Learning and Inference in Probabilistic Graphical Models

Expectation Propagation
April 28, 2010

Introduction

$$p(x) \propto \prod_i \psi_i(x)$$

Goal: Efficiently approximate intractable distributions

Features of *Expectation Propagation* (EP):

- Deterministic, iterative method for computing approximate posterior distributions
- Approximating distribution may be selected from any exponential family
- Framework for extending loopy Belief Propagation (BP):

  - *Structured approximations for greater accuracy*
  - *Inference for continuous non-Gaussian models*

Outline

Background

- Graphical models
- Exponential families

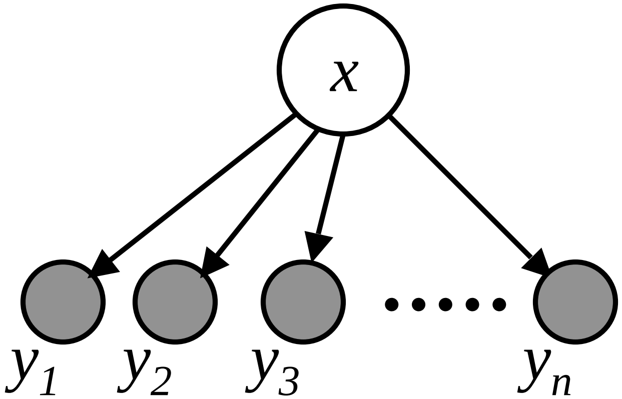
Expectation Propagation (EP)

- Assumed Density Filtering
- EP for unstructured exponential families

Connections to Belief Propagation

- BP as a fully factorized EP approximation
- Free energy interpretations
- Continuous non-Gaussian models
- Structured EP approximations

Clutter Problem



$$p_0(x) = \mathcal{N}(x; 0, 100I)$$

$$p_i(y_i|x) = (1 - w)\mathcal{N}(y_i; x, I) + w\mathcal{N}(0, 10I)$$

n independent observations from a Gaussian distribution of unknown mean x embedded in a sea of clutter

$$p(x|y_1, \ldots, y_n) \propto p_0(x) \prod_{i=1}^{n} p_i(y_i|x)$$

⟹ posterior is a mixture of 2^n Gaussians

Exponential Families

$$q(x;\theta) = \exp\left\{\sum_{\alpha} \theta_\alpha \phi_\alpha(x) - \Phi(\theta)\right\}$$

θ ⟶ *exponential (canonical) parameter vector*

$\phi_\alpha(x)$ ⟶ *potential function*

$\Phi(\theta)$ ⟶ *log partition function (normalization)*

Examples:

- Gaussian
- Poisson
- Discrete multinomial
- Factorized versions of these models

Manipulation of Exponential Families

$$q(x;\theta) = \exp\left\{\sum_\alpha \theta_\alpha \phi_\alpha(x) - \Phi(\theta)\right\}$$

Products: $q(x;\theta_1)q(x;\theta_2) \propto q(x;\theta_1+\theta_2)$

Quotients: $\dfrac{q(x;\theta_1)}{q(x;\theta_2)} \propto q(x;\theta_1-\theta_2)$

May not preserve normalizability

Projections: $\theta^* = \arg\min_\theta D\left(p(x)\;||\;q(x;\theta)\right)$

Optimal solution found via moment matching:

$$\int q(x;\theta^*)\phi_\alpha(x)\,dx = \int p(x)\phi_\alpha(x)\,dx$$

Assumed Density Filtering (ADF)

$$p(x) \propto \prod_i \psi_i(x)$$

- Choose an approximating exponential family $q(x;\theta)$
- Initialize by approximating the first compatibility function:

$$\theta^1 = \arg\min_\theta D\left(\psi_1(x) \,||\, q(x;\theta)\right)$$

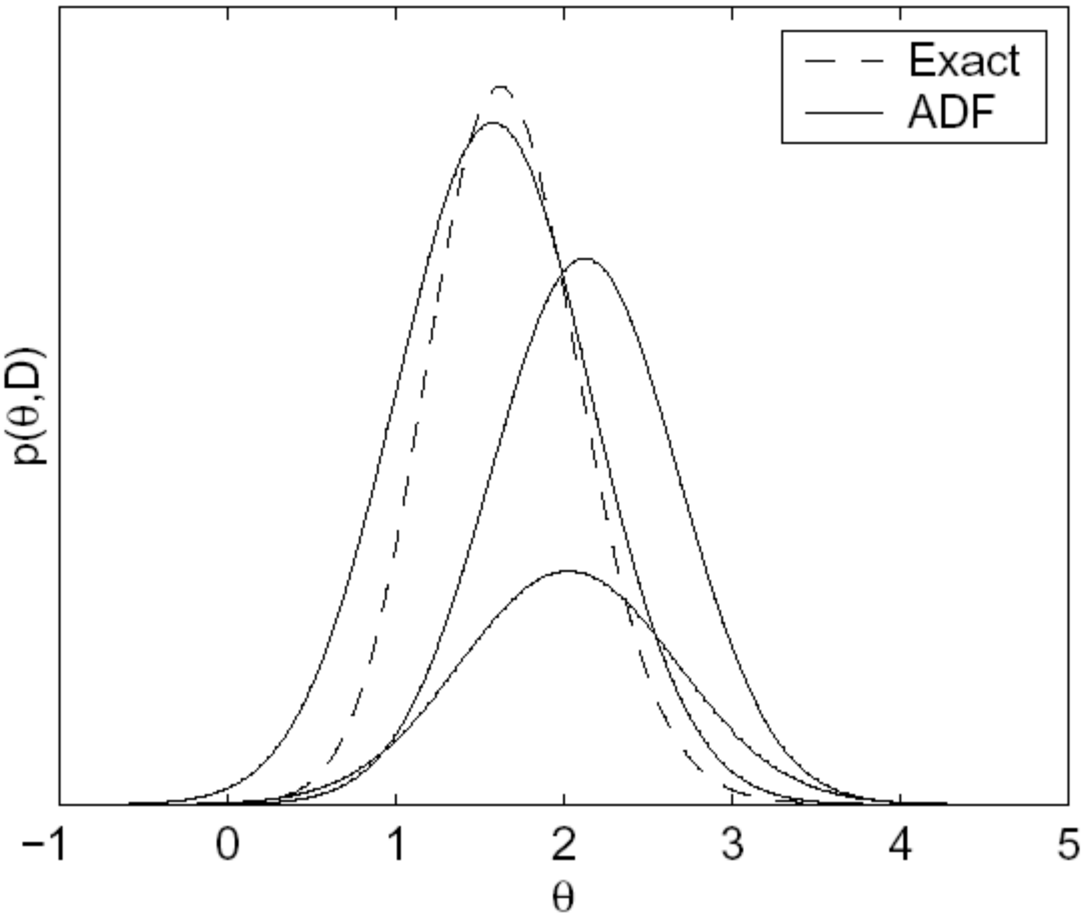
- Sequentially incorporate all other compatibilities:

$$\theta^i = \arg\min_\theta D\left(\psi_i(x) q(x;\theta^{i-1}) \,||\, q(x;\theta)\right)$$

The current best estimate $q(x;\theta^{i-1})$ *of the product distribution is used to guide the incorporation of* $\psi_i(x)$

$\Longrightarrow$ *Superior to approximating* $\psi_i(x)$ *individually*

ADF for the Clutter Problem



ADF is sensitive to the order in which compatibility functions are incorporated into the posterior

ADF as Compatibility Approximation

$$p(x) \propto \prod_i \psi_i(x)$$

$$\theta^i = \arg\min_\theta D\left(\psi_i(x) q(x; \theta^{i-1}) \mid\mid q(x; \theta)\right)$$

Standard View: Sequential approximation of the posterior

Alternate View: Sequential approximation of compatibilities

$$q(x; \theta^i) \propto m_i(x) q(x; \theta^{i-1}) \qquad m_i(x) \propto \frac{q(x; \theta^i)}{q(x; \theta^{i-1})}$$

$m_i(x) \longrightarrow$ exponential approximation to $\psi_i(x)$

member of exponential family $q(x; \theta)$

Expectation Propagation

Idea: Iterate the ADF compatibility function approximations, always using the best estimates for all but one function to improve the exponential approximation to the remaining term

Initialization:

- Choose starting values for the compatibility approximations:

$$m_i(x) = 1$$

- Initialize the corresponding posterior approximation:

$$q(x;\theta) \propto \prod_i m_i(x)$$

EP Iteration

1. Choose some $m_i(x)$ to refine.
2. Remove the effects of $m_i(x)$ from the current estimate:

$$q(x;\theta^{\setminus i}) \propto \frac{q(x;\theta)}{m_i(x)}$$

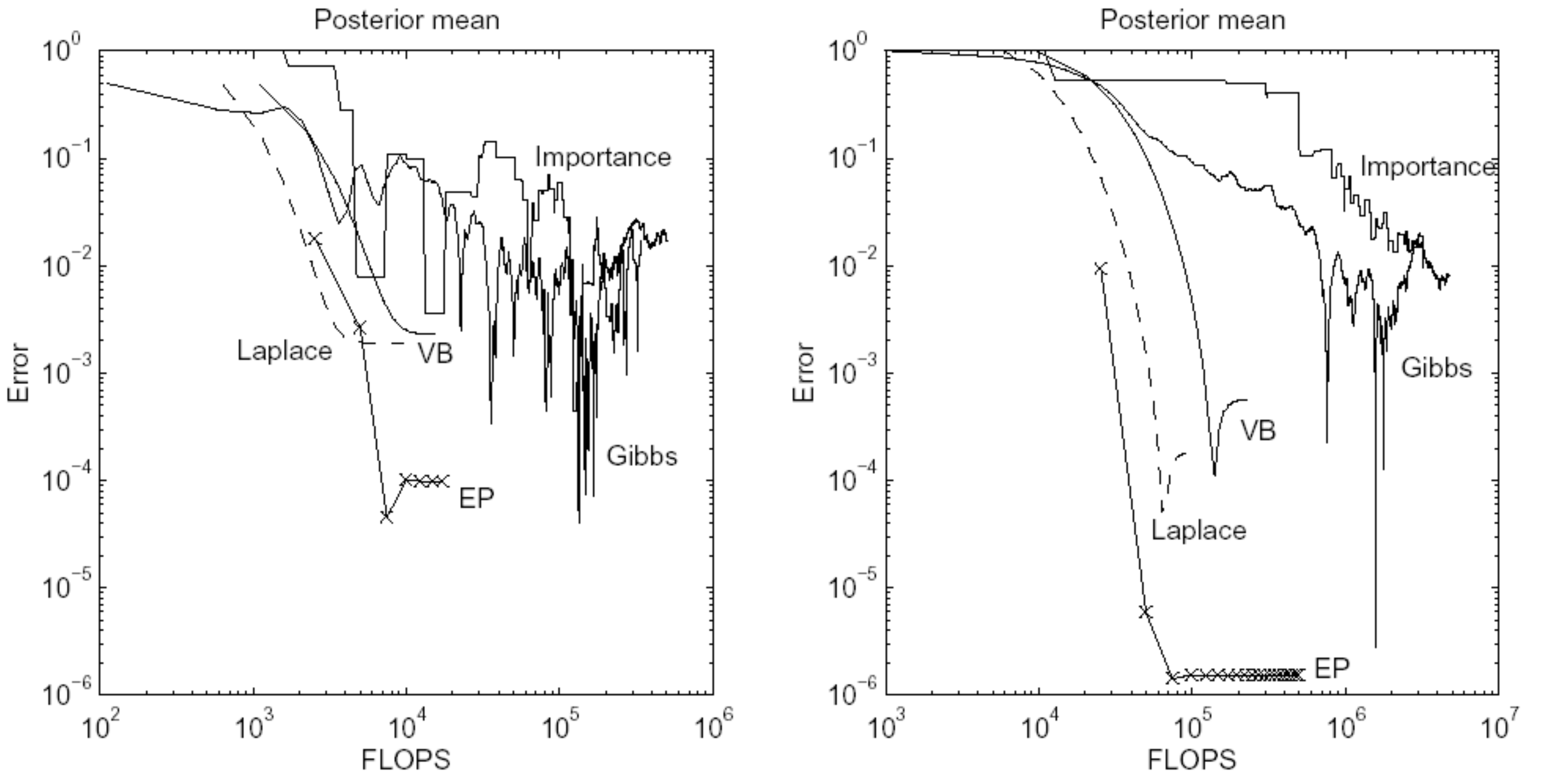
3. Update the posterior approximation to $q(x;\theta^*)$, where

$$\theta^* = \arg\min_{\theta} D\left(q(x;\theta^{\setminus i})\psi_i(x) \;||\; q(x;\theta)\right)$$

4. Refine the exponential approximation to $m_i(x)$ as

$$m_i(x) \propto \frac{q(x;\theta^*)}{q(x;\theta^{\setminus i})}$$

EP for the Clutter Problem



n = 20 n = 200

EP generally shows quite good performance, but is not guaranteed to converge

Relationship to Belief Propagation

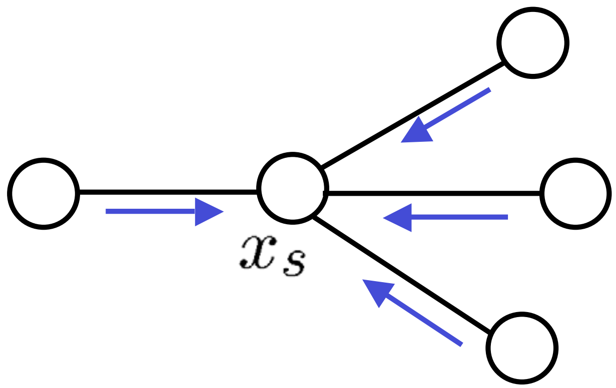
- BP is a special case of EP
- Many results characterizing BP can be extended to EP
- EP provides a mechanism for constructing improved approximations for models where BP performs poorly
- EP extends local propagation methods to many models where BP is not possible (continuous non-Gaussian)

Explore relationship for special case of pairwise MRFs:

$$p(x) = \frac{1}{Z} \prod_{(s,t) \in \mathcal{E}} \psi_{s,t}(x_s, x_t)$$

Belief Propagation

- Combine the information from all nodes in the graph through a series of local *message-passing* operations



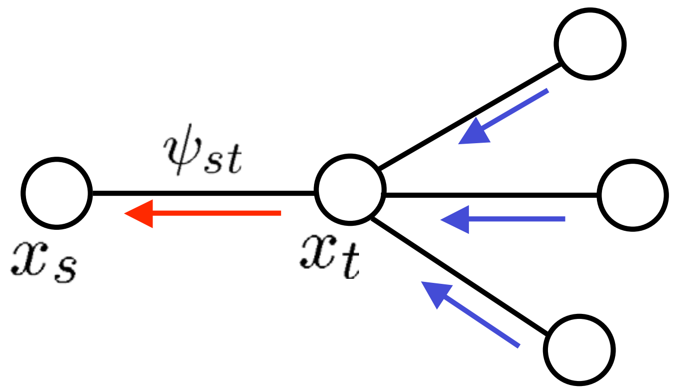
$$\widehat{p}(x_s) = \alpha \prod_{t \in \Gamma(s)} m_{ts}(x_s)$$

$\Gamma(s)$ → *neighborhood* of node *s* (adjacent nodes)

$m_{ts}(x_s)$ → *message* sent from node *t* to node *s*

("sufficient statistic" of *t*'s knowledge about *s*)

BP Message Updates



$$m_{ts}(x_s) = \alpha \int_{x_t} \psi_{s,t}(x_s, x_t) \prod_{u \in \Gamma(t) \setminus s} m_{ut}(x_t)\, dx_t$$

1. Combine incoming messages, *excluding* that from node *s*, with the local observation to form a distribution over x_t
2. Propagate this distribution from node *t* to node *s* using the pairwise interaction potential $\psi_{st}(x_s, x_t)$
3. Integrate out the effects of x_t

Fully Factorized EP Approximations

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

Each $q_s(x_s)$ can be a general discrete multinomial distribution (no restrictions other than factorization)

$$m_{s,t}(x_s, x_t) = m_{t\to s}(x_s) m_{s\to t}(x_t)$$

⟶ *Compatibility approximations in same exponential family*

Initialization:

- Initialize compatibility approximations $m_{s,t}(x_s,x_t)$
- Initialize each term in the factorized posterior approximation:

$$q_s(x_s) \propto \prod_{t\in\Gamma(s)} m_{t\to s}(x_s)$$

Factorized EP Iteration I

1. Choose some $m_{s,t}(x_s,x_t)$ to refine.

   → $m_{s,t}(x_s,x_t)$ involves only x_s and x_t, so the approximations $q_u(x_u)$ for all other nodes are unaffected by the EP update

2. Remove the effects of $m_{s,t}(x_s,x_t)$ from the current estimate:

$$q_{s\backslash t}(x_s) \propto \frac{q_s(x_s)}{m_{t\to s}(x_s)} = \prod_{u\in\Gamma(s)\backslash t} m_{u\to s}(x_s)$$

$$q_{t\backslash s}(x_t) \propto \frac{q_t(x_t)}{m_{s\to t}(x_t)} = \prod_{v\in\Gamma(t)\backslash s} m_{v\to t}(x_t)$$

Factorized EP Iteration II

3. Update the posterior approximation by determining the appropriate marginal distributions:

$$q_s(x_s) = \sum_{x_t} \psi_{s,t}(x_s, x_t) q_{s\setminus t}(x_s) q_{t\setminus s}(x_t)$$
$$q_t(x_t) = \sum_{x_s} \psi_{s,t}(x_s, x_t) q_{s\setminus t}(x_s) q_{t\setminus s}(x_t)$$

4. Refine the exponential approximation to $m_{s,t}(x_s,x_t)$ as

$$m_{t\to s}(x_s) \propto \frac{q_s(x_s)}{q_{s\setminus t}(x_s)} = \sum_{x_t} \psi_{s,t}(x_s, x_t) \prod_{v\in\Gamma(t)\setminus s} m_{v\to t}(x_t)$$
$$m_{s\to t}(x_t) \propto \frac{q_t(x_t)}{q_{t\setminus s}(x_t)} = \sum_{x_s} \psi_{s,t}(x_s, x_t) \prod_{u\in\Gamma(s)\setminus t} m_{u\to s}(x_s)$$

⟹ ***Standard BP Message Updates***

Bethe Free Energy

$$p(x) = \frac{1}{Z} \prod_{(s,t)\in\mathcal{E}} \psi_{s,t}(x_s, x_t) \prod_{s\in\mathcal{V}} \psi_s(x_s)$$

$$G(q,p) = \sum_{(s,t)\in\mathcal{E}} \int q_{s,t}(x_s, x_t) \log \frac{q_{s,t}(x_s, x_t)}{q_s(x_s) q_t(x_t) \psi_{s,t}(x_s, x_t)} \, dx_{s,t} + \sum_{s\in\mathcal{V}} \int q_s(x_s) \log \frac{q_s(x_s)}{\psi_s(x_s)} \, dx_s$$

BP: Minimize subject to marginalization constraints

$$\int q_{s,t}(x_s, x_t) \, dx_s = q_t(x_t)$$

EP: Minimize subject to expectation constraints

$$\int q_{s,t}(x_s, x_t) \phi_\alpha(x_t) \, dx_{s,t} = \int q_t(x_t) \phi_\alpha(x_t) \, dx_t$$

Implications of Free Energy Interpretation

Fixed Points

- EP has a fixed point for every product distribution *p(x)*
- Stable EP fixed points must be local *minima* of the Bethe free energy (converse does *not* hold)

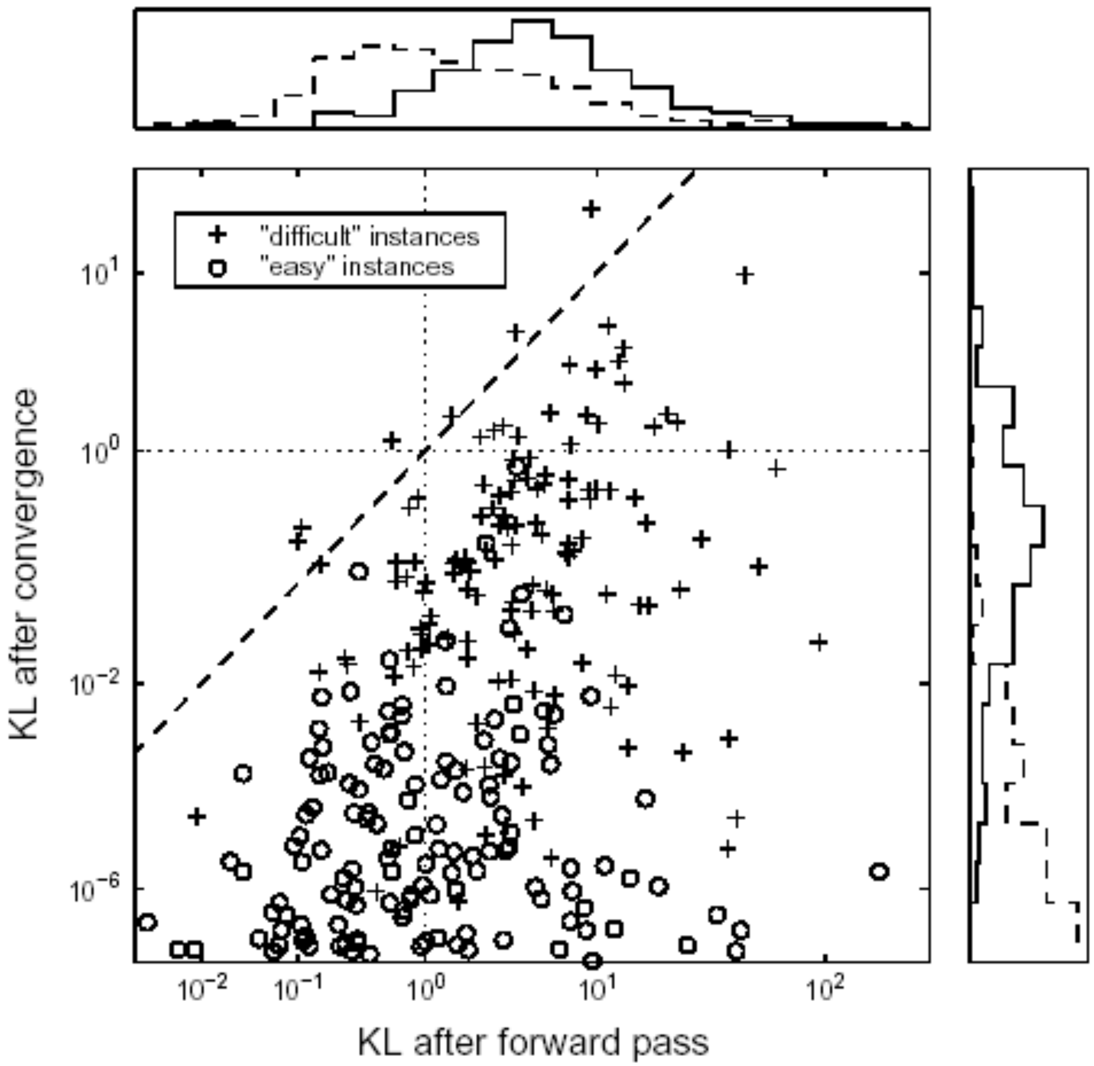
Double Loop Algorithms

- Guaranteed convergence to local minimum of Bethe
- Separate Bethe into sum of convex and concave parts:

  Outer Loop: Bound concave part linearly

  Inner Loop: Solve constrained convex minimization

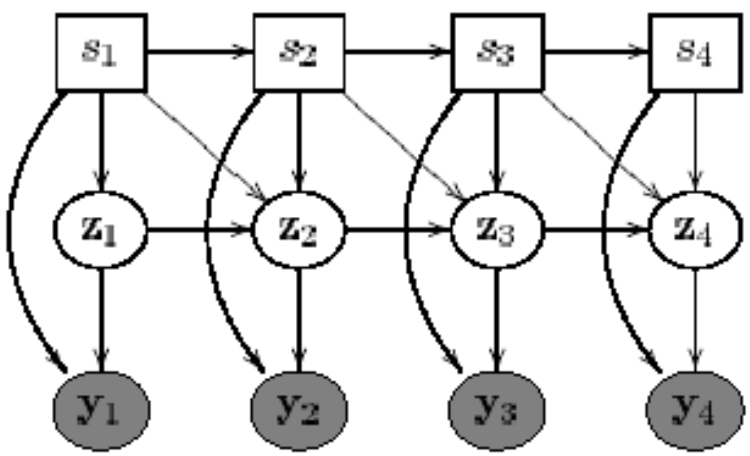
Are Double Loop Algorithms Worthwhile?



Non-Gaussian Message Passing

- Choose an approximating exponential family
- Modify the BP marginalization step to perform moment matching: construct best local exponential approximation

Switching Linear Dynamical Systems



s_t → discrete "system mode"

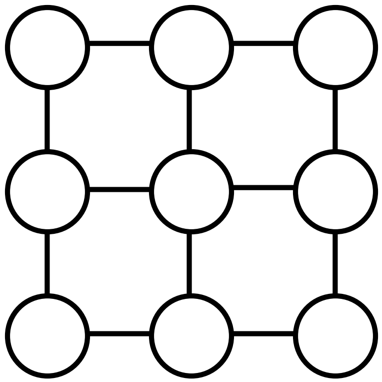
z_t → conditionally Gaussian

y_t → observation

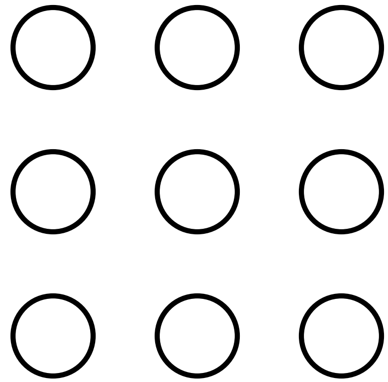
Exact Posterior: Mixture of exponentially many Gaussians

EP Approximation: Single Gaussian for each discrete state

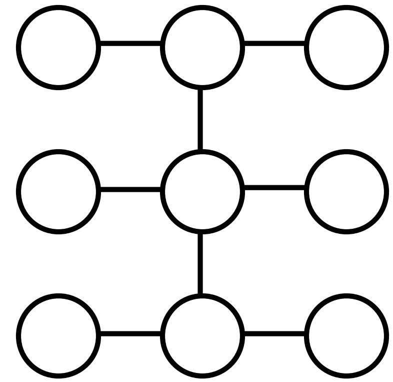
Structured EP Approximations



Original



Fully Factorized EP
(Belief Propagation)



Structured EP

- Structured EP approximations employ triangulated graphs to allow closed-form exponential family projections
- Can unify structured EP-style approximations and region based Kikuchi-style approximations in common framework
  - Every discrete EP entropy approximation has a corresponding region graph entropy and GBP algorithm *(Welling, Minka, Teh, UAI05)*
  - EP for continuous variables goes beyond GBP