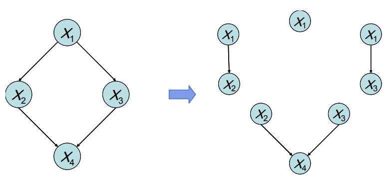
Machine Learning

Algorithms and Theory of Approximate Inference

Eric Xing

Lecture 15, August 15, 2010



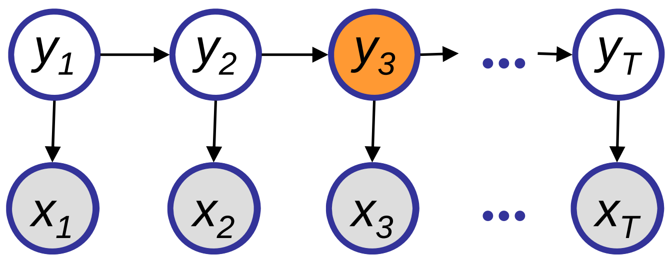
Reading:

Inference Problems

- Compute the likelihood of observed data
- Compute the marginal distribution $p(x_A)$ over a particular subset of nodes $A \subset V$
- Compute the conditional distribution $p(x_A|x_B)$ for disjoint subsets *A* and *B*
- Compute a mode of the density $\hat{x} = \arg\max_{x \in \mathcal{X}^m} p(x)$

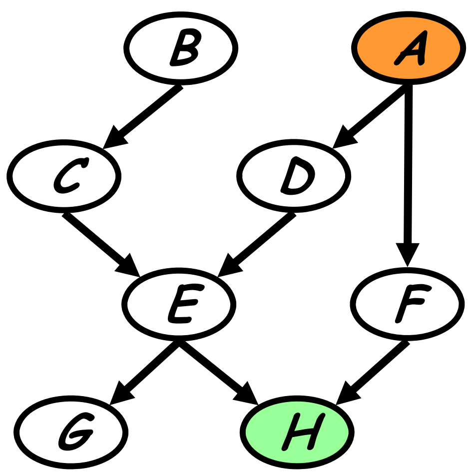
Inference in GM

- HMM



$$P(Y_3|\mathbf{x}) = ?$$

- A general BN



$$P(A|H) = ?$$

Inference Problems

- Compute the likelihood of observed data
- Compute the marginal distribution $p(x_A)$ over a particular subset of nodes $A \subset V$
- Compute the conditional distribution $p(x_A|x_B)$ for disjoint subsets *A* and *B*
- Compute a mode of the density $\hat{x} = \arg \max_{x \in \mathcal{X}^m} p(x)$

- Methods we have

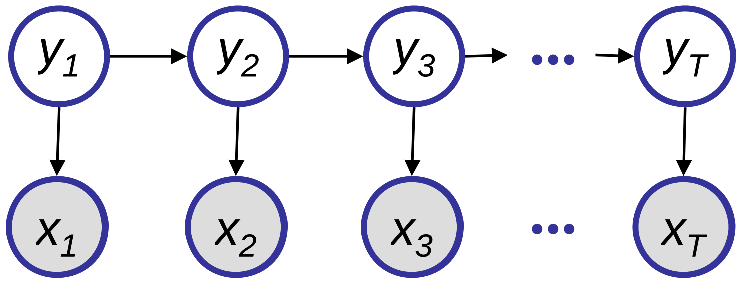
Brute force **Elimination** → **Message Passing**

(Forward-backward , Max-product /BP, Junction Tree)

Individual computations independent

Sharing intermediate terms

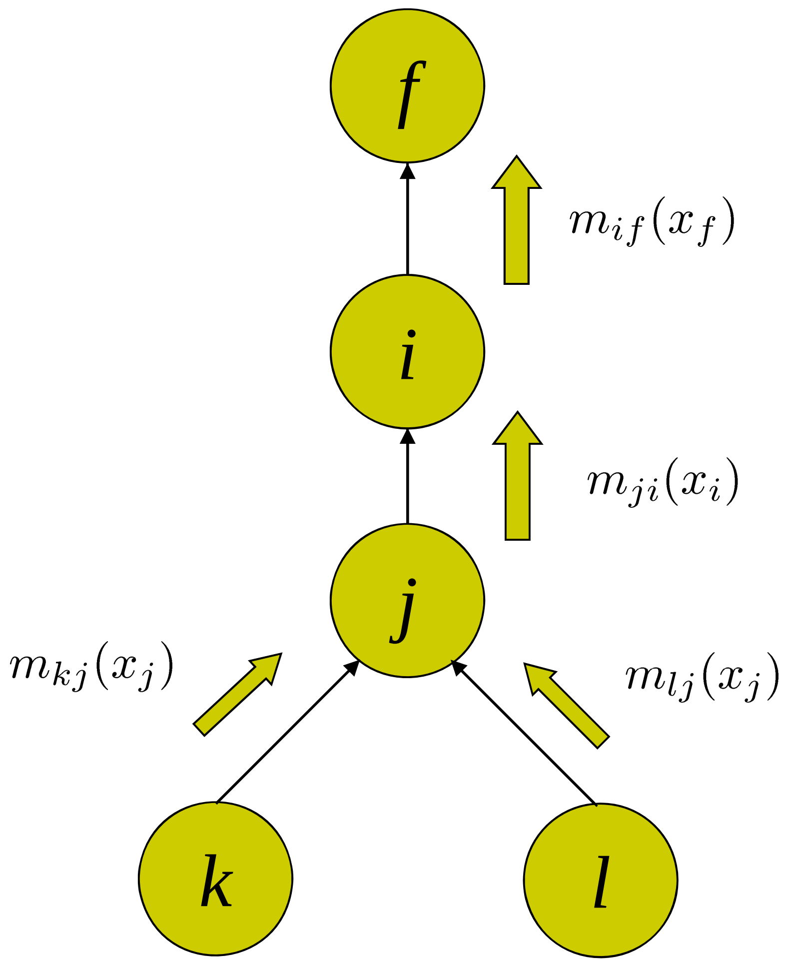
Recall forward-backward on HMM



- Forward algorithm $$\alpha_t^k = p(x_t \mid y_t^k = 1) \sum_i \alpha_{t-1}^i a_{i,k}$$
- Backward algorithm $$\beta_t^k = \sum_i a_{k,i} p(x_{t+1} \mid y_{t+1}^i = 1) \beta_{t+1}^i$$

$$P(y_t^k = 1 \mid \mathbf{x}) = \frac{P(y_t^k = 1, \mathbf{x})}{P(\mathbf{x})} = \frac{\alpha_t^k \beta_t^k}{P(\mathbf{x})}$$

Message passing for trees



Let $m_{ij}(x_i)$ denote the factor resulting from eliminating variables from bellow up to i, which is a function of x_i:

$$m_{ji}(x_i) = \sum_{x_j} \left(\psi(x_j)\psi(x_i, x_j) \prod_{k \in N(j)\setminus i} m_{kj}(x_j) \right)$$

This is reminiscent of a ***message*** sent from j to i.

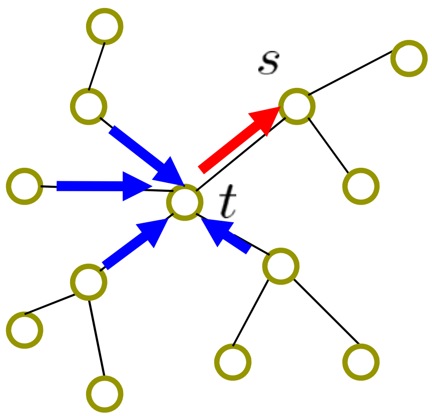
$$p(x_f) \propto \psi(x_f) \prod_{e \in N(f)} m_{ef}(x_f)$$

$m_{ij}(x_i)$ represents a "belief" of x_i from x_j!

The General Sum-Product Algorithm

- Tree-structured GMs

$$p(x_1, \cdots, x_m) = \frac{1}{Z} \prod_{s \in V} \psi_s(x_s) \prod_{(s,t) \in E} \psi_{st}(x_s, x_t)$$



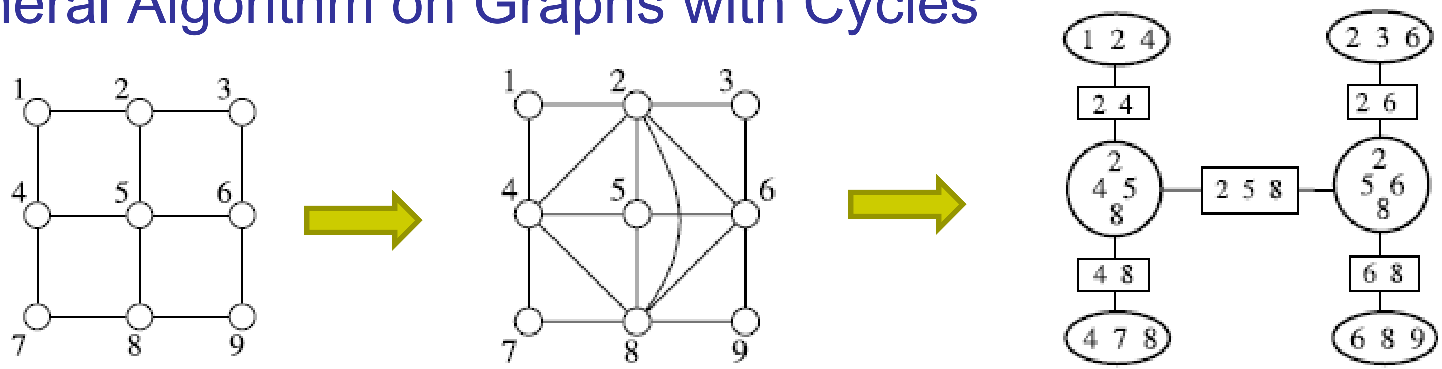
- Message Passing on Trees:

$$M_{t \to s}(x_s) \leftarrow \kappa \sum_{x'_t} \left\{ \psi_{st}(x_s, x'_t) \psi_t(x'_t) \prod_{u \in N(t) \backslash s} M_{u \to t}(x'_t) \right\}$$

  - On trees, converge to a unique fixed point after a finite number of iterations

Junction Tree Revisited

- General Algorithm on Graphs with Cycles

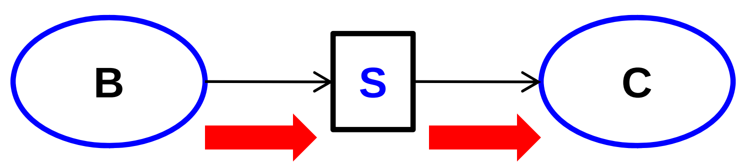


- Steps: => **Triangularization** => **Construct JTs**

=> Message Passing on Clique Trees

$$\tilde{\phi}_S(x_S) \leftarrow \sum_{x_{B\setminus S}} \phi_B(x_B)$$

$$\phi_C(x_C) \leftarrow \frac{\tilde{\phi}_S(x_S)}{\phi_S(x_S)} \phi_C(x_C)$$



Local Consistency

- Given a set of functions $\{\tau_C,\ C \in \mathcal{C}\}$ and $\{\tau_S,\ S \in \mathcal{S}\}$ associated with the cliques and separator sets

- They are locally consistent if:

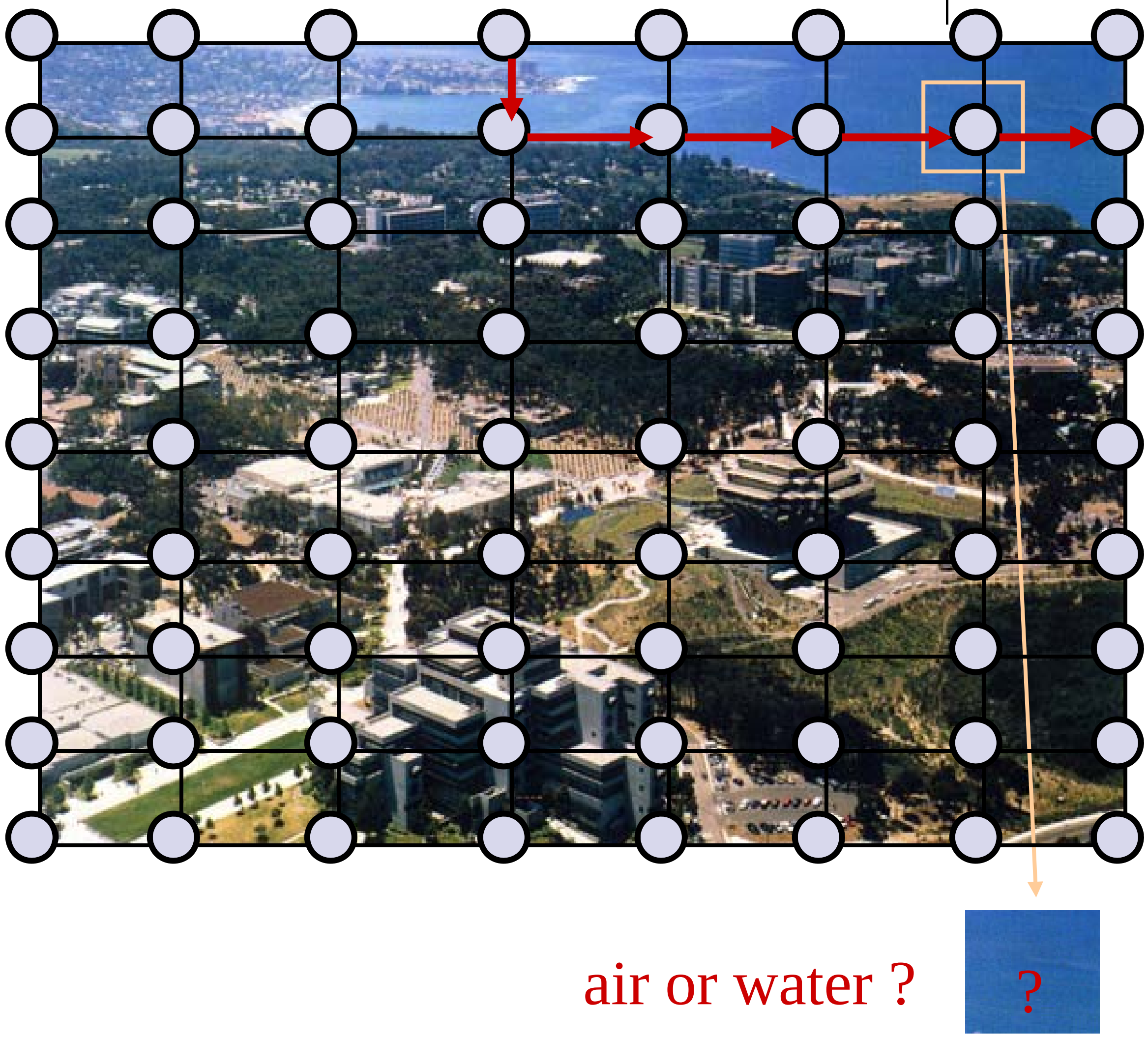
$$\sum_{x'_S} \tau_S(x'_S) = 1,\ \forall S \in \mathcal{S}$$

$$\sum_{x'_C | x'_S = x_S} \tau_C(x'_C) = \tau_S(x_S),\ \forall C \in \mathcal{C},\ S \subset C$$

- For junction trees, local consistency is equivalent to global consistency!

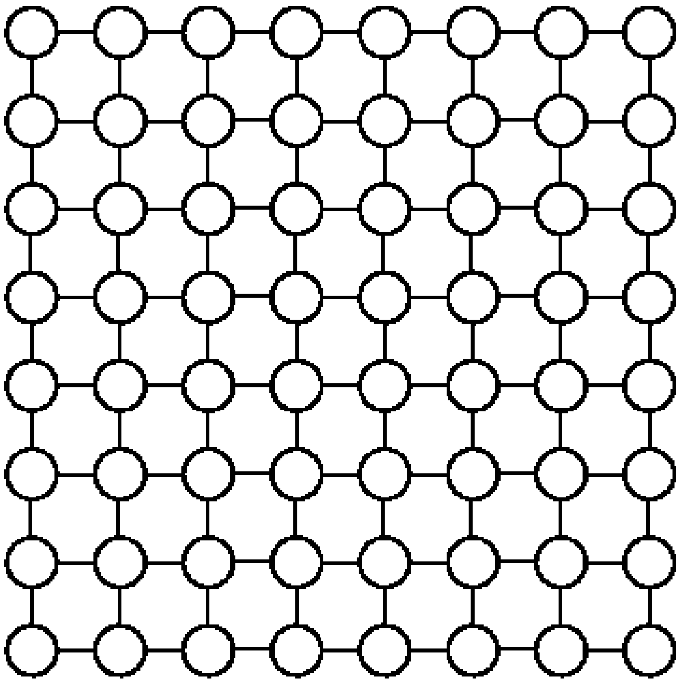
An Ising model on 2-D image

- Nodes encode hidden information (patch-identity).
- They receive local information from the image (brightness, color).
- Information is propagated though the graph over its edges.
- Edges encode 'compatibility' between nodes.



Why Approximate Inference?

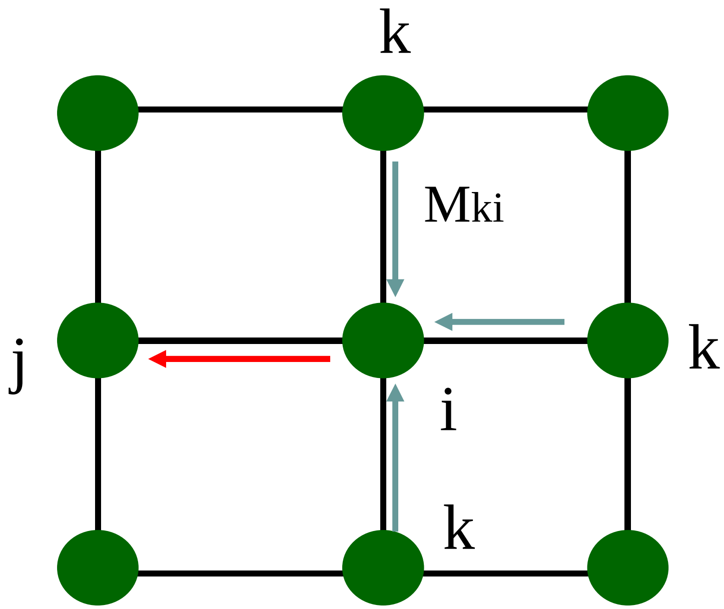
- Why can't we just run junction tree on this graph?

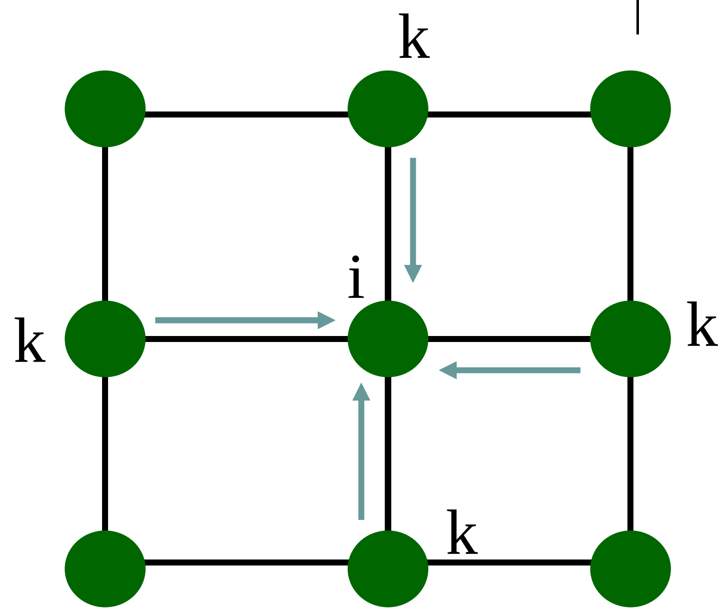


$$p(X) = \frac{1}{Z}\exp\left\{\sum_{i<j}\theta_{ij}X_iX_j + \sum_i \theta_{i0}X_i\right\}$$

- If NxN grid, tree width at least N
- N can be a huge number(~1000s of pixels)
  - If N~O(1000), we have a clique with 2^{100} entries

Solution 1: Belief Propagation on loopy graphs





- BP Message-update Rules

$$M_{i\to j}(x_j) \propto \sum_{x_i} \psi_{ij}(x_i, x_j)\psi_i(x_i)\prod_k M_{k\to i}(x_i)$$

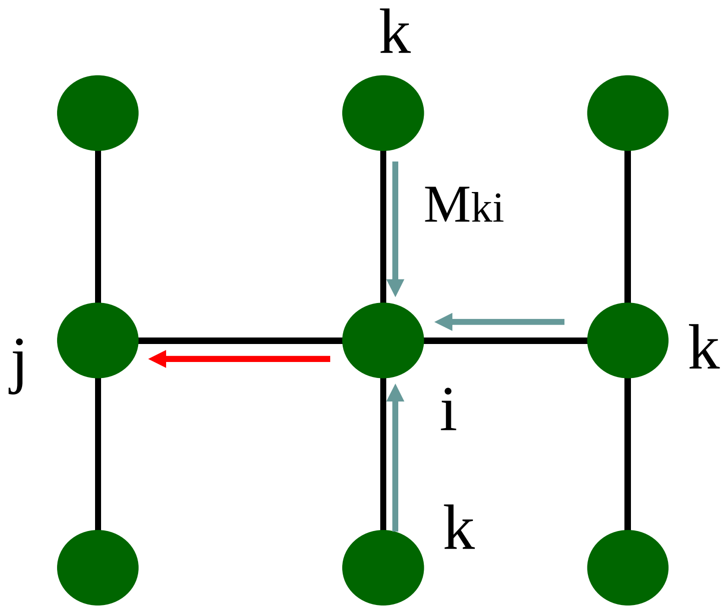
external evidence

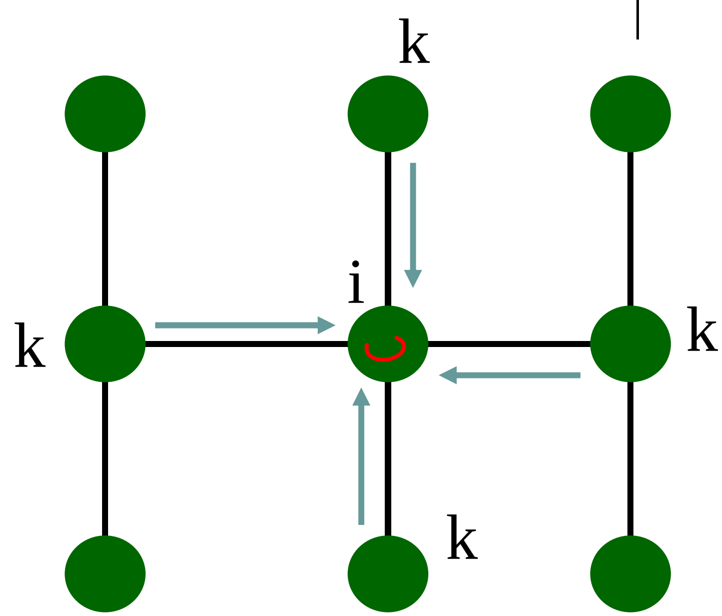
Compatibilities (interactions)

$$b_i(x_i) \propto \psi_i(x_i)\prod_k M_k(x_k)$$

- May not converge or converge to a wrong solution

Recall BP on trees





- BP Message-update Rules

$$M_{i\to j}(x_j) \propto \sum_{x_i} \psi_{ij}(x_i, x_j)\psi_i(x_i)\prod_k M_{k\to i}(x_i)$$

external evidence

Compatibilities (interactions)

$$b_i(x_i) \propto \psi_i(x_i)\prod_k M_k(x_k)$$

- BP on **trees** always converges to exact marginals

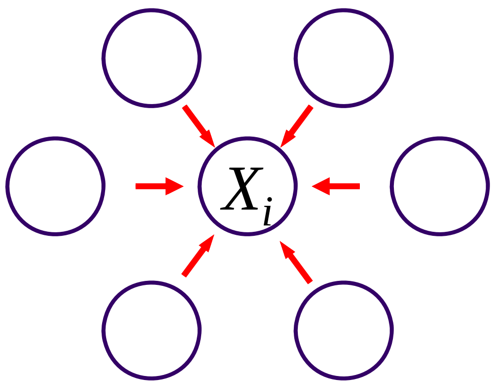
Solution 2: The naive mean field approximation

- Approximate $p(\mathbf{X})$ by fully factorized $q(\mathbf{X})=\Pi_i q_i(X_i)$

- For Boltzmann distribution $p(X)=\exp\{\sum_{i<j} q_{ij}X_iX_j+q_{io}X_i\}/Z$:

mean field equation:

$$q_i(X_i) = \exp\left\{\theta_{i0}X_i + \sum_{j\in\mathcal{N}_i}\theta_{ij}X_i\langle X_j\rangle_{q_j} + A_i\right\}$$
$$= p(X_i \mid \{\langle X_j\rangle_{q_j} : j\in\mathcal{N}_i\})$$



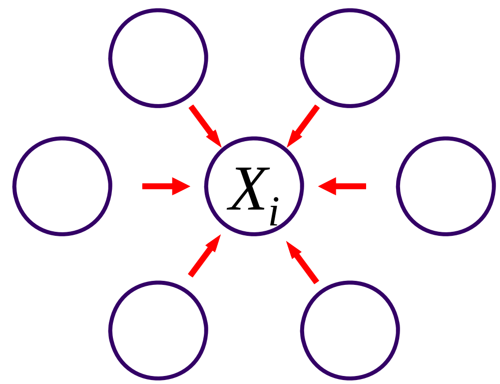
- $\langle X_j\rangle_{q_j}$ resembles a "message" sent from node j to i
- $\{\langle X_j\rangle_{q_j} : j\in\mathcal{N}_i\}$ forms the "mean field" applied to X_i from its neighborhood

Recall Gibbs sampling

- Approximate $p(\mathbf{X})$ by fully factorized $q(\mathbf{X})=\Pi_i q_i(X_i)$
- For Boltzmann distribution $p(X)=\exp\{\sum_{i<j} q_{ij}X_iX_j+q_{io}X_i\}/Z$:

Gibbs predictive distribution:

$$p(X_i \mid x_{-i}) = \exp\left\{\theta_{i0}X_i + \sum_{j\in\mathcal{N}_i}\theta_{ij}X_i x_j + A_i\right\}$$
$$= p(X_i \mid \{x_j : j \in \mathcal{N}_i\})$$



Summary So Far

- Exact inference methods are limited to tree-structured graphs
- Junction Tree methods is exponentially expensive to the tree-width
- Message Passing methods can be applied for loopy graphs, but lack of analysis!
- Mean-field is convergent, but can have local optimal
- **Where do these two algorithm come from? Do they make sense?**

Next Step …

- Develop a general theory of variational inference
- Introduce some approximate inference methods
- Provide deep understandings to some popular methods

Exponential Family GMs

- Canonical Parameterization

$$p_\theta(x_1, \cdots, x_m) = \exp\left\{\theta^\top \phi(x) - A(\theta)\right\}$$

Canonical Parameters **Sufficient Statistics** **Log-normalization Function**

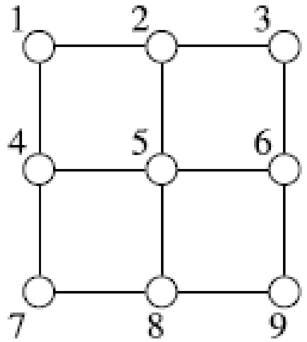
- Effective canonical parameters
- Regular family: $\Omega := \left\{\theta \in \mathbb{R}^d | A(\theta) < +\infty\right\}$

  Ω is an open set.
- Minimal representation:
  - if there does not exist a nonzero vector $a \in \mathbb{R}^d$ such that $a^\top \phi(x)$ is a constant

Examples

- Ising Model (binary r.v.: {-1, +1})

$$p_\theta(x) = \exp\left\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t - A(\theta) \right\}$$



- Gaussian MRF

$$p_\theta(x) = \exp\left\{ \sum_{s \in V} \theta_s x_s + \frac{1}{2} \mathrm{Tr}(\Theta x x^\top) - A(\theta) \right\}$$

$$\Omega = \left\{ (\theta, \Theta) \in \mathbb{R}^m \times \mathbb{R}^{m \times m} | \Theta \prec 0,\ \Theta^\top = \Theta \right\}$$

Mean Parameterization

- The mean parameter μ_α associated with a sufficient statistic $\phi_\alpha : \mathcal{X}^m \to \mathbb{R}$ is defined as

- Realizable mean parameter set

$$\mathcal{M} := \left\{ \mu \in \mathbb{R}^d \,|\, \exists p \text{ s.t. } \mathbb{E}_p[\phi_\alpha(X)] = \mu_\alpha, \ \forall \alpha \in \mathcal{I} \right\}$$

  - A convex subset of $\mathbb{R}^d$
  - Convex hull for discrete case

$$\mathcal{M} = \left\{ \mu \in \mathbb{R}^d \,|\, \sum_{x \in \mathcal{X}^m} \phi(x) p(x) = \mu, \text{ for some } p(x) \geq 0, \sum_{x \in \mathcal{X}^m} p(x) = 1 \right\}$$

$$\triangleq \text{conv}\left\{ \phi(x), x \in \mathcal{X}^m \right\}$$

  - Convex polytope when $|\mathcal{X}^m|$ is finite

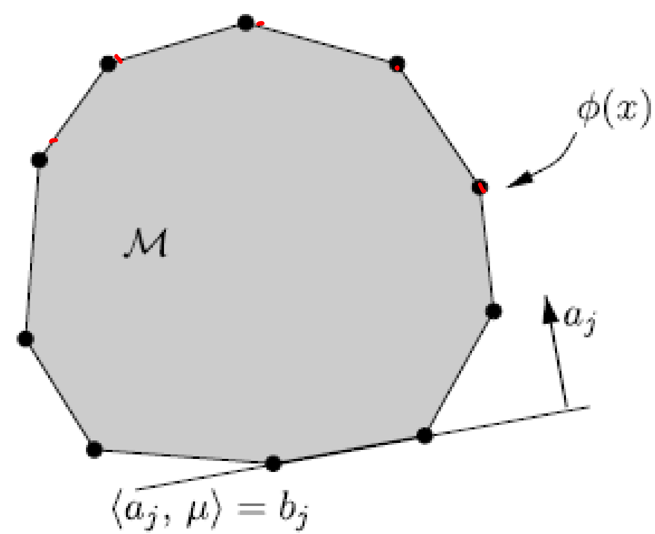
Convex Polytope

- Convex hull representation

$$\mathcal{M} = \text{conv}\left\{\phi(x), x \in \mathcal{X}^m\right\}, \text{ where } |\mathcal{X}^m| \text{ is finite.}$$

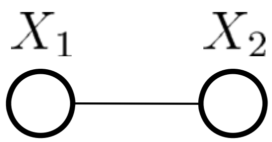
- Half-plane based representation
  - Minkowski-Weyl Theorem:
    - any polytope can be characterized by a finite collection of linear inequality constraints

$$\mathcal{M} = \left\{\mu \in \mathbb{R}^d | a_j^\top \mu \geq b_j, \ \forall j \in \mathcal{J}\right\},$$
$$\text{where } |\mathcal{J}| \text{ is finite.}$$



Example

- Two-node Ising Model

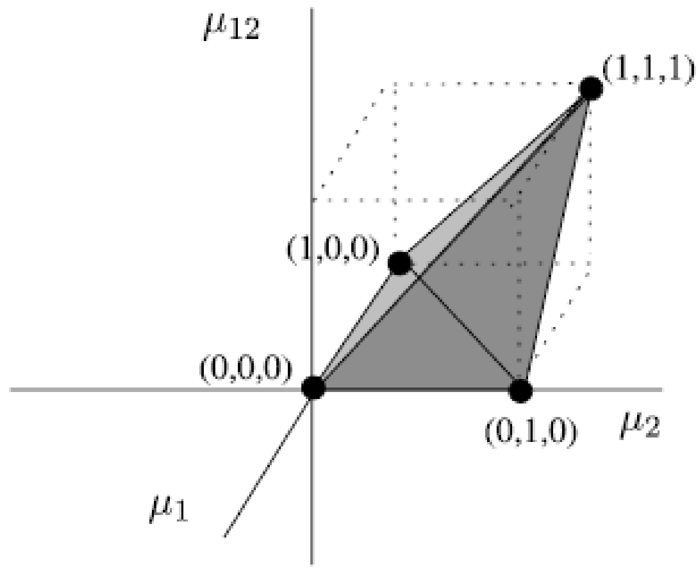


  - Convex hull representation

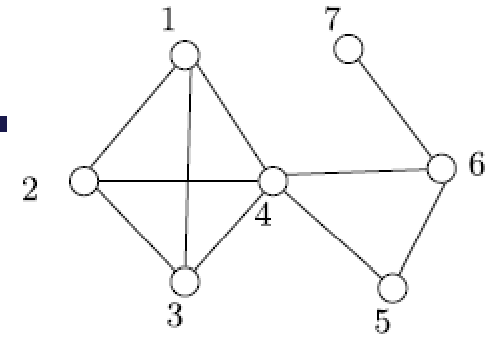
$$\mathcal{M} = \text{conv}\{(0,0,0),(1,0,0),(0,1,0),(1,1,1)\}$$

  - Half-plane representation
    - Probability Theory:

$$\mu_i \geq \mu_{12} \geq 0 \qquad 1 + \mu_{12} - \mu_1 - \mu_2 \geq 0$$



Marginal Polytope



- Canonical Parameterization

$$p_\theta(x) \propto \exp\{\sum_{v\in V}\theta_v(x_v) + \sum_{(s,t)\in E}\theta_{st}(x_s,x_t)\}$$

$$\theta_s(x_s) := \sum_j \theta_{s;j}\mathbb{I}_{s;j}(x_s) \qquad \theta_{st}(x_s,x_t) := \sum_{(j,k)}\theta_{st;jk}\mathbb{I}_{st;jk}(x_s,x_t)$$

- Mean parameterization

$$\mu_{s;j} = \mathbb{E}_p[\mathbb{I}_{s;j}(X_s)] = p(X_s = j), \quad \forall j \in \mathcal{X}_s$$

$$\mu_{st;jk} = \mathbb{E}_p[\mathbb{I}_{st;jk}(X_s,X_t)] = p(X_s = j, X_t = k), \quad \forall (j,k) \in \mathcal{X}_s \times \mathcal{X}_t$$

- Marginal distributions over nodes and edges

$$\mu_s(x_s) := \sum_{j\in\mathcal{X}_s}\mu_{s;j}\mathbb{I}_{s;j}(x_s) \qquad \mu_{st}(x_s,x_t) := \sum_{(j,k)\in\mathcal{X}_s\times\mathcal{X}_t}\mu_{st;jk}\mathbb{I}_{st;jk}(x_s,x_t)$$

- Marginal Polytope

$$\mathbb{M}(G) := \left\{\mu \in \mathbb{R}^d \,|\, \exists p \text{ with marginals } \mu_s(x_s),\ \mu_{st}(x_s,x_t)\right\}$$

Conjugate Duality

- Duality between MLE and Max-Ent:
  - For all $\mu \in \mathcal{M}^\circ$, a unique canonical parameter $\theta(\mu)$ satisfying

$$\mu = \nabla A(\theta(\mu)) = \mathbb{E}_{\theta(\mu)}[\phi(X)] \qquad A^*(\mu) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \notin \bar{\mathcal{M}} \end{cases}$$

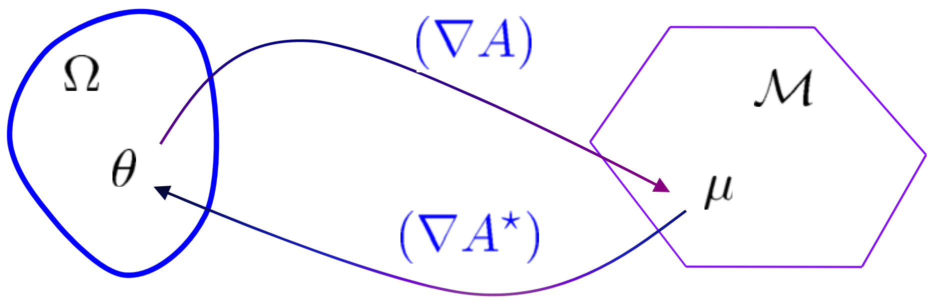
  - The log-partition function has the variational form

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\theta^\top \mu - A^*(\mu)\} \quad (*)$$

  - For all $\theta \in \Omega$, the supremum in (*) is attained uniquely at $\mu \in \mathcal{M}^\circ$ specified by the moment-matching conditions

$$\mu = \mathbb{E}_\theta[\phi(X)]$$

- Bijection for minimal exponential family

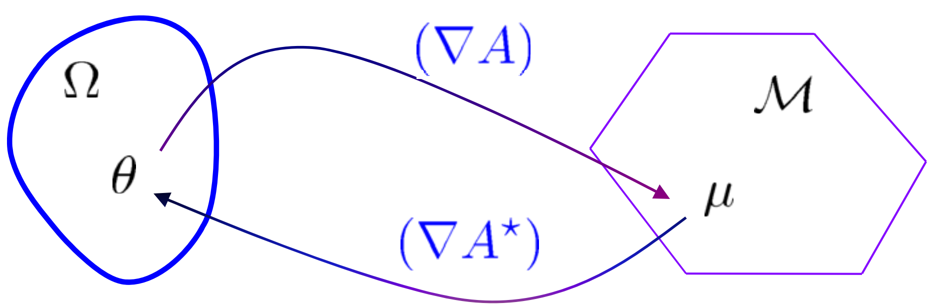


Roles of Mean Parameters

- Forward Mapping:
  - From $\theta \in \Omega$ to the mean parameters $\mu \in \mathcal{M}$
  - A fundamental class of inference problems in exponential family models

$$\sup_{\mu \in \mathcal{M}} \{\theta^\top \mu - A^*(\mu)\} \quad (*)$$

- Backward Mapping:
  - Parameter estimation to learn the unknown $\theta \in \Omega$



Example

- Bernoulli

$$\phi(x) = x, \;\; A(\theta) = \log(1+\exp(\theta)), \;\; \Omega = \mathbb{R}$$

$$A^{\star}(\mu) = \sup_{\theta \in \Omega} \{\theta^{\top}\mu - \log(1+\exp(\theta))\} \qquad (**)$$

$$\Rightarrow \quad \mu = \frac{\exp(\theta)}{1+\exp(\theta)} \quad \Big(\mu = \nabla A(\theta)\Big)$$

- If $\mu \in \mathcal{M}^{\circ} = (0,1)$ $\Rightarrow$ $\theta(\mu) = \log(\frac{\mu}{1-\mu})$ **Unique!**

$$A^{\star}(\mu) = \mu \log \mu + (1-\mu)\log(1-\mu)$$

- If $\mu \notin \bar{\mathcal{M}} = [0,1]$

No gradient stationary point in the Opt. problem (**)

$$A^{\star}(\mu) = +\infty$$

- Reverse mapping:

$$\mu = \arg\max_{\mu \in [0,1]} \{\mu^{\top}\theta - \mu\log\mu - (1-\mu)\log(1-\mu)\}$$

$$\Rightarrow \quad \mu(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)}, \quad A(\theta) = \log(1+\exp(\theta))$$ **Unique!**

Variational Inference In General

- An umbrella term that refers to various mathematical tools for optimization-based formulations of problems, as well as associated techniques for their solution

- General idea:
  - Express a quantity of interest as the solution of an optimization problem

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \theta^{\top} \mu - A^{*}(\mu) \right\} \quad (*)$$

  - The optimization problem can be relaxed in various ways
    - Approximate the functions to be optimized
    - Approximate the set over which the optimization takes place

- Goes in parallel with MCMC

A Tree-Based Outer-Bound to $\mathbb{M}(G)$

- Local Consistent (*Pseudo-*) Marginal Polytope

$$\tau := \{\tau_s,\ s \in V;\ \tau_{st},\ (s,t) \in E\}$$

$$\mathbb{L}(G) := \left\{\tau \geq 0 \,|\, \textit{normalization} \text{ and } \textit{marginalization constraints} \text{ hold.}\right\}$$

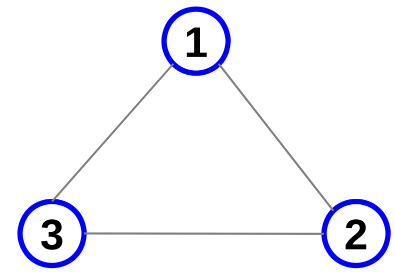
  - normalization $\sum_{x_s} \tau_s(x_s) = 1,\ \forall s \in V$
  - marginalization

$$\forall (s,t) \in E: \quad \sum_{x'_t} \tau_{st}(x_s, x'_t) = \tau_s(x_s),\ \forall x_s \in \mathcal{X}_s \quad \sum_{x'_s} \tau_{st}(x'_s, x_t) = \tau_t(x_t),\ \forall x_t \in \mathcal{X}_t$$

- Relation to $\mathbb{M}(G)$
  - $\mathbb{M}(G) \subseteq \mathbb{L}(G)$ holds for any graph
  - $\mathbb{M}(G) = \mathbb{L}(G)$ holds for tree-structured graphs

A $\mathbb{M}(G) \subset \mathbb{L}(G)$ Example

- A three node graph (binary r.v.)



$$\tau_s(x_s) := [0.5 \;\; 0.5]$$

$$\tau_{st}(x_s, x_t) := \begin{bmatrix} \beta_{st} & 0.5 - \beta_{st} \\ 0.5 - \beta_{st} & \beta_{st} \end{bmatrix}$$

- For any $\beta_{st} \in [0, 0.5]$, we have $\tau \in \mathbb{L}(G)$
- For $\beta_{12} = \beta_{23} = 0.4, \text{ and } \beta_{13} = 0.1$, we have $\tau \notin \mathbb{M}(G)$
  - an exercise?

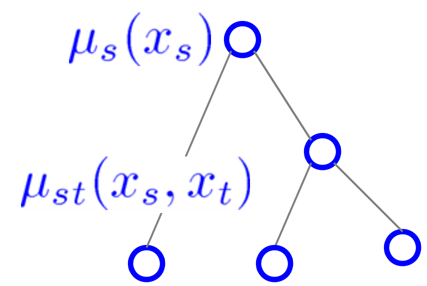
Bethe Entropy Approximation

- Approximate the negative entropy $A^{\star}(\mu)$, which doesn't has a closed-form in general graph.
- Entropy on tree (Marginals)
  - recall:

$$p_\mu = \prod_{s \in V} \mu_s(x_s) \prod_{(s,t)\in E} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}$$

  - entropy

$$H(p_\mu) = \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t)\in E} I_{st}(\mu_{st})$$



- Bethe entropy approximation (Pseudo-marginals)

$$-A^{\star}(\tau) \approx H_{\text{Bethe}}(\tau) := \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t)\in E} I_{st}(\tau_{st})$$

Bethe Variational Problem (BVP)

- We already have:
  - a convex (polyhedral) outer bound $\mathbb{L}(G)$

$$\mathbb{M}(G) \subseteq \mathbb{L}(G)$$

  - the Bethe approximate entropy

$$-A^{\star}(\tau) \approx H_{\text{Bethe}}(\tau) := \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st})$$

- Combining the two ingredients, we have

$$\max_{\tau \in \mathbb{L}(G)} \left\{ \theta^{\top} \tau + \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st}) \right\}$$

  - a simple structured problem (differentiable & constraint set is a simple polytope)
  - Max-product is the solver!

Nobel Prize in Physics (1967)

Connection to Sum-Product Alg.

- Lagrangian method for BVP:

$$\mathcal{L}(\tau, \lambda; \theta) := \theta^\top \tau + H_{\text{Bethe}}(\tau) + \sum_{s \in V} \lambda_{ss} C_{ss}(\tau)$$

$$+ \sum_{(s,t) \in E} \Big[\sum_{x_s} \lambda_{st}(x_s) C_{ts}(x_s; \tau) + \sum_{x_t} \lambda_{st}(x_t) C_{st}(x_t; \tau) \Big]$$

$$\text{where } C_{ss}(\tau) := 1 - \sum_{x_s} \tau_s(x_s), \;\; C_{st}(x_s; \tau) := \tau_s(x_s) - \sum_{x_t} \tau_{st}(x_s, x_t)$$

- Sum-product and Bethe Variational (Yedidia et al., 2002)
  - For any graph *G*, any fixed point of the sum-product updates specifies a pair of $(\tau^\star, \lambda^\star)$ such that

$$\nabla_\tau \mathcal{L}(\tau^\star, \lambda^\star; \theta) = 0, \quad \text{and} \quad \nabla_\lambda \mathcal{L}(\tau^\star, \lambda^\star; \theta) = 0$$

  - For a tree-structured MRF, the solution $(\tau^\star, \lambda^\star)$ is unique, where correspond to the exact singleton and pairwise marginal distributions of the MRF, and the optimal value of BVP is equal to $A(\theta)$

Proof

Discussions

- The connection provides a principled basis for applying the sum-product algorithm for loopy graphs

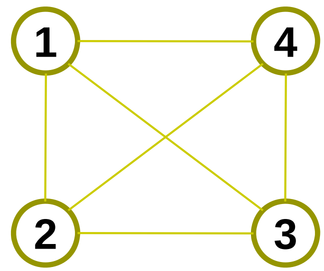
- However,
  - this connection provides no guarantees on the convergence of the sum-product alg. on loopy graphs
  - the Bethe variational problem is usually non-convex. Therefore, there are no guarantees on the global optimum
  - Generally, there are no guarantees that $A_{\text{Bethe}}(\theta)$ is a lower bound of $A(\theta)$

- However, however
  - the connection and understanding suggest a number of avenues for improving upon the ordinary sum-product alg., via progressively better approximations to the entropy function and outer bounds on the marginal polytope!

Inexactness of Bethe and Sum-Product

- From Bethe entropy approximation
  - Example $$\mu_s(x_s) = [0.5 \;\; 0.5]$$

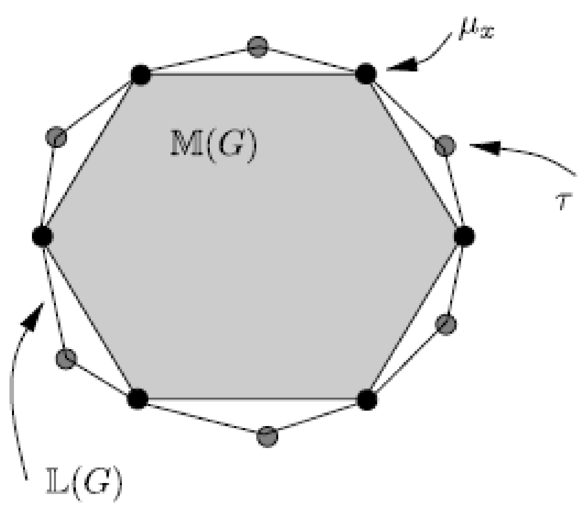
$$\mu_{st}(x_s, x_t) := \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$$



$$H_{\text{Bethe}}(\mu) = 4\log 2 - 6\log 2 = -2\log 2 < 0 \quad !!$$

$$\text{True entropy: } \log 2$$

- From pseudo-marginal outer bound
  - strict inclusion

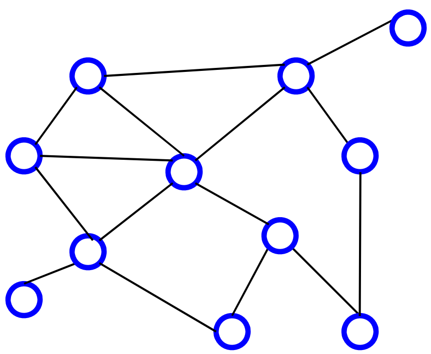


Summary of LBP

- Variational methods in general turn inference into an optimization problem
- However, both the objective function and constraint set are hard to deal with
- Bethe variational approximation is a tree-based approximation to both objective function and marginal polytope
- Belief propagation is a Lagrangian-based solver for BVP
- Generalized BP extends BP to solve the generalized hyper-tree based variational approximation problem

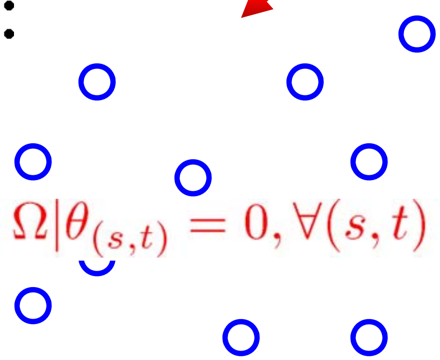
Tractable Subgraph

- Given a GM with a graph G, a subgraph F is tractable if
  - We can perform exact inference on it
- Example:

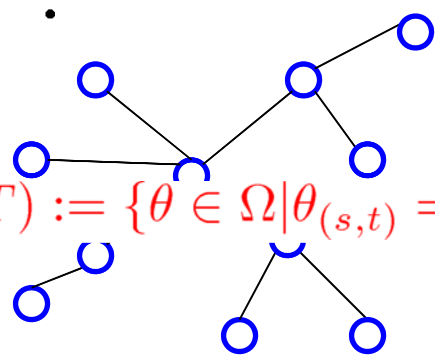


$$\Omega := \left\{\theta \in \mathbb{R}^d \,|\, A(\theta) < +\infty\right\}$$

F_0 :



T :



$$\Omega(F_0) := \{\theta \in \Omega | \theta_{(s,t)} = 0, \forall (s,t) \in E\} \quad \Omega(T) := \{\theta \in \Omega | \theta_{(s,t)} = 0 \ \forall (s,t) \notin E(T)\}$$

Mean Parameterization

- For an exponential family GM defined with graph *G* and sufficient statistics ϕ , the realizable mean parameter set

$$\mathcal{M}(G;\phi) := \left\{ \mu \in \mathbb{R}^d \middle| \ \exists p \text{ s.t. } \ \mathbb{E}_p[\phi_\alpha(X)] = \mu_\alpha, \ \forall \alpha \in \mathcal{I} \right\}$$

- For a given tractable subgraph *F*, a subset of mean parameters is of interest

$$\mathcal{M}_F(G;\phi) := \left\{ \mu \in \mathbb{R}^d \middle| \ \mu = \mathbb{E}_\theta[\phi(X)], \text{ for some } \theta \in \Omega(F) \right\}$$

- Inner Approximation

$$\mathcal{M}_F^\circ(G;\phi) \subseteq \mathcal{M}^\circ(G;\phi)$$

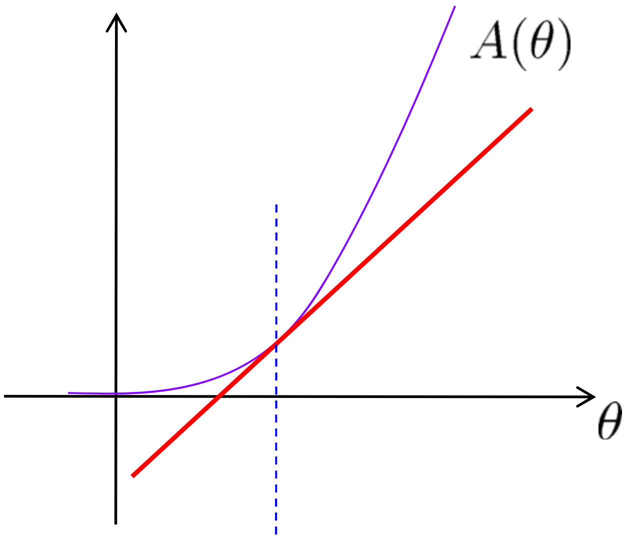
Optimizing a Lower Bound

- Any mean parameter $\mu \in \mathcal{M}^\circ$ yields a lower bound on the log-partition function

$$A(\theta) \geq \theta^\top \mu - A^\star(\mu)$$

  - Moreover, equality holds iff θ and μ are dually coupled, i.e.,

$$\mu = \mathbb{E}_\theta[\phi(X)]$$



- Proof Idea: (Jensen's Inequality)

- Optimizing the lower bound gives μ
  - This is an inference!

Mean Field Methods In General

- However, the lower bound can't explicitly evaluated in general
  - Because the dual function $A^{\star}$ typically lacks an explicit form

- Mean Field Methods
  - Approximate the lower bound

$$A_F^{\star} = A^{\star}|_{\mathcal{M}_F(G)}$$

  - Approximate the realizable mean parameter set

$$\mathcal{M}_F(G) \subseteq \mathcal{M}$$

  - The MF optimization problem

$$\max_{\mu \in \mathcal{M}_F(G)} \left\{ \theta^{\top}\mu - A_F^{\star}(\mu) \right\}$$

  - Still a lower bound?

KL-divergence

- Kullback-Leibler Divergence

$$KL(q\|p) := \mathbb{E}_q[\log \frac{q}{p}]$$

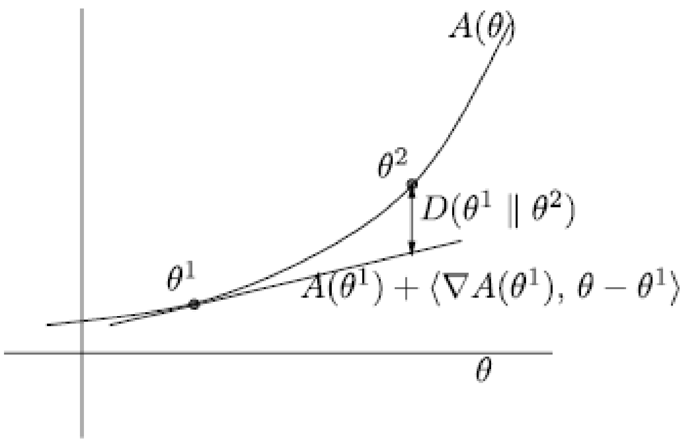
- For two exponential family distributions with the same STs:

$$KL(\theta_1\|\theta_2) = \mathbb{E}_{\theta_1}\left[\log \frac{p_{\theta_1}(x)}{p_{\theta_2}(x)}\right]$$

$$= A(\theta_2) - A(\theta_1) - \mu_1^\top(\theta_2 - \theta_1)$$ **Primal Form**

$$= A(\theta_2) + A^\star(\mu_1) - \mu_1^\top \theta_2$$ **Mixed Form**

$$= A^\star(\mu_1) - A^\star(\mu_2) - \mu_2^\top(\mu_1 - \mu_2)$$ **Dual Form**



Mean Field and KL-divergence

- Optimizing a lower bound

$$\max_{\mu \in \mathcal{M}_F(G)} \left\{ \theta^\top \mu - A_F^\star(\mu) \right\}$$

- Equivalent to minimize a KL-divergence

$$A(\theta) - (\theta^\top \mu - A_F^\star(\mu)) = KL(\mu \| \theta)$$

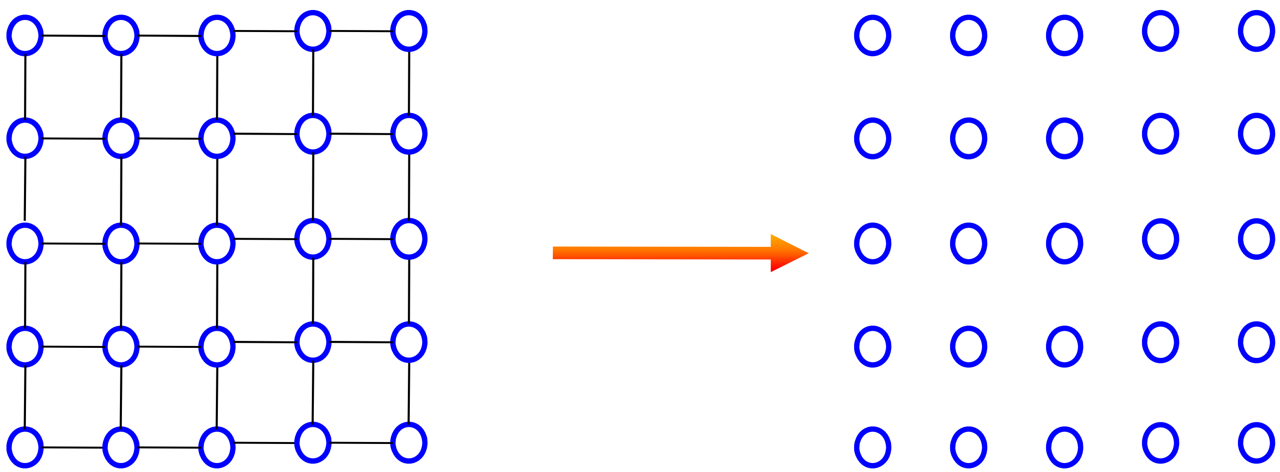
- Therefore, we are doing minimization

$$\min_{\mu \in \mathcal{M}_F(G)} KL(\mu \| \theta)$$

Naïve Mean Field

- Fully factorized variational distribution

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

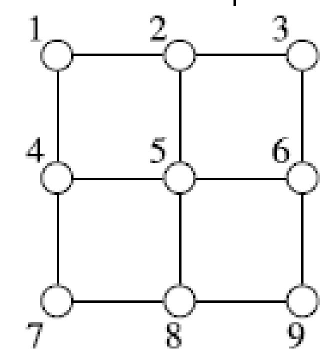


Naïve Mean Field for Ising Model

- Sufficient statistics and Mean Parameters

$$(x_s, s \in V), \text{ and } (x_s x_t, (s,t) \in E)$$

$$\mu_s = p(X_s = 1), \text{ and } \mu_{st} = p(X_s = 1, X_t = 1)$$



- Naïve Mean Field
  - Realizable mean parameter subset

$$\mathcal{M}_{F_0} = \left\{ \mu | 0 \le \mu_s \le 1 \ \forall s \in V, \text{ and } \mu_{st} = \mu_s \mu_t \ \forall (s,t) \in E \right\}$$

  - Entropy $-A^{\star}_{F_0}(\mu) = -\sum_{s \in V} [\mu_s \log \mu_s + (1-\mu_s) \log(1-\mu_s) = \sum_{s \in V} H_s(\mu_s)]$
  - Optimization Problem

$$\max_{\mu \in [0,1]^m} \left\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t + \sum_{s \in V} H_s(\mu_s) \right\}$$

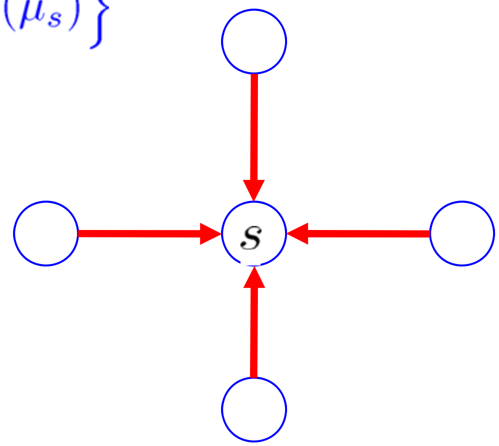
Naïve Mean Field for Ising Model

- Optimization Problem

$$\max_{\mu \in [0,1]^m} \left\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t + \sum_{s \in V} H_s(\mu_s) \right\}$$

- Update Rule

$$\mu_s \leftarrow \sigma\left(\theta_s + \sum_{t \in N(s)} \theta_{st} \mu_t\right)$$



- $\mu_t = p(X_t = 1) = \mathbb{E}_p[X_t]$ resembles "message" sent from node t to s
- $\{\mathbb{E}_p[X_t], t \in N(s)\}$ forms the "mean field" applied to s from its neighborhood

Non-Convexity of Mean Field

- Mean field optimization is always non-convex for any exponential family in which the state space $\mathcal{X}^m$ is finite

  - Finite convex hull

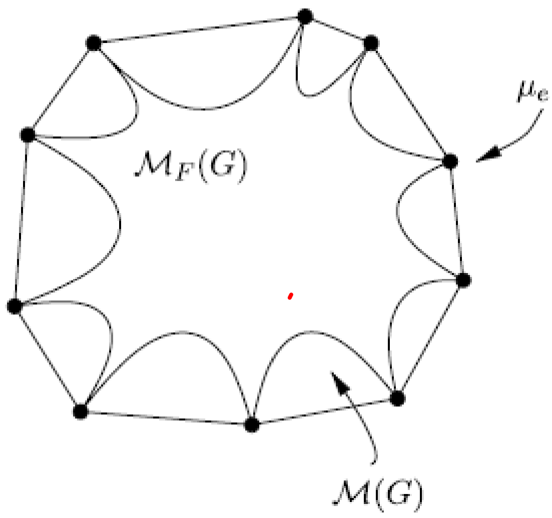
$$\mathcal{M}(G) = \mathrm{conv}\{\phi(e),\ e \in \mathcal{X}^m\}$$

  - $\mathcal{M}_F(G)$ contains all the extreme points

  - If $\mathcal{M}_F(G)$ is a convex set, then

$$\mathcal{M}_F(G) = \mathcal{M}(G)$$

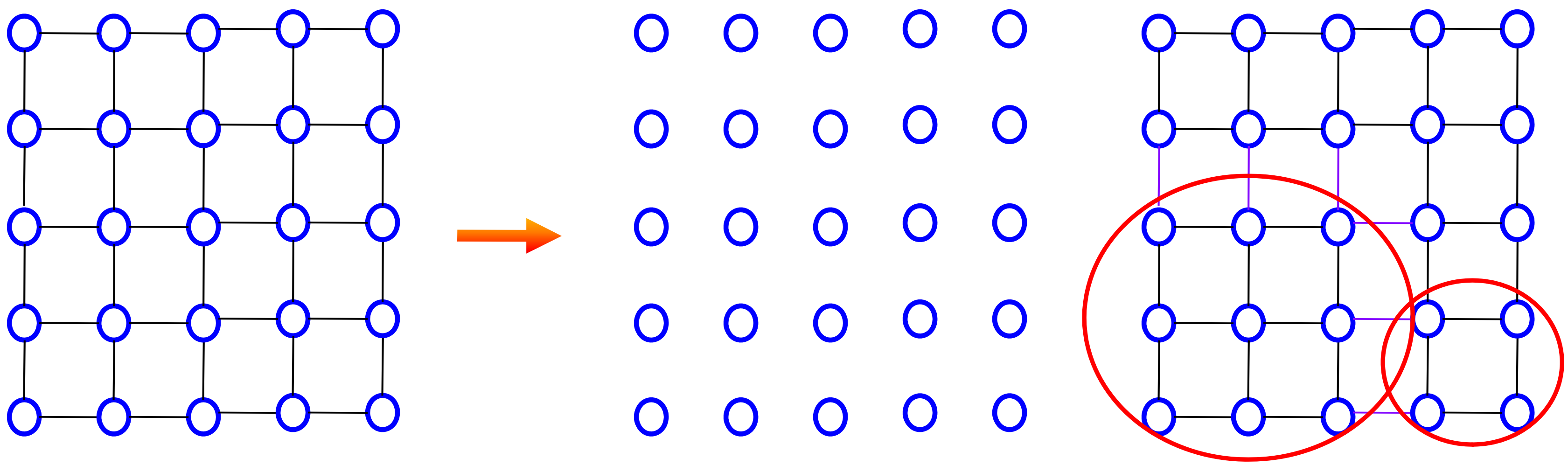
- Mean field has been used successfully



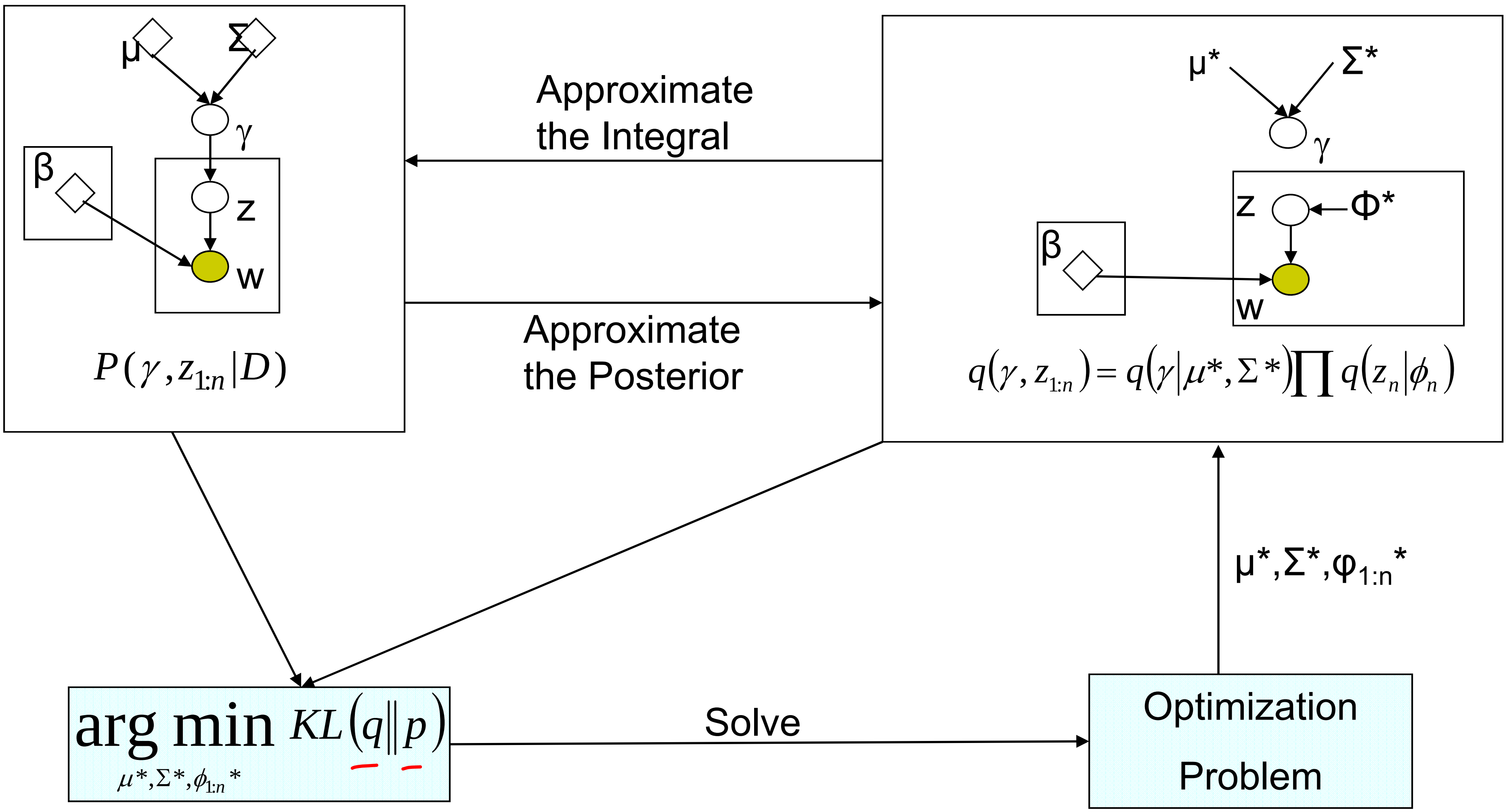
Structured Mean Field

- Mean field theory is general to any tractable sub-graphs
- Naïve mean field is based on the fully unconnected sub-graph

- Variants based on structured sub-graphs can be derived



Topic models



Variational Inference With no Tears

[Ahmed and Xing, 2006, Xing et al 2003]

- Fully Factored Distribution

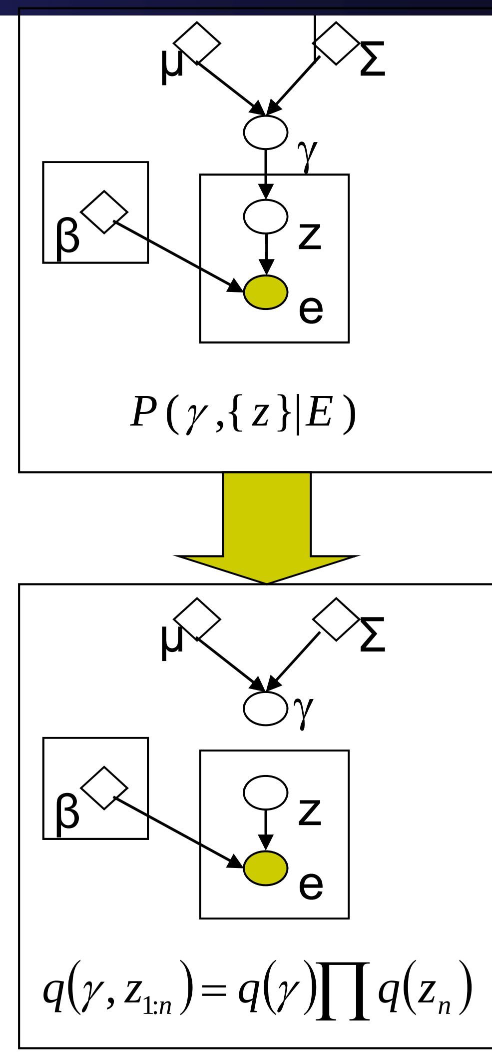
$$q(\gamma, z_{1:n}) = q(\gamma)\prod q(z_n)$$

- Fixed Point Equations

$$q_\gamma{}^*(\gamma) = P\left(\gamma \middle| \langle S_z \rangle_{q_z}, \mu, \Sigma\right) \approx N(\mu_\gamma, \Sigma_\gamma)$$

$$q_z{}^*(z) = P\left(z \middle| \langle S_\gamma \rangle_{q\gamma}, \beta_{1:k}\right) \approx \mathrm{Multi}(\theta_z)$$

Laplace approximation



Summary of GMF

- Message-passing algorithms (e.g., belief propagation, mean field) are solving approximate versions of exact variational principle in exponential families

- There are two *distinct* components to approximations:
  - Can use either **inner** or **outer** bounds to $\mathcal{M}$
  - Various approximation to the entropy function $-A^{\star}$

- BP: polyhedral outer bound and non-convex Bethe approximation
- MF: non-convex inner bound and exact form of entropy
- Kikuchi: tighter polyhedral outer bound and better entropy approximation