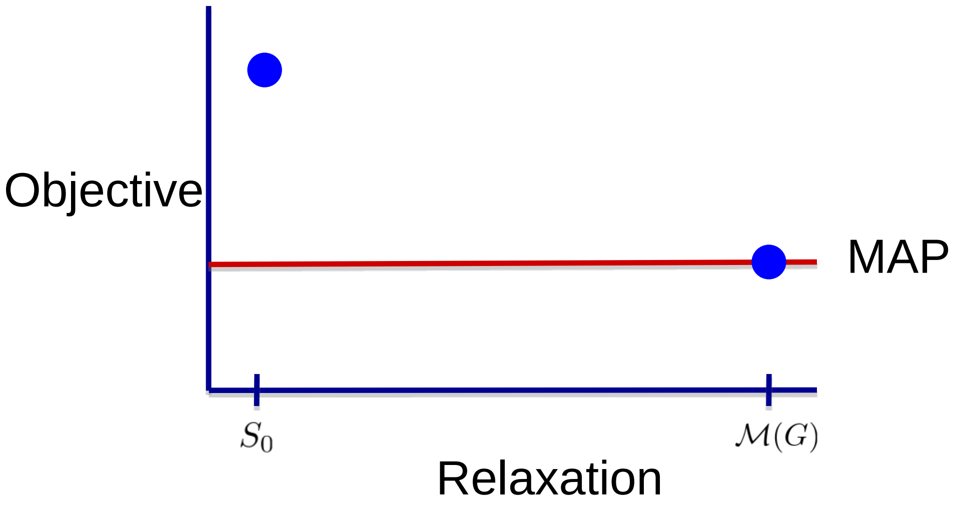


Relaxing the MAP LP

$$\max_{\boldsymbol{x}} \sum_{(i,j)\in E} \theta_{ij}(x_i, x_j) \quad \leq \quad \max_{\boldsymbol{\mu}\in S} \sum_{(i,j)\in E} \sum_{x_i, x_j} \mu_{ij}(x_i, x_j)\theta_{ij}(x_i, x_j)$$

Such that $\mathcal{M}(G) \subseteq S$



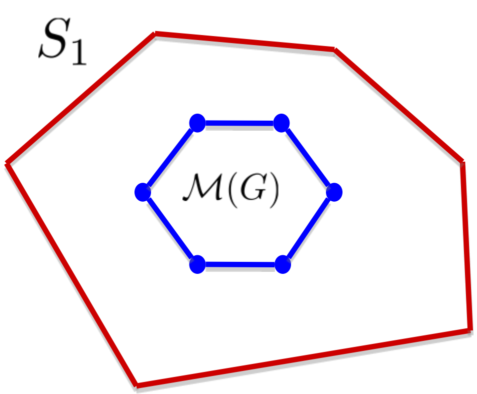


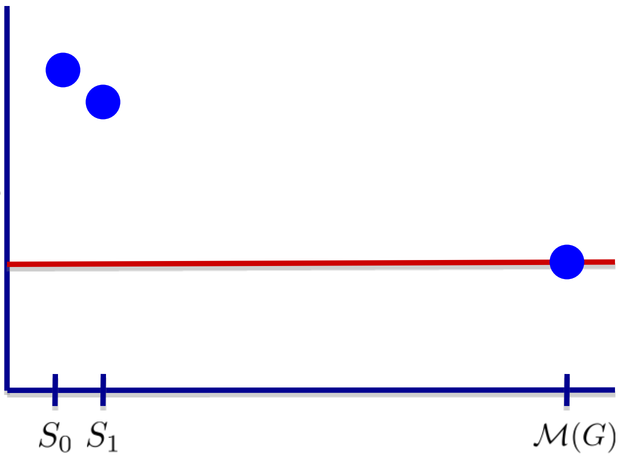
Simplest outer bound: $\sum_{x_i, x_j} \mu_{ij}(x_i, x_j) = 1$

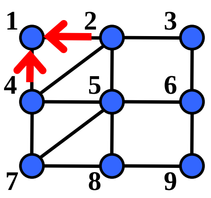
Tightening the LP

$$\max_{\boldsymbol{x}} \sum_{(i,j)\in E} \theta_{ij}(x_i, x_j) \quad \leq \quad \max_{\boldsymbol{\mu}\in S} \sum_{(i,j)\in E} \sum_{x_i, x_j} \mu_{ij}(x_i, x_j)\theta_{ij}(x_i, x_j)$$

Such that $\mathcal{M}(G) \subseteq S$







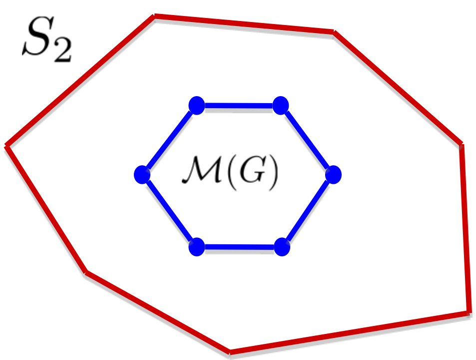
$$\sum_{x_2} \mu_{12}(x_1, x_2) = \sum_{x_4} \mu_{14}(x_1, x_4)$$

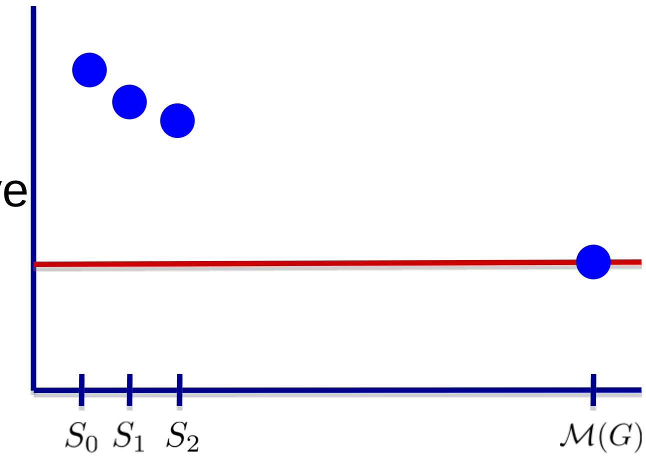
Partial pairwise consistency

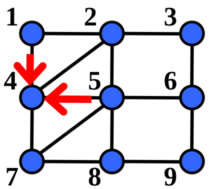
Tightening the LP

$$\max_{\boldsymbol{x}} \sum_{(i,j)\in E} \theta_{ij}(x_i, x_j) \quad \leq \quad \max_{\boldsymbol{\mu}\in S} \sum_{(i,j)\in E} \sum_{x_i, x_j} \mu_{ij}(x_i, x_j)\theta_{ij}(x_i, x_j)$$

Such that $\mathcal{M}(G) \subseteq S$







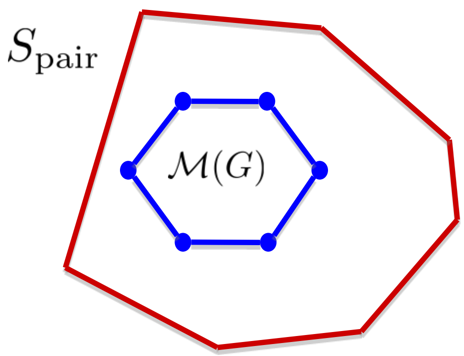
$$\sum_{x_1} \mu_{14}(x_1, x_4) = \sum_{x_5} \mu_{45}(x_4, x_5)$$

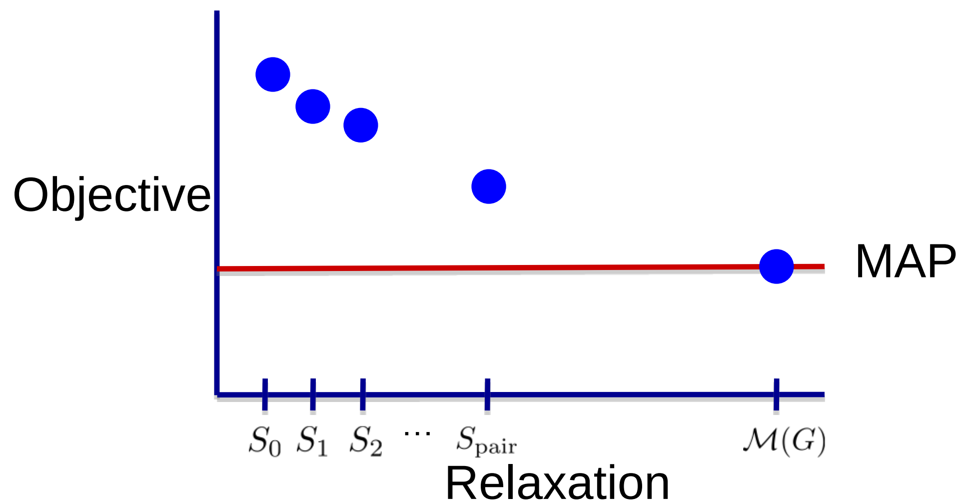
"*Partial pairwise* consistency

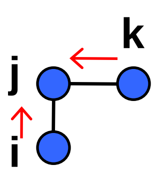
Tightening the LP

$$\max_{\boldsymbol{x}} \sum_{(i,j)\in E} \theta_{ij}(x_i, x_j) \quad \leq \quad \max_{\boldsymbol{\mu}\in S} \sum_{(i,j)\in E} \sum_{x_i, x_j} \mu_{ij}(x_i, x_j)\theta_{ij}(x_i, x_j)$$

Such that $\mathcal{M}(G) \subseteq S$







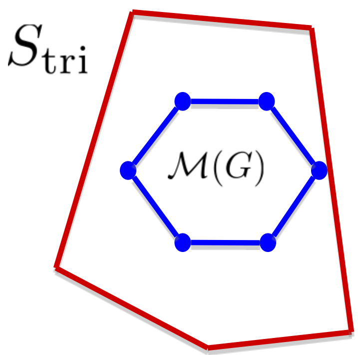
$$\sum_{x_i} \mu_{ij}(x_i, x_j) = \sum_{x_k} \mu_{ij}(x_j, x_k)$$

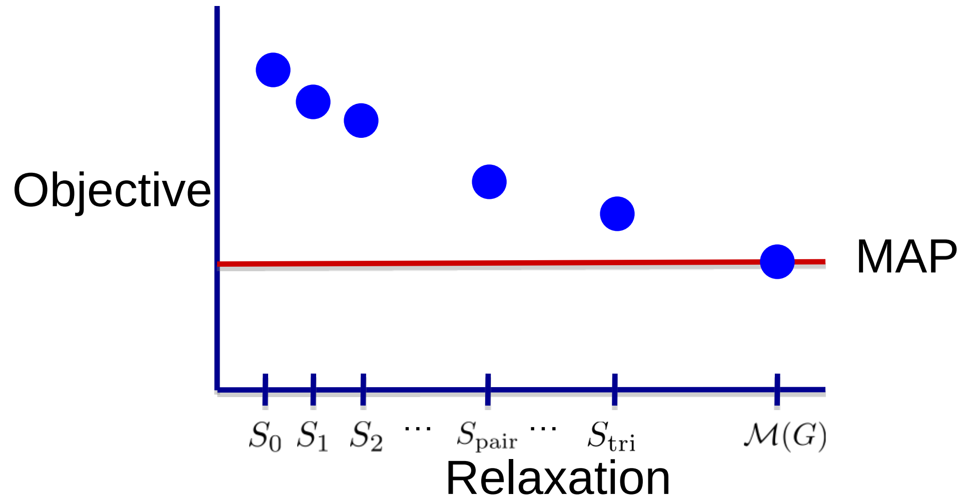
“ *Pairwise* consistency

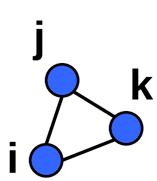
Tightening the LP

$$\max_{\boldsymbol{x}} \sum_{(i,j)\in E} \theta_{ij}(x_i, x_j) \quad \leq \quad \max_{\boldsymbol{\mu}\in S} \sum_{(i,j)\in E} \sum_{x_i, x_j} \mu_{ij}(x_i, x_j)\theta_{ij}(x_i, x_j)$$

Such that $\mathcal{M}(G) \subseteq S$







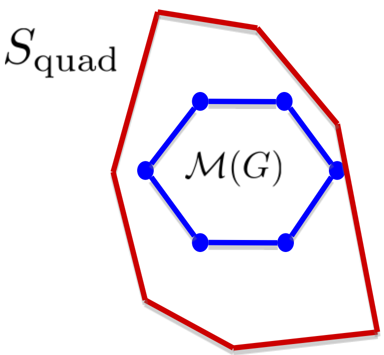
$$\sum_{x_k} \mu_{ijk}(x_i, x_j, x_k) = \mu_{ij}(x_i, x_j)$$

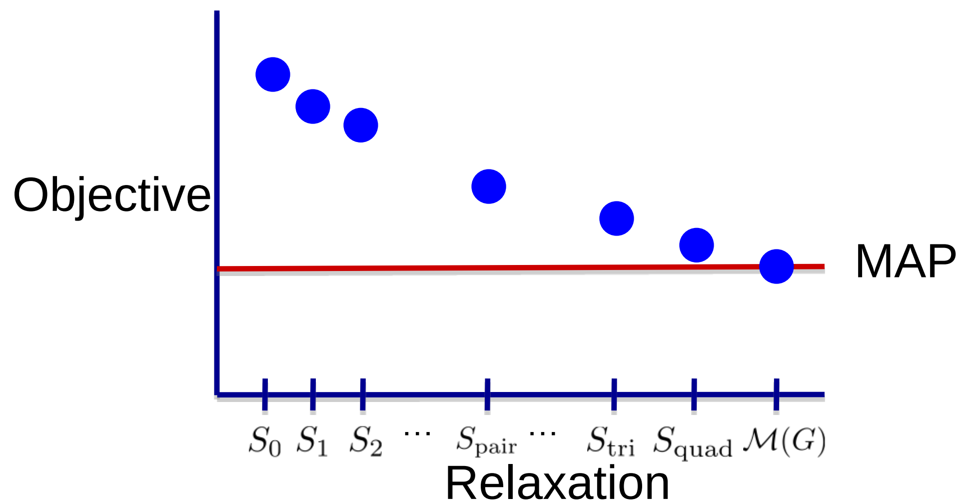
Triplet consistency

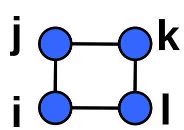
Tightening the LP

$$\max_{\boldsymbol{x}} \sum_{(i,j)\in E} \theta_{ij}(x_i, x_j) \quad \leq \quad \max_{\boldsymbol{\mu}\in S} \sum_{(i,j)\in E} \sum_{x_i,x_j} \mu_{ij}(x_i, x_j)\theta_{ij}(x_i, x_j)$$

Such that $\mathcal{M}(G) \subseteq S$







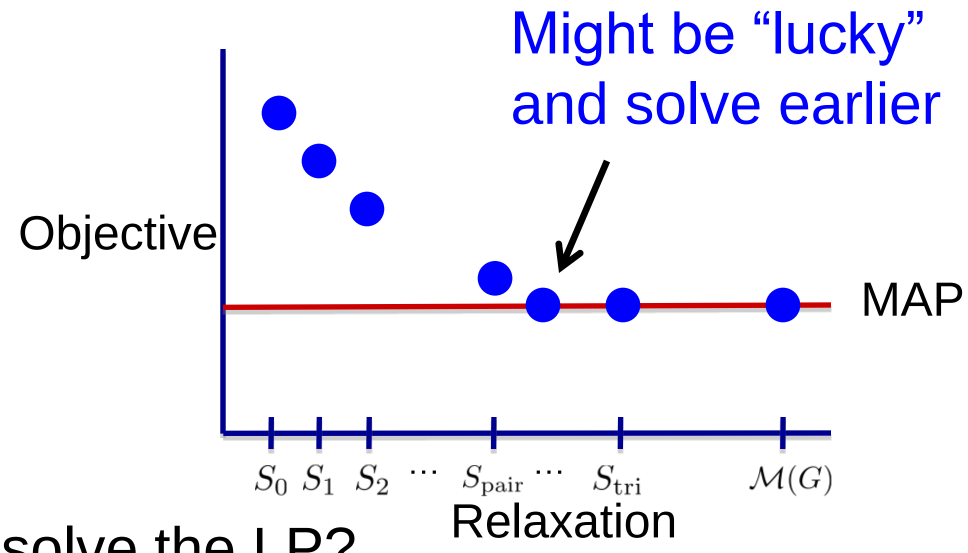
$$\sum_{x_k,x_l} \mu_{ijkl}(x_i, x_j, x_k, x_l) = \mu_{ij}(x_i, x_j)$$

“*Quadruplet* consistency

Tightening the LP

$$\max_{\boldsymbol{x}} \sum_{(i,j)\in E} \theta_{ij}(x_i, x_j) \quad \leq \quad \max_{\boldsymbol{\mu}\in S} \sum_{(i,j)\in E} \sum_{x_i,x_j} \mu_{ij}(x_i, x_j)\theta_{ij}(x_i, x_j)$$

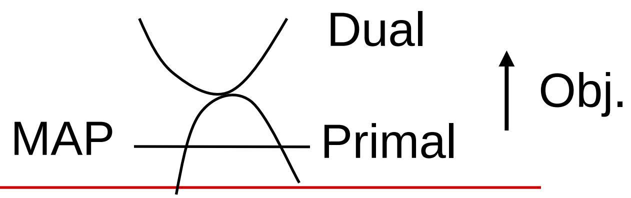
Such that $\mathcal{M}(G) \subseteq S$



Great! But…

- Can we efficiently solve the LP?
- What clusters to add?
- How do we avoid re-solving?

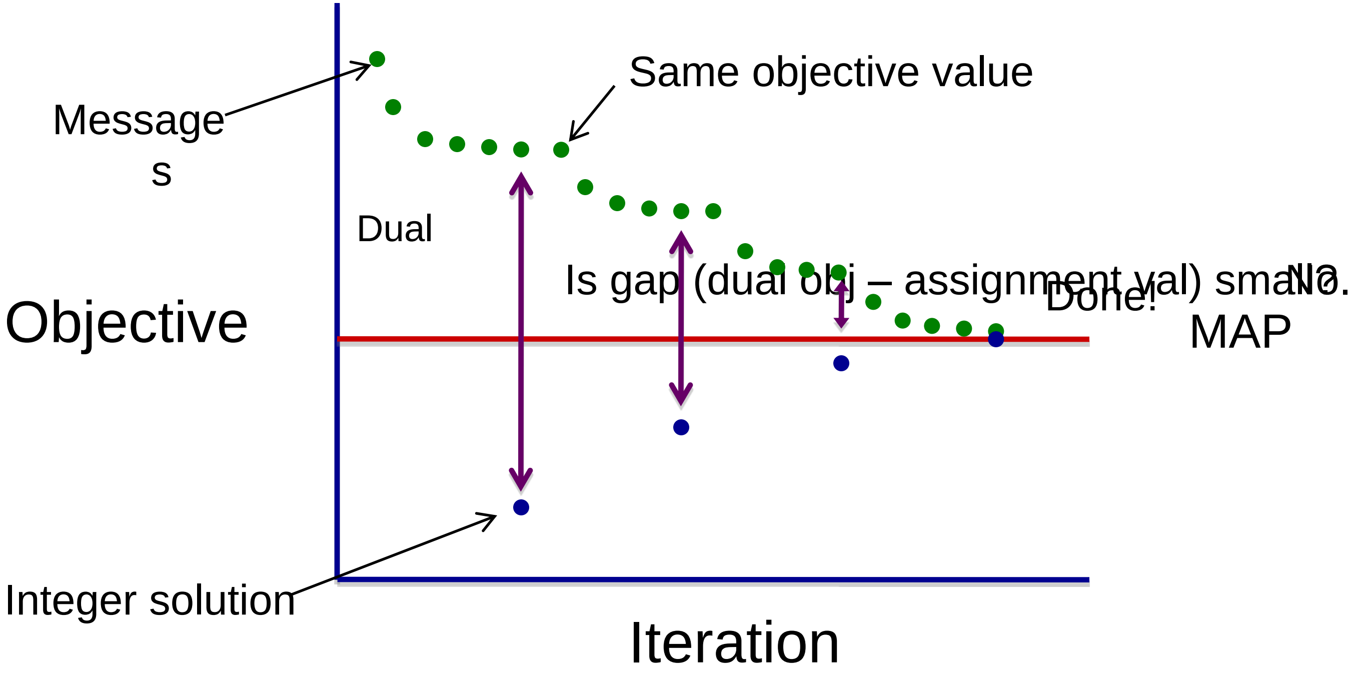
Our solution



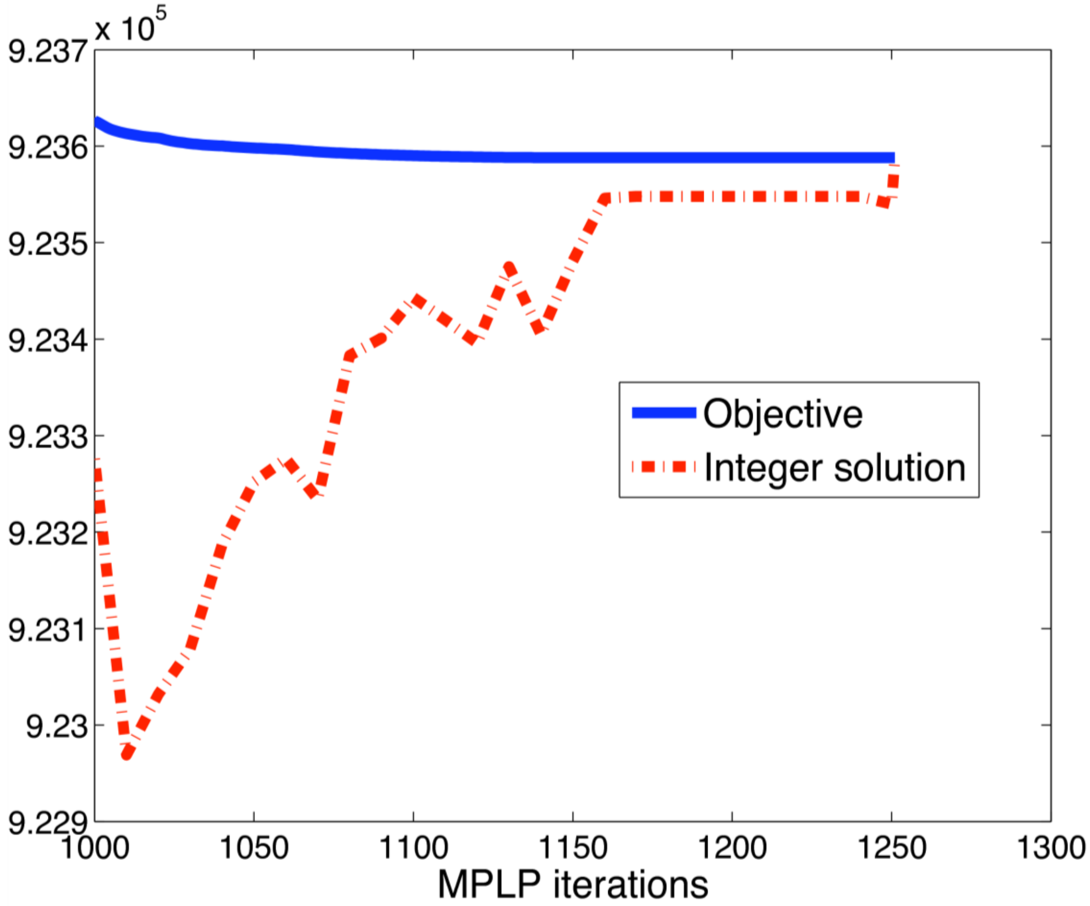
- Can we efficiently solve the LP?
  - We work in the *dual* LP (Globerson & Jaakkola ‘07)
  - Dual can be solved by efficient message-passing algorithm
  - Corresponds to coordinate-descent algorithm
- What cluster to add next?
  - We propose a greedy *bound minimization* algorithm
  - Add clusters with guaranteed improvement – upper bound gets tighter
- How do we avoid re-solving?
  - “Warm start” of new messages using the old messages

Dual algorithm

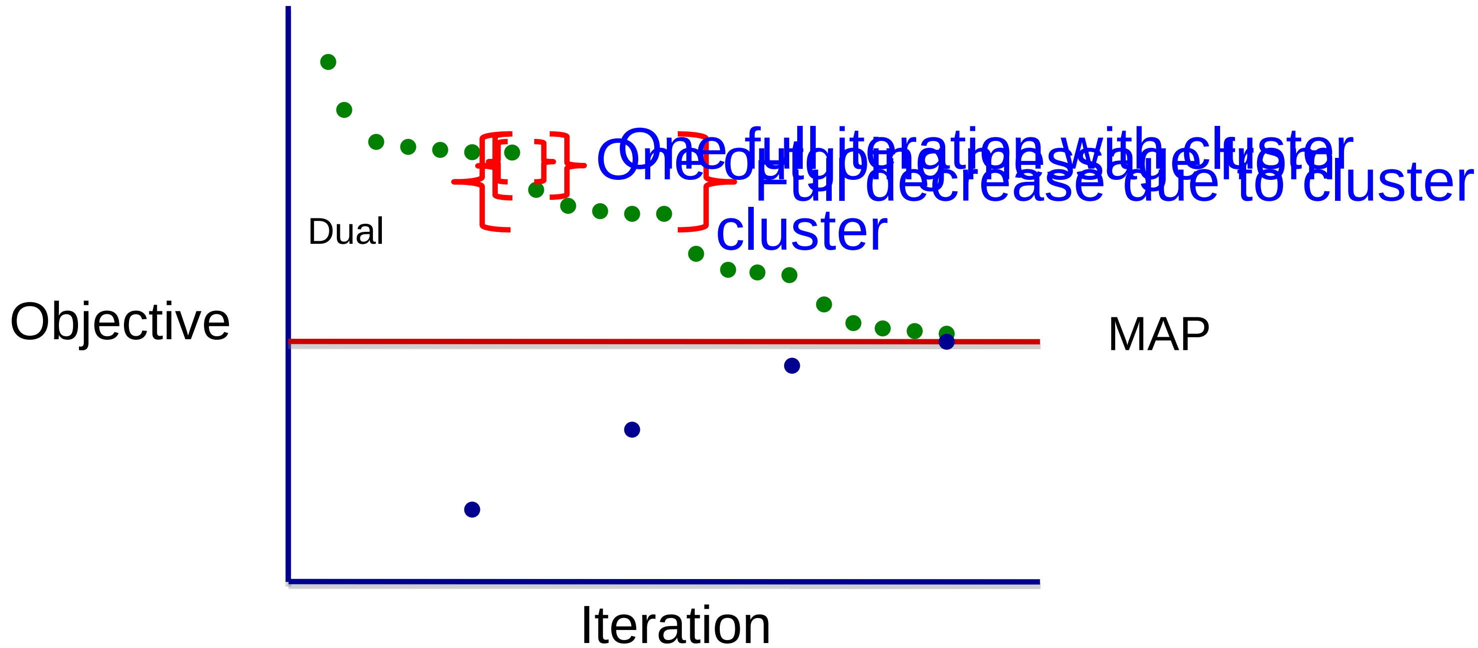
1. Run message-passing
2. Decode assignment from messages
3. Choose a cluster to add to relaxation
4. Warm start: initialize new cluster messages



Dual algorithm



What cluster to add next?



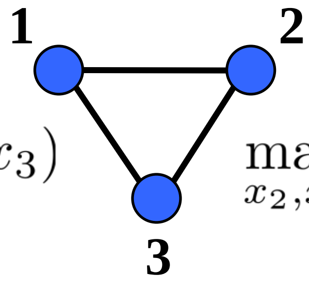
$$\sum_{e \in c} \max_{x_e} b_e(x_e) - \max_{x_c} \left[\sum_{e \in c} b_e(x_e) \right]$$

What cluster to add next?

$$\sum_{e \in c} \max_{x_e} b_e(x_e) - \max_{x_c} \left[\sum_{e \in c} b_e(x_e) \right]$$

$$\max_{x_1, x_2, x_3} [b_{12}(x_1, x_2) + b_{23}(x_2, x_3) + b_{13}(x_1, x_3)]$$

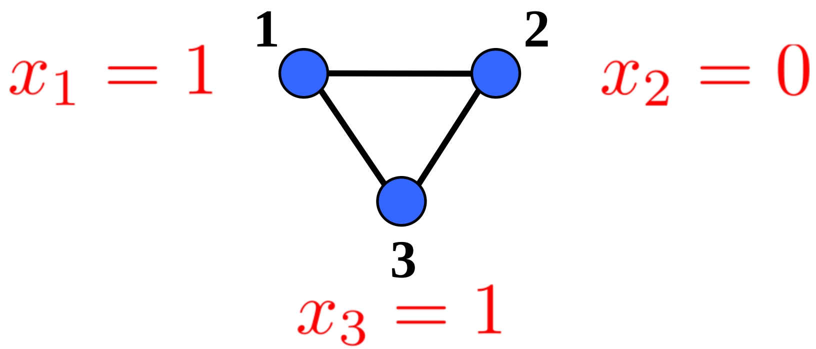
$$\max_{x_1, x_2} b_{12}(x_1, x_2)$$



$$\max_{x_1, x_3} b_{13}(x_1, x_3) \qquad \max_{x_2, x_3} b_{23}(x_2, x_3)$$

What cluster to add next?

$$\underbrace{\sum_{e \in c} \max_{x_e} b_e(x_e)}_{3*99} - \underbrace{\max_{x_c} \left[\sum_{e \in c} b_e(x_e) \right]}_{2*99-10}$$



$b_{ij}(x_i, x_j) = 99$ if $x_i \neq x_j$

$b_{ij}(x_i, x_j) = -10$ otherwise

If dual decreases, there was frustration

Related Work

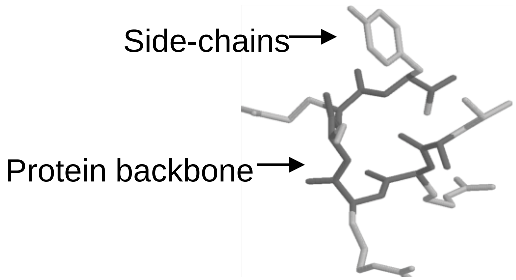
- Region-pursuit algorithm for generalized BP (Welling UAI '04)
  - Iteratively adds the regions that most change region free energy (algorithmically very similar)
  - Found that sometimes adding regions gave worse results
  - Our approach circumvents this by working with the dual LP

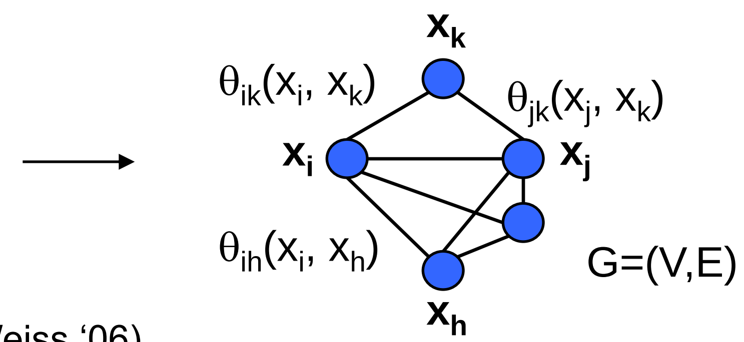
- Cutting-plane algorithm using cycle inequalities (Sontag & Jaakkola '08)
  - Selection criteria of *constraint violation* instead of *bound minimization*
  - SJ can efficiently *find* violated constraints, but *re-solving* is hard in primal

- Other dual formulations (Werner '05, Kolmogorov & Wainwright '05, Johnson et al. '07, Komodakis et al. '07, Globerson & Jaakkola '07)
- Concurrently, similar approach proposed by (Werner CVPR

Experiments: Protein design

- Given protein's 3D shape, choose amino-acids giving the most stable structure

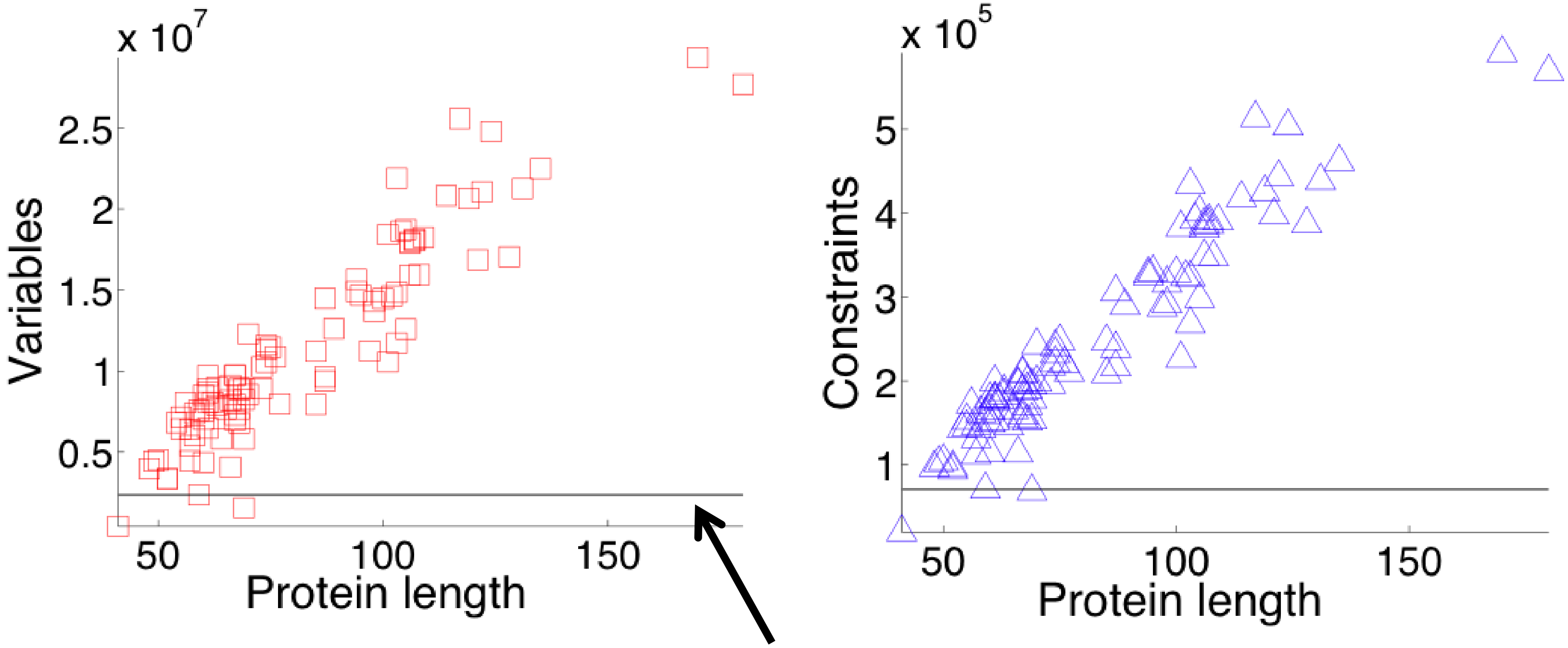




(MRFs from Yanover, Meltzer, Weiss '06)

- Each state corresponds to a choice of amino-acid and side-chain angle
- MRFs have 41-180 variables, each variable with 95-158 states
- Hard to solve
  - Very large treewidth
  - Many small cycles (20,000 triangles) and frustration

Primal LP, pairwise, is large



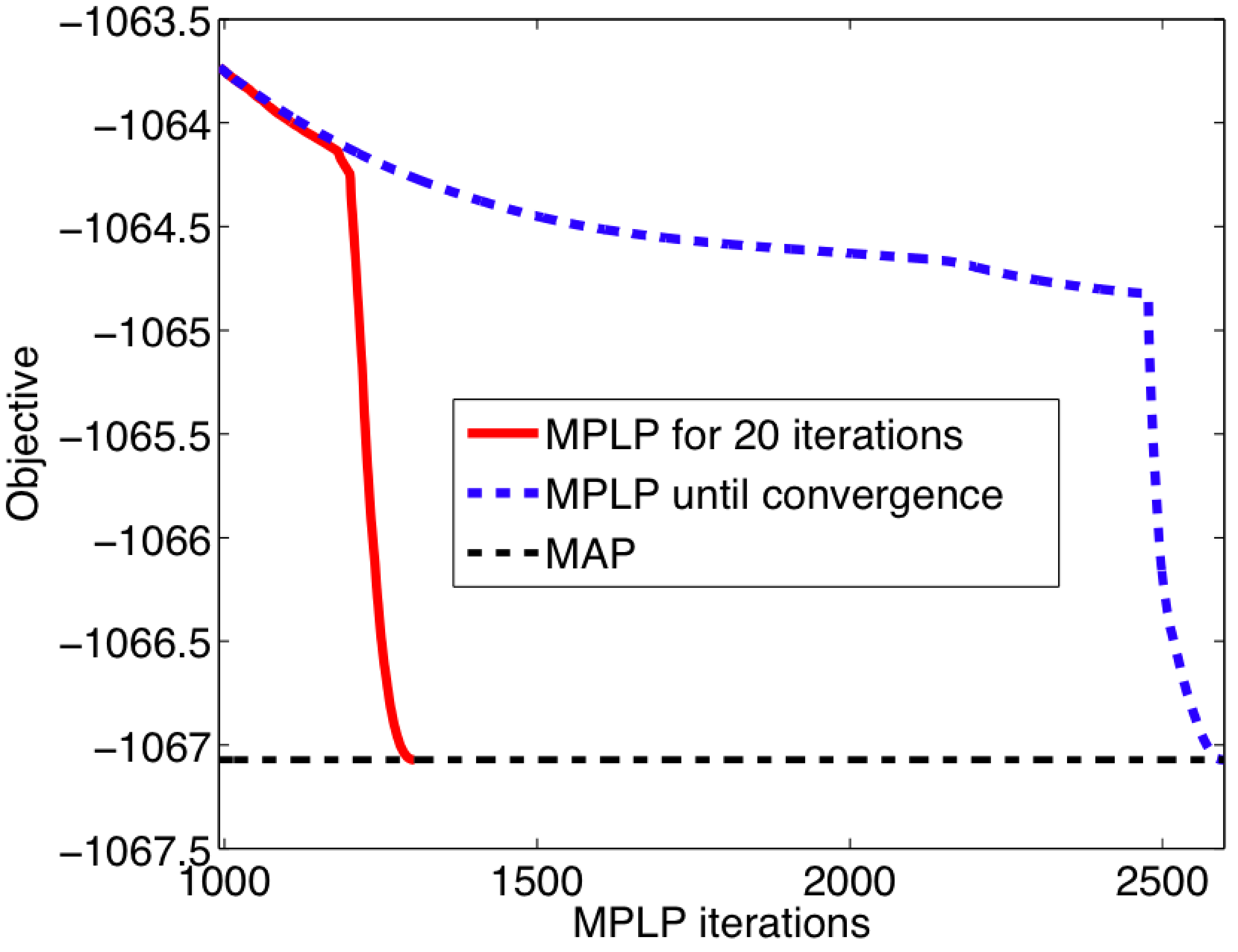
CPLEX can only run on 3:
must move to dual!

(Yanover, Meltzer, Weiss, JMLR '06)

Protein design results

- Pairwise consistency solves only 2 of the 97 proteins
(Yanover, Meltzer, Weiss, JMLR '06)
- With triplets, we solve 96 of 97 protein design problems (**!!!**)
- Between 5 and 735 triplets needed (median: 145)
  - Out of 20,000
  - Each triplet message needs >1 million computations
- 9.7 hours/problem (max: 11 days)

Faster to stop before for convergence

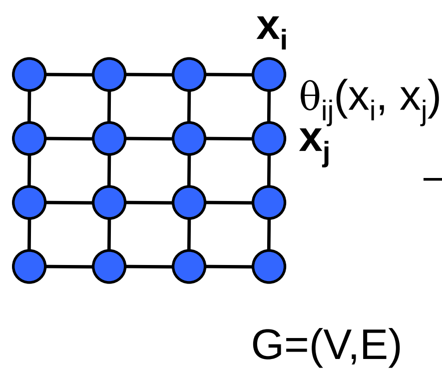


Experiments: Stereo vision

- How far away are these objects?

input: two images

(Tappen and Freeman '03)



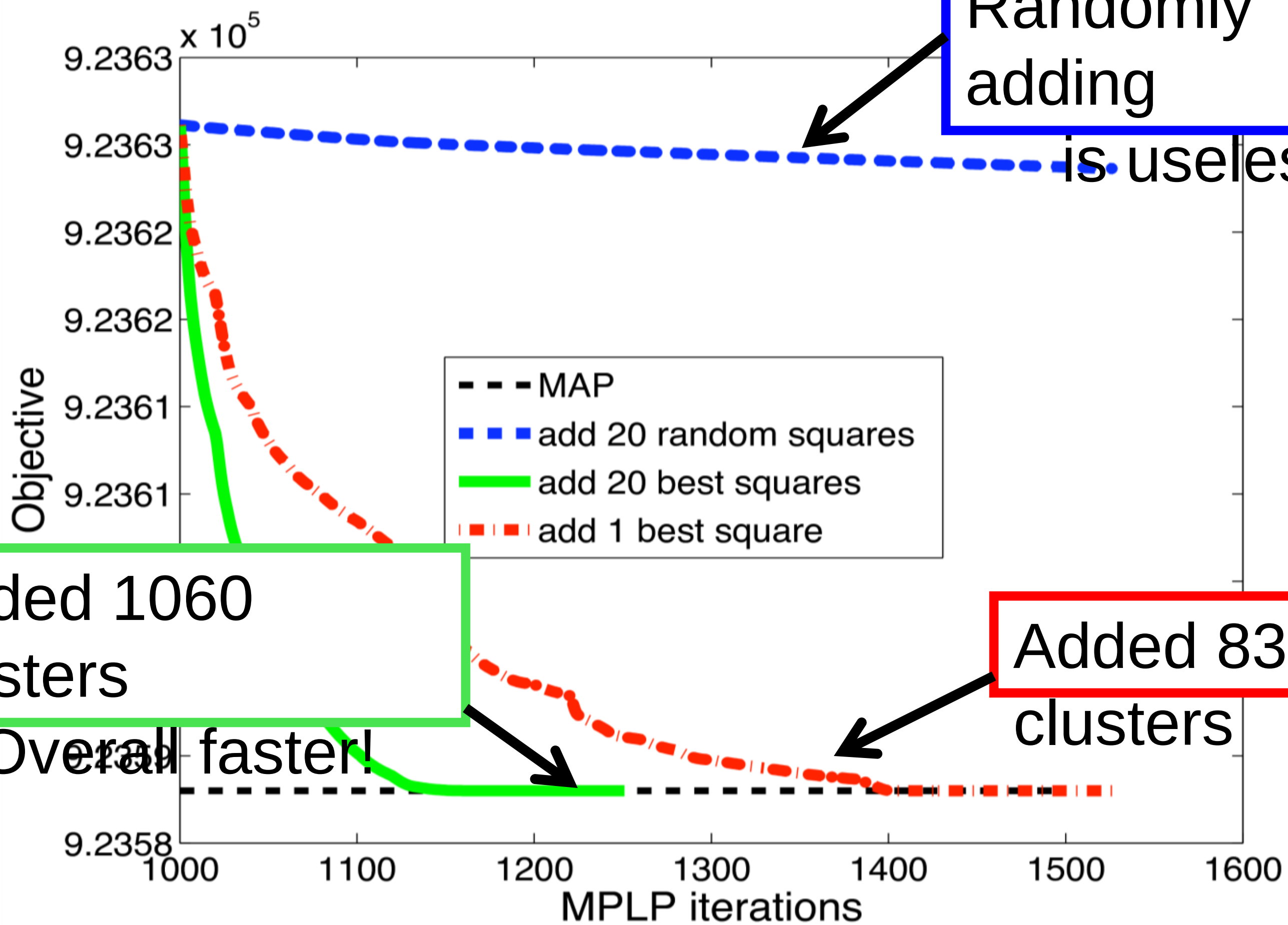
output: disparity
(MAP assignment)

- 116 x154 pixels (13,000 variables), each with 16 states
- Hard to solve
  - Treewidth is over 230
  - Many short cycles: 13,000 squares (4-cycles)
  - Non-convex potentials

Stereo vision results

- 10 images: variations on ‘Tsukuba’ sequence
- Pairwise consistency solves 6 of the 10 images
- We solve all of them

How aggressively to add clusters?



Randomly adding is useless

Added 1060 clusters

Overall faster!

Added 83 clusters

Future work

- Efficiently searching over clusters in the dual
- Structured prediction with large MRFs
- Extension to marginals and partition function
- Will soon release optimized code

Conclusions

- We give an algorithm to add clusters to message-passing
- Directly minimizes upper bound on MAP given by LP relaxation
- Using only a small number of clusters, solves some difficult real-world problems