Message-Passing Algorithms for GMRFs and Non-Linear Optimization

Jason Johnson*

Joint Work with Dmitry Malioutov, Venkat Chandrasekaran and Alan Willsky

Stochastic Systems Group, MIT

NIPS Workshop:
Approximate Bayesian Inference in Continuous Models
Whistler, British-Colombia, Canada

December 7, 2007

*jasonj@mit.edu, //ssg.mit.edu/group/jasonj.
Introduction

♠ Walk-Sum View of Inference in GMRFs.
   ◇ Describe inference as computing walk-sums.
   ◇ Convergence of iterative algorithms: Gaussian BP (GaBP), Embedded Tree Algorithm (ET).
   ◇ Convergent GaBP for non-WS models: diagonal loading with damped feedback loop.

♠ Iterative Lagrangian Relaxation (LR) Algorithm:
   ◇ Decompose model into tractable subgraphs.
   ◇ Relax constraints between replicas.
   ◇ Convergent Max-Sum diffusion algorithm.
   ◇ Related to recent developments for MAP Estimation in discrete MRFs (TRMP and LP approaches).

♠ Extensions to Non-Linear Estimation:
   ◇ Use BP/LR to implement Newton’s method for MRFs with convex potentials.
   ◇ Use Diagonal-Loading for non-convex potentials.
Gaussian MRFs

Multivariate Gaussian $x \sim N(\hat{x}, P)$:

$$p(x) \propto \exp\{-\frac{1}{2}(x - \hat{x})P^{-1}(x - \hat{x})\}$$

If Markov on $\mathcal{G}$, naturally represented in information form,

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

where $J = P^{-1}$ is a sparse matrix:

$$J_{i,j} \neq 0 \iff \{i, j\} \in \mathcal{G}.$$

Inference reduces to Gaussian elimination (GE):

$$\hat{J}_A = J_{A,A} - J_{A,B}(J_{B,B})^{-1}J_{B,A}$$

$$\hat{h}_A = h_A - J_{A,B}(J_{B,B})^{-1}h_B$$

If only $\hat{x}$ is desired, solve $J\hat{x} = h$ using GE and back-substitution.

Because of graphical fill, direct methods become impractical for many large graphs (cubic in ”tree-width” of $\mathcal{G}$).
Neumann Series and Walk-Sums

It will be convenient to rescale $x$ so that $J$ is unit diagonal. Then, $J = I - R$ where $R$ is matrix of edge-weights on $G$, given by partial correlation coefficients ($|r_{ij}| < 1$):

$$r_{i,j} = \rho(x_i, x_j | x_{V \setminus ij}) = -J_{ij}$$

Neumann series:

$$P = J^{-1} = (I - R)^{-1} = \sum_{\ell} R^{\ell}$$

converges if spectral radius $\varrho(R) < 1$.

Interpret $(R^{\ell})_{i,j}$ as *sum over walks* in $G$:

$$(R^{\ell})_{i,j} = \sum_{w:i \xrightarrow{\ell} j} \prod_{E \in w} r_E \stackrel{\Delta}{=} \sum_{w:i \xrightarrow{\ell} j} \phi(w)$$

where walks can visit nodes or cross edges multiple times and can also back-track.
Walk-Summable GMRFs

Suggests walk-sum interpretation of inference:

\[ P_{i,j} = \sum_{w: i \rightarrow j} \phi(w) \quad \text{and} \quad \hat{x}_j = \sum_{w: \ast \rightarrow j} h_\ast \phi(w) \]

where

\[ \phi(w) \triangleq \frac{\ell(w)}{\prod_{s=1}^{l(w)} r_{w(s-1),w(s)}} \]

For example:

\[
\begin{align*}
\hat{x}_1 &= h_1 + h_2 r_{12} + h_3 r_{31} + h_1 r_{12}^2 + h_2 r_{23} r_{31} + \ldots \\
P_{11} &= 1 + r_{12}^2 + r_{13}^2 + r_{12} r_{23} r_{31} + r_{13} r_{32} r_{21} + \ldots
\end{align*}
\]

If these unordered sums (over a countable infinite set) are well-defined (converge absolutely) then the model is walk-summable.
Equivalent Conditions:

- $\rho(|R|) < 1$ where $|R|_{ij} = |r_{ij}|$.

- $J' = I - |R| \succ 0$.

- Pairwise-Normalizable: $J = \sum_E J_E$, where $J_E \succ 0$.

- (Generalized) Diagonal Dominance: Exist diagonal $D > 0$ such that $J' = DJD$ is diagonally-dominant, $(J')_{ii} > \sum_j |J'_{ij}|$.

Includes valid $(J \succ 0)$ non-frustrated models, where every cycle has even number of negative edges (e.g., trees and attractive models).

Proofs use Perron-Frobenius theorem.
Walk-Sum View of GaBP*

On trees, BP has a simple walk-sum interpretation. Each message $i \rightarrow j$ is described by two walk-sums:

1. $\alpha_{ij}$: self-return walks at $i$ excluding $j$.
2. $\beta_{ij}$: $h$-rewighted walks to $i$ excluding $j$.

These walk-sums may be computed recursively, equivalent to GE/BP equations.

$$\alpha_{ij} = \left(1 - \sum_{k \neq j} r_{ki}^2 \alpha_{ki}\right)^{-1}$$

$$\beta_{ij} = \alpha_{ij} (h_i + \sum_{k \neq j} r_{ki} \beta_{ki})$$

Marginal moments:

$$P_{ii} = \left(1 - \sum_k r_{ki}^2 \alpha_{ki}\right)^{-1}$$

$$\hat{x}_i = P_{ii} (h_i + \sum_k r_{ki} \beta_{ki})$$

*Johnson, Malioutov, Willsky, '05; Malioutov et al '06.
Walk-Sums on the Computation Tree

In loopy graphs, BP computes walks-sums on the computation tree:

This implies:

1. Only "back-tracking" self-return walks are included, missing non-backtracking walks around loops.

2. All $h$-rewighted walks to $j$ are included.

Hence, BP converges to correct estimate $\hat{x}$ in WS-models. Variances converge to back-tracking walk-sums, which are incorrect.
(New) Convergent BP for Non-WS Models

Let $J = I - R$ and $M = J + \gamma I$. For $\gamma > \rho(|R|) - 1$, $M$ is WS and GaBP solves $Mx = h$.

This yields damped estimate. Add "feedback":

$$\hat{x}_{t+1} = (J + \gamma I)^{-1}(h + \gamma \hat{x}_t)$$

Converges to correct estimate for all $\gamma > 0$.

Rather than double-loop, put damped feedback directly into GaBP fixed-point equations:

$$\alpha_{ij} = (1 + \gamma - \sum_{k \neq j} r_{ki}^2 \alpha_{ki})^{-1}$$

$$\beta_{ij} = \alpha_{ij}(h_i + \sum_{k \neq j} r_{ki} \beta_{ki})$$

$$\hat{x}_i = (1 + \gamma - \sum_k r_{ki}^2 \alpha_{ki})^{-1}(h_i + \sum_k r_{ki} \beta_{ki})$$

$$h_i = (1 - \lambda)h_i + \lambda(h_i + \gamma \hat{x}_i)$$

Converges for sufficiently small $\lambda > 0$, e.g. $1/2$ usually works.
Walk-Sum View of ET*

Iterative solver using preconditioner $M_t$:

$$h_t = h - Jx_t$$
$$\hat{x}_{t+1} = \hat{x}_t + M_t^{-1}h_t$$

ET Algorithm uses embedded trees of $G$ and $O(n)$ estimation algorithms [Sudderth et al].

We show ET has a walk-sum interpretation:

$$W_{t+1}(k) = \{\text{walks to } k \text{ in } T\} \cup \cup_{i,j} W_t(i) \otimes \{(i,j) \notin T\} \otimes \{j \rightarrow k \text{ in } T\}$$

Converges to correct estimate in WS models.

*Chandrasekaran, Johnson, Willsky, '07.
Adaptive ET

Minimizes upper-bound on estimation error leading to faster convergence.

\[
\max_{\mathcal{T} \subset \mathcal{G}} \sum_{\{i,j\} \in \mathcal{T}} (|h_t(i)| + |h_t(j)|) \frac{|r_{ij}|}{1 - |r_{ij}|},
\]

Efficient solution using max-spanning tree algorithms.
Graphical Decomposition

Replicate nodes/edges of $G$ to obtain a tractable graph $G'$ comprised of small/thin components.

For example, take a 2D grid model:

Split into disconnected edges or blocks:
Split into thin strips:

![Diagram of split into thin strips]

Split into trees:

![Diagram of split into trees]

Tree-decomposition is essentially equivalent to TRMP approach of Martin Wainwright, which considers convex combinations of trees.
Gaussian Lagrangian Relaxation*

Split $\mathcal{G}$ into small clusters $V_k \subset V$, replicate variables in larger disconnected graph $\mathcal{G}'$.

\[
f(x) = -\frac{1}{2}x^T J x + h^T x
\]

\[
f'(x') = \sum_k \left[-\frac{1}{2}(x'_k)^T J_k x'_k + h_k^T x'_k\right]
\]

where $J = \sum_k J_k$, $h = \sum_k h_k$ and $J_k \succ 0$ (e.g., pairwise-normalizable).

MAP Estimation:

\[
f^* = \max f(x) = \max \{f'(x') | Ax' = 0\}
\]

Lagrangian Relaxation:

\[
g(\lambda) = \max_{x'} \{f'(x') + \lambda^T Ax'\} \geq f^*
\]

Dual Problem: $\min g(\lambda) \triangleq g^*$. Convex quadratic program with linear constraints $\Rightarrow g^* = f^*$.

*Johnson, Malioutov, Willsky, '07.
Free-Energy and Maximum-Entropy

Let $\Phi$ denote Gaussian log-normalization:

$$
\Phi(h, J) = \frac{1}{2}\{h^T J^{-1} h - \log \det J + n \log 2\pi\}
$$

Free-Energy Minimization:

$$
\min \sum_k \Phi(h_k, J_k) \\
s.t. \sum_k h_k = h, \sum_k J_k = J, J_k \succ 0
$$

Dual Problem: for fixed $\{h_k, J_k\}$

$$
\max \sum_k \{H(P_k) - \frac{1}{2} Tr(P_k J_k) + h_k^T \hat{x}_k\} \\
s.t. \hat{x}_k[S_{kl}] = \hat{x}_l[S_{kl}], \\
\bar{x}_k[S_{kl}] = \bar{x}_l[S_{kl}], P_k \succ 0
$$

where $S_{kl} = V_k \cap V_l$ and

$$
H(P) = \frac{1}{2}\{\log \det P + n \log 2\pi e\}
$$

Find splitting which achieves marginal-matching on intersections.
Quadratic Max-Sum Diffusion

For each $S$ contained in the intersection of multiple clusters $V_k$, do the following:

1. Compute $S$-Marginals $(\hat{h}_S^k, \hat{J}_S^k)$ by GE within each cluster $S \subset V_k$.

2. Average marginal information:

   $$\bar{h}_S = \frac{1}{n_S} \sum_k \hat{h}_S^k \quad \text{and} \quad \bar{J}_S = \frac{1}{n_S} \sum_k \hat{J}_S^k$$

3. Equalize Marginals:

   $$h_k = h_k + (\bar{h}_S - \hat{h}_S^k) \quad \text{and} \quad J_k = J_k + (\bar{J}_S - \hat{J}_S^k)$$

Maintains valid splitting, converges to optimal solution of LR and Free-Energy minimization.
Gaussian Example

$64 \times 64$ random-field thin-plate model (penalizes curvature).

Slower convergence for stronger correlations...
Multi-scale Reparameterization

Define a family of constrained multi-scale models that are *equivalent* to the original MRF.

Relax the cross-scale constraints and break up each level into tractable subgraphs.

**Dual Problem** minimize the MAP value over all equivalent multi-scale reparameterizations.
Generalized Gaussian LR

Allow general linear constraints on “replicas”:

\[
\tilde{x}_k \triangleq A_k x_{E_k} \text{ equal for all } E_k \in \mathcal{E}_c
\]

Relaxes to moment constraints:

\[
\tilde{\mu}_k \triangleq A_k \mu_k \text{ and } \tilde{P}_k \triangleq A_k P_k A_k^T
\]

must be equal across replicas.

Algorithm:

Compute marginal potentials:

\[
\tilde{h}_k = \tilde{P}_k^{-1} \tilde{\mu}_k, \quad \tilde{J}_k = \tilde{P}_k^{-1}
\]

Average:

\[
\bar{h} = \frac{1}{K} \sum_k \tilde{h}_k, \quad \bar{J} = \frac{1}{K} \sum_k \tilde{J}_k
\]

Reparameterize:

\[
h' \leftarrow h' + A_k^T (\bar{h} - \tilde{h}_k)
\]

\[
J' \leftarrow J' + A_k^T (\bar{J} - \tilde{J}_k) A_k
\]

Iterate over all constraints until convergence.
Multi-Scale Gaussian Example

$128 \times 128$ random-field thin membrane model (penalizes image gradient).
Newton’s Method for Smooth MRFs

Consider MRF with smooth, convex potential functions:

\[ f(x) = \sum_{E \in G} \theta_E(x_E) \]

Compute the MAP estimate \( x^* \) using Newton’s method (with back-tracking line search). Requires solving:

\[ H\delta = g \tag{1} \]

where

\[ H = \sum_{E} [\nabla^2 \theta_E]_V \quad \text{and} \quad g = \sum_{E} [\nabla \theta_E]_V \tag{2} \]

Equivalent to MAP estimate in \( G \)-normalizable GMRF. Can use GaBP or LR to solve it.

What if \( \theta \)'s aren’t convex? Levenberg-Marquardt method adds regularization term \( \gamma \| x \|^2 \), i.e., diagonal loading. Similar to stabilized GaBP, insure that regularized model is WS. Converges to local maxima.
Example: Half-Quadratic
Edge-Preserving Image Restoration

Non-Linear Thin-Membrane Model:

\[ f(x) = \sum_{\{i,j\} \in G} \phi(x_i - x_j) + \alpha \sum_i (x_i - y_i)^2 \]

where quadratic penalty \( \phi(t) = t^2 \) tends to smooth across edges. Other \( \phi \)'s have better edge-preserving properties, e.g., \( \phi_p(t) = |t|^p \) for \( 0 < p < 2 \) (convex for \( p \geq 1 \)).

Smooth approximation to \( |t|^p \):

\[ \phi_{p,\beta}(t) = (t^2 + \beta)^{p/2} \rightarrow |t|^p \quad (\beta \rightarrow 0^+) \]

Solve for several decreasing values of \( \beta > 0 \).
Convex Case ($p = 1$)
Non-Convex Case \((p = \frac{1}{2})\)

Noisy Measurements

\(p = 0.5, \beta = 0.2\)

\(p = 0.5, \beta = 0.04\)

\(p = 0.5, \beta = 0.008\)

\(p = 0.5, \beta = 0.0016\)

\(p = 0.5, \beta = 0.00032\)
Conclusion

♠ Summary:

◊ Many approaches for walk-summable models.

◊ Lagrangian Relaxation for a generalized class of models.

◊ Multiscale methods useful to accelerate convergence.

◊ These methods can be used as preconditioner for harder (non-WS, non-linear) problems.

♠ Further work:

◊ Combine inner/outer loop in non-linear method.

◊ Adaptive diagonal regularization.

◊ Obtaining variance estimates.

◊ LR Approach for Hybrid models.

◊ Continuous approaches to discrete problems.
References*


JMW. Walk-Sum Interpretation and Analysis of Gaussian Belief Propagation, NIPS 2005.

MJW. Walk-Sums and Belief Propagation in Gaussian Graphical Models. JMLR, 2006.


*J=Johnson,M=Malioutov,
C=Chandrasekaran,W=Willsky.