Graphical models and message-passing
Part II: Marginals and likelihoods

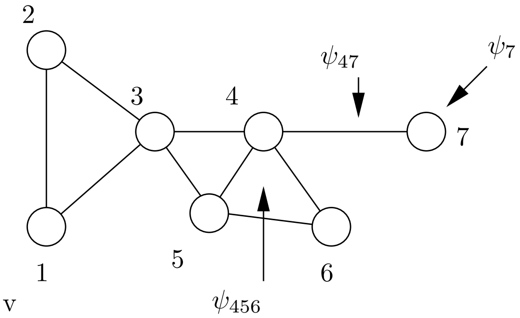
Martin Wainwright

UC Berkeley
Departments of Statistics, and EECS

Tutorial materials (slides, monograph, lecture notes) available at:
www.eecs.berkeley.edu/~wainwrig/kyoto12

September 3, 2012

Graphs and factorization



- clique C is a fully connected subset of vertices
- compatibility function ψ_C defined on variables $x_C = \{x_s, s \in C\}$
- factorization over all cliques

$$p(x_1, \ldots, x_N) = \frac{1}{Z} \prod_{C \in \mathfrak{C}} \psi_C(x_C).$$

Core computational challenges

Given an undirected graphical model (Markov random field):

$$p(x_1, x_2, \ldots, x_N) \quad = \quad \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

How to efficiently compute?

- most probable configuration (MAP estimate):

  Maximize : $$\widehat{x} = \arg \max_{\mathbf{x} \in \mathcal{X}^N} p(x_1, \ldots, x_N) \; = \; \arg \max_{\mathbf{x} \in \mathcal{X}^N} \prod_{C \in \mathcal{C}} \psi_C(x_C).$$

- the data likelihood or normalization constant

  Sum/integrate : $$Z \; = \; \sum_{x \in \mathcal{X}^N} \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

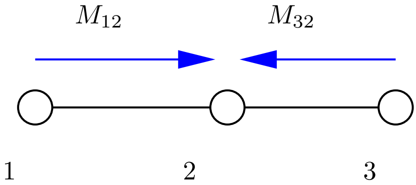
- marginal distributions at single sites, or subsets:

  Sum/integrate : $$p(X_s = x_s) = \frac{1}{Z} \sum_{x_t, \; t \neq s} \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

§1. Sum-product message-passing on trees

Goal: Compute marginal distribution at node u on a tree:

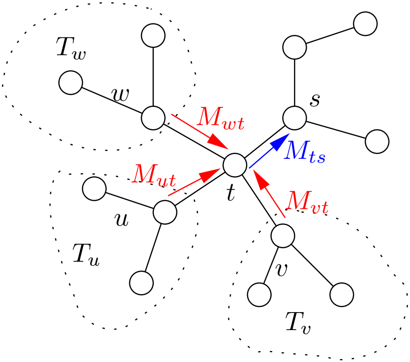
$$\widehat{x} \;=\; \arg\max_{\mathbf{x}\in\mathcal{X}^N} \left\{ \prod_{s\in V} \exp(\theta_s(x_s) \prod_{(s,t)\in E} \exp(\theta_{st}(x_s, x_t)) \right\}.$$



$$\sum_{x_1,x_2,x_3} p(\mathbf{x}) \;=\; \sum_{x_2} \left[\exp(\theta_1(x_1)) \prod_{t\in 1,3} \left\{ \sum_{x_t} \exp[\theta_t(x_t) + \theta_{2t}(x_2, x_t)] \right\} \right]$$

Putting together the pieces

Sum-product is an exact algorithm for any tree.



M_{ts} $\equiv$ message from node t to s

$\mathcal{N}(t)$ $\equiv$ neighbors of node t

Update: $\mathbf{M_{ts}(x_s)} \leftarrow \sum_{x_t' \in \mathcal{X}_t} \Big\{ \exp\Big[\theta_{st}(x_s, x_t') + \theta_t(x_t')\Big] \prod_{v \in \mathcal{N}(t)\backslash s} \mathbf{M_{vt}(x_t)} \Big\}$

Sum-marginals: $p_s(x_s;\theta) \propto \exp\{\theta_s(x_s)\} \prod_{t\in\mathcal{N}(s)} M_{ts}(x_s)$.

Summary: sum-product on trees

- converges in at most graph diameter # of iterations
- updating a single message is an $\mathcal{O}(m^2)$ operation
- overall algorithm requires $\mathcal{O}(Nm^2)$ operations

- upon convergence, yields the exact node and edge marginals:

$$p_s(x_s) \propto e^{\theta_s(x_s)} \prod_{u \in \mathcal{N}(s)} M_{us}(x_s)$$

$$p_{st}(x_s, x_t) \propto e^{\theta_s(x_s) + \theta_t(x_t) + \theta_{st}(x_s, x_t)} \prod_{u \in \mathcal{N}(s)} M_{us}(x_s) \prod_{u \in \mathcal{N}(t)} M_{ut}(x_t)$$

- messages can also be used to compute the partition function

$$Z = \sum_{x_1, \ldots, x_N} \prod_{s \in V} e^{\theta_s(x_s)} \prod_{(s,t) \in E} e^{\theta_{st}(x_s, x_t)}.$$

§2. Sum-product on graph with cycles

- as with max-product, a widely used heuristic with a long history:
  - error-control coding: Gallager, 1963
  - artificial intelligence: Pearl, 1988
  - turbo decoding: Berroux et al., 1993
  - etc..

§2. Sum-product on graph with cycles

- as with max-product, a widely used heuristic with a long history:
  - error-control coding: Gallager, 1963
  - artificial intelligence: Pearl, 1988
  - turbo decoding: Berroux et al., 1993
  - etc..

- some concerns with sum-product with cycles:
  - no convergence guarantees
  - can have multiple fixed points
  - final estimate of Z is not a lower/upper bound

§2. Sum-product on graph with cycles

- as with max-product, a widely used heuristic with a long history:
  - error-control coding: Gallager, 1963
  - artificial intelligence: Pearl, 1988
  - turbo decoding: Berroux et al., 1993
  - etc..

- some concerns with sum-product with cycles:
  - no convergence guarantees
  - can have multiple fixed points
  - final estimate of Z is not a lower/upper bound

- as before, can consider a broader class of reweighted sum-product algorithms

Tree-reweighted sum-product algorithms

Message update from node t to node s:

$$M_{ts}(x_s) \quad \leftarrow \quad \kappa \sum_{x'_t \in \mathcal{X}_t} \left\{ \exp\Big[\underbrace{\frac{\theta_{st}(x_s, x'_t)}{\rho_{st}}}_{\text{reweighted edge}} + \theta_t(x'_t) \Big] \frac{\prod_{v \in \mathcal{N}(t) \backslash s} \overbrace{\big[M_{vt}(x_t)\big]^{\rho_{vt}}}^{\text{reweighted messages}}}{\underbrace{\big[M_{st}(x_t)\big]^{(1-\rho_{ts})}}_{\text{opposite message}}} \right\}.$$

Properties:

1. Modified updates remain *distributed* and *purely local* over the graph.
2. Key differences:
   - Messages are reweighted with $\rho_{st} \in [0, 1]$.
   - Potential on edge (s, t) is rescaled by $\rho_{st} \in [0, 1]$.
   - Update involves the reverse direction edge.
3. The choice $\rho_{st} = 1$ for all edges (s, t) recovers standard update.

Bethe entropy approximation

- define local marginal distributions (e.g., for $m = 3$ states):

$$\mu_s(x_s) \;=\; \begin{bmatrix} \mu_s(0) \\ \mu_s(1) \\ \mu_s(2) \end{bmatrix} \qquad \mu_{st}(x_s, x_t) \;=\; \begin{bmatrix} \mu_{st}(0,0) & \mu_{st}(0,1) & \mu_{st}(0,2) \\ \mu_{st}(1,0) & \mu_{st}(1,1) & \mu_{st}(1,2) \\ \mu_{st}(2,0) & \mu_{st}(2,1) & \mu_{st}(2,2) \end{bmatrix}$$

- define node-based entropy and edge-based mutual information:

$$\text{Node-based entropy:} H_s(\mu_s) = -\sum_{x_s} \mu_s(x_s) \log \mu_s(x_s)$$

$$\text{Mutual information:} I_{st}(\mu_{st}) = \sum_{x_s, x_t} \mu_{st}(x_s, x_t) \log \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}.$$

- ρ-reweighted Bethe entropy

$$H_{\text{Bethe}}(\mu) = \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} \rho_{st} \; I_{st}(\mu_{st}),$$

Bethe entropy is exact for trees

- exact for trees, using the factorization:

$$p(\mathbf{x}; \theta) \quad = \quad \prod_{s \in V} \mu_s(x_s) \prod_{(s,t) \in E} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}$$

Reweighted sum-product and Bethe variational principle

Define the local constraint set

$$\mathbb{L}(G) = \Big\{\tau_s, \tau_{st} \mid \tau \geq 0, \quad \sum_{x_s} \tau_s(x_s) = 1, \quad \sum_{x_t} \tau_{st}(x_s, x_t) = \tau_s(x_s)\Big\}$$

Reweighted sum-product and Bethe variational principle

Define the local constraint set

$$\mathbb{L}(G) = \left\{\tau_s, \tau_{st} \mid \tau \geq 0, \quad \sum_{x_s} \tau_s(x_s) = 1, \quad \sum_{x_t} \tau_{st}(x_s, x_t) = \tau_s(x_s)\right\}$$

Theorem

For any choice of positive edge weights $\rho_{st} > 0$:

(a) *Fixed points of reweighted sum-product are stationary points of the Lagrangian associated with*

$$A_{Bethe}(\theta; \rho) := \max_{\tau \in \mathbb{L}(G)} \Big\{ \sum_{s \in V} \langle \tau_s, \theta_s \rangle + \sum_{(s,t) \in E} \langle \tau_{st}, \theta_{st} \rangle + H_{Bethe}(\tau; \rho) \Big\}.$$

Reweighted sum-product and Bethe variational principle

Define the local constraint set

$$\mathbb{L}(G) = \{\tau_s, \tau_{st} \mid \tau \geq 0, \quad \sum_{x_s} \tau_s(x_s) = 1, \quad \sum_{x_t} \tau_{st}(x_s, x_t) = \tau_s(x_s)\}$$

Theorem

For any choice of positive edge weights $\rho_{st} > 0$:

(a) *Fixed points of reweighted sum-product are stationary points of the Lagrangian associated with*

$$A_{Bethe}(\theta; \rho) := \max_{\tau \in \mathbb{L}(G)} \Big\{ \sum_{s \in V} \langle \tau_s, \theta_s \rangle + \sum_{(s,t) \in E} \langle \tau_{st}, \theta_{st} \rangle + H_{Bethe}(\tau; \rho) \Big\}.$$

(b) *For valid choices of edge weights $\{\rho_{st}\}$, the fixed points are unique and moreover $\log Z(\theta) \leq A_{Bethe}(\theta; \rho)$. In addition, reweighted sum-product converges with appropriate scheduling.*

Lagrangian derivation of ordinary sum-product

- let's try to solve this problem by a (partial) Lagrangian formulation
- assign a Lagrange multiplier $\lambda_{ts}(x_s)$ for each constraint $C_{ts}(x_s) := \tau_s(x_s) - \sum_{x_t} \tau_{st}(x_s, x_t) = 0$
- will enforce the normalization ($\sum_{x_s} \tau_s(x_s) = 1$) and non-negativity constraints explicitly
- the Lagrangian takes the form:

$$\mathcal{L}(\tau; \lambda) = \langle \theta, \tau \rangle + \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) + \sum_{(s,t) \in E} \Big[\sum_{x_t} \lambda_{st}(x_t) C_{st}(x_t) + \sum_{x_s} \lambda_{ts}(x_s) C_{ts}(x_s) \Big]$$

Lagrangian derivation (part II)

- taking derivatives of the Lagrangian w.r.t τ_s and τ_{st} yields

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \tau_s(x_s)} &= \theta_s(x_s) - \log \tau_s(x_s) + \sum_{t \in \mathcal{N}(s)} \lambda_{ts}(x_s) + C \\
\frac{\partial \mathcal{L}}{\partial \tau_{st}(x_s, x_t)} &= \theta_{st}(x_s, x_t) - \log \frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s)\tau_t(x_t)} - \lambda_{ts}(x_s) - \lambda_{st}(x_t) + C'
\end{aligned}
$$

- setting these partial derivatives to zero and simplifying:

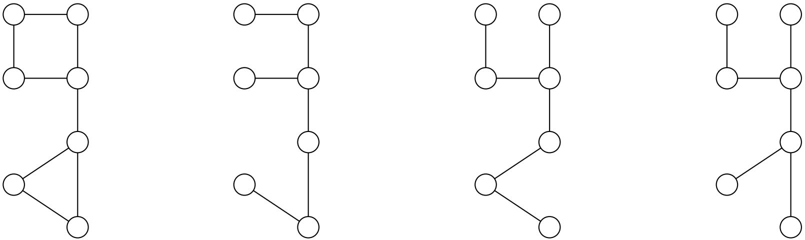
$$
\begin{aligned}
\tau_s(x_s) &\propto \exp\left\{\theta_s(x_s)\right\} \prod_{t \in \mathcal{N}(s)} \exp\left\{\lambda_{ts}(x_s)\right\} \\
\tau_s(x_s, x_t) &\propto \exp\left\{\theta_s(x_s) + \theta_t(x_t) + \theta_{st}(x_s, x_t)\right\} \times \\
& \qquad \prod_{u \in \mathcal{N}(s) \backslash t} \exp\left\{\lambda_{us}(x_s)\right\} \prod_{v \in \mathcal{N}(t) \backslash s} \exp\left\{\lambda_{vt}(x_t)\right\}
\end{aligned}
$$

- enforcing the constraint $C_{ts}(x_s) = 0$ on these representations yields the familiar update rule for the *messages* $M_{ts}(x_s) = \exp(\lambda_{ts}(x_s))$:

$$
M_{ts}(x_s) \leftarrow \sum_{x_t} \exp\left\{\theta_t(x_t) + \theta_{st}(x_s, x_t)\right\} \prod_{u \in \mathcal{N}(t) \backslash s} M_{ut}(x_t)
$$

Convex combinations of trees

Idea: Upper bound $A(\theta) := \log Z(\theta)$ with a convex combination of tree-structured problems.



$$\begin{array}{ccccccc} \theta & = & \rho(T^1)\theta(T^1) & + & \rho(T^2)\theta(T^2) & + & \rho(T^3)\theta(T^3) \\ A(\theta) & \leq & \rho(T^1)A(\theta(T^1)) & + & \rho(T^2)A(\theta(T^2)) & + & \rho(T^3)A(\theta(T^3)) \end{array}$$

$$\begin{array}{lcl} \boldsymbol{\rho} = \{\rho(T)\} & \equiv & \text{probability distribution over spanning trees} \\ \theta(T) & \equiv & \text{tree-structured parameter vector} \end{array}$$

Finding the tightest upper bound

Observation: For each fixed distribution $\boldsymbol{\rho}$ over spanning trees, there are many such upper bounds.

Goal: Find the tightest such upper bound over all trees.

Challenge: Number of spanning trees grows rapidly in graph size.

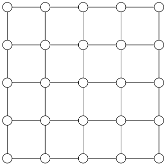
Finding the tightest upper bound

Observation: For each fixed distribution $\boldsymbol{\rho}$ over spanning trees, there are many such upper bounds.

Goal: Find the tightest such upper bound over all trees.

Challenge: Number of spanning trees grows rapidly in graph size.

Example:
On the 2-D lattice:



Grid size	# trees
9	192
16	100352
36	3.26×10^{13}
100	5.69×10^{42}

Finding the tightest upper bound

Observation: For each fixed distribution $\boldsymbol{\rho}$ over spanning trees, there are many such upper bounds.

Goal: Find the tightest such upper bound over all trees.

Challenge: Number of spanning trees grows rapidly in graph size.

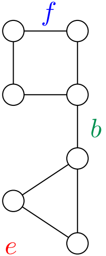
By a suitable dual reformulation, problem can be avoided:

Key duality relation:

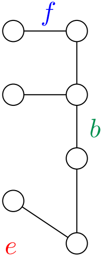
$$\min_{\sum_T \rho(T)\theta(T)=\theta} \rho(T) A(\theta(T)) = \max_{\mu \in \mathbb{L}(G)} \big\{ \langle \mu, \, \theta \rangle + H_{\text{Bethe}}(\mu; \rho_{st}) \big\}.$$

Edge appearance probabilities

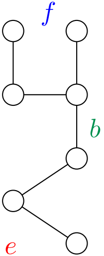
Experiment: What is the probability ρ_e that a given edge $e \in E$ belongs to a tree T drawn randomly under $\boldsymbol{\rho}$?



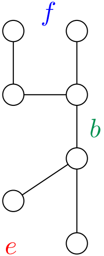
(a) Original



(b) $\rho(T^1) = \frac{1}{3}$



(c) $\rho(T^2) = \frac{1}{3}$



(d) $\rho(T^3) = \frac{1}{3}$

In this example: $\rho_b = 1$; $\rho_e = \frac{2}{3}$; $\rho_f = \frac{1}{3}$.

The vector $\boldsymbol{\rho_e} = \{ \rho_e \mid e \in E \}$ must belong to the *spanning tree polytope.* (Edmonds, 1971)

Why does entropy arise in the duality?

Due to a deep correspondence between two problems:

Maximum entropy density estimation

Maximize entropy $\qquad H(p) = -\sum_{\mathbf{x}} p(x_1, \ldots, x_N) \log p(x_1, \ldots, x_N)$

subject to expectation constraints of the form

$$\sum_{\mathbf{x}} p(\mathbf{x}) \phi_\alpha(\mathbf{x}) = \widehat{\mu}_\alpha.$$

Why does entropy arise in the duality?

Due to a deep correspondence between two problems:

Maximum entropy density estimation

Maximize entropy $\quad H(p) = -\sum_{\mathbf{x}} p(x_1, \ldots, x_N) \log p(x_1, \ldots, x_N)$

subject to expectation constraints of the form

$$\sum_{\mathbf{x}} p(\mathbf{x}) \phi_\alpha(\mathbf{x}) = \widehat{\mu}_\alpha.$$

Maximum likelihood in exponential family

Maximize likelihood of parameterized densities

$$p(x_1, \ldots, x_N; \theta) = \exp\Big\{\sum_\alpha \theta_\alpha \phi_\alpha(x) - A(\theta)\Big\}.$$

Conjugate dual functions

- conjugate duality is a fertile source of variational representations
- any function f can be used to define another function f^* as follows:

$$f^*(v) \quad := \quad \sup_{u \in \mathbb{R}^n} \big\{ \langle v,\, u \rangle - f(u) \big\}.$$

- easy to show that f^* is always a convex function
- how about taking the "dual of the dual"? I.e., what is $(f^*)^*$?
- when f is well-behaved (convex and lower semi-continuous), we have $(f^*)^* = f$, or alternatively stated:

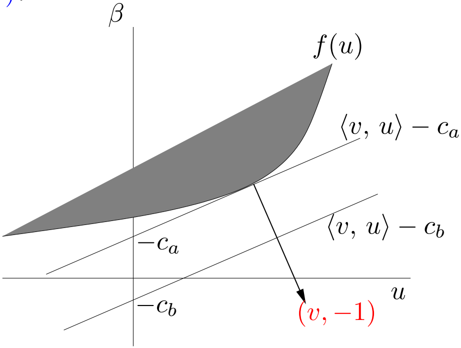
$$f(u) \quad = \quad \sup_{v \in \mathbb{R}^n} \big\{ \langle u,\, v \rangle - f^*(v) \big\}$$

Geometric view: Supporting hyperplanes

Question: Given all hyperplanes in $\mathbb{R}^n \times \mathbb{R}$ with normal $(v, -1)$, what is the intercept of the one that supports $\mathrm{epi}(f)$?

Epigraph of f:

$\mathrm{epi}(f) := \{(u, \beta) \in \mathbb{R}^{n+1} \mid f(u) \leq \beta\}$.



Analytically, we require the smallest $c \in \mathbb{R}$ such that:

$$\langle v,\, u\rangle - c \quad \leq \quad f(u) \quad \text{for all} \quad u \in \mathbb{R}^n$$

By re-arranging, we find that this optimal c^* is the dual value:

$$c^* \quad = \quad \sup_{u \in \mathbb{R}^n} \big\{\langle v,\, u\rangle - f(u)\big\}.$$

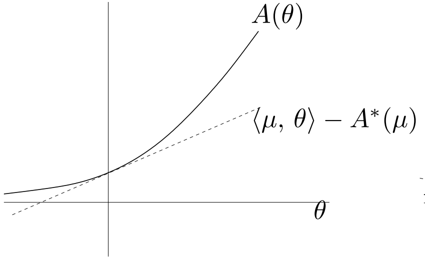
Example: Single Bernoulli

Random variable $X \in \{0, 1\}$ yields exponential family of the form:

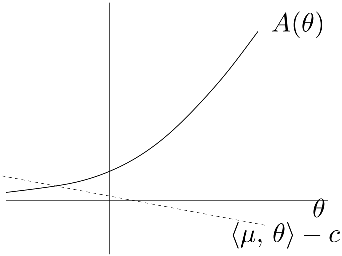
$$p(x;\theta) \propto \exp\big\{\theta\, x\big\} \quad \text{with} \quad A(\theta) = \log\big[1 + \exp(\theta)\big].$$

Let's compute the dual $A^*(\mu) := \sup_{\theta \in \mathbb{R}} \big\{\mu\theta - \log[1 + \exp(\theta)]\big\}$.

(Possible) stationary point: $\mu = \exp(\theta)/[1 + \exp(\theta)]$.



(a) Epigraph supported

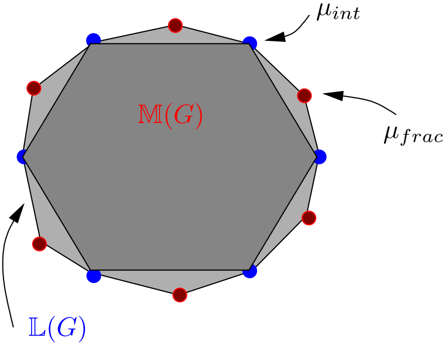


(b) Epigraph *cannot* be supported

We find that: $$A^*(\mu) = \begin{cases} \mu \log \mu + (1 - \mu) \log(1 - \mu) & \text{if } \mu \in [0, 1] \\ +\infty & \text{otherwise.} \end{cases}$$

Leads to the variational representation: $A(\theta) = \max_{\mu \in [0,1]} \big\{\mu \cdot \theta - A^*(\mu)\big\}$.

Geometry of Bethe variational problem



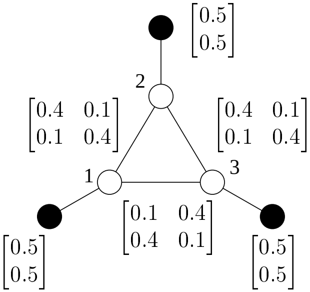
- belief propagation uses a *polyhedral outer approximation* to $\mathbb{M}(G)$:
  - for any graph, $\mathbb{L}(G) \supseteq \mathbb{M}(G)$.
  - equality holds $\iff$ G is a tree.

Natural question: Do BP fixed points ever fall outside of the marginal polytope $\mathbb{M}(G)$?

Illustration: Globally inconsistent BP fixed points

Consider the following assignment of pseudomarginals τ_s, τ_{st}:

Locally consistent (pseudo)marginals



- can verify that $\tau \in \mathbb{L}(G)$, and that τ is a fixed point of belief propagation (with all constant messages)
- however, τ is globally inconsistent

Note: More generally: for any τ in the interior of $\mathbb{L}(G)$, can construct a distribution with τ as a BP fixed point.

High-level perspective: A broad class of methods

- message-passing algorithms (e.g., mean field, belief propagation) are solving approximate versions of exact variational principle in exponential families
- there are two *distinct* components to approximations:
  (a) can use either inner or outer bounds to $\mathbb{M}$
  (b) various approximations to entropy function $-A^*(\mu)$

Refining one or both components yields better approximations:

- BP: polyhedral outer bound and non-convex Bethe approximation
- Kikuchi and variants: tighter polyhedral outer bounds and better entropy approximations (e.g.,Yedidia et al., 2002)
- Expectation-propagation: better outer bounds and Bethe-like entropy approximations (Minka, 2002)