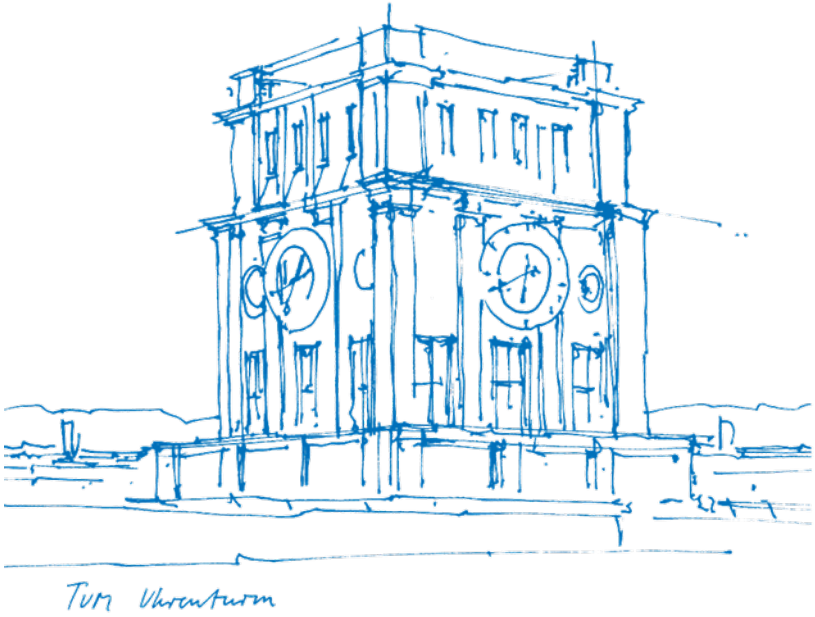
III : Inference on Graphical Models

Tao Wu, Yuesong Shen, Zhenzhang Ye

Computer Vision & Artificial Intelligence
Technical University of Munich



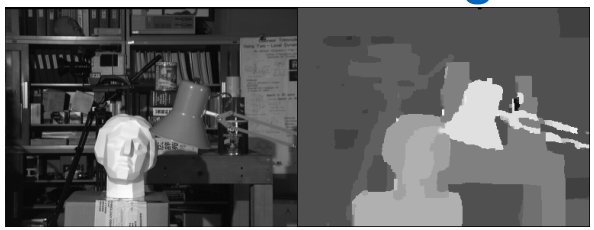
Motivation

– Many computer vision tasks boil down to inference on graphical models.

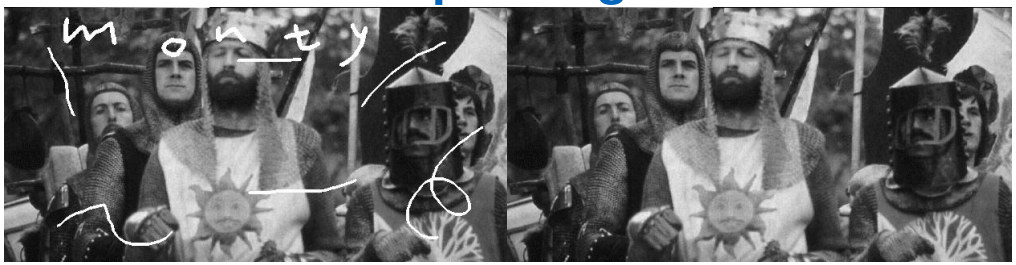
Denoising

Optical flow

Stereo matching



Inpainting



Super-resolution

1. **Probabilistic inference**: compute marginal distribution

$$p(y) = \sum_x p(y, x).$$

2. **MAP inference**: compute maximum of conditional distribution

$$\arg\max_y p(y|x).$$

Exact Inference

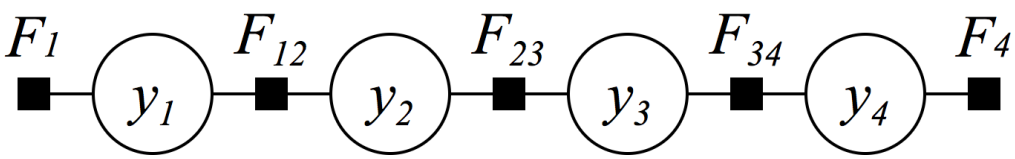
Outline of the Section

- Basic idea: Variable elimination.
- Junction tree algorithm on arbitrary MRFs.
- Belief propagation on tree factor graphs.

Example: Marginal Query on a "Chain" MRF

Joint distribution represented by MRF:

$$p(y_1, y_2, y_3, y_4) = \frac{1}{Z}\phi_1(y_1) \cdot \phi_{12}(y_1, y_2) \cdot \phi_{23}(y_2, y_3) \cdot \phi_{34}(y_3, y_4) \cdot \phi_4(y_4).$$



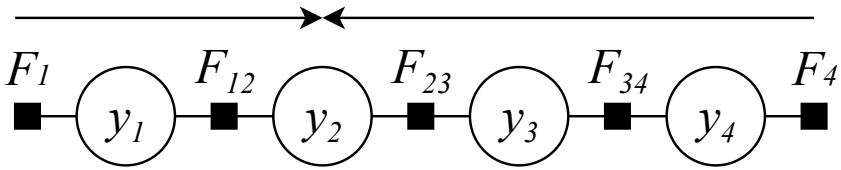
Query about marginal distribution $p(y_2) = ?$

Variable Elimination

Apply **variable elimination** (VE) to the marginal query:

$$
\begin{aligned}
p(y_2) &= \sum_{y_1,y_3,y_4} p(y_1,y_2,y_3,y_4) \\
&= \sum_{y_1,y_3,y_4} \frac{1}{Z}\phi_1(y_1)\phi_{12}(y_1,y_2)\phi_{23}(y_2,y_3)\phi_{34}(y_3,y_4)\phi_4(y_4) \\
&= \frac{1}{Z}\underbrace{\sum_{y_1}\Big(\phi_1(y_1)\phi_{12}(y_1,y_2)\Big)}_{=:\, m_{1\to 2}(y_2)}\sum_{y_3}\Big(\phi_{23}(y_2,y_3)\underbrace{\sum_{y_4}\big(\phi_{34}(y_3,y_4)\phi_4(y_4)\big)}_{=:\, m_{4\to 3}(y_3)}\Big) \\
&= \frac{1}{Z}m_{1\to 2}(y_2)\underbrace{\sum_{y_3}\Big(\phi_{23}(y_2,y_3)m_{4\to 3}(y_3)\Big)}_{=:\, m_{3\to 2}(y_2)} \\
&= \frac{1}{Z}m_{1\to 2}(y_2)m_{3\to 2}(y_2), \\
Z &= \sum_{y_2} m_{1\to 2}(y_2)m_{3\to 2}(y_2).
\end{aligned}
$$

Variable Elimination and Beyond



- This algorithm is called **sum-product** VE.
- Sum-product VE yields *exact* inference (of one node marginal) on any *tree-structured factor graph*.
- Observed nodes (a.k.a. *evidence*) can be introduced as reduced factors.
- A similar algorithm can be derived for MAP inference – simply switch all "sum" to "max". The resulting algorithm is called **max-product** VE.
- We shall consider two different extensions beyond VE:

1. Inference on arbitrary MRFs? $\rightsquigarrow$ **Junction tree algorithm**.
2. Compute all node/factor marginals at one shot? $\rightsquigarrow$ **Belief propagation**.

Junction Tree

- For an undirected graph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, the **junction tree** of $\mathcal{H}$ is a tree $\mathcal{T}$ s.t.
  1. The nodes of $\mathcal{T}$ consist of the *maximal cliques* of $\mathcal{H}$.
  2. The edge S_{ij} between two nodes C_i, C_j of $\mathcal{T}$ (i.e. two maximal cliques of $\mathcal{H}$) is given by $S_{ij} = C_i \cap C_j$ (known as the *running intersection property*).
- $\mathcal{H}$ is **triangulated** if every cycle of length ≥ 4 has a *chord*. (A chord is an edge that is not part of the cycle but connects two vertices of the cycle.)
- <u>Theorem</u> [Lauritzen '96]: A graph has a junction tree iff it is triangulated.

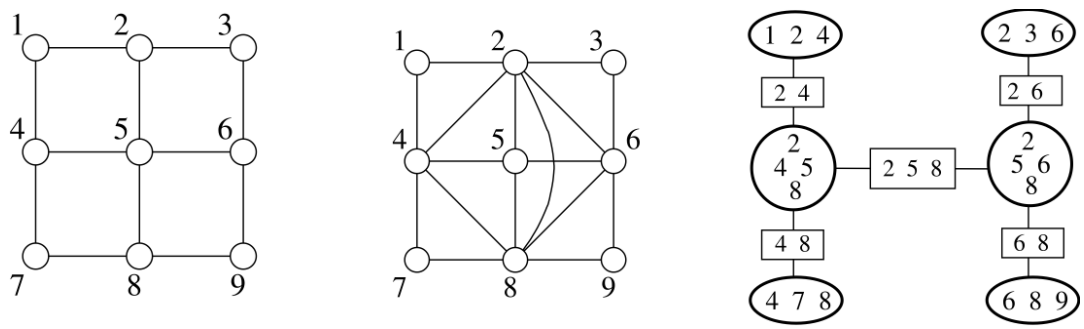
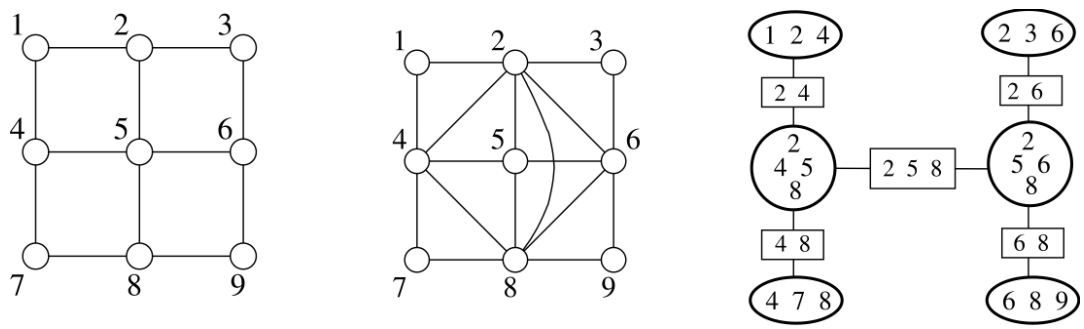


Figure:[1] (a) Original graph; (b) Triangulation of (a); (c) Junction tree for (b).

[1] Wainwright and Jordan, "Graphical Models, Exponential Families, and Variational Inference".

Junction Tree Algorithm (Sketch)



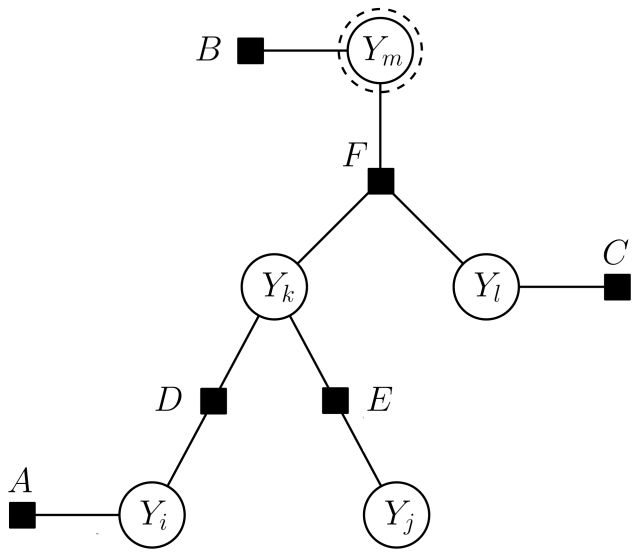
Sum-product message passing on a junction tree $\mathcal{T}$ appears like:

$$m_{C_i \to C_j}(y_{C_j \cap C_i}) = \sum_{y_{C_i \setminus C_j}} \phi_{C_i}(y_{C_i}) \prod_{C_k \in \mathrm{nbr}_{\mathcal{T}}(C_i) \setminus \{C_j\}} m_{C_k \to C_i}(y_{C_i \cap C_k}).$$

Overall **junction tree algorithm** for exact inference on an arbitrary MRF:

1. Given an MRF with cycles, triangulate it by adding edges as necessary.
2. Form a junction tree $\mathcal{T}$ for the triangulated MRF.
3. Run VE on the junction tree $\mathcal{T}$.

Belief Propagation on Tree Factor Graphs[2]



- Factor graph $\mathcal{G} = (\mathcal{V}, \mathcal{F}, \mathcal{E})$: assumed to be a tree.
- Neighbors of a variable or factor node:

$$\text{nbr}_{\mathcal{G}}(i) = \{F \in \mathcal{F} : (i, F) \in \mathcal{E}\},$$
$$\text{nbr}_{\mathcal{G}}(F) = \{i \in \mathcal{V} : (i, F) \in \mathcal{E}\}.$$

- (Log-domain) energies: $E_F(y_F) = -\log \phi_F(y_F)$.

[2] Illustrations for BP are extracted from Nowozin & Lampert, 2011.

BP: Leaf-to-Root Stage

0. Pick $Y_r \in \mathcal{V}$ as the tree root (e.g. Y_m in the figure).

1a. Schedule the leaf-to-root messages.

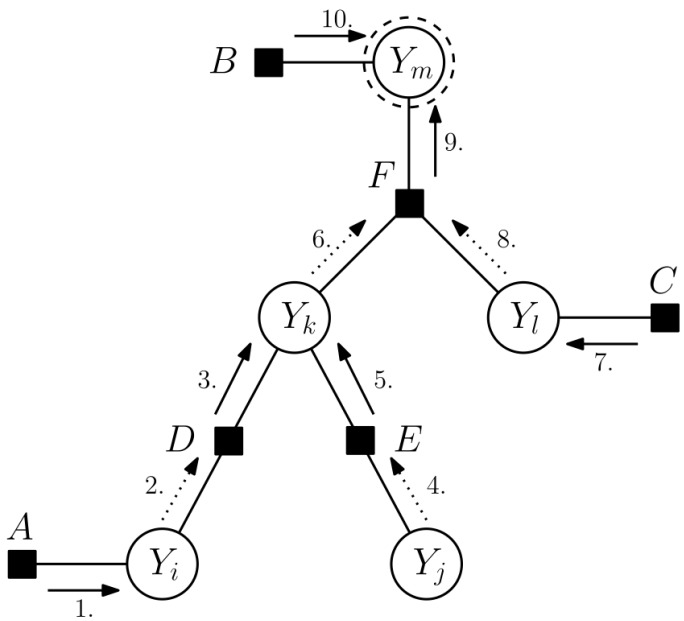


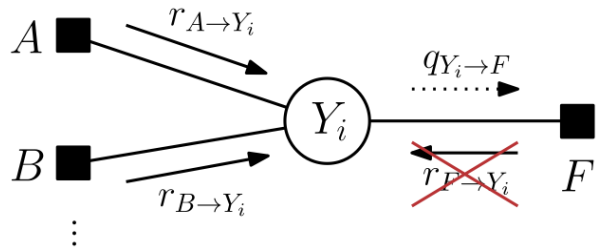
Figure: Belief propagation: leaf-to-root stage.

1b. Compute all leaf-to-root messages (detailed in the next slide).

BP: Compute Messages

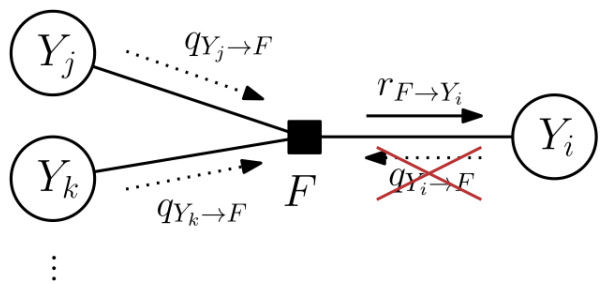
- Compute variable-to-factor message:

$$q_{i\to F}(y_i) = \sum_{F' \in \mathrm{nbr}_{\mathcal{G}}(i)\setminus\{F\}} r_{F'\to i}(y_i).$$



- Compute factor-to-variable message:

$$r_{F\to i}(y_i) = \log \sum_{y_{F\setminus\{i\}}} \exp\Big(- E_F(y_F) + \sum_{i' \in \mathrm{nbr}_{\mathcal{G}}(F)\setminus\{i\}} q_{i'\to F}(y_{i'})\Big).$$



BP: Compute the Partition Function

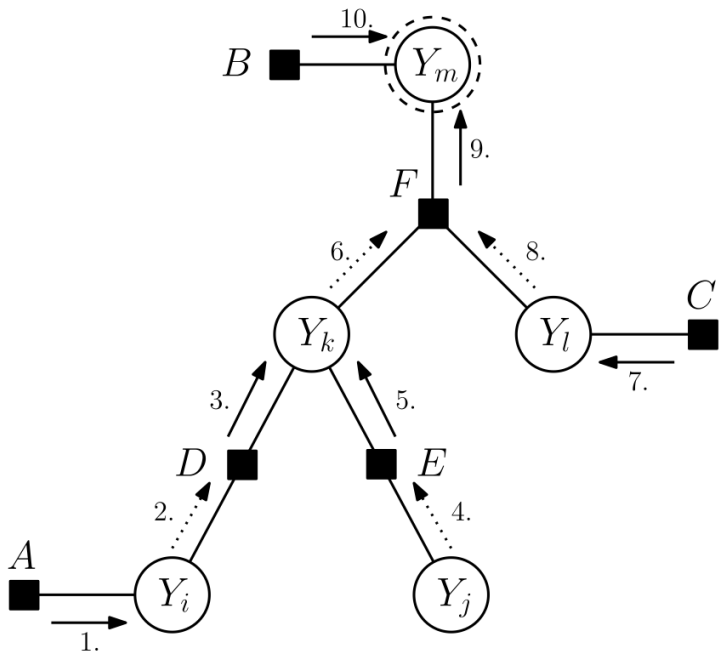


Figure: Belief propagation: leaf-to-root stage.

1c. Compute the log partition function:

$$\log Z = \log \sum_{y_r} \exp\Big(\sum_{F \in \mathrm{nbr}_{\mathcal{G}}(r)} r_{F \to r}(y_r) \Big).$$

BP: Root-to-Leaf Stage

2a. Schedule the root-to-leaf messages.

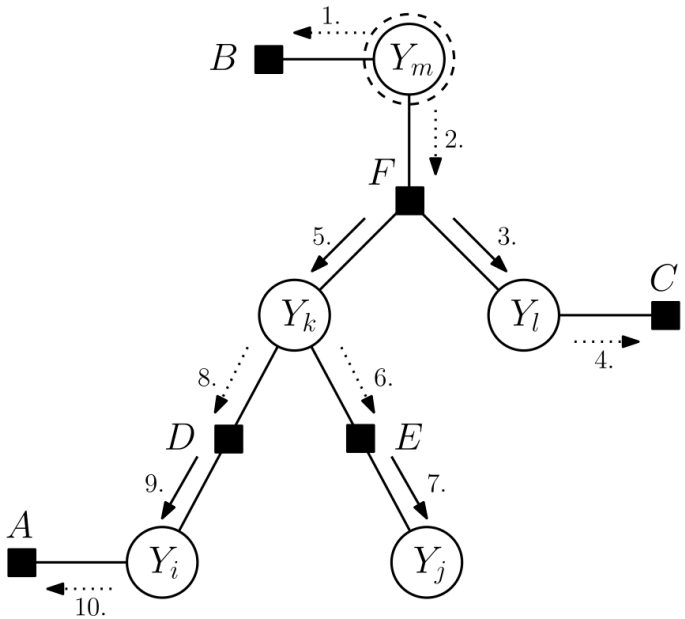


Figure: Belief propagation: root-to-leaf stage.

2b. Compute the root-to-leaf messages using the same formulas on page 12.

BP: Compute Factor / Variable Marginals

2c. Alongside Step 2b, combine messages and compute factor marginals:

$$\mu_F(y_F) := p(y_F) = \exp\Big(- E_F(y_F) + \sum_{i \in \mathrm{nbr}_{\mathcal{G}}(F)} q_{i \to F}(y_i) - \log Z \Big),$$

as well as variable marginals:

$$\mu_i(y_i) := p(y_i) = \exp\Big(\sum_{F \in \mathrm{nbr}_{\mathcal{G}}(i)} r_{F \to i}(y_i) - \log Z \Big).$$

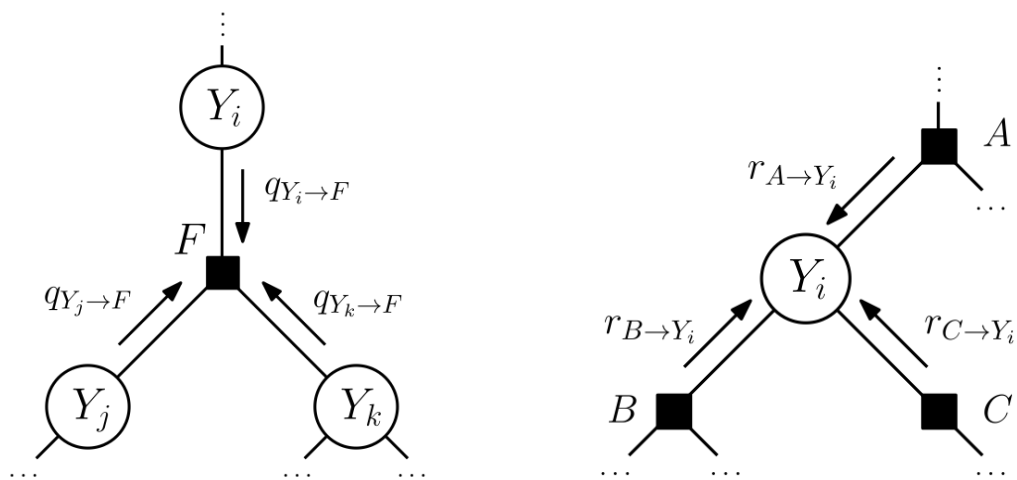


Figure: (left) Factor marginal; (right) Variable marginal.

BP on Pairwise MRFs (as exercise)

For a pairwise MRF $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, the joint distribution is factorized by

$$p(y) = \exp\Big(- \sum_{i \in \mathcal{V}} E_i(y_i) - \sum_{(i,j) \in \mathcal{E}} E_{ij}(y_i, y_j) - \log Z \Big).$$

BP on such pairwise MRF can be simplified:

- Variable-to-variable message is computed by

$$m_{i \to j}(y_j) = \log \sum_{y_i} \exp\Big(- E_i(y_i) - E_{ij}(y_i, y_j) + \sum_{k \in \mathrm{nbr}_{\mathcal{H}}(i) \setminus \{j\}} m_{k \to i}(y_i) \Big).$$

- Variable marginal is computed by

$$\mu_i(y_i) = \exp\Big(- E_i(y_i) + \sum_{k \in \mathrm{nbr}_{\mathcal{H}}(i)} m_{k \to i}(y_i) - \log Z \Big).$$

Further Reading

- Koller & Friedman, Chapters 9, 10.
- Murphy, Chapter 20.
- Nowozin & Lampert, Section 3.1.

Approximate Inference

Outline of this Section

- Basic idea: Variational inference.
- Mean field (MF) method.
- Loopy belief propagation (LBP).

Approximation by Tractable Distributions

- Goal: probabilistic inference on joint distribution $p(y)$ represented by *general* MRF (i.e. possibly with loops).
- Instead of tackling the inference on p directly, we first seek for an approximation q within a family $\mathcal{Q}$ consisting of "tractable" distributions:

$$q^* = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}\left(q \,|\, p\right).$$

- The **Kullback-Leibler (KL) divergence** (a.k.a. *relative entropy*) between two distributions q, p (assuming p is a positive distribution) is defined by

$$\mathrm{KL}\left(q \,|\, p\right) = \sum_y q(y) \log \frac{q(y)}{p(y)}.$$

- Basic properties of KL:
  1. $\mathrm{KL}\left(q \,|\, p\right) = 0$ iff $p = q$.
  2. $\mathrm{KL}\left(q \,|\, p\right) \geq 0 \; \forall q, p$.
  3. $\mathrm{KL}\left(\cdot \,|\, \cdot\right)$ is not symmetric. Nor does it satisfy the triangle inequality.

Preliminaries to Variational Inference

- Represented by a factor graph $\mathcal{G} = (\mathcal{V}, \mathcal{F}, \mathcal{E})$, p takes the form

$$p(y) = \exp\Big(- \sum_{F \in \mathcal{F}} E_F(y_F) - \log Z \Big).$$

- Plug p into KL divergence $\rightsquigarrow$

$$\begin{aligned} \mathrm{KL}\,(q \,|\, p) &= \sum_y q(y) \log \frac{q(y)}{p(y)} = \sum_y q(y) \log q(y) - \sum_y q(y) \log p(y) \\ &= -H(q) + \sum_{F \in \mathcal{F}} \sum_{y_F} \mu_F[q](y_F) E_F(y_F) + \log Z. \end{aligned}$$

- $H(q)$ is the **entropy** of distribution q.
- $\mu_F[q]$ is the marginal distribution of q over variables Y_F.
- $F_{\text{Gibbs}}(q; p) := \mathrm{KL}\,(q \,|\, p) - \log Z = -H(q) + \sum_{F \in \mathcal{F}} \sum_{y_F} \mu_F[q](y_F) E_F(y_F)$ is called the **Gibbs free energy**.
- $\mathrm{KL}\,(q \,|\, p) \geq 0 \;\Rightarrow\; \log Z$ is lower bounded by $-F_{\text{Gibbs}}(q; p)$.

Mean Field Approximation

In (naive) **mean field** method, $\mathcal{Q}$ consists of q factorized by only unaries:

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

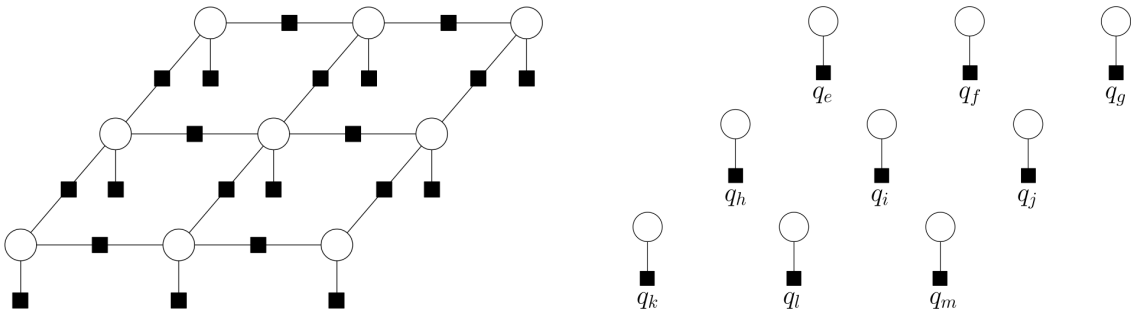


Figure: (left) Original factor graph; (right) (Naive) mean field approximation.

- Such q is "tractable" because $\{q_i(y_i)\}$ provide variable marginals.
- Quick facts: $$H(q) = \sum_{i \in \mathcal{V}} H(q_i) = -\sum_{i \in \mathcal{V}} \sum_{y_i} q_i(y_i) \log q_i(y_i),$$
$$\mu_F[q](y_F) = \prod_{i \in \mathrm{nbr}_{\mathcal{G}}(F)} q_i(y_i).$$

Mean Field (MF) Approximation

Derivation of MF approximation:

$$
\begin{aligned}
q^* &= \arg\min_{q \in \mathcal{Q}} \mathrm{KL}\left(q \,|\, p\right) = \arg\min_{q \in \mathcal{Q}} F(q; p) \\
&= \arg\min_{q \in \mathcal{Q}} -H(q) + \sum_{F \in \mathcal{F}} \sum_{y_F} \mu_F[q](y_F) E_F(y_F) \\
&= \arg\min_{\{q_i\}_{i \in \mathcal{V}}} \sum_{i \in \mathcal{V}} \sum_{y_i} q_i(y_i) \log q_i(y_i) + \sum_{F \in \mathcal{F}} \sum_{y_F} \left(\prod_{i \in \mathrm{nbr}_{\mathcal{G}}(F)} q_i(y_i) \right) E_F(y_F).
\end{aligned}
$$

Each q_i lies in the probability simplex Δ_i, i.e.

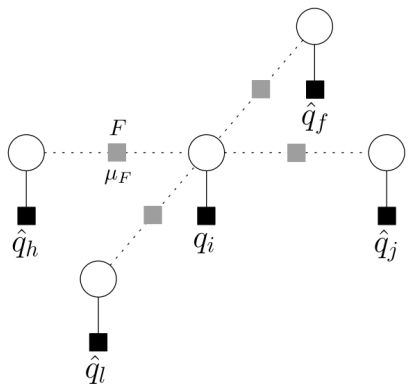
$$
\begin{aligned}
q_i(y_i) &\geq 0 \quad \forall y_i, \\
\sum_{y_i} q_i(y_i) &= 1.
\end{aligned}
$$

The optimization can be resolved by *coordinate descent* (next slide).

MF Update Formula

For each block q_i, fix $\widehat{q}_{i'}(y_{i'}) = q_{i'}(y_{i'})\ \forall i' \neq i$ and solve:

$$q_i^* = \arg\min_{q_i \in \Delta_i} \sum_{y_i} q_i(y_i) \log q_i(y_i) + \sum_{F \in \mathrm{nbr}_{\mathcal{G}}(i)} \sum_{y_F} \left(\prod_{i' \in \mathrm{nbr}_{\mathcal{G}}(F) \setminus \{i\}} \widehat{q}_{i'}(y_{i'}) \right) q_i(y_i) E_F(y_F).$$



We obtain an analytical solution via Lagrange multiplier λ for $\sum_{y_i} q_i^*(y_i) = 1$:

$$q_i^*(y_i) = \exp\left(-1 - \sum_{F \in \mathrm{nbr}_{\mathcal{G}}(i)} \sum_{y_{F \setminus \{i\}}} \left(\prod_{i' \in \mathrm{nbr}_{\mathcal{G}}(F) \setminus \{i\}} \widehat{q}_{i'}(y_{i'}) \right) E_F(y_F) + \lambda \right)$$

$$\propto \exp\left(- \sum_{F \in \mathrm{nbr}_{\mathcal{G}}(i)} \sum_{y_{F \setminus \{i\}}} \left(\prod_{i' \in \mathrm{nbr}_{\mathcal{G}}(F) \setminus \{i\}} \widehat{q}_{i'}(y_{i'}) \right) E_F(y_F) \right).$$

Some Remarks on MF

- The term $\prod_{i' \in \mathrm{nbr}_{\mathcal{G}}(F) \setminus \{i\}} \widehat{q}_{i'}(y_{i'})$ is taken to be 1 if $\mathrm{nbr}_{\mathcal{G}}(F) \setminus \{i\} = \emptyset$.
- For a pairwise MRF $\mathcal{H}$, the MF update rule can be simplified as

$$q_i^*(y_i) \propto \exp\left(- E_i(y_i) - \sum_{j \in \mathrm{nbr}_{\mathcal{H}}(i)} \sum_{y_j} \widehat{q}_j(y_j) E_{ij}(y_i, y_j) \right).$$

- MF is an iterative procedure which converges to a *locally optimal* solution q^*.
- Upon convergence, $\{q_i^*\}$ directly provide (approximate) variable marginals.
- The tractable family $\mathcal{Q}$ can be more sophisticated than factorizations of unaries in naive mean field. $\rightsquigarrow$ *Structured mean field* approximation.

From Belief Propagation to Loopy Belief Propagation

- Previously we have seen how belief propagation works on tree factor graphs.
- We can use similar update rules to derive **loopy belief propagation** (LBP).
- Although LBP does not guarantee the convergence (if at all) to the true marginal, it often performs well and is widely used in practice[3].
- In the following, we first present the LBP algorithm and then interpret it from perspective of variational inference.

[3] Murphy et al., "Loopy Belief Propagation for Approximate Inference: An Empirical Study".

Loopy Belief Propagation

On a factor graph $\mathcal{G} = (\mathcal{V}, \mathcal{F}, \mathcal{E})$, LBP proceeds as follows.

0. Initialize all variable-to-factor messages: $q_{i \to F}(y_i) = 0$. Then iterate:
1. Compute all factor-to-variable messages:

$$r_{F \to i}(y_i) = \log \sum_{y_{F \setminus \{i\}}} \exp\Big(- E_F(y_F) + \sum_{i' \in \mathrm{nbr}_{\mathcal{G}}(F) \setminus \{i\}} q_{i' \to F}(y_{i'}) \Big).$$

2. Compute all (normalized) variable-to-factor messages:

$$\bar{q}_{i \to F}(y_i) = \sum_{F' \in \mathrm{nbr}_{\mathcal{G}}(i) \setminus \{F\}} r_{F' \to i}(y_i),$$

$$\delta_{i \to F} = \log \sum_{y_i} \exp\Big(\bar{q}_{i \to F}(y_i) \Big),$$

$$q_{i \to F}(y_i) = \bar{q}_{i \to F}(y_i) - \delta_{i \to F}.$$

Loopy Belief Propagation (cont'd)

3. Compute all factor marginals:

$$\mu_F(y_F) \propto \exp\Big(- E_F(y_F) + \sum_{i \in \mathrm{nbr}_{\mathcal{G}}(F)} q_{i \to F}(y_i) \Big).$$

4. Compute all variable marginals:

$$\mu_i(y_i) \propto \exp\Big(\sum_{F \in \mathrm{nbr}_{\mathcal{G}}(i)} r_{F \to i}(y_i) \Big).$$

Differences compared to BP:

- The normalization constants in the computation of marginals differ at each factor/variable.
- The log partition function is not directly available, but it can be approximated by the **Bethe free energy**:

$$\begin{aligned} -\log Z \approx F_{\mathrm{Bethe}}(\mu; p) := & \sum_{i \in \mathcal{V}} (1 - |\,\mathrm{nbr}_{\mathcal{G}}(i)|) \sum_{y_i} \mu_i(y_i) \log \mu_i(y_i) \\ & + \sum_{F \in \mathcal{F}} \sum_{y_F} \mu_F(y_F) \Big(E_F(y_F) + \log \mu_F(y_F) \Big). \end{aligned}$$

Interpretation of LBP

On a pairwise MRF $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, LBP can be interpreted as an attempt to solve:

$$\underset{\{\mu_i\}_{i\in\mathcal{V}},\, \{\mu_{ij}\}_{(i,j)\in\mathcal{E}}}{\text{minimize}} \sum_{i\in\mathcal{V}} (1 - |\,\text{nbr}_{\mathcal{H}}(i)|) \sum_{y_i} \mu_i(y_i) \log \mu_i(y_i)$$
$$+ \sum_{(i,j)\in\mathcal{E}} \sum_{y_i, y_j} \mu_{ij}(y_i, y_j) \Big(E_{ij}(y_i, y_j) + \log \mu_{ij}(y_i, y_j) \Big)$$

$$\text{subject to } \mu_i(y_i) \geq 0,\ \mu_{ij}(y_i, y_j) \geq 0,\ \sum_{y_i} \mu_i(y_i) = 1,\ \sum_{y_i} \mu_{ij}(y_i, y_j) = \mu_j(y_j).$$

- The constraints impose *local consistency* between node marginals $\{\mu_i\}$ and edge marginals $\{\mu_{ij}\}$.
- However, $\{\mu_i\}, \{\mu_{ij}\}$ under these constraints are may not be marginals of any joint distribution on $\mathcal{H}$ (i.e. outer approximation of *marginal polytope*).
- The solution for the optimization, if exists, has an analytical form (derived via Lagrangian multipliers), from which one can recover LBP updates.
- An amazing theory on variational inference arise in this context — we point those interested to the "monster" paper [Jordan & Wainwright, 2008].

LBP vs. MF

(+) (Naive) MF optimizes over only variable marginals; LBP optimizes over variable and factor marginals under local consistency constraints.

(+) LBP does exact inference on factor graphs without loops; MF is exact on a strict subclass of factor graphs, on which all true factor marginals are factorized by $\mu_F(y_F) = \prod_{i \in \mathrm{nbr}_{\mathcal{G}}(F)} \mu_i(y_i)$ (hence an inner approximation of marginal polytope).

(+) While both being approximate inference techniques, LBP tends to be more accurate than MF in practice.

(−) MF provides a lower bound of the log partition function (given by negative Gibbs free energy), while LBP does not.

(−) Compared to LBP, it is easier to extend MF to distributions other than discrete and Gaussian, due to the simplicity of working with only variable marginals.

Further Reading

- Murphy, Chapters 21, 22.
- Nowozin & Lampert, Sections 3.2, 3.3.
- Jordan & Wainwright, Chapters 4, 5.
- Koller & Friedman, Chapter 11.