Stochastic processes on graphs with cycles: Approximate inference and bounds

Speaker: Martin Wainwright, MIT

mjwain@mit.edu

Workshop on Information Theory at MSRI

February 27, 2002

Joint work with: Tommi Jaakkola & Alan Willsky

tommi@ai.mit.edu willsky@mit.edu

Outline

1. Introduction and background
   (a) Stochastic processes on graphs
   (b) Different graphical formalisms
   (c) Junction tree representation
2. Approximate inference as reparameterization
   (a) Theoretical results on belief propagation: fixed points, invariance, error analysis
   (b) Extensions to more advanced approximations: Cluster variational and relatives
3. Bounding the log partition function
4. Conclusions and future directions

§1. Introduction and Background

Stochastic processes defined by graphs arise in a variety of fields:

coding theory: various graphical codes including LDPCs, turbo codes (e.g., Gallager, 1963; Luby et al. 1998, McEliece et al., 1998)

statistical physics: models of gases, magnets, crystals (e.g, Ising model; Potts model)

artificial intelligence: neural network models; medical diagnosis; robotics (e.g, Pearl, 1988; Jordan et al., 1999)

statistics: log-linear models; maximum entropy; Markov random fields (e.g., Hammersley & Clifford, 1973; Darroch et al., 1980)

image processing and computer vision: Markov image models; Gibbs sampler (e.g., Woods, 1978; Geman & Geman, 1984)

network information theory: e.g., broadcast channel; MAC (e.g., Cover, 1972; El Gamal & Cover, 1980; Csiszár & Körner, 1980)

Set-up for graphical models

- graph $\mathcal{G}$ $\begin{cases} \text{set of nodes } \mathcal{V} = \{1, \ldots, N\} \\ \text{set of edges } \mathcal{E} \end{cases}$
- place at each node $s \in \mathcal{V}$ a random variable x_s taking values in the space $\mathcal{X}$ (e.g., $\mathcal{X} = \mathbb{R}$; $\mathcal{X} = \{0, 1, \ldots, m-1\}$)
- overall sample space $\mathcal{X}^N$ is set of all $N-$vectors

$$\mathbf{x} \triangleq \{ x_s \mid s \in \mathcal{V} \}$$

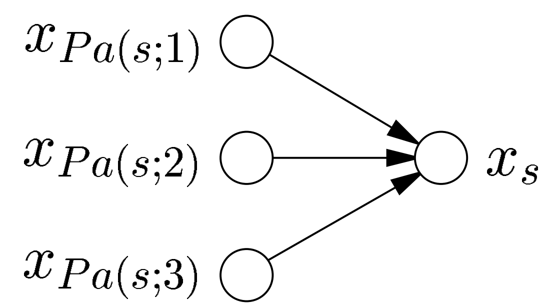
- will consider probability distributions $p(\mathbf{x})$ that are constrained by graph structure

Directed versus undirected edges

(a) **Directed graphs**

Full distribution specified as the product of conditional distributions over x_s given the set of its parents:

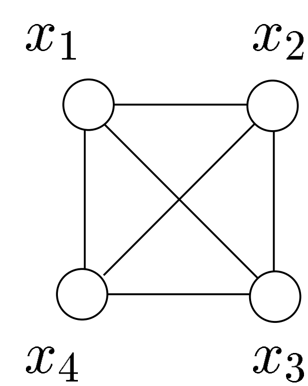
$$\mathbf{x}_{Pa(s)} = \{ \, x_t \mid t \text{ is parent of } s \, \}$$



(b) **Undirected graphs**

Full distribution specified as the product of compatibility functions $\psi_{\mathcal{C}}(\mathbf{x}_{\mathcal{C}})$ over variables in cliques:

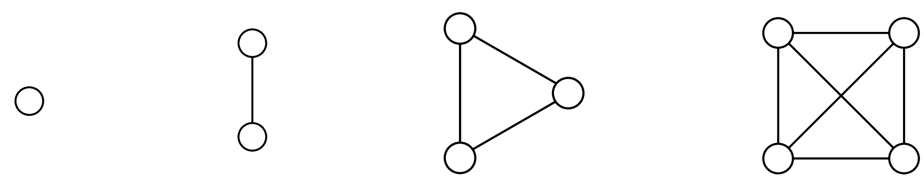
$$\mathbf{x}_{\mathcal{C}} = \{ \, x_t \mid t \in \mathcal{C} \, \}$$



$\psi_{1234}(x_1, x_2, x_3, x_4)$

Notation for undirected graphs

- *clique*: a fully connected subset $\mathcal{C}$ of $\mathcal{V}$; i.e., for all $s, t \in \mathcal{C}, (s,t) \in \mathcal{E}$



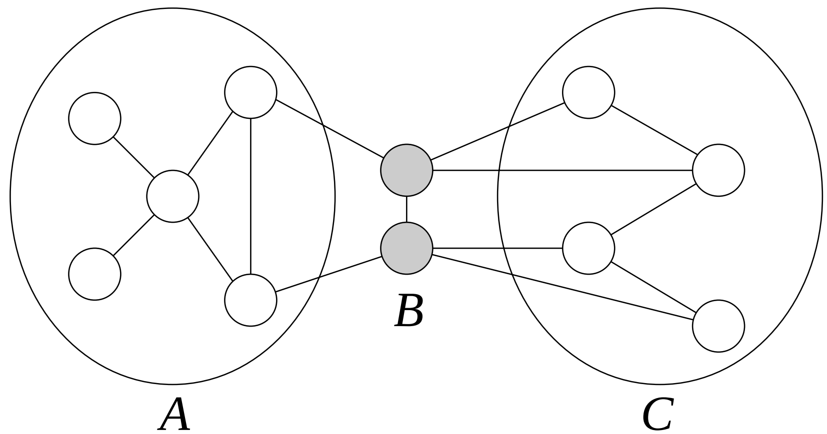
- *maximal clique*: a clique not properly contained within any other clique
- *compatibility function*: $\psi_{\mathcal{C}} : \mathcal{X}^N \to \mathbb{R}$ depending only on a limited subvector $\mathbf{x}_{\mathcal{C}} = \{ \, x_s \mid s \in \mathcal{C} \, \}$

  E.g. for binary $\mathbf{x}$, compatibility function on 2-clique $\{s, t\}$:

$$\psi_{st}(x_s, x_t) = \begin{pmatrix} \psi_{st}(0,0) & \psi_{st}(0,1) \\ \psi_{st}(1,0) & \psi_{st}(1,1) \end{pmatrix}$$

Graph separation and Markov

- stochastic processes $\mathbf{x}$ of interest are *Markov* with respect to the graph



Markov property: $\mathbf{x}_{A|B} \perp \mathbf{x}_{C|B}$ if B separates A from C.

Note: The notation $\mathbf{x}_{A|B} \perp \mathbf{x}_{C|B}$ means that $\mathbf{x}_A$ is conditionally independent of $\mathbf{x}_C$ given $\mathbf{x}_B$.

Hammersley-Clifford theorem

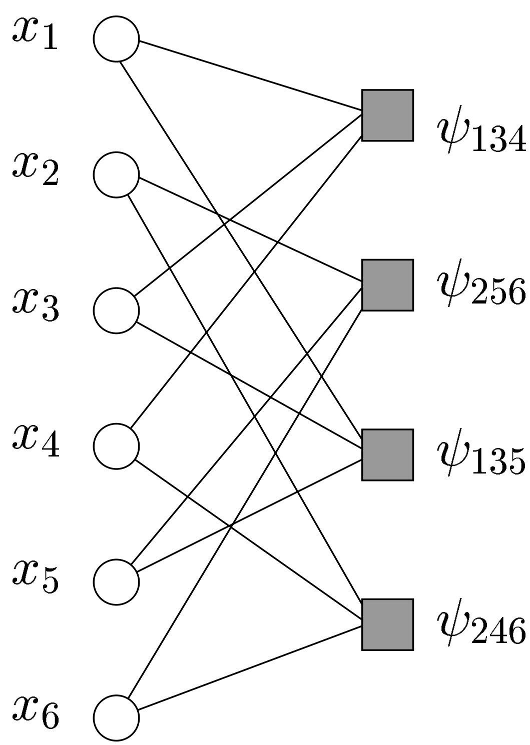
Consider stochastic process $\mathbf{x}$ on $\mathcal{G}$ such that $p(\mathbf{e}) > 0 \;\; \forall\, \mathbf{e} \in \mathcal{X}^N$.

$$\underbrace{\mathbf{x} \;\text{ is }\; \text{Markov} \;\; \text{w.r.t}\; \mathcal{G}}_{\text{Markov property}} \quad \Longleftrightarrow \quad \underbrace{p(\mathbf{x}) = \frac{1}{Z} \prod_{\mathcal{C} \in \mathbf{C}} \psi_{\mathcal{C}}(\mathbf{x})}_{\text{Factorization of distribution}}$$

Remarks:

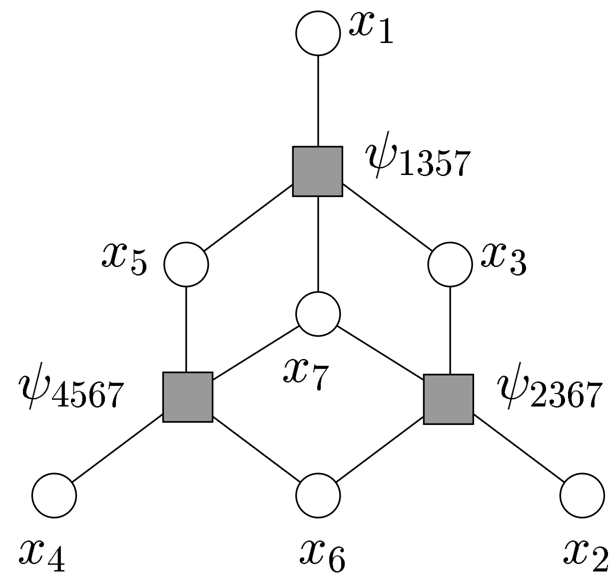
1. The *partition function* $Z = \sum_{\mathbf{x} \in \mathcal{X}^N} \prod_{\mathcal{C}} \psi_{\mathcal{C}}(\mathbf{x})$ is the normalizing constant.
2. There are a variety of proofs of this result (Hammersley & Clifford, 1973; Grimmett, 1973; Besag, 1974; Clifford, 1990).

Tanner graphs and factor graphs



(a) Tanner graph

(2,3) LDPC



(b) Factor graph

(7,4) Hamming code

Estimation or inference

- given observations $\mathbf{y} = \{\, y_s \mid s \in \mathcal{V}\}$ specified by measurement density

$$p(\mathbf{y}|\mathbf{x}) = \prod_s \psi_s(x_s; y_s)$$

- by Bayes' rule:

$$p(\mathbf{x}|\mathbf{y}) \;=\; \frac{1}{Z'} \prod_s \psi_s(x_s; y_s) \prod_{\mathcal{C}} \psi_{\mathcal{C}}(\mathbf{x}_{\mathcal{C}})$$

- this conditional density is central to various estimation problems:

(a) MAP estimate $\widehat{\mathbf{x}}_{MAP} = \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y})$

(b) node marginal distributions $p(x_s|\mathbf{y}) = \sum_{\mathbf{x}' \text{ s.t } x'_s = x_s} p(\mathbf{x}'|\mathbf{y})$

Examples of inference problems

1. **Decoding of graphical codes:** vector $\mathbf{y}$ represents bits received from noisy channel; $\mathbf{x}$ represents the codeword.

   (a) $\widehat{\mathbf{x}}_{MAP} = \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y})$ minimizes word error rate.

   (b) $\widehat{x}_s = \begin{cases} 1 & \text{if} \quad p(x_s = 1 \,|\, \mathbf{y}) > 0.5 \\ 0 & \text{otherwise} \end{cases}$ minimizes symbol error rate.

2. **Image denoising:** vector $\mathbf{x}$ is a representation of the image (e.g., pixels, wavelets); vector $\mathbf{y}$ is a noise-corrupted version of the image $\mathbf{x}$.

3. **Medical diagnosis:** vector $\mathbf{y}$ represents the observed constellation of symptoms; $\mathbf{x}$ represents the underlying disease.

Algorithms for trees

- for graphs without cycles, exploit the partial ordering of nodes in scale — i.e., dynamic programming on trees
- this leads to direct, recursive algorithms for inference:

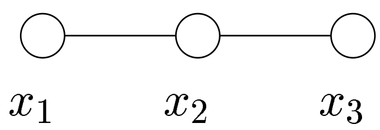
  (a) computation of $\widehat{\mathbf{x}}_{MAP}$: *max-product/min-sum algorithm*
  (generalization of Viterbi algorithm)

  (b) computation of marginals $p(x_s \,|\, \mathbf{y})$: *sum-product algorithm*, also known as *belief propagation*.
  (generalization of BCJR; Kalman-RTS; $\alpha - \beta$ algorithm etc.)

- more generally, similar algorithms apply to any commutative semi-ring (Verdú & Poor, 1987; Aji & McEliece, 2001)

Alternative high-level view of inference

- consider a very simple example: the Markov chain



- HC theorem gives a representation of the form:
  $p(\mathbf{x}) = \frac{1}{Z}\,\psi_1(x_1)\,\psi_2(x_2)\,\psi_3(x_3)\,\psi_{12}(x_1, x_2)\,\psi_{23}(x_2, x_3)$

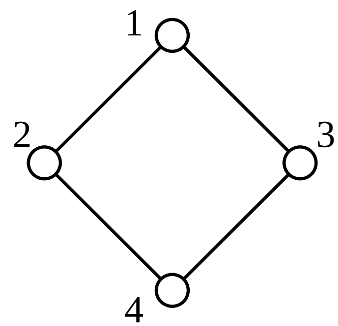
- think of inference (i.e., computing marginals) as converting from the $\{\psi_s,\, \psi_{st}\}$-representation to the more familiar form:

$$\begin{aligned} p(\mathbf{x}) &= p(x_1)\;\; p(x_2 \,|\, x_1)\;\; p(x_3 \,|\, x_2) \\ &= p(x_1)\, p(x_2)\, p(x_3) \left[\frac{p(x_1, x_2)}{p(x_1)\, p(x_2)}\right] \left[\frac{p(x_2, x_3)}{p(x_2)\, p(x_3)}\right] \end{aligned}$$

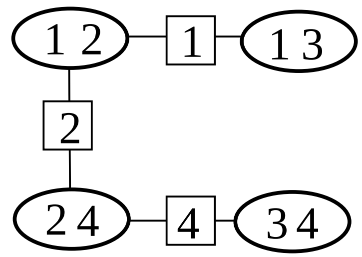
What to do for graphs with cycles?

Idea: Cluster nodes within cliques of graph with cycles to form a *clique tree*. Run a standard tree algorithm on this clique tree.

Caution: A naive approach will fail.



(a)



(b)

Need to enforce consistency between the copy of x_3 in cluster $\{1, 3\}$ and that in $\{3, 4\}$.

Running intersection and junction trees

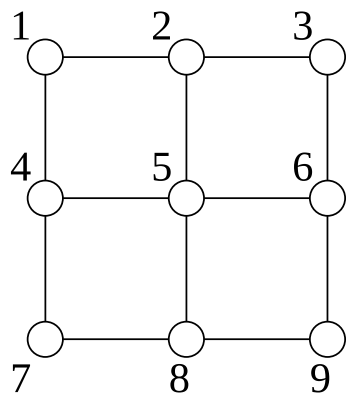
Definition: A clique tree satisfies the *running intersection property* if for any two clique nodes $\mathcal{C}_1$ and $\mathcal{C}_2$, all nodes on the unique path joining them contain the intersection $\mathcal{C}_1 \cap \mathcal{C}_2$.

A clique tree with this property is known as a *junction tree*.

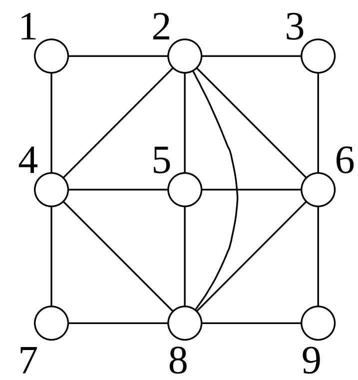
Definition: A graph $\mathcal{G}$ is triangulated means that every cycle of length 4 or greater has a chord.

Proposition: A graph $\mathcal{G}$ has a junction tree if and only if it is triangulated. (Lauritzen, 1996)

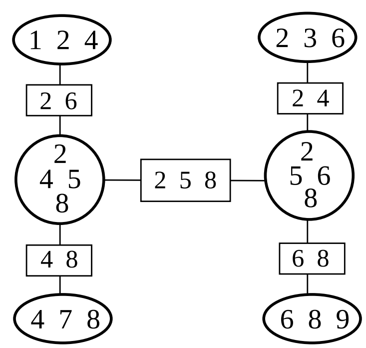
Illustration of junction tree



(a) Original graph



(b) Triangulated graph $\widetilde{\mathcal{G}}$



(c) Junction tree

Junction tree for exact inference

Algorithm: (Lauritzen & Spiegelhalter, 1988)

1. Given an undirected graph $\mathcal{G}$, form a triangulated graph $\widetilde{\mathcal{G}}$ by adding edges as necessary.
2. Form a junction tree of "super-nodes" by clustering together all nodes within each maximal clique.
3. Run standard inference algorithms on the resulting tree.

Note: Separator sets are formed by the intersections of cliques adjacent in the junction tree.

Junction tree representation

Junction tree representation guarantees that $p(\mathbf{x})$ can be factored as:

$$p(\mathbf{x}) = \frac{\prod_{\mathcal{C} \in \mathbf{C}_{\text{max}}} p(\mathbf{x}_{\mathcal{C}})}{\prod_{\mathcal{S} \in \mathbf{C}_{\text{sep}}} p(\mathbf{x}_{\mathcal{S}})}$$

where

$\mathbf{C}_{\text{max}} \equiv$ set of all maximal cliques in *triangulated* graph $\widetilde{\mathcal{G}}$

$\mathbf{C}_{\text{sep}} \equiv$ set of all separator sets (intersections of adjacent cliques)

Special case for tree:

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

§2. Approximate inference as reparameterization

- belief propagation (BP) is a message-passing algorithm for computing approximate marginals
- it is an exact method for trees, but approximate for graphs with cycles
- important in a variety of applications:

(a) coding theory: turbo codes and low-density parity check codes (e.g., Gallager, 1963; McEliece et al., 1998; McKay, 1998)

(b) artificial intelligence (e.g., Pearl, 1988; Murphy & Weiss, 2001)

(c) computer vision and statistical image processing (e.g., Freeman et. al., 1999, Frey et al., 2001)

Previous and current work on BP

- certain special cases well-understood:
  - (a) single loops
    (Aji et al., 1997; Anderson & Hladnik, 1998; Weiss, 1997, 2000)
  - (b) Gaussians on arbitrary graphs
    (Rusmevichientong & Van Roy, 2000; Weiss & Freeman, 2000)
- geometric approach to turbo decoding (Richardson, 2000)
- variational formulation as minimizing Bethe free energy
  (Yedidia, Freeman & Weiss, 2000)
- better algorithms for minimizing Bethe free energy
  (Yuille, 2001; Welling & Teh, 2001)
- more advanced approximations (Yedidia et al., 2000; Minka, 2001)

Our approach

BP fixed points are stationary points of F_{Bethe} (Yedidia et al., 2000):

$$F_{\text{Bethe}}(\{\mathbf{T}\}) \quad \approx \quad F_{\text{true}}(\{\mathbf{T}\})$$

We want to understand how:

$$\arg\min_{\{\mathbf{T}\}} F_{\text{Bethe}}(\{\mathbf{T}\}) \quad \approx \quad \arg\min_{\{\mathbf{T}\}} F_{\text{true}}(\{\mathbf{T}\})$$

New viewpoint in terms of reparameterization:

1. leads to novel characterization of the fixed points
2. gives analytical expression and bounds on the approximation error for an arbitrary graph

Notation

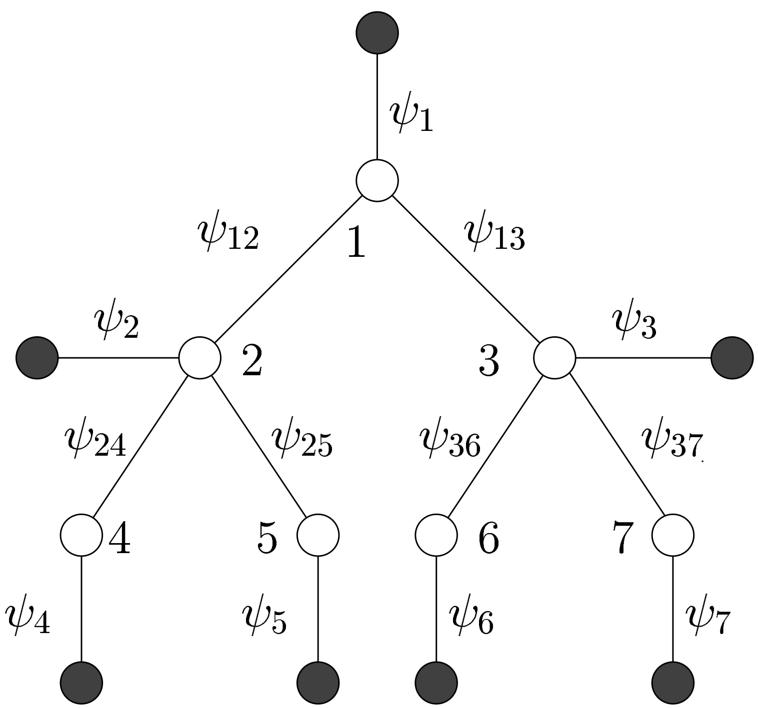
- with a few caveats, no loss of generality in restricting attention to pairwise MRFs: graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ such that edges are maximal cliques
  (**Note:** Our analysis extends to higher order cliques.)
- consider probability distribution over the discrete random vector $\mathbf{x} \in \mathcal{X}^N$:

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

Goal: Compute (approximations to) single-node marginal distributions:

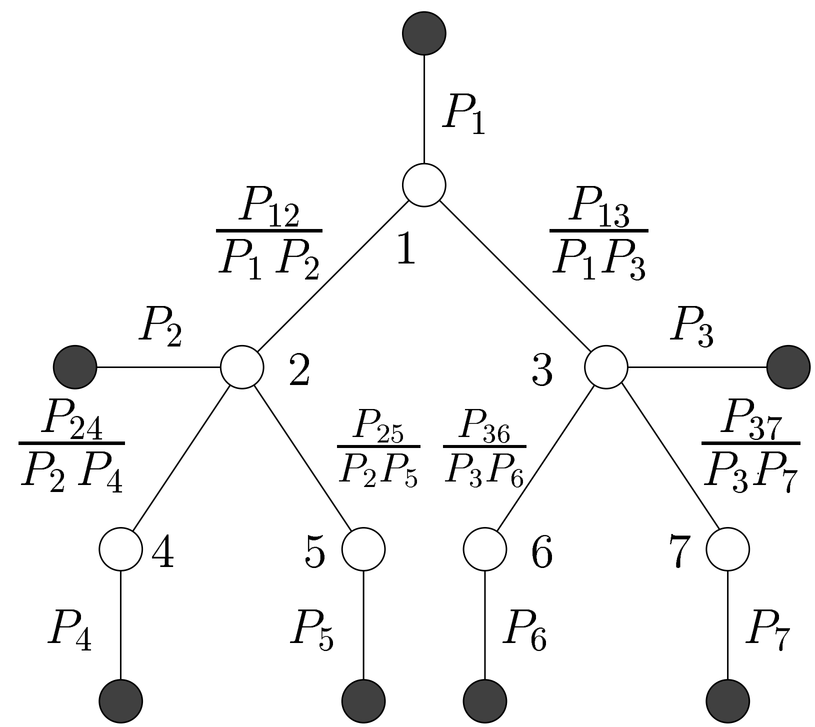
$$p(x_s) = \sum_{\mathbf{x}' \text{ s.t } x'_s = x_s} p(\mathbf{x}')$$

Tree estimation as reparameterization



(a) Initial parameterization

$p(\mathbf{x}) = \frac{1}{Z} \prod_s \psi_s(x_s) \prod_{(s,t)} \psi_{st}(x_s, x_t)$

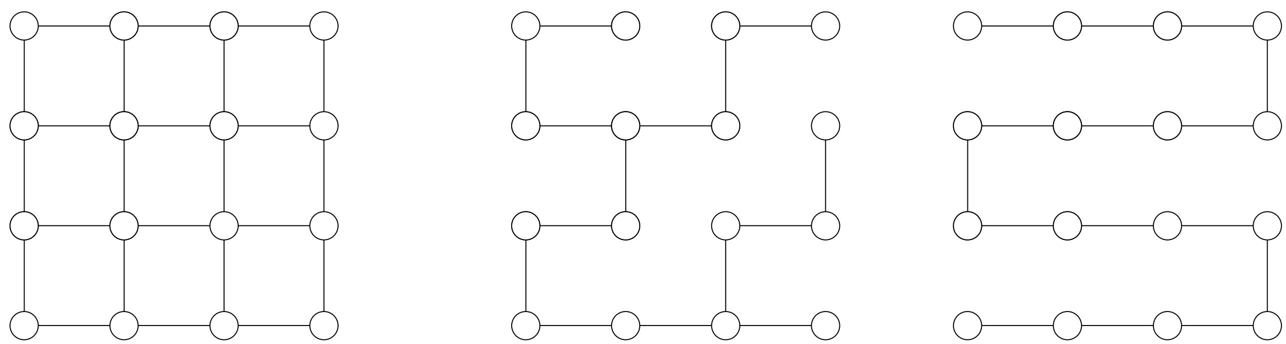


(b) Desired parameterization

$p(\mathbf{x}) = \prod_s P_s(x_s) \prod_{(s,t)} \frac{P_{st}(x_s, x_t)}{P_s(x_s)\, P_t(x_t)}$

Embedded spanning trees

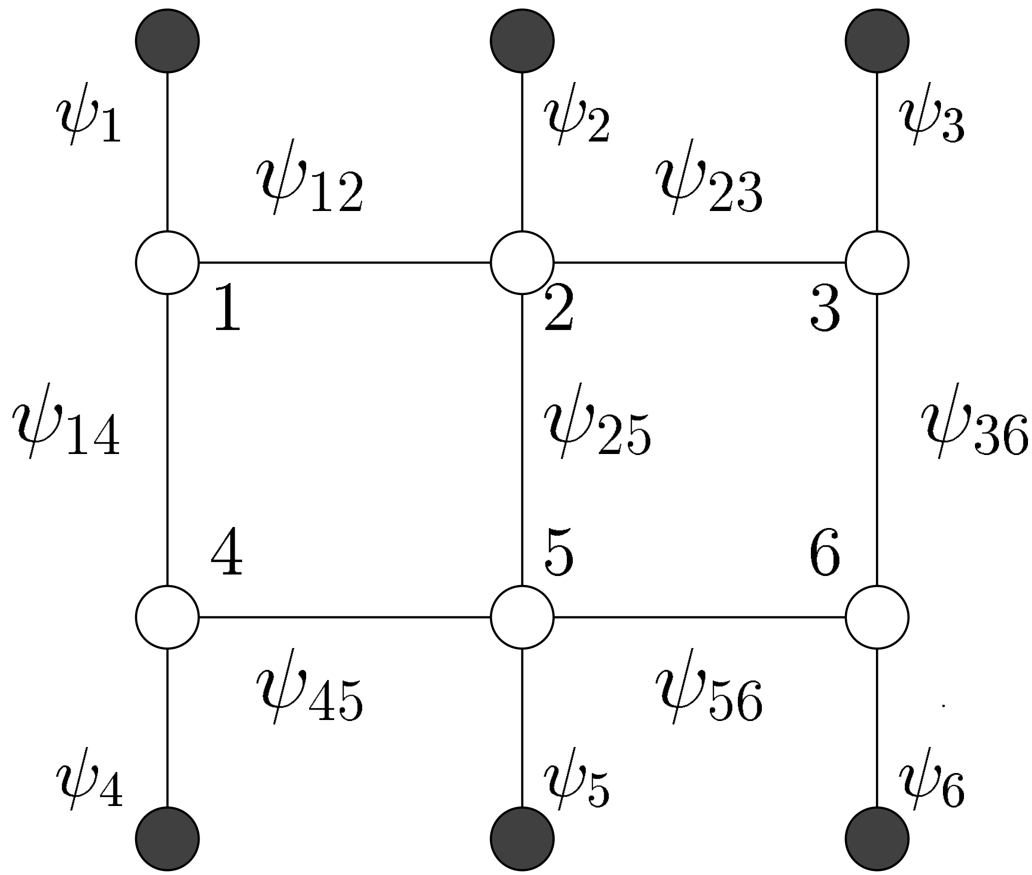
Observation: A graph with cycles has a (typically) large number of spanning trees.



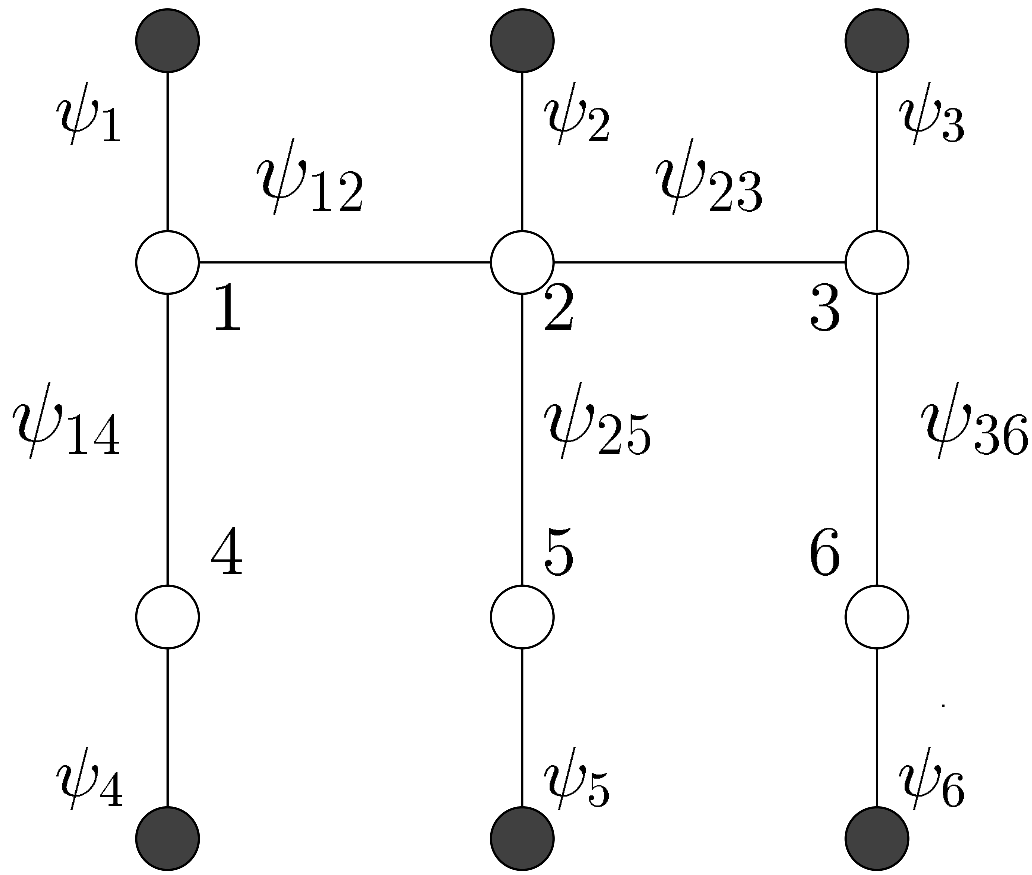
(a) Original graph $\mathcal{G}$ (b) Tree $\mathcal{T}^1$ (c) Tree $\mathcal{T}^2$

Strategy: Define and study modified problems on spanning trees. Let $\mathcal{T}^i$ denote a spanning tree with edge set $\mathcal{E}(\mathcal{T}^i) \equiv \mathcal{E}^i$.

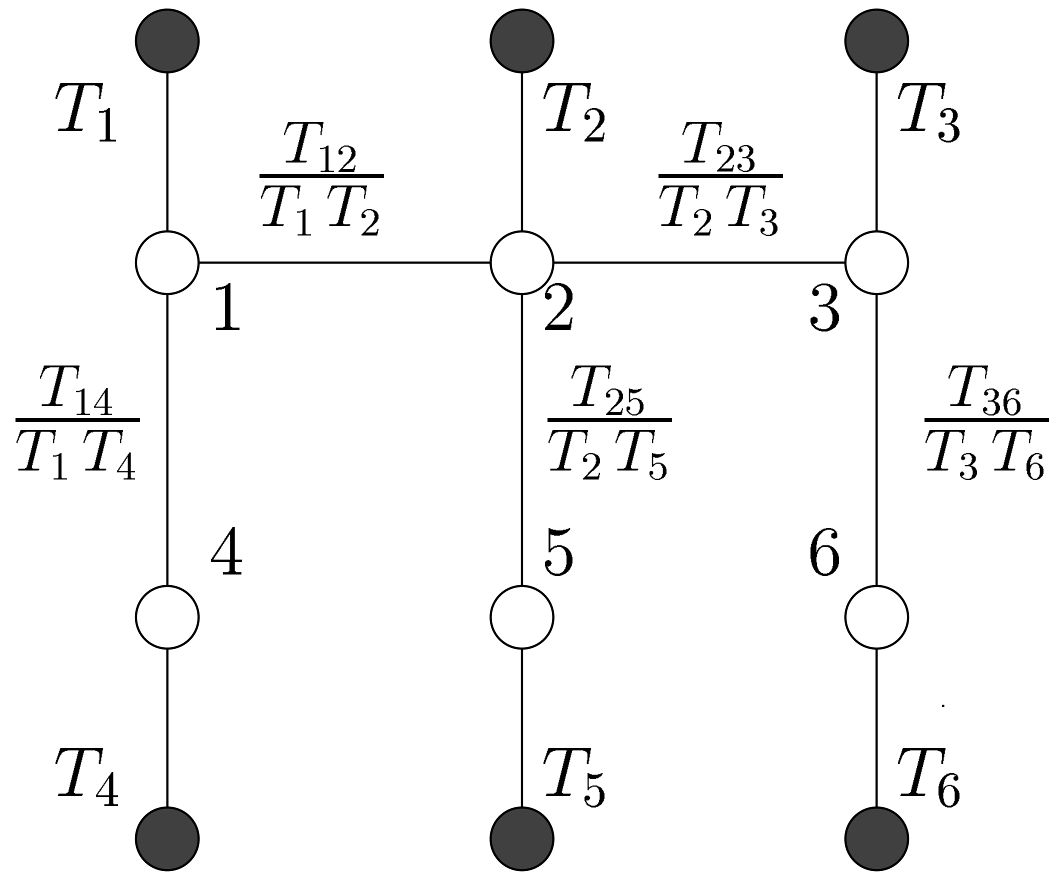
(1) Start with distribution $p(\mathbf{x}; \psi)$ on full graph.



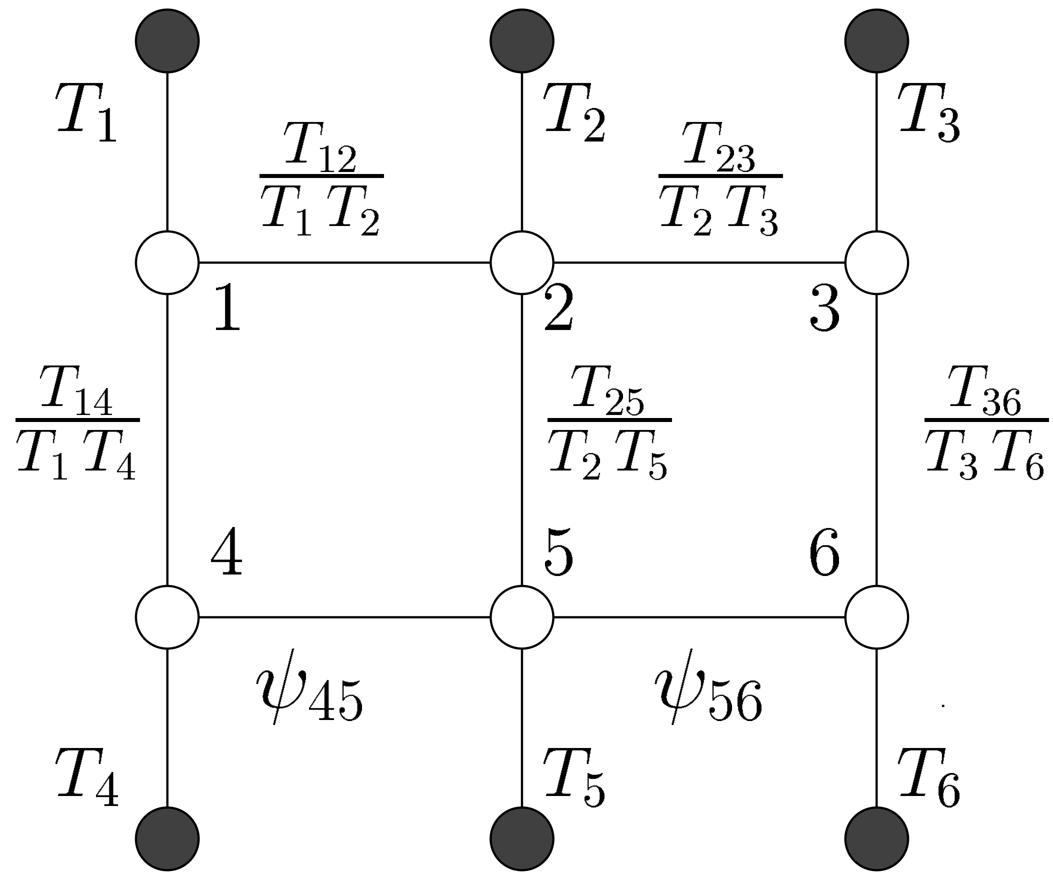
(2) Isolate components corresponding to spanning tree.



(3) Perform tree reparameterization update.



(4) Reinstate removed potentials.



Set-up for tree reparameterization (TRP)

Let $\mathbf{T}^n = \{ T_s^n, \, T_{st}^n \}$ be a vector of pseudomarginals at single nodes and edges.

Key parameterization:

$$p(\mathbf{x}; \mathbf{T}^n) \quad = \quad \frac{1}{Z(\mathbf{T}^n)} \prod_{s \in \mathcal{V}} T_s^n(x_s) \prod_{(s,t) \in \mathcal{E}} \frac{T_{st}^n(x_s, x_t)}{\left(\sum_{x'_s} T_{st}^n(x'_s, x_t)\right) \left(\sum_{x'_t} T_{st}^n(x_s, x'_t)\right)}$$

TRP is a sequence of functional updates $\mathbf{T}^n \mapsto \mathbf{T}^{n+1}$.

Tree decomposition: Given a set of tree edges $\mathcal{E}(\mathcal{T})$, break $p(\mathbf{x}; \mathbf{T}^n)$ into a product of two terms:

Tree terms: $$p^i(\mathbf{x}; \mathbf{T}^n) \quad = \quad \prod_{s \in \mathcal{V}} T_s^n \prod_{(s,t) \in \mathcal{E}(\mathcal{T})} \frac{T_{st}^n}{\left(\sum_{x_s} T_{st}^n\right) \left(\sum_{x_t} T_{st}^n\right)}$$

Residual: $$r^i(\mathbf{x}; \mathbf{T}^n) \quad = \quad \prod_{(s,t) \in \mathcal{E}/\mathcal{E}(\mathcal{T})} \frac{T_{st}^n}{\left(\sum_{x_s} T_{st}^n\right) \left(\sum_{x_t} T_{st}^n\right)}$$

TRP algorithm

1. Initialize $p(\mathbf{x}; \mathbf{T}^0)$ in terms of $\{\psi_s, \psi_{st}\}$:

$$\begin{aligned} T_s^0(x_s) &= \kappa\, \psi_s(x_s) \prod_{t \in \mathcal{N}(s)} \Big[\sum_{x'_t} \psi_{st}(x_s, x'_t)\psi_t(x'_t)\Big] \\ T_{st}^0(x_s, x_t) &= \kappa\, \psi_{st}(x_s, x_t)\psi_s(x_s)\psi_t(x_t) \end{aligned}$$

   Note that $p(\mathbf{x}; \mathbf{T}^0) \equiv p(\mathbf{x}; \boldsymbol{\psi})$.

2. Isolate $p^{i(n)}(\mathbf{x}; \mathbf{T}^n)$ corresponding to spanning tree $\mathcal{T}^{i(n)}$. Perform updates on tree:

$$\begin{aligned} T_{st}^{n+1}(x_s, x_t) &= \sum_{\mathbf{x}' \text{ s.t } x'_s = x_s,\ x'_t = x_t} p^{i(n)}(\mathbf{x}'; \mathbf{T}^n) \quad \forall\, (s,t) \in \mathcal{E}^{i(n)} \\ T_{st}^{n+1}(x_s, x_t) &= T_{st}^n(x_s, x_t) \quad \forall\, (s,t) \in \mathcal{E}/\mathcal{E}^{i(n)} \end{aligned}$$

Constraint sets and cost functions

The set of valid $\mathbf{T}$ satisfy the *local edge-wise* marginalization constraints:

$$\mathbb{C} \quad \triangleq \quad \Big\{\mathbf{T} \;\Big|\; \sum_{x'_s} T_s(x'_s) = 1 \;;\; \sum_{x'_s} T_{st}(x'_s, x_t) = T_t(x_t) \;\text{ for } (s,t) \in \mathcal{E}\Big\}$$

Use cost (closely related to Bethe free energy) that approximates the KL divergence between $p(\mathbf{x};\mathbf{T})$ and $p(\mathbf{x};\mathbf{U})$:

$$G(\mathbf{T};\mathbf{U}) = \sum_{s\in\mathcal{V}} G^s(T_s;U_s) + \sum_{(s,t)\in\mathcal{E}} G^{st}(T_{st};U_{st})$$

where

$$\begin{aligned} G^s(T_s;U_s) &= \sum_{x_s} T_s(x_s)\log[T_s(x_s)/U_s(x_s)] \\ G^{st}(T_{st};U_{st}) &= \sum_{x_s,x_t} T_{st}\Big\{\log[T_{st}/(\sum_{x_s} T_{st})(\sum_{x_t} T_{st})] - \log[U_{st}/(\sum_{x_s} U_{st})(\sum_{x_t} U_{st})]\Big\} \end{aligned}$$

TRP as successive projection method

- consider the set of $\mathbf{T}$ consistent on tree $\mathcal{T}^i$:

$$\mathbb{C}^i \quad \triangleq \quad \{\mathbf{T} \mid \sum_{x'_s} T_s(x'_s) = 1 \; ; \; \sum_{x'_s} T_{st}(x'_s, x_t) = T_t(x_t) \text{ for } (s,t) \in \mathcal{E}(\mathcal{T}^i)\}$$

  where $\mathcal{E}(\mathcal{T}^i) \subset \mathcal{E}$

- note that $\mathbb{C}^i \supset \mathbb{C}$, and that $\cap_i \mathbb{C}^i = \mathbb{C}$ whenever $\cup_i \mathcal{E}(\mathcal{T}^i) = \mathcal{E}$.

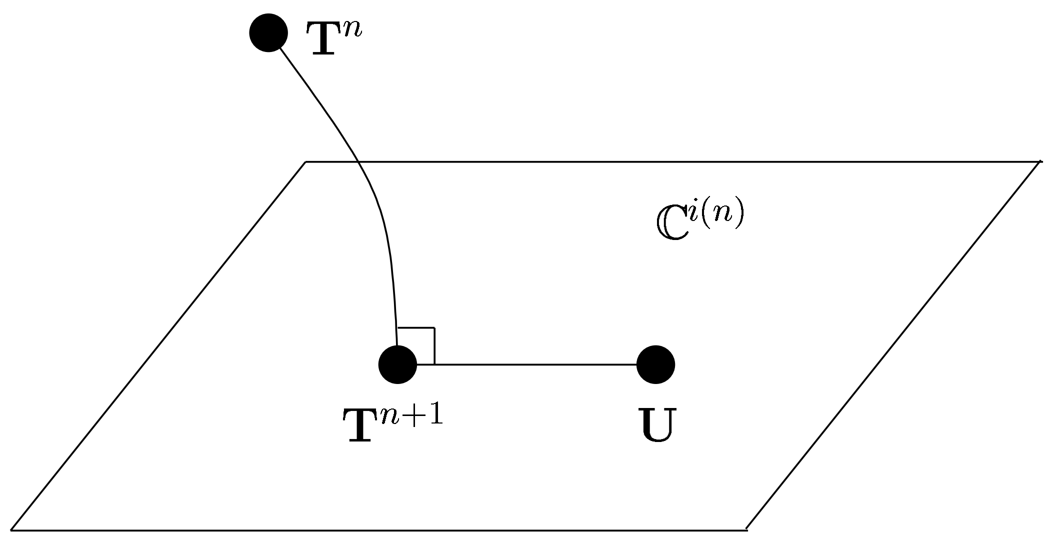
- TRP can be viewed as analogous to a successive projection technique for attempting to minimize $G(\mathbf{T}; \mathbf{T}^0)$ subject to the constraint $\mathbf{T} \in \cap \mathbb{C}^i$.

- each iteration entails a "projection" onto the constraint set $\mathbb{C}^{i(n)}$ associated with tree $\mathcal{T}^{i(n)}$.

Pythagorean relation

Proposition: At each iteration $n = 0, 1, 2\ldots$ and for all $\mathbf{U} \in \mathbb{C}^{i(n)}$:

$$G(\mathbf{U}; \mathbf{T}^n) = G(\mathbf{U}; \mathbf{T}^{n+1}) + G(\mathbf{T}^{n+1}; \mathbf{T}^n)$$

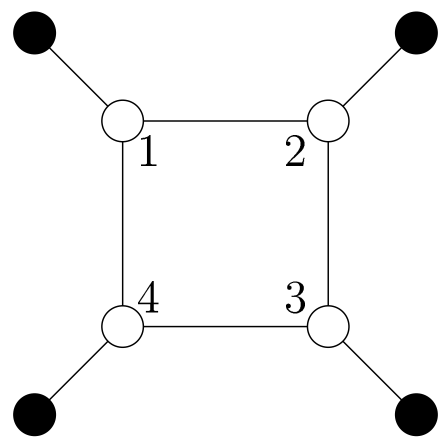


Remarks:

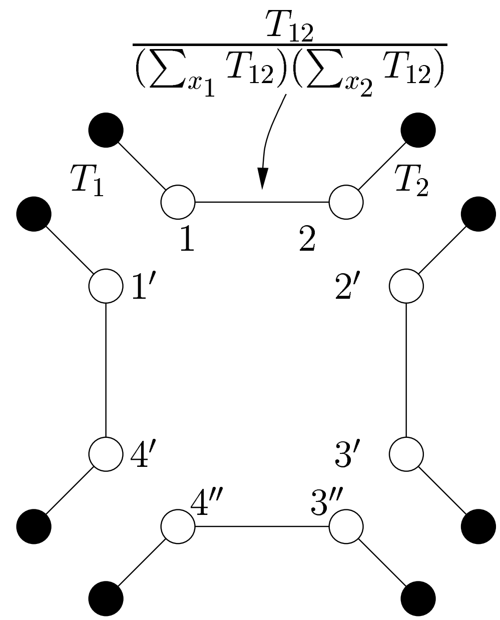
(a) Cost function G plays a role analogous to the squared Euclidean distance (hence Pythagorean).

(b) Similar relation holds for Bregman distances (e.g., KL divergence and information geometry). (Csiszár, 1975; Amari, 1982)

BP as reparameterization over two-node trees



(a) Original graph



(b) Two-node trees

BP can be reformulated as a very local form of reparameterization over 2-node trees.

Invariance of distribution

We initialize at $\mathbf{T}^0$ such that $p(\mathbf{x}; \mathbf{T}^0) \equiv p(\mathbf{x}; \boldsymbol{\psi})$.

At each iteration, we use the decomposition:

$$p(\mathbf{x}; \mathbf{T}^n) \quad \propto \quad \overbrace{p^i(\mathbf{x}; \mathbf{T}^n)}^{\text{tree terms}} \quad \underbrace{r^i(\mathbf{x}; \mathbf{T}^n)}_{\text{residual terms}}$$

Theorem:
Distribution on graph with cycles is invariant under the updates $\mathbf{T}^n \mapsto \mathbf{T}^{n+1}$. That is,

$$p(\mathbf{x}; \mathbf{T}^n) \equiv p(\mathbf{x}; \mathbf{T}^0) \qquad \text{for all } n = 1, 2, \ldots$$

Any limit point $\mathbf{T}^*$ is also a reparameterization in this sense.

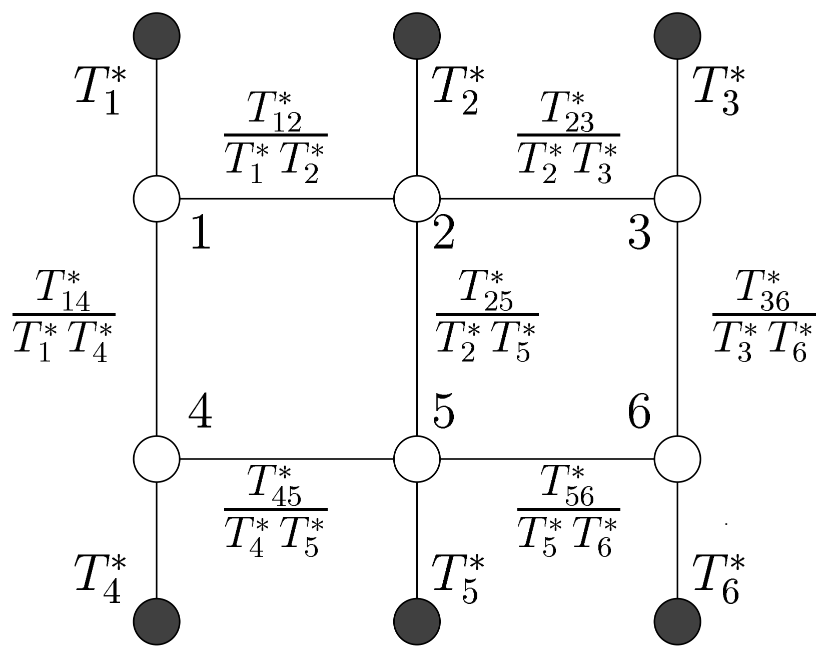
Remarks on invariance theorem

1. Invariance also holds for BP (when suitably reformulated in the reparameterization form).

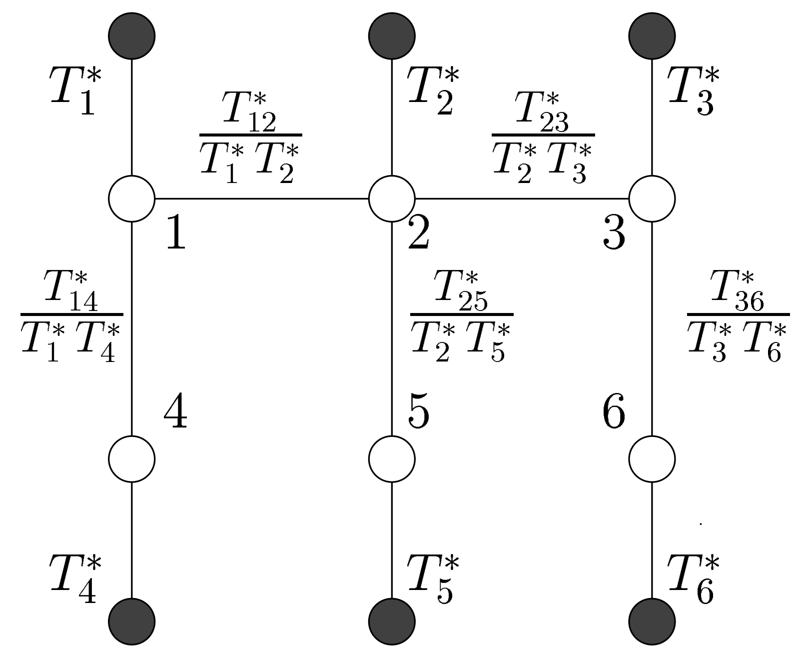
2. Any local minimum of Bethe free energy, regardless of the algorithm used to obtain it, is a reparameterization in this sense.

3. Special property of TRP/BP algorithms: *all* iterates (not just the fixed points) are reparameterizations of the original distribution.

Fixed point condition



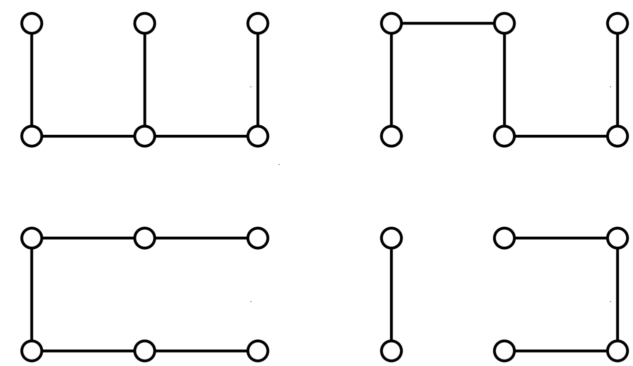
(a) Graph with cycles



(b) Tree consistency ($\mathcal{T}$-consistent)

Remarks on fixed pt. theorem

1. We are guaranteed that fixed point $\mathbf{T}^*$ is $\mathcal{T}-$consistent on *any* tree (or forest) embedded within the graph.



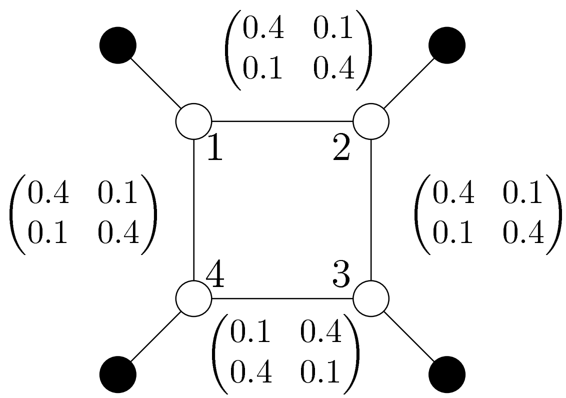
2. Fixed point characterization applies to any local minimum of Bethe free energy (regardless of the algorithm.)

3. The existence of such a $\mathcal{T}$-consistent reparameterization is obvious for a tree; more interesting for a graph with cycles.

4. The pseudomarginals $\mathbf{T} = \{T_s^*,\ T_{st}^*\}$, though $\mathcal{T}$-consistent, may not be consistent with any distribution globally on $\mathcal{G}$.

Illustration of global inconsistency

Consider the following assignments on the single cycle (MacKay et al., 2001):



$$\psi_s = [0.5 \quad 0.5]'$$

$$\psi_{12}, \psi_{23}, \psi_{14} \propto \begin{pmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{pmatrix}$$

$$\psi_{34} \propto \begin{pmatrix} 0.1 & 0.4 \\ 0.4 & 0.1 \end{pmatrix}$$

Can show:

1. The parameterization $\mathbf{T} = \{\psi_s, \psi_{st}\}$ is a BP/TRP fixed point.
2. However, the corresponding pseudomarginal vector $\mathbf{T}$ is *not* globally consistent with any distribution (Markov or otherwise).

Consequences of invariance and fixed pt. characterization

A. Geometric insight; links to information geometry
(Amari, 1982; Csiszár, 1975)

B. Strong restrictions on when TRP/BP can be exact
(there are cases other than trees!)

C. Elementary proof of exactness of means in Gaussian BP
(Weiss & Freeman, 2000; Rusmevichientong & Van Roy, 2000)

D. Error analysis

Analysis of BP approximation error

Previous results on error in special cases:

(a) exact expression for a single cycle (Weiss, 2000)

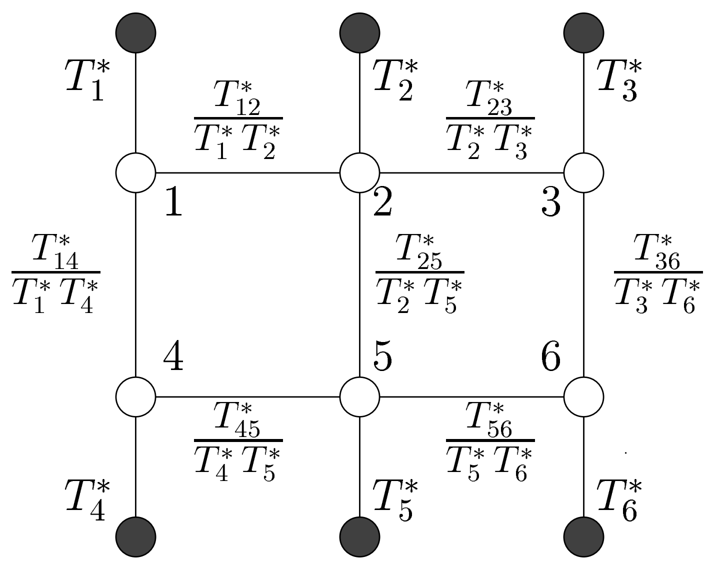
(b) approximate expression for turbo decoding (Richardson, 2000)

We give an exact expression and computable bounds for the error on an arbitrary graph with cycles.

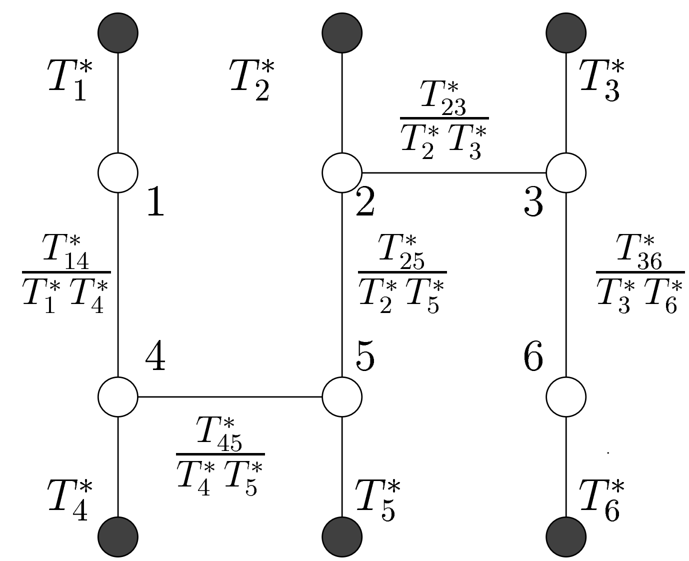
Key properties in our analysis are:

1. The quantities $\{T_s^*\}$ have *two* distinct interpretations:
   (a) TRP/BP approximations to the true marginals P_s on graph with cycles
   (b) consistent single node marginals of distribution defined on any spanning tree
2. From invariance of distribution, $p(\mathbf{x}; \boldsymbol{\psi}) = p(\mathbf{x}; \mathbf{T}^*)$

Consequences



(a) Original $p(\mathbf{x}; \mathbf{T}^*)$



(b) Consistent tree distribution

Exact marginals $\{P_s\}$ on graph with cycles are related to TRP/BP approximations $\{T_s^*\}$ by a perturbation — namely, removing edges.

Exact expression for error

Recall the decomposition of $p(\mathbf{x}; \mathbf{T}^*)$:

$$p(\mathbf{x}; \mathbf{T}^*) \quad = \quad \frac{1}{Z(\mathbf{T}^*)} \overbrace{p^i(\mathbf{x}; \mathbf{T}^*)}^{\text{tree-structured distribution}} \underbrace{r^i(\mathbf{x}; \mathbf{T}^*)}_{\text{residual terms}}$$

Following exact expression is starting point for deriving bounds:

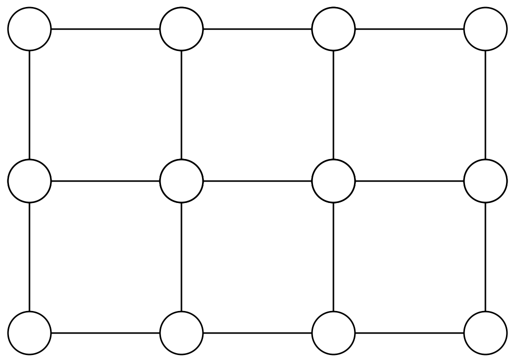
$$P_{s;j} - T^*_{s;j} \quad = \quad \mathbb{E}_{p^i(\mathbf{x};\mathbf{T}^*)} \left[\left\{ \frac{r^i(\mathbf{x}; \mathbf{T}^*)}{Z(\mathbf{T}^*)} - 1 \right\} \delta(x_s = j) \right]$$

Extensions to more advanced approximations

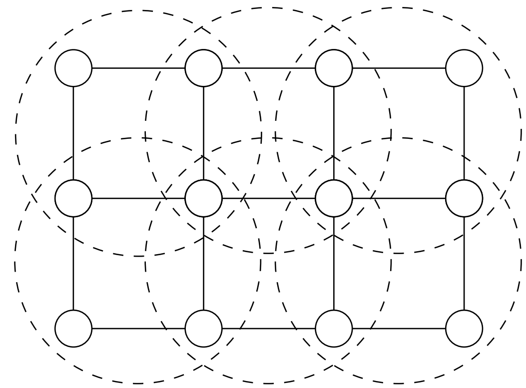
- techniques that exploit more structure than BP have been proposed:
  (a) Kikuchi and related methods (Yedidia et al., 2000)
  (b) expectation-propagation updates (Minka, 2001)

- our analysis carries over to these more advanced methods:
  (a) the idea of reparameterization is applicable
  (b) invariance of the distribution under updates
  (c) characterization of the fixed points, and error analysis

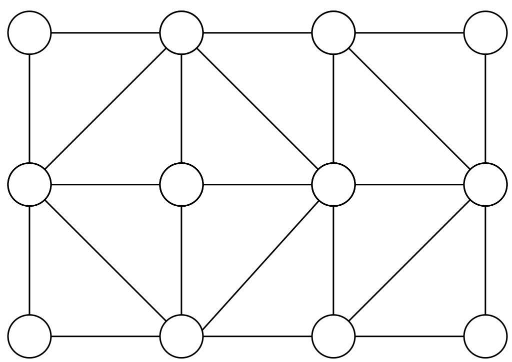
Illustration for Kikuchi approximation



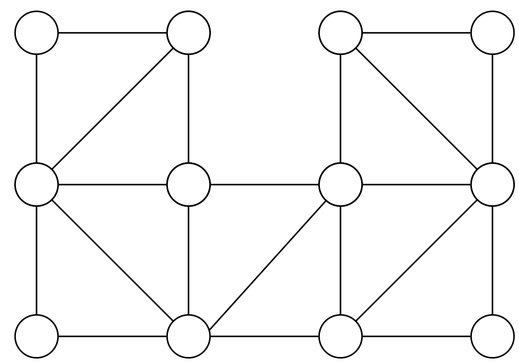
(a) Original graph



(b) Kikuchi 4-plaque clustering



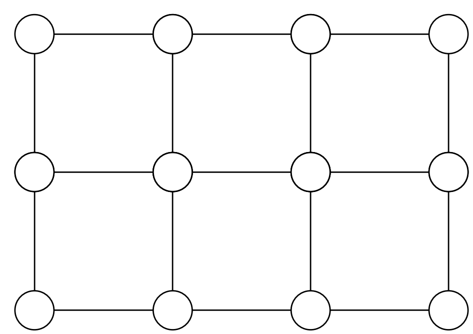
(c) Partial triangulation



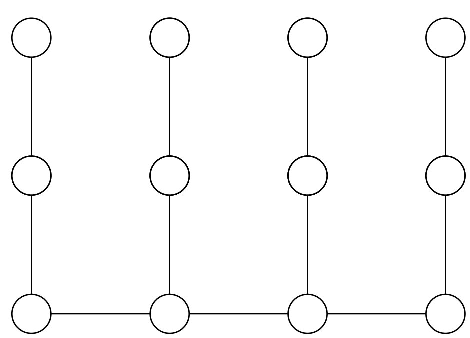
(d) Hypertree (treewidth 2)

Maximal subgraphs

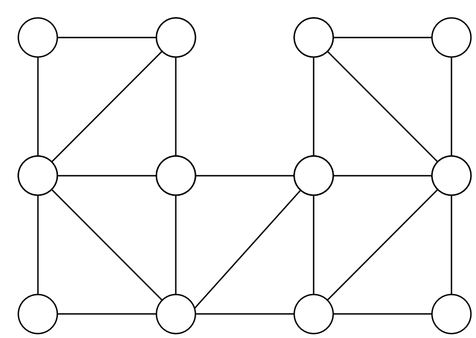
Question: What are the largest subgraphs over which the set of pseudomarginals $\mathbf{T}^*$ is guaranteed to be globally consistent?



(a) Original graph



(b) Spanning trees (Bethe)



(c) Width 2 hypertree (Kikuchi)

Implications for iterative decoding?

- most work on BP decoding (e.g., Luby et al., 2001; Richardson et al., 2001) has two key features:
  (a) entails averaging over an ensemble of codes
  (b) asymptotic in code length

- our work applies to BP decoding for a *fixed, finite-length* code:
  (a) recall that bitwise optimal (ML) decoding of a binary code is based on the sign of the log likelihood ratio $\log \frac{p(x_s=1;\mathbf{T}^*)}{p(x_s=0;\mathbf{T}^*)}$
  (b) BP decoding is based on the sign of modified likelihood ratio

$$\log \frac{p(x_s = 1;\ \Pi^{\mathcal{T}}(\mathbf{T}^*))}{p(x_s = 0;\ \Pi^{\mathcal{T}}(\mathbf{T}^*))}$$

Here $p(\mathbf{x}; \Pi^{\mathcal{T}}(\mathbf{T}^*))$ denotes a tree-structured distribution. In fact, this log likelihood is equal for *any* tree embedded within $\mathcal{G}$.

Possible research directions

- are there intermediate size codes/graphs for which BP log likelihood ratio is guaranteed (w.h.p) to have the same sign as the optimal LLR?

- enhancing BP approximations (post hoc) by including higher-order terms — i.e., partially accounting for presence of cycles

- uses in reliability-based decoding (e.g., Fossorier, 2001)

Note: If a tree-based updates are used, then bounds on the error can still be obtained *prior* to BP convergence.

§3. Bounds on the log partition function

Question: What is wrong with the Bethe/Kikuchi free energies?

- usually not convex (multiple local minima; convergence issues)
- do not give bounds on the log partition function

Bounding the partition function is important for various problems:

- obtaining bounds on marginals and likelihood ratios
- large deviations analysis (error exponents)
- bounds on rate distortion and capacity

Bounds based on convex combinations of trees

- a new class of upper bounds on the log partition function based on convex combinations of (hyper)trees
- leads to "convexified" Bethe/Kikuchi free energies

Notation:

- let $\mathfrak{T}$ denote the set of spanning trees of $\mathcal{G}$
  (typically, a large set; e.g., for the complete graph K_N, $|\mathfrak{T}| = N^{N-2}$)
- let $\vec{\boldsymbol{\mu}} = \{\, \mu(\mathcal{T}) \mid \mathcal{T} \in \mathfrak{T} \,\}$ be a probability distribution over all spanning trees of the graph.
- for each edge $e \in \mathcal{E}$, let $\mu_e = \Pr_{\vec{\boldsymbol{\mu}}}\{\, e \in \mathcal{T} \,\}$ be the *edge appearance probability*.
- let $\mathbb{T}(\mathcal{G})$ be the valid set of $\boldsymbol{\mu_e} = \{\, \mu_e \mid e \in \mathcal{E} \,\}$; this is the *spanning tree polytope* (Edmonds, 1971).

Convexified Bethe free energy

Consider the distribution:

$$\begin{aligned} p(\mathbf{x}) &= \frac{1}{Z(\boldsymbol{\psi})} \prod_{s \in \mathcal{V}} \psi_s(x_s) \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t) \\ Z(\boldsymbol{\psi}) &= \sum_{\mathbf{x} \in \mathcal{X}^N} \Big[\prod_{s \in \mathcal{V}} \psi_s(x_s) \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t) \Big] \end{aligned}$$

Let $\boldsymbol{\mu_e} \in \mathbb{T}(\mathcal{G})$ be arbitrary. Bounds on $\log Z(\boldsymbol{\psi})$ are based on the following function:

$$\begin{aligned} \mathcal{F}(\mathbf{T}; \boldsymbol{\mu_e}; \boldsymbol{\psi}) \triangleq -\sum_{s \in \mathcal{V}} H_s(T_s) + \sum_{(s,t) \in \mathcal{E}} \mu_{st} I_{st}(T_{st}) \\ - \sum_{s \in \mathcal{V}} \mathbb{E}_{T_s}[\log \psi_s] - \sum_{(s,t) \in \mathcal{E}} \mathbb{E}_{T_{st}}[\log \psi_{st}] \end{aligned}$$

$$\begin{aligned} H_s(T_s) &\triangleq \text{entropy of node marginal } T_s(x_s) \\ I_{st}(T_{st}) &\triangleq \text{mutual information under joint } T_{st}(x_s, x_t) \end{aligned}$$

Theorem: For all $\boldsymbol{\mu_e} \in \mathbb{T}(\mathcal{G})$:

(a) The quantity $\mathcal{F}(\mathbf{T}; \boldsymbol{\mu_e}; \boldsymbol{\psi})$ is convex as a function of $\mathbf{T}$.

(b) The log partition function is bounded above as

$$\log Z(\boldsymbol{\psi}) \quad \leq \quad -\min_{\mathbf{T} \in \mathbb{C}} \mathcal{F}(\mathbf{T}; \boldsymbol{\mu_e}; \boldsymbol{\psi})$$

where

$$\mathbb{C} \quad \triangleq \quad \{\, \mathbf{T} \mid \sum_{x'_s} T_s(x'_s) = 1 \,;\, \sum_{x'_s} T_{st}(x'_s, x_t) = T_t(x_t) \text{ for } (s,t) \in \mathcal{E} \,\}$$

Note:

1. Note that when $\boldsymbol{\mu_e} = \mathbf{1}$, the function $\mathcal{F}(\mathbf{T}; \mathbf{1}; \boldsymbol{\psi})$ is equivalent to the Bethe free energy.

   Catch: The vector $\mathbf{1} \in \mathbb{T}(\mathcal{G})$ only when $\mathcal{G}$ is actually a tree.

2. As with Bethe free energy and BP; the optimizing arguments $\widehat{\mathbf{T}}$ can be taken as approximations to the marginals.

   Advantages: Unique global min. can be found by convex programming.

Rough sketch of proof

- based on ideas from convex analysis and information geometry
- the log partition function is convex; its Legendre dual is the negative entropy function
- the entropy of a pairwise MRF depends *only* on the single-node and pairwise marginals $\mathbf{P} = \{ P_s, P_{st} \}$
- given a tree $\mathcal{T}$ embedded within $\mathcal{G}$, we have:

$$H(\mathbf{P}) \;\leq\; H(\Pi^{\mathcal{T}}(\mathbf{P})) \;=\; \sum_{s \in \mathcal{V}} H_s(P_s) - \sum_{(s,t) \in \mathcal{E}(\mathcal{T})} I_{st}(P_{st})$$

- take convex combinations:

$$H(\mathbf{P}) \;\leq\; \mathbb{E}_{\vec{\boldsymbol{\mu}}}\big[H(\Pi^{\mathcal{T}}(\mathbf{P}))\big] \;=\; \sum_{s \in \mathcal{V}} H_s(P_s) - \sum_{(s,t) \in \mathcal{E}(\mathcal{T})} \mu_{st}\, I_{st}(P_{st})$$

Further remarks on upper bounds

1. Stationary conditions for variational problem (optimal $\widehat{\mathbf{T}}$) are *very similar* to tree-based consistency conditions of TRP/BP.

2. Consider optimizing $\mathcal{F}(\mathbf{T}; \boldsymbol{\mu_e}; \boldsymbol{\psi})$ over both $\mathbf{T} \in \mathbb{C}$ *and* $\boldsymbol{\mu_e} \in \mathbb{T}(\mathcal{G})$. I.e., find the best distribution over spanning trees.

   Facts: Exists a unique global minimum; can be found efficiently (involves solving maximum weight spanning tree problems).

3. Extensions to more advanced approximations (e.g., Kikuchi) by considering distributions over hypertrees of the graph.

Summary

- reparameterization perspective leads to theoretical insights on a hierarchy of approximations (from BP upwards)
  (a) invariance of distribution
  (b) consistency-based characterization of fixed points
  (c) exact expression and computable bounds on the error
- new class of upper bounds on the log partition function based on convex combinations of (hyper)trees

Contact information

Martin Wainwright

`mjwain@mit.edu`

Papers at: http://ssg.mit.edu/group/mjwain/mjwain.shtml

References

[1] S.M. Aji and R.J. McEliece. The generalized distributive law. *IEEE Trans. Info. Theory*, 46:325–343, March 2000.

[2] S. Amari. Differential geometry of curved exponential families — curvatures and information loss. *Annals of Statistics*, 10(2):357–385, 1982.

[3] J. B. Anderson and S. M. Hladnik. Tailbiting map decoders. *IEEE Sel. Areas Comm.*, 16:297–302, February 1998.

[4] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Stat. Soc. Series B*, 36:192–236, 1974.

[5] P. Clifford. Markov random fields in statistics. In G.R. Grimmett and D. J. A. Welsh, editors, *Disorder in physical systems*. Oxford Science Publications, 1990.

[6] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.

[7] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability,*, 3(1):146–158, Feb. 1975.

[8] J. Darroch, S. Lauritzen, and T. Speed. Markov fields and log-linear interaction models for contingency tables. *Annals of Statistics*, 8(3):522–539, 1980.

[9] W. T. Freeman D.J.C. MacKay, J. S. Yedidia and Y. Weiss. A conversation about the Bethe free energy and sum-product algorithm. Technical Report TR2001-18, Mitsubishi Electric Research Labs, May 2001. Available at `http://www.merl.com/papers/TR2001-18/`.

[10] J. Edmonds. Matroids and the greedy algorithm. *Mathematical Programming*, 1:127–136, 1971.

[11] A. El Gamal and T. Cover. Multiple user information theory. *Proceedings of the IEEE*, 68(12):1466–1483, December 1980.

[12] M. P. C. Fossorier. Iterative reliability-based decoding of low-density parity check codes. *IEEE Transactions on Information Theory*, pages 908–917, May 2001.

[13] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *Intl. J. Computer Vision*, 40(1):25–47, 2000.

[14] B. Frey, R. Koetter, and N. Petrovic. Very loopy belief propagation for unwrapping phase images. In *NIPS 14*. MIT Press, 2001. To appear.

[15] R. G. Gallager. *Low-density parity check codes*. MIT Press, Cambridge, MA, 1963.

[16] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Pat. Anal. Mach. Intell.*, 6:721–741, 1984.

[17] G. R. Grimmett. A theorem about random fields. *Bulletin of the London Mathematical Society*, 5:81–84, 1973.

[18] E. Ising. Beitrag zur theorie der ferromagnetismus. *Zeitschrift für Physik*, 31:253–258, 1925.

[19] M. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. In *Learning in graphical models*, pages 105–161. MIT Press, 1999.

[20] F. Kschischang and B. Frey. Iterative decoding of compound codes by probability propagation in graphical models. *IEEE Sel. Areas Comm.*, 16(2):219–230, February 1998.

[21] S. L. Lauritzen. *Graphical models.* Oxford University Press, Oxford, 1996.

[22] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society B*, 50:155–224, January 1988.

[23] M. Luby, M. Mitzenmacher, M. A. Shokrollahi, and D. Spielman. Improved low-density parity-check codes using irregular graphs and belief propagation. In *Proceedings 1998 International Symposium on Information Theory*, page 117. IEEE, 1998.

[24] M. Luby, M. Mitzenmacher, M. A. Shokrollahi, and D. Spielman. Improved low-density parity check codes using irregular graphs. *IEEE Trans. Info. Theory*, 47:585–598, February 2001.

[25] D.J.C. MacKay. Good error-correcting codes based on very sparse matrices. *IEEE Trans. Info. Theory*, 45(2):399–431, 1999.

[26] R.J. McEliece, D.J.C. McKay, and J.F. Cheng. Turbo decoding as an instance of Pearl's belief propagation algorithm. *IEEE Jour. Sel. Communication*, 16(2):140–152, February 1998.

[27] T. P. Minka. *A family of algorithms for approximate Bayesian inference.* PhD thesis, MIT Media Lab, January 2001.

[28] K. Murphy and Y. Weiss. The factored frontier algorithm for approximate inference in DBNs. In *Uncertainty in Artificial Intelligence*, volume 11, 2001.

[29] J. Pearl. *Probabilistic reasoning in intelligent systems*. Morgan Kaufman, San Mateo, 1988.

[30] T. Richardson. The geometry of turbo-decoding dynamics. *IEEE Trans. Info. Theory*, 46(1):9–23, January 2000.

[31] T. Richardson, A. Shokrollahi, and R. Urbanke. Design of capacity-approaching irregular low-density parity check codes. *IEEE Trans. Info. Theory*, 47:619–637, February 2001.

[32] T. Richardson and R. Urbanke. The capacity of low-density parity check codes under message-passing decoding. *IEEE Trans. Info. Theory*, 47:599–618, February 2001.

[33] P. Rusmevichientong and B. Van Roy. An analysis of turbo decoding with Gaussian densities. In *NIPS 12*, pages 575–581. MIT Press, 2000.

[34] R. M. Tanner. A recursive approach to low complexity codes. *IEEE Trans. Info. Theory*, 81(5):533–547, September 1981.

[35] S. Verdu and H. V. Poor. Abstract dynamic programming models under commutativity conditions. *SIAM J. Control and Optimization*, 25(4):990–1006, July 1987.

[36] M. J. Wainwright, T. Jaakkola, and A. S. Willsky. Tree-based reparameterization for approximate estimation on graphs with cycles. LIDS Tech. report P-2510: available at `http://ssg.mit.edu/group/mjwain/mjwain.shtml`, May 2001.

[37] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. Tree-based reparameterization for approximate inference on loopy graphs. In *NIPS 14*. MIT Press, 2002. To appear; Preprint available at http://ssg.mit.edu/group/mjwain/mjwain.shtml.

[38] M.J. Wainwright. *Stochastic processes on graphs with cycles: geometric and variational approaches*. PhD thesis, MIT, Laboratory for Information and Decision Systems, January 2002.

[39] Y. Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Computation*, 12:1–41, 2000.

[40] Y. Weiss and W. T. Freeman. Correctness of belief propagation in Gaussian graphical models of arbitrary topology. In *NIPS 12*, pages 673–679. MIT Press, 2000.

[41] M. Welling and Y. Teh. Belief optimization: A stable alternative to loopy belief propagation. In *Uncertainty in Artificial Intelligence*, July 2001.

[42] J.W. Woods. Markov image modeling. *IEEE Transactions on Automatic Control*, 23:846–850, October 1978.

[43] J. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In *NIPS 13*, pages 689–695. MIT Press, 2001.

[44] A. Yuille. A double-loop algorithm to minimize the Bethe and Kikuchi free energies. *Neural Computation*, To appear, 2001.