

Recall, using the info/canonical form of a Gaussian  $\rightarrow$  P.D.

$$x \sim N(\mu, \Sigma)$$

$$\Sigma = \Lambda^{-1} \quad \mu = \Lambda^{-1} h$$

$$P(x) \propto \exp(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu))$$

$$x \sim N^{-1}(h, \Lambda)$$

$$\Lambda = \Sigma^{-1} \quad h = \Lambda \mu$$

$$P(x) \propto \exp(-\frac{1}{2} x^T \Lambda x + h^T x)$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

$$x_1 \sim N(\mu_1, \Sigma_{11})$$

$$x_1 \perp\!\!\!\perp x_2 \Leftrightarrow \Sigma_{12} = 0$$

(uncorr  $\Rightarrow$   $\perp\!\!\!\perp$ )

$\perp\!\!\!\perp$   
Indep

Marginalizing

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N^{-1}\left(\begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}\right)$$

$$x_1 \sim N^{-1}(h_1 - \Lambda_{12} \Lambda_{22}^{-1} \Lambda_{21}, \Lambda_{11})$$

Conditioning

$$P(x_1 | x_2) \sim N(\mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2), \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21})$$

$$P(x_1 | x_2) \sim N^{-1}(h^1 - \Lambda_{12} x_2, \Lambda_{11})$$

$\perp\!\!\!\perp$  Conditional  $\perp\!\!\!\perp$

$$x_i - x_{\text{rest}} - x_j \Leftrightarrow \Lambda_{ij} = 0$$

So, we have a gaussian GM:

$$P(x) = \prod_{i \in V} \exp(-\frac{1}{2} x_i^T J_{ii} x_i + h_i^T x_i) \prod_{(i,j) \in E} \exp(-x_i^T J_{ij} x_j)$$

$$G = (V, E) : (i, j) \in E \iff J_{ij} \neq 0$$

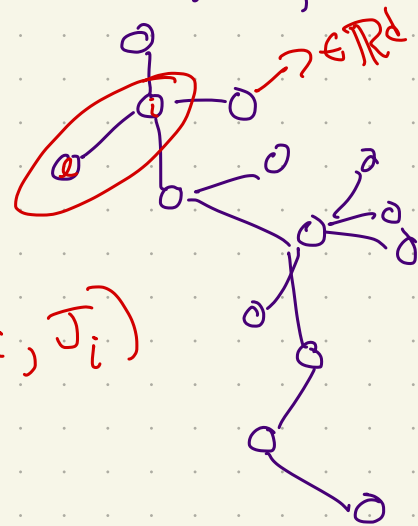
Generally, BP will be used to condition on info from other parts of the graph. Eg.  $P(x_1 | x_2 = z)$ .

Let's start with a tree  $G$ . This'll look like an elimination alg

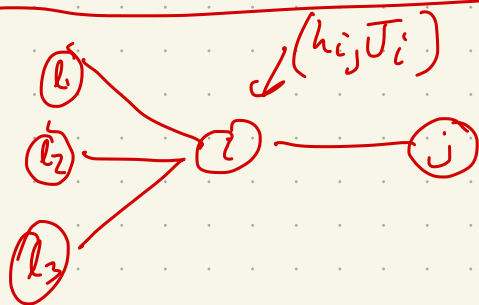
$$N^{-1}(h_{l \rightarrow i}, J_{l \rightarrow i}) \sim m_{li}(x_i) \quad (h_l, J_l) \quad J_{li} \quad (h_i, J_i)$$

Given  $\begin{pmatrix} x_l \\ x_i \end{pmatrix} \sim N^{-1} \left( \begin{bmatrix} h_l \\ 0 \end{bmatrix}, \begin{pmatrix} J_l & J_{li} \\ J_{li}^T & 0 \end{pmatrix} \right)$

$$x_i \sim N^{-1}(h_{l \rightarrow i}, J_{l \rightarrow i}) = N^{-1} \left( -J_{li} J_l^{-1} h_l, -J_{li} J_l^{-1} J_{li} \right)$$



$$M_{i \rightarrow j}(x_j) = N^{-1}(h_{i \rightarrow j}, J_{i \rightarrow j})$$



$$m_{i \rightarrow j}(x_j) \sim N^{-1}(h_{i \rightarrow j}, J_{i \rightarrow j})$$

Parallel Update

$$h_{i \rightarrow j} \sim N^{-1} \left( -J_{ji} \left( J_i + \sum_{l \in \partial i \setminus \{j\}} J_{l \rightarrow i} \right)^{-1} (h_i + \sum_{l \in \partial i \setminus \{j\}} h_{l \rightarrow i}) \right),$$

$$J_{i \rightarrow j} = J_{ji} \left( J_i + \sum_{l \in \partial i \setminus \{j\}} J_{l \rightarrow i} \right)^{-1} J_{ij}$$

Decision/Marginal

$$\hat{h}_i = h_i + \sum_{l \in \partial i \setminus \{j\}} \hat{h}_{l \rightarrow i}$$

$$\hat{J}_i = J_{ii} + \sum_{l \in \partial i} J_{li}$$

$$\hat{x}_i \sim N^{-1}(\hat{h}_i, \hat{J}_{ii}) = N(\hat{J}_{ii}^{-1} \hat{h}_i, \hat{J}_{ii}^{-1})$$

$\downarrow O(d^3)$   $\rightarrow O(d^2)$

$T$  steps of BP takes  $O(d^3 |E| \cdot T)$   $\nearrow n$  for a tree  
 inverting  $T \in \mathbb{R}^{2n \times 2n}$  takes  $O(d^3 n^3)$   $\nearrow \log n$

Alt version of GBP

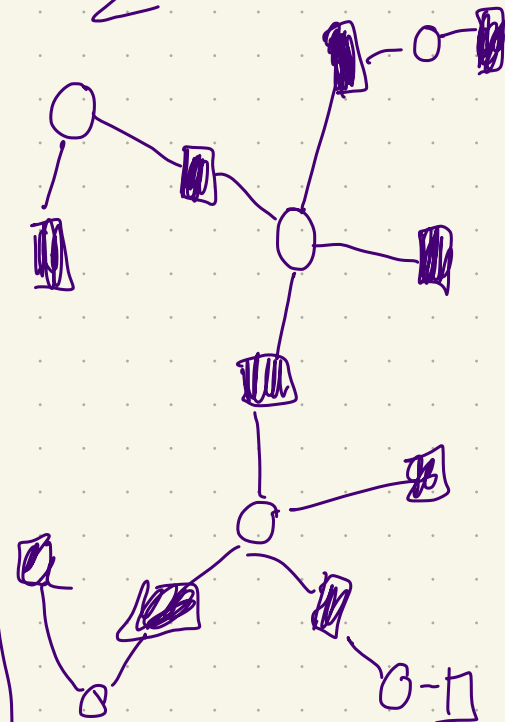
$$h_{i \rightarrow j} = h_i - \sum_{k \in \partial i \setminus j} T_{ik} T_{k \rightarrow i}^{-1} h_{k \rightarrow i}$$

$$T_{i \rightarrow j} = T_{ii} - \sum_{k \in \partial i \setminus j} T_{ik} T_{k \rightarrow i}^{-1} T_{ki}$$

$$\hat{h}_i = h_i - \sum_{k \in \partial i} T_{ik} T_{k \rightarrow i}^{-1} h_{k \rightarrow i}$$

$$\hat{T}_i = T_i - \sum_{k \in \partial i} T_{ik} T_{k \rightarrow i}^{-1} T_{ki}$$

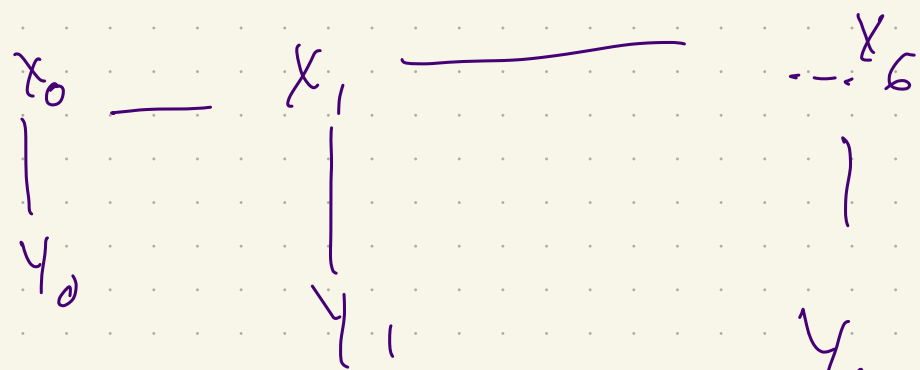
$\Sigma \Rightarrow$



\* Maximization is equivalent  $(\vec{\mu}_1, \dots, \vec{\mu}_n) = \vec{\mu}$

Q:  $J = \begin{bmatrix} \end{bmatrix}$  How can we check if  $J > 0$ ?  
 $- O(n^3)$  i

6 Hidden Markov Models (2)  $\exists$  suff conditions  $\Rightarrow J > 0$ ?



$\hookrightarrow$  Linear Dynamical Systems (Kalman Filtering)

$$\begin{aligned} & x_0 \sim N(0, \Sigma_0) \\ & x_{t+1} = Ax_t + Bv_t \end{aligned} \quad \left. \begin{array}{l} \text{observe } y_t \\ y_t \in \mathbb{R}^J \\ \text{noise } w_t \sim N(0, w) \end{array} \right\} \quad y_t = (x_t + w_t)$$

$x_t$  : state  $\in \mathbb{R}^d$   
 $A \in \mathbb{R}^{d \times d}$  state trans. matrix

Process noise  $\begin{cases} v_t \in \mathbb{R}^p \sim N(0, V) \\ B \in \mathbb{R}^{d \times p} \end{cases}$

GGM

$$x_0 \sim N(0, \Sigma_0)$$

$$x_{t+1} | x_t \sim N(Ax_t, H = BVB^T)$$

$$y_t | x_t \sim N(x_t, W)$$

Factorization:

$$P_Y(x) = \frac{1}{Z} \exp\left(-\frac{1}{2} x_0^T \overset{\downarrow P(x_0)}{\Sigma_0} x_0\right) \exp\left[-\frac{1}{2} (x_1 - Ax_0)^T \overset{\downarrow P(x_2|x_0)}{H^{-1}} (x_1 - Ax_0)\right]$$

$$\exp\left(-\frac{1}{2} y_0 - (x_0)^T W^{-1} (y_0 - (x_0))\right)$$

$$\nearrow P_{y_0|x_0}$$

Information form:

$$= \frac{1}{Z} \prod_{i=0}^n \exp\left(-\frac{1}{2} x_i^T J_i x_i + h_i^T\right)$$



# Gaussian graphical models

- belief propagation naturally extends to continuous distributions by replacing summations to integrals

$$\nu_{i \rightarrow j}(x_i) = \prod_{k \in \partial i \setminus j} \int \psi_{ik}(x_i, x_k) \nu_{k \rightarrow i}(x_k) dx_k$$

- integration can be intractable for general functions
- however, for Gaussian graphical models for jointly Gaussian random variables, we can avoid explicit integration by exploiting algebraic structure, which yields efficient inference algorithms



# Multivariate jointly Gaussian random variables

four definitions of a **Gaussian random vector**  $x \in \mathbb{R}^n$ :  $x$  is **Gaussian** iff

1.  $x = Au + b$  for standard i.i.d. Gaussian random vector  $u \sim \mathcal{N}(0, \mathbf{I})$
2.  $y = a^T x$  is Gaussian for all  $a \in \mathbb{R}^n$
3. **covariance form**: the probability density function is

$$\mu(x) = \frac{1}{(2\pi)^{n/2} |\Lambda|^{1/2}} \exp \left\{ -\frac{1}{2} (x - m)^T \Lambda^{-1} (x - m) \right\}$$

denoted as  $x \sim \mathcal{N}(m, \Lambda)$  with mean  $m = \mathbb{E}[x]$  and covariance matrix  $\Lambda = \mathbb{E}[(x - m)(x - m)^T]$  (for some positive definite  $\Lambda$ ).

4. **information form**: the probability density function is

$$\mu(x) \propto \exp \left\{ -\frac{1}{2} x^T J x + h^T x \right\}$$

denoted as  $x \sim \mathcal{N}^{-1}(h, J)$  with *potential vector*  $h$  and *information (or precision) matrix*  $J$  (for some positive definite  $J$ )

- note that  $J = \Lambda^{-1}$  and  $h = \Lambda^{-1}m = Jm$
- $x$  can be non-Gaussian and the marginals still Gaussian

- consider two operations on the following Gaussian random vector

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \right) = \mathcal{N}^{-1} \left( \begin{bmatrix} h_1 \\ h_2 \end{bmatrix}, \begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix} \right)$$

- marginalization** is easy to compute when  $x$  is in covariance form

$$x_1 \sim \mathcal{N}(m_1, \Lambda_{11})$$

for  $x_1 \in \mathbb{R}^{d_1}$ , one only needs to read the corresponding entries of dimensions  $d_1$  and  $d_1^2$  but complicated when  $x$  is in information form

$$x_1 \sim \mathcal{N}^{-1}(h', J')$$

where  $J' = \Lambda_{11}^{-1} = \left( \begin{bmatrix} \mathbb{I} & 0 \end{bmatrix} J^{-1} \begin{bmatrix} \mathbb{I} \\ 0 \end{bmatrix} \right)^{-1}$  and

$$h' = J' m_1 = \left( \begin{bmatrix} \mathbb{I} & 0 \end{bmatrix} J^{-1} \begin{bmatrix} \mathbb{I} \\ 0 \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbb{I} & 0 \end{bmatrix} J^{-1} h$$

- we will prove that  $h' = h_1 - J_{12} J_{22}^{-1} h_2$  and  $J' = J_{11} - J_{12} J_{22}^{-1} J_{21}$
- what is wrong in computing the marginal with the above formula?  
for  $x_1 \in \mathbb{R}^{d_1}$  and  $x_2 \in \mathbb{R}^{d_2}$  and  $d_1 \ll d_2$ , inverting  $J_{22}$  requires runtime  $O(d_2^{2.8074})$  (Strassen algorithm)

• Proof of  $J' = \Lambda_{11}^{-1} = J_{11} - J_{12}J_{22}^{-1}J_{21}$

- ▶  $J'$  is called **Schur complement** of the block  $J_{22}$  of the matrix  $J$
- ▶ useful matrix identity

$$\begin{bmatrix} \mathbf{I} & -BD^{-1} \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} \mathbf{I} & 0 \\ -D^{-1}C & \mathbf{I} \end{bmatrix} = \begin{bmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{bmatrix}$$

$$\begin{aligned} \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} &= \begin{bmatrix} \mathbf{I} & 0 \\ -D^{-1}C & \mathbf{I} \end{bmatrix} \begin{bmatrix} (A - BD^{-1}C)^{-1} & 0 \\ 0 & D^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -BD^{-1} \\ 0 & \mathbf{I} \end{bmatrix} \\ &= \begin{bmatrix} (A - BD^{-1}C)^{-1} & -S^{-1}BD^{-1} \\ -D^{-1}CS^{-1} & D^{-1} + D^{-1}CS^{-1}BD^{-1} \end{bmatrix} \end{aligned}$$

where  $S = A - BD^{-1}C$

- ▶ since  $\Lambda = J^{-1}$ ,

$$\Lambda = \begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix}^{-1} = \begin{bmatrix} (J_{11} - J_{12}J_{22}^{-1}J_{21})^{-1} & -S^{-1}J_{12}J_{22}^{-1} \\ -J_{22}^{-1}J_{21}S^{-1} & J_{22}^{-1} + J_{22}^{-1}J_{21}S^{-1}J_{12}J_{22}^{-1} \end{bmatrix}$$

where  $S = J_{11} - J_{12}J_{22}^{-1}J_{21}$ , which gives

$$\Lambda_{11} = (J_{11} - J_{12}J_{22}^{-1}J_{21})^{-1}$$

hence,

$$J' = \Lambda_{11}^{-1} = J_{11} - J_{12}J_{22}^{-1}J_{21}$$



- Proof of  $h' = J'm_1 = h_1 - J_{12}J_{22}^{-1}h_2$

► notice that since

$$\Lambda = \begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix}^{-1} = \begin{bmatrix} S^{-1} & -S^{-1}J_{12}J_{22}^{-1} \\ -J_{22}^{-1}J_{21}S^{-1} & J_{22}^{-1} + J_{22}^{-1}J_{21}S^{-1}J_{12}J_{22}^{-1} \end{bmatrix}$$

where  $S = J_{11} - J_{12}J_{22}^{-1}J_{21}$ , we know from  $m = \Lambda h$  that

$$m_1 = \begin{bmatrix} S^{-1} & -S^{-1}J_{12}J_{22}^{-1} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix}$$

since  $J' = S$ , we have

$$h' = J'm_1 = \begin{bmatrix} \mathbb{I} & -J_{12}J_{22}^{-1} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix}$$

□

- **conditioning** is easy to compute when  $x$  is in information form

$$x_1|x_2 \sim \mathcal{N}^{-1}\left(h_1 - J_{12}x_2, J_{11}\right)$$

**proof:** treat  $x_2$  as a constant to get

$$\begin{aligned}\mu(x_1|x_2) &\propto \mu(x_1, x_2) \\ &\propto \exp\left\{-\frac{1}{2}\begin{bmatrix}x_1^T & x_2^T\end{bmatrix}\begin{bmatrix}J_{11} & J_{12} \\ J_{21} & J_{22}\end{bmatrix}\begin{bmatrix}x_1 \\ x_2\end{bmatrix} + \begin{bmatrix}h_1^T & h_2^T\end{bmatrix}\begin{bmatrix}x_1 \\ x_2\end{bmatrix}\right\} \\ &\propto \exp\left\{-\frac{1}{2}(x_1^T J_{11} x_1 + 2x_2^T J_{21} x_1) + h_1^T x_1\right\} \\ &= \exp\left\{-\frac{1}{2}x_1^T J_{11} x_1 + (h_1 - J_{12}x_2)^T x_1\right\}\end{aligned}$$

but complicated when  $x$  is in covariance form

$$x_1|x_2 \sim \mathcal{N}(m', \Lambda')$$

where  $m' = m_1 + \Lambda_{12}\Lambda_{22}^{-1}(x_2 - m_2)$  and  $\Lambda' = \Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}$

# Gaussian graphical model

**theorem 1.** For  $x \sim \mathcal{N}(m, \Lambda)$ ,  $x_i$  and  $x_j$  are independent if and only if  $\Lambda_{ij} = 0$

Q. for what other distribution does uncorrelation imply independence?

**theorem 2.** For  $x \sim \mathcal{N}^{-1}(h, J)$ ,  $x_i \perp x_{V \setminus \{i, j\}} \perp x_j$  if and only if  $J_{ij} = 0$

Q. is it obvious?

- graphical model representation of Gaussian random vectors
  - ▶  $J$  encodes the pairwise Markov independencies
  - ▶ obtain Gaussian graphical model by adding an edge whenever  $J_{ij} \neq 0$

$$\begin{aligned}\mu(x) &\propto \exp \left\{ -\frac{1}{2} x^T J x + h^T x \right\} \\ &= \prod_{i \in V} \underbrace{e^{-\frac{1}{2} x_i^T J_{ii} x_i + h_i^T x_i}}_{\psi_i(x_i)} \prod_{(i,j) \in E} \underbrace{e^{-\frac{1}{2} x_i^T J_{ij} x_j}}_{\psi_{ij}(x_i, x_j)}\end{aligned}$$

- ▶ is pairwise Markov property enough?
- ▶ Is pairwise Markov Random Field enough?

**problem:** compute marginals  $\mu(x_i)$  when  $G$  is a tree

- ▶ messages and marginals are Gaussian, completely specified by mean and variance
- ▶ simple algebra to compute integration

## Gaussian belief propagation on trees

- initialize messages on the leaves as Gaussian (each node has  $x_i$  which can be either a scalar or a vector)

$$\nu_{i \rightarrow j}(x_i) = \psi_i(x_i) = e^{-\frac{1}{2} x_i^T J_{ii} x_i + h_i^T x_i} \sim \mathcal{N}^{-1}(h_{i \rightarrow j}, J_{i \rightarrow j})$$

- therefore, messages are also Gaussian  $\nu_{i \rightarrow j}(x_i) \sim \mathcal{N}^{-1}(h_{i \rightarrow j}, J_{i \rightarrow j})$
- completely specified by two parameters: mean and variance

- Gaussian belief propagation

$$h_{i \rightarrow j} = h_i - \sum_{k \in \partial i \setminus j} J_{ik} J_{k \rightarrow i}^{-1} h_{k \rightarrow i}$$

$$J_{i \rightarrow j} = J_{ii} - \sum_{k \in \partial i \setminus j} J_{ik} J_{k \rightarrow i}^{-1} J_{ki}$$

- marginal can be computed as  $x_i \sim \mathcal{N}^{-1}(\hat{h}_i, \hat{J}_i)$

$$\hat{h}_i = h_i - \sum_{k \in \partial i} J_{ik} J_{k \rightarrow i}^{-1} h_{k \rightarrow i}$$

$$\hat{J}_i = J_{ii} - \sum_{k \in \partial i} J_{ik} J_{k \rightarrow i}^{-1} J_{ki}$$

- for  $x_i \in \mathbb{R}^d$  Gaussian BP requires  $O(n \cdot d^3)$  operations on a tree
  - ▶ matrix inversion can be computed in  $O(d^3)$  (e.g., Gaussian elimination)
- if we naively invert the information matrix  $J_{22}$  of the entire graph

$$x_1 \sim \mathcal{N}^{-1}(h_1 - J_{12} J_{22}^{-1} h_2, J_{11} - J_{12} J_{22}^{-1} J_{21})$$

requires  $O((nd)^3)$  operations



- MAP configuration

- ▶ for Gaussian random vectors, mean is the mode

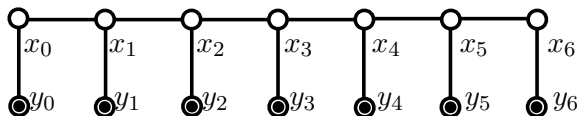
$$\max_x \exp \left\{ -\frac{1}{2}(x - m)^T \Lambda^{-1}(x - m) \right\}$$

taking the gradient of the exponent

$$\frac{\partial}{\partial x} \left\{ -\frac{1}{2}(x - m)^T \Lambda^{-1}(x - m) \right\} = -\Lambda^{-1}(x - m)$$

hence the mode  $x^* = m$

# Gaussian hidden Markov models



- Gaussian HMM

- ▶ states  $x_t \in \mathbb{R}^d$
- ▶ state transition matrix  $A \in \mathbb{R}^{d \times d}$
- ▶ process noise  $v_t \in \mathbb{R}^p$  and  $\sim \mathcal{N}(0, V)$  for some  $V \in \mathbb{R}^{p \times p}$ ,  $B \in \mathbb{R}^{d \times p}$

$$x_{t+1} = Ax_t + Bv_t$$

$$x_0 \sim \mathcal{N}(0, \Lambda_0)$$

- ▶ observation  $y_t \in \mathbb{R}^{d'}$ ,  $C \in \mathbb{R}^{d' \times d}$
- ▶ observation noise  $w_t \sim \mathcal{N}(0, W)$  for some  $R \in \mathbb{R}^{d' \times d'}$

$$y_t = Cx_t + w_t$$

- in summary, for  $H = BVB^T$

$$\begin{aligned}x_0 &\sim \mathcal{N}(0, \Lambda_0) \\ x_{t+1}|x_t &\sim \mathcal{N}(Ax_t, \textcolor{red}{H}) \\ y_t|x_t &\sim \mathcal{N}(Cx_t, W)\end{aligned}$$

- factorization

$$\begin{aligned}\mu(x, y) &= \mu(x_0)\mu(y_0|x_0)\mu(x_1|x_0)\mu(y_1|x_1)\cdots \\ &\propto \exp\left(-\frac{1}{2}x_0^T\Lambda_0^{-1}x_0\right)\exp\left(-\frac{1}{2}(y_0-Cx_0)^TW^{-1}(y_0-Cx_0)\right) \\ &\quad \exp\left(-\frac{1}{2}(x_1-Ax_0)^TH^{-1}(x_1-Ax_0)\right)\cdots \\ &= \prod_{k=0}^t \psi_k(x_k) \prod_{k=1}^t \psi_{k-1,k}(x_{k-1}, x_k) \prod_{k=0}^t \phi_k(y_k) \prod_{k=0}^t \phi_{k,k}(x_k, y_k)\end{aligned}$$

- factorization

$$\mu(x, y) \propto \prod_{k=0}^t \psi_k(x_k) \prod_{k=1}^t \psi_{k-1,k}(x_{k-1}, x_k) \prod_{k=0}^t \phi_k(y_k) \prod_{k=0}^t \phi_{k,k}(x_k, y_k)$$

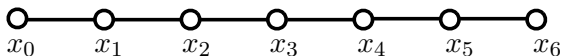
$$\log \psi_k(x_k) = \begin{cases} -\frac{1}{2}x_0^T \underbrace{(A_0^{-1} + C^T W^{-1} C + A^T H^{-1} A)}_{\equiv J_0} x_0 & k=0 \\ -\frac{1}{2}x_k^T \underbrace{(H^{-1} + C^T W^{-1} C + A^T H^{-1} A)}_{\equiv J_k} x_k & 0 < k < t \\ -\frac{1}{2}x_t^T \underbrace{(H^{-1} + C^T W^{-1} C)}_{\equiv J_t} x_t & k=t \end{cases}$$

$$\log \psi_{k-1,k}(x_{k-1}, x_k) = x_{k-1}^T \underbrace{H^{-1} A}_{\equiv L_k} x_k$$

$$\log \phi_k(y_k) = -\frac{1}{2} y_k^T W^{-1} y_k$$

$$\log \phi_{k,k}(x_k, y_k) = x_k^T \underbrace{C^T W^{-1}}_{\equiv M_k} y_k$$

- **problem:** given observations  $y$  estimate hidden states  $x$



$$\mu(x|y) \propto \prod_{k=0}^t \exp \left\{ -\frac{1}{2} x_k^T J_k x_k + x_k^T \underbrace{M_k y_k}_{h_k} \right\} \prod_{k=1}^t \exp \left\{ -x_k^T \underbrace{(-L_k)}_{J_{k,k-1}} x_{k-1} \right\}$$

- use Gaussian BP to compute marginals for this Gaussian graphical model on a line

- ▶ initialize

$$J_{0 \rightarrow 1} = J_0, \quad h_{0 \rightarrow 1} = h_0$$

$$J_{6 \rightarrow 5} = J_6, \quad h_{6 \rightarrow 5} = h_6$$

- ▶ forward update

$$J_{i \rightarrow i+1} = J_i - L_i J_{i-1 \rightarrow i}^{-1} L_i^T$$

$$h_{i \rightarrow i+1} = h_i - L_i J_{i-1 \rightarrow i}^{-1} h_{i-1 \rightarrow i}$$

- ▶ backward update

$$J_{i \rightarrow i-1} = J_{i+1} - L_{i+1} J_{i+1 \rightarrow i}^{-1} L_{i+1}^T$$

$$h_{i \rightarrow i-1} = h_{i+1} - L_{i+1} J_{i+1 \rightarrow i}^{-1} h_{i+1 \rightarrow i}$$

- compute marginals

$$\tilde{J}_i = J_i - L_i J_{i-1 \rightarrow i}^{-1} L_i^T - L_{i+1} J_{i+1 \rightarrow i}^{-1} L_{i+1}^T$$

$$\tilde{h}_i = h_i - L_i J_{i-1 \rightarrow i}^{-1} h_{i-1 \rightarrow i} - L_{i+1} J_{i+1 \rightarrow i}^{-1} h_{i+1 \rightarrow i}$$

- the marginal is

$$x_i \sim \mathcal{N}(\tilde{J}_i^{-1} \tilde{h}_i, \tilde{J}_i^{-1})$$

## Correctness

- there is little theoretical understanding of loopy belief propagation (except for graphs with a single loop)
- perhaps surprisingly, loopy belief propagation (if it converges) gives the correct **mean** of Gaussian graphical models even if the graph has loops (convergence of the variance is not guaranteed)
- **Theorem** [Weiss, Freeman 2001, Rusmevichientong, Van Roy 2001]  
If Gaussian belief propagation *converges*, then the expectations are computed correctly: let

$$\hat{m}_i^{(\ell)} \equiv (\hat{J}_i^{(\ell)})^{-1} \hat{h}_i^{(\ell)}$$

where  $\hat{m}_i^{(\ell)}$  = belief propagation expectation after  $\ell$  iterations

$\hat{J}_i^{(\ell)}$  = belief propagation information matrix after  $\ell$  iterations

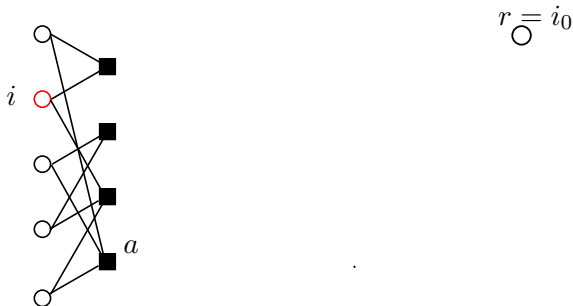
$\hat{h}_i^{(\ell)}$  = belief propagation precision after  $\ell$  iterations and if

$\hat{m}_i^{(\infty)} \triangleq \lim_{\ell \rightarrow \infty} \hat{m}_i^{(\ell)}$  exists, then

$$\hat{m}_i^{(\infty)} = m_i$$

## A detour: Computation tree

- what is  $\hat{m}_i^{(\ell)}$ ?
- **computation tree**  $\text{CT}_G(i; \ell)$  is the tree of  $\ell$ -steps non-reversing walks on  $G$  starting at  $i$ .

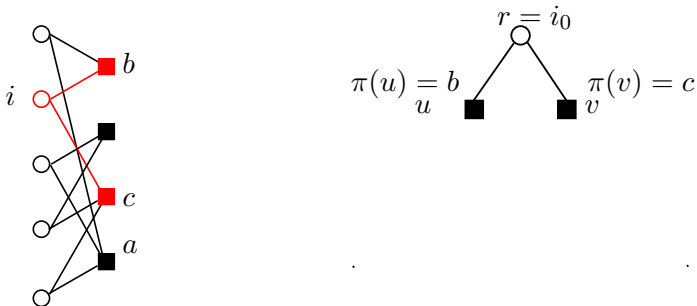


- $i, j, k, \dots, a, b, \dots$  for nodes in  $G$  and  $r, s, t, \dots$  for nodes in  $\text{CT}_G(i; \ell)$
- potentials  $\psi_i$  and  $\psi_{ij}$  are copied to  $\text{CT}_G(i; \ell)$
- each node (edge) in  $G$  corresponds to multiple nodes (edges) in  $\text{CT}_G(i; \ell)$ .
- natural projection  $\pi : \text{CT}_G(i; \ell) \rightarrow G$ , e.g.,  $\pi(t) = \pi(s) = j$



## A detour: Computation tree

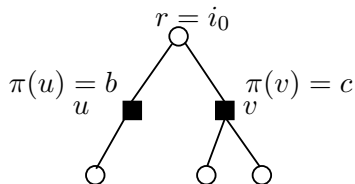
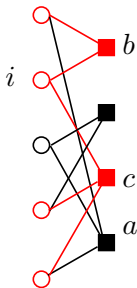
- what is  $\hat{m}_i^{(\ell)}$ ?
- **computation tree**  $\text{CT}_G(i; \ell)$  is the tree of  $\ell$ -steps non-reversing walks on  $G$  starting at  $i$ .



- $i, j, k, \dots, a, b, \dots$  for nodes in  $G$  and  $r, s, t, \dots$  for nodes in  $\text{CT}_G(i; \ell)$
- potentials  $\psi_i$  and  $\psi_{ij}$  are copied to  $\text{CT}_G(i; \ell)$
- each node (edge) in  $G$  corresponds to multiple nodes (edges) in  $\text{CT}_G(i; \ell)$ .
- natural projection  $\pi : \text{CT}_G(i; \ell) \rightarrow G$ , e.g.,  $\pi(t) = \pi(s) = j$

## A detour: Computation tree

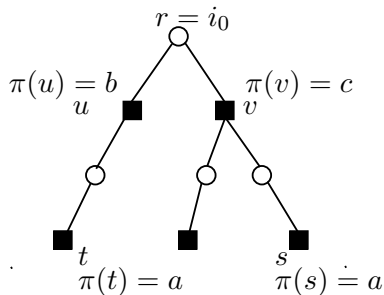
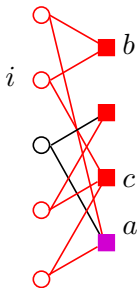
- what is  $\hat{m}_i^{(\ell)}$ ?
- **computation tree**  $\text{CT}_G(i; \ell)$  is the tree of  $\ell$ -steps non-reversing walks on  $G$  starting at  $i$ .



- $i, j, k, \dots, a, b, \dots$  for nodes in  $G$  and  $r, s, t, \dots$  for nodes in  $\text{CT}_G(i; \ell)$
- potentials  $\psi_i$  and  $\psi_{ij}$  are copied to  $\text{CT}_G(i; \ell)$
- each node (edge) in  $G$  corresponds to multiple nodes (edges) in  $\text{CT}_G(i; \ell)$ .
- natural projection  $\pi : \text{CT}_G(i; \ell) \rightarrow G$ , e.g.,  $\pi(t) = \pi(s) = j$

## A detour: Computation tree

- what is  $\hat{m}_i^{(\ell)}$ ?
- **computation tree**  $\text{CT}_G(i; \ell)$  is the tree of  $\ell$ -steps non-reversing walks on  $G$  starting at  $i$ .



- $i, j, k, \dots, a, b, \dots$  for nodes in  $G$  and  $r, s, t, \dots$  for nodes in  $\text{CT}_G(i; \ell)$
- potentials  $\psi_i$  and  $\psi_{ij}$  are copied to  $\text{CT}_G(i; \ell)$
- each node (edge) in  $G$  corresponds to multiple nodes (edges) in  $\text{CT}_G(i; \ell)$ .
- natural projection  $\pi : \text{CT}_G(i; \ell) \rightarrow G$ , e.g.,  $\pi(t) = \pi(s) = j$

## What is $\hat{m}_i^{(\ell)}$ ?

- **Claim 1.**  $\hat{m}_i^{(\ell)}$  is  $\hat{m}_r^{(\ell)}$ , which is the expectation of  $x_r$  w.r.t. Gaussian model on  $\text{CT}_G(i; \ell)$ 
  - ▶ **proof of claim 1.** by induction over  $\ell$ .
  - ▶ idea: BP ‘does not know’ whether it is operating on  $G$  or on  $\text{CT}_G(i; \ell)$

- recall that for Gaussians, mode of  $-\frac{1}{2}x^T Jx + h^T x$  is the mean  $m$ , hence

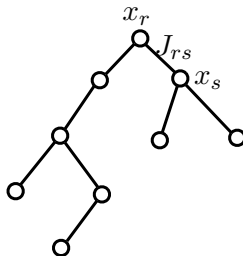
$$Jm = h$$

and since  $J$  is invertible (due to positive definiteness),  $m = J^{-1}h$ .

- locally,  $m$  is the unique solution that satisfies all of the following series of equations for all  $i \in V$

$$J_{ii}m_i + \sum_{j \in \partial i} J_{ij}m_j = h_i$$

- similarly, for a Gaussian graphical model on  $\text{CT}_G(i; \ell)$



the estimated mean  $\hat{m}^{(\ell)}$  is exact on a tree. Precisely, since the width of the tree is at most  $2\ell$ , the BP updates on  $\text{CT}_G(i; \ell)$  converge to the correct marginals for  $t \geq 2\ell$  and satisfy

$$J_{rr}\hat{m}_r^{(t)} + \sum_{s \in \partial r} J_{rs}\hat{m}_s^{(t)} = h_r$$

where  $r$  is the root of the computation tree. In terms of the original information matrix  $J$  and potential  $h$

$$J_{\pi(r), \pi(r)}\hat{m}_r^{(t)} + \sum_{s \in \partial r} J_{\pi(r), \pi(s)}\hat{m}_s^{(t)} = h_{\pi(r)}$$

since we copy  $J$  and  $h$  for each edge and node in  $\text{CT}_G(i; \ell)$ .

- ▶ note that on the computation tree  $\text{CT}_G(i, ; \ell)$ ,  $\hat{m}_r^{(t)} = \hat{m}_r^{(\ell)}$  for  $t \geq \ell$  since the root  $r$  is at most distance  $\ell$  away from any node.
- ▶ similarly, for a neighbor  $s$  of the root  $r$ ,  $\hat{m}_s^{(t)} = \hat{m}_s^{(\ell+1)}$  for  $t \geq \ell + 1$  since  $s$  is at most distance  $\ell + 1$  away from any node.
- ▶ hence we can write the above equation as

$$J_{\pi(r), \pi(r)} \hat{m}_r^{(\ell)} + \sum_{s \in \partial r} J_{\pi(r), \pi(s)} \hat{m}_s^{(\ell+1)} = h_{\pi(r)} \quad (1)$$

if the BP fixed point converges then

$$\lim_{\ell \rightarrow \infty} \hat{m}_i^{(\ell)} = \hat{m}_i^{(\infty)}$$

we claim that  $\lim_{\ell \rightarrow \infty} \hat{m}_r^{(\ell)} = \hat{m}_{\pi(r)}^{(\infty)}$ , since

$$\lim_{\ell \rightarrow \infty} \hat{m}_r^{(\ell)} = \lim_{\ell \rightarrow \infty} \hat{m}_{\pi(r)}^{(\ell)} \quad \text{by Claim 1.}$$

$$= \hat{m}_{\pi(r)}^{(\infty)} \quad \text{by the convergence assumption}$$

we can generalize this argument (without explicitly proving it in this lecture) to claim that in the computation tree  $\text{CT}_G(i; \ell)$  if we consider a neighbor  $s$  of the root  $r$ ,

$$\lim_{\ell \rightarrow \infty} \hat{m}_s^{(\ell+1)} = \hat{m}_{\pi(s)}^{(\infty)}$$

# Convergence

from Eq. (1), we have

$$J_{\pi(r),\pi(r)}\hat{m}_r^{(\ell)} + \sum_{s \in \partial r} J_{\pi(r),\pi(s)}\hat{m}_s^{(\ell+1)} = h_{\pi(r)}$$

taking the limit  $\ell \rightarrow \infty$ ,

$$J_{\pi(r),\pi(r)}\hat{m}_{\pi(r)}^{(\infty)} + \sum_{s \in \partial r} J_{\pi(r),\pi(s)}\hat{m}_{\pi(s)}^{(\infty)} = h_{\pi(r)}$$

hence, BP is exact on the original graph with loops assuming convergence, i.e. BP is correct:

$$\begin{aligned} J_{i,i}\hat{m}_i^{(\infty)} + \sum_{j \in \partial i} J_{i,j}\hat{m}_j^{(\infty)} &= h_i \\ J\hat{m}^{(\infty)} &= h \end{aligned}$$



# What have we achieved?

- complexity?
- convergence?
- **correlation decay**: the influence of leaf nodes on the computation tree decreases as iterations increase
- understanding BP in a broader class of graphical models (loopy belief propagation)
- help clarify the empirical performance results (e.g. Turbo codes)

# Gaussian Belief Propagation (GBP)

- **Sufficient conditions** for convergence and correctness of GBP
  - ▶ Rusmevichientong and Van Roy (2001), Wainwright, Jaakkola, Willsky (2003) : if means converge, then they are correct
  - ▶ Weiss and Freeman (2001): if the information matrix is diagonally dominant, then GBP converges
  - ▶ convergence known for trees, attractive, non-frustrated, and diagonally dominant Gaussian graphical models
  - ▶ Malioutov, Johnson, Willsky (2006): **walk-summable** graphical models converge (this includes all of the known cases above)
  - ▶ Moallemi and Van roy (2006): if **pairwise normalizable** then consensus propagation converges