7. Gaussian graphical models

- Gaussian graphical models
- Gaussian belief propagation
- Kalman filtering
- Example: consensus propagation
- Convergence and correctness

Gaussian graphical models

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

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

- 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 Λ).
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

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

  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

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

  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

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

$\square$

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

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

proof: treat x_2 as a constant to get

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

but complicated when x is in covariance form

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

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–$x_{V\setminus\{i,j\}}$–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 $J_{ij} \neq 0$

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

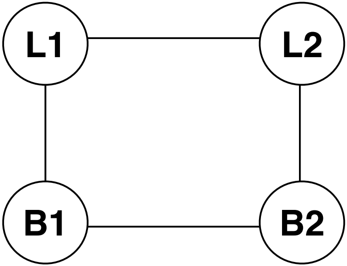
  - is pairwise Markov property enough?
  - Is pairwise Markov Random Field enough?

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

example: heredity of head dimensions [Frets 1921]

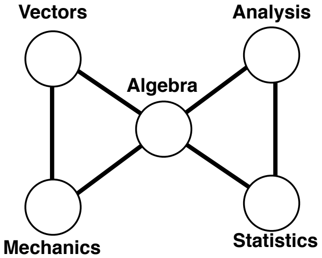
- estimated mean and covariance of four dimensional vector (L_1, B_1, L_2, B_2)
- lengths and breadths of first and second born sons are measured
- 25 samples
- analyses by [Whittaker 1990] support the following Gaussian graphical model



example: mathematics scores [Whittaker 1990]

- Examination scores of 88 students in 5 subjects
- empirical information matrix (diagonal and above) covariance (below diagonal)

	Mechanics	Vectors	Algebra	Analysis	Statistics
Mechanics	5.24	-2.44	-2.74	0.01	-0.14
Vectors	0.33	10.43	-4.71	-0.79	-0.17
Algebra	0.23	0.28	26.95	-7.05	-4.70
Analysis	0.00	0.08	0.43	9.88	-2.02
Statistics	0.02	0.02	0.36	0.25	6.45



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)

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

where $h_{i\to j} = h_i$ and $J_{i\to j} = J_{ii}$

- update messages assuming $\nu_{k\to i}(x_k) \sim \mathcal{N}^{-1}(h_{k\to i}, J_{k\to i})$

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

- evaluating the integration ($=$ marginalizing Gaussian)

$$\begin{aligned}
\int \psi_{ik}(x_i, x_k)\nu_{k\to i}(x_k)\, dx_k &= \int e^{-x_i^T J_{ki} x_k - \frac{1}{2}x_k^T J_{k\to i} x_k + h_{k\to i}^T x_k}\, dx_k \\
&= \int \exp\left\{ -\frac{1}{2}[x_i^T \;\; x_k^T] \begin{bmatrix} 0 & J_{ik} \\ J_{ik} & J_{k\to i} \end{bmatrix} \begin{bmatrix} x_i \\ x_k \end{bmatrix} + [0 \; h_{k\to i}^T] \begin{bmatrix} x_i \\ x_k \end{bmatrix} \right\} dx_k \\
&\sim \mathcal{N}^{-1}\big(-J_{ik}J_{k\to i}^{-1}h_{k\to i}, -J_{ik}J_{k\to i}^{-1}J_{ki} \big)
\end{aligned}$$

since this is evaluating the marginal of x_i for $(x_i, x_k) \sim \mathcal{N}^{-1}\left(\begin{bmatrix} 0 \\ h_{k\to i} \end{bmatrix}, \begin{bmatrix} 0 & J_{ik} \\ J_{ik} & J_{k\to i} \end{bmatrix} \right)$.

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

$$\begin{aligned} h_{i\to j} &= h_i - \sum_{k\in\partial i\setminus j} J_{ik} J_{k\to i}^{-1} h_{k\to i} \\ J_{i\to j} &= J_{ii} - \sum_{k\in\partial i\setminus j} J_{ik} J_{k\to i}^{-1} J_{ki} \end{aligned}$$

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

$$\begin{aligned} \hat{h}_i &= h_i - \sum_{k\in\partial i} J_{ik} J_{k\to i}^{-1} h_{k\to i} \\ \hat{J}_i &= J_{ii} - \sum_{k\in\partial i} J_{ik} J_{k\to i}^{-1} J_{ki} \end{aligned}$$

- 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 $O\big((nd)^3\big)$ operations

- connections to **Gaussian elimination**
  - one way to view Gaussian BP is that given J and h it computes

$$m = J^{-1}h$$

  - this implies that for any positive-definite matrix A with tree structure, we can use Gaussian BP to solve for x

$$Ax = b \qquad\qquad (x = A^{-1}b)$$

  - **example:** Gaussian elimination

$$\begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$$

$$\begin{bmatrix} 4 - \frac{2}{3} \cdot 2 & 2 - \frac{2}{3} \cdot 3 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 - \frac{2}{3} \cdot 3 \\ 3 \end{bmatrix}$$

$$\begin{bmatrix} \frac{8}{3} & 0 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

  - Gaussian elimination that exploits tree structure by eliminating from the leaves is equivalent as Gaussian BP

- MAP configuration
  - for Gaussian random vectors, mean is the mode

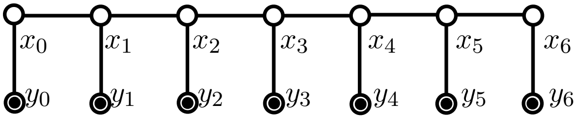
$$\max_x \ \exp\Big\{-\frac{1}{2}(x-m)^T\Lambda^{-1}(x-m)\Big\}$$

taking the gradient of the exponent

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

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 $V \in \mathbb{R}^{p \times p}$, $B \in \mathbb{R}^{d \times p}$

$$\begin{aligned} x_{t+1} &= Ax_t + Bv_t \\ x_0 &\sim \mathcal{N}(0, \Lambda_0) \end{aligned}$$

  - 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 $H = BVB^T$

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

- factorization

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

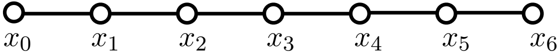
- factorization

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

$$\log \psi_k(x_k) = \begin{cases} -\frac{1}{2} x_0^T (\underbrace{\Lambda_0^{-1} + C^T W^{-1} C + A^T H^{-1} A}_{\equiv\ J_0}) x_0 & k = 0 \\ -\frac{1}{2} x_k^T (\underbrace{H^{-1} + C^T W^{-1} C + A^T H^{-1} A}_{\equiv\ J_k}) x_k & 0 < k < t \\ -\frac{1}{2} x_t^T (\underbrace{H^{-1} + C^T W^{-1} C}_{\equiv\ J_t}) x_t & k = t \end{cases}$$

$$\begin{aligned} \log \psi_{k-1,k}(x_{k-1},x_k) &= x_k^T \underbrace{H^{-1} A}_{\equiv\ L_k} x_{k-1} \\ \log \phi_k(y_k) &= -\frac{1}{2} y_k^T W^{-1} y_k \\ \log \phi_{k,k}(x_k,y_k) &= x_k^T \underbrace{C^T W^{-1}}_{\equiv\ M_k} y_k \end{aligned}$$

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



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

- use Gaussian BP to compute marginals for this Gaussian graphical model on a line
  - initialize
$$\begin{aligned} J_{0\to 1} &= J_0, \qquad & h_{0\to 1} &= h_0 \\ J_{6\to 5} &= J_6, \qquad & h_{6\to 5} &= h_6 \end{aligned}$$
  - forward update
$$\begin{aligned} J_{i\to i+1} &= J_i - L_i J_{i-1\to i}^{-1} L_i^T \\ h_{i\to i+1} &= h_i - L_i J_{i-1\to i}^{-1} h_{i-1\to i} \end{aligned}$$
  - backward update
$$\begin{aligned} J_{i\to i-1} &= J_i - L_{i+1} J_{i+1\to i}^{-1} L_{i+1}^T \\ h_{i\to i-1} &= h_i - L_{i+1} J_{i+1\to i}^{-1} h_{i+1\to i} \end{aligned}$$

- compute marginals

$$\begin{aligned}
\hat{J}_i &= J_i - L_i J_{i-1\to i}^{-1} L_i^T - L_{i+1} J_{i+1\to i}^{-1} L_{i+1}^T \\
\hat{h}_i &= h_i - L_i J_{i-1\to i}^{-1} h_{i-1\to i} - L_{i+1} J_{i+1\to i}^{-1} h_{i+1\to i}
\end{aligned}$$

- the marginal is

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

Kalman filtering (1959)

- important problem in control
- provides a different perspective on Gaussian HMMs
- **problem:** linear quadratic estimation (LQE)
  - minimize the quadratic loss:

$$L(x, \widehat{x}(y)) \quad = \quad \sum_k (\widehat{x}(y)_k - x_k)^2 = \left(\widehat{x}(y) - x\right)^T \left(\widehat{x}(y) - x\right)$$

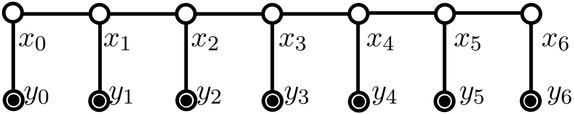
  - since x is random, we minimize the expected loss

$$\mathbb{E}\big[L(x, \widehat{x}(y))|y\big] \quad = \quad \widehat{x}(y)^T \widehat{x}(y) + \mathbb{E}[x^T x|y] - 2\widehat{x}(y)^T \mathbb{E}[x|y]$$

  - taking the gradient w.r.t $\widehat{x}(y)$ and setting it equal to zero yields

$$2\widehat{x}(y) - 2\mathbb{E}[x|y] \quad = \quad 0$$

  - minimum mean-squared error estimate is $\widehat{x}^*(y) = \mathbb{E}[x|y]$



- Linear dynamical systems with Gaussian noise (= 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 $V \in \mathbb{R}^{p\times p}$, $B \in \mathbb{R}^{d\times p}$

$$\begin{aligned} x_{t+1} &= Ax_t + Bv_t \\ x_0 &\sim \mathcal{N}(0, \Lambda_0) \end{aligned}$$

  - 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 $R \in \mathbb{R}^{d'\times d'}$

$$y_t = Cx_t + w_t$$

  - all noise are independent

Conditioning on observed output

- we use notations

$$
\begin{aligned}
x_{t|s} &= \mathbb{E}[x_t | y_0, \cdots, y_s] \\
\Sigma_{t|s} &= \mathbb{E}[(x_t - x_{t|s})(x_t - x_{t|s})^T | y_0, \cdots, y_s]
\end{aligned}
$$

  - the random variable $x_t | y_0, \cdots, y_s$ is Gaussian with mean $x_{t|s}$ and covariance $\Sigma_{t|s}$
  - $x_{t|s}$ is the minimum mean-square error estimate of x_t given $y_0, \cdots, y_s$
  - $\Sigma_{t|s}$ is the covariance of the error of the estimate $x_{t|s}$
- we focus on two state estimation problems:
  - finding $x_{t|t}$, i.e., estimating the current state based on the current and past observations
  - finding $x_{t+1|t}$, i.e., predicting the next state based on the current and past observations
- **Kalman filter** is a clever method for computing $x_{t|t}$ and $x_{t+1|t}$ recursively

Measurement update

- let's find $x_{t|t}$ and $\Sigma_{t|t}$ in terms of $x_{t|t-1}$ and $\Sigma_{t|t-1}$
- let $Y_{t-1} = (y_0, \cdots, y_{t-1})$, then

$$y_t|Y_{t-1} = Cx_t|Y_{t-1} + w_t|Y_{t-1} = Cx_t|Y_{t-1} + w_t$$

  since w_t and Y_{t-1} are independent
- so $x_t|Y_{t-1}$ and $y_t|Y_{t-1}$ are jontly Gaussian with mean and covariance

$$\begin{bmatrix} x_{t|t-1} \\ C\, x_{t|t-1} \end{bmatrix}, \quad \begin{bmatrix} \Sigma_{t|t-1} & \Sigma_{t|t-1}C^T \\ C\Sigma_{t|t-1} & C\Sigma_{t|t-1}C^T + W \end{bmatrix}$$

- now use standard formula for conditoining Gaussian random vector to get mean and variance of

$$\big(x_t|Y_{t-1}\big) \,\big|\, \big(y_t|Y_{t-1}\big)$$

which is exactly the same as $x_t|Y_t$

$$\begin{aligned} x_{t|t} &= x_{t|t-1} + \Sigma_{t|t-1}C^T(C\Sigma_{t|t-1}C^T + W)^{-1}(y_t - Cx_{t|t-1}) \\ \Sigma_{t|t} &= \Sigma_{t|t-1} - \Sigma_{t|t-1}C^T(C\Sigma_{t|t-1}C^T + W)^{-1}C\Sigma_{t|t-1} \end{aligned}$$

- this recursively defines $x_{t|t}$ and $\Sigma_{t|t}$ in terms of $x_{t|t-1}$ and $\Sigma_{t|t-1}$
- this is called *measurement update* since it gives our updated estimate of x_t based on the measurement y_t becoming available

Time update

- now we increment time using $x_{t+1} = Ax_t + Bv_t$
- condition on Y_t to get

$$x_{t+1}|Y_t \;=\; Ax_t|Y_t + Bv_t|Y_t \;=\; Ax_t|Y_t + Bv_t$$

since v_t is independent of Y_t

- therefore $x_{t+1|t} = Ax_{t|t}$ and

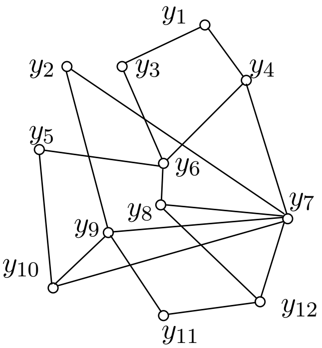
$$\begin{aligned}
\Sigma_{t+1|t} &= \mathbb{E}\big[(x_{t+1|t} - x_{t+1})(x_{t+1|t} - x_{t+1})^T\big] \\
&= \mathbb{E}\big[(Ax_{t|t} - Ax_t - Bv_t)(Ax_{t|t} - Ax_t - Bv_t)^T\big] \\
&= A\Sigma_{t|t}A^T + BVB^T
\end{aligned}$$

Kalman filter

- **Kalman filter:**
  - measurement update and time update together give a recursion
  - start with $x_{0|-1} = 0$ and $\Sigma_{0|-1} = \Lambda_0$
  - apply measurement update to get $x_{0|0}$ and $\Sigma_{0|0}$
  - apply time update to get $x_{1|0}$ and $\Sigma_{1|0}$
  - repeat ...
- we have an *efficient* recursion to compute $x_{t|t} = \arg\min_x \mathbb{E}\big[(x_t - x)^T(x_t - x)|y_0 \ldots, y_t\big]$
- notice there is no backward update as in Gaussian BP, because we are interested in real time estimation: estimate current state given observations so far

Example #1: Consensus propagation

[Moallemi and Van Roy, 2006]



Observations of the 'state of the world': $y_1, y_2, \dots, y_n$
Objective: compute at each node the mean

$$\overline{y} \equiv \frac{1}{n} \sum_{i=1}^{n} y_i$$

Bottle-neck: communication allowed on the edges

Graphical model approach

- define a Gaussian graphical model $\mu_y(x)$ on G with parameters $y = (y_1, y_2, \ldots, y_n)$ such that

$$\mathbb{E}_\mu\{x_i\} = \overline{y}$$

- equivalently, define J and h such that $m = J^{-1}h = \overline{y}\mathbf{1}$
- of course we could define $J = \mathbf{I}$ and $h = \overline{y}\mathbf{1}$

$$\begin{aligned} \mu_y(x) &= \frac{1}{(2\pi)^{n/2}} \exp\Big\{ -\frac{1}{2} \|x - \overline{y}\,\mathbf{1}\|_2^2 \Big\} \\ &= \prod_{i \in V} \frac{1}{(2\pi)^{1/2}} \exp\Big\{ -\frac{1}{2} (x_i - \overline{y})^2 \Big\} \end{aligned}$$

...but this does not address the problem.

- use a Gaussian graphical model

$$\mu_y(x) \;=\; \frac{1}{Z}\,\exp\Big\{ -\frac{\gamma}{2}\sum_{(i,j)\in E}(x_i - x_j)^2 - \frac{1}{2}\sum_{i\in V}(x_i - y_i)^2 \Big\}$$

and solve for x (hoping that the solution m is close to $\bar{y}\mathbf{1}$)

$$m \;=\; \arg\min_{x\in\mathbb{R}^n}\;\; \frac{\gamma}{2}\sum_{(i,j)\in E}(x_i - x_j)^2 + \frac{1}{2}\sum_{i\in V}(x_i - y_i)^2$$

- **intuition:** as $\gamma \to \infty$, $x_i \approx x_j$ for all $i, j \in V$. Hence $x_i \approx \overline{x}$:

$$m \;=\; \arg\min_{\xi\in\mathbb{R}}\Big\{ \sum_{i\in V}(\xi - y_i)^2 \Big\} \;=\; \frac{1}{n}\sum_{i\in V} y_i$$

Graph Laplacian

- **Laplacian** of a graph is defined as

$$(\mathcal{L}_G)_{ij} = \begin{cases} -1 & \text{if } (i,j) \in E, \\ \deg_G(i) & \text{if } i = j. \end{cases}$$

- for $x \in \mathbb{R}^n$, we have $\langle x, \mathcal{L}_G\, x \rangle = \frac{1}{2} \sum_{(i,j)\in E} (x_i - x_j)^2$. In particular,
  - $\mathcal{L}_G \succeq 0$
  - $\mathcal{L}_G \mathbf{1} = 0$
  - If G is connected, then the null space of $\mathcal{L}_G$ has dimension one

- rewriting things,

$$\begin{aligned}\mu_y(x) &= \frac{1}{Z}\exp\Big\{-\frac{\gamma}{2}\sum_{(i,j)\in E}(x_i-x_j)^2-\frac{1}{2}\sum_{i\in V}(x_i-y_i)^2\Big\}\\ &= \frac{1}{Z}\exp\Big\{-\frac{\gamma}{2}\langle x,\mathcal{L}_G\,x\rangle-\frac{1}{2}\,\|x-y\|_2^2\Big\}\end{aligned}$$

- if we compute $\mathbb{E}_\mu\{x\}$ by taking the derivative and setting it to zero, we get

$$-\gamma\,\mathcal{L}_G\,x-x+y=0$$

and

$$\begin{aligned}\mathbb{E}_\mu\{x\} &= (I+\gamma\mathcal{L}_G)^{-1}y\\ &\overset{\gamma\to\infty}{\to} \mathbf{1}\mathbf{1}^T y=\overline{y}\,\mathbf{1}\end{aligned}$$

- can we compute this using Gaussian belief propagation (in a distributed fashion)?

Consensus Propagation

Belief Propagation

let $\Delta_i = \text{deg}_G(i)$, then

$$\mu_y(x) \;=\; \frac{1}{Z}\,\exp\left\{\gamma \sum_{(i,j)\in E} x_i x_j - \frac{1}{2}\sum_{i\in V}(1+\gamma\Delta_i)x_i^2 + \sum_{i\in V} y_i x_i\right\},$$

$$J_{ii} = 1+\gamma\Delta_i, \qquad J_{ij} = -\gamma\,, \qquad h_i = y_i\,.$$

$$J_{i\to j}^{(t+1)} \;=\; 1+\gamma\Delta_i - \sum_{k\in\partial i\setminus j}\frac{\gamma^2}{J_{k\to i}^{(t)}}\,,$$

$$h_{i\to j}^{(t+1)} \;=\; y_i + \sum_{k\in\partial i\setminus j}\frac{\gamma}{J_{k\to i}^{(t)}}\,h_{k\to i}^{(t)}\,.$$

Consensus Propagation

Redefine

$$
\begin{aligned}
K_{i\to j}^{(t)} &= -\gamma + J_{i\to j}^{(t)} \\
m_{i\to j}^{(t)} &= \frac{h_{i\to j}^{(t)}}{K_{i\to j}^{(t)}}
\end{aligned}
$$

$$
\begin{aligned}
K_{i\to j}^{(t+1)} &= 1 + \sum_{k\in\partial i\setminus j} \frac{K_{k\to i}^{(t)}}{1+\gamma^{-1}K_{k\to i}^{(t)}}\,, \\
m_{i\to j}^{(t+1)} &= \frac{y_i + \sum_{k\in\partial i\setminus j} \frac{K_{k\to i}^{(t)}}{1+\gamma^{-1}K_{k\to i}^{(t)}}\, m_{k\to i}^{(t)}}{1+\sum_{k\in\partial i\setminus j} \frac{K_{k\to i}^{(t)}}{1+\gamma^{-1}K_{k\to i}^{(t)}}}
\end{aligned}
$$

Interpretation? $K_{i\to j}$ as size of population and $m_{i\to j}$ as population mean

From a quadratic MRF to a covariance matrix

- given a quadratic Markov random field parametrized by h and J

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

  it is a valid Gaussian distribution only if J is positive definite
- a symmetric matrix J is positive definite if and only if
  0. $x^T J x > 0$ for all $x \in \mathbb{R}^n$
  1. all eigen values are positive

  proof $\Rightarrow$ $\lambda = \frac{x^T J x}{\|x\|^2} > 0$

  proof $\Leftarrow$ $x^T J x = x^T U D U^T x = \tilde{x}^T D \tilde{x} = \sum_i \tilde{x}_i^2 D_{ii} > 0$

  2 has a Cholesky decomposition: there exists a (unique) lower triangular matrix L with strictly positive diagonal entries such that $J = L^T L$

  proof $\Rightarrow$

  proof $\Leftarrow$ $x^T J x = x^T L^T L x = \|Lx\|^2 > 0$

  3 satisfies Sylvester's criterion: leading principal minors are all positive (a kth leading principal minor of a matrix J is the determinant of its upper left k by k sub-matrix)
- there is no simple way to check if J is positive definite

Sufficient conditions

- toy example on 2×2 symmetric matrices

$$\begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

- what is the sufficient and necessary condition for positive definiteness?

Sufficient conditions

- **sufficient condition 1.** J is positive definite if it is **diagonally dominant**, i.e.,

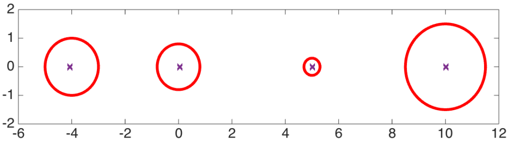
$$\sum_{j \neq i} |J_{ij}| < J_{ii}$$

  - proof by Gershgorin's circle theorem

- **[Gershgorin's circle theorem]** every eigenvalue of $A \in \mathbb{R}^{n \times n}$ lies within at least one of the Gershgorin discs, defined for each $i \in [n]$ as

$$D_i \;\equiv\; \Big\{\, x \in \mathbb{R} \,\Big|\, |x - J_{ii}| \leq \sum_{j \neq i} |J_{ij}| \,\Big\}$$

$$\begin{bmatrix} 10 & 0.5 & 0.5 & 0.5 \\ 0 & 5 & 0.1 & 0.2 \\ 0.3 & 0 & 0 & 0.5 \\ 0 & 0.5 & 0.5 & -4 \end{bmatrix}$$



- **Corollary.** diagonally dominant matrices are positive definite

- **proof (for a amore general complex valued matrix A).** consider an eigen value $\lambda \in \mathbb{C}$ and an eigen vector $x \in \mathbb{C}^n$ such that

$$Ax = x\lambda$$

let i denote the index of the maximum magnitude entry of x such that $|x_i| \geq |x_j|$ for all $j \neq i$, then it follows that

$$\sum_{j \in [n]} A_{ij} x_j = \lambda x_i$$

and

$$\sum_{j \neq i} A_{ij} x_j = (\lambda - A_{ii}) x_i$$

dividing both sides by x_i gives

$$|\lambda - A_{ii}| \; = \; \Big| \frac{\sum_{j \neq i} A_{ij} x_j}{x_i} \Big| \; \leq \; \sum_{j \neq i} \Big| \frac{A_{ij} x_j}{x_i} \Big| \; \leq \; \sum_{j \neq i} |A_{ij}| = R_i$$

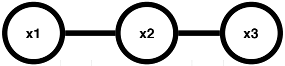
- when there is an overlap, it is possible to have an empty disc, for example $\begin{bmatrix} 0 & 1 \\ 4 & 0 \end{bmatrix}$ and $\begin{bmatrix} 1 & -2 \\ 1 & -1 \end{bmatrix}$ have eigen values $\{-2, 2\}$ and $\{i, -i\}$
- **theorem.** if a union of k discs is disjoint from the union on the rest of $n-k$ discs, then the former union contains k eigen values and the latter $n-k$.
- **proof.** let
$$B(t) \triangleq (1-t)\mathrm{diag}(A) + t(A)$$
for $t \in [0, 1]$, and note that eigen values of $B(t)$ are continuous in t. $B(0)$ has eigen values at the center of the discs and the eigen values $\{\lambda(t)_i\}_{i\in[n]}$ of $B(t)$ move from this center as t increases, but by continuity the k eigen values of the first union of discs can not escape the expanding union of discs
- counter example? computational complexity?

Sufficient conditions

- **sufficient condition 2.** J is positive definite if it is **pairwise normalizable**, i.e., if there exists compatibility functions ψ_i's and ψ_{ij}'s such that $J_{ii} = \sum_{j \in \partial i} a_{ij}^{(ii)}$,

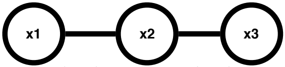
$$\begin{aligned} -\log \psi_i(x_i) &= x_i^T a_i x_i + b_i^T x_i \\ -\log \psi_{ij}(x_i, x_j) &= x_i^T a_{ij}^{(ii)} x_i + x_j^T a_{ij}^{(jj)} x_j + x_i^T a_{ij}^{(ij)} x_j \end{aligned}$$

  we have $a_i > 0$ for all i and $\begin{bmatrix} a_{ij}^{(ii)} & \frac{1}{2} a_{ij}^{(ij)} \\ \frac{1}{2} a_{ij}^{(ij)} & a_{ij}^{(jj)} \end{bmatrix}$ is PSD for all 2×2 minors
  - follows from $\int f(x) g(x)\, dx \le \int |f(x)|\, dx \int |g(x)|\, dx$



	x1	x2
x1	2	-1
x2	-1	3

	x2	x3
x2	1	2
x3	2	3



	x1	x2
x1	2	-1
x2	-1	2

	x2	x3
x2	2	2
x3	2	3

- counter example? computational complexity?

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

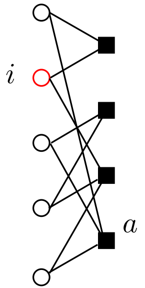
$$\hat{m}_i^{(\ell)} \equiv (\hat{J}_i^{(\ell)})^{-1}\hat{h}_i^{(\ell)}$$

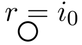
  where $\hat{m}_i^{(\ell)}$ = belief propagation expectation after ℓ iterations
  $\hat{J}_i^{(\ell)}$ = belief propagation information matrix after ℓ iterations
  $\hat{h}_i^{(\ell)}$ = belief propagation precision after ℓ iterations and if
  $\hat{m}_i^{(\infty)} \triangleq \lim_{\ell\to\infty} \hat{m}_i^{\infty}$ exists, then

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

A detour: Computation tree

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

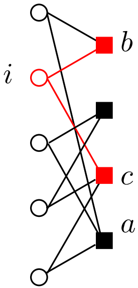


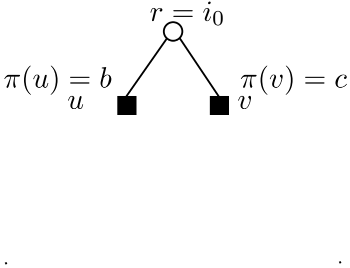


- $i, j, k, \ldots, a, b, \ldots$ for nodes in G and $r, s, t, \ldots$ for nodes in $\mathsf{CT}_G(i;\ell)$
- potentials ψ_i and ψ_{ij} are copied to $\mathsf{CT}_G(i;\ell)$
- each node (edge) in G corresponds to multiple nodes (edges) in $\mathsf{CT}_G(i;\ell)$.
- natural projection $\pi : \mathsf{CT}_G(i;\ell)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

A detour: Computation tree

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

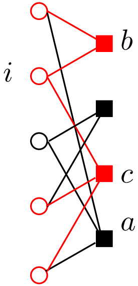


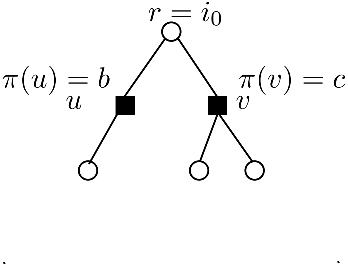


- $i, j, k, \ldots, a, b, \ldots$ for nodes in G and $r, s, t, \ldots$ for nodes in $\mathsf{CT}_G(i;\ell)$
- potentials ψ_i and ψ_{ij} are copied to $\mathsf{CT}_G(i;\ell)$
- each node (edge) in G corresponds to multiple nodes (edges) in $\mathsf{CT}_G(i;\ell)$.
- natural projection $\pi : \mathsf{CT}_G(i;\ell)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

A detour: Computation tree

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

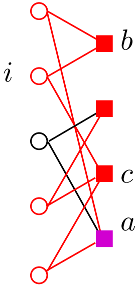


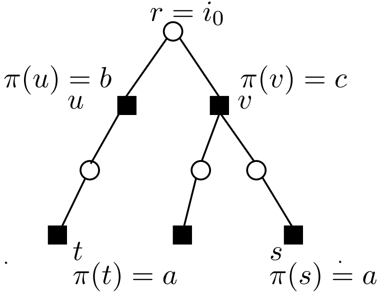


- $i, j, k, \ldots, a, b, \ldots$ for nodes in G and $r, s, t, \ldots$ for nodes in $\mathsf{CT}_G(i;\ell)$
- potentials ψ_i and ψ_{ij} are copied to $\mathsf{CT}_G(i;\ell)$
- each node (edge) in G corresponds to multiple nodes (edges) in $\mathsf{CT}_G(i;\ell)$.
- natural projection $\pi : \mathsf{CT}_G(i;\ell)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

A detour: Computation tree

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





- $i, j, k, \ldots, a, b, \ldots$ for nodes in G and $r, s, t, \ldots$ for nodes in $\mathsf{CT}_G(i;\ell)$
- potentials ψ_i and ψ_{ij} are copied to $\mathsf{CT}_G(i;\ell)$
- each node (edge) in G corresponds to multiple nodes (edges) in $\mathsf{CT}_G(i;\ell)$.
- natural projection $\pi : \mathsf{CT}_G(i;\ell) \to G$, 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 $\mathsf{CT}_G(i;\ell)$
  - **proof of claim 1.** by induction over ℓ.
  - idea: BP 'does not know' whether it is operating on G or on $\mathsf{CT}_G(i;\ell)$

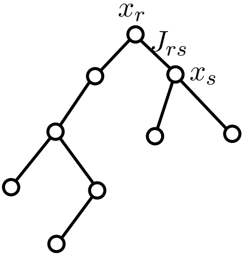
- recall that for Gaussians, mode of $-\frac{1}{2}x^T Jx + h^T x$ is the mean m, hence

$$Jm = h$$

  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 $\mathsf{CT}_G(i;\ell)$



the estimated mean $\hat{m}^{(\ell)}$ is exact on a tree. Precisely, since the width of the tree is at most 2ℓ, the BP updates on $\mathsf{CT}_G(i;\ell)$ 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 $\mathsf{CT}_G(i;\ell)$.

- note that on the computation tree $\mathrm{CT}_G(i,;\ell)$, $\hat{m}_r^{(t)} = \hat{m}_r^{(\ell)}$ for $t \geq \ell$ since the root r is at most distance ℓ 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

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

if the BP fixed point converges then

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

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

$$\begin{aligned}
\lim_{\ell\to\infty} \hat{m}_r^{(\ell)} &= \lim_{\ell\to\infty} \hat{m}_{\pi(r)}^{(\ell)} && \text{by Claim 1.} \\
&= \hat{m}_{\pi(r)}^{(\infty)} && \text{by the convergence assumption}
\end{aligned}$$

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

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

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 $\ell \to \infty$,

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

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