ECE 6504: Advanced Topics in Machine Learning

Probabilistic Graphical Models and Large-Scale Learning

Topics

- Markov Random Fields: **MAP** Inference
  - Max-Product Message Passing
  - Integer Programming, LP formulation
  - Dual Decomposition

Readings: KF 13.1-5, Barber 5.1,28.9

Dhruv Batra

Virginia Tech

Administrativia

- HW1 Solutions
  - Released
  - Grades almost done too

- Project Presentations
  - When: April 22, 24
  - Where: in class
  - 5 min talk
    - Main results
    - Semester completion 2 weeks out from that point so nearly finished results expected
    - Slides due: April 21 11:55pm

Recap of Last Time

Message Passing

- Variables/Factors "talk" to each other via messages:

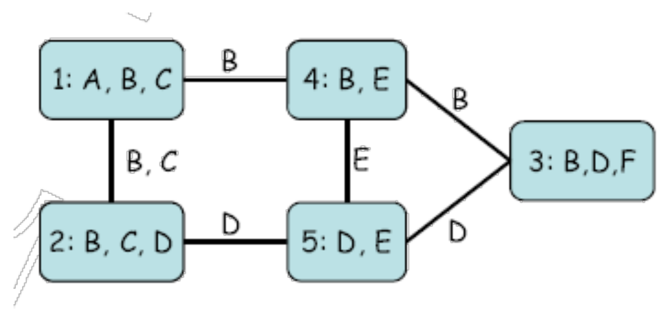
"I (variable X_3) think that you (variable X_2):

belong to state 1 with confidence 0.4
belong to state 2 with confidence 10
belong to state 3 with confidence 1.5"

Image Credit: James Coughlan

Generalized BP

- Initialization:
  - Assign each factor ϕ to a cluster $\alpha(\phi)$, Scope$[\phi] \subseteq \mathbf{C}_{\alpha(\phi)}$
  - Initialize cluster: $\psi_i^0(\mathbf{C}_i) \propto \prod_{\phi:\alpha(\phi)=i} \phi$
  - Initialize messages: $\delta_{j\to i} = 1$



- While not converged, send messages:

$$\delta_{i\to j}(\mathbf{S}_{ij}) \propto \sum_{\mathbf{C}_i - \mathbf{S}_{ij}} \psi_i^0(\mathbf{C}_i) \prod_{k\in\mathcal{N}(i)-j} \delta_{k\to i}(\mathbf{S}_{ik})$$

- Belief:
  - On board

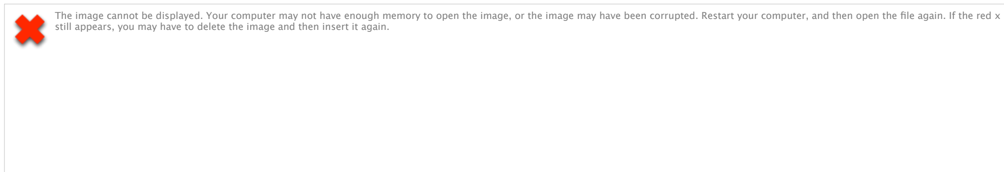
What is Variational Inference?

- A class of methods for approximate inference
  - And parameter learning
  - And approximating integrals basically..

- Key idea
  - Reality is complex
  - Instead of performing approximate computation in something complex
  - Can we perform exact computation in something “simple”?
  - Just need to make sure the simple thing is “close” to the complex thing.

- Key Problems
  - What is close?
  - How do we measure closeness when we can’t perform operations on the complex thing?

Variational Approximate Inference



- Choose a family of approximating distributions which is tractable. The simplest [Mean Field] Approximation:

$$q(x) = \prod_{s \in \mathcal{V}} q_s(x_s)$$

- Measure the quality of approximations. Two possibilities:

$$D(p \,||\, q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \qquad D(q \,||\, p) = \sum_x q(x) \log \frac{q(x)}{p(x)}$$

- Find the approximation minimizing this distance

D(p||q) for mean field – KL the right way

- D(p||q)=

- Trivially minimized by setting $q_i(x_i) = p_i(x_i)$
- Doesn't provide a computational method…

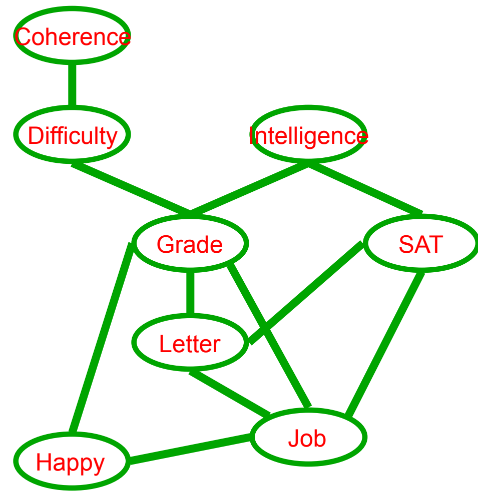
D(q||p) for mean field – KL the reverse direction

- D(q||p)=

$$D(q||p) = \sum_x q(x)\log q(x) - \sum_x q(x)\log p(x)$$

Reverse KL & The Partition Function

- D(q||p):
  - p is Markov net P_F



- **Theorem**: $$\log Z = F[p,q] + D(q||p)$$

- Where "Gibbs Free Energy":

$$F[p,q] = H_q(\mathcal{X}) + \mathbb{E}_q\left[\sum_c \log \psi_c(X_c)\right]$$

$$= H_q(\mathcal{X}) + \mathbb{E}_q\left[\text{Score}(\mathcal{X})\right]$$

$$= H_q(\mathcal{X}) + \sum_c \sum_{x_c} q(x_c)\theta(x_c)$$

Understanding Reverse KL, Free Energy & The Partition Function

$$\log Z = F[p,q] + D(q||p)$$

$$F[p,q] = H_q(\mathcal{X}) + \mathbb{E}_q\left[\sum_c \log \psi_c(X_c)\right]$$

- Maximizing Energy Functional ⇔ Minimizing Reverse KL

- **Theorem**: Energy Function is lower bound on partition function
  - Maximizing energy functional corresponds to search for tight lower bound on partition function

Mean Field Equations

$$F[p,q] = H_q(\mathcal{X}) + \mathbb{E}_q\left[\sum_c \log \psi_c(X_c)\right]$$

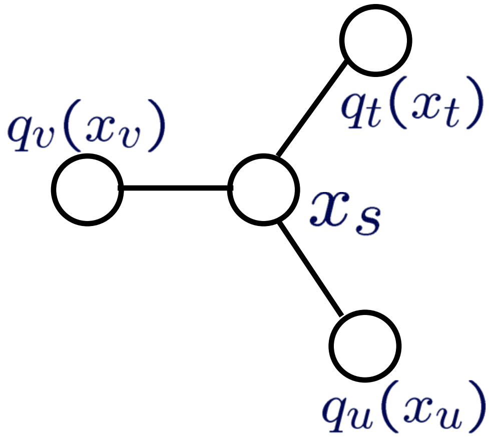
$$H(q) = \sum_{s\in\mathcal{V}} H_s(q_s) = -\sum_{s\in\mathcal{V}}\sum_{x_s} q_s(x_s)\log q_s(x_s)$$

$$\sum_c \sum_{x_c} q_c(x_c)\theta(x_c) = \sum_i \sum_{x_i} q_i(x_i)\theta_i(x_i) + \sum_{(i,j)\in E}\sum_{x_i}\sum_{x_j} q_i(x_i)q_j(x_j)\theta_{ij}(x_i,x_j)$$

- Add Lagrange multipliers to enforce $\sum_{x_s} q_s(x_s) = 1$

- Taking derivatives and simplifying, we find a set of fixed point equations:

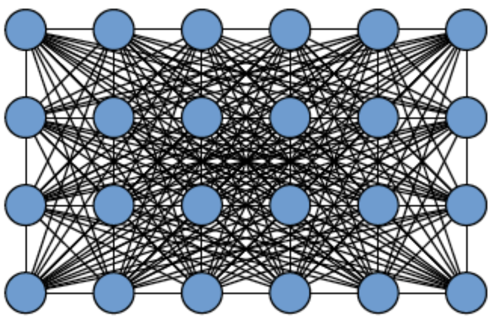
$$q_i(x_i) \propto \psi_i(x_i) \prod_{j\in N(i)} \exp\left\{\sum_{x_j} \theta_{ij}(x_i,x_j)q_j(x_j)\right\}$$



- Updating one marginal at a time gives convergent coordinate descent

Fully connected CRF

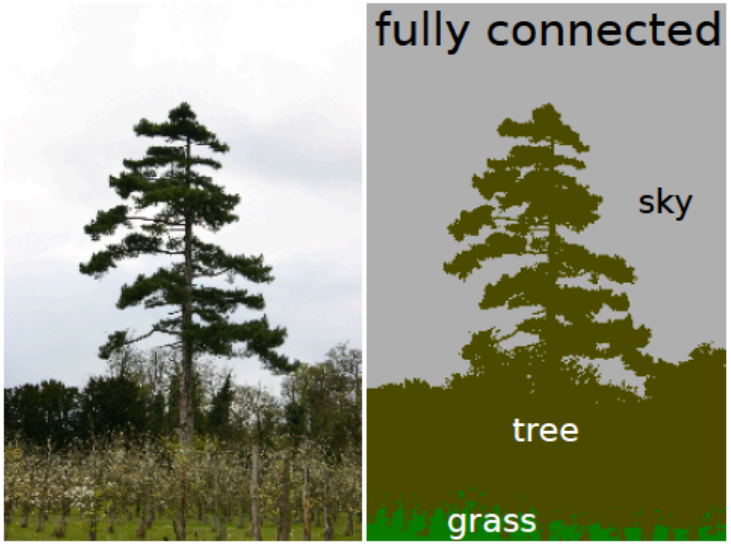
$$E(\mathbf{x}) = \sum_i \underbrace{\psi_u(x_i)}_{\text{unary term}} + \sum_i \sum_{j>i} \underbrace{\psi_p(x_i, x_j)}_{\text{pairwise term}}$$



- Every node is connected to every other node
  - Connections weighted differently

Fully connected CRF

$$E(\mathbf{x}) = \sum_i \underbrace{\psi_u(x_i)}_{\text{unary term}} + \sum_i \sum_{j>i} \underbrace{\psi_p(x_i, x_j)}_{\text{pairwise term}}$$



- Long-range interactions
- No more shrinking bias

What you need to know about variational methods

- Structured Variational method:
  - select a form for approximate distribution
  - minimize reverse KL

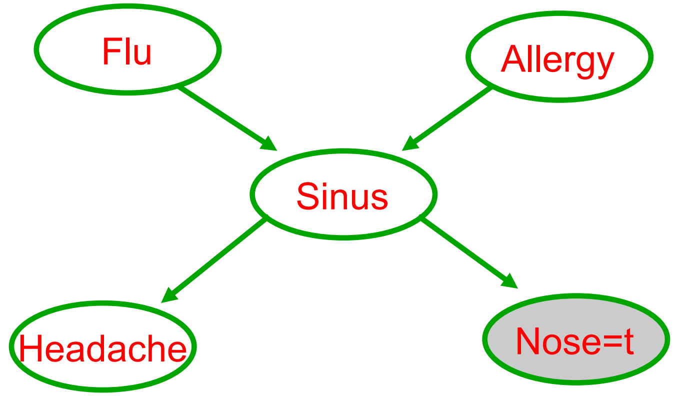
- Equivalent to maximizing energy functional
  - searching for a tight lower bound on the partition function

- Many possible models for *Q*:
  - independent (mean field)
  - structured as a Markov net
  - cluster variational
- Several subtleties outlined in the book

Plan for today

- MRF Inference
  - (Specialized) MAP Inference
    - Integer Programming Formulation
    - Linear Programming Relaxation
    - Dual Decomposition

Possible Queries



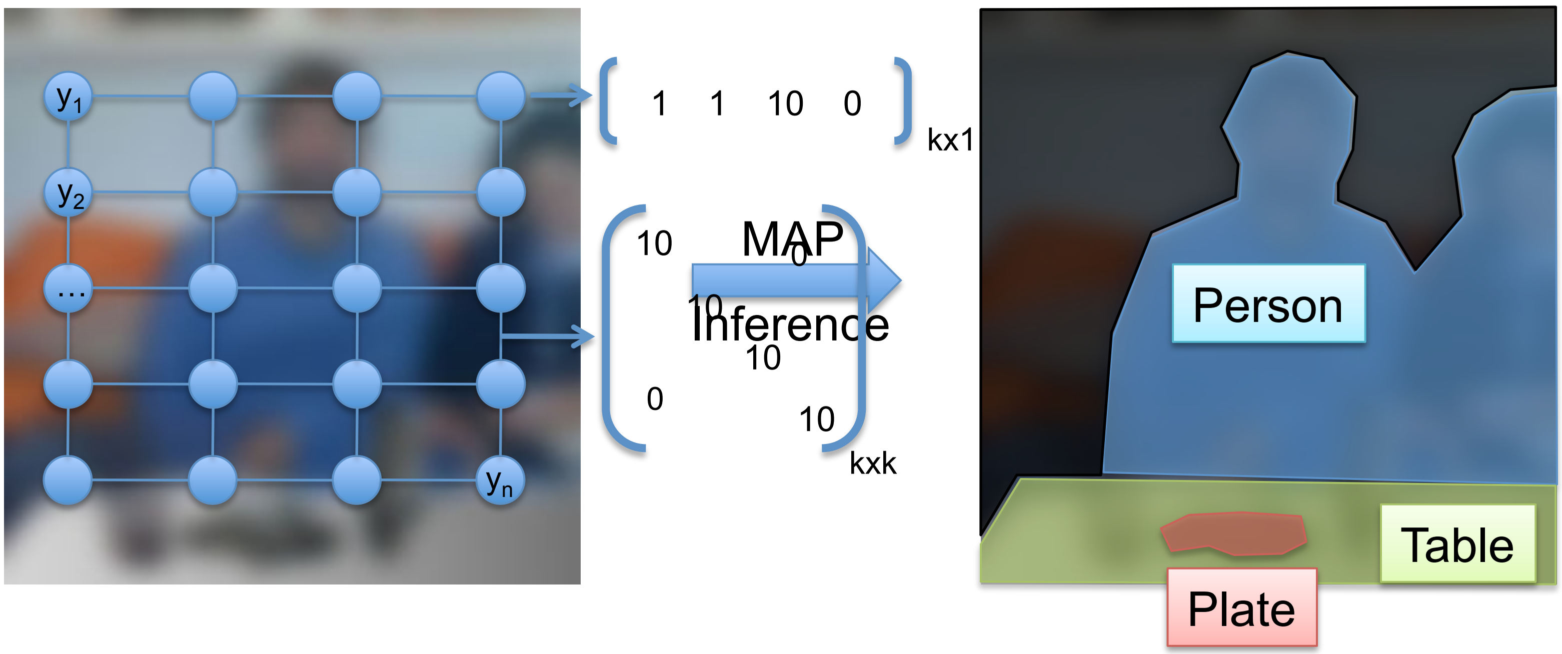
- Evidence: **E**=**e** (e.g. N=t)
- Query variables of interest **Y**

- Conditional Probability: P(**Y** | **E**=**e**)
  - E.g. P(F,A | N=t)
  - Special case: Marginals P(F)

- Maximum a Posteriori: argmax P(All variables | **E**=**e**)
  - argmax_{f,a,s,h} P(f,a,s,h | N = t)

- Marginal-MAP: argmax_y P(**Y** | **E**=**e**)
  - = argmax_{y} Σ_o P(**Y**=**y**, **O**=**o** | **E**=**e**)

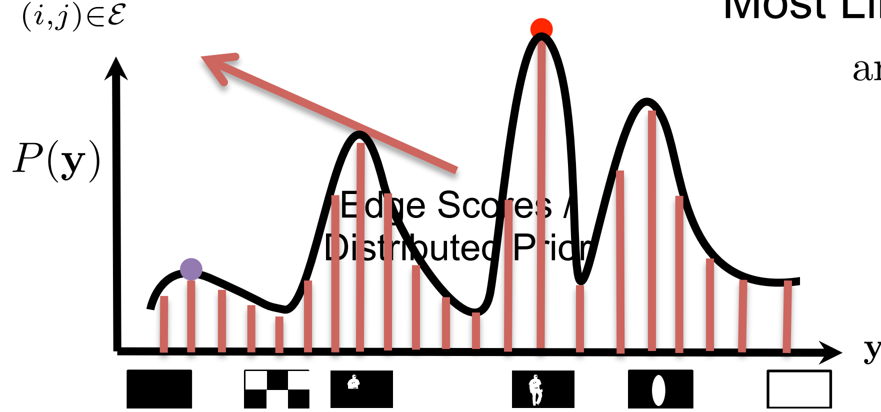
MAP Inference



$$S(\mathbf{y}) = \sum_{i \in \mathcal{V}} \theta_i(y_i) + \sum_{(i,j) \in \mathcal{E}} \theta_{ij}(y_i, y_j)$$

$$P(\mathbf{y}) = \frac{1}{Z} e^{S(\mathbf{y})}$$

Node Scores / Local Rewards

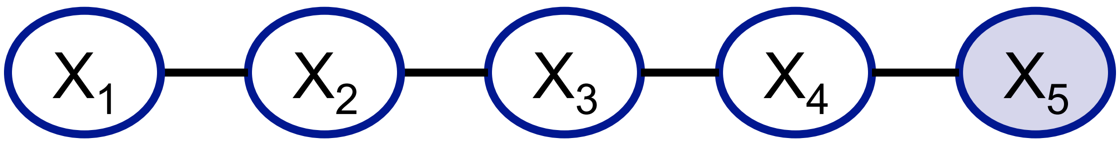


Most Likely Assignment

$$\underset{\mathbf{y}}{\operatorname{argmax}}\ S(\mathbf{y})$$

Example

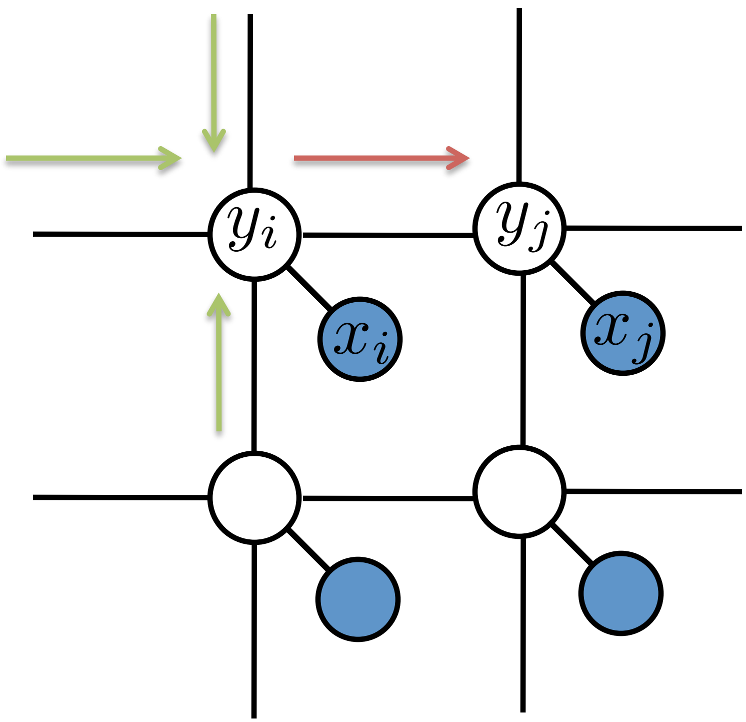
- Chain MRF



- Max-Product VE steps on board

Loopy BP on Pairwise Markov Nets

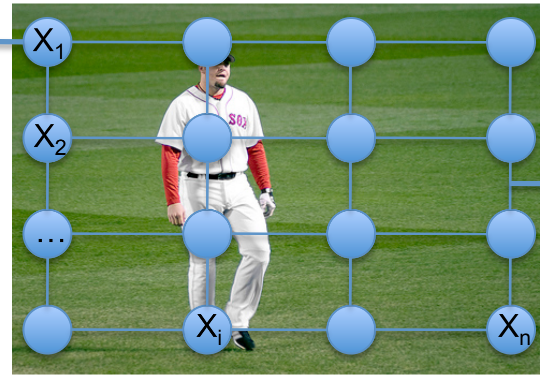
$$\delta_{i \to j}(y_j) = \sum_{y_i} \phi_i(y_i)\phi_{ij}(y_i, y_j) \prod_{k \in \mathcal{N}(i)-j} \delta_{k \to i}(y_i)$$



MAP in Pairwise MRFs

$$G = (\mathcal{V}, \mathcal{E})$$

$$\begin{bmatrix} \theta_{ij}(1,1) & & \theta_{ij}(1,k) \\ & \cdots & \\ \theta_{ij}(k,1) & \cdots & \theta_{ij}(k,k) \end{bmatrix} \quad \begin{bmatrix} \theta_1(1) \\ \cdots \\ \theta_1(k) \end{bmatrix}_{kx1}$$



$$\begin{bmatrix} \theta_{ij}(1,1) & & \theta_{ij}(1,k) \\ & \cdots & \\ \theta_{ij}(k,1) & \cdots & \theta_{ij}(k,k) \end{bmatrix}_{kxk} \quad \begin{bmatrix} \theta_n(1) \\ \cdots \\ \theta_n(k) \end{bmatrix}$$

- Over-Complete Representation

$$\boxed{x_1} \quad \cdots \quad \boxed{x_n} \quad \boxed{(x_1, x_2)} \quad \cdots \quad \boxed{(x_{n-1}, x_n)}$$

$$\boldsymbol{\theta} = \big(\theta_1(1)\ldots\theta_1(k) \quad \theta_n(1)\ldots\theta_n(k) \quad \theta_{12}(1,1)\ldots\theta_{12}(k,k) \quad \theta_{n-1,n}(1,1)\ldots\theta_{n-1,n}(k,k)\big)$$

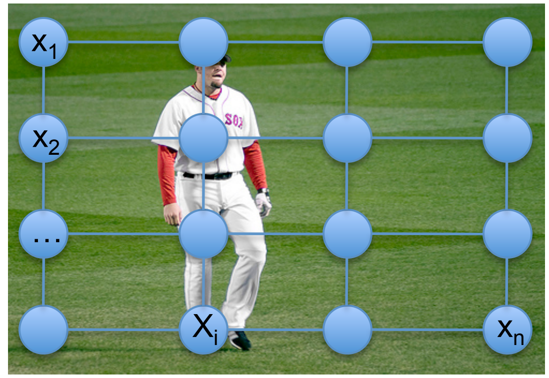
$$\mu_1(1)\ldots\mu_1(k) \qquad \mu_n(1)\ldots\mu_n(k)$$

$$x_i \Rightarrow \begin{bmatrix} \cdot \\ \mu_i(s) \\ \cdot \\ \cdot \end{bmatrix}_{kx1}$$

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}_{x_i = 1} \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}_{x_i = 2} \qquad \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

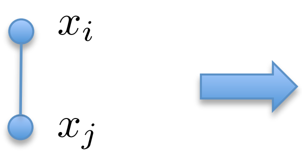
MAP in Pairwise MRFs

$G = (\mathcal{V}, \mathcal{E})$



- Over-Complete Representation

$$
\begin{array}{ccccccc}
 & \boxed{x_1} & \cdots & \boxed{x_n} & \boxed{(x_1, x_2)} & \cdots & \boxed{(x_{n-1}, x_n)} \\
\boldsymbol{\theta} = & \big[\theta_1(1)\ldots\theta_1(k) & & \theta_n(1)\ldots\theta_n(k) & \theta_{12}(1,1)\ldots\theta_{12}(k,k) & & \theta_{n-1,n}(1,1)\ldots\theta_{n-1,n}(k,k)\big] \\
\boldsymbol{\mu}_{\mathbf{x}} = & \big[\mu_1(1)\ldots\mu_1(k) & & \mu_n(1)\ldots\mu_n(k) & \mu_{12}(1,1)\ldots\mu_{12}(k,k) & & \mu_{n-1,n}(1,1)\ldots\mu_{n-1,n}(k,k)\big]
\end{array}
$$



$$
\begin{bmatrix} \vdots \\ \mu_{ij}(s,t) \\ \vdots \end{bmatrix}_{k^2\text{x}1}
\qquad
\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}
\begin{array}{l} x_i = 1 \\ x_j = 1 \end{array}
\qquad
\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}
\begin{array}{l} x_i = 1 \\ x_j = 2 \end{array}
$$

$$S(\mathbf{x}) = \boldsymbol{\theta} \cdot \boldsymbol{\mu}_{\mathbf{x}}$$

MAP in Pairwise MRFs

- Integer Program

$$\max_{\boldsymbol{\mu}} \quad \boldsymbol{\theta}^T \boldsymbol{\mu}$$

$$\left.\begin{aligned} \mu_i(s) &\in \{0,1\} \\ \mu_{ij}(s,t) &\in \{0,1\} \end{aligned}\right\} \quad \text{Indicator Variables}$$

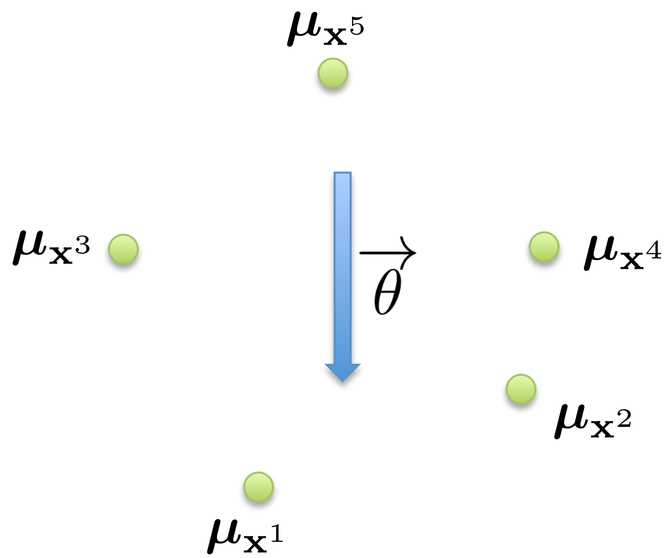
$$\left.\begin{aligned} \sum_s \mu_i(s) &= 1 \\ \sum_{s,t} \mu_{i,j}(s,t) &= 1 \end{aligned}\right\} \quad \text{Unique Label}$$

$$\left.\sum_s \mu_{ij}(s,t) = \mu_j(t) \right\} \quad \text{Consistent Assignments}$$

MAP in Pairwise MRFs

- MAP Integer Program

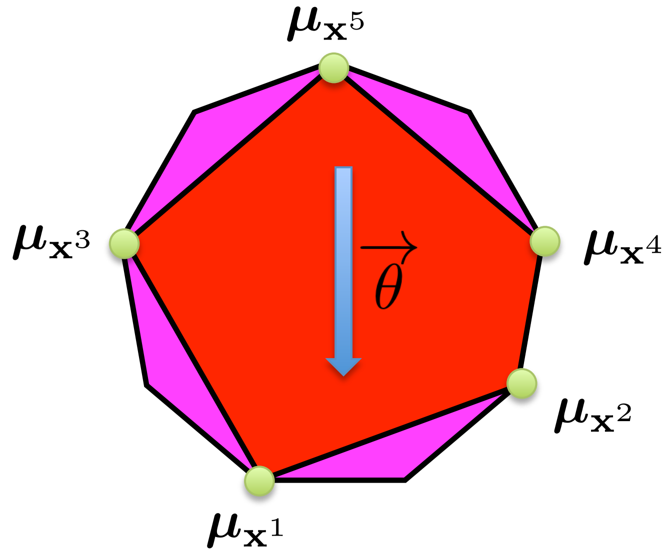
$$
\begin{aligned}
\max_{\boldsymbol{\mu}} \quad & \boldsymbol{\theta}^T \boldsymbol{\mu} \\
s.t. \quad & A\boldsymbol{\mu} = b \\
& \boldsymbol{\mu}(\cdot) \in \{0, 1\}
\end{aligned}
$$



MAP in Pairwise MRFs

- MAP Linear Program

$$\begin{aligned} \max_{\boldsymbol{\mu}} \quad & \boldsymbol{\theta}^T \boldsymbol{\mu} \\ s.t. \quad & A\boldsymbol{\mu} = b \\ & \boldsymbol{\mu}(\cdot) \in [0, 1] \end{aligned}$$



- Properties
  - If LP-opt is integral, MAP is found
  - LP always integral for trees
  - Efficient message-passing schemes for solving LP

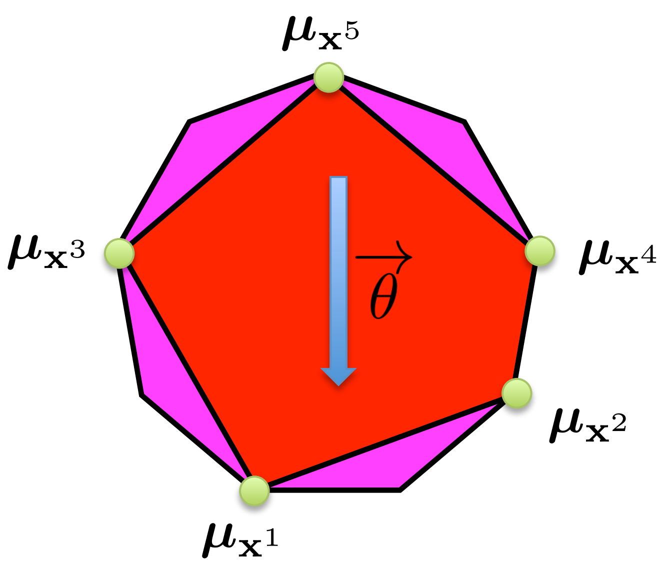
MAP in Pairwise MRFs

- Compare MAP LP to Variational Inference
  - On board
  - Difference in entropy term (objective)
  - Family of Q distributions

MAP in Pairwise MRFs

- MAP Linear Program

$$\max_{\boldsymbol{\mu}} \quad \boldsymbol{\theta}^T \boldsymbol{\mu}$$
$$s.t. \quad A\boldsymbol{\mu} = b$$
$$\boldsymbol{\mu}(\cdot) \in [0, 1]$$

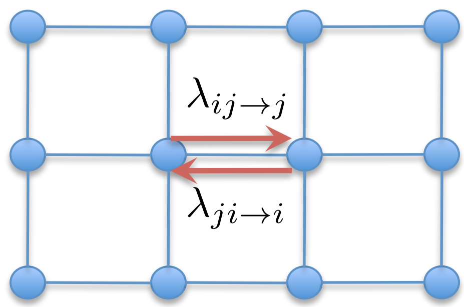


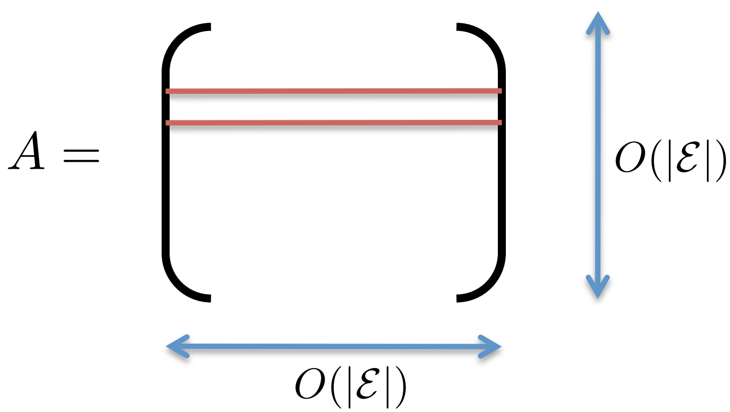
$$A = \underbrace{\left[\qquad \right]}_{O(|\mathcal{E}|)} \; O(|\mathcal{E}|)$$

Off-the-shelf solvers
CPLEX
Mosek
etc

LP Relaxation

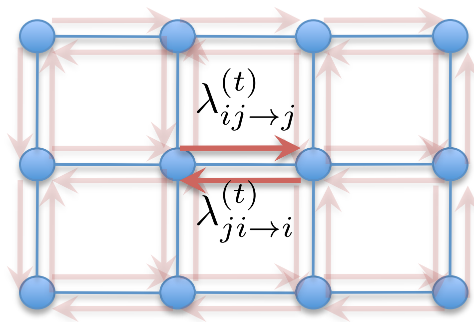
- Block Co-ordinate / Sub-gradient Descent on Dual



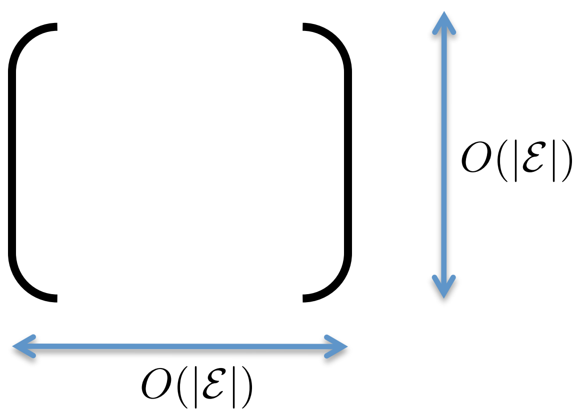


LP Relaxation

- Block Co-ordinate / Sub-gradient Descent on Dual

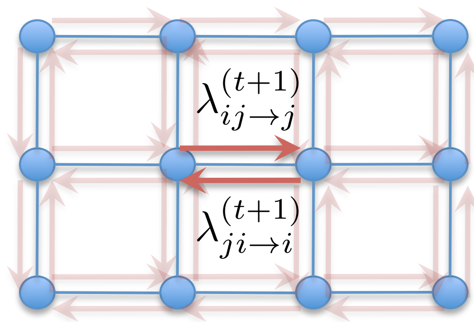


$A =$



LP Relaxation

- Block Co-ordinate / Sub-gradient Descent on Dual



Distributed Message-Passing

Still inefficient!