

# Combinatorial Optimization in Computer Vision (IN2245)

Frank R. Schmidt  
Csaba Domokos

Winter Semester 2015/2016

10. Tree-reweighted Message Passing & Mean Field Methods . . . . .	2
<b>Tree-reweighted message passing</b>	<b>3</b>
Introduction . . . . .	4
Canonical overcomplete representation . . . . .	5
Revisit the Max-sum algorithm . . . . .	6
Revisit the Max-sum algorithm . . . . .	7
Reparameterization . . . . .	8
Normal form . . . . .	9
LP relaxation . . . . .	10
Convex combinations of trees . . . . .	11
Convex combinations of trees . . . . .	12
New Tree-reweighting message passing . . . . .	13
Weak tree agreement . . . . .	14
TRW-S algorithm . . . . .	15
<b>Mean Field methods</b>	<b>16</b>
KL divergence . . . . .	17

Motivation . . . . . 18  
Mean Field methods . . . . . 19  
Gibbs inequality . . . . . 20  
Naive mean field . . . . . 21  
Naive mean field . . . . . 22  
Naive Mean Field . . . . . 23  
Optimization . . . . . 24  
Lagrange multipliers . . . . . 25  
Lagrange multipliers . . . . . 26  
Update equation . . . . . 27  
Semantic segmentation . . . . . 28  
Energy functions . . . . . 29  
Inference . . . . . 30  
Literature. . . . . 31

## 10. Tree-reweighted Message Passing & Mean Field Methods

2 / 31

### Tree-reweighted message passing

3 / 31

#### Introduction

Assume an undirected (pairwise) graphical model  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  with the following energy function:

$$E(\mathbf{y}) = \text{const} + \sum_{i \in \mathcal{V}} E_i(y_i) + \sum_{(i,j) \in \mathcal{E}} E_{ij}(y_i, y_j). \quad (1)$$

For each  $i \in \mathcal{V}$ , let  $Y_i$  be a random variable taking values from a (finite) set  $\mathcal{Y}_i$ , therefore  $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n$ .

Let us introduce the following notations

- $E_i(a) \triangleq \theta_{i;a}$ , which is a vector of size  $|\mathcal{Y}_i|$ .
- $E_{ij}(a, b) \triangleq \theta_{ij;ab}$ , which is a vector of size  $|\mathcal{Y}_i \times \mathcal{Y}_j|$ . Note that  $\theta_{ij;ab} \equiv \theta_{ji;ab}$ .

One can consider

$$\theta = \{\theta_\alpha \mid \alpha \in \mathcal{I}\} \in \mathbb{R}^d$$

as a vector, where  $\mathcal{I} = \{\text{const}\} \cup \{(i; a)\} \cup \{(ij; ab)\}$ .

### Canonical overcomplete representation

The energy function (1) can be written (with equivalent notations) as

$$E(\mathbf{y}; \theta) = \theta_{\text{const}} + \sum_{i \in \mathcal{V}} \theta_{i; y_i} + \sum_{(i,j) \in \mathcal{E}} \theta_{ij; y_i y_j} .$$

We introduce a mapping  $\phi : \mathcal{Y} \rightarrow \mathbb{R}^d$  so that

$$E(\mathbf{y}; \theta) = \langle \theta, \phi(\mathbf{y}) \rangle = \sum_{\alpha \in \mathcal{I}} \theta_{\alpha} \phi_{\alpha}(\mathbf{y}) .$$

The mapping  $\phi$  is called the **canonical overcomplete representation** consists of the following functions  $\phi_{\alpha} : \mathcal{Y} \rightarrow \mathbb{R}$ :

$$\begin{aligned} \phi_{\text{const}}(\mathbf{y}) &= 1 \\ \phi_{i;a}(\mathbf{y}) &= \llbracket y_i = a \rrbracket \\ \phi_{ij;ab}(\mathbf{y}) &= \llbracket y_i = a, y_j = b \rrbracket . \end{aligned}$$

## Revisit the Max-sum algorithm

Reminder: the *Max-sum algorithm* solves the following optimization problem:

$$y^* \in \operatorname{argmax}_{y \in \mathcal{Y}} p(y) = \operatorname{argmax}_{y \in \mathcal{Y}} \frac{1}{Z} \exp \left( \sum_{F \in \mathcal{F}} -E_F(y_F) \right) = \operatorname{argmin}_{y \in \mathcal{Y}} \sum_{F \in \mathcal{F}} E_F(y_F) .$$

It maintains messages  $M_{ij} = \{M_{ij;a} \mid a \in \mathcal{Y}_j\}$  for each  $(i, j) \in \mathcal{E}$ , where

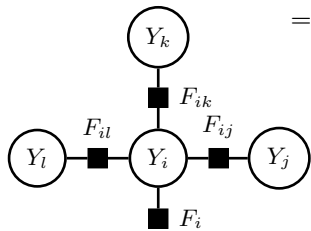
$$M_{ij;b} := \min_{a \in \mathcal{Y}_i} \left\{ \theta_{i;a} + \sum_{(s,i) \in \mathcal{E}, s \neq j} M_{si;a} \right\} + \theta_{ij;ab} .$$

$M = \{M_{ij}\}$  denotes the vector of all messages.

## Revisit the Max-sum algorithm

Assuming the following factor graph, let us calculate the message  $r_{F_{ij} \rightarrow Y_j}$ .

$$\begin{aligned}
 r_{F_{ij} \rightarrow Y_j}(y_j) &= \min_{y_i \in \mathcal{Y}_i} \{E_{ij}(y_i, y_j) + q_{Y_i \rightarrow F_{ij}}(y_i)\} \\
 &= \min_{y_i \in \mathcal{Y}_i} \left\{ E_{ij}(y_i, y_j) + \sum_{F \in M(i) \setminus \{F_{ij}, F_i\}} r_{F \rightarrow Y_i}(y_i) + r_{F_i \rightarrow Y_i}(y_i) \right\} \\
 &= \min_{y_i \in \mathcal{Y}_i} \left\{ (E_i(y_i) + \sum_{F \in M(i) \setminus \{F_{ij}, F_i\}} r_{F \rightarrow Y_i}(y_i)) + E_{ij}(y_i, y_j) \right\} \\
 &= \min_{y_i \in \mathcal{Y}_i} \left\{ (\theta_{i; y_i} + \sum_{s \in N(F) \setminus \{i, j\}, F \in M(i)} M_{s; y_i}) + \theta_{ij; y_i y_j} \right\} = M_{ij; y_j} .
 \end{aligned}$$



## Reparameterization

Assuming two parameterization  $\theta$  and  $\bar{\theta}$ , if they define the same energy function, i.e.  $E(\mathbf{y}; \theta) = E(\mathbf{y}; \bar{\theta})$  for all  $\mathbf{y} \in \mathcal{Y}$ , denoted by  $\theta \equiv \bar{\theta}$ , then  $\theta$  is called a **reparameterization** of  $\bar{\theta}$ .

Note that this condition does not necessarily imply that  $\theta = \bar{\theta}$ . Indeed, any message vector  $M = \{M_{st}\}$  defines reparameterization  $\theta = \bar{\theta}[M]$  as follows:

$$\begin{aligned}\theta_i &= \bar{\theta}_i + \sum_{(i,j) \in \mathcal{E}} M_{ij} \\ \theta_{ij;ab} &= \bar{\theta}_{ij;ab} - M_{ij;b} - M_{ji;a} \\ \theta_{\text{const}} &= \bar{\theta}_{\text{const}}\end{aligned}$$

In belief propagation (BP) we can alternatively store the reparameterization  $\theta = \bar{\theta}[M]$  instead of  $\bar{\theta}$  and  $M$ . Namely, sending a message from node  $i$  to  $j$  is equivalent to reparameterizing vectors  $\theta_i$  and  $\theta_{ij}$ .

## Normal form

A message for an edge  $(i, j) \in \mathcal{E}$  is called **valid** if any update does not change  $M_{ij}$ . A message for  $(i, j) \in \mathcal{E}$  is valid iff

$$\min_{a \in \mathcal{Y}_i} \{\theta_{i;a} + \theta_{ij;ab}\} = \text{const}_{ij} \quad \forall b \in \mathcal{Y}_j .$$

That is a message from  $s$  to  $t$  does not change  $\theta_{ij}$  and  $\theta_i$ . We say that  $\theta$  is in a **normal form** if all messages are valid. Minimum value of the energy is given by  $\Phi(\theta) = \min_{\mathbf{y} \in \mathcal{Y}} E(\mathbf{y}; \theta)$  and the min-marginals for nodes and edges are given by

$$\Phi_{i;a}(\theta) = \min_{\mathbf{y} \in \mathcal{Y}, y_i=a} E(\mathbf{y}; \theta) \quad \text{and} \quad \Phi_{ij;ab}(\theta) = \min_{\mathbf{y} \in \mathcal{Y}, y_i=a, y_j=b} E(\mathbf{y}; \theta) .$$

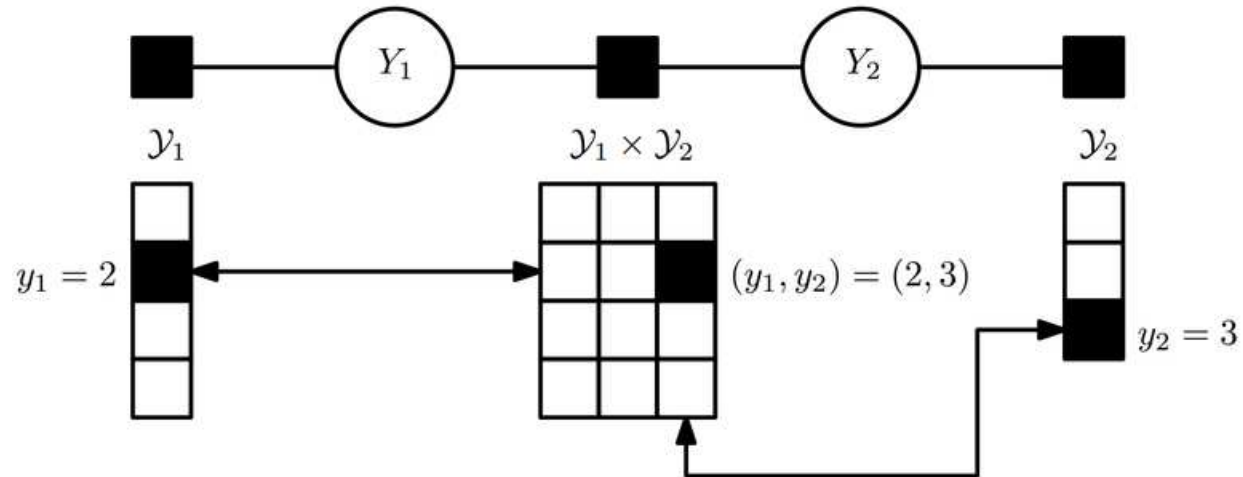
For a tree-structured graph the values  $\theta_{i;a}$  and  $\theta_{ij;ab}$  for vector  $\theta$  in a normal correspond to min-marginals (up to a constant):

$$\begin{aligned}\Phi_{i;a}(\theta) &= \theta_{i;a} && + \text{const}_i \\ \Phi_{ij;ab}(\theta) &= \{\theta_{i;a} + \theta_{ij;ab} + \theta_{j;b}\} && + \text{const}_{ij}\end{aligned} \tag{2}$$

## LP relaxation

In general, energy minimization (1) is NP-hard. Therefore, one can focused on approximation, such as **linear programming relaxation**. The constraint set is as follows:

$$\mathcal{L}(\mathcal{G}) = \left\{ \tau \in \mathbb{R}_+^d \mid \tau_{\text{const}} = 1, \sum_{a \in \mathcal{V}} \tau_{i;a} = 1, \sum_{a \in \mathcal{V}} \tau_{ij;ab} = \tau_{j;b} \right\}$$



The following minimization problem yields a lower bound on  $\Phi(\bar{\theta})$ :

$$\min_{\tau \in \mathcal{L}(\mathcal{G})} \langle \bar{\theta}, \tau \rangle .$$

(3)



### Convex combinations of trees

We need to introduce some notation. Let  $\mathcal{T}$  be a collection of trees in graph  $\mathcal{G}$  and  $\rho^T > 0$ ,  $T \in \mathcal{T}$  be some distribution on  $\mathcal{T}$ . We assume that each edge is covered by at least one tree.

For a given tree  $T = (\mathcal{V}^T, \mathcal{E}^T)$  we define a set of indices associated with vertices and edges in the tree:

$$\mathcal{I}^T = \{\text{const}\} \cup \{(i; a) \mid i \in \mathcal{V}^T\} \cup \{(ij; ab) \mid (i, j) \in \mathcal{E}^T\}.$$

To each tree  $T \in \mathcal{T}$ , we associate an energy parameter  $\theta^T$  belonging to the following linear constraint set:

$$\mathcal{A}^T = \{\theta^T \in \mathbb{R}^d \mid \theta_\alpha^T = 0 \forall \alpha \in \mathcal{I} \setminus \mathcal{I}^T\}.$$

By concatenating all of the tree vectors, we get a vector  $\boldsymbol{\theta} = \{\theta^T \mid T \in \mathcal{T}\} \in \mathbb{R}^{d \times |\mathcal{T}|}$  belonging to the constraint set

$$\mathcal{A} = \{\boldsymbol{\theta} \in \mathbb{R}^{d \times |\mathcal{T}|} \mid \theta^T \in \mathcal{A}^T \text{ for all } T \in \mathcal{T}\}.$$

## Convex combinations of trees

Consider function  $\Phi_\rho : \mathcal{A} \rightarrow \mathbb{R}$  defined as follows:

$$\Phi_\rho(\boldsymbol{\theta}) = \sum_T \rho^T \Phi(\theta^T) = \sum_T \rho^T \min_{\mathbf{y} \in \mathcal{Y}} \langle \theta^T, \phi(\mathbf{y}) \rangle .$$

Let  $\bar{\boldsymbol{\theta}} = \sum_T \rho^T \theta^T$ , then

$$\begin{aligned} \Phi_\rho(\boldsymbol{\theta}) &= \sum_T \rho^T \Phi(\theta^T) = \mathbb{E}[\Phi(\theta^T)] \leq \Phi(\mathbb{E}[\theta^T]) = \min_{\mathbf{y} \in \mathcal{Y}} \langle \mathbb{E}[\theta^T], \phi(\mathbf{y}) \rangle \\ &= \min_{\mathbf{y} \in \mathcal{Y}} \langle \sum_T \rho^T \theta^T, \phi(\mathbf{y}) \rangle = \min_{\mathbf{y} \in \mathcal{Y}} \langle \bar{\boldsymbol{\theta}}, \phi(\mathbf{y}) \rangle = \Phi(\bar{\boldsymbol{\theta}}) . \end{aligned}$$

To get the tightest bound we can consider the following maximization problem:

$$\max_{\boldsymbol{\theta} \in \mathcal{A}, \sum_T \rho^T \theta^T = \bar{\boldsymbol{\theta}}} \Phi_\rho(\boldsymbol{\theta}) . \quad (4)$$

## New Tree-reweighing message passing

**Theorem 1.** *Minimization problem (3) is the dual to maximization problem (4). Strong duality holds, so their optimal values coincide.*

$$\min_{\tau \in \mathcal{L}(\mathcal{G})} \langle \bar{\boldsymbol{\theta}}, \tau \rangle \quad \leftrightarrow \quad \max_{\boldsymbol{\theta} \in \mathcal{A}, \sum_T \rho^T \theta^T = \bar{\boldsymbol{\theta}}} \Phi_\rho(\boldsymbol{\theta})$$

The maximization problem (4) is modified by replacing the constraint as

$$\max_{\boldsymbol{\theta} \in \mathcal{A}, \sum_T \rho^T \theta^T \equiv \bar{\boldsymbol{\theta}}} \Phi_\rho(\boldsymbol{\theta}) . \quad (5)$$

**Theorem 2.** *The optimal value of problem (5) equals to the optimal value of problem (4).*

The goal of the reparameterization step is to make sure that the algorithm satisfies the min-marginal property.

### Weak tree agreement

Let  $\text{OPT}^T(\theta^T)$  be the set of optimal configurations for parameter  $\theta^T$  and  $\text{OPT}(\theta) = \{\text{OPT}^T(\theta^T) \mid T \in \mathcal{T}\} \in (2^{\mathcal{Y}})^{|\mathcal{T}|}$ . For two collections  $\mathbb{S}, \tilde{\mathbb{S}} \in (2^{\mathcal{Y}})^{|\mathcal{T}|}$ , we write  $\mathbb{S} \subseteq \tilde{\mathbb{S}}$  if  $\mathbb{S}^T \subseteq \tilde{\mathbb{S}}^T$  for every tree  $T$ .

$\mathbb{S}$  is **consistent** if it satisfies the following three conditions:

1. For every tree  $T$  set  $\mathbb{S}^T$  is non-empty.
2. If node  $i$  is contained in trees  $T$  and  $T'$ , then for all  $\mathbf{y} \in \mathbb{S}^T$  there exists configuration  $\mathbf{y}' \in \mathbb{S}^{T'}$  which agrees with  $\mathbf{y}$  on node  $i$ , i.e.  $y_i = y'_i$ .
3. If edge  $(i, j)$  is contained in trees  $T$  and  $T'$ , then for all  $\mathbf{y} \in \mathbb{S}^T$  there exists configuration  $\mathbf{y}' \in \mathbb{S}^{T'}$  which agrees with  $\mathbf{y}$  on nodes  $i$  and  $j$ , i.e.  $y_i = y'_i, y_j = y'_j$ .

Vector  $\theta = \{\theta^T\} \in \mathcal{A}$  is said to satisfy the **weak tree agreement condition** if there exists collection  $\mathbb{S} \subseteq \text{OPT}(\theta)$  which is consistent.

If a vector  $\theta$  satisfies the WTA condition, then the TRW-S algorithm will not make any progress, i.e. it will not increase function  $\Phi_\rho$ .

### TRW-S algorithm

0. Initialize  $\theta$  so that  $\theta \in \mathcal{A}$  and  $\sum_T \rho^T \theta^T \equiv \bar{\theta}$ .
1. Select some order for nodes and edges in  $\mathcal{V} \cup \mathcal{E}$ . For each element  $\omega \in \mathcal{V} \cup \mathcal{E}$  find all trees  $\mathcal{T}_\omega \subseteq \mathcal{T}$  containing  $\omega$ . If there is more than one tree, then do the following:
  - (a) For all trees  $T \in \mathcal{T}_\omega$  reparameterize  $\theta^T$  such that values  $\theta_{i;a}^T$  (if  $\omega = i$  is a node) or  $\theta_{i;a}^T + \theta_{ij;ab}^T \theta_{j;b}^T$  (if  $\omega = (i, j)$  is an edge) give correct min-marginals for tree  $T$ .
  - (b) "Averaging operation":  
 If  $\omega = i$  is a node then set  $\theta_i^T := \frac{1}{\rho_i} \sum_{T \in \mathcal{T}_i} \rho^T \theta_i^T$  for trees  $T \in \mathcal{T}_i$   
 If  $\omega = (i, j)$  is an edge then set  $\theta_i^T, \theta_{ij}^T, \theta_j^T$  for trees  $T \in \mathcal{T}_{ij}$  so that

$$(\theta_{i;a}^T + \theta_{ij;ab}^T + \theta_{j;b}^T) = \frac{1}{\rho_{ij}} \sum_{T \in \mathcal{T}_{ij}} (\theta_{i;a}^T + \theta_{ij;ab}^T + \theta_{j;b}^T)$$

2. Check whether a stopping criterion is satisfied; if yes, terminate, otherwise go to step 1.

**KL divergence**

Assume two discrete probability distributions  $P$  and  $Q$ . One way to measure the *difference* between  $P$  and  $Q$  is to calculate the **Kullback–Leibler (KL) divergence** (a.k.a. relative entropy) defined as

$$\begin{aligned} D_{\text{KL}}(P\|Q) &= \sum_i P(i) \log \frac{P(i)}{Q(i)} = \sum_i P(i) \log P(i) - \sum_i P(i) \log Q(i) \\ &= \mathbb{E}_P[\log P(i)] - \mathbb{E}_P[\log Q(i)]. \end{aligned}$$

It is defined iff  $Q(i) = 0$  implies  $P(i) = 0$ , for all  $i$ . If  $P(i) = 0$ , then the  $i$ th term is interpreted as 0. The KL divergence is always non-negative, moreover  $D_{\text{KL}}(P\|Q) = 0$  iff  $P = Q$  *almost everywhere*.

*Interpretation* (Information Theory): it is the amount of information lost when  $Q$  is used to approximate  $P$ . It measures the expected number of extra bits required to code samples from  $P$  using a code optimized for  $Q$  rather than the code optimized for  $P$ .

**Motivation**

For general (discrete) factor graph models, performing *probabilistic inference* is hard. Assume we are given an **intractable** distribution  $p(y | x)$ . We consider an **approximate distribution**  $q(y)$ , which is tractable, for  $p(y | x)$ .

One way of finding the best approximating distribution is to pose it as an **optimization problem** over probability distributions: given a distribution  $p(y | x)$  and a family  $Q$  of tractable distributions  $q \in Q$  on  $\mathcal{Y}$ , we want to solve

$$\begin{aligned} q^* \in \operatorname{argmin}_{q \in Q} D_{\text{KL}}(q(y)\|p(y | x)) &= \operatorname{argmin}_{q \in Q} \sum_{y \in \mathcal{Y}} q(y) \log \frac{q(y)}{p(y | x)} \\ &= \operatorname{argmin}_{q \in Q} \left\{ \underbrace{\sum_{y \in \mathcal{Y}} q(y) \log q(y)}_{-H(q)} - \sum_{y \in \mathcal{Y}} q(y) \log p(y | x) \right\}. \end{aligned}$$

The term  $-\sum_{y \in \mathcal{Y}} q(y) \log q(y) \triangleq H(q)$  is called the **entropy** of the distribution  $q$ .



## Mean Field methods

$$\begin{aligned}
 D_{\text{KL}}(q(y) \| p(y | x)) &= -H(q) - \sum_{y \in \mathcal{Y}} q(y) \log p(y | x) \\
 &= -H(q) - \sum_{y \in \mathcal{Y}} q(y) \log \frac{1}{Z(x)} \prod_{F \in \mathcal{F}} \exp(-E_F(y_F; x_F)) \\
 &= -H(q) + \sum_{y \in \mathcal{Y}} q(y) \sum_{F \in \mathcal{F}} E_F(y_F; x_F) + \log Z(x) \\
 &= -H(q) + \sum_{F \in \mathcal{F}} \sum_{y_F \in \mathcal{Y}_F} \underbrace{\sum_{\substack{y' \in \mathcal{Y}, \\ y'_F = y_F}} q(y)}_{\mu_{F, y_F}(q)} E_F(y_F; x_F) + \log Z(x) \\
 &= -H(q) + \sum_{F \in \mathcal{F}} \sum_{y_F \in \mathcal{Y}_F} \mu_{F, y_F}(q) E_F(y_F; x_F) + \log Z(x),
 \end{aligned}$$

where  $\mu_{F, y_F}(q) = \sum_{y' \in \mathcal{Y}, y'_F = y_F} q(y)$  are the marginals of  $q$ .

## Gibbs inequality

If the set  $Q$  is rich enough to contain a close approximation to  $p(y | x)$  and we succeed at finding it, then the marginals of  $q^*$  will provide a good approximation to the true marginals of  $p(y | x)$  that are intractable to compute.

**Gibbs inequality** provides a lower bound on the log *partition function*.

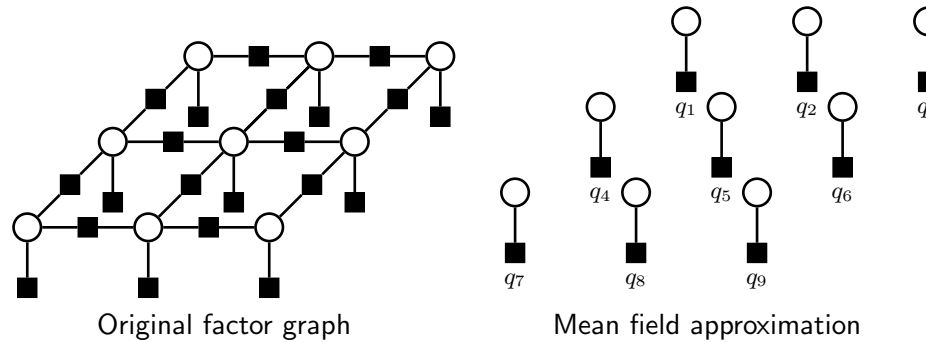
$$\begin{aligned}
 0 \leq D_{\text{KL}}(q(y) \| p(y | x)) &= -H(q) + \sum_{F \in \mathcal{F}} \sum_{y_F \in \mathcal{Y}_F} \mu_{F, y_F}(q) E_F(y_F; x_F) + \log Z(x) \\
 \log Z(x) &\geq H(q) - \sum_{F \in \mathcal{F}} \sum_{y_F \in \mathcal{Y}_F} \mu_{F, y_F}(q) E_F(y_F; x_F).
 \end{aligned}$$

## Naive mean field

Take a set  $Q$  as the set of all distributions in the form:

$$q(y) = \prod_{i \in \mathcal{V}} q_i(y_i) .$$

For example, in case of the following factor graph:



## Naive mean field

Set  $Q$  consists of all distributions in the form:

$$q(y) = \prod_{i \in \mathcal{V}} q_i(y_i) .$$

Marginals  $\mu_{F, y_F}$  take the form

$$\mu_{F, y_F}(q) = \sum_{\substack{y' \in \mathcal{V}, \\ y'_F = y_F}} q(y) = q_{N(F)}(y_F) = \prod_{i \in N(F)} q_i(y_i) .$$

Entropy  $H(q)$  decomposes as

$$H(q) = \sum_{i \in \mathcal{V}} H_i(q_i) = - \sum_{i \in \mathcal{V}} \sum_{y_i \in \mathcal{Y}_i} q_i(y_i) \log q_i(y_i) .$$

*Proof.* Exercise. □

## Naive Mean Field

Putting it together,

$$\begin{aligned} q^* &\in \operatorname{argmin}_{q \in Q} D_{\text{KL}}(q(y) \| p(y | x)) \\ &= \operatorname{argmin}_{q \in Q} \left\{ -H(q) + \sum_{F \in \mathcal{F}} \sum_{y_F \in \mathcal{Y}_F} \mu_{F, y_F}(q) E_F(y_F; x_F) + \log Z(x) \right\} \\ &= \operatorname{argmax}_{q \in Q} \left\{ H(q) - \sum_{F \in \mathcal{F}} \sum_{y_F \in \mathcal{Y}_F} \mu_{F, y_F}(q) E_F(y_F; x_F) \right\} \\ &= \operatorname{argmax}_{q \in Q} \left\{ - \sum_{i \in \mathcal{V}} \sum_{y_i \in \mathcal{Y}_i} q_i(y_i) \log q_i(y_i) - \sum_{F \in \mathcal{F}} \sum_{y_F \in \mathcal{Y}_F} \left( \prod_{i \in N(F)} q_i(y_i) \right) E_F(y_F; x_F) \right\}. \end{aligned}$$

Optimizing over  $Q$  means to optimize over all  $q_i$  such that  $q_i(y_i) \geq 0$  and  $\sum_{y_i \in \mathcal{Y}_i} q_i(y_i) = 1$  for all  $i \in \mathcal{V}$ .



## Optimization

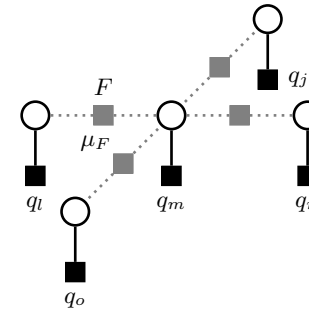
$$\operatorname{argmax}_{q \in Q} \left\{ - \sum_{i \in \mathcal{V}} \sum_{y_i \in \mathcal{Y}_i} q_i(y_i) \log q_i(y_i) - \sum_{F \in \mathcal{F}} \sum_{y_F \in \mathcal{Y}_F} \left( \prod_{i \in N(F)} q_i(y_i) \right) E_F(y_F; x_F) \right\}.$$

The entropy term is concave and the second term is non-concave due to products of variables occurring in the expression. Therefore solving this non-concave maximization problem globally is hard in general.

*Remedy:* **block coordinate ascent**

We hold all variables fixed except for a single block  $q_m$ , then we obtain a tractable concave maximization problem

→ closed-form update for each  $q_m$ .



## Lagrange multipliers

To obtain closed form solution, we define the *Lagrangian function*:

$$L(q_i, \lambda) = \left\{ - \sum_{i \in \mathcal{V}} \sum_{y_i \in \mathcal{Y}_i} q_i(y_i) \log q_i(y_i) - \sum_{F \in \mathcal{F}} \sum_{y_F \in \mathcal{Y}_F} \left( \prod_{i \in N(F)} q_i(y_i) \right) E_F(y_F; x_F) + \lambda \left( \sum_{y_i \in \mathcal{Y}_i} q_i(y_i) - 1 \right) \right\}.$$

Setting the derivatives of  $L$  w.r.t.  $q_i$  to 0, we obtain

$$\frac{\partial L}{\partial q_i(y_i)} = 0 = -(\log q_i(y_i) + 1) - \sum_{\substack{F \in M(i) \\ y'_F \in \mathcal{Y}_F, \\ y'_i = y_i}} \sum_{j \in N(F) \setminus \{i\}} \left( \prod_{j \in N(F) \setminus \{i\}} \hat{q}_j(y_j) \right) E_F(y_F; x_F) + \lambda$$

$$q_i^*(y_i) = \exp \left( -1 - \sum_{\substack{F \in M(i) \\ y'_F \in \mathcal{Y}_F, \\ y'_i = y_i}} \sum_{j \in N(F) \setminus \{i\}} \left( \prod_{j \in N(F) \setminus \{i\}} \hat{q}_j(y_j) \right) E_F(y_F; x_F) + \lambda \right).$$

## Lagrange multipliers

$\lambda$  can be calculated as follows.

$$\sum_{y_i \in \mathcal{Y}_i} q_i(y_i) = \sum_{y_i \in \mathcal{Y}_i} \exp \left( -1 - \sum_{F \in M(i)} \sum_{\substack{y'_F \in \mathcal{Y}_F, \\ y'_i = y_i}} \left( \prod_{j \in N(F) \setminus \{i\}} \hat{q}_j(y_j) \right) E_F(y_F; x_F) + \lambda \right)$$
$$\exp(1 - \lambda) = \underbrace{\sum_{y_i \in \mathcal{Y}_i} \exp \left( - \sum_{F \in M(i)} \sum_{y'_F \in \mathcal{Y}_F, y'_i = y_i} \left( \prod_{j \in N(F) \setminus \{i\}} \hat{q}_j(y_j) \right) E_F(y_F; x_F) \right)}_{Z_i(x_F)}$$
$$\lambda - 1 = -\log Z_i(x_F),$$

where  $Z_i(x_F)$  is a normalizing constant for  $q_i$ .

### Update equation

By substituting, we obtain the obtain equation for the Naive Mean Field method

$$\begin{aligned} q_i^*(y_i) &= \exp \left( - \sum_{F \in M(i)} \sum_{\substack{y'_F \in \mathcal{Y}_F, \\ y'_i = y_i}} \left( \prod_{j \in N(F) \setminus \{i\}} \hat{q}_j(y_j) \right) E_F(y_F; x_F) - \log Z_i(x_F) \right) \\ &= \frac{1}{Z_i(x_F)} \exp \left( - \sum_{F \in M(i)} \sum_{\substack{y'_F \in \mathcal{Y}_F, \\ y'_i = y_i}} \left( \prod_{j \in N(F) \setminus \{i\}} \hat{q}_j(y_j) \right) E_F(y_F; x_F) \right). \end{aligned}$$

## Semantic segmentation

Krähenbühl and Koltun proposed an efficient approximate inference in fully connected CRF model by applying *Naive Mean Field* approach.

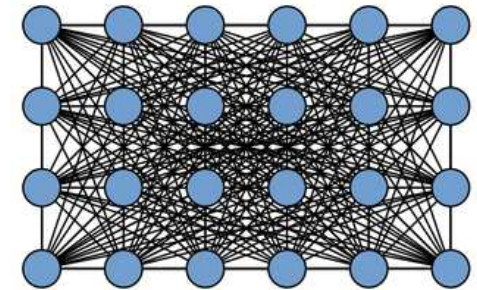
**Semantic segmentation:** assign a label from the set of labels  $\mathcal{L} = \{l_1, l_2, \dots, l_k\}$  for each pixel on the image regarding their semantic meaning.



For each pixel on the image a random variable is assigned taking a value from  $\mathcal{L}$ . A fully connected pairwise CRF model  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  is considered, where the corresponding energy function is given by

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

where  $\mathcal{E} = \{(i, j) \in \mathcal{V} \times \mathcal{V} \mid i < j\}$ .



## Energy functions

- **Unary energies**  $E_i(y_i)$  are computed independently for each pixel as  $E_i(y_i) = -\log P_i(y_i)$  measures the degree of disagreement between labelling  $y_i$  and the image at pixel  $i$ .
- **Pairwise energies** (so-called **contrast sensitive Potts-model**), measuring the extent to which the labelling  $y$  is not piecewise smooth, have the form ( $p_i$  and  $I_i$  denote the pixel coordinates and intensity, resp.)

$$\begin{aligned}
 E_{ij}(y_i, y_j) &= \mathbb{1}[y_i \neq y_j] \sum_m w^{(m)} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) \\
 &= \mathbb{1}[y_i \neq y_j] \sum_m w^{(m)} \exp\left(-\frac{1}{2}(\mathbf{f}_i - \mathbf{f}_j)^T \mathbf{\Lambda}^{(m)} (\mathbf{f}_i - \mathbf{f}_j)\right) \\
 &= \mathbb{1}[y_i \neq y_j] \left\{ w^{(1)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right) \right. \\
 &\quad \left. + w^{(2)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right) \right\}.
 \end{aligned}$$

The parameters  $\theta_\alpha$ ,  $\theta_\beta$  and  $\theta_\gamma$  are estimated on a set of training images.

## Inference

The inference is based on Naive Mean Field approximation, where the update equation is given by

$$q_i(y_i) = \frac{1}{Z_i} \exp \left\{ -E_i(y_i) - \sum_{l' \in \mathcal{L}} \mathbb{1}[y_i \neq y_{l'}] \sum_{m=1}^K w^{(m)} \sum_{j \in \mathcal{V}} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) q_j(l') \right\}.$$

The inference is performed in average 0.2 seconds for 500.000 variables (in contrast to 36 hours).

*The main idea:* the message passing step can be expressed as a convolution with a Gaussian kernel  $G_{\mathbf{\Lambda}^{(m)}}$  in feature space:

$$\sum_{j \in \mathcal{V}} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) q_j(l) - q_i(l) = [G_{\mathbf{\Lambda}^{(m)}} * q(l)](\mathbf{f}_i) - q_i(l).$$

Note that the convolution sums over all variables, while message passing does not sum over  $q_i$ . This convolution can be efficiently calculated in  $\mathcal{O}(|\mathcal{V}|)$  time (instead of  $\mathcal{O}(|\mathcal{V}|^2)$ ).



## Literature

- Martin J. Wainwright, Tommi S. Jaakkola, and Alan S. Willsky. **MAP Estimation Via Agreement on Trees: Message-Passing and Linear Programming.** In *IEEE Transactions on Information Theory*, vol. 51(11), pp. 3697–3717, November 2005.
- Vladimir Kolmogorov. **Convergent Tree-reweighted Message Passing for Energy Minimization.** In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28(10), pp. 1568-1583, October 2006.
- Sebastian Nowozin and Christoph H. Lampert. **Structured Prediction and Learning in Computer Vision.** In *Foundations and Trends in Computer Graphics and Vision*, Volume 6, Number 3-4. Note: Chapter 3.
- Daphne Koller and Nir Friedman. **Probabilistic Graphical Models: Principles and Techniques.** The MIT Press, 2009. Note: Chapter 11.
- Philipp Krähenbühl and Vladlen Koltun. **Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials.** In *Proceedings of Advances in Neural Information Processing Systems*, pp. 109–117, Granada, Spain, Dec 2011. MIT Press.