Georgia Institute of Technology

CSIP
Center for Signal & Image Processing

ECE 8803

Approximate Inference in Graphical Models

Module 8: Part A

Variational Inference via Mean Field Method

Faramarz Fekri

Center for Signal and Information Processing

Approximate Inference

- Particle (Sampling) methods (Already Covered)
- Global Approximate Inference (also known as Variational Inference) (view: Inference as an optimization Problem)
  1. Generalized (Loopy) Belief Propagation (GBP)
  2. Propagation with approximate messages (known as: Expectation Propagation -EP)
  3. Structured variational approximations (Main Algorithm: Mean-Field Methods)

Read Chapter 11 of K&F

General Strategy for a Hard inference Problem in P(X)

- Define a class of simpler distributions **Q**
- Search for a particular instance in **Q** that is "close" to P
  - All methods we will discuss optimize the same target function for measuring the similarity between Q and P
- Answer queries using inference in Q rather than P.
- By relaxing the constraints and/or approximating the objective, we can trade accuracy for speed.

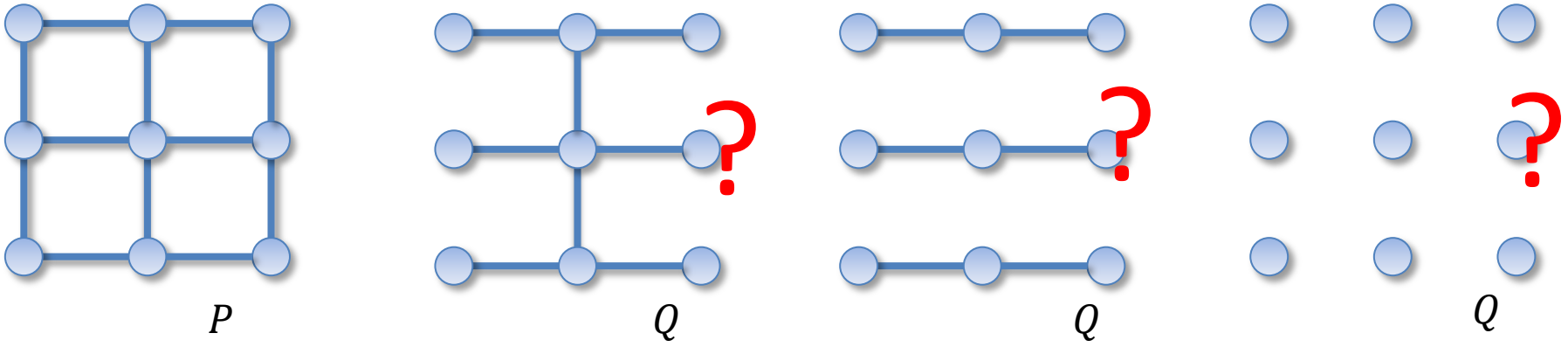
Varitional inference

- Loops are inflicting challenges for exact inference
- Junction tree can still results in large cliques (eg. Grid)
- Variational inference: approximate distribution with loopy graphs by using simpler distribution
  - This reduces inference to an optimization problem.

Key Questions in Variational Inference

- Which approximating structure is good?

mean field approximation



- How to measure the goodness of the approximation of $Q(X_1,\ldots,X_n)$ to the original $P(X_1,\ldots,X_n)$?
- $||Q - P||_F^2$?
- Relative Entropy (KL-divergence) $KL(Q \,||\, P)$?
- How do we compute the new parameters?

$$\text{Optimization } Q^* = \text{argmin}_Q \, KL(Q||P)$$

KL-divergence: Similarity of two distributions

- Kullback-Leibler (KL) divergence is often used
  - $KL(P||Q) = \sum_x P(X) \ln \frac{P(X)}{Q(X)} \geq 0$
  - $KL(P||Q) = \int_{-\infty}^{\infty} P(X) ln \frac{P(X)}{Q(X)} dX \geq 0$

- $D(P||Q) = 0$ iff $P = Q$

- Not symmetric: P determines where difference is important
  - $P(X) = 0$ and $Q(X) \neq 0$, $P(X) ln \frac{P(X)}{Q(X)} = 0 ln 0 = 0$
  - $P(X) \neq 0$ and $Q(X) = 0$, $P(X) \ln \frac{P(X)}{Q(X)} = \epsilon ln \frac{\epsilon}{0} = \infty$

Which KL-divergence to use?

- Recall, we wish to find simple Q that approximates P: $argmin_Q\, KL(P||Q)$
- But computing KL($P||Q$) requires inference:

$$KL(P||Q) = \sum_x P(X) \ln \frac{P(X)}{Q(X)} = \sum_x P(X) \ln P(X) - \sum_x P(X) \ln Q(X)$$

$no\ Q\ in\ it, ignore$

$cross\ entropy$

Given $Q \propto \prod_i \Psi(X_i)$,
it is equal to
$\sum_i \sum_x P(X) \ln \Psi(X_i) + const.$

We need to solve an inference problem: $\sum_{x_1,\ldots,x_n} P(X_1, \ldots, X_n) \ln \Psi(X_i)$

using **P(X)** which is a difficult.

Reverse KL-divergence: D(Q||P)

$$KL(Q||P) = \sum_x Q(X) \ln \frac{Q(X)}{P(X)} = \sum_x Q(X) \ln Q(X) - \sum_x Q(X) \ln P(X)$$

$\sum_x Q(X) \ln Q(X)$: -(Entropy by Q): $-H(Q)$

$\sum_x Q(X) \ln P(X)$: Cross Entropy

Let: $Q = \frac{1}{Z} \prod_i \Psi(X_i)$, $P = \frac{1}{Z'} \prod_{D_i} \Psi(D_i)$

$$-H(Q) = \frac{1}{Z} \sum_i \sum_x (\prod_j \Psi(X_j) \ln \Psi(X_i)) + \text{ some constant}$$

$$\sum_x Q(X) \ln P(X) = \frac{1}{Z} \sum_x (\prod_j \Psi(X_j) \sum_{D_i} \ln \Psi(D_i) - ln\, Z') + \text{ constant}$$

- Computing H(Q) is easy in a fully factorized distribution Q.
- Computing the cross entropy is feasible as Q is a factorized distribution.

High-Level Road Map: Energy Functional

- Suppose we want to approximate P with Q
  - Represent P by factors F $P_F(\mathbf{U}) = \frac{1}{Z}\prod_{\phi\in F}\phi(\mathbf{U}_\phi)$
  - Distance metric? – Many ways, but let's use relative entropy (aka KL-divergence)

$$D(Q\,\|\,P_F) = E_Q\left[\ln\frac{Q}{P_F}\right]$$

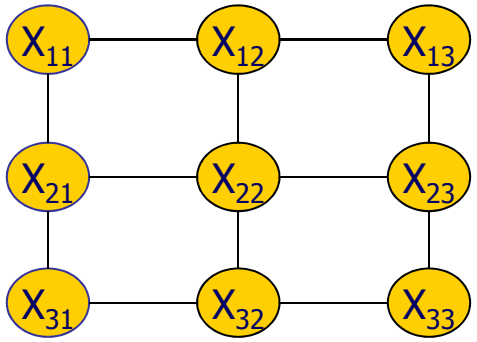
  - Define the energy functional $F[P_F, Q] = \sum_{\phi\in F} E_Q[\ln\phi] + H_Q(\mathbf{U})$
- Then, we can show that $D(Q\,\|\,P_F) = \ln Z - F[P_F, Q]$

(See proof in two slides later.)

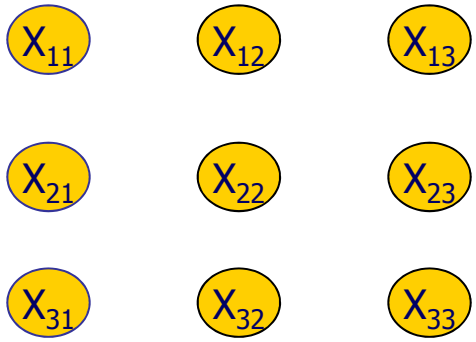
- Minimizing $D(Q||P_F)$ is equivalent to maximizing $F[P_F,Q]$
- $\ln Z \geq F[P_F,Q]$ (since $D(Q||P_F)\geq 0$)

Mean Field Approximation: General Idea

- Basic idea: The problem is viewed as maximizing the energy functional $F[P_{F'},Q]$
- Define a distribution Q over clique potentials
- Select a simple family of distributions **Q**
- Find $Q \in \mathbf{Q}$ that maximizes $F[P_F,Q]$



P_F – Markov grid network



Q – Mean field network

- $Q(x) \propto \prod \Psi(X_i)$, hence Q loses information in distribution P_F, but approximation is computationally easier.
  - Every query in Q is easier to calculate.

Reverse KL & Energy function

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{D_i} \Psi(D_i)$$

$$KL(Q||P) = \sum_x Q(X) \ln \frac{Q(X)}{P(X)} = \sum_x Q(X) \ln Q(X) - \sum_x Q(X) \ln P(X)$$

$$= \sum_x Q(X) \ln Q(X) - \sum_{D_i} \sum_x Q(X) \ln \Psi(D_i) + \ln Z$$

Energy function: $-F(P, Q)$ log partition function

$$\ln Z = F(P, Q) + KL(Q||P)$$

Equivalent to Maximize this

nonnegative Minimize it

lnZ does not depend on Q. Hence, minimizing the relative entropy KL(Q||P) is equivalent to maximizing the energy functional F (P, Q).

- Ln $Z \geq F(P,Q)$, energy function is a lower bound of log partition function. We maximize $F(P,Q)$ to find a tight lower bound on Ln Z (i.e., minimum KL-divergence).
- Z is the probability of the evidence in directed GM.

Optimization for Mean Field

- $max_Q\ F(P,Q) = \max_Q \sum_{D_i} E_Q[\ln \Psi(D_i)]\ + H(Q)$
  - $subject\ to\ \forall i, \sum_{x_i} Q_i(X_i) = 1\ , and\ Q_i(X_i) \geq 0$

- Constrained optimization
  - Add λ, form Lagrangian multipliers
  - Take derivative, set to zeros

Objective function is concave, H is strictly concave, but the parameter set Q itself is not convex set (it is in a product form) for our factorized family of Q, i.e., Non-convex Optimization.

- Q is a stationary point of mean field approximation iff $\forall i$:

$$Q_i(X_i) = \frac{1}{Z_i} \exp\left(\sum_{D_j : X_i \in D_j} E_Q\left[\ln \Psi\left(D_j\right)\right] \right)$$

Excluding $Q_i(X_i)$

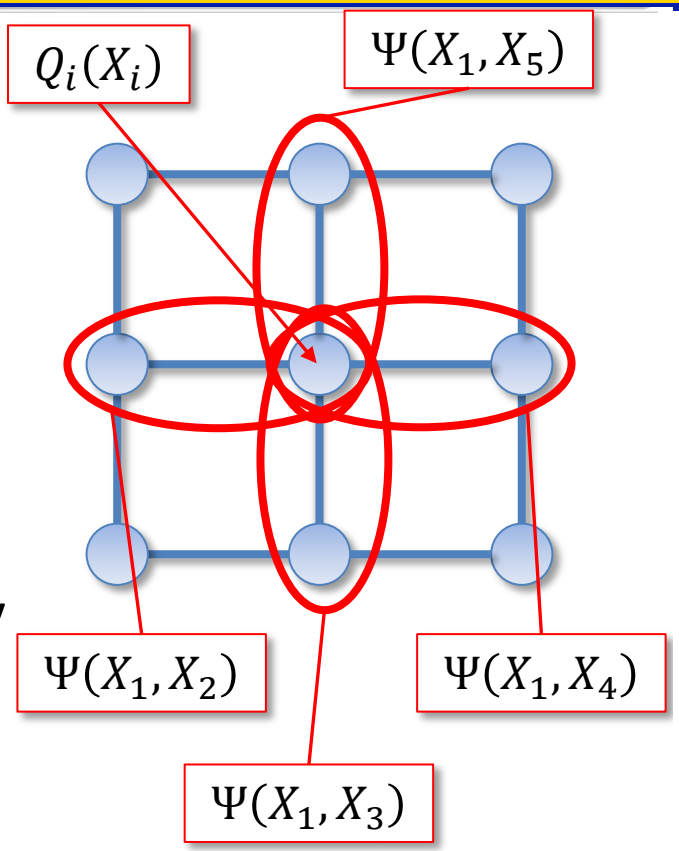
(Proof: Chapter 11 Koller, pages 451-452, see the proof in the continuous variable case in later slides)

Example: Mean-Field Fix Point Equation

- Q_i only needs to consider factors that intersect X_i

$$Q_i(X_i) = \frac{1}{Z_i} \exp\left(\sum_{D_j : X_i \in D_j} E_Q[\ln \Psi(D_j)] \right)$$

Excluding $Q_i(X_i)$



- It is like fixing the variables in the Markov blanket to their average (mean)
- Then renormalize
- Eg.

$$Q_1(X_1) = \frac{1}{Z_1} \exp\left(E_{Q(X_2)}[\ln \Psi(X_1, X_2)] + E_{Q(X_3)}[\ln \Psi(X_1, X_3)] + E_{Q(X_4)}[\ln \Psi(X_1, X_4)] + E_{Q(X_5)}[\ln \Psi(X_1, X_5)] \right)$$

Iterative Approach for Finding Stationary Point

- Initialize Q (eg., randomly or judiciously)
- Set all variables to unprocessed
- Pick an unprocessed variable X_i
  - Update Q_i

$$Q_i(X_i) = \frac{1}{Z_i} \exp\left(\sum_{D_j : X_i \in D_j} E_Q\left[\ln \Psi(D_j)\right] \right)$$

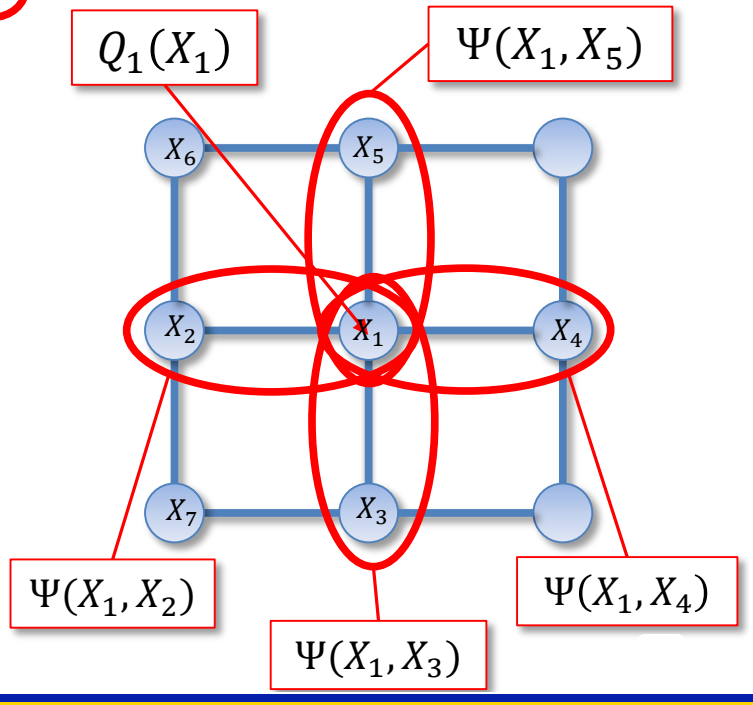
Excluding $Q_i(X_i)$

  - Set variable X_i as processed
  - If Q_i has changed
    - Then set neighbors of X_i to unprocessed
- Guaranteed to converge (to local maxima)

Deriving Mean Field Fixed Point Condition (I)

- Approximate $P(X) \propto \prod_j \Psi(D_j)$ with $Q(X) = \prod_i Q_i(X_i)$
  - $s.t.\ \forall X_i,\ Q_i(X_i) > 0, \int Q_i(X_i) dX_i = 1$
  - $D(Q||P) = -\int Q(X) \ln P(X) dX + \int Q(X) \ln Q(X) dX + const.$
  - $= -\int \prod_i Q_i(X_i) (\sum_j \ln \Psi(D_j)) dX + \int \prod_i Q_i(X_i) (\sum_i \ln Q_i(X_i)) dX$
  - $= -\sum_j \int \prod_{i \in D_j} Q_i(X_i) \ln \Psi(D_j) dX_{D_j} + \sum_i \int Q_i(X_i) \ln Q_i(X_i) dX_i$

E.g., $\int \prod_i Q_i(X_i) \ln \Psi(X_1, X_5)\ dX =$
$\int Q_1(X_1) Q_5(X_5) \ln \Psi(X_1, X_5) dX_1 dX_5$



-Song

Deriving Mean Field Fixed Point Condition (II)

- Let L be the Lagrangian function

$$L = -\sum_j \int \prod_{i \in D_j} Q_i(X_i) \ln \Psi(D_j) dX_{D_j} + \sum_i \int Q_i(X_i) \ln Q_i(X_i) dX_i - \sum_i \beta_i (1 - \int Q_i(X_i)) dX_i$$

- If Q is a maximum for the original constrained problem, then there exists β_i such that (Q, β_i) is a stationary point for the Lagrange function (set derivative equal to 0).

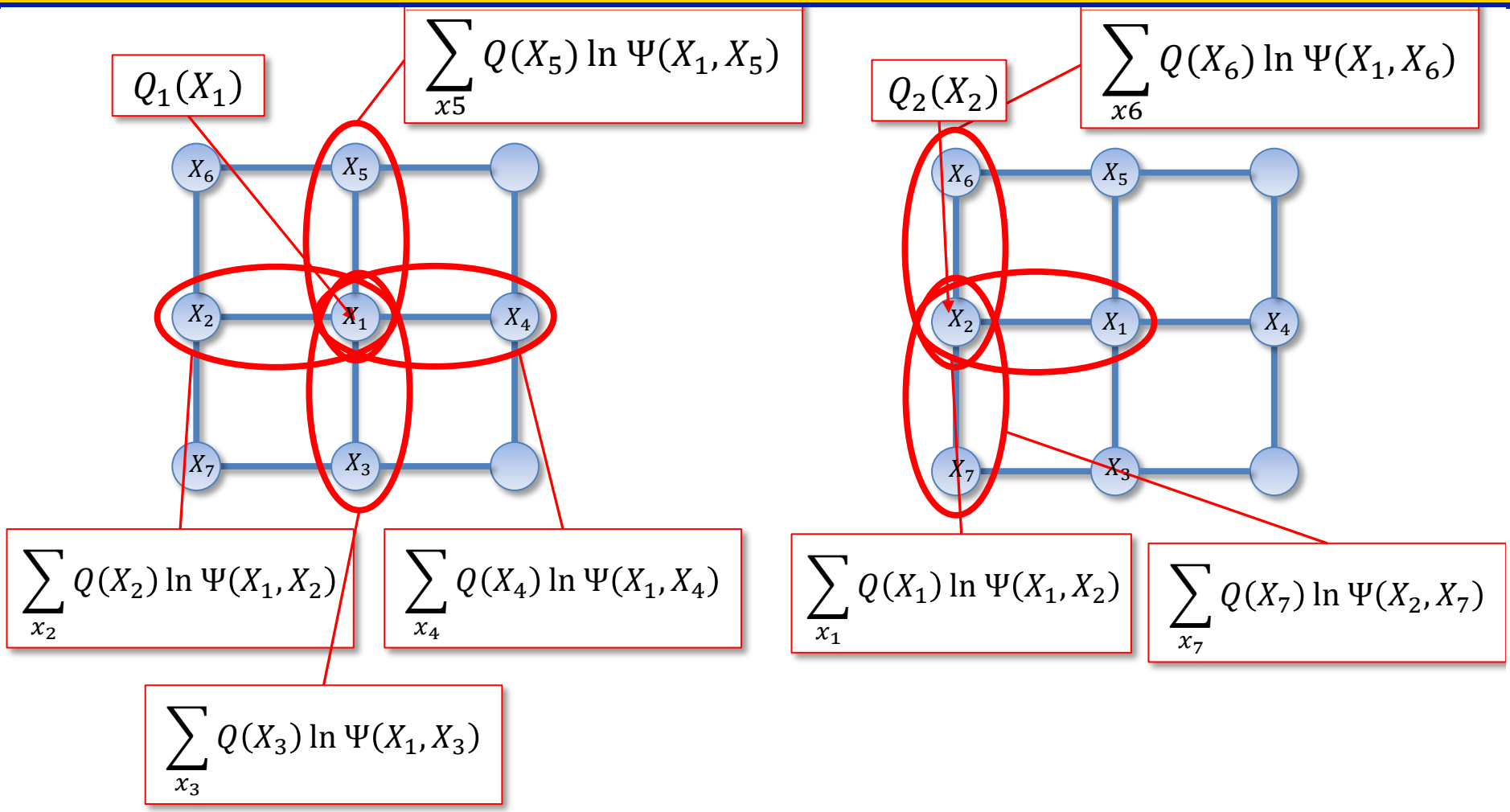
$$\frac{\partial L}{\partial Q_i(X_i)} = -\sum_{j: i \in D_j} \int \prod_{k \in D_j \setminus i} Q_k(X_k) \ln \Psi(D_j)) + \ln Q_i(X_i) - \beta_i = -1$$

- $$Q_i(X_i) = \exp\left(\sum_{j: i \in D_j} \int \prod_{k \in D_j \setminus i} Q_k(X_k) \ln \Psi(D_j)) + \beta_i'\right) =$$
$$\frac{1}{Z_i} \exp\left(\sum_{j: i \in D_j} \int \prod_{k \in D_j \setminus i} Q_k(X_k) \ln \Psi(D_j))\right)$$

Excluding $Q_i(X_i)$

$$Q_i(X_i) = \frac{1}{Z_i} \exp\left(\sum_{D_j: X_i \in D_j} E_Q[\ln \Psi(D_j)]\right)$$

Example: Mean-Field Fix Point Equation



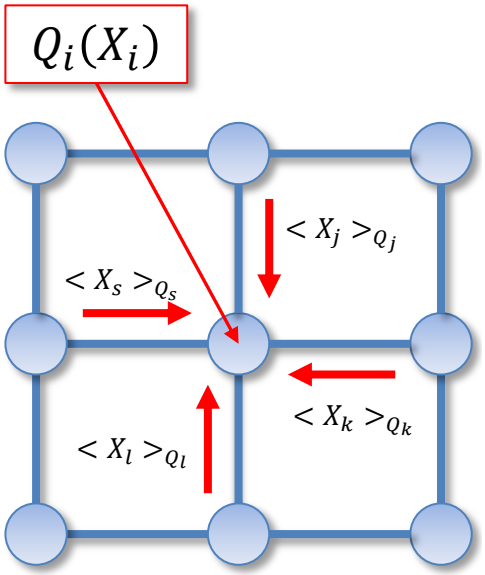
$$Q_1(X_1) = \frac{1}{Z_1} \exp\left(E_{Q(X_2)}[\ln \Psi(X_1, X_2)] + E_{Q(X_3)}[\ln \Psi(X_1, X_3)] + E_{Q(X_4)}[\ln \Psi(X_1, X_4)] + E_{Q(X_5)}[\ln \Psi(X_1, X_5)]\right)$$

$$Q_2(X_2) = \frac{1}{Z_1} \exp\left(E_{Q(X_1)}[\ln \Psi(X_1, X_2)] + E_{Q(X_6)}[\ln \Psi(X_2, X_6)] + E_{Q(X_7)}[\ln \Psi(X_2, X_7)]\right)$$

Example: Pair-Wise MN

- Mean field equation
  - $Q_i(X_i) = \frac{1}{Z_i} \exp\left(\sum_{D_j : X_i \in D_j} E_Q\left[\ln \Psi(D_j)\right]\right)$

Excluding $Q_i(X_i)$



- Pairwise Markov random field
  - $P(X_1, \ldots, X_n) \propto \exp\left(\sum_{ij} \theta_{ij} X_i X_j + \sum_i \theta_i X_i\right)$
  - $= \prod_{ij} \exp(\theta_{ij} X_i X_j) \prod_i \exp(\theta_i X_i)$

$\Psi(X_i, X_j)$
$\ln \Psi(X_i, X_j) = \theta_{ij} X_i X_j$

$\Psi(X_i)$
$\ln \Psi(X_i) = \theta_i X_i$

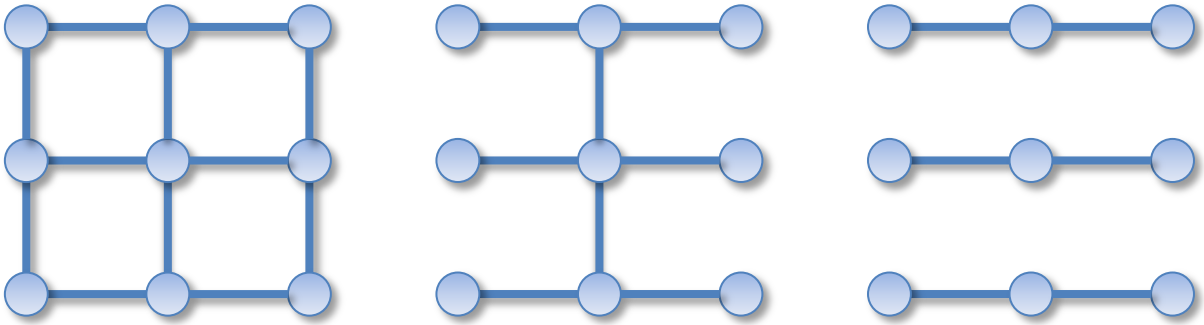
- The mean field equation has simple form:
  - $Q_i(X_i) = \frac{1}{Z_i} \exp\left(\sum_{j \in N(i)} \sum_{x_j} \theta_{ij} X_i X_j Q_j(X_j) + \theta_i X_i\right)$
  - $= \frac{1}{Z_i} \exp\left(\sum_{j \in N(i)} \theta_{ij} X_i < X_j >_{Q_j} + \theta_i X_i\right)$

It is like fixing variables in the Markov blanket to their average (mean)

General (Structured) Approximation

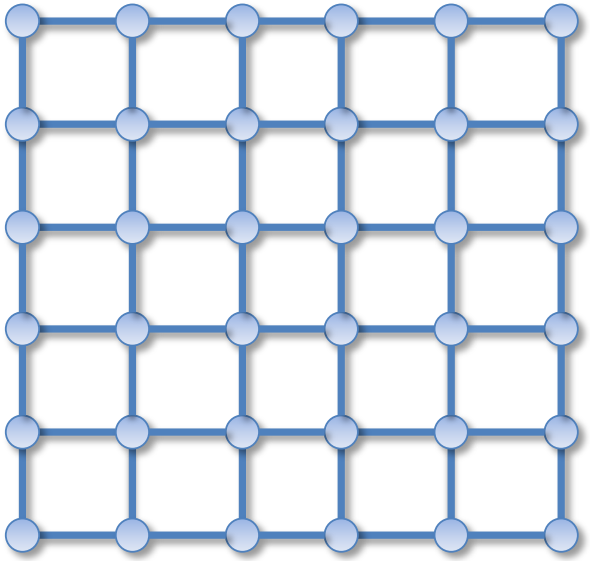
- Mean field method is naïve approximation (used factorized Q).
  - It is simple and easy to implement, hence still used very often.

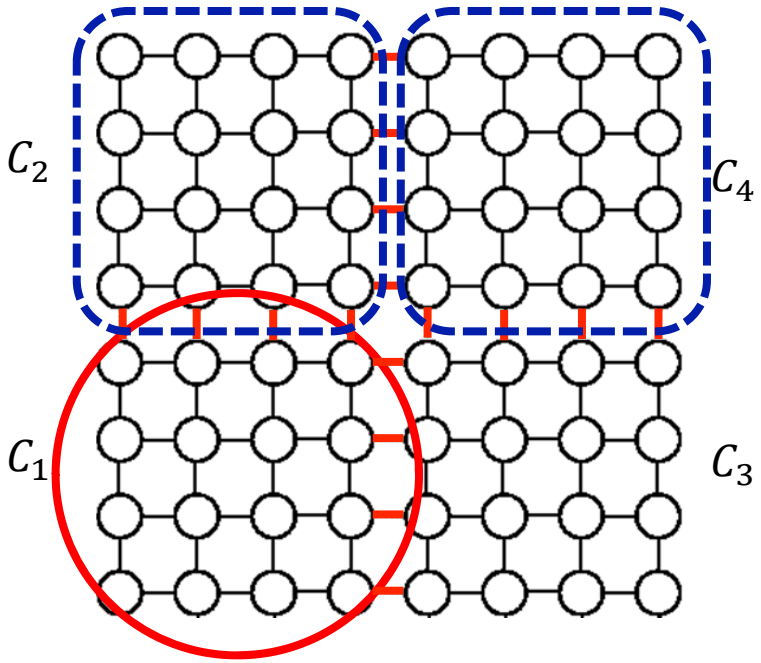
- One may consider more general form for Q:
  - While still exact inference is doable over Q



Generalized Mean Field: Stationary Point

- Solution for Q: $Q(C_i) \propto \exp\left(\sum_{D_j : C_i \cap D_j \neq \emptyset} E_Q[\ln \Psi(D_j)]\right)$





$$Q(C_1) \propto \exp\left(\sum_{i \in C_1} \theta_i X_i + \sum_{(ij) \in E, i \in C_1, j \in C_1} \theta_{ij} X_i X_j + \sum_k \sum_{i \in C_1, j \in C_k, j \in MB(C_1)} \theta_{ij} X_i < X_j >_{Q(C_k)}\right)$$

Node potential within C_1

Edge potential within C_1

Mean of variables in Markov blanket

Edge potential across C_1 and C_k