Convergence Analysis of Belief Propagation for Pairwise Linear Gaussian Models

Jian Du[†]

Joint Work with Shaodan Ma[◊], Yik-Chung Wu[ǂ], Soummya Kar[†], and José M. F. Moura[†]

Carnegie Mellon University[†]
University of Macau[◊]
The University of Hong Kong[ǂ]

Acknowledgements: NSF under grants CCF-1513936.

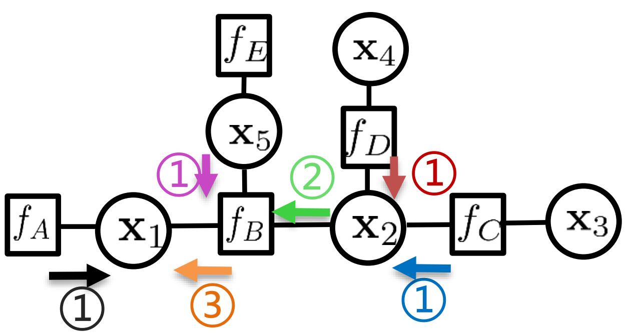
Electrical & Computer ENGINEERING

Belief Propagation (BP) on Trees

- **Computing *marginal distributions/modes* efficiently by exploiting the distributive law.**

$$q(\mathbf{x}_1) \propto \int \dots \int f_A(\mathbf{x}_1) f_B(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_5) f_C(\mathbf{x}_2, \mathbf{x}_3) f_D(\mathbf{x}_2, \mathbf{x}_4) \mathrm{d}\mathbf{x}_2 \mathrm{d}\mathbf{x}_3 \mathrm{d}\mathbf{x}_4 \mathrm{d}\mathbf{x}_5 \mathrm{d}\mathbf{x}_6$$

$$= \underbrace{f_A(\mathbf{x}_1)}_{①: m_{f_A \to 1}(\mathbf{x}_1)} \underbrace{\int_{\mathbf{x}_2} \int_{\mathbf{x}_5} f_B(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_5) \underbrace{f_E(\mathbf{x}_5)}_{①: m_{5 \to f_B}(\mathbf{x}_5)} \underbrace{\underbrace{\left(\int_{\mathbf{x}_4} f_C(\mathbf{x}_2, \mathbf{x}_3) \mathrm{d}\mathbf{x}_3 \right)}_{①: m_{f_C \to 2}(\mathbf{x}_2)} \underbrace{\left(\int_{\mathbf{x}_4} f_D(\mathbf{x}_2, \mathbf{x}_4) \mathrm{d}\mathbf{x}_4 \right)}_{①: m_{f_D \to 2}(\mathbf{x}_2)}}_{②: m_{2 \to f_B}(\mathbf{x}_2)} \mathrm{d}\mathbf{x}_5 \mathrm{d}\mathbf{x}_2}_{③: m_{f_B \to 1}(\mathbf{x}_1)}$$



- **Message from *variable* to *factor*:**

  ① ②

- **Message from *factor* to *variable*:**

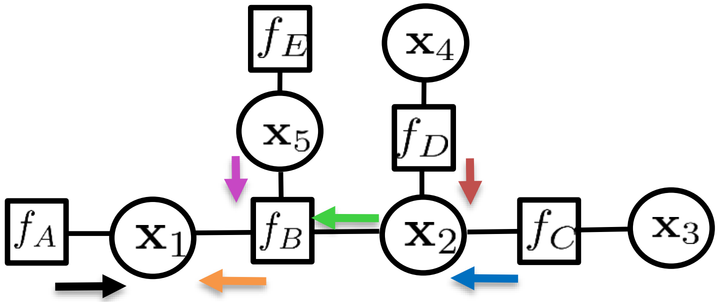
  ① ① ① ③

- **Marginal distribution:**

  $q(\mathbf{x}_1) \propto ① \times ③$

Message Definition

- **Message from *variable* to *factor*:**



$$m_{j\to f_n}(\mathbf{x}_j) := p(\mathbf{x}_j) \prod_{f_k \in \mathcal{B}(j)\setminus f_n} m_{f_k \to j}(\mathbf{x}_j)$$

$$\text{e.g., } m_{2\to f_B}(\mathbf{x}_2) = m_{f_D\to 2}(\mathbf{x}_2) m_{f_C\to 2}(\mathbf{x}_2)$$

- **Message from *factor* to *variable*:**

$$m_{f_n\to i}(\mathbf{x}_i) := \int \cdots \int f_n \prod_{j\in\mathcal{B}(f_n)\setminus i} m_{j\to f_n}(\mathbf{x}_j) \mathrm{d}\{\mathbf{x}_j\}_{j\in(f_n)\setminus i}$$

$$\text{e.g., } m_{f_B\to 1}(\mathbf{x}_1) = \int\int f_B m_{2\to f_B}(\mathbf{x}_2) m_{5\to f_B}(\mathbf{x}_5) \mathrm{d}\mathbf{x}_2 \mathrm{d}\mathbf{x}_5$$

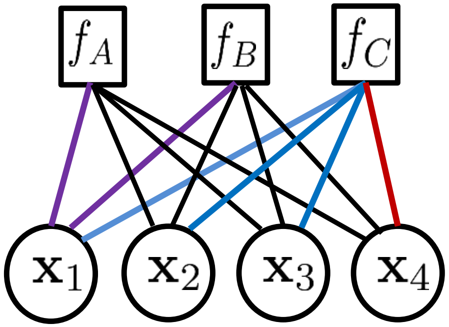
- **Marginal distribution**

$$b(\mathbf{x}_i) \propto \prod\nolimits_{f_n\in\mathcal{B}(\mathbf{x}_i)} m_{f_n\to i}(\mathbf{x}_i)$$

$$\text{e.g., } b(\mathbf{x}_1) \propto m_{f_A\to 1}(\mathbf{x}_1) m_{f_B\to 1}(\mathbf{x}_1)$$

BP on Graph with Loops

- **Distributive law may NOT be exploited on graph with loops**



$$q(\mathbf{x}_1) \propto \int \dots \int f_A(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) f_B(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) \times f_C(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) \mathrm{d}\mathbf{x}_2 \mathrm{d}\mathbf{x}_3 \mathrm{d}\mathbf{x}_4$$

- **Use same message updating rule *in parallel***
  - **Message from *variable* to *factor*:**

$$m^{(l)}_{j \to f_n}(\mathbf{x}_j) = p(\mathbf{x}_j) \prod_{f_k \in \mathcal{B}(j) \setminus f_n} m^{(l-1)}_{f_k \to j}(\mathbf{x}_j)$$

  - **Message from *factor* to *variable*:**

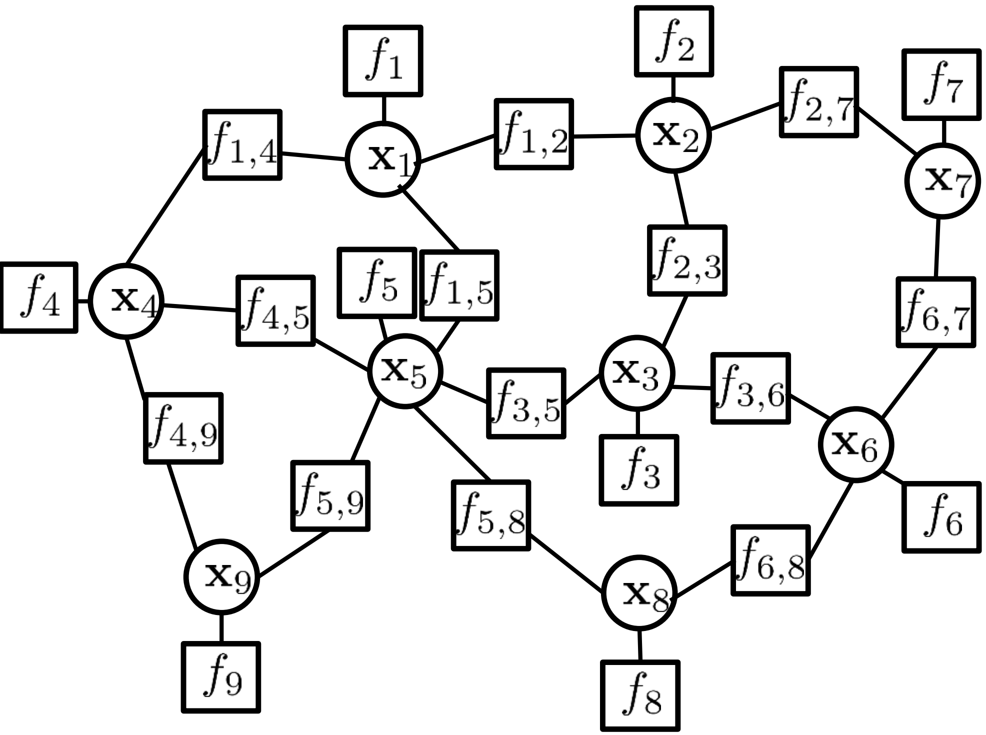
$$m^{(l)}_{f_n \to i}(\mathbf{x}_i) = \int \cdots \int f_n \prod_{j \in \mathcal{B}(f_n) \setminus i} m^{(l)}_{j \to f_n}(\mathbf{x}_j) \mathrm{d}\{\mathbf{x}_j\}_{j \in (f_n) \setminus i}$$

  - **Approximate marginal distribution:**

$$b^{(l)}(\mathbf{x}_i) \propto \prod_{f_n \in \mathcal{B}(\mathbf{x}_i)} m^{(l-1)}_{f_n \to i}(\mathbf{x}_i)$$

Will $b^{(l)}$ converge?
Where will it converge?
Convergence rate?

Pairwise Models: GMRF & Linear Gaussian Model



1) GMRF

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

$$f_i(x_i) = \exp\left(-\frac{1}{2}J_{i,i}x_i^2 + h_i x_i\right)$$

$$f_{i,j}(x_i, x_j) = \exp\left(-x_i J_{i,j} x_j\right)$$

2) Pairwise Linear Gaussian Model

$$f_i(\mathbf{x}_j) \sim \mathcal{N}(\mathbf{x}_j | \mathbf{0}, \mathbf{W}_j)$$

$$f_{i,j}(\mathbf{x}_i, \mathbf{x}_j) = \mathcal{N}\left(\mathbf{y}_{i,j} | \mathbf{A}_{j,i}\mathbf{x}_i + \mathbf{A}_{i,j}\mathbf{x}_j, \mathbf{R}_{i,j}\right)$$

Electrical & Computer ENGINEERING

BP on GMRF

- **A joint Gaussian distribution function can always be written as:**

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

$$= \prod_{i\in\mathcal{V}} \underbrace{\exp\left(-\frac{1}{2}J_{i,i}x_i^2 1 + h_i x_i\right)}_{\triangleq f_i(x_i)} \prod_{(i,j)\in\mathcal{E}_{\text{MRF}}} \underbrace{\exp\left(-x_i J_{i,j} x_j\right)}_{\triangleq f_{i,j}(x_{i,j})}$$

"Pairwise Model"

- **A sufficient convergence condition, given by the spectrum radius is obtained:**

$$\rho(|\mathbf{I} - \mathbf{J}|) < 1 \Leftrightarrow \text{Walk-summable (BP converges on GMRF)}$$

Malioutov, Johnson and Willksy, Journal of Machine Learning Research 2006.

BP in Linear Gaussian Model

- **In a general connected network, the local observations at every node $n \in \mathcal{V}$ are in the form of**

$$\underbrace{\mathbf{y}_{i,j}}_{\text{local observation}} = \mathbf{A}_{j,i}\mathbf{x}_i + \underbrace{\mathbf{A}_{i,j}}_{\text{known coefficient}}\mathbf{x}_j + \mathbf{z}_{i,j} \quad \begin{cases} \mathbf{z}_{i,j} \sim \mathcal{N}(\mathbf{z}_{i,j}|\mathbf{0}, \mathbf{R}_{i,j}) \\ p(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{x}_i|\mathbf{0}, \mathbf{W}_i) \end{cases}$$

- **The joint posterior distribution of $[\mathbf{x}_1, \ldots \mathbf{x}_{|\mathcal{V}|}]^T$**

$$p(\mathbf{x})\, p(\mathbf{y}|\mathbf{x}) = \prod_{i\in\mathcal{V}} \underbrace{p(\mathbf{x}_i)}_{\triangleq f_i} \prod_{i\in\mathcal{V}} \underbrace{p(\mathbf{y}_{i,j}|\mathbf{x}_i, \mathbf{x}_j, \{i,j\} \in \mathcal{E}_{\text{Net}})}_{\triangleq f_{i,j}}.$$

- **Applications for distributed estimation:**

  -distributed power state estimation,

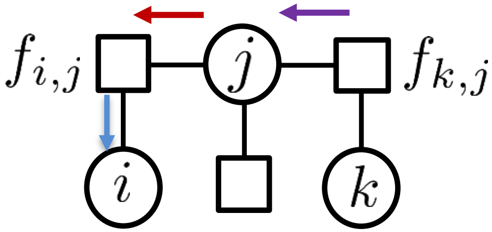
  -distributed localization/synchronization, etc.

Du, Ma, Wu, Kar, and Moura, arXiv preprint arXiv:1611.02010 submitted to Journal of Machine Learning Research.

BP Updating Equation in Linear Gaussian Model

- The general expression for message updating from *variable node* to *factor node* is

$$\left[\mathbf{C}^{(\ell)}_{j\to f_{i,j}}\right]^{-1} = \mathbf{W}_j^{-1} + \sum_{f_{k,j}\in\mathcal{B}(j)\setminus f_{i,j}} \left[\mathbf{C}^{(\ell-1)}_{f_{k,j}\to j}\right]^{-1}$$



$$\mathbf{v}^{(\ell)}_{j\to f_{i,j}} = \mathbf{C}^{(\ell)}_{j\to f_{i,j}}\left[\sum_{f_{k,j}\in\mathcal{B}(j)\setminus f_{i,j}} \left[\mathbf{C}^{(\ell-1)}_{f_{k,j}\to j}\right]^{-1}\mathbf{v}^{(\ell-1)}_{f_{k,j}\to j}\right]$$

- The message from *factor node* to *variable node* is

$$\left[\mathbf{C}^{(\ell)}_{f_{i,j}\to i}\right]^{-1} = \mathbf{A}^T_{j,i}\left[\mathbf{R}_{i,j} + \mathbf{A}_{i,j}\mathbf{C}^{(\ell)}_{j\to f_{i,j}}\mathbf{A}^T_{i,j}\right]^{-1}\mathbf{A}_{j,i}.$$

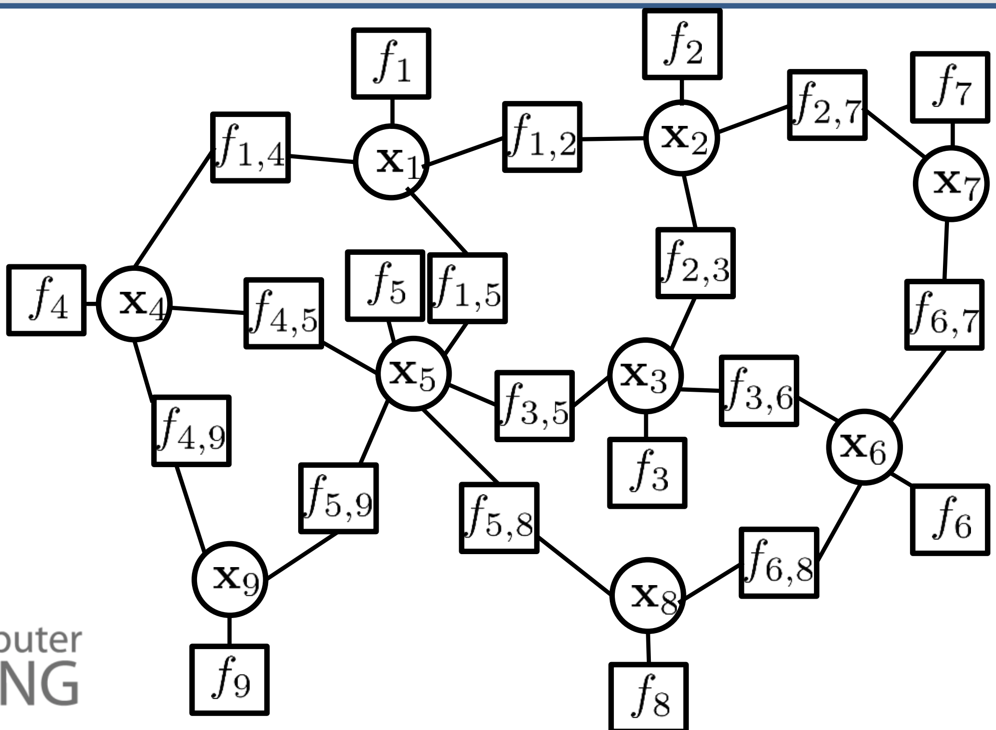
$$\mathbf{v}^{(\ell)}_{f_{i,j}\to i} = \underbrace{\mathbf{A}^T_{j,i}\left[\mathbf{R}_{i,j} + \mathbf{A}_{i,j}\mathbf{C}^{(\ell)}_{j\to f_{i,j}}\mathbf{A}^T_{i,j}\right]^{-1}}_{\text{like a Kalman Gain}}\Big(\underbrace{\mathbf{y}_{i,j}}_{\text{local innovation}} - \underbrace{\mathbf{A}_{i,j}\mathbf{v}^{(\ell)}_{j\to f_{i,j}}}_{\text{cooperative innovation}}\Big)$$

like a Kalman Gain — local innovation — cooperative innovation

Convergence Property

$$[\mathbf{C}^{(\ell)}_{f_{i,j}\to i}]^{-1} = \mathbf{A}^T_{j,i}\Big[\mathbf{R}_{i,j} + \mathbf{A}_{i,j}\big[\mathbf{W}^{-1}_j + \sum_{f_{k,j}\in\mathcal{B}(j)\backslash f_{i,j}} [\mathbf{C}^{(\ell-1)}_{f_{k,j}\to j}]^{-1}\big]^{-1}\mathbf{A}^T_{i,j}\Big]^{-1}\mathbf{A}_{j,i}$$

Theorem 1. The matrix sequence $\left\{\left[\mathbf{C}^{(0)}_{f_{i,j}\to i}\right]^{-1}\right\}_{l=0,1,\ldots}$ converges to a unique positive definite matrix for any initial covariance matrix $\left[\mathbf{C}^{(0)}_{f_{i,j}\to i}\right]^{-1} \succeq \mathbf{0}$ for all $i,j\in\mathcal{V}$



Convergence Rate

- **Convergence rate with respect to part metric:**

Part (Birkhoff) Metric: For arbitrary square matrices $\mathbf{X}$ and $\mathbf{Y}$ with the same dimension, if there exists $\alpha \geq 1$ such that $\alpha\mathbf{X} \succeq \mathbf{Y} \succeq \alpha^{-1}\mathbf{X}$, $\mathbf{X}$ and $\mathbf{Y}$ are called the parts, and $\mathrm{d}\left(\mathbf{X}, \mathbf{Y}\right) \triangleq \inf\left\{\log\alpha : \alpha\mathbf{X} \succeq \mathbf{Y} \succeq \alpha^{-1}\mathbf{X}, \alpha \geq 1\right\}$ defines a metric called the part metric.

$$\mathrm{d}\left(\mathbf{C}^{(\ell)}, \mathbf{C}^{*}\right) < \beta^{\ell}\mathrm{d}\left(\mathbf{C}^{(0)}, \mathbf{C}^{*}\right), \quad 0 < \beta < 1.$$

Theorem 2. With the initial covariance matrix set to be an arbitrary p.s.d. matrix, i.e., $\left[\mathbf{C}^{(0)}_{f_{i,j}\to i}\right]^{-1} \succeq \mathbf{0}$, the sequence $\left\{\left[\mathbf{C}^{(0)}_{f_{i,j}\to i}\right]^{-1}\right\}_{l=0,1,\ldots}$ converges at a geometric rate with respect to the part metric.

Convergence Rate

- **Convergence rate with respect to part metric:**

$$\mathrm{d}\left(\mathbf{C}^{(\ell)}, \mathbf{C}^{*}\right) < \beta^{\ell}\mathrm{d}\left(\mathbf{C}^{(0)}, \mathbf{C}^{*}\right), \quad 0 < \beta < 1.$$

- **From Part metric to monotone norm**

$$||\mathbf{J}^{(\ell)} - \mathbf{J}^{*}|| \le \left(2\exp\left\{\mathrm{d}\left(\mathbf{J}^{(\ell)}, \mathbf{J}^{*}\right)\right\} - \exp\left\{-\mathrm{d}\left(\mathbf{J}^{(\ell)}, \mathbf{J}^{*}\right)\right\} - 1\right)\min\left\{||\mathbf{J}^{(\ell)}||, ||\mathbf{J}^{*}||\right\}.$$

$$||\mathbf{J}^{(\ell)} - \mathbf{J}^{*}|| < 2\zeta\exp\left\{c^{\ell}d_0\right\}, \quad c < 1.$$

Theorem 3. With the initial covariance matrix set to be an arbitrary p.s.d. matrix, i.e., $\left[\mathbf{C}^{(0)}_{f_n \to i}\right]^{-1} \succeq \mathbf{0}$, the sequence $\left\{\mathbf{C}^{(\ell)}\right\}_{l=0,1,\ldots}$ converges at a **double exponential rate** in terms of the monotone norm.

Convergence Property

- **Updating equation**

$$\mathbf{v}^{(\ell)}_{j\to f_{i,j}} = \mathbf{b}_{j\to f_{i,j}} - \mathbf{C}^{*}_{j\to f_{i,j}} \sum_{f_{k,j}\in\mathcal{B}(j)\backslash f_{i,j}} \mathbf{C}^{*}_{f_{kj}\to j}\mathbf{M}_{k,j}\mathbf{A}_{j,k}\mathbf{v}^{(\ell)}_{k\to f_{k,j}},$$

$$\mathbf{v}^{(\ell)} = -\mathbf{Q}\mathbf{v}^{(\ell-1)} + \mathbf{b}.$$

Theorem 4. The belief mean converges to the optimal value if and only if $\rho(\mathbf{Q}) < 1$. The matrix $\mathbf{Q}$ satisfies $\mathbf{v}^{(\ell)} = \mathbf{Q}\mathbf{v}^{(\ell-1)} + \mathbf{b}$ with $\mathbf{v}^{(\ell)}$ being a vector containing all the $\mathbf{v}^{(\ell)}_{j\to f_n}$.

- **A *distributed* sufficient convergence condition**

Theorem 5. The belief mean converges to the optimal value if the spectrum radious of block diagonal (each block's dim. equals the corresponding variable's dim.) of $\mathbf{Q}\mathbf{Q}^T$ is smaller than 1.

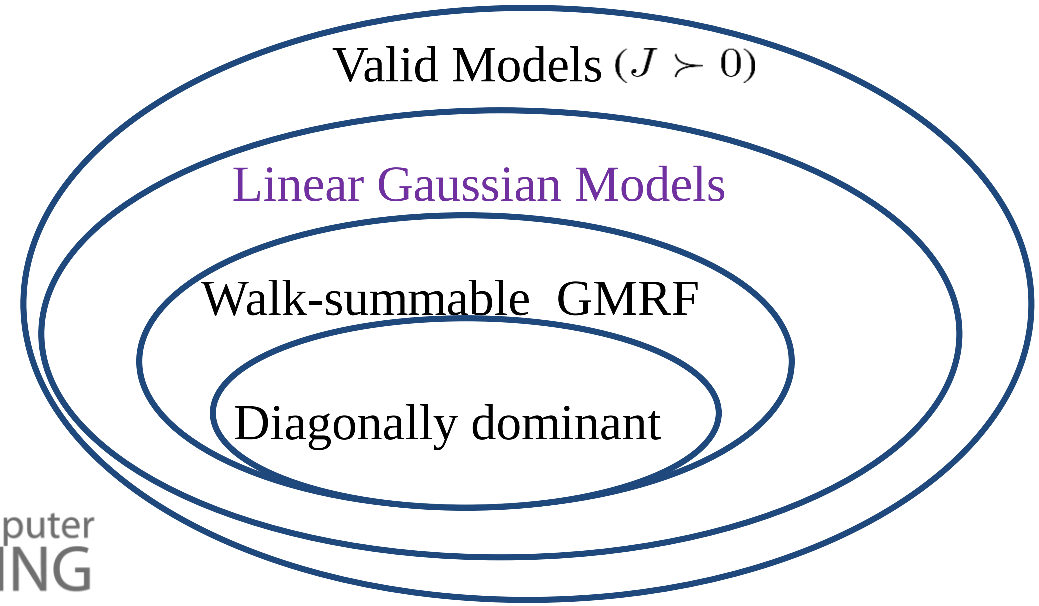
LGM subsumes Walk-Summable GMRF (I)

- **A joint Gaussian distribution function can always be written as:**

$$p(\mathbf{x}) \propto \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{J}\mathbf{x} + \mathbf{h}^T\mathbf{x}\right) = \prod_{i\in\mathcal{V}} \underbrace{\exp\left(-\frac{1}{2}J_{i,i}x_i^2 1 + h_i x_i\right)}_{\triangleq f_i(x_i)} \prod_{(i,j)\in\mathcal{E}_{\text{MRF}}} \underbrace{\exp\left(-x_i J_{i,j} x_j\right)}_{\triangleq f_{i,j}(x_{i,j})}.$$

$\rho(|\mathbf{I}-\mathbf{J}|) < 1 \Leftrightarrow$ Walk-summable (BP converges on GMRF)

- **Converged linear Gaussian model subsumes walk-summable GMRF .**



LGM subsumes Walk-Summable GMRF (II)

- **Distributive law may NOT be exploited on graph with loops**

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

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

$$= \exp\left\{-\frac{1}{2}\left(\mathbf{V}^T\mathbf{x}\right)^T\left(\mathbf{V}^T\mathbf{x}\right) - \frac{1}{2}\left(\omega\mathbf{x}^T\mathbf{x} - 2\mathbf{h}^T\mathbf{x}\right)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\sum_{n=1}^{M}\left(V_{n,n_i}x_{n_i} + V_{n,n_j}x_{n_j}\right)^2 - \frac{1}{2}\sum_{n=1}^{M}\omega(x_n - \frac{h_n}{\omega})^2\right\}.$$

- Let $0 < \omega < 1$ and ω is smaller than the minimum eigenvalue of $\mathbf{I} - |\mathbf{R}|$.

$$q(\mathbf{x}) \propto \prod_{n=1}^{M}\underbrace{\mathcal{N}(x_n|\frac{1}{\omega}h_n, \frac{1}{\omega})}_{p(x_n)}\prod_{n=1}^{M}\underbrace{\mathcal{N}(0|V_{n,n_i}x_{n_i} + V_{n,n_j}x_{n_j}, 1)}_{f_{n_i,n_j}}.$$

Electrical & Computer ENGINEERING

Moallemi and Roy, TIT 2009

LGM subsumes Walk-Summable GMRF (III)

- **H-Matrices**

Definition H-Matrices (Boman et al. 2015): A matrix $\mathbf{X}$ is an H-matrix if all eigenvalues of the matrix $H(\mathbf{X})$, where $[H(\mathbf{X})]_{i,i} = |\mathbf{X}_{i,i}|$, and $[H(\mathbf{X})]_{i,j} = -|\mathbf{X}_{i,j}|$ have positive real parts.

Factor width at most 2 factorization (Boman et al. 2015): A symmetric H-matrix $\mathbf{X}$ that has non-negative diagonals can always be factorized as $\mathbf{X} = \mathbf{V}\mathbf{V}^T$, where $\mathbf{V}$ is a real matrix with each column of $\mathbf{V}$ containing at most 2 non-zeros.

$(1-\omega)\mathbf{I} - \mathbf{R}$ is an H matrix.

$(1-\omega)\mathbf{I} - \mathbf{R} = \mathbf{J} - \omega\mathbf{I} = \mathbf{V}\mathbf{V}^T$, where each column of $\mathbf{V}$ contains at most 2 non-zeros.

Conclusion

- **Existing convergence analysis of GMRF can not be used for distributed inference in linear Gaussian models.**
- **Belief covariance of BP in pairwise linear Gaussian models**
  - **converges to a unique positive definite matrix for arbitrary positive definite initial value.**
  - **converges at a doubly exponential rate.**
- **Necessary and sufficient convergence condition of belief mean**
- **The convergence condition of pairwise linear Gaussian models subsumes walk-summable GMRF.**

Full Version: Du, Ma, Wu, Kar, and Moura, arXiv preprint arXiv:1611.02010

Thank you!

$$\mathbf{J} = \begin{bmatrix} 1 & \frac{1}{3\sqrt{2}} & \frac{1}{\sqrt{3}} & \frac{\sqrt{2}}{3} \\ \frac{1}{3\sqrt{2}} & 1 & 0 & \frac{1}{3} \\ \frac{1}{\sqrt{3}} & 0 & 1 & \frac{1}{\sqrt{6}} \\ \frac{\sqrt{2}}{3} & \frac{1}{3} & \frac{1}{\sqrt{6}} & 1 \end{bmatrix}.$$

The eigenvalues of $\mathbf{I} - |\mathbf{R}|$ to 4 decimal places are -0.0754, 0.9712, 1.4780, and 1.6262. According to the walk-summable definition in, it is non walk-summable and the convergence condition is inconclusive as to whether Gaussian BP converges.

References

- J. Du, S. Ma, Y.C. Wu, S. Kar, and J.M.F. Moura, "Convergence analysis of distributed inference with vector-valued Gaussian belief propagation". https://arxiv.org/abs/1611.02010
- J. Du, S. Kar, and J.M.F. Moura "Distributed convergence verification for Gaussian belief propagation," in Asilomar Conference on Signals, Systems, and Computers, Oct. 29 - Nov. 1, 2017, Asilomar, Pacific Grove, CA.
- J. Du and Y.C. Wu, "Network-Wide Distributed Carrier Frequency Offsets Estimation and Compensation," IEEE Transactions on Signal Processing, vol. 61, no. 23, pp. 5868-5877, Dec. 2013.
- J. Du and Y.C. Wu, "Distributed Clock Skew and Offset Estimation in Wireless Sensor Networks: Asynchronous Algorithm and Convergence Analysis," IEEE Transactions on Wireless Communications, vol. 12, no. 11, pp. 5908-5917, Nov. 2013.
- D. M. Malioutov, J. K. Johnson, and A. S. Willsky. Walk-sums and belief propagation in Gaussian graphical models. Journal of Machine Learning Research, 7(2):2031-2064, 2006.
- C. C. Moallemi and B. Van Roy. Convergence of min-sum message passing for quadratic optimization. IEEE Trans. Information Theory, 55(5):2413{2423, 2009