Linear Programming Relaxations for Graphical Models

Tommi Jaakkola (MIT) , Amir Globerson (Hebrew Univ.)

Based on joint work with
M. Collins, M. Fromer, T. Koo,
M. Meila, O. Meshi, A. Rush, D. Sontag

Outline

- Inference with linear programs
  - problems, examples
  - geometry, polytopes, constraints
  - relaxations, consequences
  - dual decomposition, properties
- Solving linear programs
  - sub-gradient, coordinate descent
- Learning with LPs
  - structured prediction
  - coordinate descent, cutting plane
  - pseudo-max

Inference problems

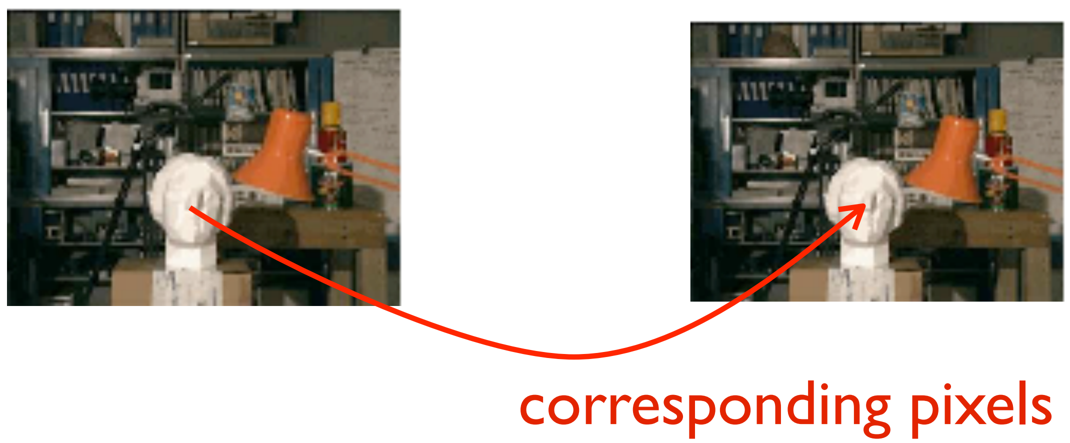
- Computer vision
  - e.g., image segmentation, stereopsis, scene analysis
- Computational biology
  - e.g., drug design, molecular structure prediction
- Natural language processing
  - e.g., parsing, machine translation, information extraction
- Medical informatics
  - e.g., automated diagnosis
- Exploratory analysis
  - e.g., learning Bayesian networks
- etc.

Inference problems

- Computer vision
  - e.g., image segmentation, **stereopsis**, scene analysis
- Computational biology
  - e.g., drug design, **molecular structure prediction**
- Natural language processing
  - e.g., **parsing**, machine translation, information extraction
- Medical informatics
  - e.g., automated diagnosis
- Exploratory analysis
  - e.g., learning **Bayesian networks**
- etc.

Inference problems

- Computer vision
  - e.g., image segmentation, **stereopsis**, scene analysis

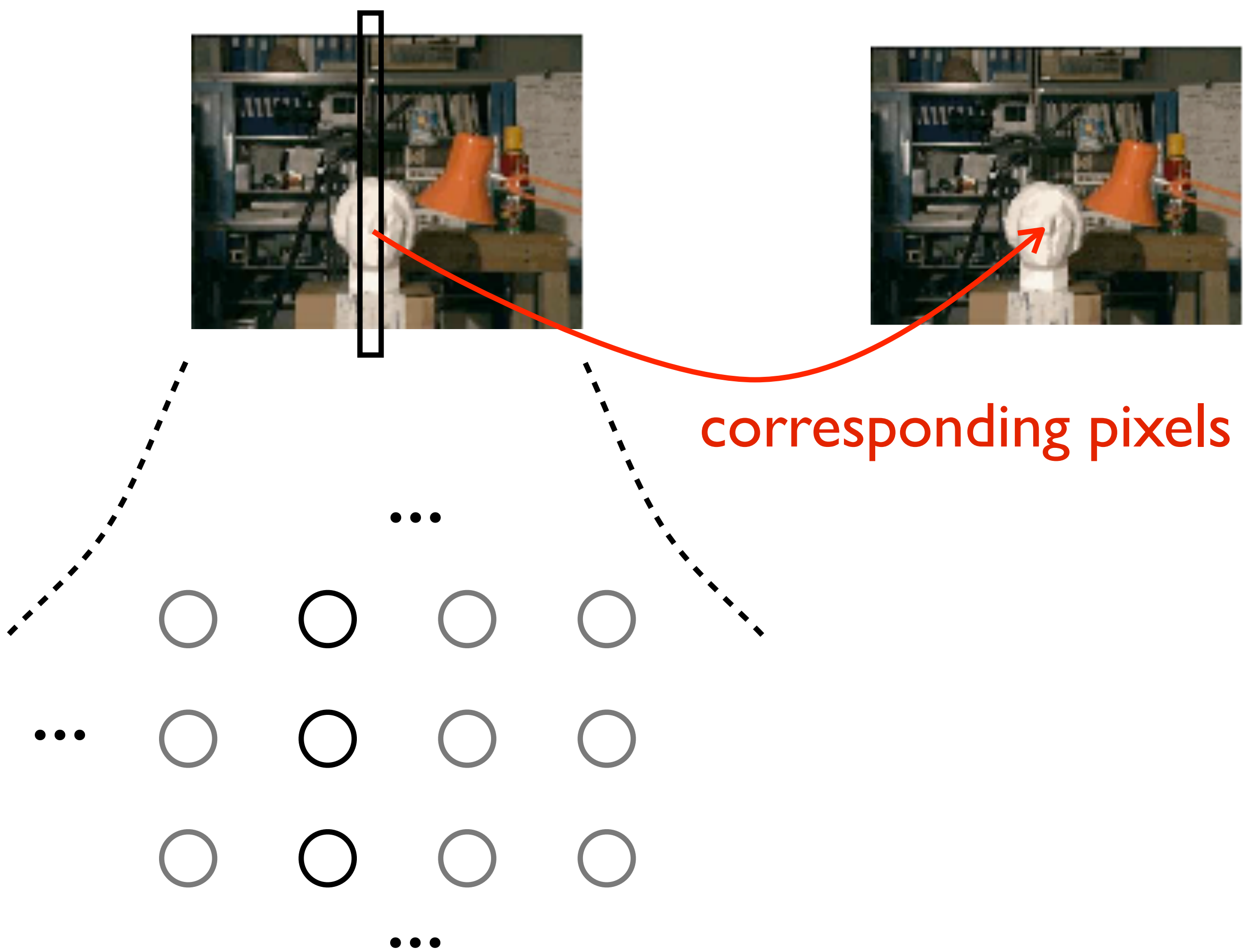


(Tappen & Freeman '03)

Goal: recover the depth of each pixel in the image

Inference problems

- Computer vision
  - e.g., image segmentation, **stereopsis**, scene analysis

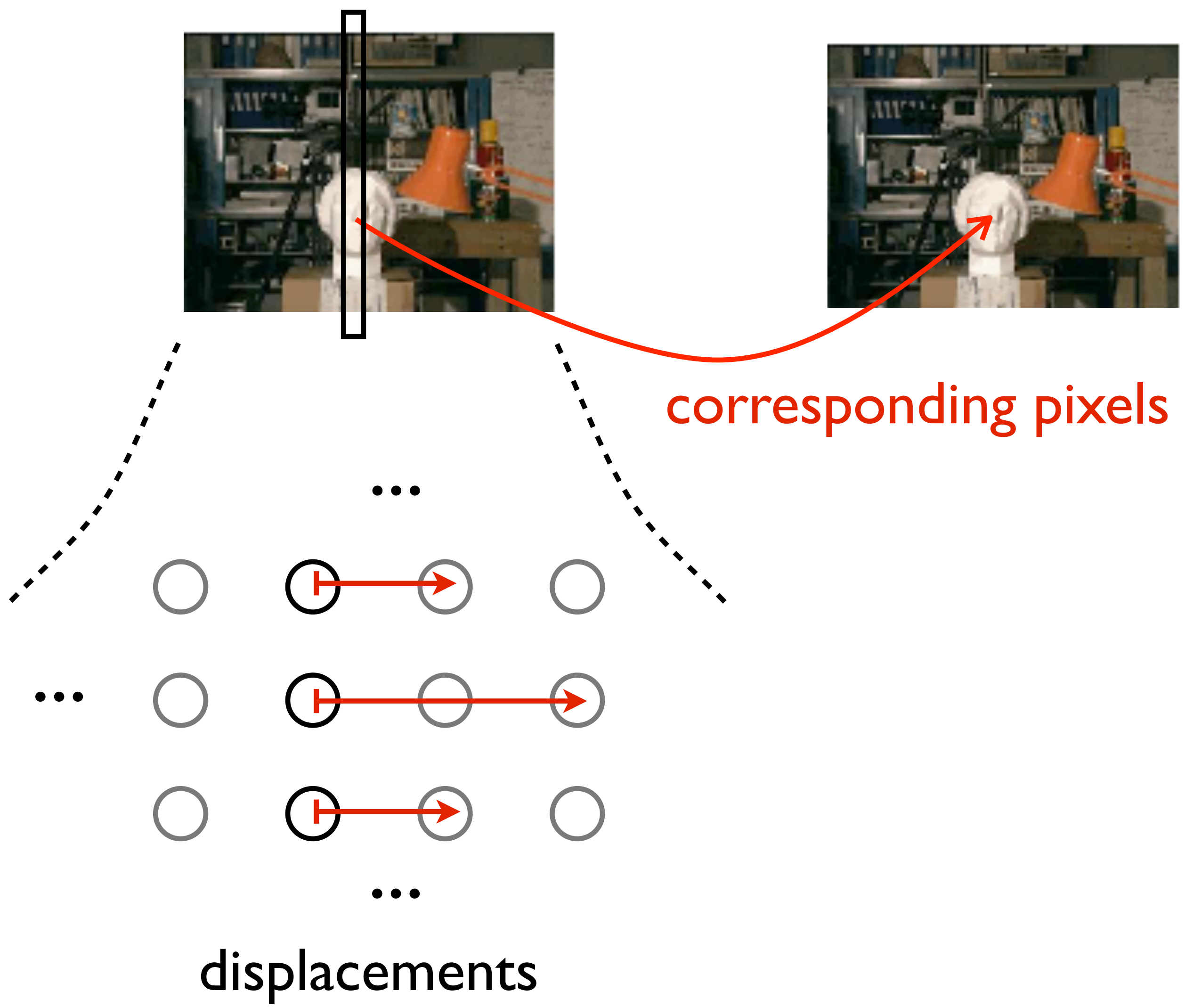


(Tappen & Freeman ’03)

Goal: recover the depth of each pixel in the image

Inference problems

- Computer vision
  - e.g., image segmentation, **stereopsis**, scene analysis

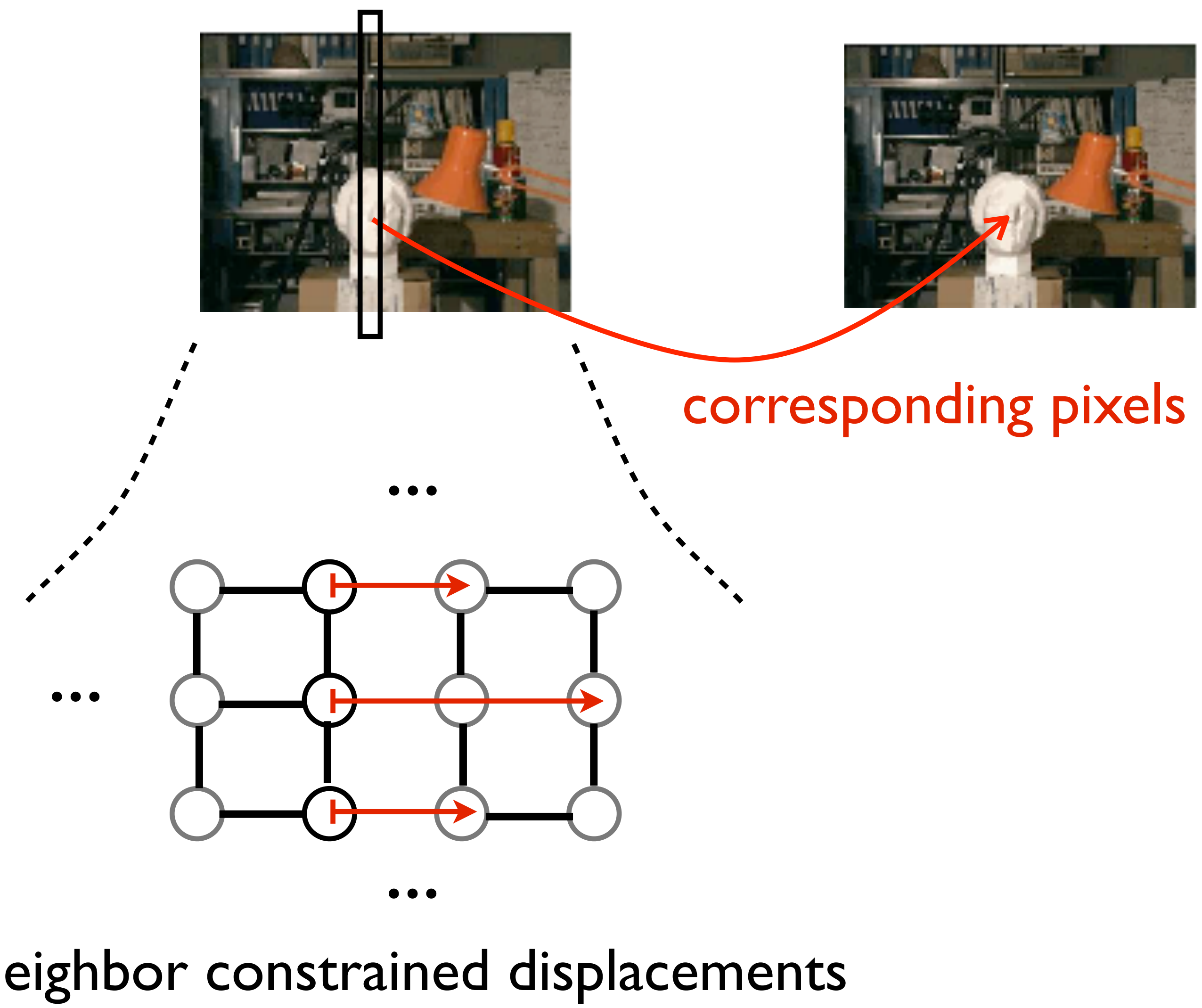


(Tappen & Freeman '03)

Goal: recover the depth of each pixel in the image

Inference problems

- Computer vision
  - e.g., image segmentation, **stereopsis**, scene analysis

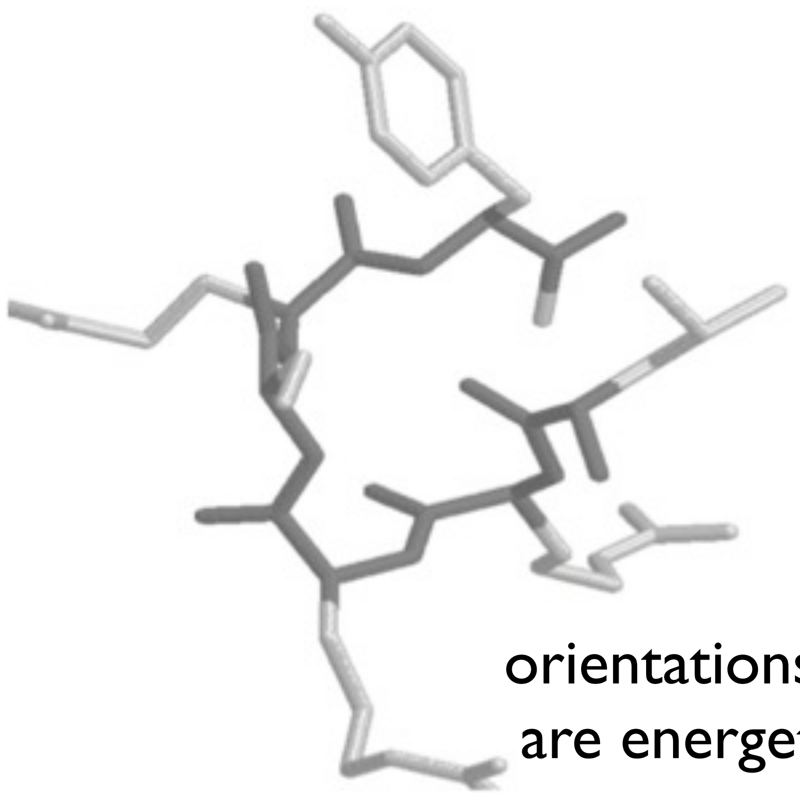


(Tappen & Freeman ’03)

Goal: recover the depth of each pixel in the image

Inference problems

- Computational biology
  - e.g., drug design, **molecular structure prediction**



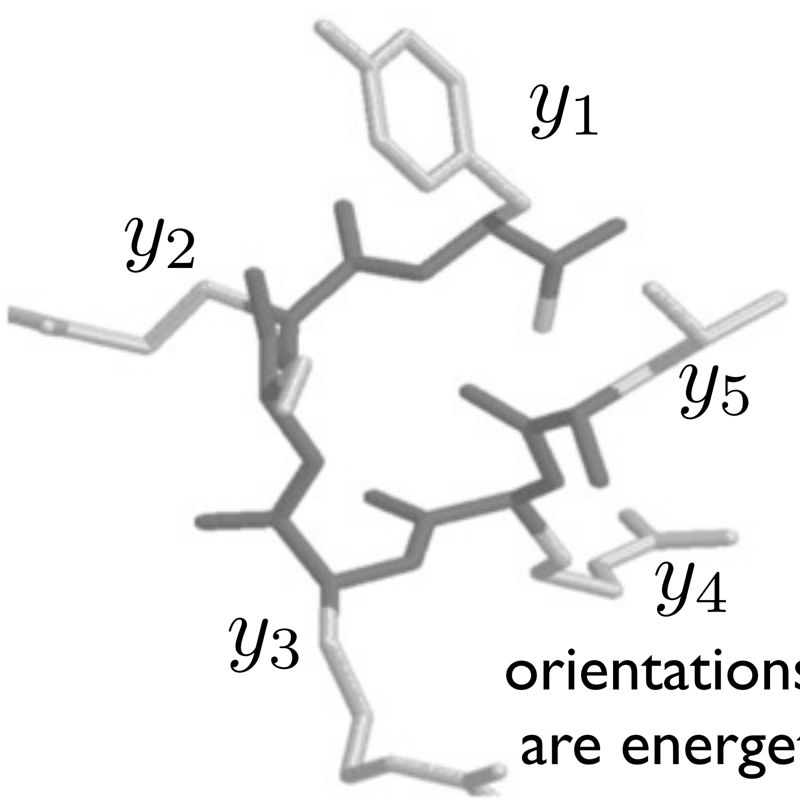
orientations of side chains
are energetically coupled

protein

Goal: recover energetically optimal side-chain orientations

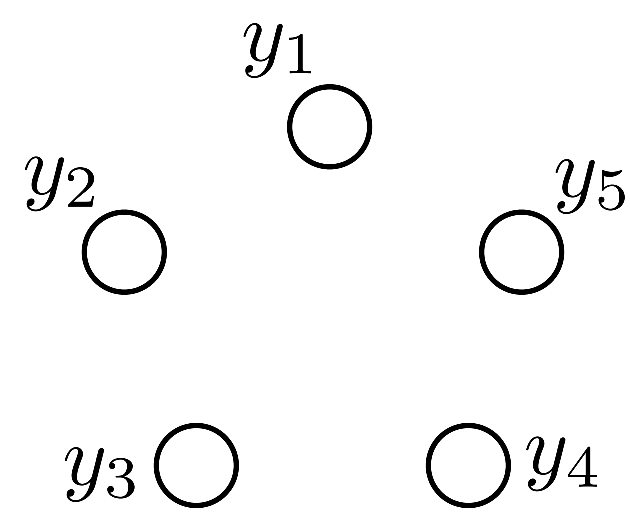
Inference problems

- Computational biology
  - e.g., drug design, **molecular structure prediction**



orientations of side chains
are energetically coupled

protein

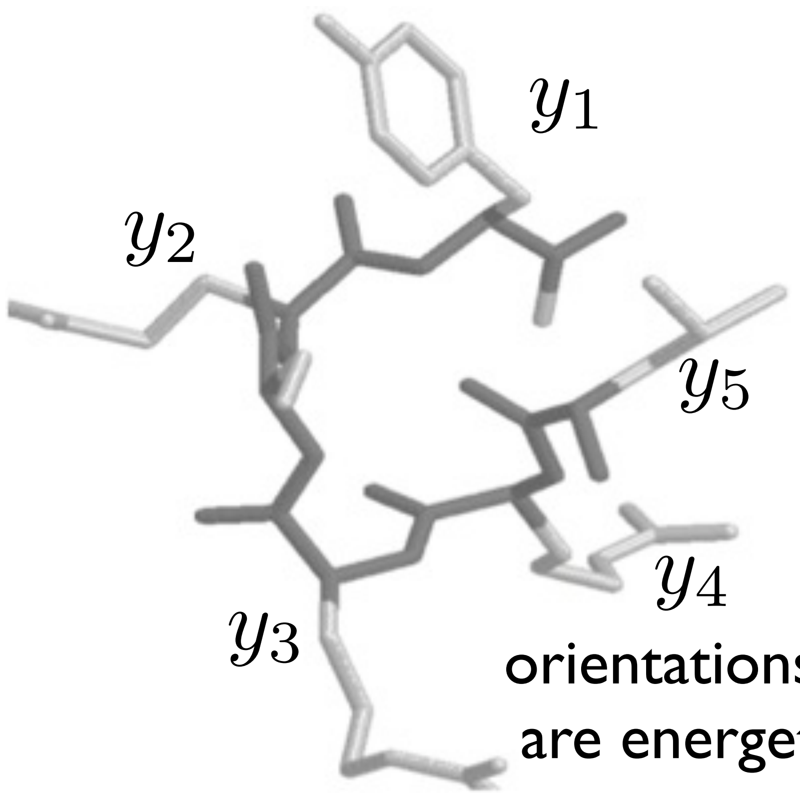


orientation angles as variables,

Goal: recover energetically optimal side-chain orientations

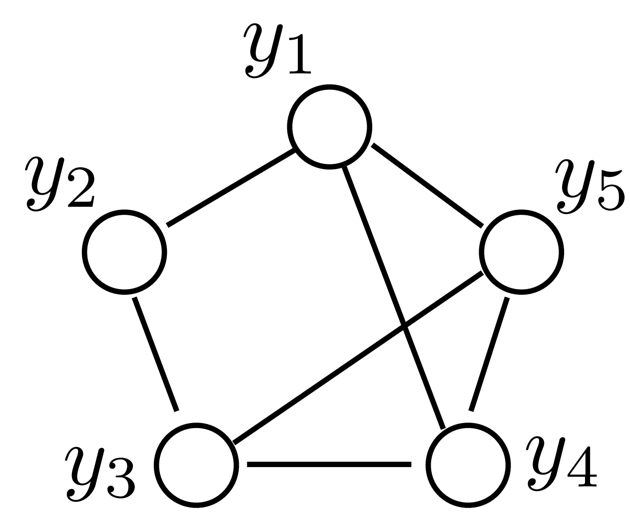
Inference problems

- Computational biology
  - e.g., drug design, **molecular structure prediction**



orientations of side chains are energetically coupled

protein

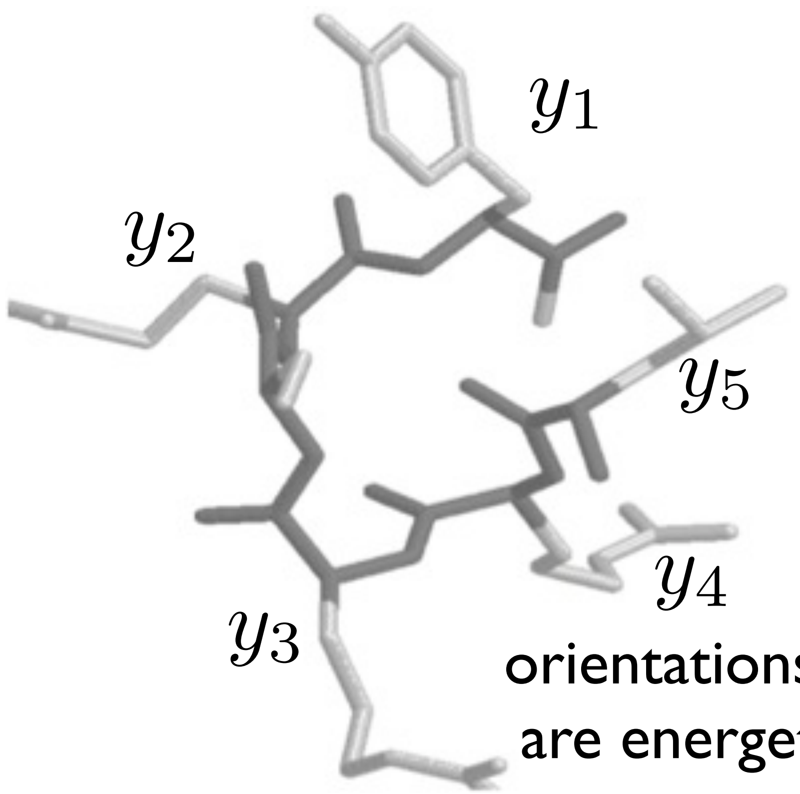


orientation angles as variables, densely coupled

Goal: recover energetically optimal side-chain orientations

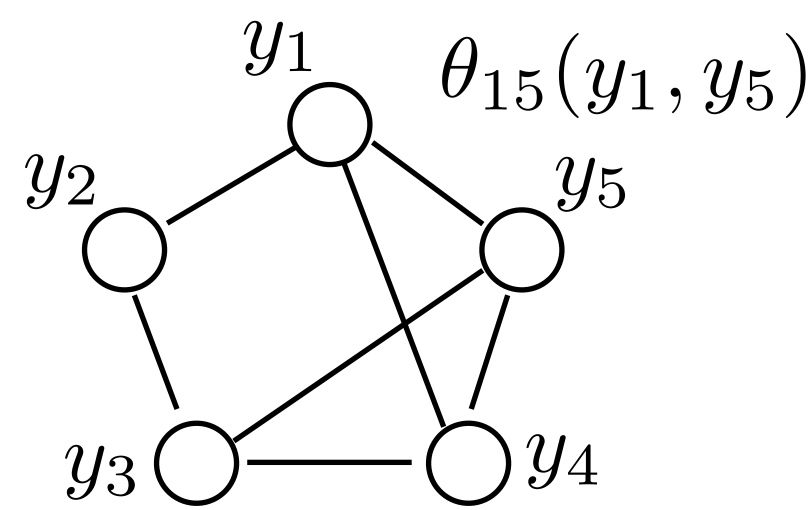
Inference problems

- Computational biology
  - e.g., drug design, **molecular structure prediction**



orientations of side chains are energetically coupled

protein

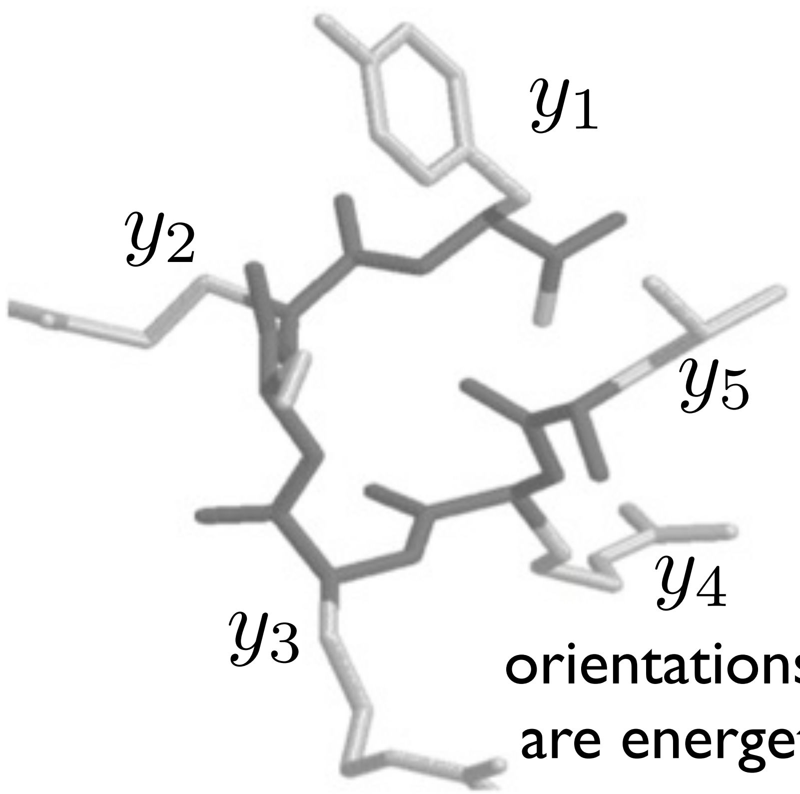


orientation angles as variables, densely coupled

Goal: recover energetically optimal side-chain orientations

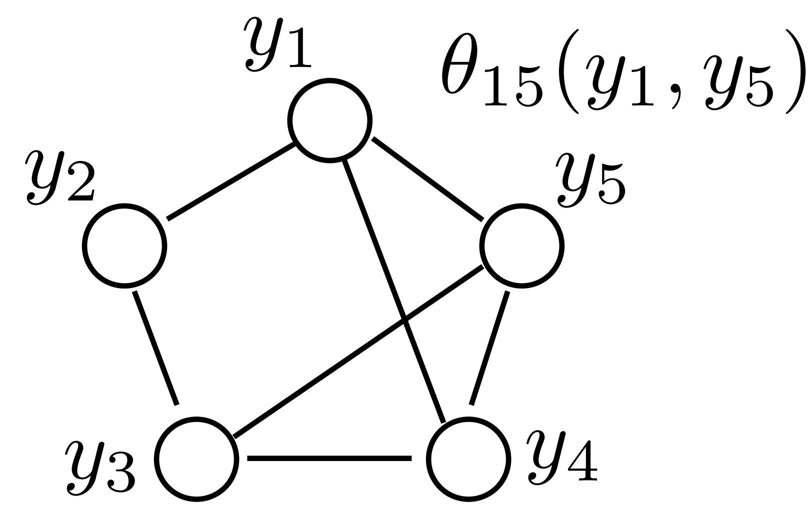
Inference problems

- Computational biology
  - e.g., drug design, **molecular structure prediction**



orientations of side chains are energetically coupled

protein



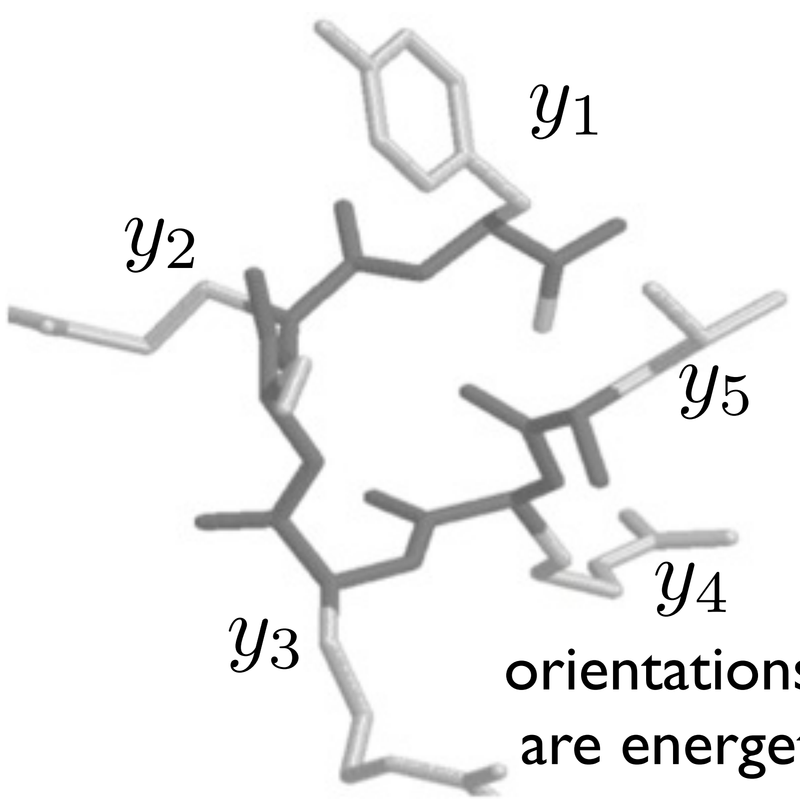
orientation angles as variables, densely coupled

$$P(y) = \frac{1}{Z} \exp\{ \sum_{(i,j) \in E} \theta_{ij}(y_i, y_j) \}$$

Goal: recover energetically optimal side-chain orientations

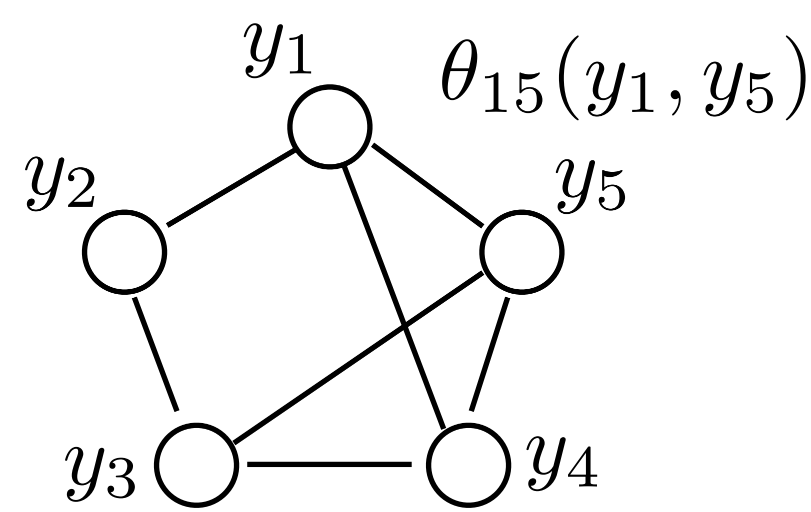
Inference problems

- Computational biology
  - e.g., drug design, **molecular structure prediction**



orientations of side chains are energetically coupled

protein



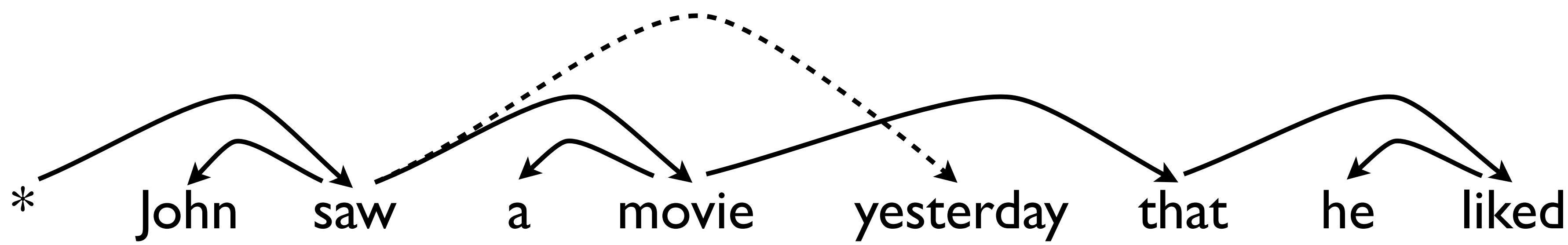
orientation angles as variables, densely coupled

$$\max_y \Big\{ \sum_{(i,j)\in E} \theta_{ij}(y_i, y_j) \Big\}$$

Goal: recover energetically optimal side-chain orientations

Inference problems

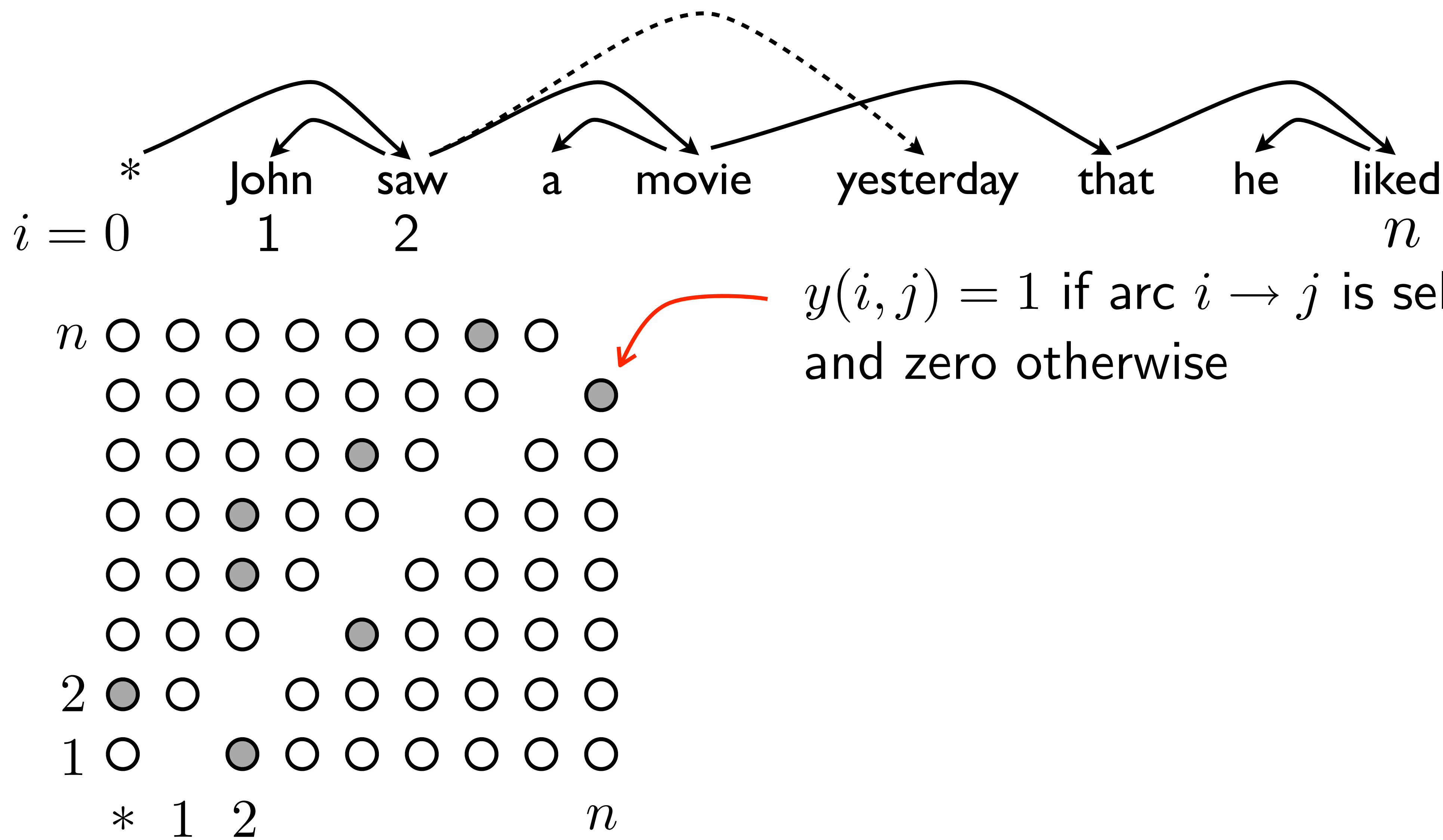
- Natural language processing
  - e.g., **dependency parsing**, machine translation, information extraction



Goal: find the highest scoring parse for any given sentence

Inference problems

- Natural language processing
  - e.g., **dependency parsing**, machine translation, information extraction

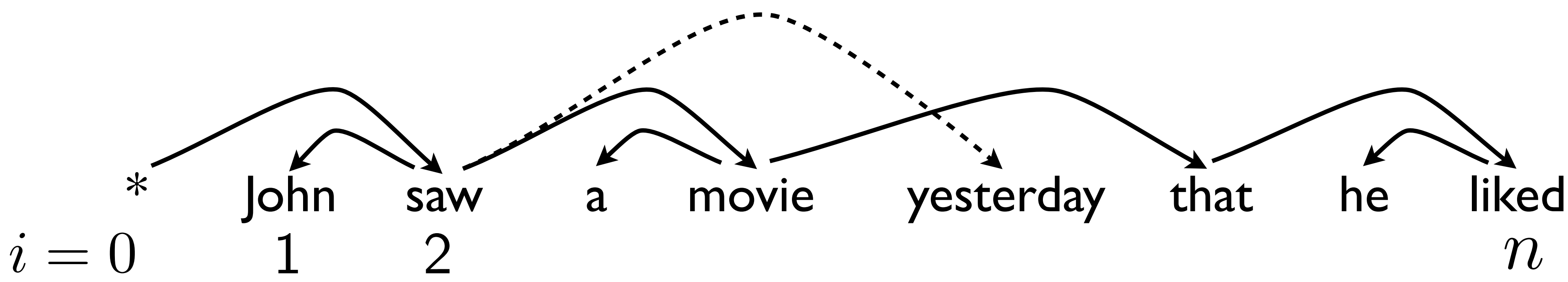


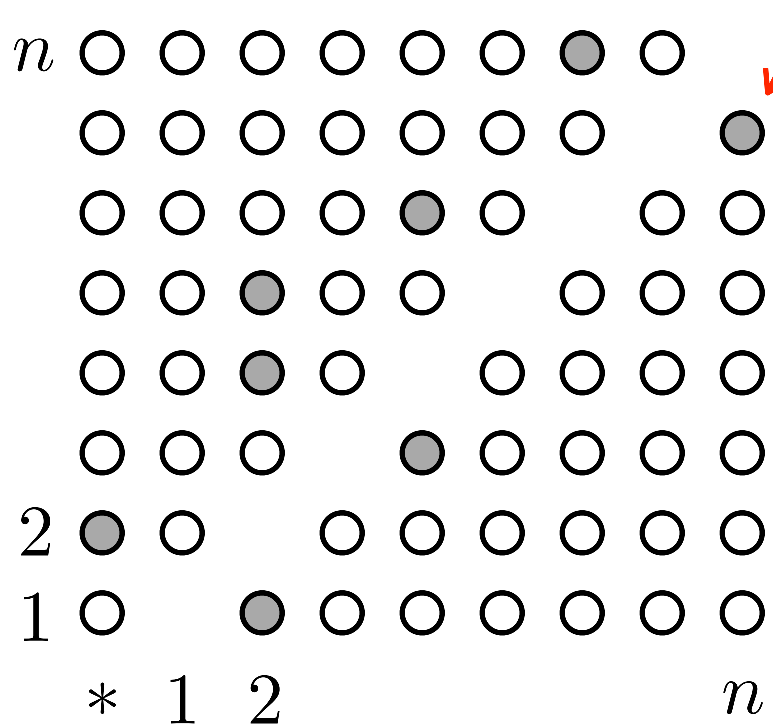
$y(i,j) = 1$ if arc $i \to j$ is selected and zero otherwise

Goal: find the highest scoring parse for any given sentence

Inference problems

- Natural language processing
  - e.g., **dependency parsing**, machine translation, information extraction





$y(i,j) = 1$ if arc $i \to j$ is selected and zero otherwise

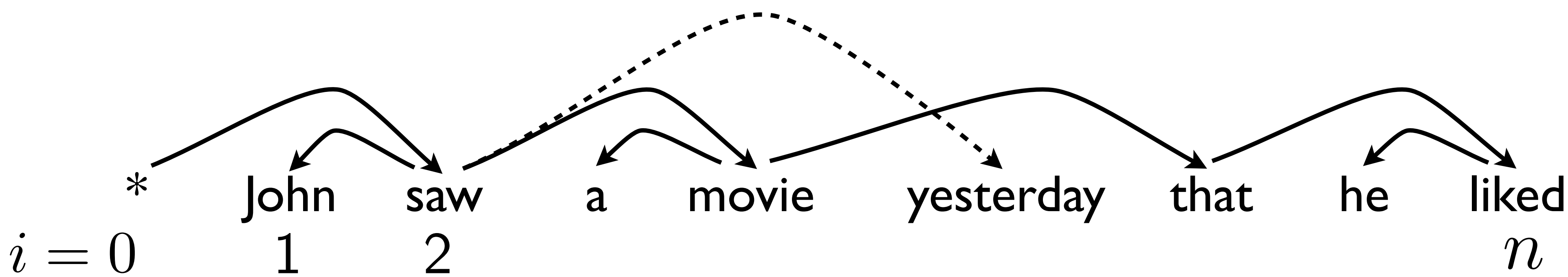
$$\max_y \left\{ \sum_{i,j} y(i,j)\theta(i,j) + \theta_T(y) \right\}$$

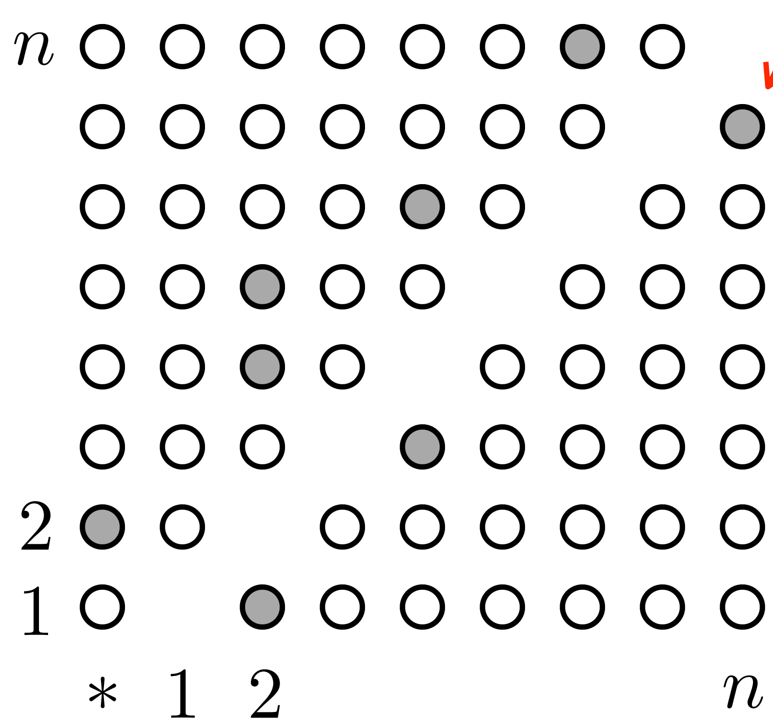
has to be a tree

Goal: find the highest scoring parse for any given sentence

Inference problems

- Natural language processing
  - e.g., **dependency parsing**, machine translation, information extraction





$y(i,j) = 1$ if arc $i \rightarrow j$ is selected and zero otherwise

$$\max_y \Big\{ \sum_i \theta_i(y(i,\cdot)) + \theta_T(y) \Big\}$$

has to be a tree

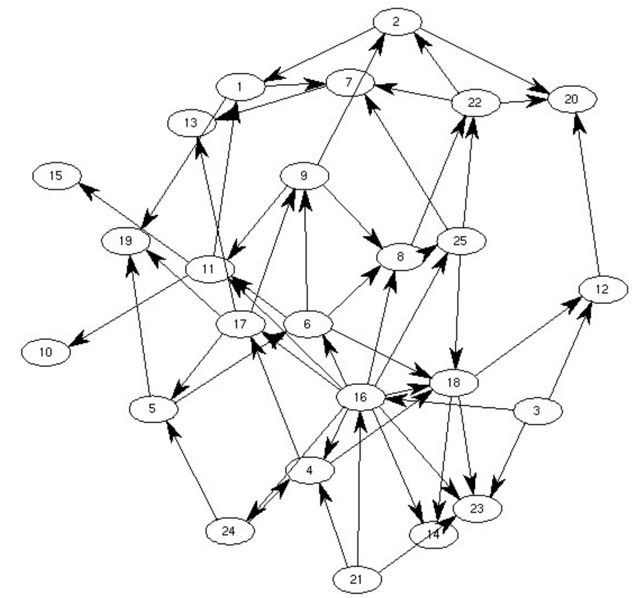
useful scoring functions couple modifier (arc) choices

Goal: find the highest scoring parse for any given sentence

Inference problems

- Exploratory analysis
  - e.g., **learning Bayesian networks**

```
2 2 0 1 1 2 0 0 2 0 2 ...
0 2 2 2 2 2 0 1 1 2 2
1 1 0 1 0 1 1 1 1 1 0
2 2 0 1 0 2 0 0 2 0 2
0 1 0 1 1 1 0 1 1 1 0
          ...
```

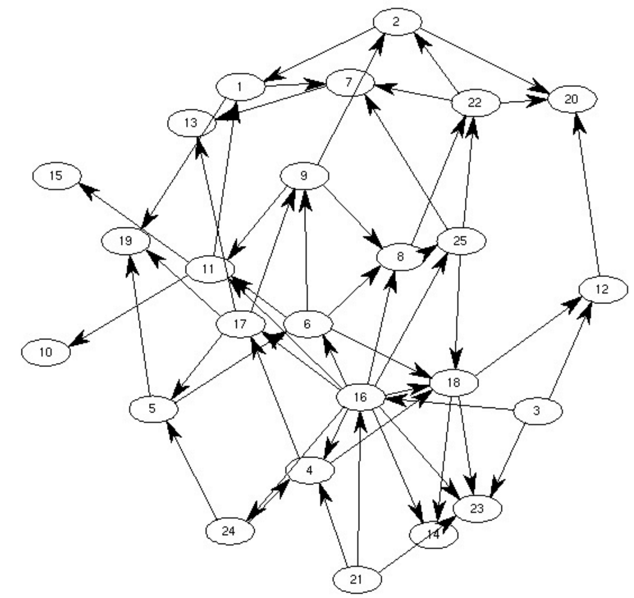


Goal: find the highest scoring acyclic graph

Inference problems

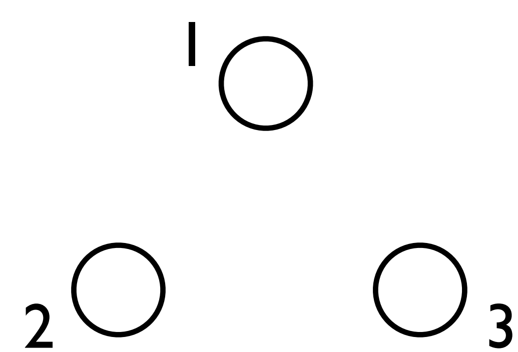
- Exploratory analysis
  - e.g., **learning Bayesian networks**

```
2 2 0 1 1 2 0 0 2 0 2 ...
0 2 2 2 2 2 0 1 1 2 2
1 1 0 1 0 1 1 1 1 1 0
2 2 0 1 0 2 0 0 2 0 2
0 1 0 1 1 1 0 1 1 1 0
          ...
```



$$y_1 \in \begin{array}{c} \emptyset \\ \{2\} \\ \{3\} \\ \{2,3\} \end{array}$$

$$y_2 \in \begin{array}{c} \emptyset \\ \{1\} \\ \{3\} \\ \{1,3\} \end{array} \qquad y_3 \in \begin{array}{c} \emptyset \\ \{1\} \\ \{2\} \\ \{1,2\} \end{array}$$

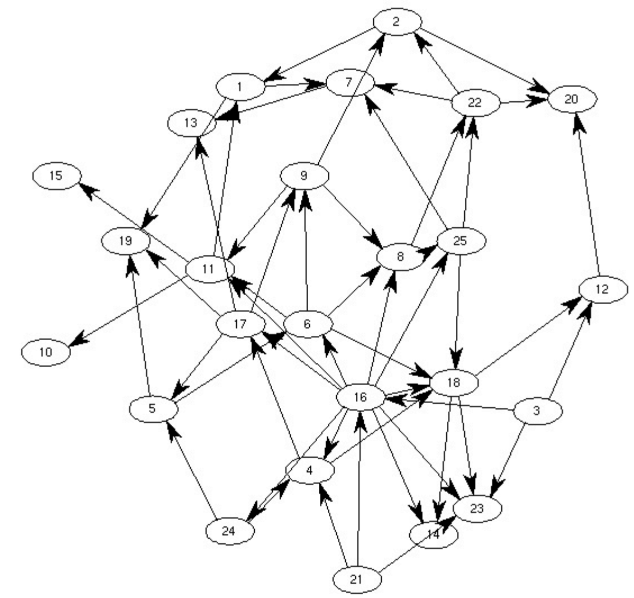


Goal: find the highest scoring acyclic graph

Inference problems

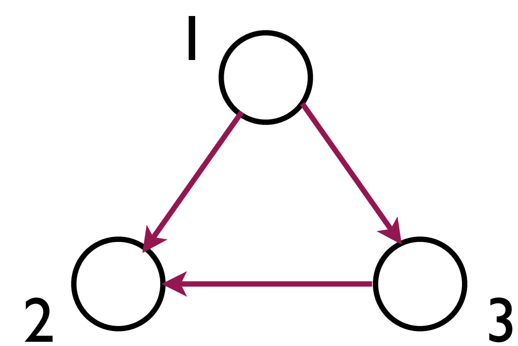
- Exploratory analysis
  - e.g., **learning Bayesian networks**

```
2 2 0 1 1 2 0 0 2 0 2 ...
0 2 2 2 2 2 0 1 1 2 2
1 1 0 1 0 1 1 1 1 1 0
2 2 0 1 0 2 0 0 2 0 2
0 1 0 1 1 1 0 1 1 1 0
          ...
```



$$y_1 \in \begin{array}{c} \emptyset \\ \{2\} \\ \{3\} \\ \{2,3\} \end{array}$$

$$y_2 \in \begin{array}{c} \emptyset \\ \{1\} \\ \{3\} \\ \{1,3\} \end{array} \qquad y_3 \in \begin{array}{c} \emptyset \\ \{1\} \\ \{2\} \\ \{1,2\} \end{array} \qquad \Longrightarrow$$

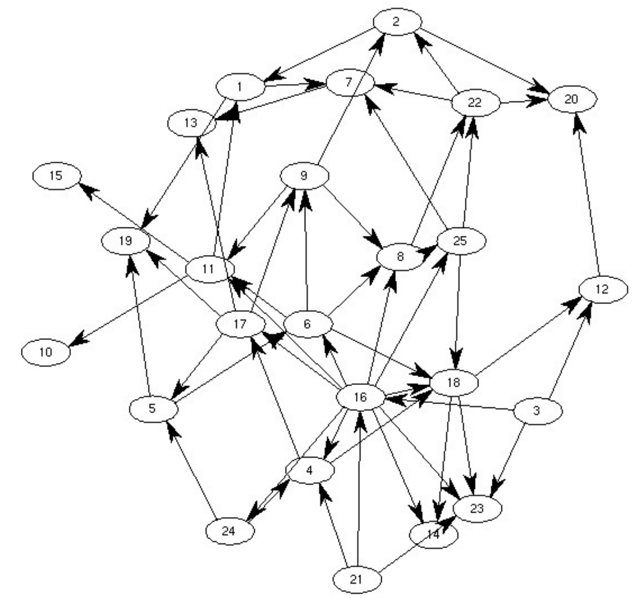


Goal: find the highest scoring acyclic graph

Inference problems

- Exploratory analysis
  - e.g., **learning Bayesian networks**

```
2 2 0 1 1 2 0 0 2 0 2 ...
0 2 2 2 2 2 0 1 1 2 2
1 1 0 1 0 1 1 1 1 1 0
2 2 0 1 0 2 0 0 2 0 2
0 1 0 1 1 1 0 1 1 1 0
          ...
```



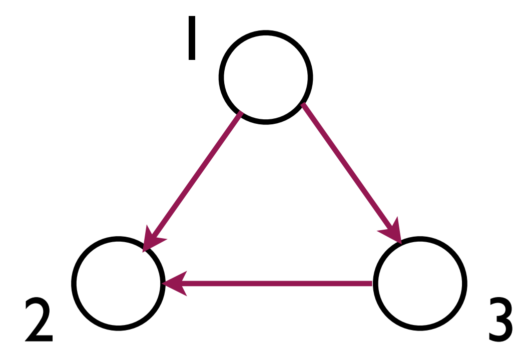
$\theta_1(\emptyset)$

$$y_1 \in \begin{array}{c} \emptyset \\ \{2\} \\ \{3\} \\ \{2,3\} \end{array}$$

$$y_2 \in \begin{array}{c} \emptyset \\ \{1\} \\ \{3\} \\ \{1,3\} \end{array} \quad \theta_2(\{1,3\})$$

$$y_3 \in \begin{array}{c} \emptyset \\ \{1\} \\ \{2\} \\ \{1,2\} \end{array} \quad \theta_3(\{1\})$$

$\Rightarrow$

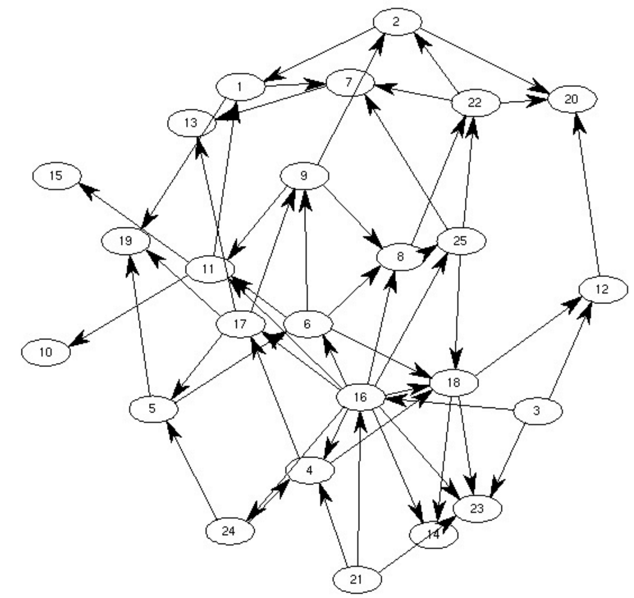


Goal: find the highest scoring acyclic graph

Inference problems

- Exploratory analysis
  - e.g., **learning Bayesian networks**

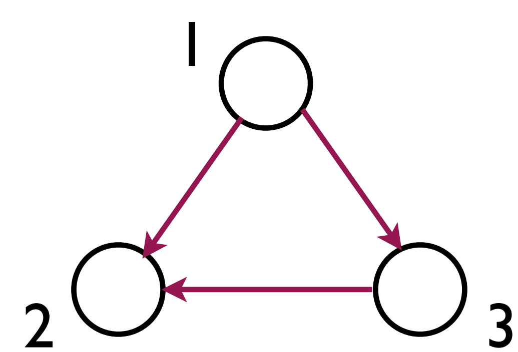
```
2 2 0 1 1 2 0 0 2 0 2 ...
0 2 2 2 2 2 0 1 1 2 2
1 1 0 1 0 1 1 1 1 1 0
2 2 0 1 0 2 0 0 2 0 2
0 1 0 1 1 1 0 1 1 1 0
          ...
```



$$y_1 \in \begin{array}{l} \emptyset \;\; \theta_1(\emptyset) \\ \{2\} \\ \{3\} \\ \{2,3\} \end{array}$$

$$y_2 \in \begin{array}{l} \emptyset \\ \{1\} \\ \{3\} \\ \{1,3\} \;\; \theta_2(\{1,3\}) \end{array} \qquad y_3 \in \begin{array}{l} \emptyset \\ \{1\} \;\; \theta_3(\{1\}) \\ \{2\} \\ \{1,2\} \end{array}$$

$\Rightarrow$



$$\max_y \Big\{ \sum_i \theta_i(y_i) + \theta_{DAG}(y) \Big\}$$

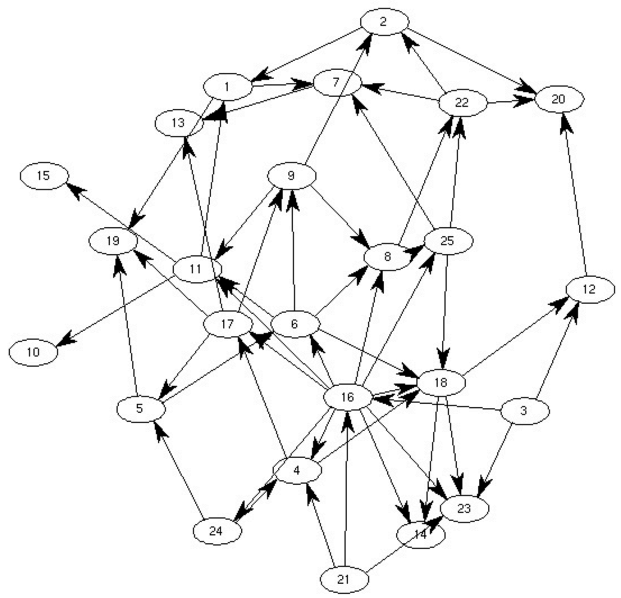
has to be a DAG

Goal: find the highest scoring acyclic graph

Inference problems

- Exploratory analysis
  - e.g., **learning Bayesian networks**

```
2 2 0 1 1 2 0 0 2 0 2 ...
0 2 2 2 2 2 0 1 1 2 2
1 1 0 1 0 1 1 1 1 1 0
2 2 0 1 0 2 0 0 2 0 2
0 1 0 1 1 1 0 1 1 1 0
          ...
```

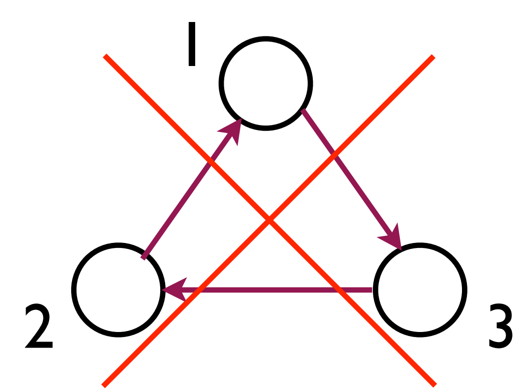


$$y_1 \in \begin{matrix} \emptyset \\ \{2\} \\ \{3\} \\ \{2,3\} \end{matrix} \quad \theta_1(\{2\})$$

$$y_2 \in \begin{matrix} \emptyset \\ \{1\} \\ \{3\} \\ \{1,3\} \end{matrix} \quad \theta_2(\{1,3\})$$

$$y_3 \in \begin{matrix} \emptyset \\ \{1\} \\ \{2\} \\ \{1,2\} \end{matrix} \quad \theta_3(\{1\})$$

$$\Rightarrow$$



$$\max_y \Big\{ \sum_i \theta_i(y_i) + \theta_{DAG}(y) \Big\}$$

has to be a DAG

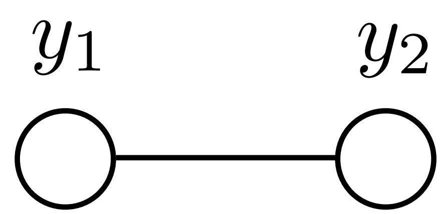
Goal: find the highest scoring acyclic graph

Inference problems

- MAP inference problems are often provably hard
  - protein design (Pierce et al. ’02)
  - dependency parsing with sibling scoring (McDonald et al. ’07)
  - structure learning of Bayesian networks (Chickering ’96, etc.)
  - etc.
- But “typical” instances of these problems may be solvable substantially faster
- We will discuss here (dual) linear programming (LP) relaxations for solving “typical” instances of such problems

Markov Random Fields

- Let's start with a simple Markov model over binary variables



$$y_i \in \{0, 1\}$$

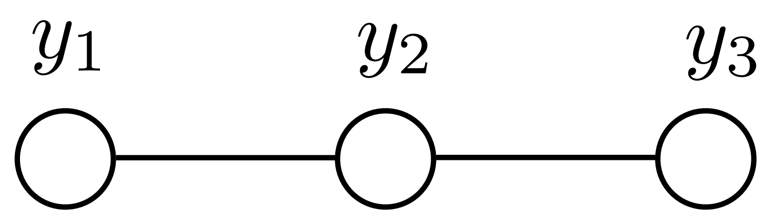
$$P(y;\theta) = \frac{1}{Z(\theta)} \exp\left\{\theta_{12}(y_1, y_2)\right\}$$

where

$$\theta = \begin{bmatrix} \theta_{12}(0,0) \\ \theta_{12}(1,0) \\ \theta_{12}(0,1) \\ \theta_{12}(1,1) \end{bmatrix}$$

Markov Random Fields

- Let's start with a simple Markov model over binary variables



$$y_i \in \{0, 1\}$$

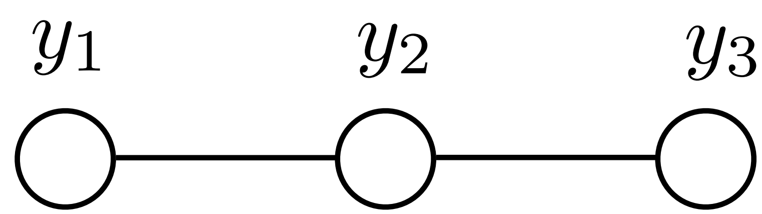
$$P(y; \theta) = \frac{1}{Z(\theta)} \exp\left\{\theta_{12}(y_1, y_2) + \theta_{23}(y_2, y_3)\right\}$$

where

$$\theta = \begin{bmatrix} \theta_{12}(0,0) \\ \theta_{12}(1,0) \\ \theta_{12}(0,1) \\ \theta_{12}(1,1) \\ \theta_{23}(0,0) \\ \theta_{23}(1,0) \\ \theta_{23}(0,1) \\ \theta_{23}(1,1) \end{bmatrix}$$

Markov Random Fields

- Let's start with a simple Markov model over binary variables



$$y_i \in \{0, 1\}$$

- The MAP problem we wish to solve is

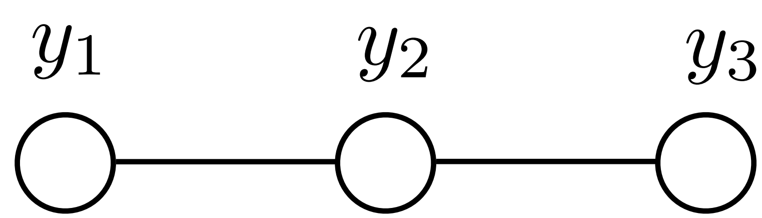
$$(y_1^*, y_2^*, y_3^*) = \arg\max_{y_1, y_2, y_3} \{\theta_{12}(y_1, y_2) + \theta_{23}(y_2, y_3)\}$$

where

$$\theta = \begin{bmatrix} \theta_{12}(0,0) \\ \theta_{12}(1,0) \\ \theta_{12}(0,1) \\ \theta_{12}(1,1) \\ \theta_{23}(0,0) \\ \theta_{23}(1,0) \\ \theta_{23}(0,1) \\ \theta_{23}(1,1) \end{bmatrix}$$

Markov Random Fields

- Let’s start with a simple Markov model over binary variables



$$y_i \in \{0, 1\}$$

- The MAP problem we wish to solve is

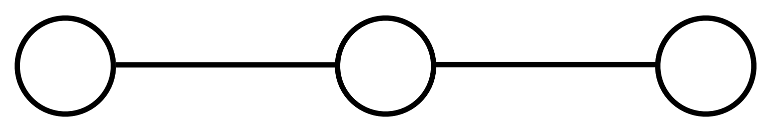
$$(y_1^*, y_2^*, y_3^*) = \arg\max_{y_1, y_2, y_3} \{\theta_{12}(y_1, y_2) + \theta_{23}(y_2, y_3)\}$$

where

$$\theta_{12}(1,0) + \theta_{23}(0,1) = \begin{bmatrix} \theta_{12}(0,0) \\ \theta_{12}(1,0) \\ \theta_{12}(0,1) \\ \theta_{12}(1,1) \\ \theta_{23}(0,0) \\ \theta_{23}(1,0) \\ \theta_{23}(0,1) \\ \theta_{23}(1,1) \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

Markov Random Fields

- Let's start with a simple Markov model over binary variables



$$y_i \in \{0, 1\}$$

- The MAP problem we wish to solve is

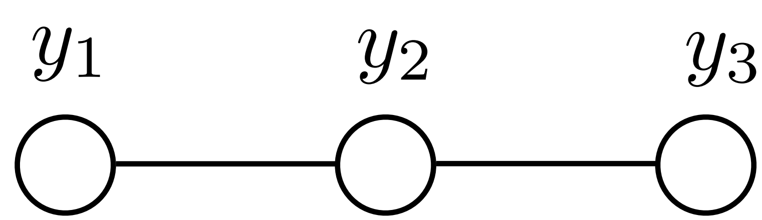
$$(y_1^*, y_2^*, y_3^*) = \arg\max_{y_1, y_2, y_3} \{\theta_{12}(y_1, y_2) + \theta_{23}(y_2, y_3)\}$$

where

$$\theta_{12}(y_1, y_2) + \theta_{23}(y_2, y_3) = \begin{bmatrix} \theta_{12}(0,0) \\ \theta_{12}(1,0) \\ \theta_{12}(0,1) \\ \theta_{12}(1,1) \\ \theta_{23}(0,0) \\ \theta_{23}(1,0) \\ \theta_{23}(0,1) \\ \theta_{23}(1,1) \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

Markov Random Fields

- Let's start with a simple Markov model over binary variables



$$y_i \in \{0, 1\}$$

- The MAP problem we wish to solve is

$$(y_1^*, y_2^*, y_3^*) = \arg\max_{y_1, y_2, y_3} \{\theta_{12}(y_1, y_2) + \theta_{23}(y_2, y_3)\}$$

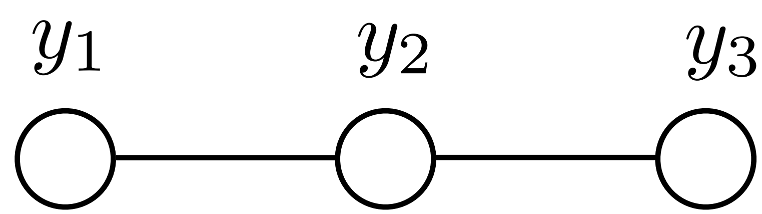
where

$$\theta_{12}(y_1, y_2) + \theta_{23}(y_2, y_3) = \begin{bmatrix} \theta_{12}(0,0) \\ \theta_{12}(1,0) \\ \theta_{12}(0,1) \\ \theta_{12}(1,1) \\ \theta_{23}(0,0) \\ \theta_{23}(1,0) \\ \theta_{23}(0,1) \\ \theta_{23}(1,1) \end{bmatrix} \cdot \begin{bmatrix} 1[y_1 = 0 \wedge y_2 = 0] \\ 1[y_1 = 1 \wedge y_2 = 0] \\ 1[y_1 = 0 \wedge y_2 = 1] \\ 1[y_1 = 1 \wedge y_2 = 1] \\ 1[y_2 = 0 \wedge y_3 = 0] \\ 1[y_2 = 1 \wedge y_3 = 0] \\ 1[y_2 = 0 \wedge y_3 = 1] \\ 1[y_2 = 1 \wedge y_3 = 1] \end{bmatrix}$$

$$\theta \quad \cdot \quad \phi(y)$$

Markov Random Fields

- Let's start with a simple Markov model over binary variables



$$y_i \in \{0, 1\}$$

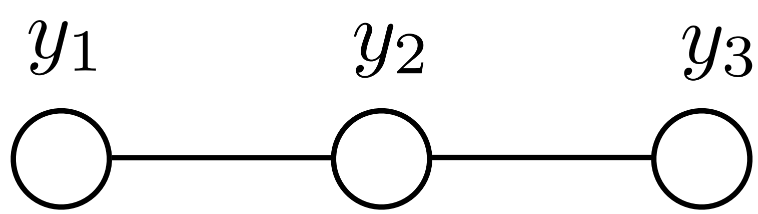
- The MAP problem we wish to solve is

$$(y_1^*, y_2^*, y_3^*) = \arg\max_{y_1, y_2, y_3} \left\{\theta_{12}(y_1, y_2) + \theta_{23}(y_2, y_3)\right\}$$

where $\quad \theta_{12}(y_1, y_2) + \theta_{23}(y_2, y_3) = \theta \cdot \phi(y)$

Markov Random Fields

- Let's start with a simple Markov model over binary variables



$y_i \in \{0, 1\}$

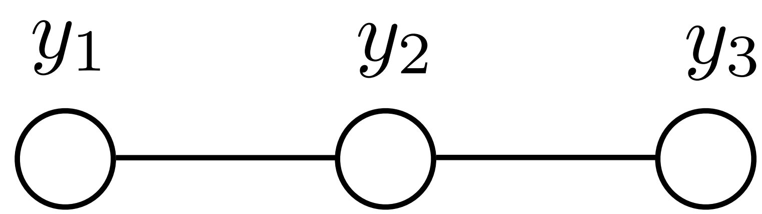
- The MAP problem we wish to solve is

$$y^* = \arg\max_y \left\{ \theta \cdot \phi(y) \right\}$$

where $\quad \theta_{12}(y_1, y_2) + \theta_{23}(y_2, y_3) = \theta \cdot \phi(y)$

Markov Random Fields

- Let's start with a simple Markov model over binary variables

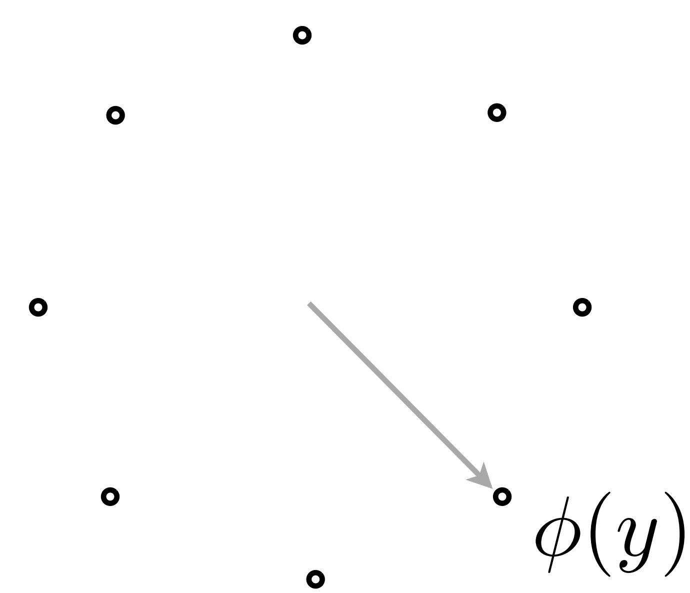


$y_i \in \{0, 1\}$

- The MAP problem we wish to solve is

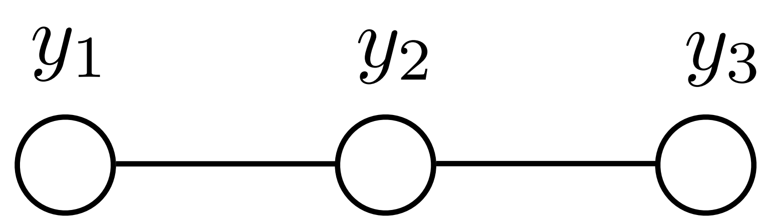
$$y^* = \arg\max_{y} \left\{ \theta \cdot \phi(y) \right\}$$

where $\quad \theta_{12}(y_1, y_2) + \theta_{23}(y_2, y_3) = \theta \cdot \phi(y)$



Markov Random Fields

- Let's start with a simple Markov model over binary variables

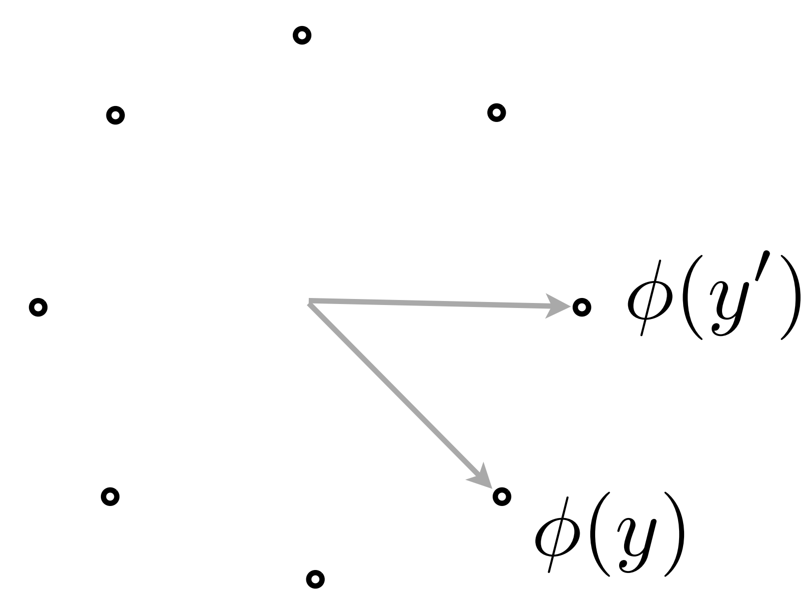


$$y_i \in \{0, 1\}$$

- The MAP problem we wish to solve is

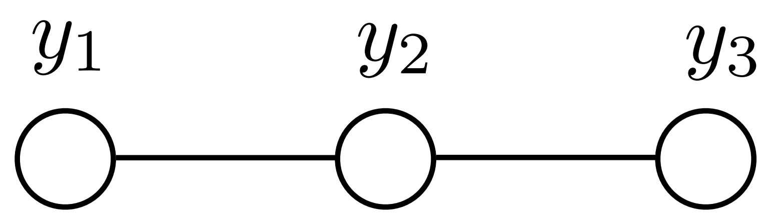
$$y^* = \arg\max_y \{\theta \cdot \phi(y)\}$$

where $\theta_{12}(y_1, y_2) + \theta_{23}(y_2, y_3) = \theta \cdot \phi(y)$



Markov Random Fields

- Let's start with a simple Markov model over binary variables

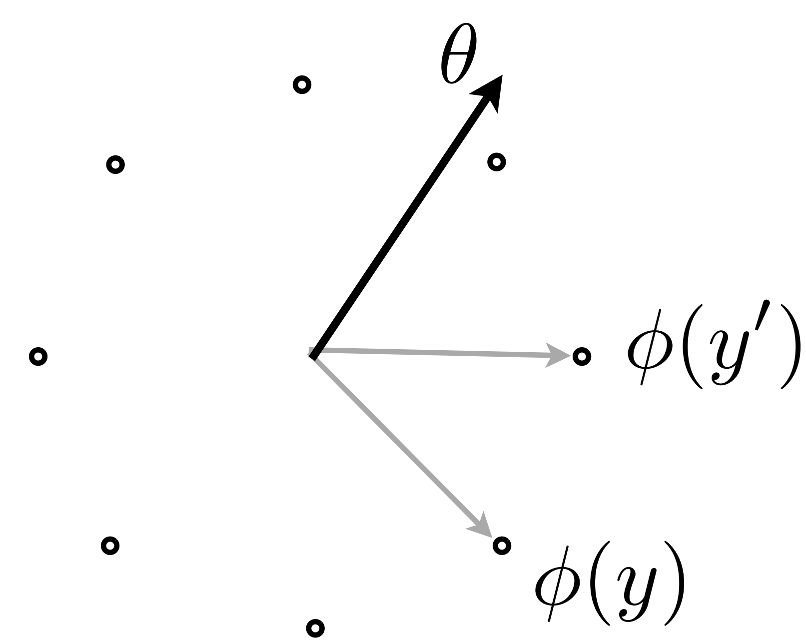


$y_i \in \{0, 1\}$

- The MAP problem we wish to solve is

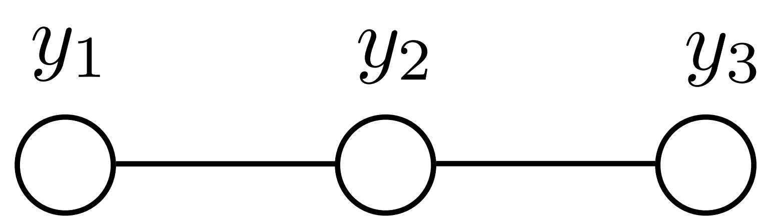
$$y^* = \arg\max_y \left\{ \theta \cdot \phi(y) \right\}$$

where $\theta_{12}(y_1, y_2) + \theta_{23}(y_2, y_3) = \theta \cdot \phi(y)$



MAP assignment

- Let's start with a simple Markov model over binary variables



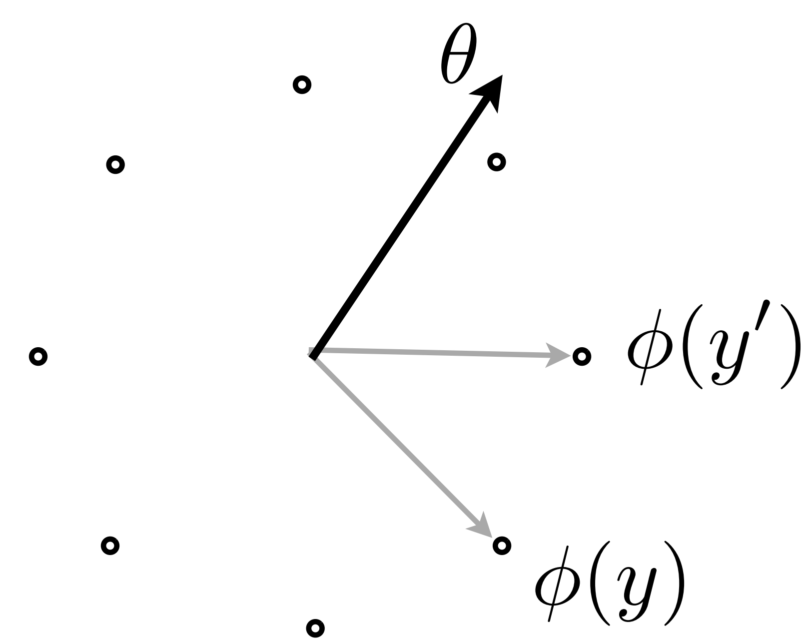
$$y_i \in \{0, 1\}$$

- The MAP problem we wish to solve is

$$y^* = \arg\max_y \{\theta \cdot \phi(y)\}$$

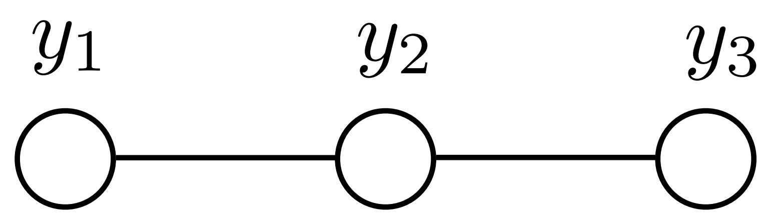
where $\theta_{12}(y_1, y_2) + \theta_{23}(y_2, y_3) = \theta \cdot \phi(y)$

$$\max_y \{\theta \cdot \phi(y)\} =$$



MAP assignment

- Let's start with a simple Markov model over binary variables



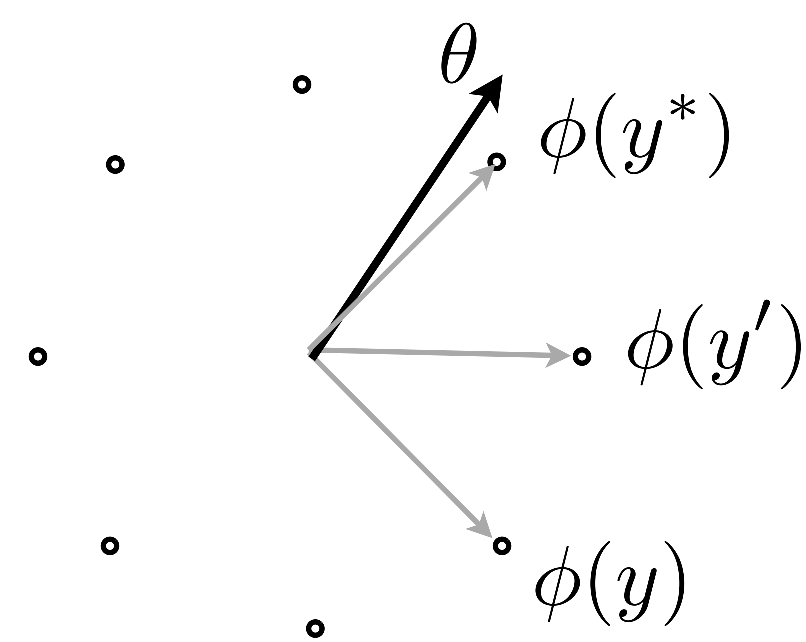
$y_i \in \{0, 1\}$

- The MAP problem we wish to solve is

$$y^* = \arg\max_y \{\theta \cdot \phi(y)\}$$

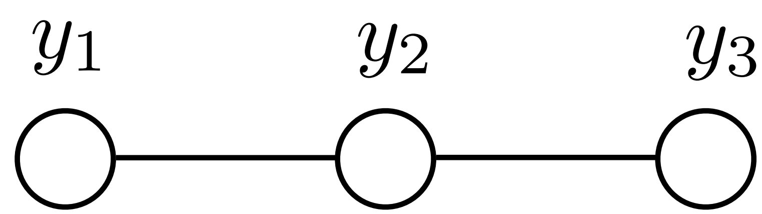
where $\theta_{12}(y_1, y_2) + \theta_{23}(y_2, y_3) = \theta \cdot \phi(y)$

$$\max_y \{\theta \cdot \phi(y)\} = \theta \cdot \phi(y^*)$$



MAP assignment

- Let's start with a simple Markov model over binary variables



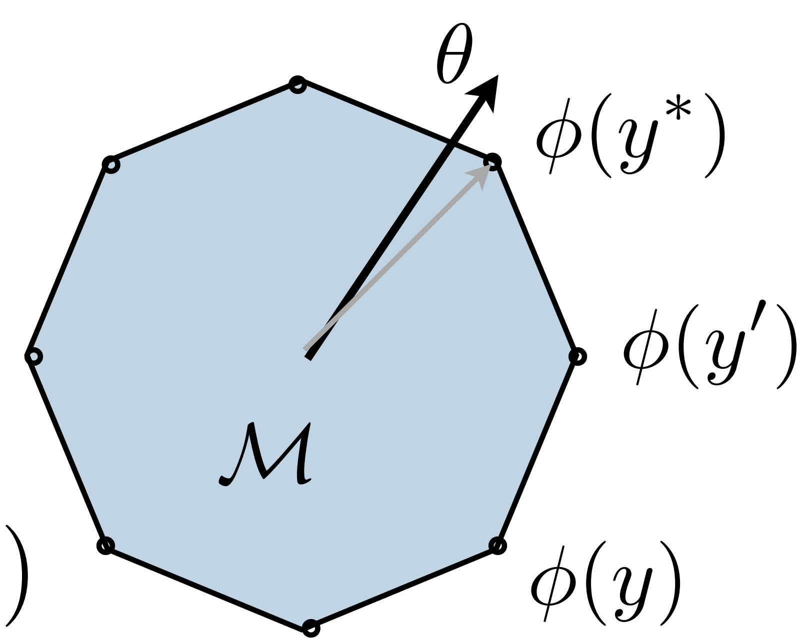
$$y_i \in \{0, 1\}$$

- The MAP problem we wish to solve is

$$y^* = \arg\max_y \{\theta \cdot \phi(y)\}$$

where $\theta_{12}(y_1, y_2) + \theta_{23}(y_2, y_3) = \theta \cdot \phi(y)$

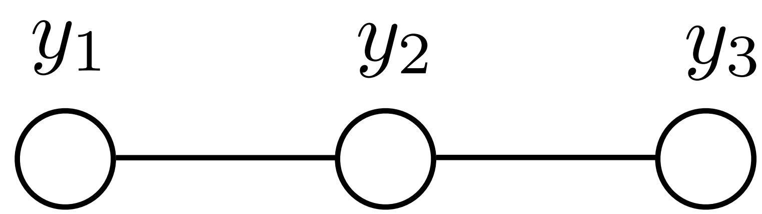
$$\max_y \{\theta \cdot \phi(y)\} = \theta \cdot \phi(y^*)$$



$$\mathcal{M} = \text{conv}\big(\{\phi(y)\}\big)$$

MAP assignment

- Let's start with a simple Markov model over binary variables



$y_i \in \{0, 1\}$

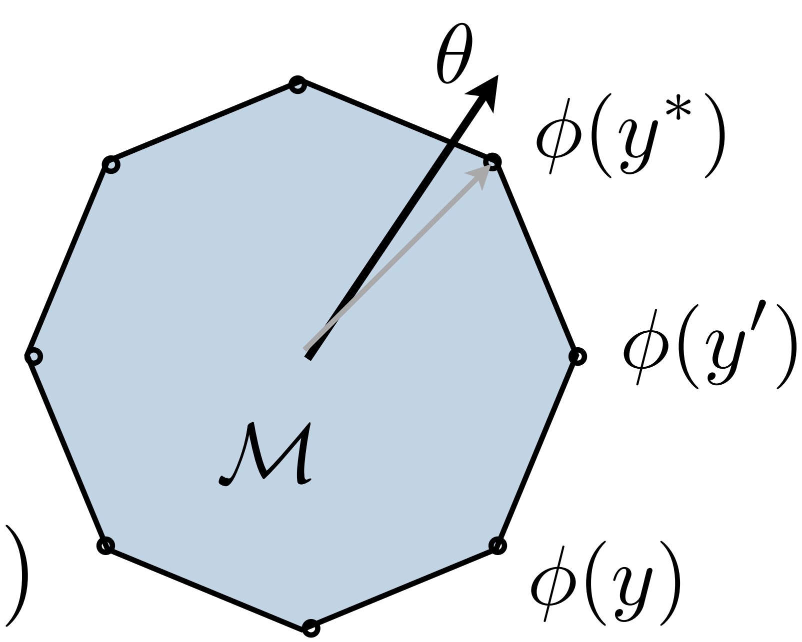
- The MAP problem we wish to solve is

$$y^* = \arg\max_y \left\{ \theta \cdot \phi(y) \right\}$$

where $\quad \theta_{12}(y_1, y_2) + \theta_{23}(y_2, y_3) = \theta \cdot \phi(y)$

$$\max_y \left\{ \theta \cdot \phi(y) \right\} = \theta \cdot \phi(y^*)$$

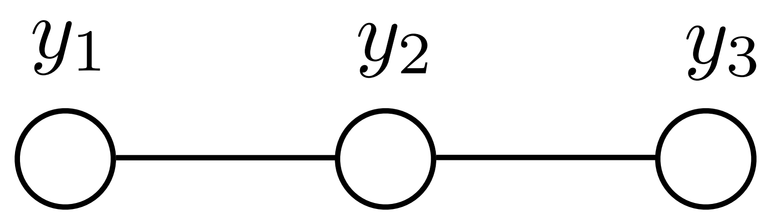
$$= \max_{\mu \in \mathcal{M}} \left\{ \theta \cdot \mu \right\}$$



$$\mathcal{M} = \text{conv}\Big(\{\phi(y)\} \Big)$$

MAP assignment as a LP

- Let's start with a simple Markov model over binary variables



$y_i \in \{0, 1\}$

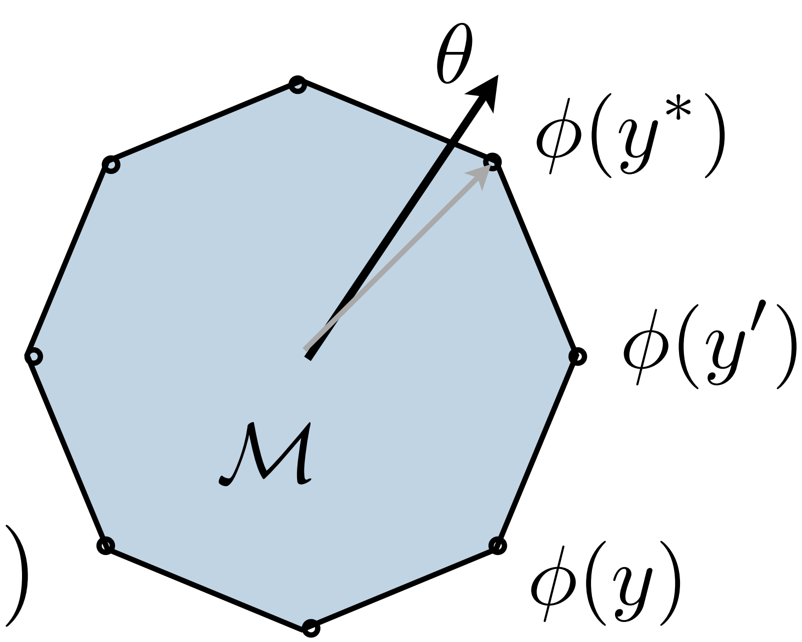
- The MAP problem we wish to solve is

$$y^* = \arg\max_y \left\{\theta \cdot \phi(y)\right\}$$

where $\quad \theta_{12}(y_1, y_2) + \theta_{23}(y_2, y_3) = \theta \cdot \phi(y)$

$$\max_y \left\{\theta \cdot \phi(y)\right\} = \ \theta \cdot \phi(y^*)$$

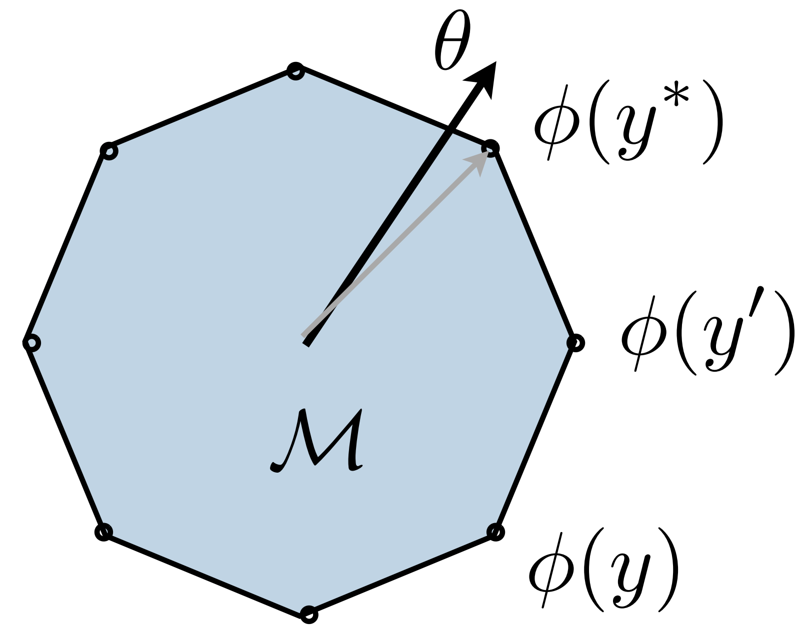
$$= \underbrace{\max_{\mu \in \mathcal{M}}}_{\text{convex set}} \underbrace{\left\{\theta \cdot \mu\right\}}_{\text{linear objective}}$$



$$\mathcal{M} = \operatorname{conv}\big(\,\{\phi(y)\}\,\big)$$

A more algebraic view

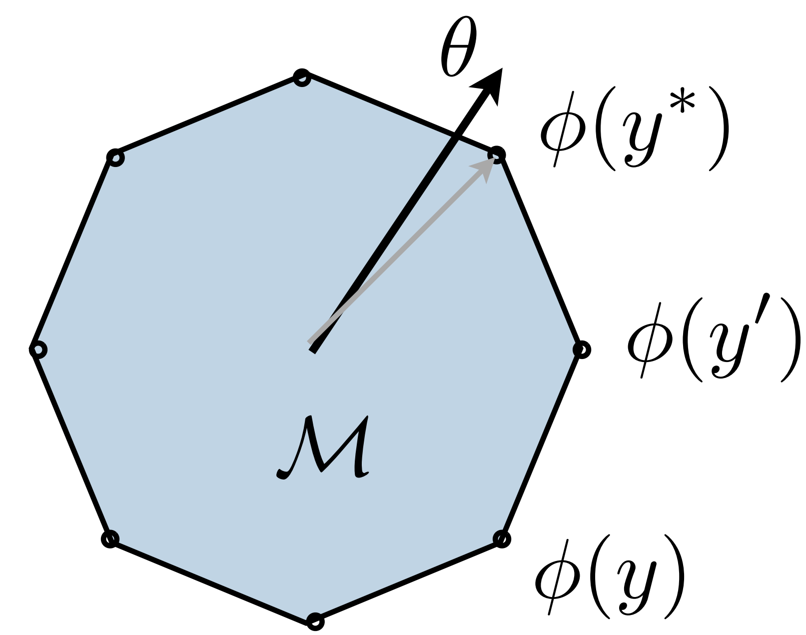
$$\max_{y} \left\{ \theta \cdot \phi(y) \right\} =$$



$$\mathcal{M} = \mathrm{conv}\big(\{\phi(y)\} \big)$$

A more algebraic view

$$\max_y \left\{ \theta \cdot \phi(y) \right\} = \max_p \left\{ \sum_y p(y) \left[\theta \cdot \phi(y) \right] \right\}$$

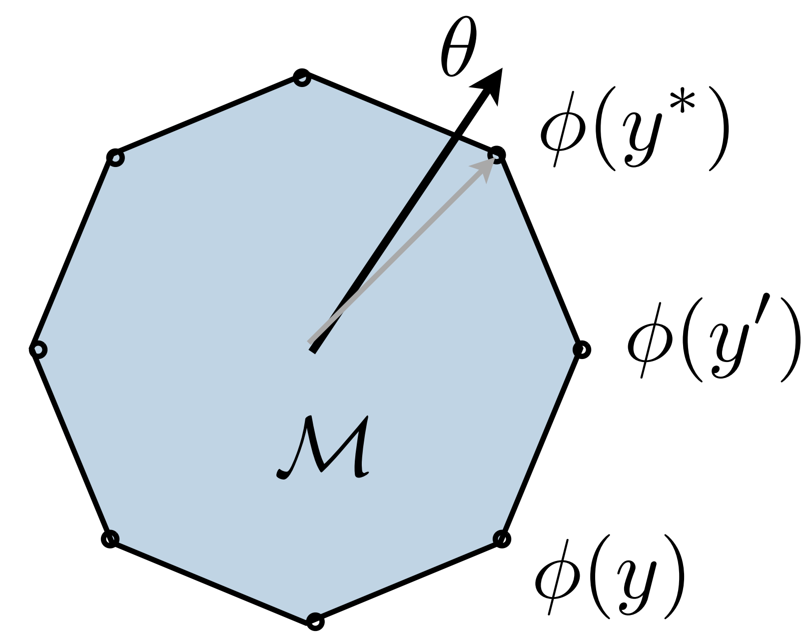


$$\mathcal{M} = \mathrm{conv}\Big(\{\phi(y)\} \Big)$$

A more algebraic view

$$\max_y \{\theta \cdot \phi(y)\} = \max_p \{ \sum_y p(y) \, [\theta \cdot \phi(y)] \}$$

$$= \max_p \{\theta \cdot [\sum_y p(y)\phi(y)]\}$$



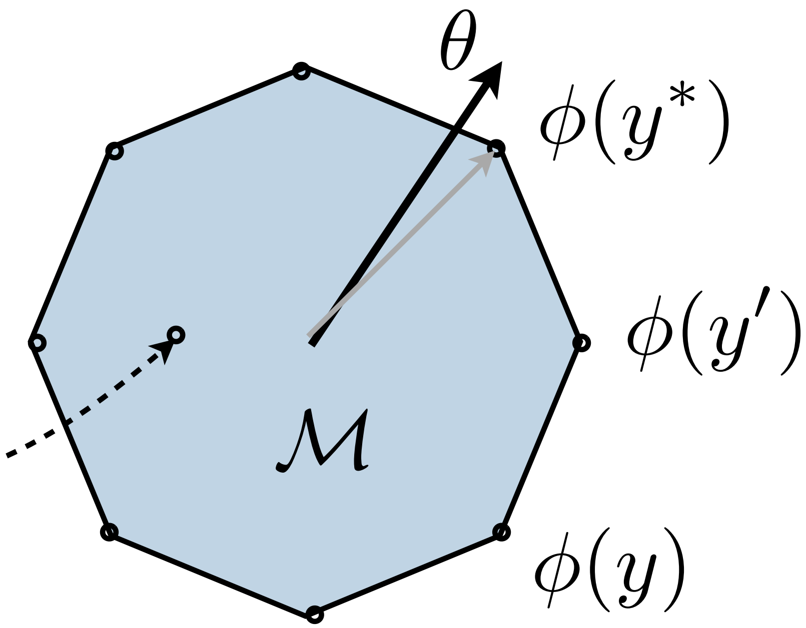
$$\mathcal{M} = \mathrm{conv}\big(\{\phi(y)\} \big)$$

A more algebraic view

$$\max_y \left\{\theta \cdot \phi(y)\right\} = \max_p \left\{ \sum_y p(y) \left[\theta \cdot \phi(y)\right]\right\}$$

$$= \max_p \left\{\theta \cdot \left[\sum_y p(y)\phi(y)\right]\right\}$$

$$= \mu$$



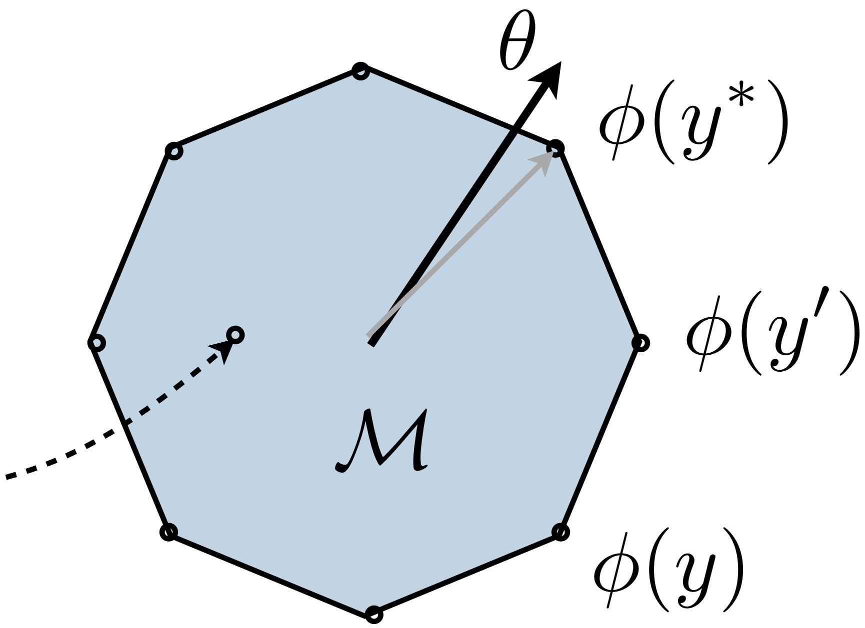
$$\mathcal{M} = \mathrm{conv}\big(\{\phi(y)\}\big)$$

A more algebraic view

$$\max_y \{\theta \cdot \phi(y)\} = \max_p \Big\{ \sum_y p(y)\, [\theta \cdot \phi(y)] \Big\}$$

$$= \max_p \Big\{\theta \cdot \big[\underbrace{\sum_y p(y)\phi(y)}_{=\mu}\big]\Big\}$$

$$= \max_{\mu \in \mathcal{M}} \{\theta \cdot \mu\}$$



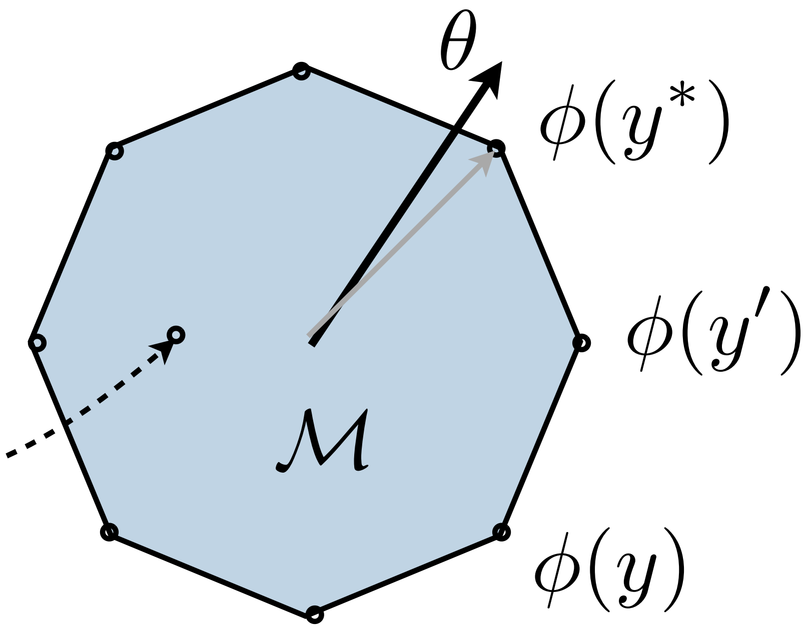
$$\mathcal{M} = \mathrm{conv}\big(\, \{\phi(y)\}\, \big)$$

A more algebraic view

$$\max_y \{\theta \cdot \phi(y)\} = \max_p \{ \sum_y p(y)\, [\theta \cdot \phi(y)] \}$$

$$= \max_p \{\theta \cdot [\underbrace{\sum_y p(y)\phi(y)}_{= \mu}]\}$$

$$= \max_{\mu \in \mathcal{M}} \{\theta \cdot \mu\}$$

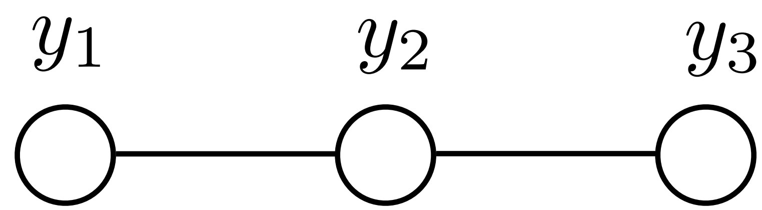


$$\mathcal{M} = \text{conv}\big(\,\{\phi(y)\}\,\big)$$

- What are the expected statistics $\mu = \sum_y p(y)\phi(y)$?
- The convex hull $\mathcal{M}$ is defined on the basis of an exponentially many vertexes $\{\phi(y)\}$. Can we specify it more compactly?

Expected statistics

- Let's go back to our simple MRF



$y_i \in \{0, 1\}$

$$\mu = \sum_y p(y)\phi(y)$$

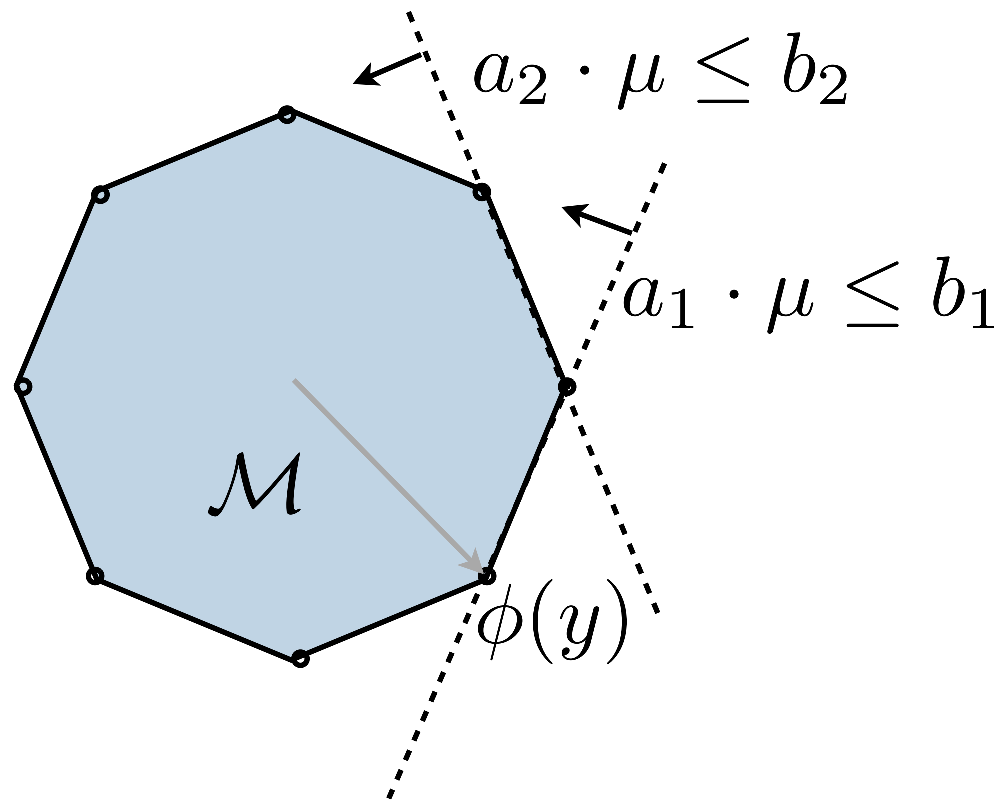
$$= \sum_y p(y) \begin{bmatrix} 1[y_1 = 0 \wedge y_2 = 0] \\ 1[y_1 = 1 \wedge y_2 = 0] \\ 1[y_1 = 0 \wedge y_2 = 1] \\ 1[y_1 = 1 \wedge y_2 = 1] \\ 1[y_2 = 0 \wedge y_3 = 0] \\ 1[y_2 = 1 \wedge y_3 = 0] \\ 1[y_2 = 0 \wedge y_3 = 1] \\ 1[y_2 = 1 \wedge y_3 = 1] \end{bmatrix} = \begin{bmatrix} \mu_{12}(0,0) \\ \mu_{12}(1,0) \\ \mu_{12}(0,1) \\ \mu_{12}(1,1) \\ \mu_{23}(0,0) \\ \mu_{23}(1,0) \\ \mu_{23}(0,1) \\ \mu_{23}(1,1) \end{bmatrix}$$

a vector of marginal probabilities

Marginal polytope

- According to the Weyl's theorem, convex hull of a finite set of vectors (here $\phi(y)$'s) can be written in terms of a set of linear constraints, i.e., as a polytope

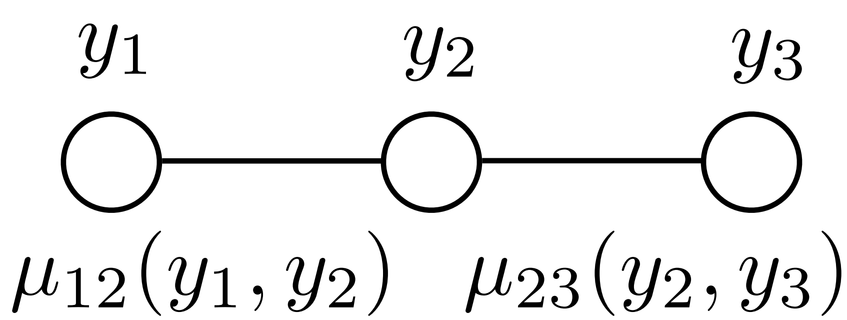
$$\mathcal{M} = \mathrm{conv}\big(\,\{\phi(y)\}\,\big)$$
$$= \{\,\mu : \; A\mu \leq b\,\}$$



- what are the linear constraints?
- how many constraints do we need to specify $\mathcal{M}$?

Linear constraints: examples

- The marginal polytope for tree structured models is defined by simple constraints on the edge marginals

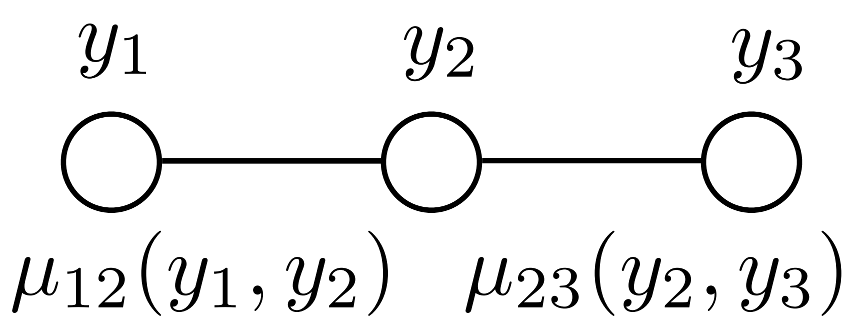


$$y_i \in \{0, 1\}$$

$$\mu = \begin{bmatrix} \mu_{12}(0,0) \\ \mu_{12}(1,0) \\ \mu_{12}(0,1) \\ \mu_{12}(1,1) \\ \mu_{23}(0,0) \\ \mu_{23}(1,0) \\ \mu_{23}(0,1) \\ \mu_{23}(1,1) \end{bmatrix}$$

Linear constraints: examples

- The marginal polytope for tree structured models is defined by simple constraints on the edge marginals



$y_i \in \{0, 1\}$

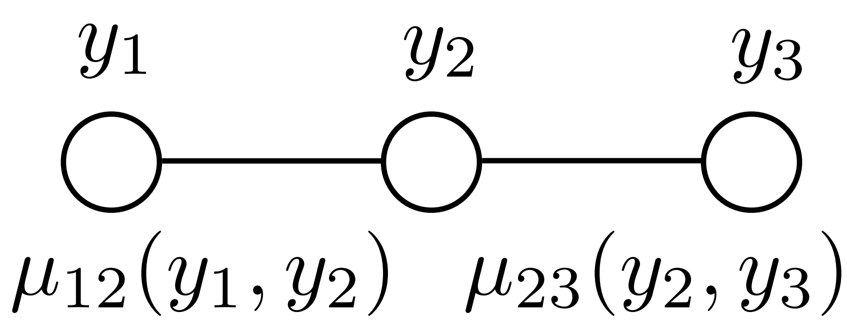
$$\mu = \begin{bmatrix} \mu_{12}(0,0) \\ \mu_{12}(1,0) \\ \mu_{12}(0,1) \\ \mu_{12}(1,1) \\ \mu_{23}(0,0) \\ \mu_{23}(1,0) \\ \mu_{23}(0,1) \\ \mu_{23}(1,1) \end{bmatrix}$$

$$(1) \quad \mu_{ij}(y_i, y_j) \geq 0,$$

non-negative

Linear constraints: examples

- The marginal polytope for tree structured models is defined by simple constraints on the edge marginals



$y_i \in \{0, 1\}$

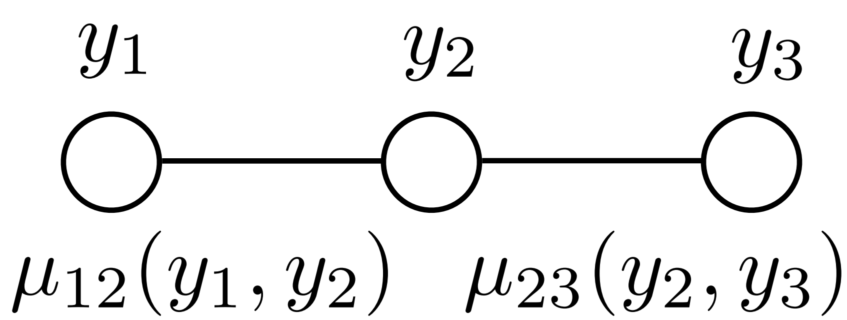
$$\mu = \begin{bmatrix} \mu_{12}(0,0) \\ \mu_{12}(1,0) \\ \mu_{12}(0,1) \\ \mu_{12}(1,1) \\ \mu_{23}(0,0) \\ \mu_{23}(1,0) \\ \mu_{23}(0,1) \\ \mu_{23}(1,1) \end{bmatrix}$$

$(1) \quad \mu_{ij}(y_i, y_j) \geq 0,$ non-negative

$(2) \quad \sum_{y_i, y_j} \mu_{ij}(y_i, y_j) = 1,$ normalized

Linear constraints: examples

- The marginal polytope for tree structured models is defined by simple constraints on the edge marginals



$$y_i \in \{0, 1\}$$

$$\mu = \begin{bmatrix} \mu_{12}(0,0) \\ \mu_{12}(1,0) \\ \mu_{12}(0,1) \\ \mu_{12}(1,1) \\ \mu_{23}(0,0) \\ \mu_{23}(1,0) \\ \mu_{23}(0,1) \\ \mu_{23}(1,1) \end{bmatrix}$$

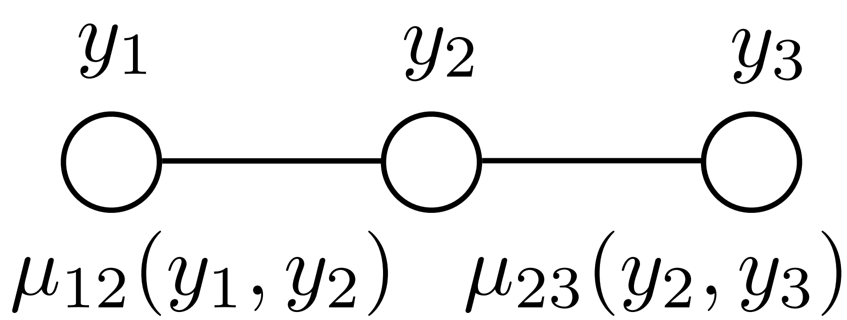
$$(1) \quad \mu_{ij}(y_i, y_j) \geq 0,$$ non-negative

$$(2) \quad \textstyle\sum_{y_i, y_j} \mu_{ij}(y_i, y_j) = 1,$$ normalized

$$(3) \quad \textstyle\sum_{y_i} \mu_{ij}(y_i, y_j) = \sum_{y_k} \mu_{jk}(y_j, y_k)$$ consistent

Linear constraints: examples

- The marginal polytope for tree structured models is defined by simple constraints on the edge marginals



$y_i \in \{0, 1\}$

$$\mu = \begin{bmatrix} \mu_{12}(0,0) \\ \mu_{12}(1,0) \\ \mu_{12}(0,1) \\ \mu_{12}(1,1) \\ \mu_{23}(0,0) \\ \mu_{23}(1,0) \\ \mu_{23}(0,1) \\ \mu_{23}(1,1) \end{bmatrix}$$

$$\mu = \{\mu_{ij}(y_i, y_j)\} \in \mathcal{M}_L \text{ iff}$$

$$(1) \quad \mu_{ij}(y_i, y_j) \geq 0,$$ non-negative

$$(2) \quad \sum_{y_i, y_j} \mu_{ij}(y_i, y_j) = 1,$$ normalized

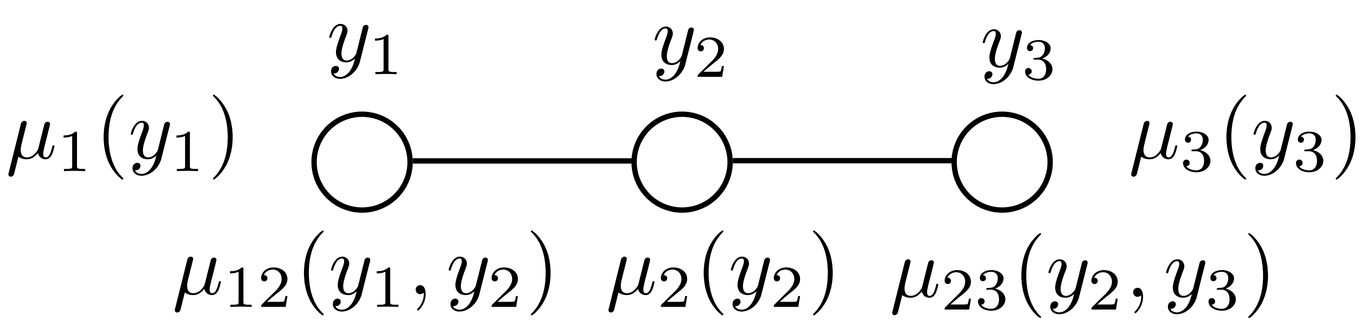
$$(3) \quad \sum_{y_i} \mu_{ij}(y_i, y_j) = \sum_{y_k} \mu_{jk}(y_j, y_k)$$ consistent

"Local marginal polytope", "pairwise relaxation"

- We will prove shortly that, indeed, for trees $\mathcal{M}_L = \mathcal{M}$

Linear constraints: examples

- It is often convenient to add extra single node statistics



$y_i \in \{0, 1\}$

$$\mu = \{\mu_i(y_i),\ \mu_{ij}(y_i, y_j)\} \in \mathcal{M}_L \ \text{ iff}$$

$$(1) \quad \mu_{ij}(y_i, y_j) \geq 0, \qquad \text{non-negative}$$

$$(2) \quad \sum_{y_i} \mu_i(y_i) = 1, \qquad \text{normalized}$$

$$(3) \quad \begin{aligned} \sum_{y_i} \mu_{ij}(y_i, y_j) &= \mu_j(y_j) \\ \sum_{y_j} \mu_{ij}(y_i, y_j) &= \mu_i(y_i) \end{aligned} \qquad \text{consistent}$$

$$\mu = \begin{bmatrix} \mu_1(0) \\ \mu_1(1) \\ \mu_2(0) \\ \mu_2(1) \\ \mu_3(0) \\ \mu_3(1) \\ \mu_{12}(0,0) \\ \mu_{12}(1,0) \\ \mu_{12}(0,1) \\ \mu_{12}(1,1) \\ \mu_{23}(0,0) \\ \mu_{23}(1,0) \\ \mu_{23}(0,1) \\ \mu_{23}(1,1) \end{bmatrix}$$

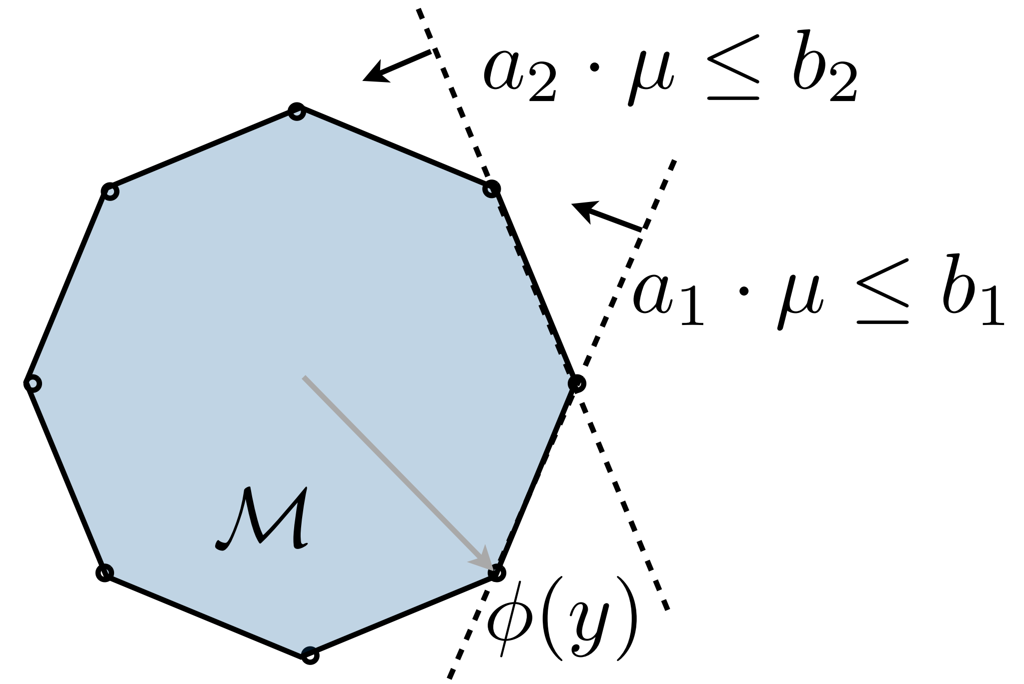
(same polytope, embedded in higher dimensions)

“Proof” for trees

- Consider tree structured models where the expected statistics include single node marginals (for simplicity)

$$\mu = \{\mu_i(y_i),\ \mu_{ij}(y_i, y_j)\} \in \mathcal{M}_L \ \text{ iff}$$

$$\begin{array}{ll} (1) & \mu_{ij}(y_i, y_j) \geq 0, \\ (2) & \sum_{y_i} \mu_i(y_i) = 1, \\ (3) & \sum_{y_i} \mu_{ij}(y_i, y_j) = \mu_j(y_j) \\ & \sum_{y_j} \mu_{ij}(y_i, y_j) = \mu_i(y_i) \end{array}$$



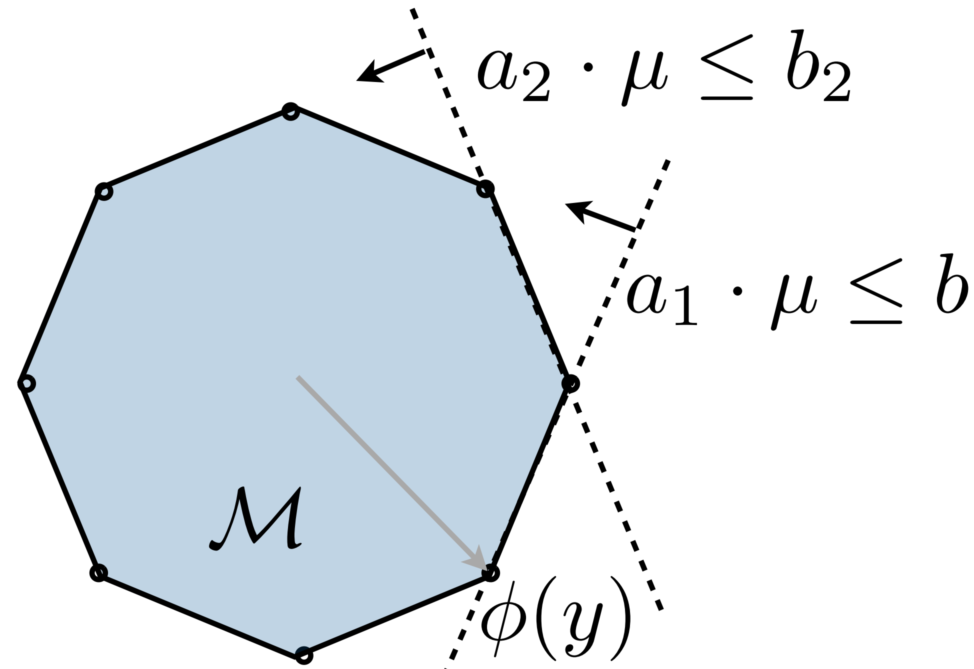
- clearly any $\mu \in \mathcal{M}$ also belongs to $\mathcal{M}_L$ $\quad \mathcal{M} \subseteq \mathcal{M}_L$

"Proof" for trees

- Consider tree structured models where the expected statistics include single node marginals (for simplicity)

$\mu = \{\mu_i(y_i),\ \mu_{ij}(y_i, y_j)\} \in \mathcal{M}_L$ iff

$$\begin{array}{ll} (1) & \mu_{ij}(y_i, y_j) \geq 0, \\ (2) & \sum_{y_i} \mu_i(y_i) = 1, \\ (3) & \sum_{y_i} \mu_{ij}(y_i, y_j) = \mu_j(y_j) \\ & \sum_{y_j} \mu_{ij}(y_i, y_j) = \mu_i(y_i) \end{array}$$



- clearly any $\mu \in \mathcal{M}$ also belongs to $\mathcal{M}_L$ $\mathcal{M} \subseteq \mathcal{M}_L$
- any $\mu \in \mathcal{M}_L$ gives rise to a tree distribution

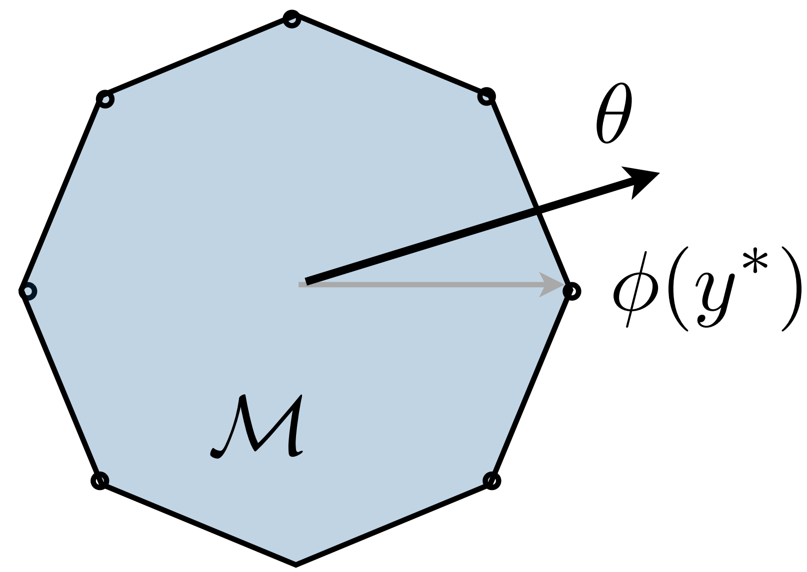
$$P(y) = \prod_i \mu_i(y_i) \prod_{(i,j) \in E} \frac{\mu_{ij}(y_i, y_j)}{\mu_i(y_i)\mu_j(y_j)}$$

such that $\sum_y P(y)\phi(y) = \mu \in \mathcal{M}$ $\mathcal{M}_L \subseteq \mathcal{M}$

Excluding vertexes (for trees)

- For trees, pairwise relaxation suffices to define $\mathcal{M}$

$$
\begin{aligned}
&(1) \quad \mu_{ij}(y_i, y_j) \geq 0, \\
&(2) \quad \textstyle\sum_{y_i} \mu_i(y_i) = 1, \\
&(3) \quad \textstyle\sum_{y_i} \mu_{ij}(y_i, y_j) = \mu_j(y_j) \\
&\qquad\;\; \textstyle\sum_{y_j} \mu_{ij}(y_i, y_j) = \mu_i(y_i)
\end{aligned}
$$

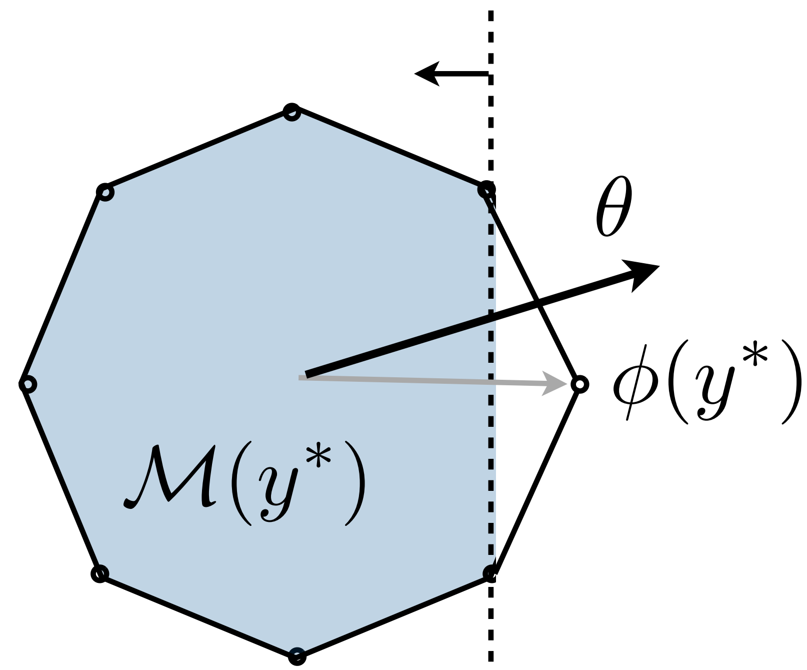


- What if we needed to exclude a specific (e.g., the maximizing) vertex?

Excluding vertexes (for trees)

- For trees, pairwise relaxation suffices to define $\mathcal{M}$

$$
\begin{aligned}
&(1) \quad \mu_{ij}(y_i, y_j) \geq 0, \\
&(2) \quad \textstyle\sum_{y_i} \mu_i(y_i) = 1, \\
&(3) \quad \textstyle\sum_{y_i} \mu_{ij}(y_i, y_j) = \mu_j(y_j) \\
&\qquad\;\; \textstyle\sum_{y_j} \mu_{ij}(y_i, y_j) = \mu_i(y_i)
\end{aligned}
$$

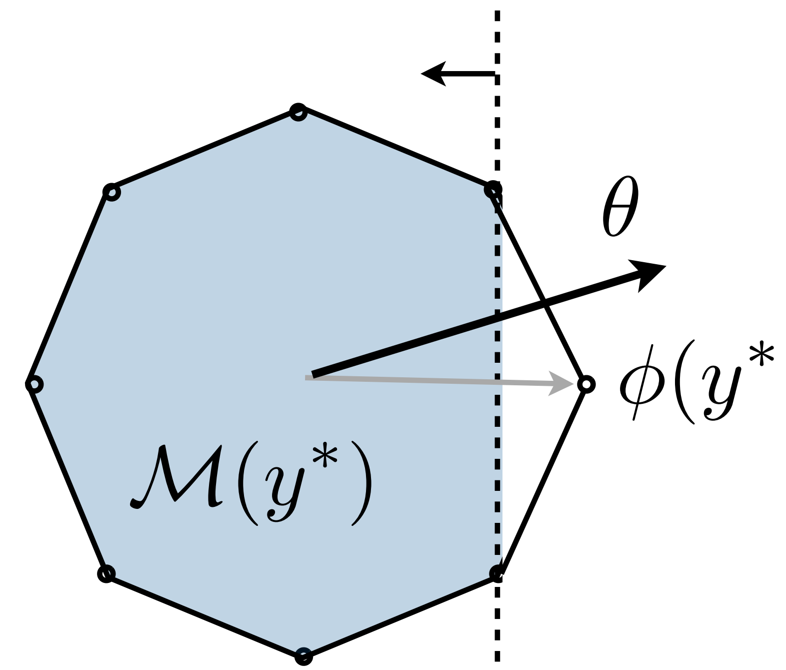


- What if we needed to exclude a specific (e.g., the maximizing) vertex?

Excluding vertexes (for trees)

- For trees, pairwise relaxation suffices to define $\mathcal{M}$

$$
\begin{array}{ll}
(1) & \mu_{ij}(y_i, y_j) \geq 0, \\
(2) & \sum_{y_i} \mu_i(y_i) = 1, \\
(3) & \sum_{y_i} \mu_{ij}(y_i, y_j) = \mu_j(y_j) \\
 & \sum_{y_j} \mu_{ij}(y_i, y_j) = \mu_i(y_i)
\end{array}
$$



- What if we needed to exclude a specific (e.g., the maximizing) vertex?

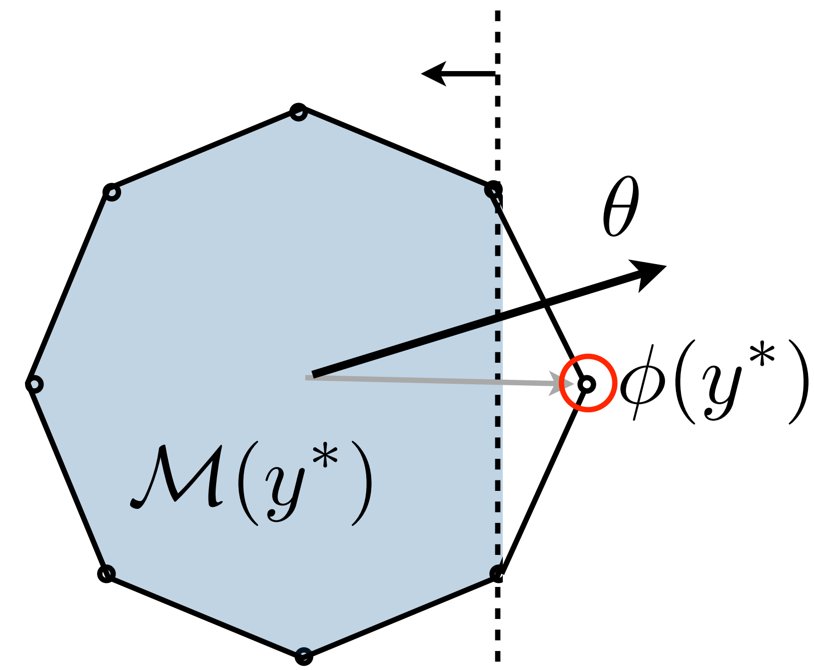
$$(4) \quad \sum_i (1 - d_i)\mu_i(y_i^*) + \sum_{(i,j)\in E} \mu_{ij}(y_i^*, y_j^*) \leq 0$$

(Fromer et al. '09)

Excluding vertexes (for trees)

- For trees, pairwise relaxation suffices to define $\mathcal{M}$

$$
\begin{aligned}
&(1) \quad \mu_{ij}(y_i, y_j) \geq 0, \\
&(2) \quad \textstyle\sum_{y_i} \mu_i(y_i) = 1, \\
&(3) \quad \textstyle\sum_{y_i} \mu_{ij}(y_i, y_j) = \mu_j(y_j) \\
&\qquad\;\; \textstyle\sum_{y_j} \mu_{ij}(y_i, y_j) = \mu_i(y_i)
\end{aligned}
$$



- What if we needed to exclude a specific (e.g., the maximizing) vertex?

$$(4) \quad \underbrace{\sum_i (1-d_i)\underbrace{\mu_i(y_i^*)}_{=1} + \sum_{(i,j)\in E} \underbrace{\mu_{ij}(y_i^*, y_j^*)}_{=1}}_{=1} \leq 0$$

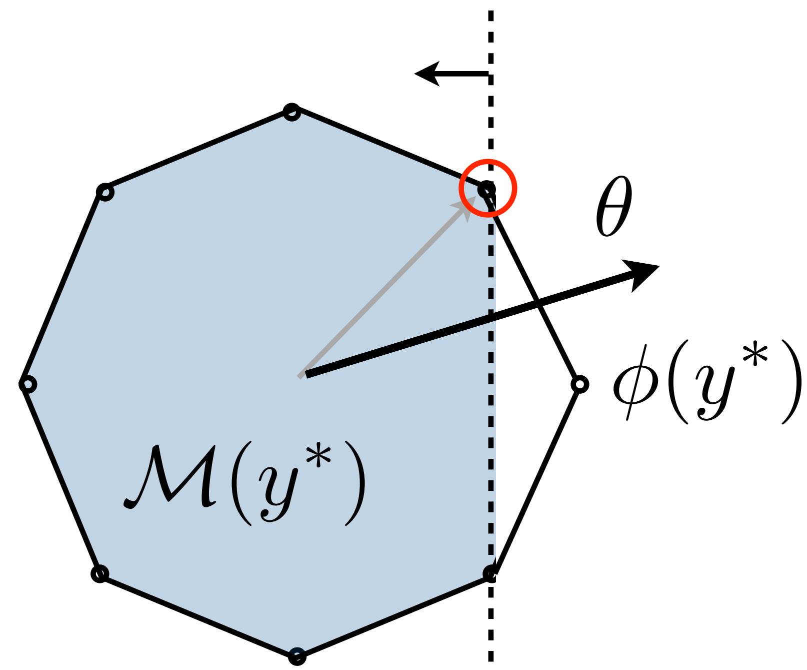
(Fromer et al. '09)

if $\mu = \phi(y^*)$ $\quad\Rightarrow$ violated!

Excluding vertexes (for trees)

- For trees, pairwise relaxation suffices to define $\mathcal{M}$

$$
\begin{array}{ll}
(1) & \mu_{ij}(y_i, y_j) \geq 0, \\
(2) & \sum_{y_i} \mu_i(y_i) = 1, \\
(3) & \sum_{y_i} \mu_{ij}(y_i, y_j) = \mu_j(y_j) \\
 & \sum_{y_j} \mu_{ij}(y_i, y_j) = \mu_i(y_i)
\end{array}
$$



- What if we needed to exclude a specific (e.g., the maximizing) vertex?

$$(4) \quad \underbrace{\sum_i (1 - d_i)\mu_i(y_i^*) + \sum_{(i,j)\in E} \mu_{ij}(y_i^*, y_j^*)}_{=0} \leq 0$$

(Fromer et al. '09)

if $\mu =$ neighbor of $\phi(y^*)$ $\quad = 0 \quad \Rightarrow$ tight!

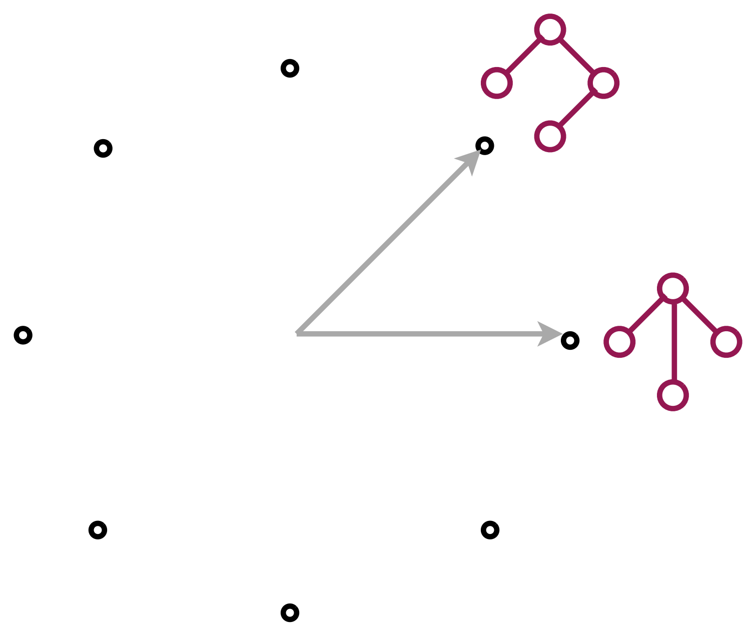
The spanning tree polytope

- Suppose our MAP problem involves finding the maximum weight spanning tree

$$y_{ij} = 1 \ \text{ if edge } (i,j) \text{ is selected and zero otherwise}$$

$$\theta_T(y) = \begin{cases} 0, & \text{if } y \text{ specifies a spanning tree} \\ -\infty, & \text{otherwise} \end{cases}$$

$$\max_y \Big\{ \sum_{(i,j)} y_{ij}\theta_{ij} + \theta_T(y) \Big\}$$



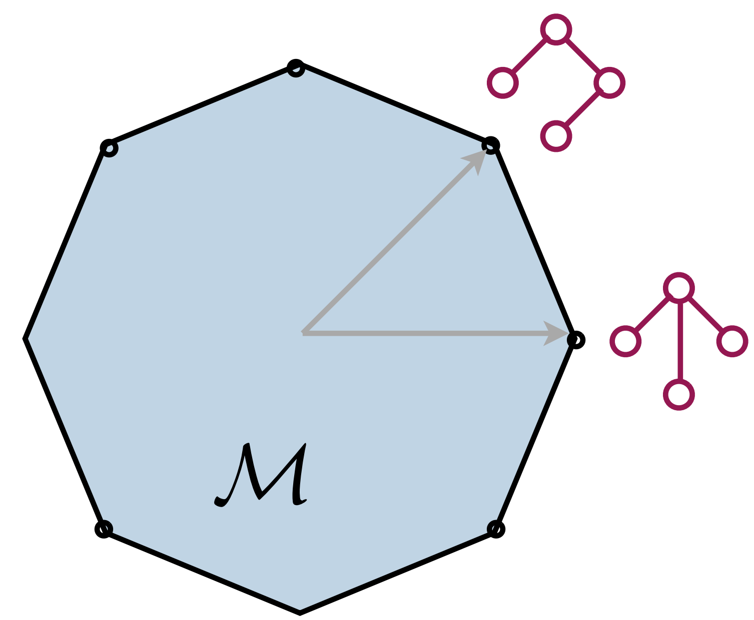
The spanning tree polytope

- Suppose our MAP problem involves finding the maximum weight spanning tree

$$y_{ij} = 1 \text{ if edge } (i,j) \text{ is selected and zero otherwise}$$

$$\theta_T(y) = \begin{cases} 0, & \text{if } y \text{ specifies a spanning tree} \\ -\infty, & \text{otherwise} \end{cases}$$

$$\max_y \Big\{ \sum_{(i,j)} y_{ij}\theta_{ij} + \theta_T(y) \Big\} = \max_{\mu \in \mathcal{M}} \{\theta \cdot \mu\}$$



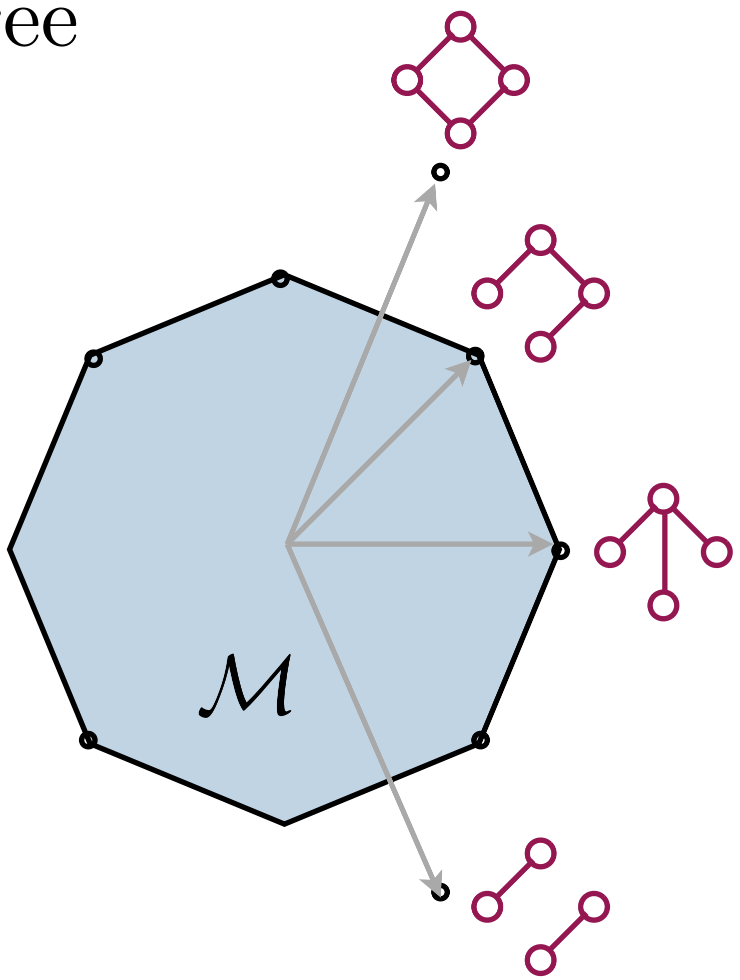
The spanning tree polytope

- Suppose our MAP problem involves finding the maximum weight spanning tree

$$y_{ij} = 1 \text{ if edge } (i,j) \text{ is selected and zero otherwise}$$

$$\theta_T(y) = \begin{cases} 0, & \text{if } y \text{ specifies a spanning tree} \\ -\infty, & \text{otherwise} \end{cases}$$

$$\max_y \Big\{ \sum_{(i,j)} y_{ij}\theta_{ij} + \theta_T(y) \Big\} = \max_{\mu \in \mathcal{M}} \{\theta \cdot \mu\}$$



The spanning tree polytope

- Suppose our MAP problem involves finding the maximum weight spanning tree

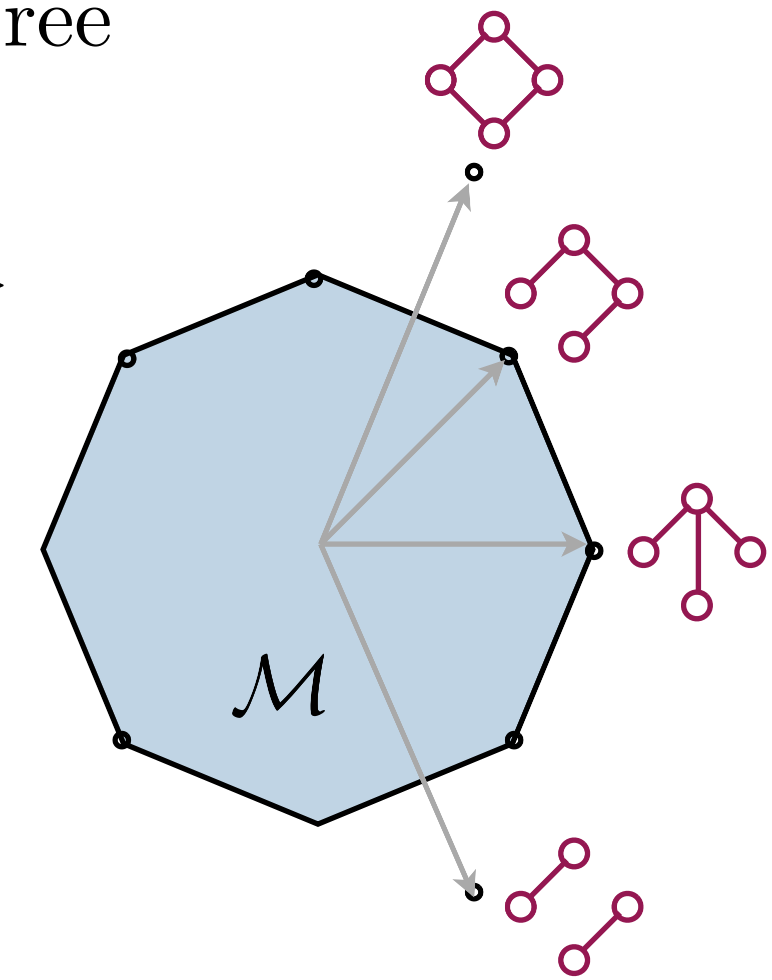
$$y_{ij} = 1 \text{ if edge } (i,j) \text{ is selected and zero otherwise}$$

$$\theta_T(y) = \begin{cases} 0, & \text{if } y \text{ specifies a spanning tree} \\ -\infty, & \text{otherwise} \end{cases}$$

$$\max_y \Big\{ \sum_{(i,j)} y_{ij}\theta_{ij} + \theta_T(y) \Big\} = \max_{\mu \in \mathcal{M}} \{\theta \cdot \mu\}$$

$\mu \in \mathcal{M}$ iff

$$\begin{array}{ll} a) & \mu_{ij} \geq 0 \\ b) & \sum_{(i,j)} \mu_{ij} = n - 1 \\ c) & \sum_{(i,j) \in E(S)} \mu_{ij} \leq |S| - 1, \ \forall S \subset V \end{array}$$



The spanning tree polytope

- Suppose our MAP problem involves finding the maximum weight spanning tree

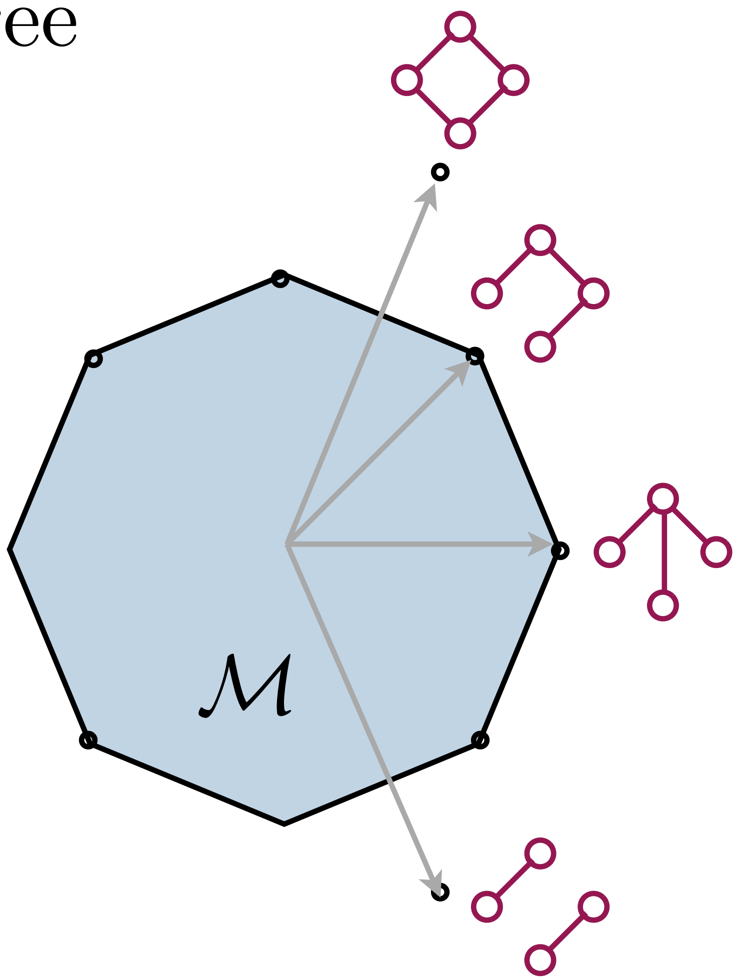
$$y_{ij} = 1 \text{ if edge } (i,j) \text{ is selected and zero otherwise}$$

$$\theta_T(y) = \begin{cases} 0, & \text{if } y \text{ specifies a spanning tree} \\ -\infty, & \text{otherwise} \end{cases}$$

$$\max_y \Big\{ \sum_{(i,j)} y_{ij}\theta_{ij} + \theta_T(y) \Big\} = \max_{\mu \in \mathcal{M}} \{\theta \cdot \mu\}$$

$\mu \in \mathcal{M}$ iff

$$\begin{array}{ll} a) & \mu_{ij} \geq 0 \\ b) & \sum_{(i,j)} \mu_{ij} = n-1 \\ c) & \sum_{(i,j) \in E(S)} \mu_{ij} \leq |S| - 1, \ \forall S \subset V \end{array}$$



- There are exponentially many constraints!

The spanning tree polytope

- Suppose our MAP problem involves finding the maximum weight spanning tree

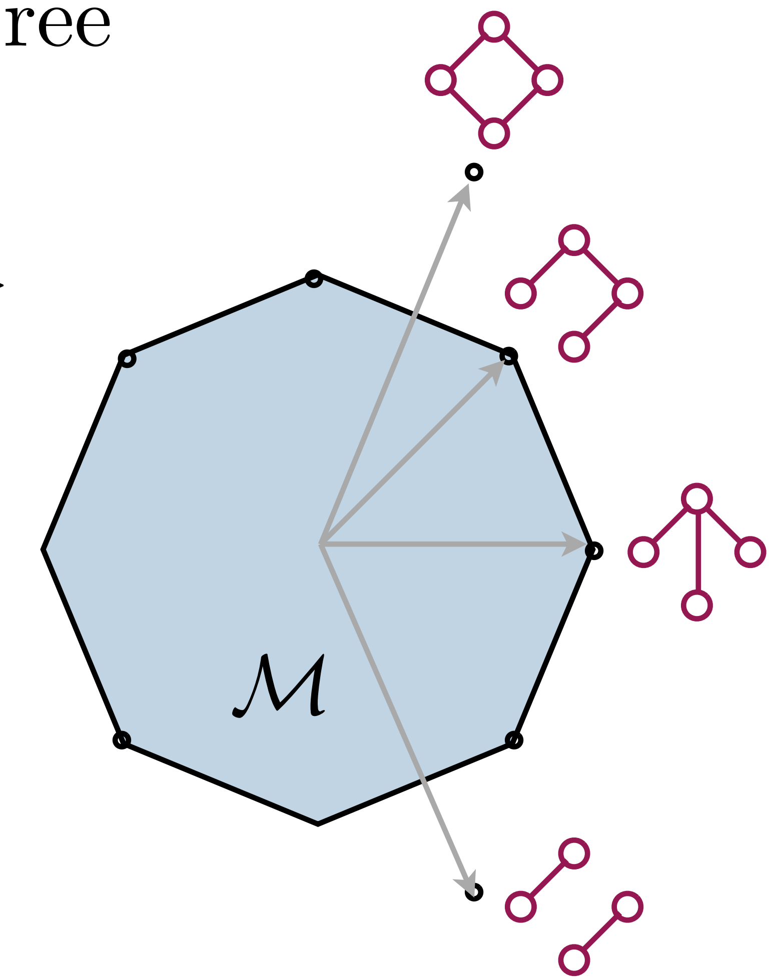
$$y_{ij} = 1 \text{ if edge } (i,j) \text{ is selected and zero otherwise}$$

$$\theta_T(y) = \begin{cases} 0, & \text{if } y \text{ specifies a spanning tree} \\ -\infty, & \text{otherwise} \end{cases}$$

$$\max_y \Big\{ \sum_{(i,j)} y_{ij}\theta_{ij} + \theta_T(y) \Big\} = \max_{\mu \in \mathcal{M}} \{\theta \cdot \mu\}$$

$\mu \in \mathcal{M}$ iff

$$\begin{array}{ll} a) & \mu_{ij} \geq 0 \\ b) & \sum_{(i,j)} \mu_{ij} = n-1 \\ c) & \sum_{(i,j)\in E(S)} \mu_{ij} \leq |S| - 1, \ \forall S \subset V \end{array}$$



- There are exponentially many constraints!
- Polynomial description exists via lift & project (e.g., Yannakakis '90)

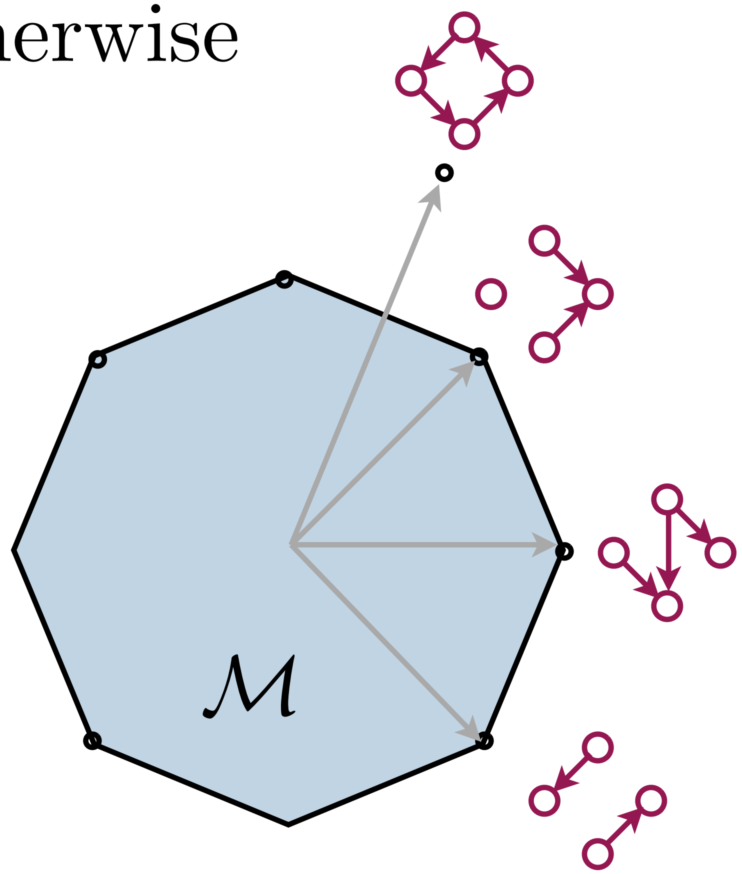
The acyclic subgraph polytope

- The problem is considerably harder if we consider arc selections that form acyclic graphs (cf. structure learning)

$$y_{ij} = 1 \text{ if arc } i \to j \text{ is selected and zero otherwise}$$

$$\theta_{DAG}(y) = \begin{cases} 0, & \text{if } y \text{ specifies a DAG} \\ -\infty, & \text{otherwise} \end{cases}$$

$$\max_y \left\{ \sum_{i,j} y_{ij}\theta_{ij} + \theta_{DAG}(y) \right\} = \max_{\mu \in \mathcal{M}} \{\theta \cdot \mu\}$$



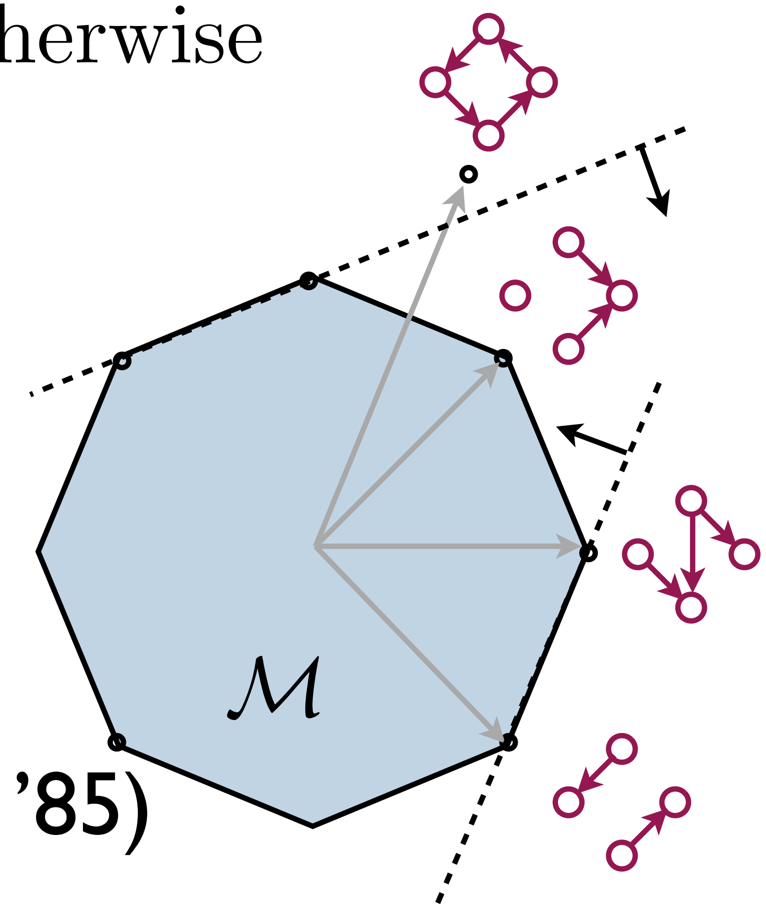
The acyclic subgraph polytope

- The problem is considerably harder if we consider arc selections that form acyclic graphs (cf. structure learning)

$$y_{ij} = 1 \ \text{ if arc } i \to j \text{ is selected and zero otherwise}$$

$$\theta_{DAG}(y) = \begin{cases} 0, & \text{if } y \text{ specifies a DAG} \\ -\infty, & \text{otherwise} \end{cases}$$

$$\max_y \Big\{ \sum_{i,j} y_{ij}\theta_{ij} + \theta_{DAG}(y) \Big\} = \max_{\mu \in \mathcal{M}} \{\theta \cdot \mu\}$$

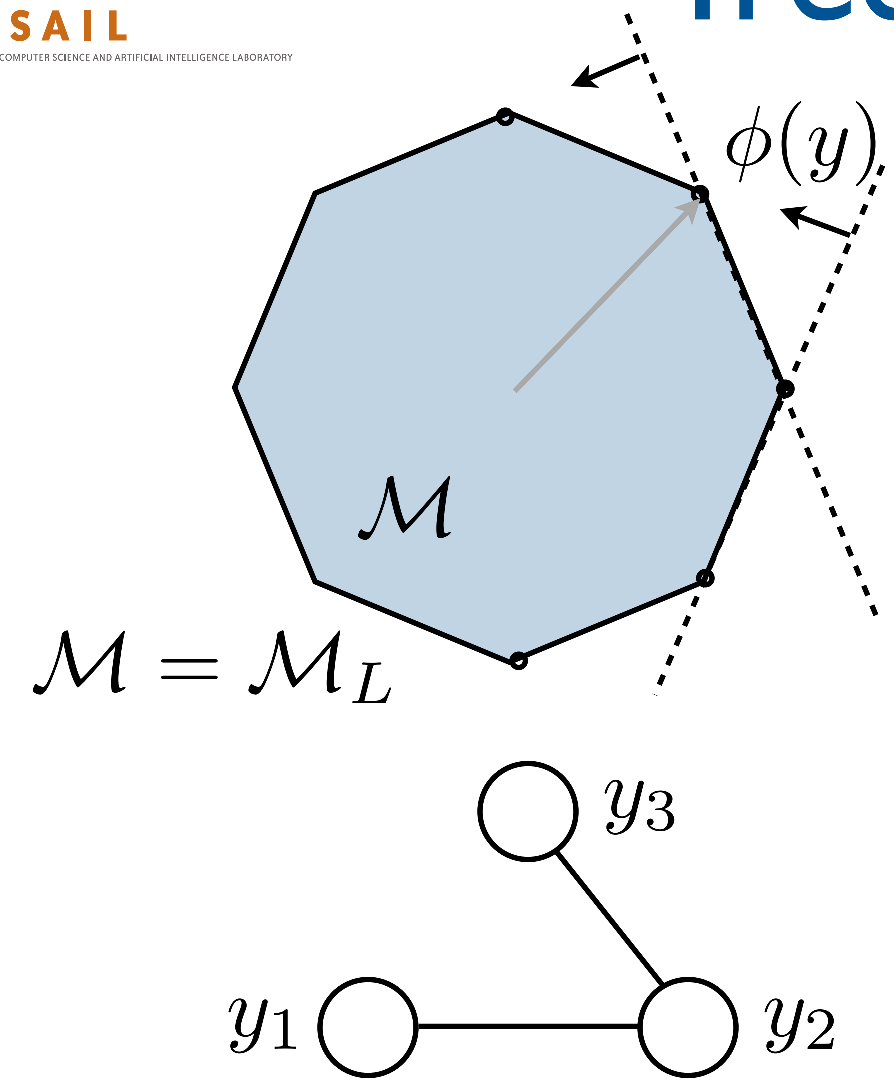


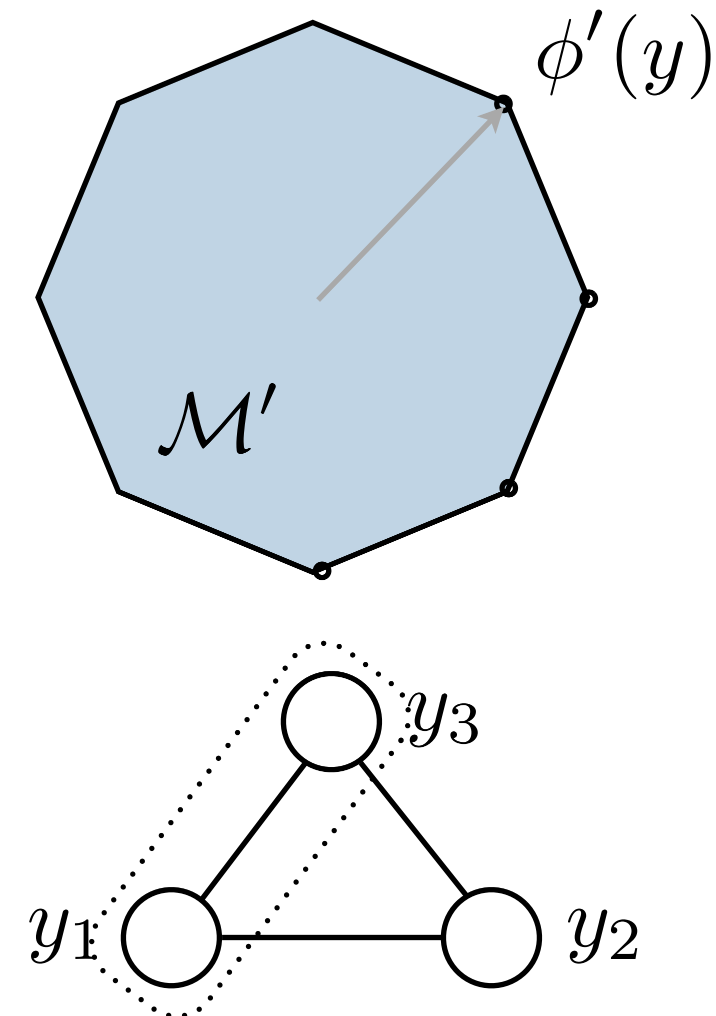
- Many but not all of the facet defining inequalities are known (e.g., Grötschel et al. '85)

$$\mu_{ij} \geq 0, \quad \sum_{i \to j \in C} \mu_{ij} \leq |C| - 1, \ \ \forall \text{ cycles } C$$

"cycle inequalities"

Trees vs. non-trees

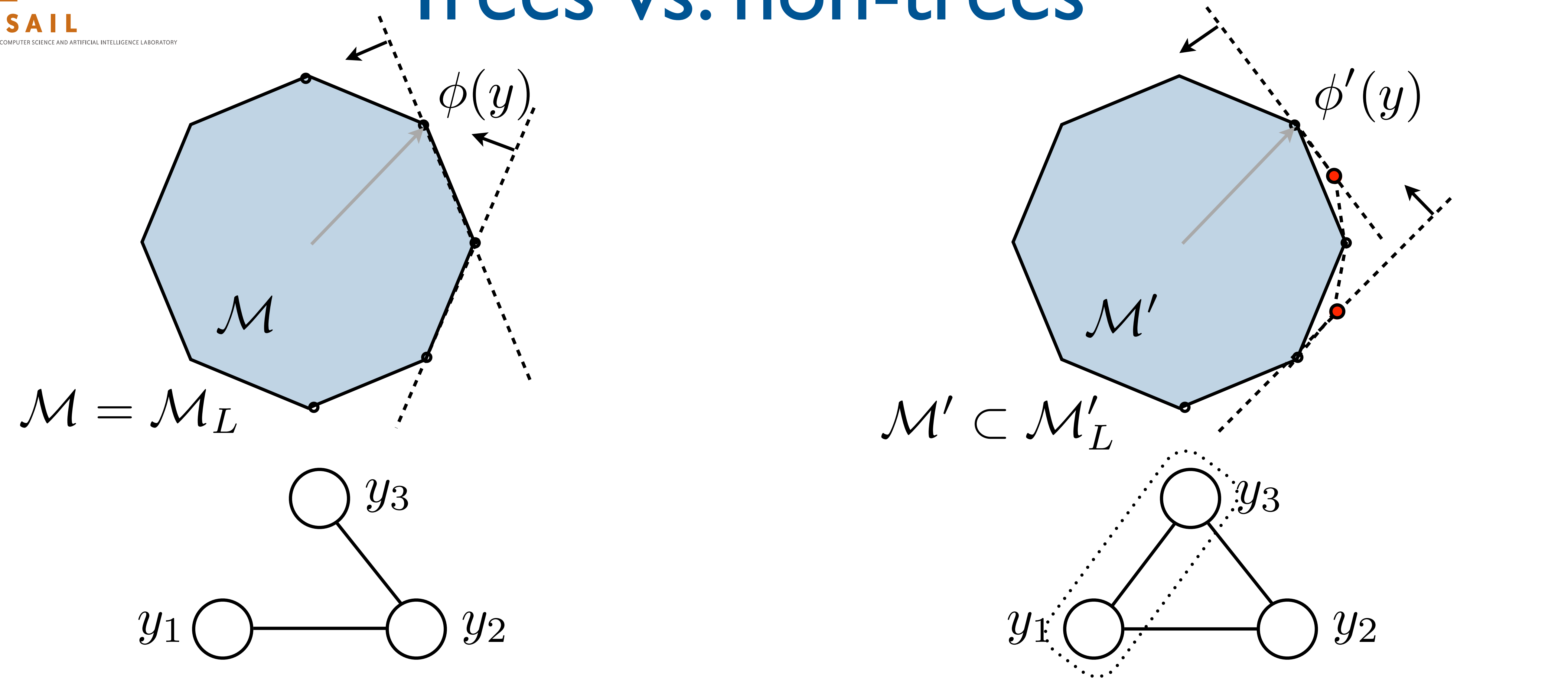




$$\mu = \{\mu_i(y_i),\ \mu_{ij}(y_i, y_j)\} \in \mathcal{M}_L \ \text{ iff}$$

$$
\begin{array}{ll}
(1) & \mu_{ij}(y_i, y_j) \geq 0, \\
(2) & \sum_{y_i} \mu_i(y_i) = 1, \\
(3) & \sum_{y_i} \mu_{ij}(y_i, y_j) = \mu_j(y_j) \\
 & \sum_{y_j} \mu_{ij}(y_i, y_j) = \mu_i(y_i)
\end{array}
$$

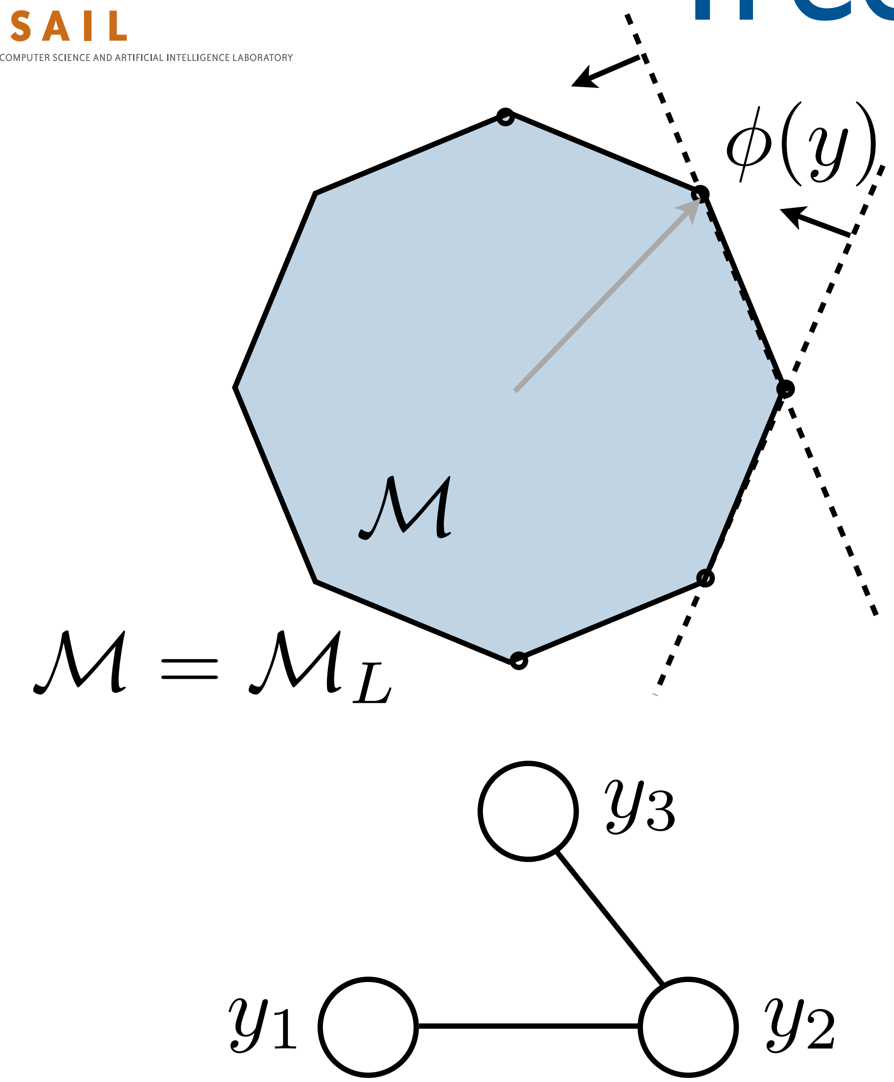
Trees vs. non-trees

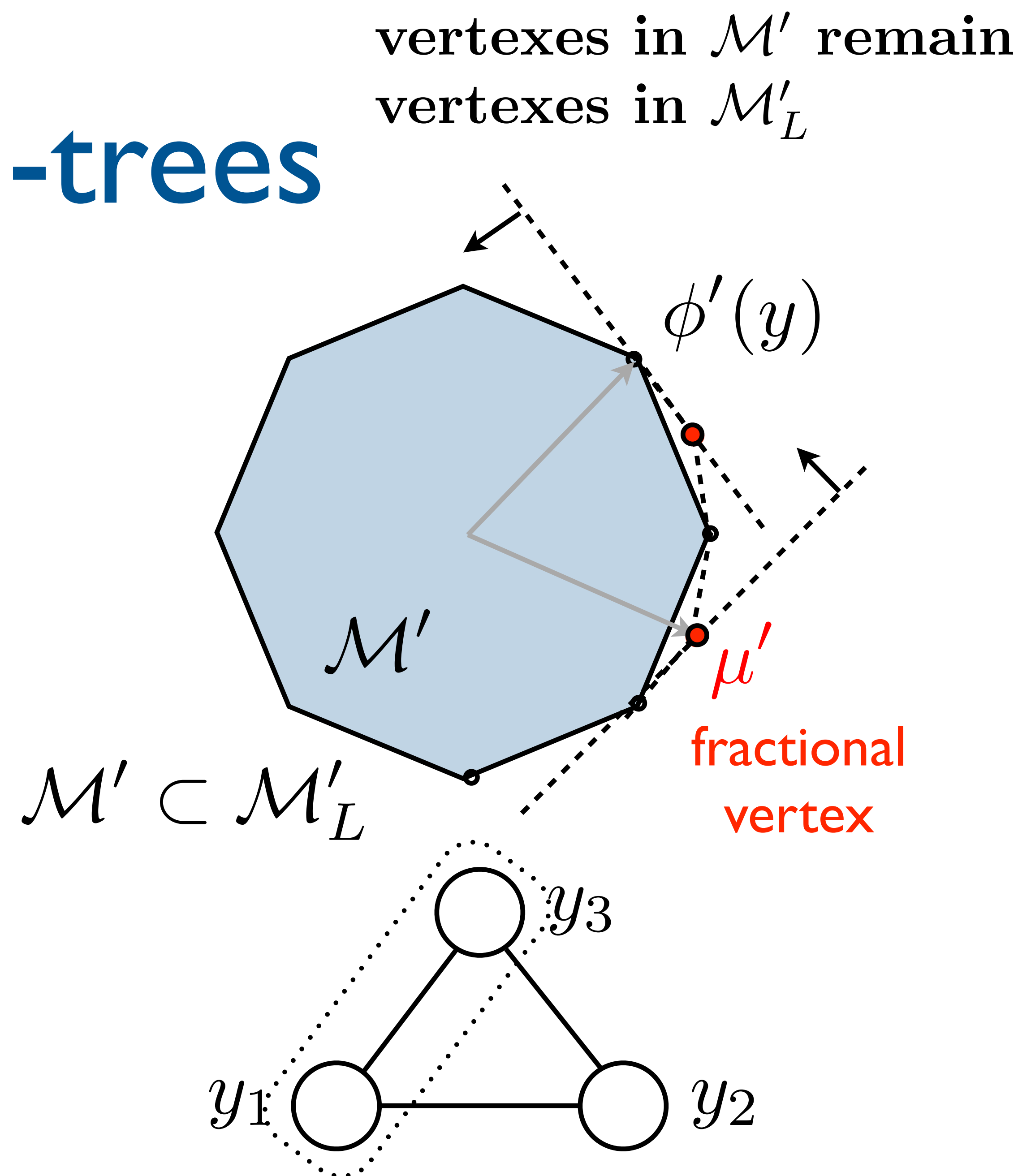


$$\mu = \{\mu_i(y_i),\ \mu_{ij}(y_i, y_j)\} \in \mathcal{M}_L \ \text{ iff}$$

$$\begin{array}{ll}
(1) & \mu_{ij}(y_i, y_j) \geq 0, \\
(2) & \sum_{y_i} \mu_i(y_i) = 1, \\
(3) & \sum_{y_i} \mu_{ij}(y_i, y_j) = \mu_j(y_j) \\
 & \sum_{y_j} \mu_{ij}(y_i, y_j) = \mu_i(y_i)
\end{array}$$

Trees vs. non-trees

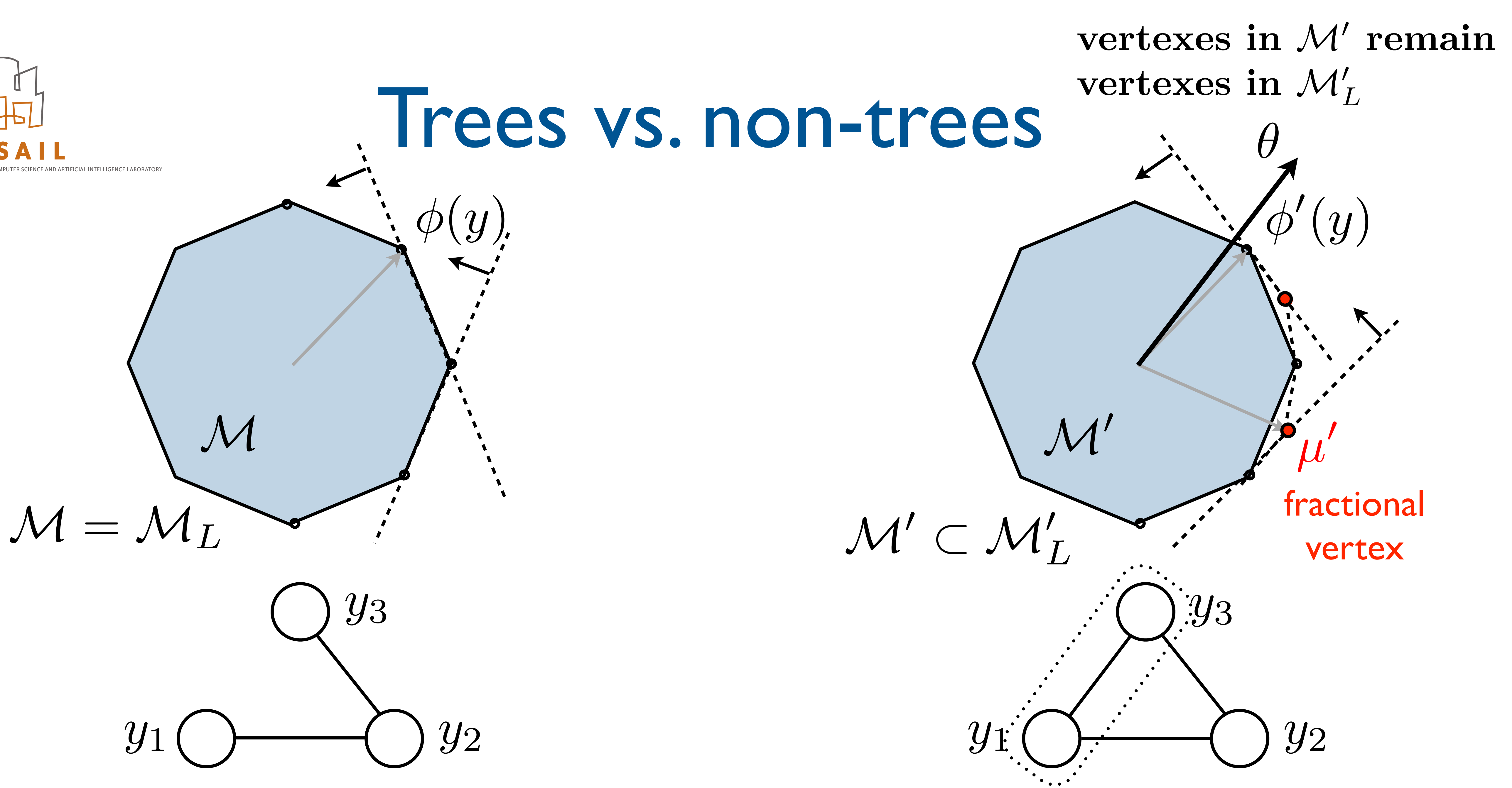




$$\mu = \{\mu_i(y_i),\ \mu_{ij}(y_i, y_j)\} \in \mathcal{M}_L \ \text{ iff}$$

$$\begin{aligned}
&(1) \quad \mu_{ij}(y_i, y_j) \geq 0, \\
&(2) \quad \textstyle\sum_{y_i} \mu_i(y_i) = 1, \\
&(3) \quad \textstyle\sum_{y_i} \mu_{ij}(y_i, y_j) = \mu_j(y_j) \\
&\qquad\ \ \textstyle\sum_{y_j} \mu_{ij}(y_i, y_j) = \mu_i(y_i)
\end{aligned}$$

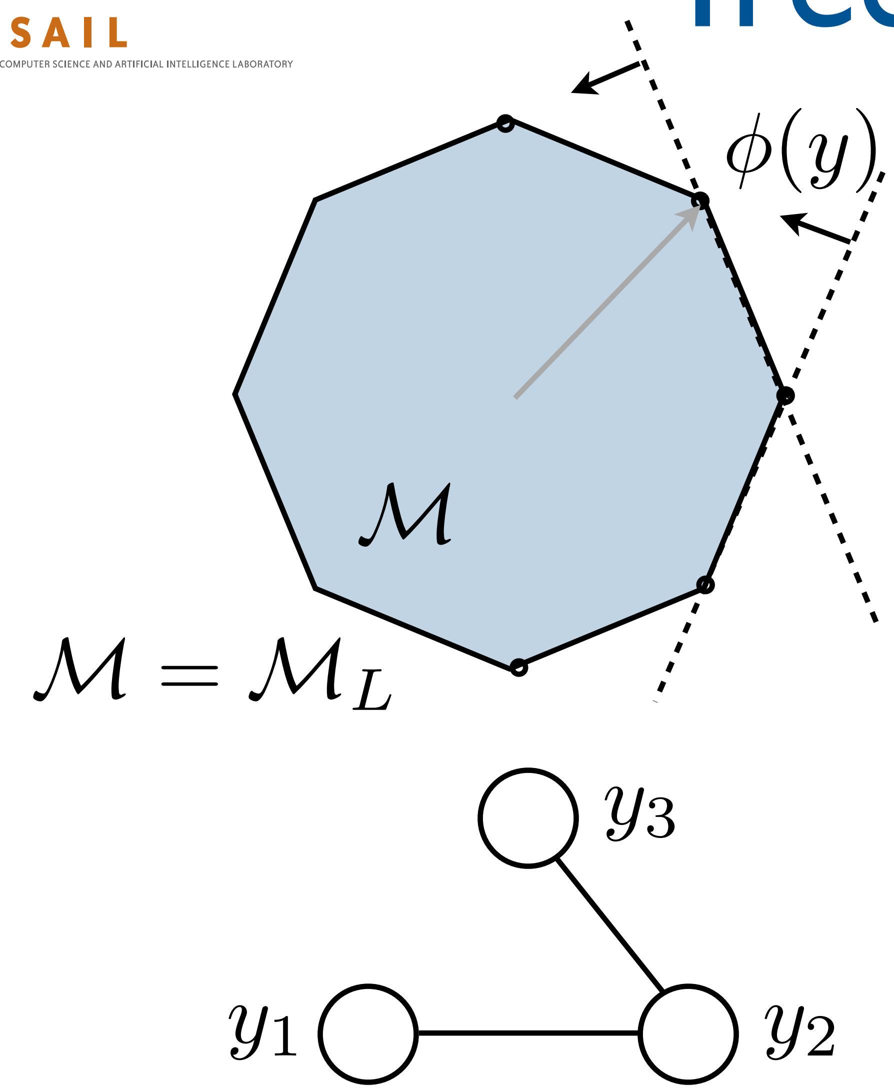
Trees vs. non-trees

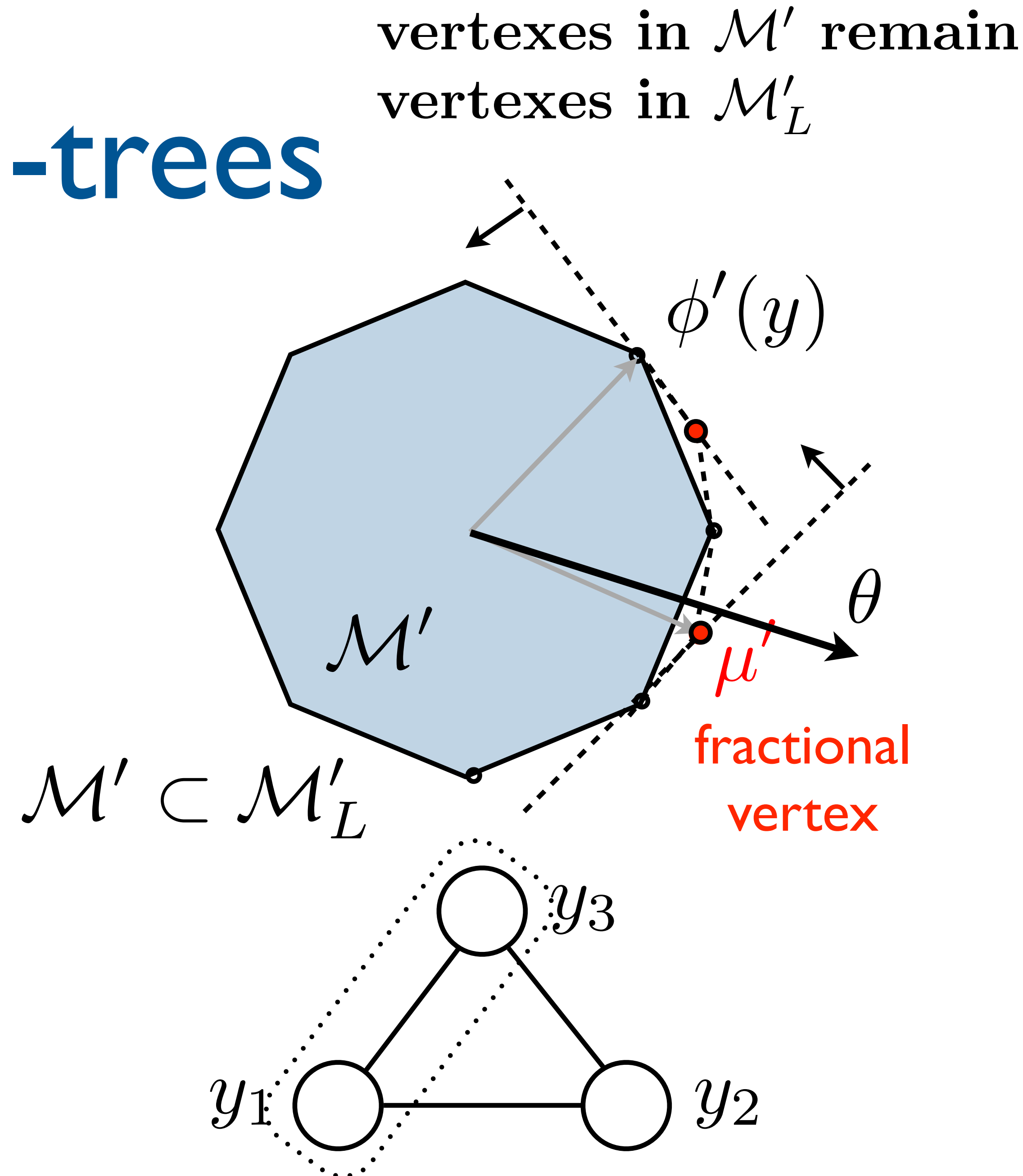


$$\mu = \{\mu_i(y_i),\ \mu_{ij}(y_i, y_j)\} \in \mathcal{M}_L \ \text{ iff}$$

$$
\begin{aligned}
&(1) \quad \mu_{ij}(y_i, y_j) \geq 0, \\
&(2) \quad \textstyle\sum_{y_i} \mu_i(y_i) = 1, \\
&(3) \quad \textstyle\sum_{y_i} \mu_{ij}(y_i, y_j) = \mu_j(y_j) \\
&\qquad\ \ \textstyle\sum_{y_j} \mu_{ij}(y_i, y_j) = \mu_i(y_i)
\end{aligned}
$$

Trees vs. non-trees





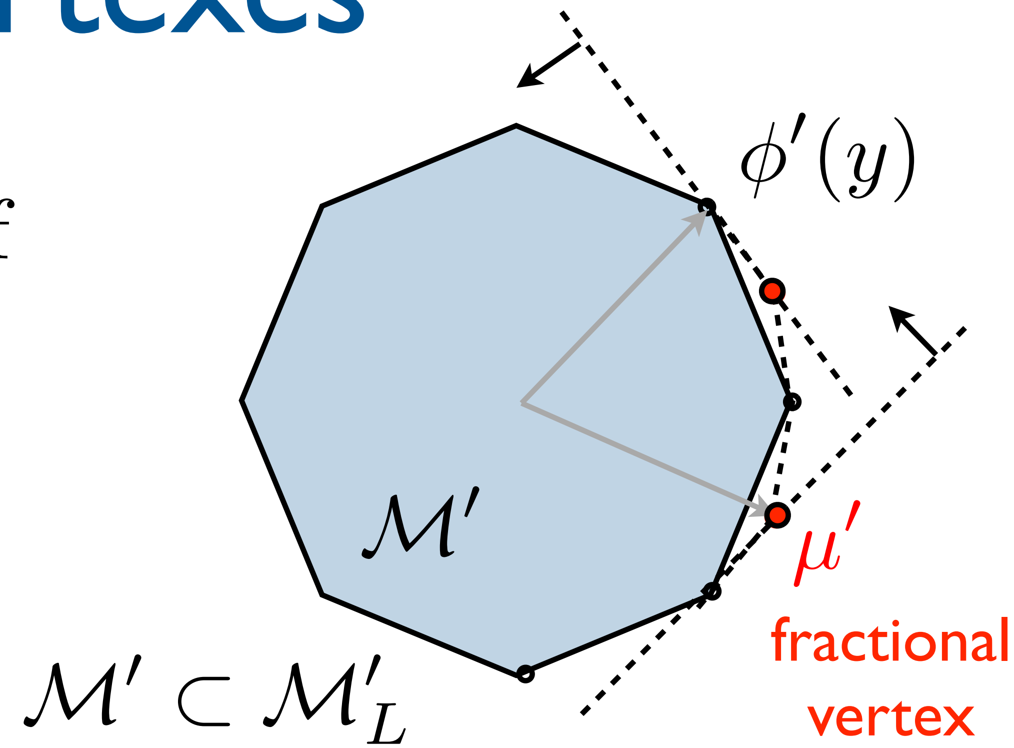
$$\mu = \{\mu_i(y_i),\ \mu_{ij}(y_i, y_j)\} \in \mathcal{M}_L \ \text{ iff}$$

$$
\begin{aligned}
&(1) \quad \mu_{ij}(y_i, y_j) \geq 0, \\
&(2) \quad \textstyle\sum_{y_i} \mu_i(y_i) = 1, \\
&(3) \quad \textstyle\sum_{y_i} \mu_{ij}(y_i, y_j) = \mu_j(y_j) \\
&\qquad\ \textstyle\sum_{y_j} \mu_{ij}(y_i, y_j) = \mu_i(y_i)
\end{aligned}
$$

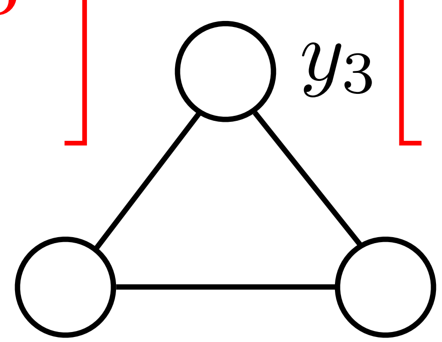
Fractional vertexes

$$\mu = \{\mu_i(y_i),\, \mu_{ij}(y_i, y_j)\} \in \mathcal{M}_L \ \text{ iff}$$

$$\begin{aligned}
&(1) \quad \mu_{ij}(y_i, y_j) \geq 0,\\
&(2) \quad \textstyle\sum_{y_i} \mu_i(y_i) = 1,\\
&(3) \quad \textstyle\sum_{y_i} \mu_{ij}(y_i, y_j) = \mu_j(y_j)\\
&\qquad\ \ \textstyle\sum_{y_j} \mu_{ij}(y_i, y_j) = \mu_i(y_i)
\end{aligned}$$



$\mu' =$



y_1: $\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$, y_2: $\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$, y_3: $\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$

(y_1, y_3): $\begin{bmatrix} 0 & 0.5 \\ 0.5 & 0 \end{bmatrix}$ "disagreement form"

(y_2, y_3): $\begin{bmatrix} 0 & 0.5 \\ 0.5 & 0 \end{bmatrix}$ "disagreement form"

(y_1, y_2): $\begin{bmatrix} 0 & 0.5 \\ 0.5 & 0 \end{bmatrix}$ "disagreement form"

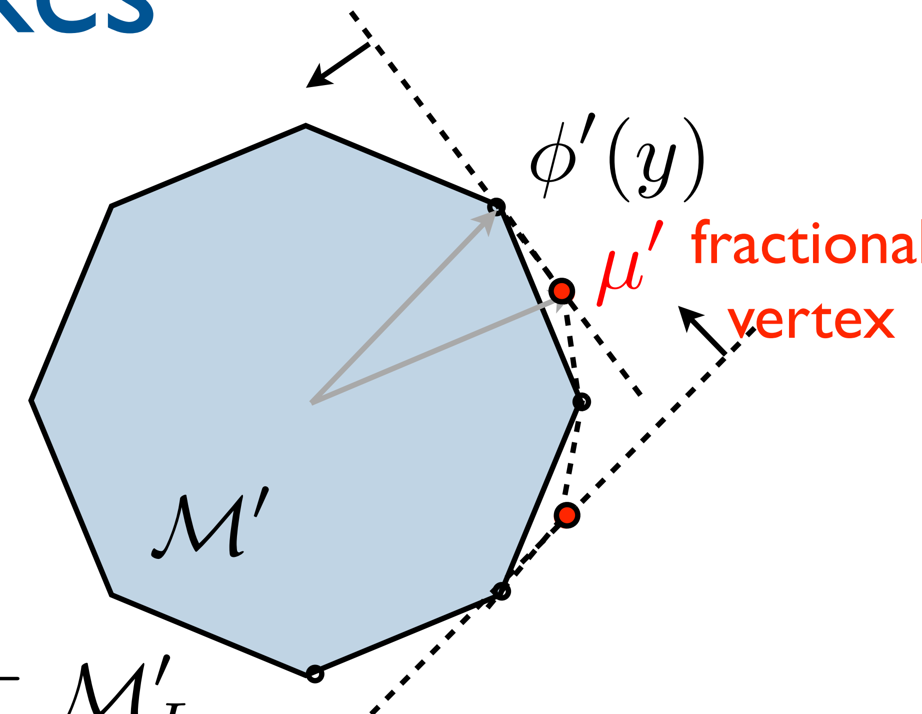
these satisfy (1)-(3) but no distribution can generate these marginals

(cf. Deza & Laurent '96)

Fractional vertexes

$$\mu = \{\mu_i(y_i),\, \mu_{ij}(y_i, y_j)\} \in \mathcal{M}_L \;\text{ iff}$$

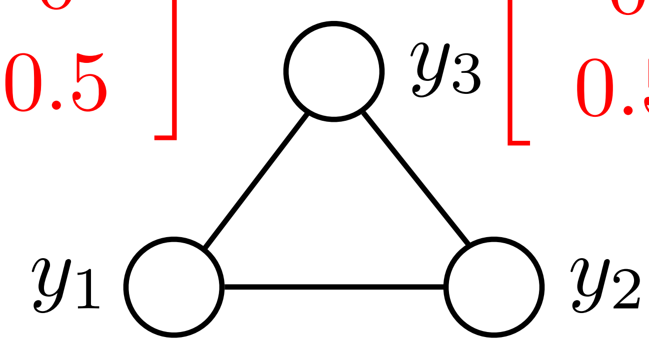
$$\begin{aligned}
&(1) \quad \mu_{ij}(y_i, y_j) \geq 0, \\
&(2) \quad \textstyle\sum_{y_i} \mu_i(y_i) = 1, \\
&(3) \quad \textstyle\sum_{y_i} \mu_{ij}(y_i, y_j) = \mu_j(y_j) \\
&\qquad\;\; \textstyle\sum_{y_j} \mu_{ij}(y_i, y_j) = \mu_i(y_i)
\end{aligned}$$



$\mathcal{M}' \subset \mathcal{M}'_L$

"agreement form" $\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$ "disagreement form"

$\mu' =$ $\begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$ $\begin{bmatrix} 0 & 0.5 \\ 0.5 & 0 \end{bmatrix}$



$\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$ $\begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$ $\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$

"agreement form"

these satisfy (1)-(3) but no distribution can generate these marginals

Persistency

- **Theorem** in pairwise MRF models, any partially integral solution resulting from pairwise relaxation can be extended to an optimal solution provided that

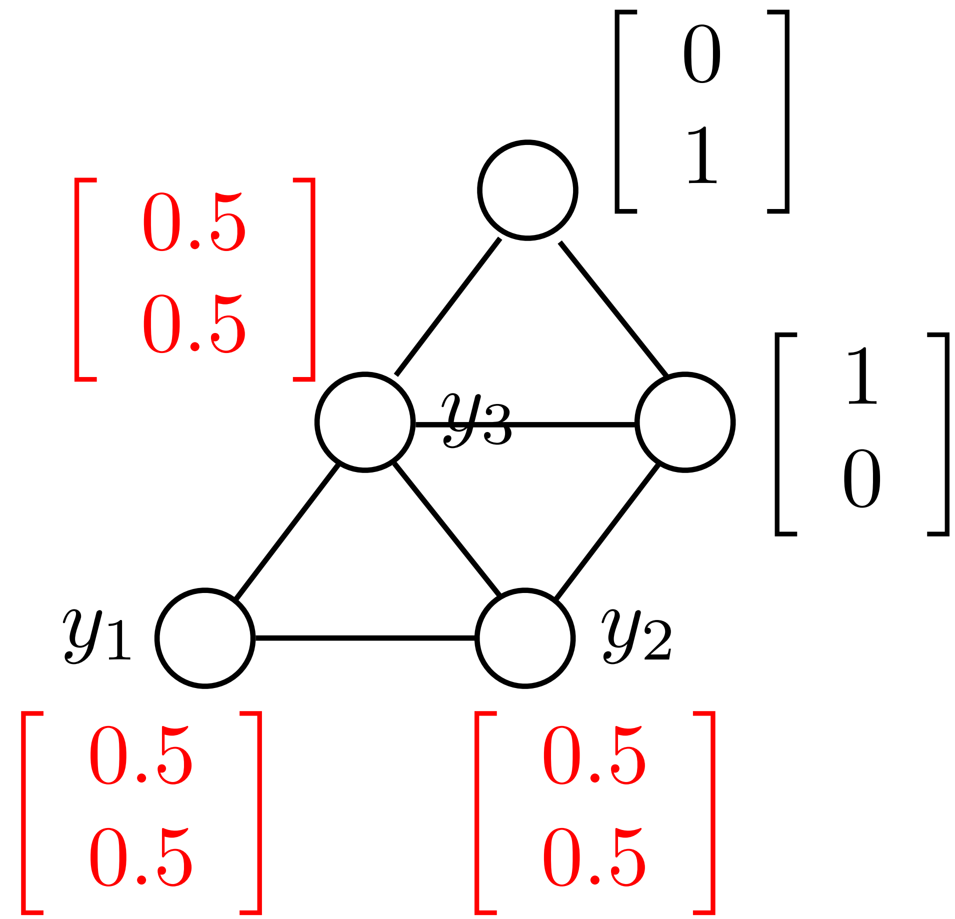
$$\mu_i(y_i) > 0 \;\; \forall y_i$$

holds for all neighbors of integral nodes

(cf. Nemhauser et al. '75, Kolmogorov et al. '05, Weiss et al. '08)

$$\mu = \{\mu_i(y_i),\, \mu_{ij}(y_i, y_j)\} \in \mathcal{M}_L \;\text{ iff}$$

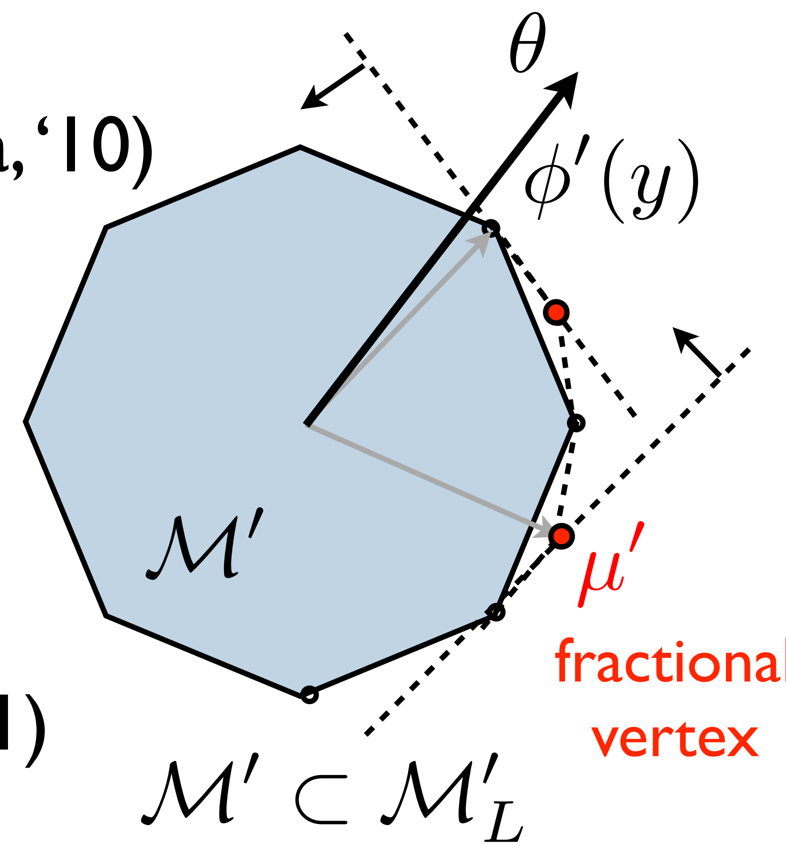
$$\begin{array}{ll} (1) & \mu_{ij}(y_i, y_j) \geq 0, \\ (2) & \sum_{y_i} \mu_i(y_i) = 1, \\ (3) & \sum_{y_i} \mu_{ij}(y_i, y_j) = \mu_j(y_j) \\ & \sum_{y_j} \mu_{ij}(y_i, y_j) = \mu_i(y_i) \end{array}$$



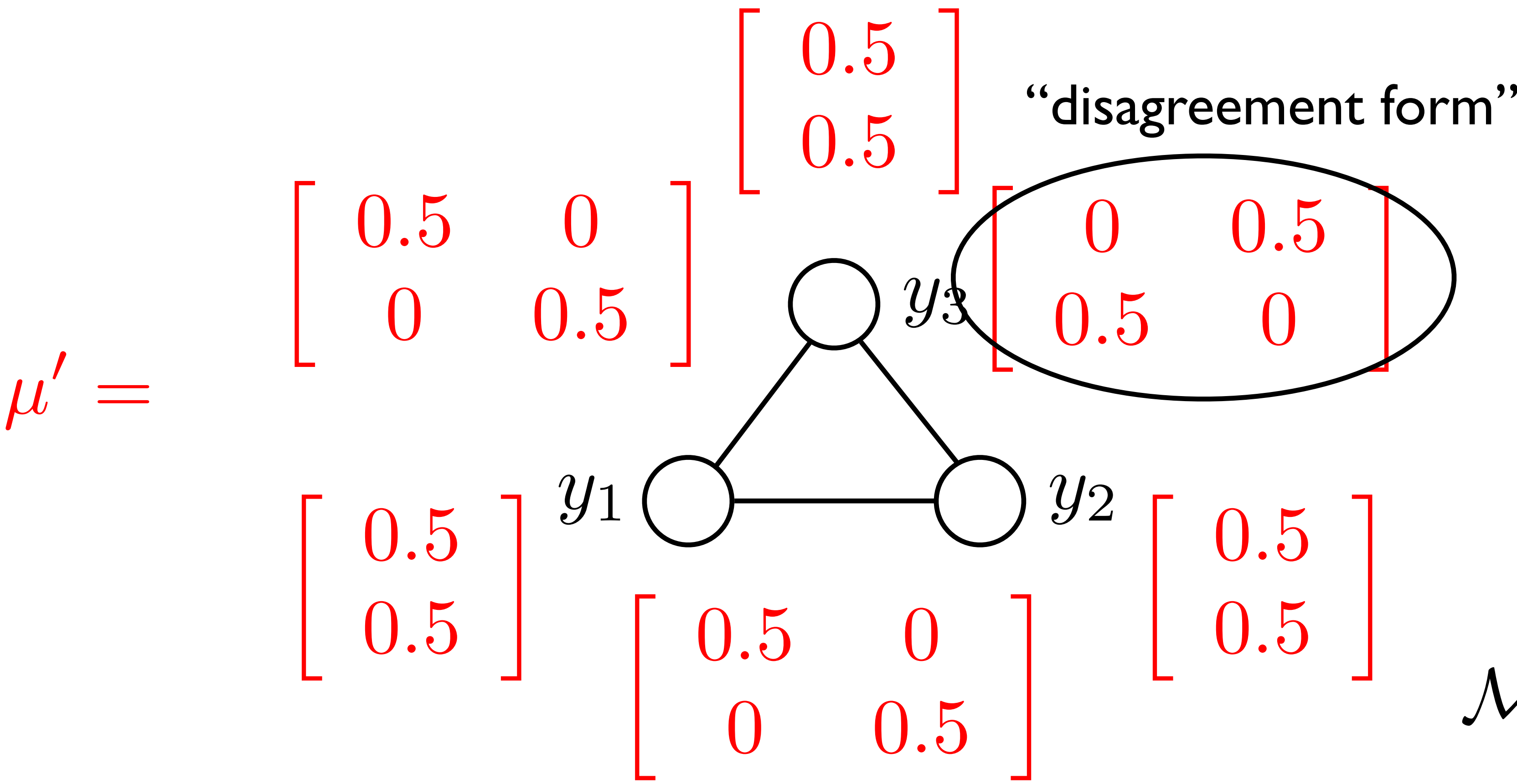
(pairwise marginals not shown)

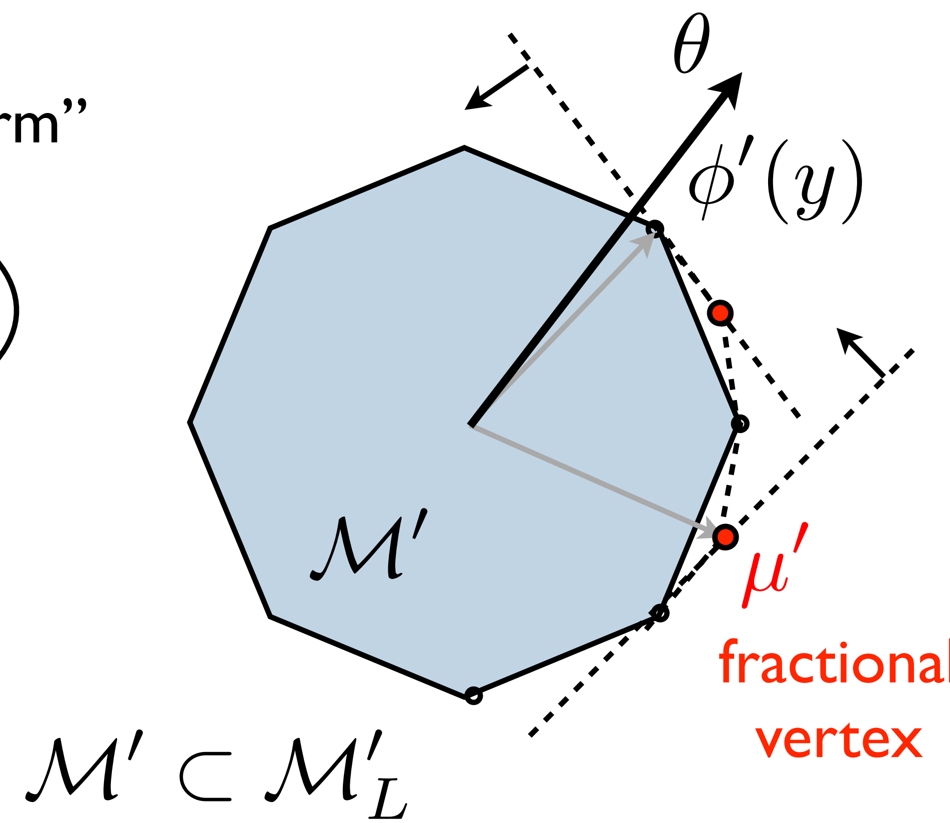
Preventing fractional vertexes

✓ • Structural constraints
  - e.g., trees, planar, perfect graphs (Jebara, '10)
- Parameter restrictions
  - e.g., attractive potentials
- Tighter relaxations
  - e.g., Gomory, Sherali-Adams ('90), Lovasz and Schrijver ('91), Lasserre ('01)



Parameter restrictions





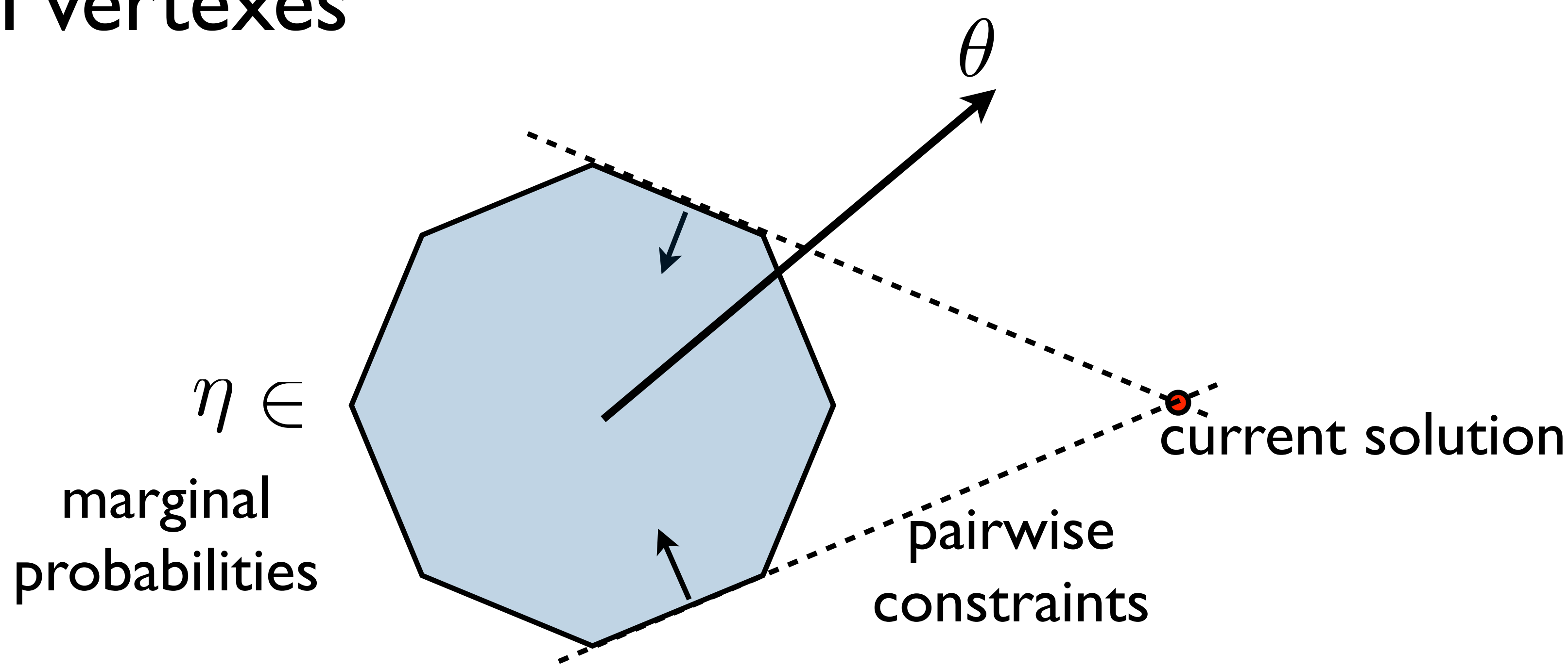
- In the pairwise relaxation, a fractional vertex has to have at least one pairwise marginal in a "disagreement" form.
- We can eliminate such marginals from the maximizing set by constraining the parameters

$$\theta_{ij}(1,1) + \theta_{ij}(0,0) \geq \theta_{ij}(1,0) + \theta_{ij}(0,1)$$

(a.k.a. attractive potentials)

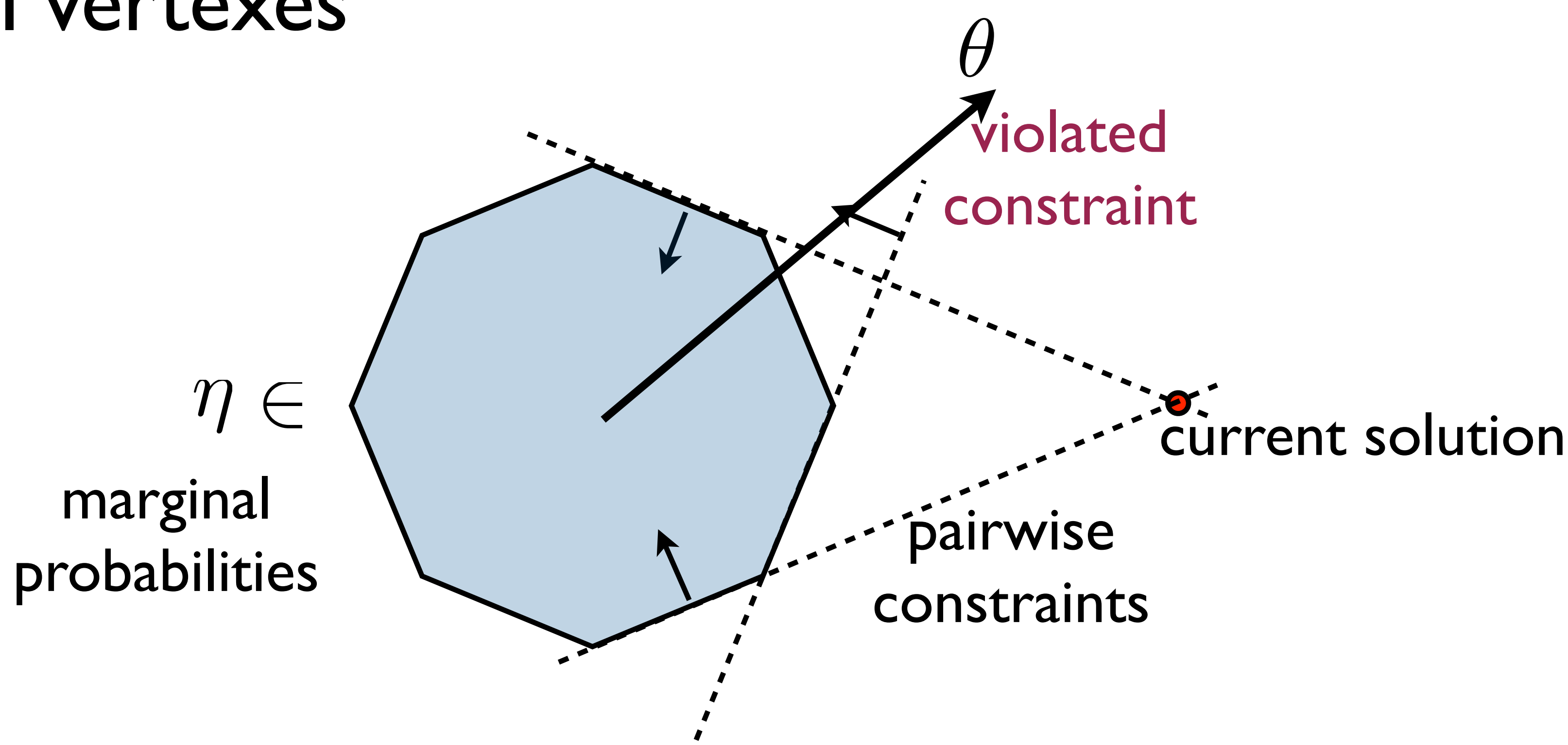
Tighter relaxations

- We can iteratively add additional constraints to remove fractional vertexes



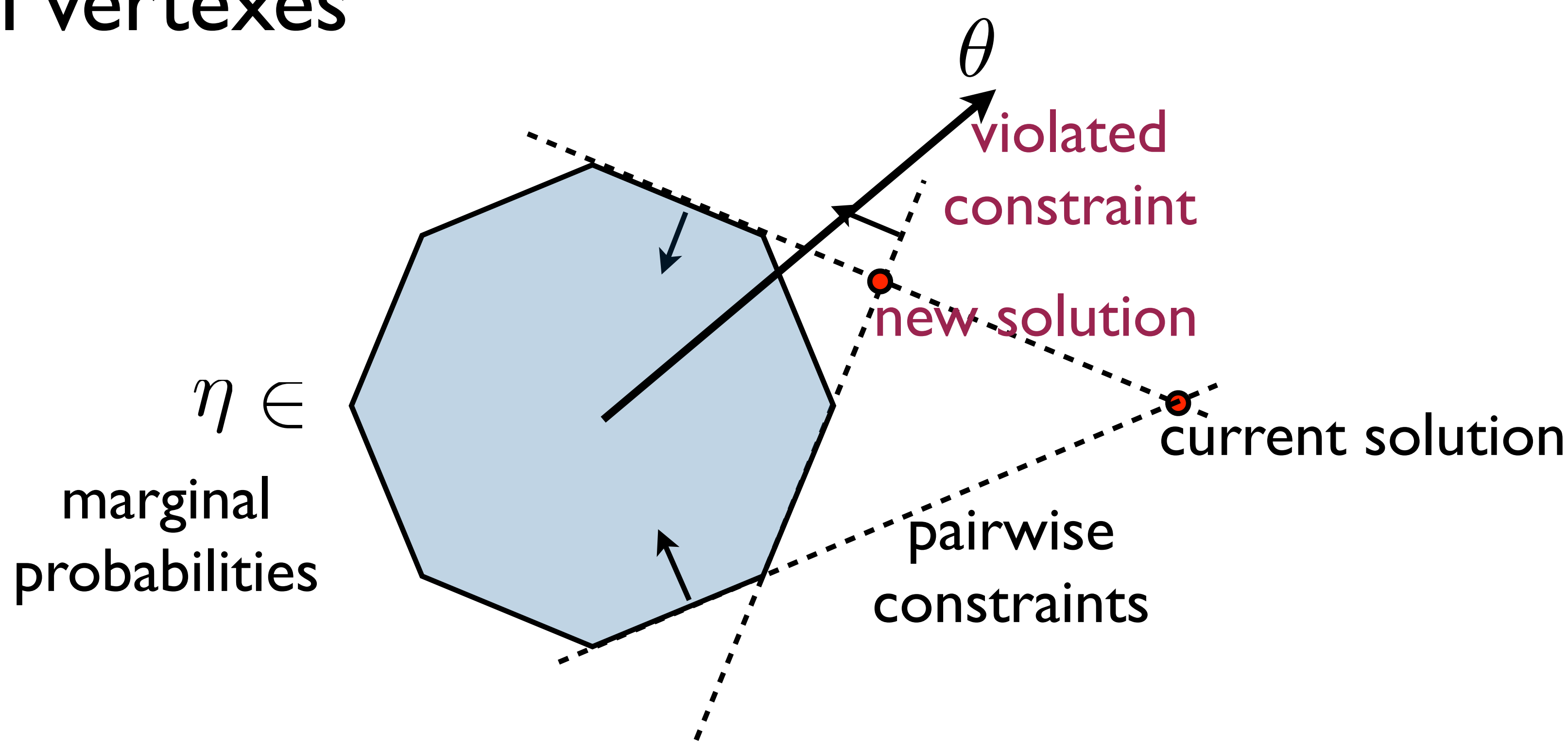
Tighter relaxations

- We can iteratively add additional constraints to remove fractional vertexes



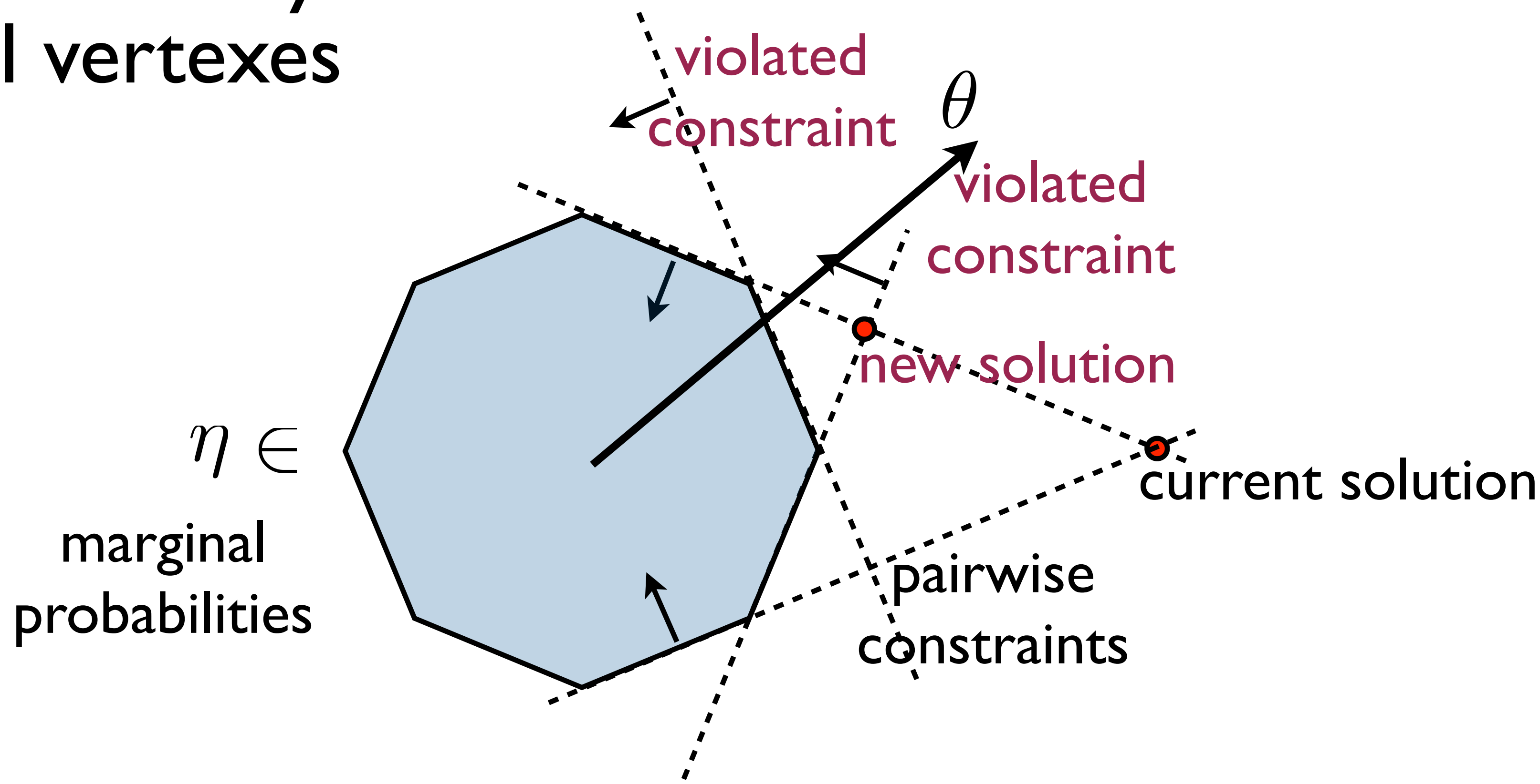
Tighter relaxations

- We can iteratively add additional constraints to remove fractional vertexes



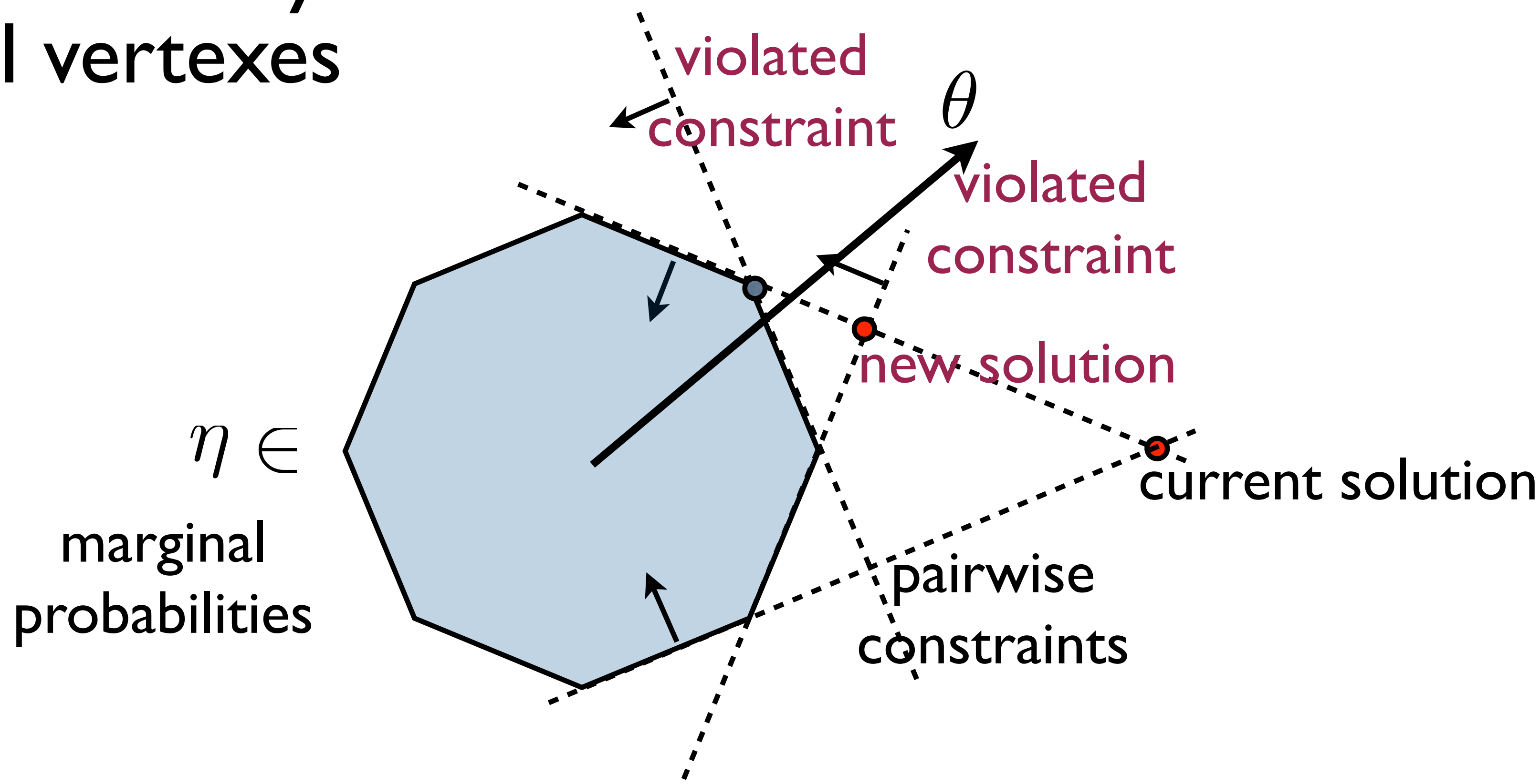
Tighter relaxations

- We can iteratively add additional constraints to remove fractional vertexes



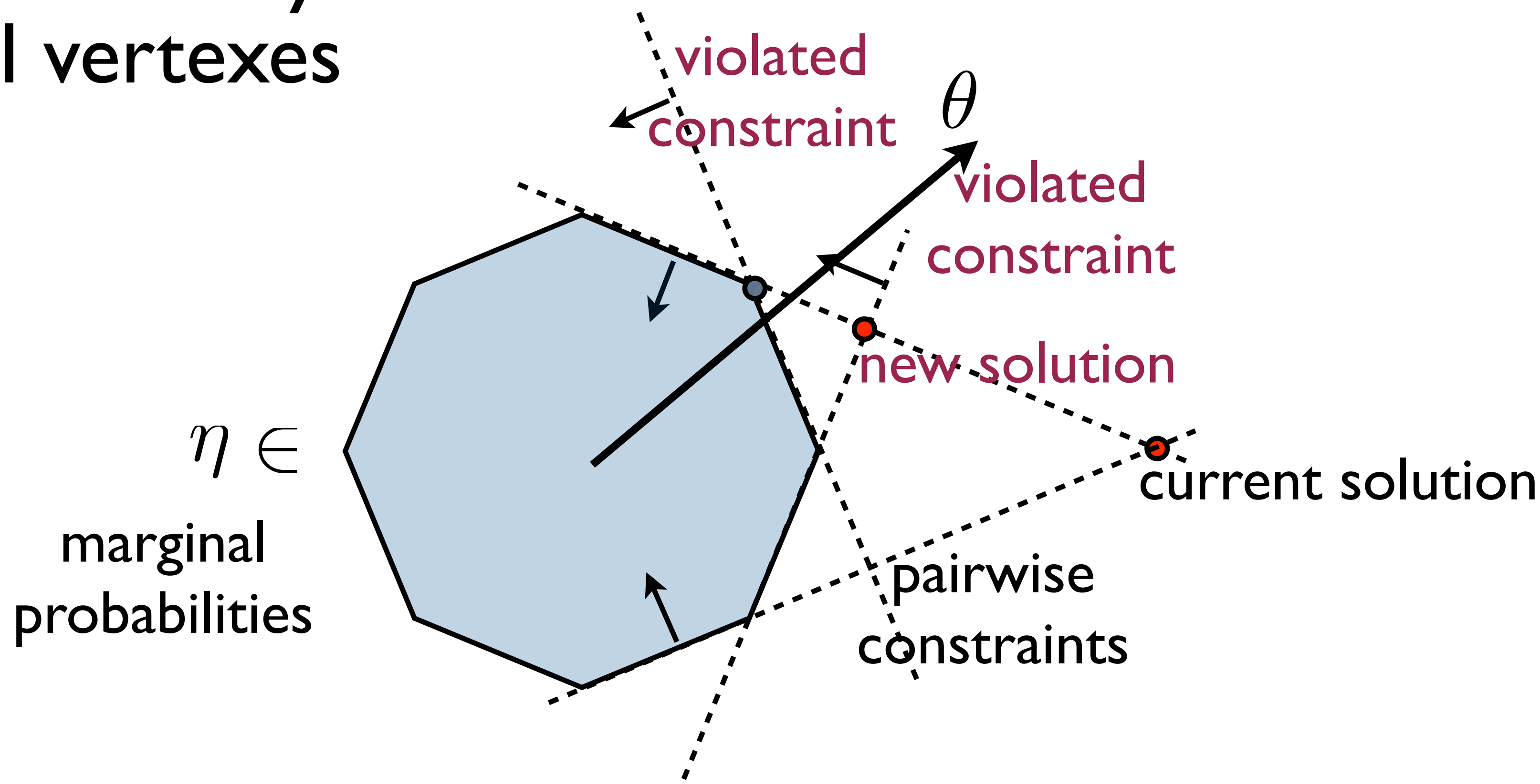
Tighter relaxations

- We can iteratively add additional constraints to remove fractional vertexes



Tighter relaxations

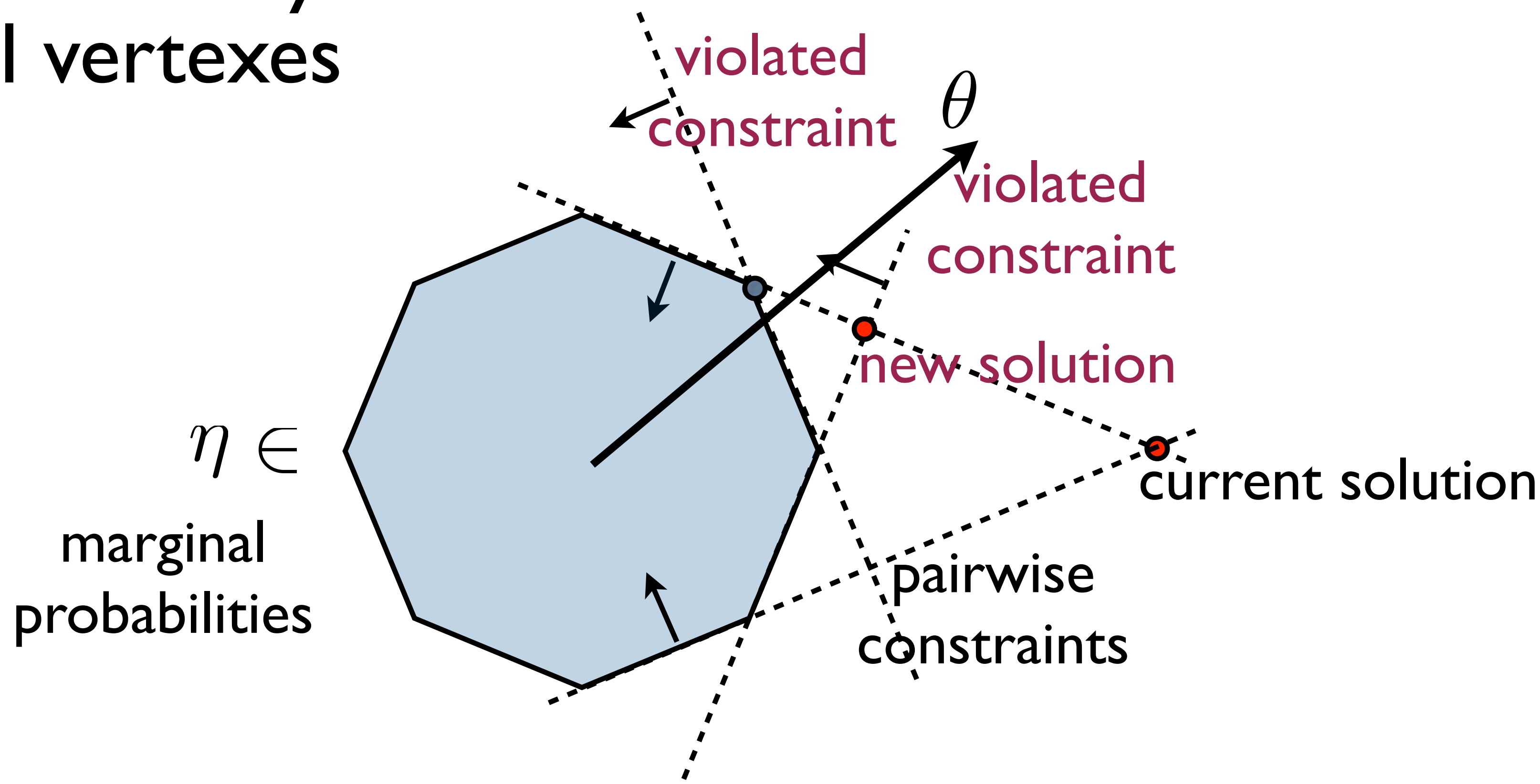
- We can iteratively add additional constraints to remove fractional vertexes



$$\mathcal{M} \subseteq \mathcal{M}^{(k)} \subset \ldots \subset \mathcal{M}^{(1)} \subset \mathcal{M}_L$$

Tighter relaxations

- We can iteratively add additional constraints to remove fractional vertexes

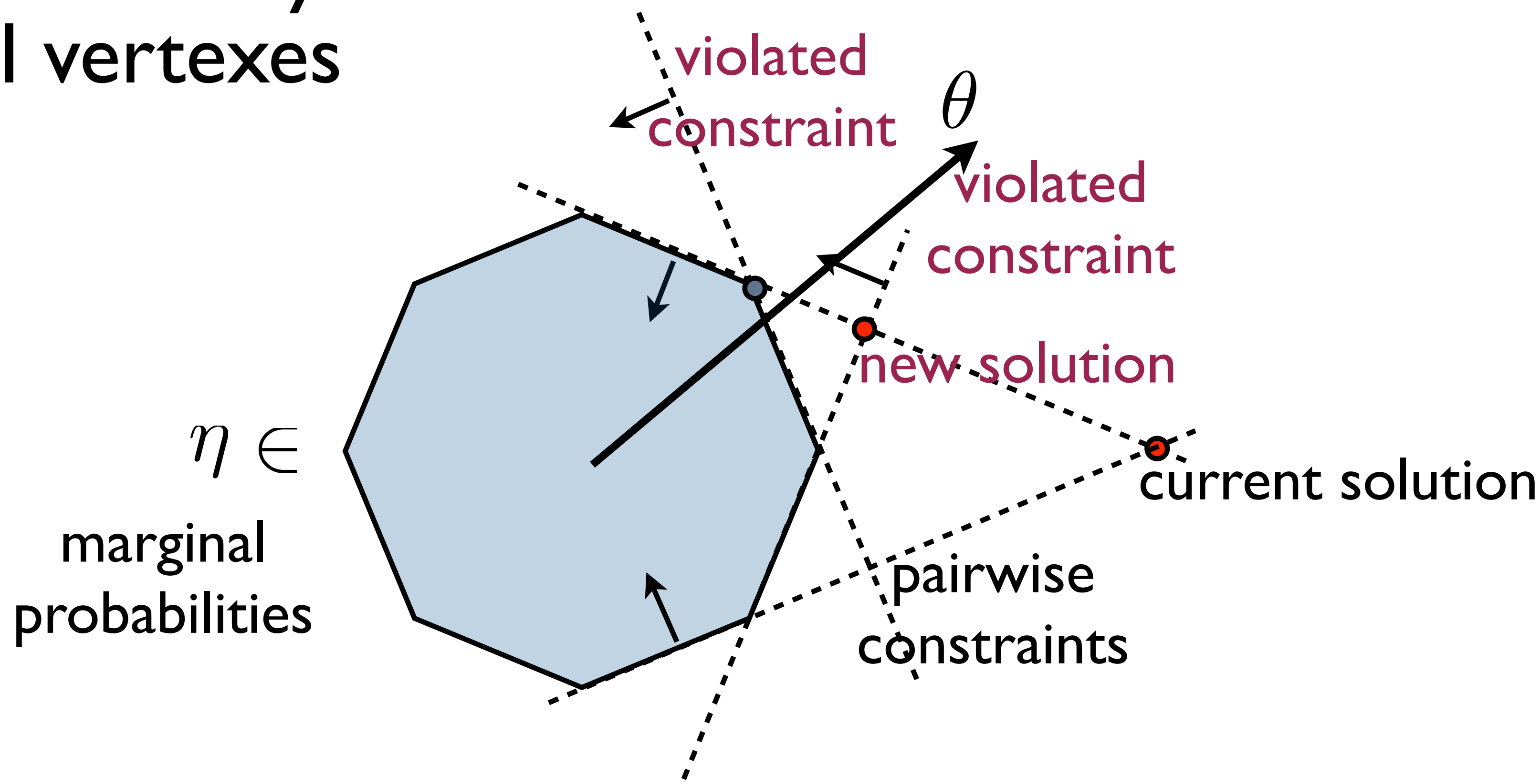


$$\mathcal{M} \subseteq \mathcal{M}^{(k)} \subset \ldots \subset \mathcal{M}^{(1)} \subset \mathcal{M}_L$$

- What are the additional constraints? How to find them?

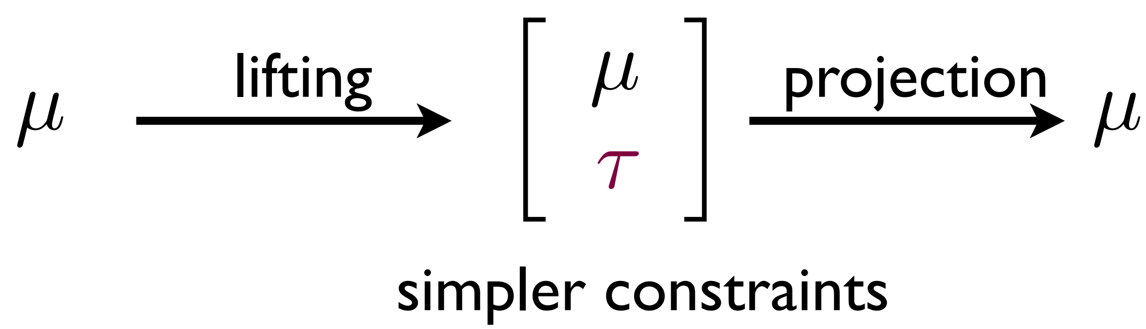
Tighter relaxations

- We can iteratively add additional constraints to remove fractional vertexes



$$\mathcal{M} \subseteq \mathcal{M}^{(k)} \subset \ldots \subset \mathcal{M}^{(1)} \subset \mathcal{M}_L$$

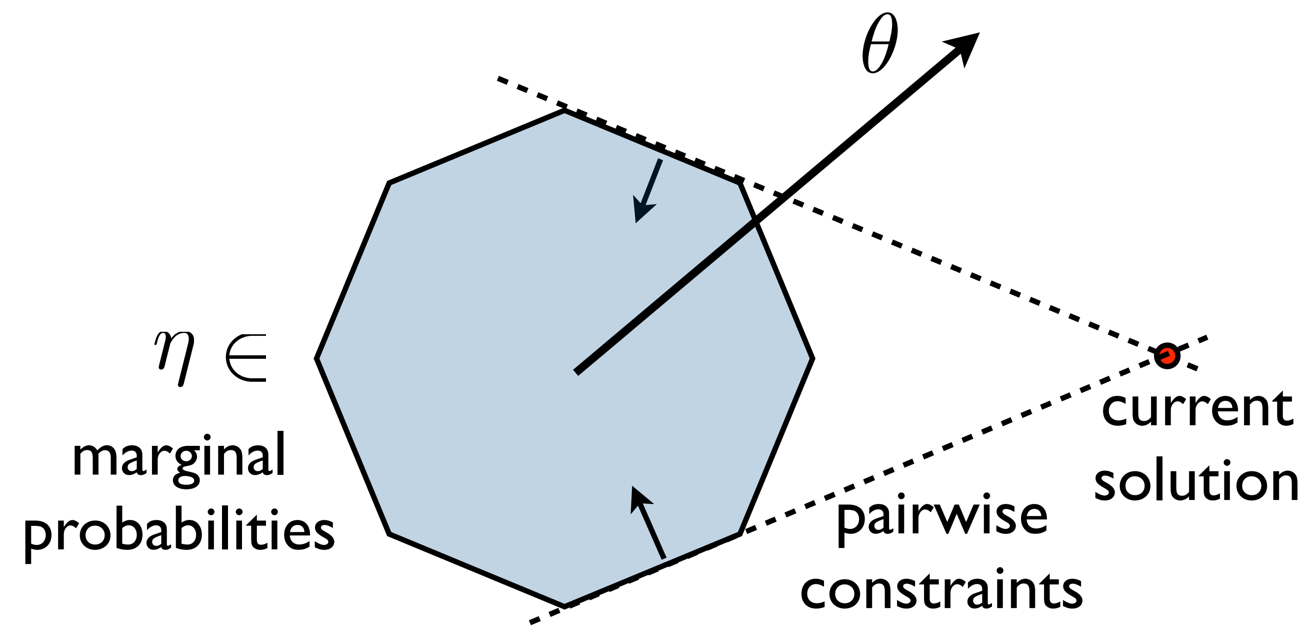
- What are the additional constraints? How to find them?

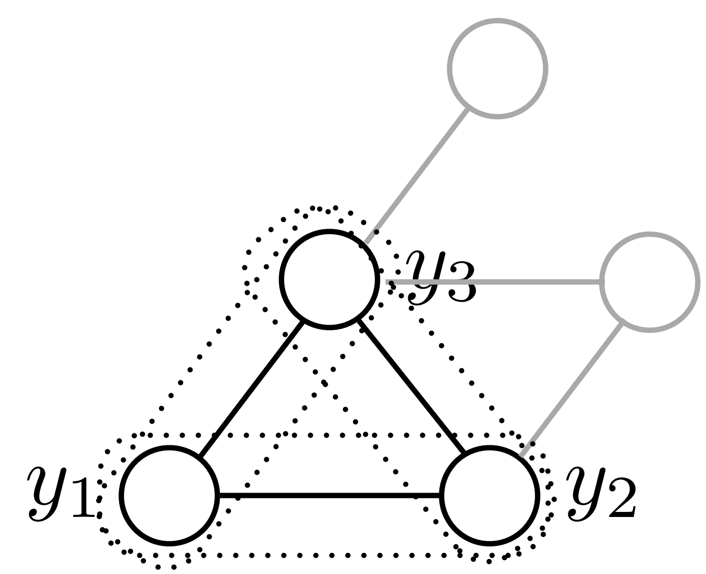


Tighter relaxations via lifting

- We can eliminate some fractional vertexes by requiring that any triplet of pairwise marginals come from a valid distribution

$$
\begin{aligned}
&(1) \quad \mu_{ij}(y_i, y_j) \geq 0, \\
&(2) \quad \textstyle\sum_{y_i} \mu_i(y_i) = 1, \\
&(3) \quad \textstyle\sum_{y_i} \mu_{ij}(y_i, y_j) = \mu_j(y_j) \\
&\qquad\; \textstyle\sum_{y_j} \mu_{ij}(y_i, y_j) = \mu_i(y_i)
\end{aligned}
$$

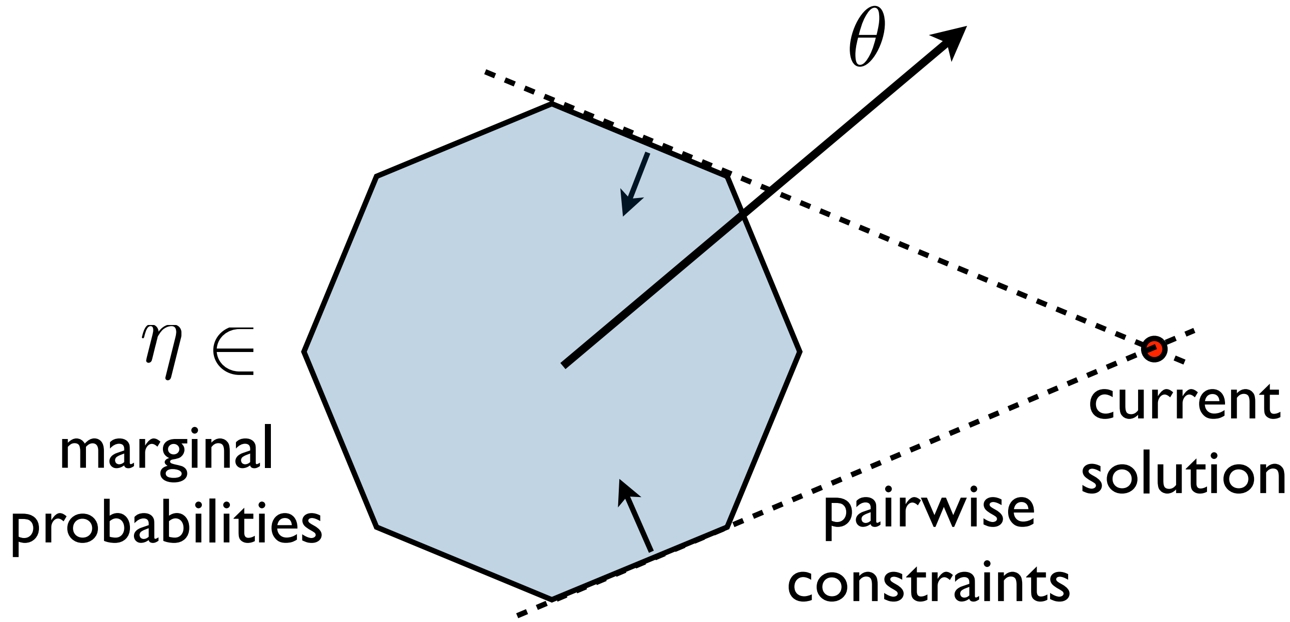




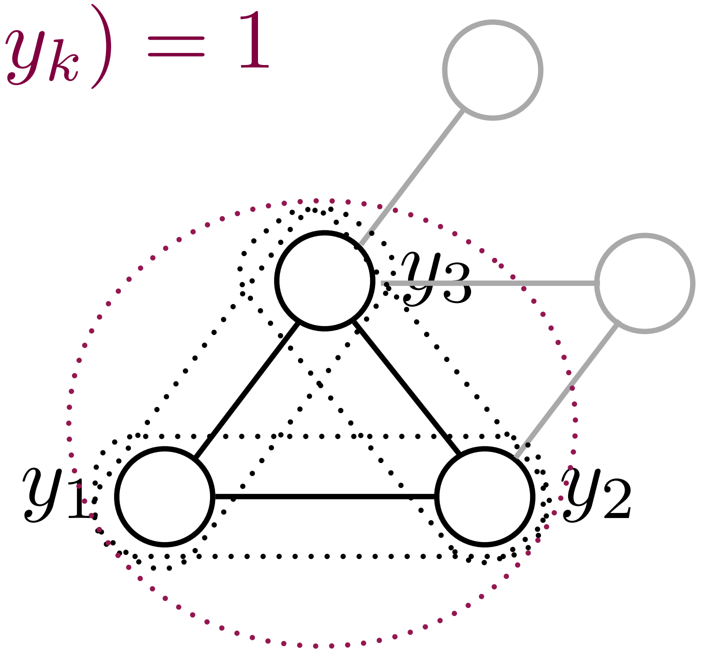
Tighter relaxations via lifting

- We can eliminate some fractional vertexes by requiring that any triplet of pairwise marginals come from a valid distribution

$$
\begin{array}{ll}
(1) & \mu_{ij}(y_i, y_j) \geq 0, \\
(2) & \sum_{y_i} \mu_i(y_i) = 1, \\
(3) & \sum_{y_i} \mu_{ij}(y_i, y_j) = \mu_j(y_j) \\
 & \sum_{y_j} \mu_{ij}(y_i, y_j) = \mu_i(y_i)
\end{array}
$$



$$
\begin{array}{ll}
(4) & \tau_{ijk}(y_i, y_j, y_k) \geq 0, \ \sum_{y_i, y_j, y_k} \tau_{ijk}(y_i, y_j, y_k) = 1 \\
(5) & \sum_{y_k} \tau_{ijk}(y_i, y_j, y_k) = \mu_{ij}(y_i, y_j) \\
 & \sum_{y_j} \tau_{ijk}(y_i, y_j, y_k) = \mu_{ik}(y_i, y_k) \\
 & \sum_{y_i} \tau_{ijk}(y_i, y_j, y_k) = \mu_{jk}(y_j, y_k)
\end{array}
$$



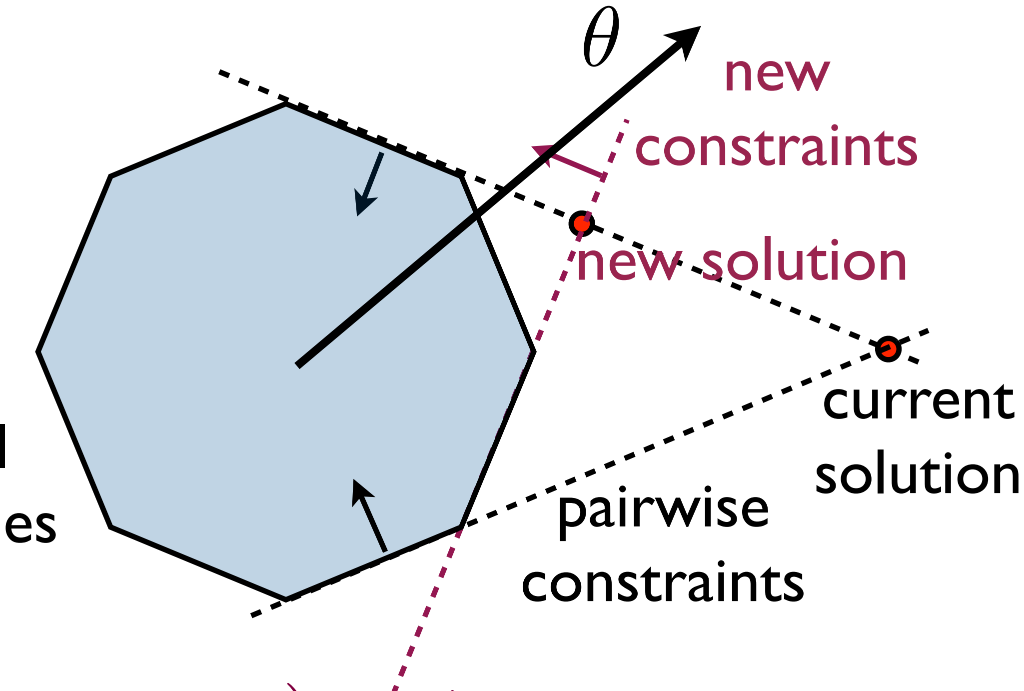
Tighter relaxations via lifting

- We can eliminate some fractional vertexes by requiring that any triplet of pairwise marginals come from a valid distribution

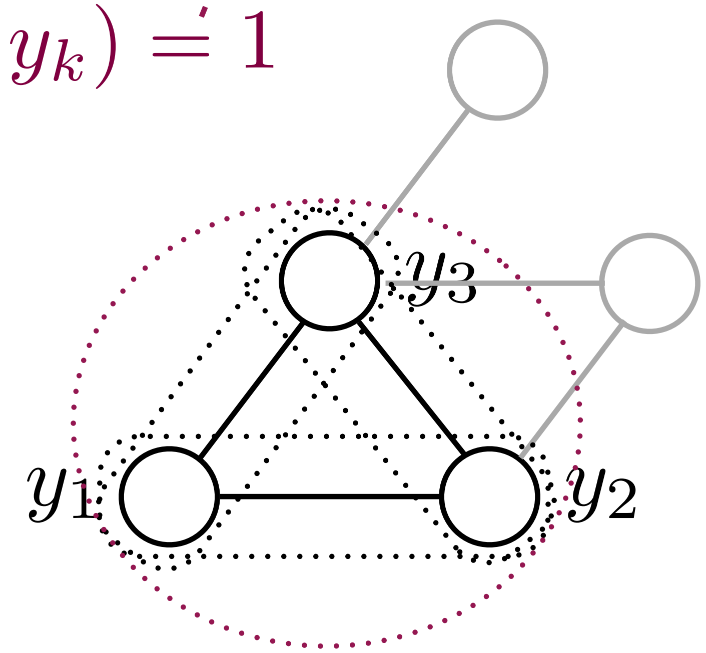
$$
\begin{array}{ll}
(1) & \mu_{ij}(y_i, y_j) \geq 0, \\
(2) & \sum_{y_i} \mu_i(y_i) = 1, \\
(3) & \sum_{y_i} \mu_{ij}(y_i, y_j) = \mu_j(y_j) \\
 & \sum_{y_j} \mu_{ij}(y_i, y_j) = \mu_i(y_i)
\end{array}
$$

$\eta \in$

marginal probabilities



$$
\begin{array}{ll}
(4) & \tau_{ijk}(y_i, y_j, y_k) \geq 0, \;\; \sum_{y_i, y_j, y_k} \tau_{ijk}(y_i, y_j, y_k) = 1 \\
(5) & \sum_{y_k} \tau_{ijk}(y_i, y_j, y_k) = \mu_{ij}(y_i, y_j) \\
 & \sum_{y_j} \tau_{ijk}(y_i, y_j, y_k) = \mu_{ik}(y_i, y_k) \\
 & \sum_{y_i} \tau_{ijk}(y_i, y_j, y_k) = \mu_{jk}(y_j, y_k)
\end{array}
$$



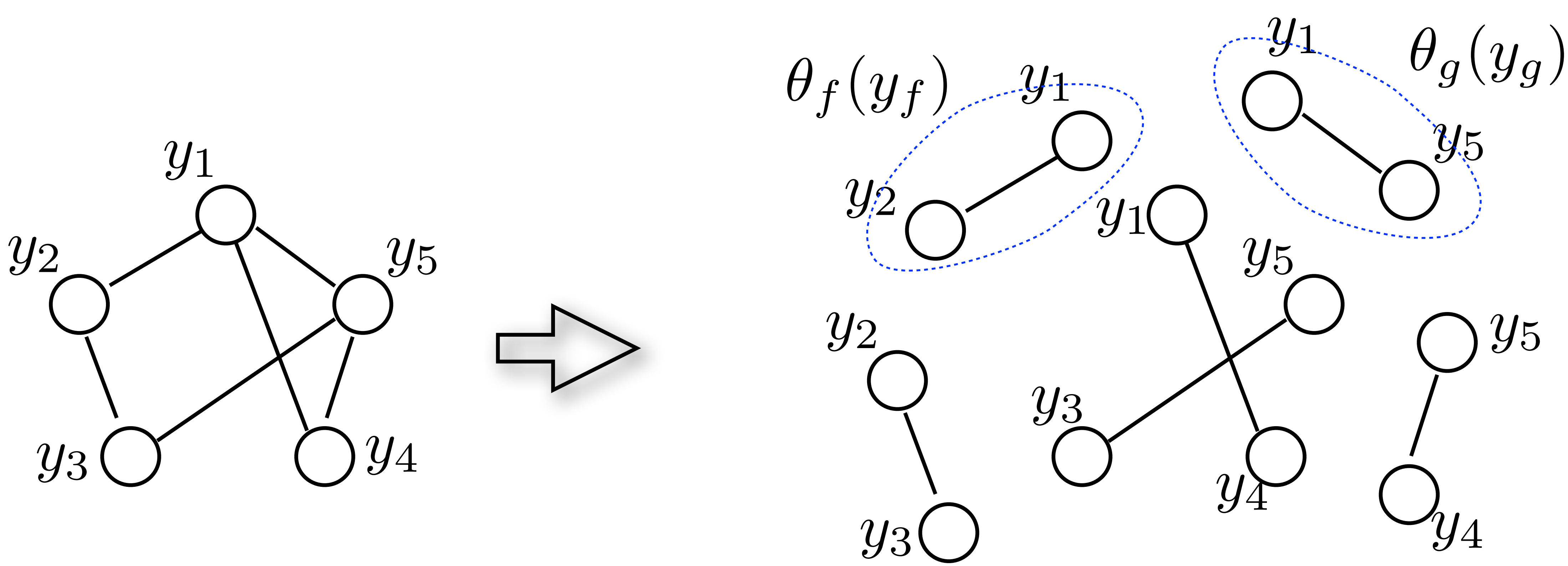
Dual decomposition

- There are many possible ways of preventing fractional vertexes, including
  - ✓ - Structural constraints
    - e.g., trees, planar, perfect graphs (Jebara, '10)
  - ✓ - Parameter restrictions
    - e.g., attractive potentials
  - ✓ - Tighter relaxations
    - e.g., Gomory, Sherali-Adams ('90), Lovasz and Schrijver ('91), Lasserre ('01)
- These are key ingredients in decomposition methods that break the original problem into exactly solvable sub-problems (cf. Guignard, Fisher, '80's)
  - e.g., Wainwright et al. '02+, Johnson et al. '07, Kommodakis '07+, Sontag et al. '08+, etc.

Decomposition

- We can always break the original model into pieces that would be exactly solvable in isolation

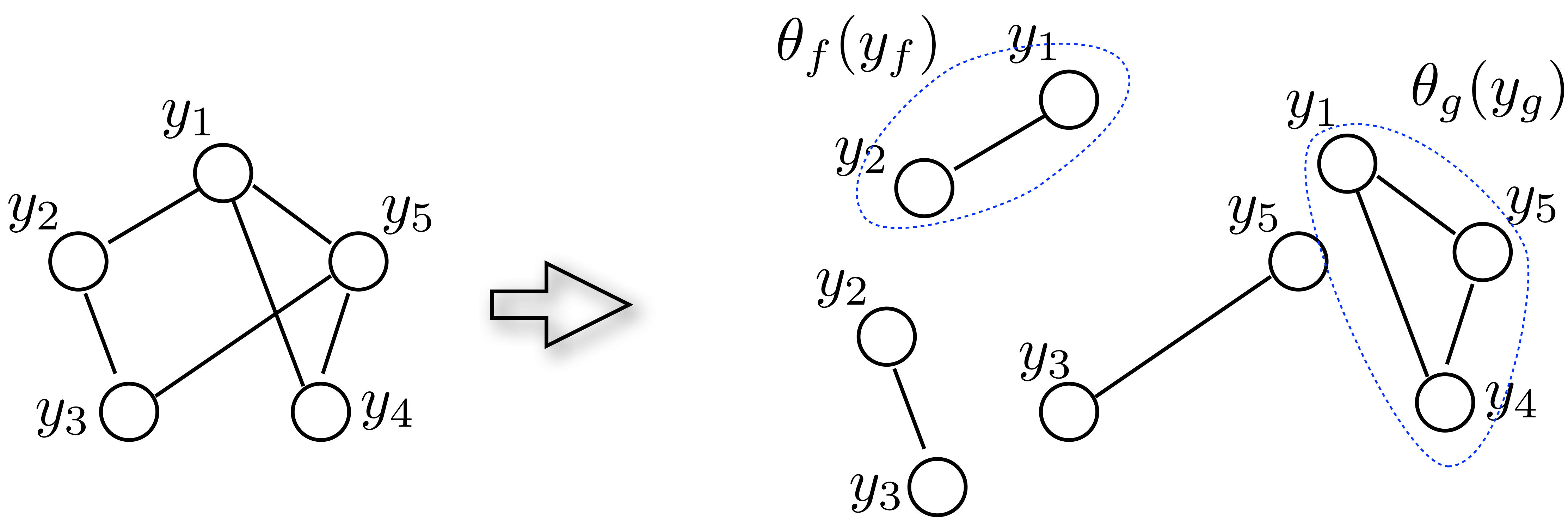
$$\sum_{(i,j)\in E} \theta_{ij}(y_i, y_j) = \sum_{f\in F} \theta_f(y_f)$$



Decomposition

- We can always break the original model into pieces that would be exactly solvable in isolation

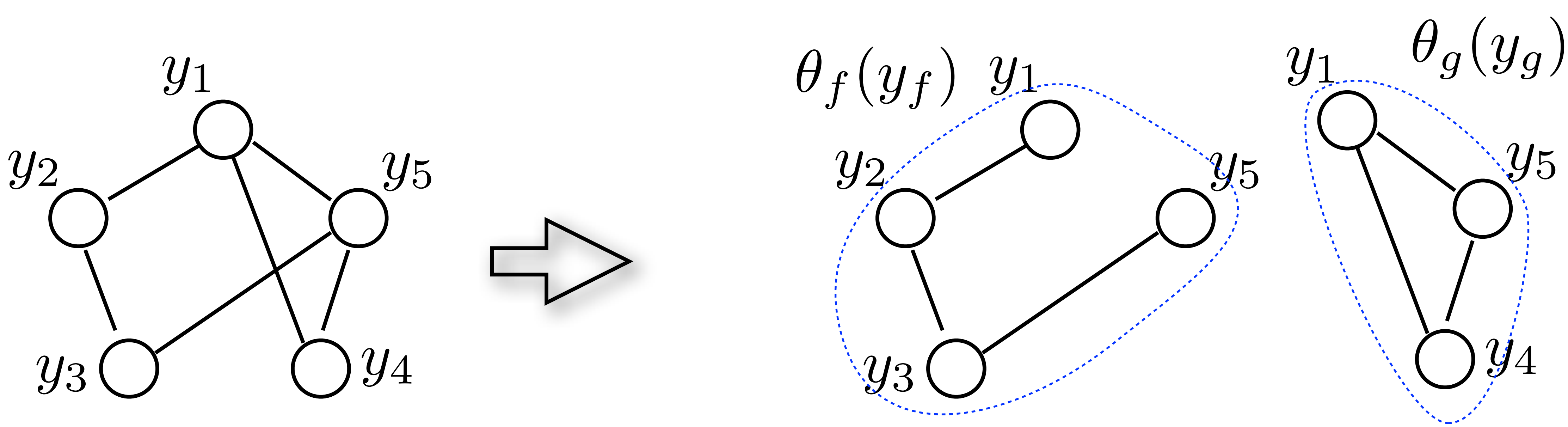
$$\sum_{(i,j)\in E} \theta_{ij}(y_i, y_j) = \sum_{f\in F} \theta_f(y_f)$$



Decomposition

- We can always break the original model into pieces that would be exactly solvable in isolation

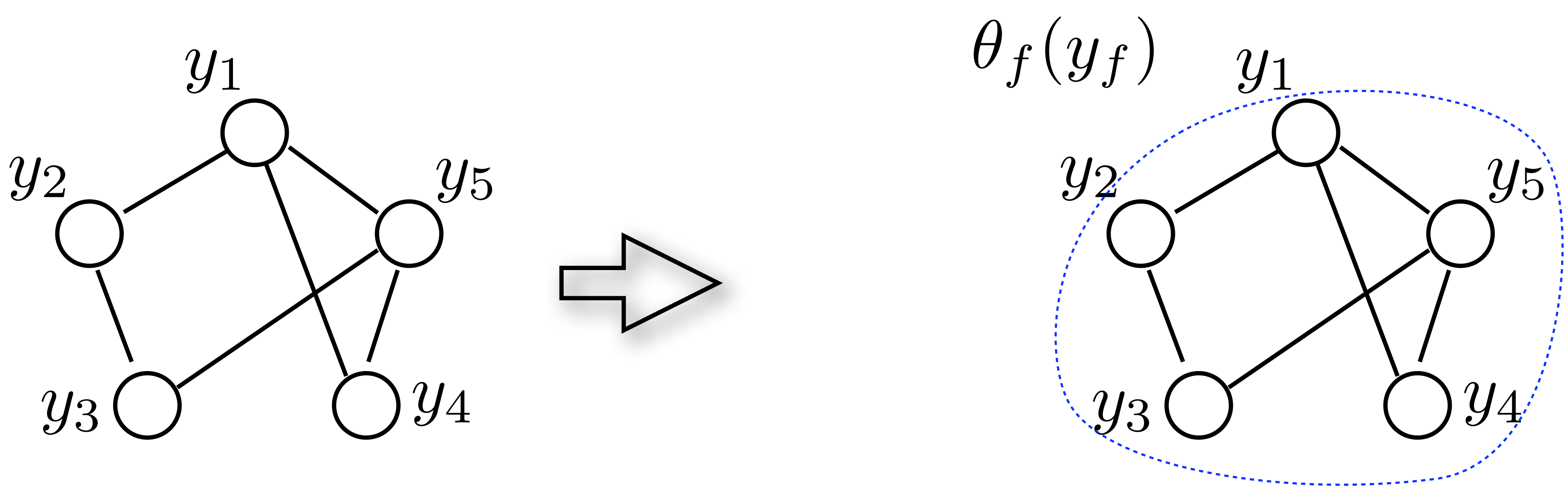
$$\sum_{(i,j)\in E} \theta_{ij}(y_i, y_j) = \sum_{f\in F} \theta_f(y_f)$$



Decomposition

- We can always break the original model into pieces that would be exactly solvable in isolation

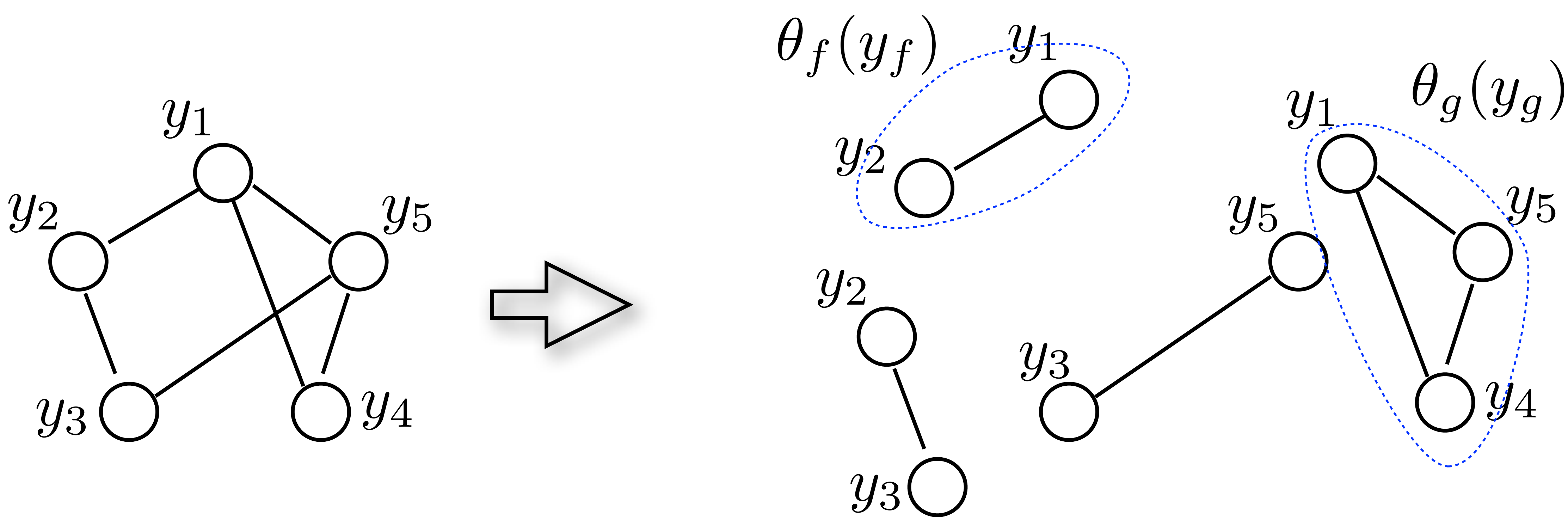
$$\sum_{(i,j)\in E} \theta_{ij}(y_i, y_j) = \qquad \theta_f(y_f)$$



Decomposition

- We can always break the original model into pieces that would be exactly solvable in isolation

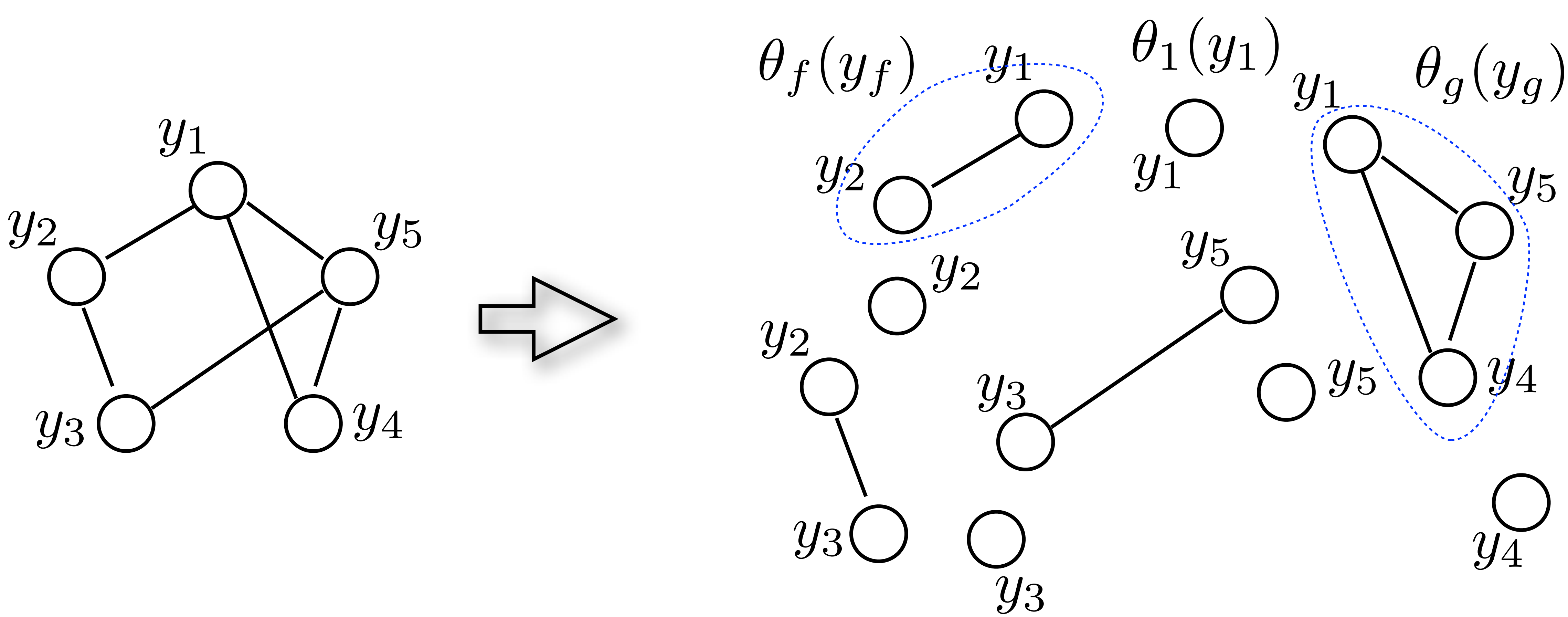
$$\sum_{(i,j)\in E} \theta_{ij}(y_i, y_j) = \sum_{f\in F} \theta_f(y_f)$$



Decomposition

- We can always break the original model into pieces that would be exactly solvable in isolation

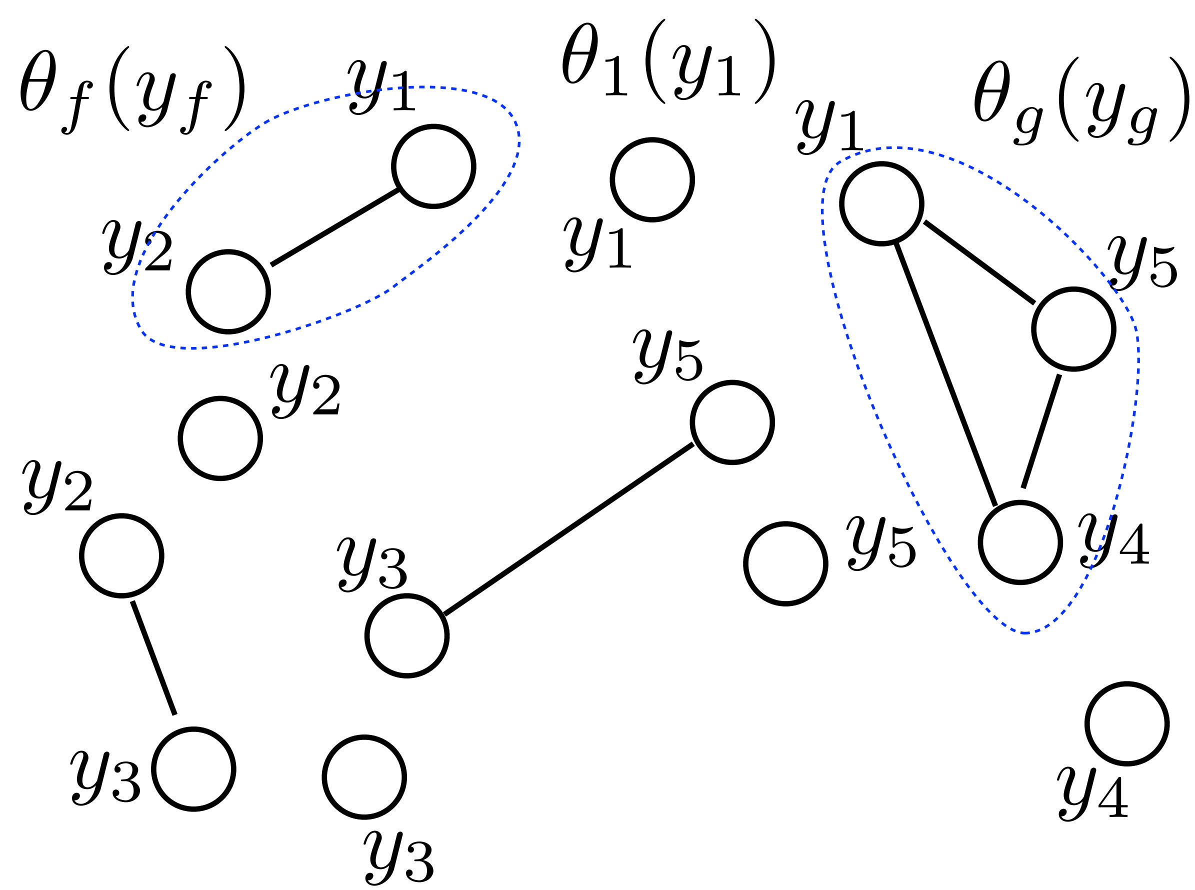
$$\sum_{(i,j)\in E} \theta_{ij}(y_i, y_j) = \sum_i \theta_i(y_i) + \sum_{f\in F} \theta_f(y_f)$$



Decomposition

- We can always break the original model into pieces that would be exactly solvable in isolation

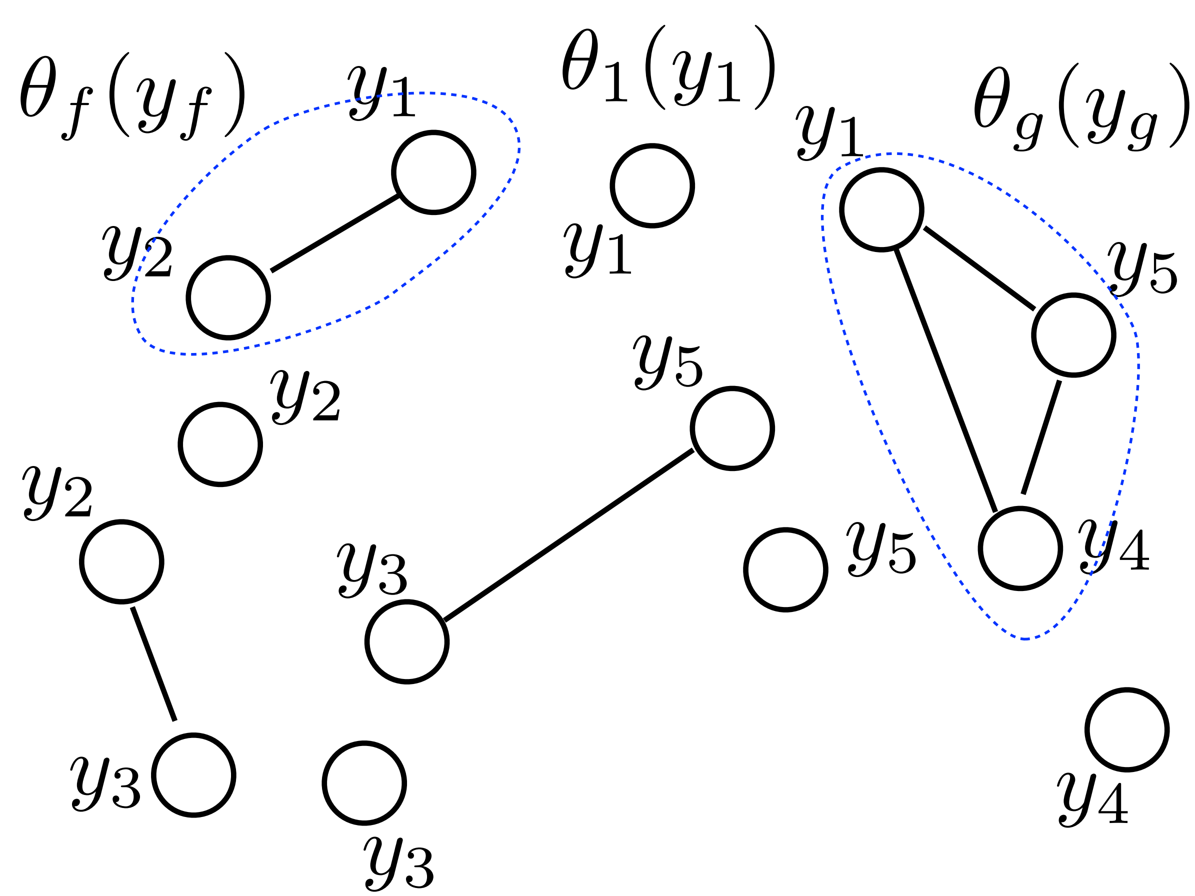
$$\sum_i \theta_i(y_i) + \sum_{f \in F} \theta_f(y_f)$$



Decomposition

- We can always break the original model into pieces that would be exactly solvable in isolation

$$\sum_i \theta_i(y_i) + \sum_{f \in F} \theta_f(y_f) = \theta \cdot \phi(y)$$

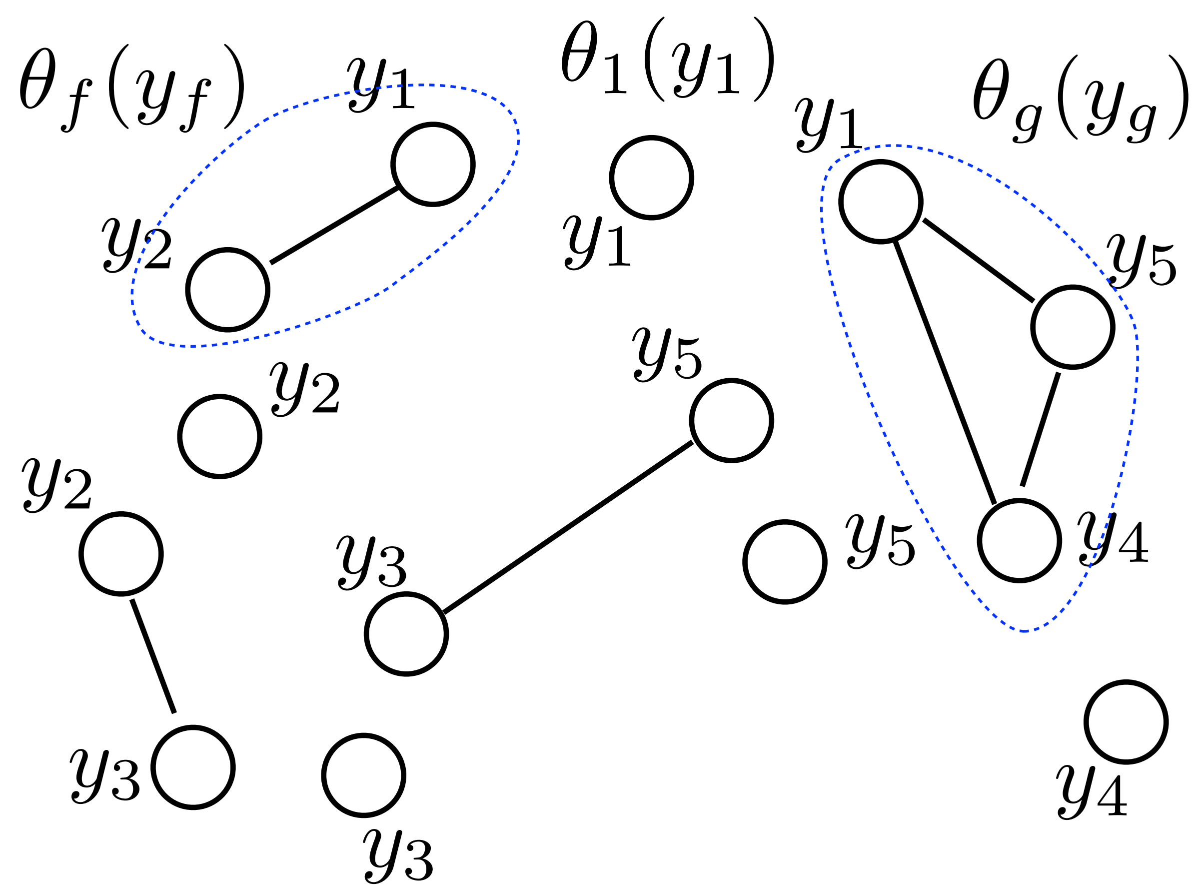


Decomposition, LP relaxation

- The decomposition leads to a natural component-wise LP relaxation

$$\max_y \Big\{ \sum_i \theta_i(y_i) + \sum_{f \in F} \theta_f(y_f) \Big\} = \max_{\mu \in \mathcal{M}} \{\theta \cdot \mu\}$$

$$= \max_{\mu \in \mathcal{M}} \Big\{ \sum_i \sum_{y_i} \mu_i(y_i)\theta_i(y_i) + \sum_{f \in F} \sum_{y_f} \mu_f(y_f)\theta_f(y_f) \Big\}$$



Decomposition, LP relaxation

- The decomposition leads to a natural component-wise LP relaxation

$$\max_y \Big\{ \sum_i \theta_i(y_i) + \sum_{f \in F} \theta_f(y_f) \Big\} \leq \max_{\mu \in \mathcal{M}_L} \{\theta \cdot \mu\}$$

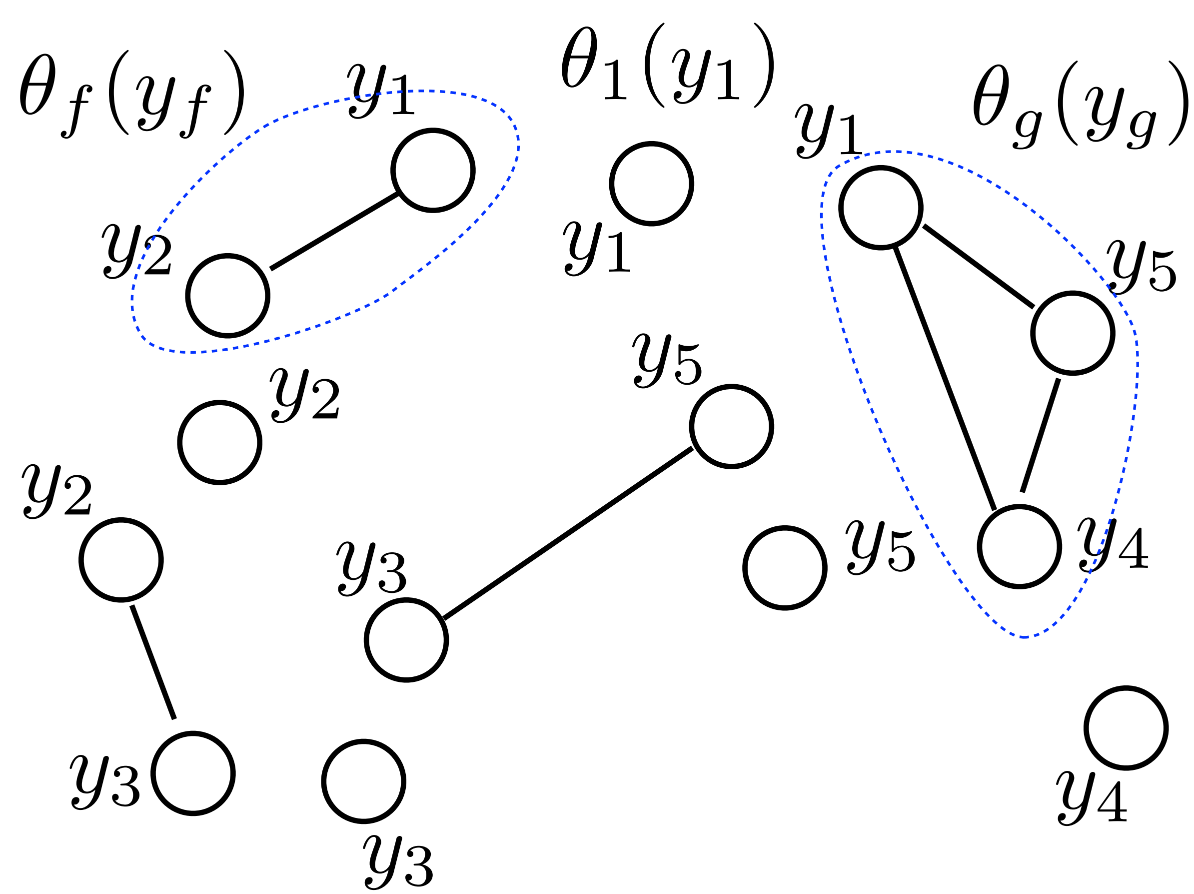
$$= \max_{\mu \in \mathcal{M}_L} \Big\{ \sum_i \sum_{y_i} \mu_i(y_i)\theta_i(y_i) + \sum_{f \in F} \sum_{y_f} \mu_f(y_f)\theta_f(y_f) \Big\}$$

$\mu = \{\mu_i(y_i), \mu_f(y_f)\} \in \mathcal{M}_L$ iff

$\mu_i(y_i) \geq 0, \;\; \mu_f(y_f) \geq 0$

$$\sum_{y_i} \mu_i(y_i) = 1$$

$$\sum_{y_{f \setminus i}} \mu_f(y_f) = \mu_i(y_i), \;\; \text{if } i \in f$$

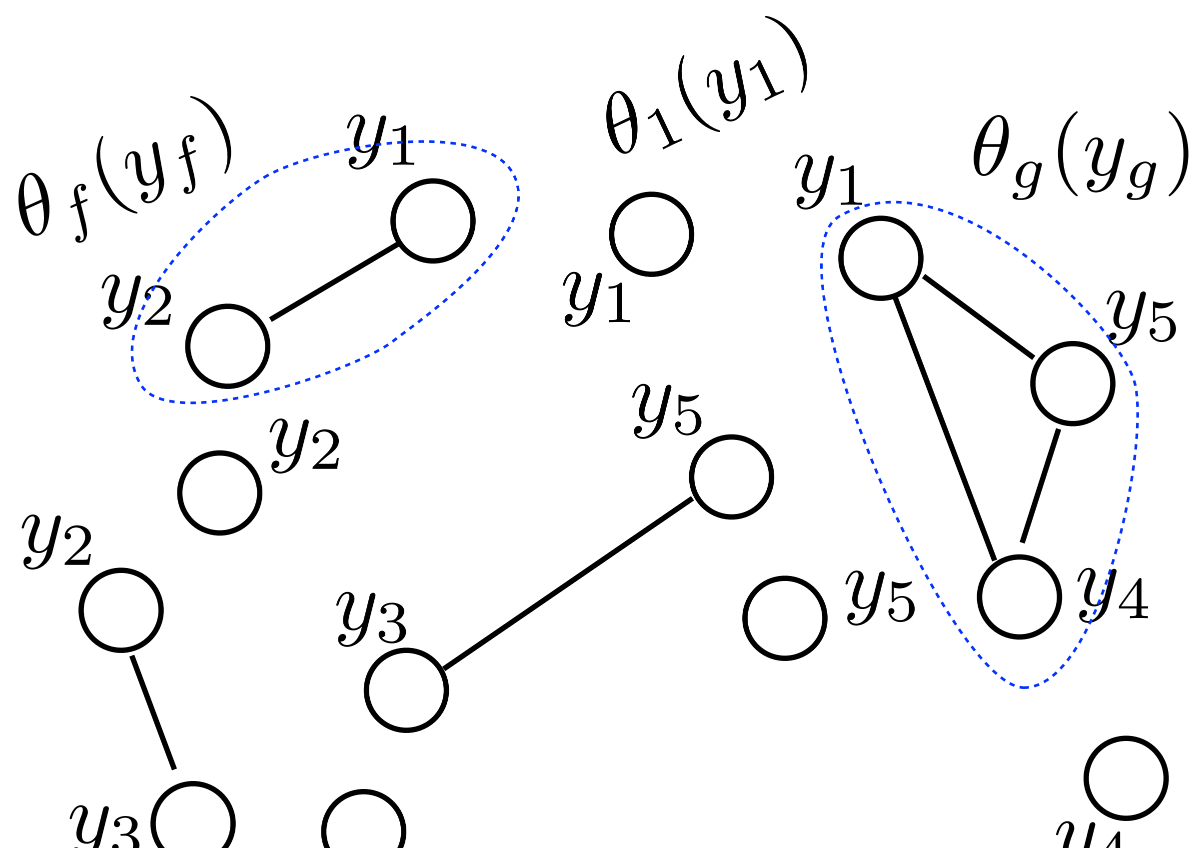


Dual decomposition

- We can think of the relaxation as enforcing that the components agree about the maximizing assignment

$$\max_y \Big\{ \sum_i \theta_i(y_i) + \sum_{f \in F} \theta_f(y_f) \Big\}$$

$$\leq \sum_i \max_{y_i} \{\theta_i(y_i)\} + \sum_{f \in F} \max_{y_f} \{\theta_f(y_f)\}$$

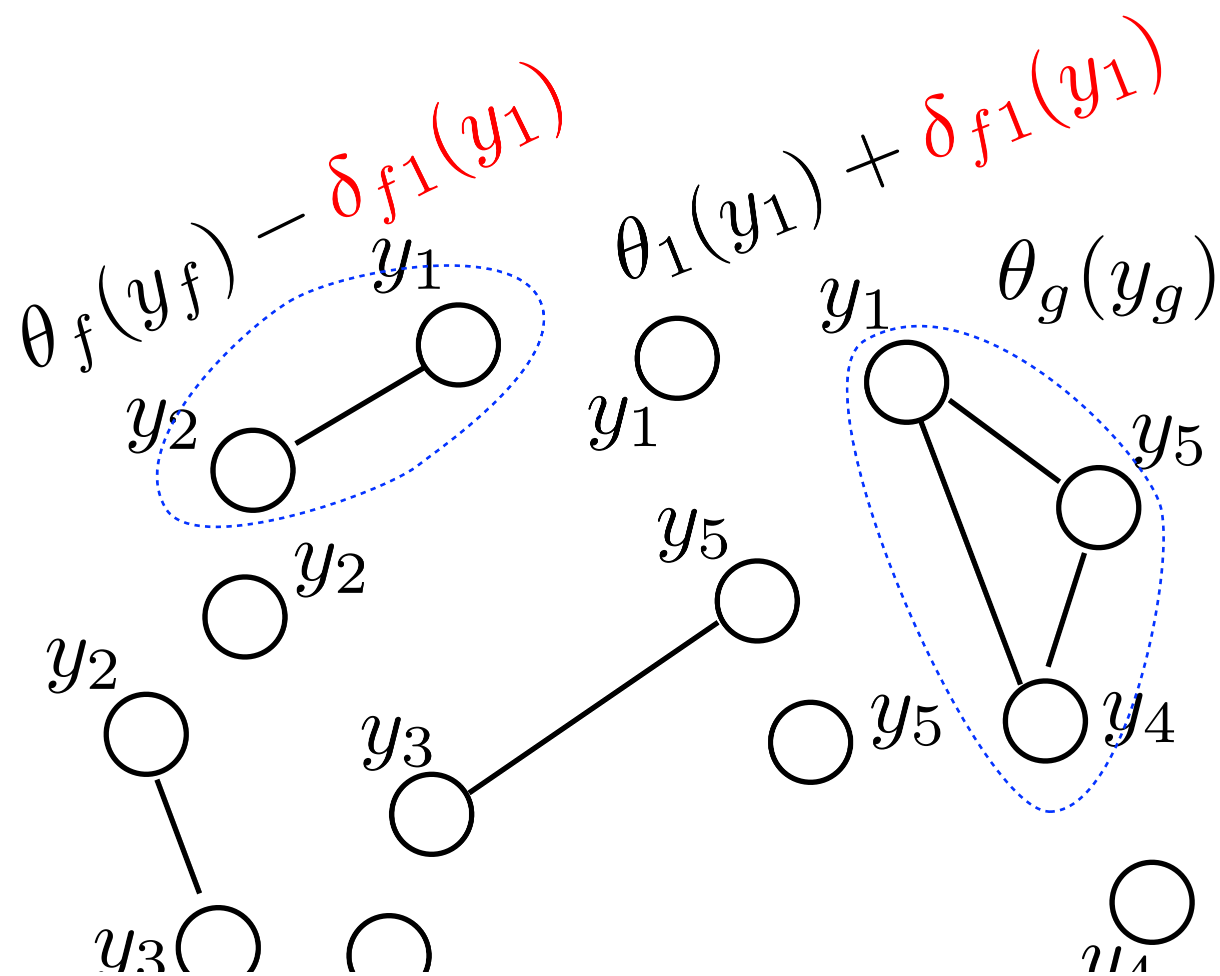


Dual decomposition

- We can think of the relaxation as enforcing that the components agree about the maximizing assignment

$$\max_{y} \Big\{ \sum_{i} \theta_i(y_i) + \sum_{f \in F} \theta_f(y_f) \Big\}$$
$$\leq \sum_{i} \max_{y_i} \{\theta_i(y_i)\} + \sum_{f \in F} \max_{y_f} \{\theta_f(y_f)\}$$

- Agreements can be enforced via Lagrange multipliers associated with the equality constraints



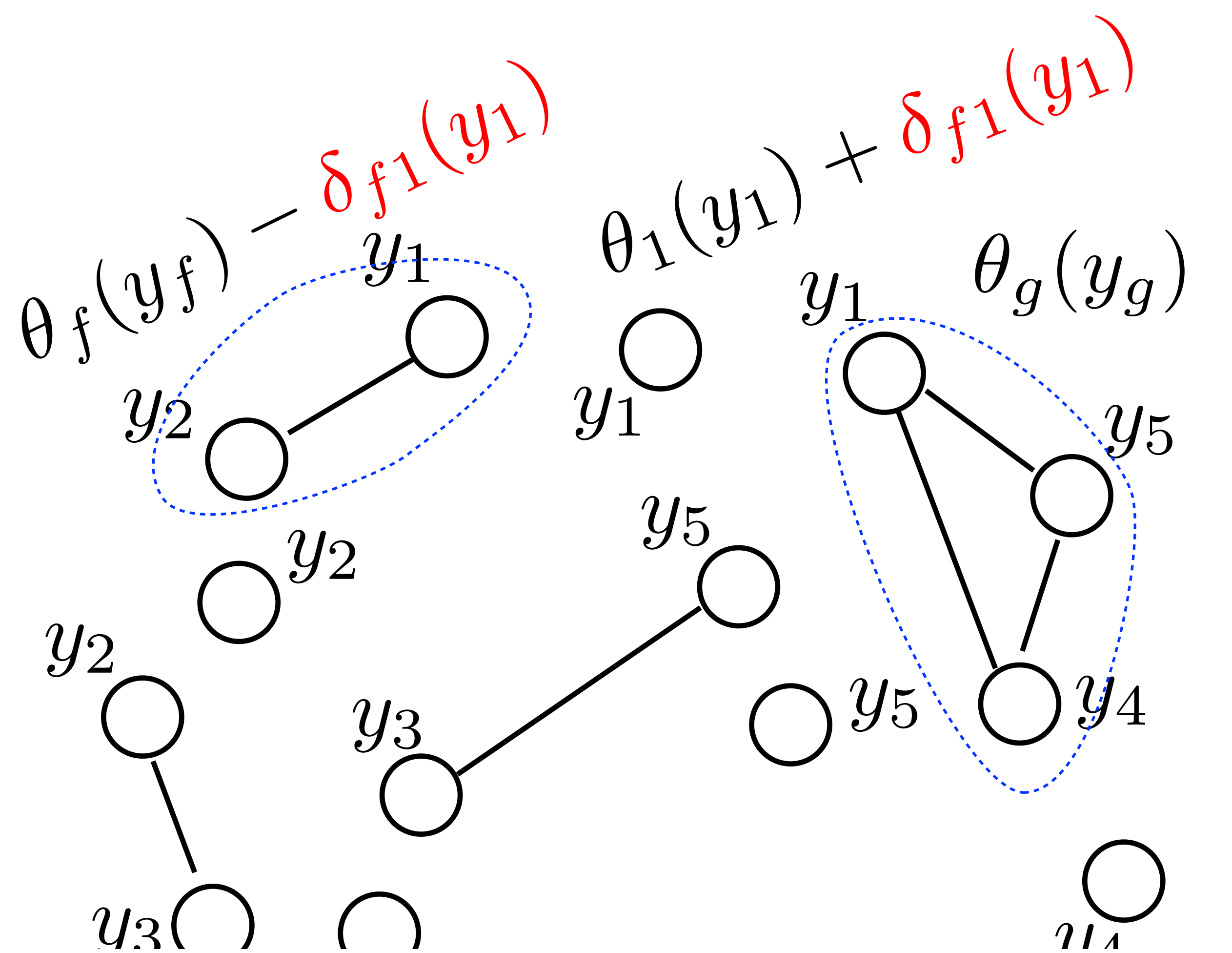
Dual decomposition

- We can think of the relaxation as enforcing that the components agree about the maximizing assignment

$$\max_{y} \Big\{ \sum_i \theta_i(y_i) + \sum_{f \in F} \theta_f(y_f) \Big\}$$

$$\le \sum_i \max_{y_i} \Big\{ \theta_i(y_i) + \sum_{f : i \in f} \delta_{fi}(y_i) \Big\} + \sum_{f \in F} \max_{y_f} \Big\{ \theta_f(y_f) - \sum_{i \in f} \delta_{fi}(y_i) \Big\}$$

- Agreements can be enforced via Lagrange multipliers associated with the equality constraints



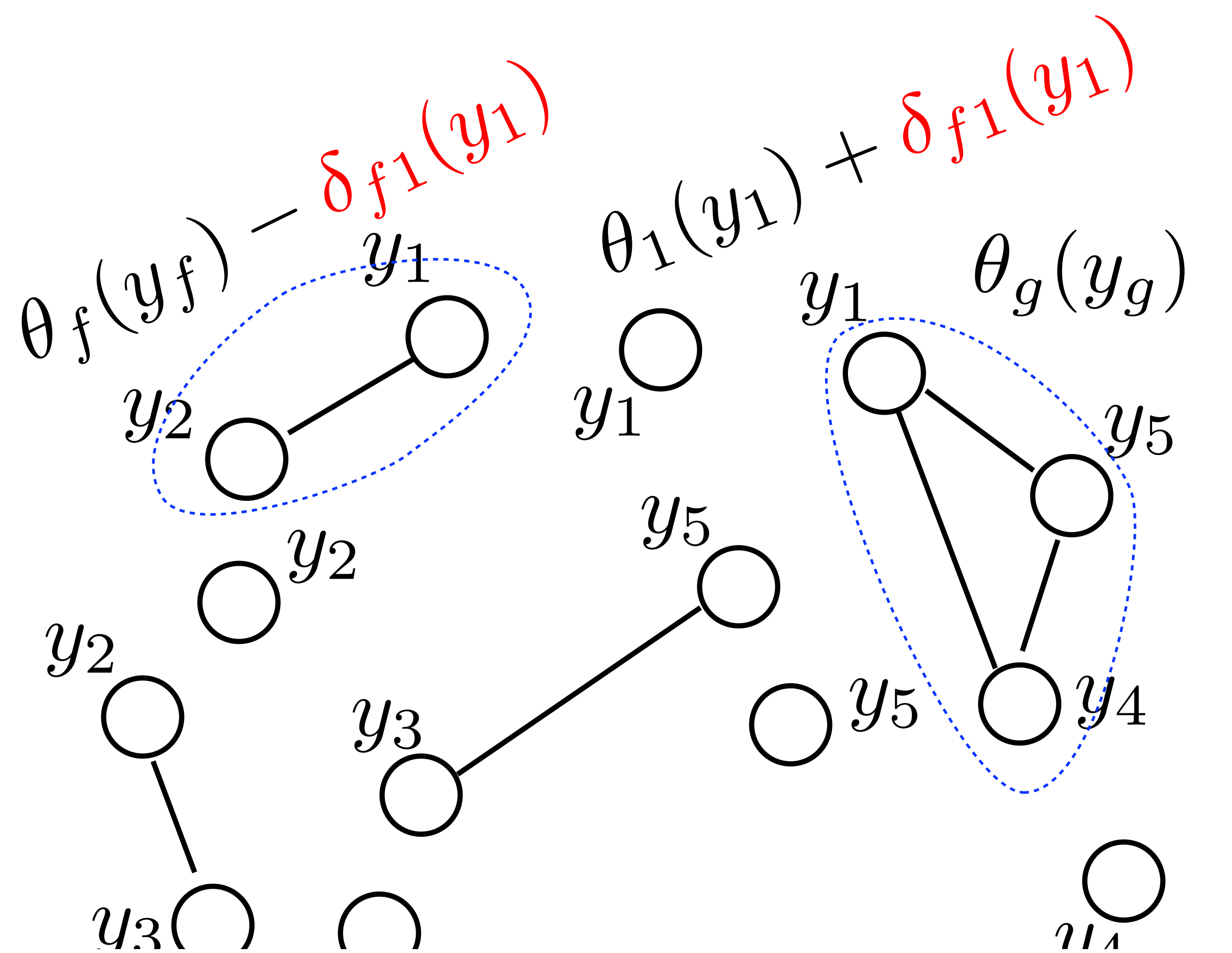
Dual decomposition

- We can think of the relaxation as enforcing that the components agree about the maximizing assignment

$$\max_y \Big\{ \sum_i \theta_i(y_i) + \sum_{f \in F} \theta_f(y_f) \Big\}$$

$$\leq \sum_i \max_{y_i} \Big\{ \theta_i(y_i) + \sum_{f: i \in f} \delta_{fi}(y_i) \Big\} + \sum_{f \in F} \max_{y_f} \Big\{ \theta_f(y_f) - \sum_{i \in f} \delta_{fi}(y_i) \Big\}$$

- Agreements can be enforced via Lagrange multipliers associated with the equality constraints
- Lagrange multipliers re-parameterize the model



Dual decomposition

- We minimize the dual objective

$$\max_y \Big\{ \sum_i \theta_i(y_i) + \sum_{f\in F} \theta_f(y_f) \Big\}$$

$$\leq \sum_i \max_{y_i} \Big\{ \theta_i(y_i) + \sum_{f:i\in f} \delta_{fi}(y_i) \Big\} + \sum_{f\in F} \max_{y_f} \Big\{ \theta_f(y_f) - \sum_{i\in f} \delta_{fi}(y_i) \Big\}$$

with respect to Lagrange multipliers associated with equality constraints

- this is an unconstrained convex minimization problem (cf. constrained primal maximization)
- minimization seeks to enforce agreements among the sub-problems
- a number of distributed algorithms have been developed for this purpose (e.g., Werner '07, Globerson et al., '08) **(2nd part)**

Weak duality

dual = $\sum_i \max_{y_i}\{\theta_i(y_i) + \sum_{f:i\in f} \delta_{fi}(y_i)\} + \sum_{f\in F} \max_{y_f}\{\theta_f(y_f) - \sum_{i\in f}\delta_{fi}(y_i)\}$

primal constraints:

Weak duality

dual = $\sum_i \max_{y_i} \{ \theta_i(y_i) + \sum_{f:i \in f} \delta_{fi}(y_i) \} + \sum_{f \in F} \max_{y_f} \{ \theta_f(y_f) - \sum_{i \in f} \delta_{fi}(y_i) \}$

$$\geq \sum_i \sum_{y_i} \mu_i(y_i) \{ \theta_i(y_i) + \sum_{f:i \in f} \delta_{fi}(y_i) \} +$$

primal constraints:

$$\mu_i(y_i) \geq 0, \; \sum_{y_i} \mu_i(y_i) = 1$$

Weak duality

dual = $\sum_i \max_{y_i} \{\theta_i(y_i) + \sum_{f:i\in f} \delta_{fi}(y_i)\} + \sum_{f\in F} \max_{y_f} \{\theta_f(y_f) - \sum_{i\in f} \delta_{fi}(y_i)\}$

$$\geq \sum_i \sum_{y_i} \mu_i(y_i) \{\theta_i(y_i) + \sum_{f:i\in f} \delta_{fi}(y_i)\} +$$

$$\sum_{f\in F} \sum_{y_f} \mu_f(y_f) \{\theta_f(y_f) - \sum_{i\in f} \delta_{fi}(y_i)\}$$

primal constraints:

$$\mu_i(y_i) \geq 0, \ \sum_{y_i} \mu_i(y_i) = 1 \qquad \mu_f(y_f) \geq 0, \ \sum_{y_f} \mu_f(y_f) = 1$$

Weak duality

$$\text{dual} = \sum_i \max_{y_i}\{\theta_i(y_i) + \sum_{f:i\in f} \delta_{fi}(y_i)\} + \sum_{f\in F} \max_{y_f}\{\theta_f(y_f) - \sum_{i\in f}\delta_{fi}(y_i)\}$$

$$\geq \sum_i \sum_{y_i} \mu_i(y_i)\{\theta_i(y_i) + \sum_{f:i\in f} \delta_{fi}(y_i)\} +$$

$$\sum_{f\in F} \sum_{y_f} \mu_f(y_f)\{\theta_f(y_f) - \sum_{i\in f}\delta_{fi}(y_i)\}$$

primal constraints:

$$\mu_i(y_i) \geq 0, \ \sum_{y_i} \mu_i(y_i) = 1 \qquad \mu_f(y_f) \geq 0, \ \sum_{y_f} \mu_f(y_f) = 1$$

$$\sum_{y_{f\setminus i}} \mu_f(y_f) = \mu_i(y_i), \ \forall i \in f$$

Weak duality

$$\text{dual} = \sum_i \max_{y_i} \{\theta_i(y_i) + \sum_{f:i\in f} \delta_{fi}(y_i)\} + \sum_{f\in F} \max_{y_f} \{\theta_f(y_f) - \sum_{i\in f} \delta_{fi}(y_i)\}$$

$$\geq \sum_i \sum_{y_i} \mu_i(y_i) \{\theta_i(y_i) + \sum_{f:i\in f} \cancel{\delta_{fi}(y_i)}\} +$$

$$\sum_{f\in F} \sum_{y_f} \mu_f(y_f) \{\theta_f(y_f) - \sum_{i\in f} \cancel{\delta_{fi}(y_i)}\}$$

primal constraints:

$$\mu_i(y_i) \geq 0, \ \sum_{y_i} \mu_i(y_i) = 1 \qquad \mu_f(y_f) \geq 0, \ \sum_{y_f} \mu_f(y_f) = 1$$

$$\sum_{y_{f\setminus i}} \mu_f(y_f) = \mu_i(y_i), \ \forall i \in f$$

Weak duality

$$\text{dual} = \sum_i \max_{y_i}\{\theta_i(y_i) + \sum_{f:i\in f}\delta_{fi}(y_i)\} + \sum_{f\in F}\max_{y_f}\{\theta_f(y_f) - \sum_{i\in f}\delta_{fi}(y_i)\}$$

$$\geq \sum_i \sum_{y_i}\mu_i(y_i)\{\theta_i(y_i) + \cancel{\sum_{f:i\in f}\delta_{fi}(y_i)}\} +$$

$$\sum_{f\in F}\sum_{y_f}\mu_f(y_f)\{\theta_f(y_f) - \cancel{\sum_{i\in f}\delta_{fi}(y_i)}\}$$

$$= \sum_i \sum_{y_i}\mu_i(y_i)\theta_i(y_i) + \sum_{f\in F}\sum_{y_f}\mu_f(y_f)\theta_f(y_f) \quad = \text{primal}$$

primal constraints:

$$\mu_i(y_i) \geq 0, \ \sum_{y_i}\mu_i(y_i) = 1 \qquad \mu_f(y_f) \geq 0, \ \sum_{y_f}\mu_f(y_f) = 1$$

$$\sum_{y_{f\setminus i}}\mu_f(y_f) = \mu_i(y_i), \ \forall i \in f$$

Weak duality

dual = $\sum_i \max_{y_i} \{\theta_i(y_i) + \sum_{f:i\in f} \delta_{fi}(y_i)\} + \sum_{f\in F} \max_{y_f} \{\theta_f(y_f) - \sum_{i\in f} \delta_{fi}(y_i)\}$

$$\geq \sum_i \sum_{y_i} \mu_i(y_i)\{\theta_i(y_i) + \sum_{f:i\in f} \cancel{\delta_{fi}(y_i)}\} +$$

$$\sum_{f\in F} \sum_{y_f} \mu_f(y_f)\{\theta_f(y_f) - \sum_{i\in f} \cancel{\delta_{fi}(y_i)}\}$$

$$= \sum_i \sum_{y_i} \mu_i(y_i)\theta_i(y_i) \quad + \quad \sum_{f\in F} \sum_{y_f} \mu_f(y_f)\theta_f(y_f)$$ = primal

primal constraints:

$$\mu_i(y_i) \geq 0,\ \sum_{y_i} \mu_i(y_i) = 1 \qquad \mu_f(y_f) \geq 0,\ \sum_{y_f} \mu_f(y_f) = 1$$

$$\sum_{y_{f\setminus i}} \mu_f(y_f) = \mu_i(y_i),\ \forall i \in f$$

Strong duality

- Min dual value = max primal value

$$\sum_i \max_{y_i} \{\theta_i(y_i) + \sum_{f:i\in f} \delta_{fi}(y_i)\} + \sum_{f\in F} \max_{y_f} \{\theta_f(y_f) - \sum_{i\in f} \delta_{fi}(y_i)\}$$

$$= \sum_i \sum_{y_i} \mu_i(y_i) \{\theta_i(y_i) + \sum_{f:i\in f} \delta_{fi}(y_i)\} +$$

$$\sum_{f\in F} \sum_{y_f} \mu_f(y_f) \{\theta_f(y_f) - \sum_{i\in f} \delta_{fi}(y_i)\}$$

- At the primal-dual optimum, equality follows from complementary slackness

$$\mu_i(y_i) > 0 \ \Rightarrow \ y_i \in \operatorname*{argmax}_{y_i} \{\theta_i(y_i) + \sum_{f:i\in f} \delta_{fi}(y_i)\}$$

$$\mu_f(y_f) > 0 \ \Rightarrow \ y_f \in \operatorname*{argmax}_{y_f} \{\theta_f(y_f) - \sum_{i\in f} \delta_{fi}(y_i)\}$$

Dual properties

- The dual objective

$$\max_y \Big\{ \sum_i \theta_i(y_i) + \sum_{f \in F} \theta_f(y_f) \Big\}$$

$$\leq \sum_i \max_{y_i} \Big\{ \theta_i(y_i) + \sum_{f: i \in f} \delta_{fi}(y_i) \Big\} + \sum_{f \in F} \max_{y_f} \Big\{ \theta_f(y_f) - \sum_{i \in f} \delta_{fi}(y_i) \Big\}$$

Dual properties

- The dual objective

$$\max_{y} \Big\{ \sum_{i} \theta_i(y_i) + \sum_{f \in F} \theta_f(y_f) \Big\}$$

$$\leq \sum_i \max_{y_i} \Big\{ \theta_i(y_i) + \sum_{f:i \in f} \delta_{fi}(y_i) \Big\} + \sum_{f \in F} \max_{y_f} \Big\{ \theta_f(y_f) - \sum_{i \in f} \delta_{fi}(y_i) \Big\}$$

- Suppose we find a maximizing assignment $\hat{y}$ that is the same for all the terms

$$\hat{y}_i \in \operatorname*{argmax}_{y_i} \Big\{ \theta_i(y_i) + \sum_{f:i \in f} \theta_{fi}(y_i) \Big\} \qquad \forall i$$

$$\hat{y}_f \in \operatorname*{argmax}_{y_f} \Big\{ \theta_f(y_f) - \sum_{i \in f} \theta_{fi}(y_f) \Big\} \qquad \forall f \in F$$

Dual properties

- The dual objective

$$\max_y \Big\{ \sum_i \theta_i(y_i) + \sum_{f \in F} \theta_f(y_f) \Big\}$$

$$\le \sum_i \cancel{\max_{y_i}} \Big\{ \theta_i(\hat{y}_i) + \sum_{f:i \in f} \delta_{fi}(\hat{y}_i) \Big\} + \sum_{f \in F} \cancel{\max_{y_f}} \Big\{ \theta_f(\hat{y}_f) - \sum_{i \in f} \delta_{fi}(\hat{y}_i) \Big\}$$

- Suppose we find a maximizing assignment $\hat{y}$ that is the same for all the terms

$$\hat{y}_i \in \operatorname{argmax}_{y_i} \Big\{ \theta_i(y_i) + \sum_{f:i \in f} \theta_{fi}(y_i) \Big\} \qquad \forall i$$

$$\hat{y}_f \in \operatorname{argmax}_{y_f} \Big\{ \theta_f(y_f) - \sum_{i \in f} \theta_{fi}(y_f) \Big\} \qquad \forall f \in F$$

Dual properties

- The dual objective

$$\max_y \Big\{ \sum_i \theta_i(y_i) + \sum_{f \in F} \theta_f(y_f) \Big\}$$

$$\leq \sum_i \Big\{ \theta_i(\hat{y}_i) + \sum_{f:i \in f} \delta_{fi}(\hat{y}_i) \Big\} + \sum_{f \in F} \Big\{ \theta_f(\hat{y}_f) - \sum_{i \in f} \delta_{fi}(\hat{y}_i) \Big\}$$

- Suppose we find a maximizing assignment $\hat{y}$ that is the same for all the terms

$$\hat{y}_i \in \operatorname*{argmax}_{y_i} \Big\{ \theta_i(y_i) + \sum_{f:i \in f} \delta_{fi}(y_i) \Big\} \qquad \forall i$$

$$\hat{y}_f \in \operatorname*{argmax}_{y_f} \Big\{ \theta_f(y_f) - \sum_{i \in f} \delta_{fi}(y_i) \Big\} \qquad \forall f \in F$$

Dual properties

- The dual objective

$$\max_y \Big\{ \sum_i \theta_i(y_i) + \sum_{f \in F} \theta_f(y_f) \Big\}$$

reparameterization

$$\leq \sum_i \Big\{ \theta_i(\hat{y}_i) + \cancel{\sum_{f:i \in f} \delta_{fi}(\hat{y}_i)} \Big\} + \sum_{f \in F} \Big\{ \theta_f(\hat{y}_f) - \cancel{\sum_{i \in f} \delta_{fi}(\hat{y}_i)} \Big\}$$

- Suppose we find a maximizing assignment $\hat{y}$ that is the same for all the terms

$$\hat{y}_i \in \operatorname*{argmax}_{y_i} \Big\{ \theta_i(y_i) + \sum_{f:i \in f} \delta_{fi}(y_i) \Big\} \qquad \forall i$$

$$\hat{y}_f \in \operatorname*{argmax}_{y_f} \Big\{ \theta_f(y_f) - \sum_{i \in f} \delta_{fi}(y_i) \Big\} \qquad \forall f \in F$$

Dual properties

- The dual objective

$$\max_y \Big\{ \sum_i \theta_i(y_i) + \sum_{f \in F} \theta_f(y_f) \Big\}$$

$$\leq \sum_i \theta_i(\hat{y}_i) + \sum_{f \in F} \theta_f(\hat{y}_f)$$

- Suppose we find a maximizing assignment $\hat{y}$ that is the same for all the terms

$$\hat{y}_i \in \operatorname*{argmax}_{y_i} \Big\{ \theta_i(y_i) + \sum_{f : i \in f} \delta_{fi}(y_i) \Big\} \qquad \forall i$$

$$\hat{y}_f \in \operatorname*{argmax}_{y_f} \Big\{ \theta_f(y_f) - \sum_{i \in f} \delta_{fi}(y_i) \Big\} \qquad \forall f \in F$$

Dual properties

- The dual objective

$$\max_y \Big\{ \sum_i \theta_i(y_i) + \sum_{f \in F} \theta_f(y_f) \Big\}$$

$$= \sum_i \theta_i(\hat{y}_i) + \sum_{f \in F} \theta_f(\hat{y}_f) \qquad \Rightarrow \hat{y} \text{ is a MAP assignment !}$$

- Suppose we find a maximizing assignment $\hat{y}$ that is the same for all the terms

$$\hat{y}_i \in \operatorname*{argmax}_{y_i} \Big\{ \theta_i(y_i) + \sum_{f : i \in f} \delta_{fi}(y_i) \Big\} \qquad \forall i$$

$$\hat{y}_f \in \operatorname*{argmax}_{y_f} \Big\{ \theta_f(y_f) - \sum_{i \in f} \delta_{fi}(y_i) \Big\} \qquad \forall f \in F$$

Dual properties

- We minimize the dual objective

$$\max_y \Big\{ \sum_i \theta_i(y_i) + \sum_{f \in F} \theta_f(y_f) \Big\}$$

$$\leq \sum_i \max_{y_i} \Big\{ \theta_i(y_i) + \sum_{f : i \in f} \delta_{fi}(y_i) \Big\} + \sum_{f \in F} \max_{y_f} \Big\{ \theta_f(y_f) - \sum_{i \in f} \delta_{fi}(y_i) \Big\}$$

with respect to Lagrange multipliers associated with equality constraints

- **Theorem (certificate)**: if the components agree about an assignment, the assignment is optimal

Dual properties

- We minimize the dual objective

$$\max_y \Big\{ \sum_i \theta_i(y_i) + \sum_{f \in F} \theta_f(y_f) \Big\}$$

$$\leq \sum_i \max_{y_i} \Big\{ \theta_i(y_i) + \sum_{f : i \in f} \delta_{fi}(y_i) \Big\} + \sum_{f \in F} \max_{y_f} \Big\{ \theta_f(y_f) - \sum_{i \in f} \delta_{fi}(y_i) \Big\}$$

with respect to Lagrange multipliers associated with equality constraints

- **Theorem (certificate)**: if the components agree about an assignment, the assignment is optimal
- We can obtain such an agreement **only** if the corresponding LP relaxation is tight **and** the Lagrange multipliers are set optimally

Dual decomposition: summary

$$\max_y \Big\{ \sum_i \theta_i(y_i) + \sum_{f \in F} \theta_f(y_f) \Big\}$$

$$\leq \sum_i \max_{y_i} \Big\{ \theta_i(y_i) + \sum_{f : i \in f} \delta_{fi}(y_i) \Big\} + \sum_{f \in F} \max_{y_f} \Big\{ \theta_f(y_f) - \sum_{i \in f} \delta_{fi}(y_i) \Big\}$$

- we try to solve the MAP problem by encouraging exactly solvable sub-problems to agree
- the dual objective is an *unconstrained* convex optimization problem (cf. primal)
- certificate of optimality = primal integrality
- the dual value always upper bounds the MAP value (cf. learning)
- dual value can guide search for tighter relaxations (e.g., Sontag et al. '08)