Marginal polytopes of graphical models: Linear programs, max-product, and variational relaxation

Martin Wainwright
Department of Statistics
Department of Electrical Engineering and Computer Science
UC Berkeley, CA
Email: `wainwrig@{stat,eecs}.berkeley.edu`

Based on joint works with:

Tommi Jaakkola, Alan Willsky (MIT)

Michael Jordan (Univ. California, Berkeley)

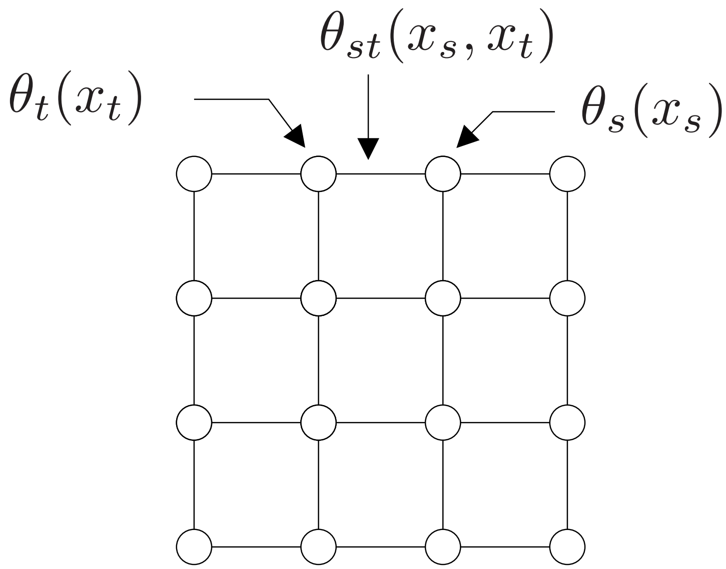
Vladimir Kolmogorov (Univ. College London)

Alekh Agarwal, Pradeep Ravikumar (Univ. California, Berkeley)

Introduction

- **max/sum-product message-passing:**
  - "divide and conquer": based on factorization/Markov properties
  - exact for decomposable; approximate for general graphs
  - now standard in various fields (e.g., statistics, statistical machine learning, statistical physics, computer vision, computational biology....)
- **convex relaxations (LP, SOCP, SDP etc.):**
  - "relax" a hard combinatorial problem into a simple convex one
  - standard method in computer science, operations research, polyhedral combinatorics
- notion of **marginal polytope**:
  - geometric object associated with any undirected graphical model
  - complexity critically determined by graph topology
  - yields fruitful connections between message-passing and LP relaxation

MAP optimization in undirected graphical models



- undirected graph $G = (V, E)$
- $X_s \equiv$ random variable at node s taking values $x_s \in \mathcal{X}_s$
- $\theta_s(x_s) \equiv$ observation term
- $\theta_{st}(x_s, x_t) \equiv$ coupling term

- overall distribution decomposes additively on graph cliques:

$$p(\mathbf{x}; \theta) \quad \propto \quad \exp\Big\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \Big\}$$

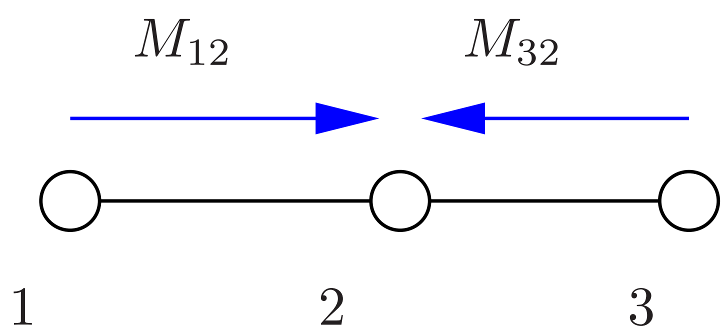
- mode or maximum a posteriori (MAP) estimate:

$$\widehat{\mathbf{x}} \quad \in \arg\max_{\mathbf{x} \in \mathcal{X}^N} \Big\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \Big\}.$$

Max-product on trees

Goal: Compute most probable configuration on a tree:

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

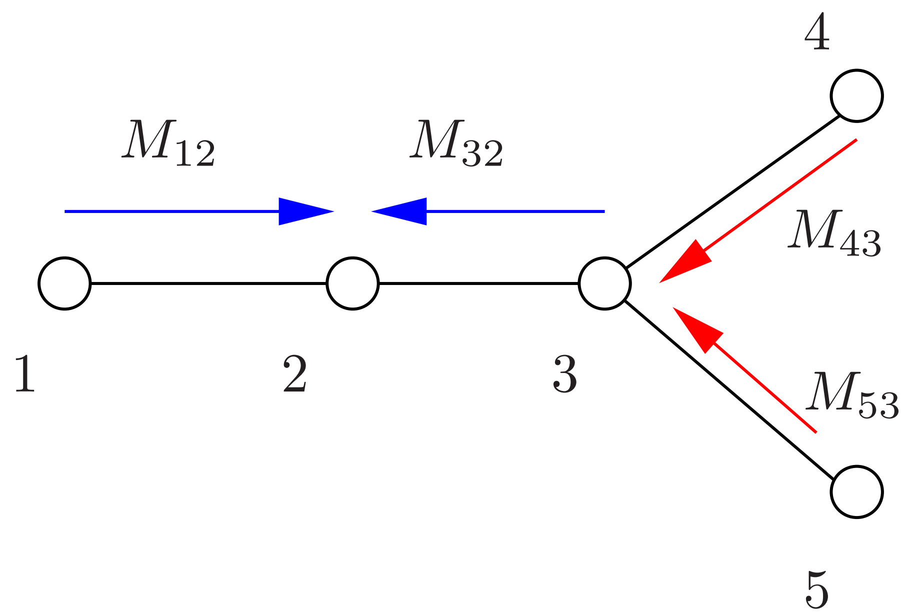


$$\max_{x_1,x_2,x_3} p(\mathbf{x}) \;=\; \max_{x_1} \left[\exp(\theta_1(x_1)) \prod_{t\in\{1,3\}} \left\{ \max_{x_t} \exp[\theta_t(x_t) + \theta_{2t}(x_2, x_t)] \right\} \right]$$

Max-product strategy: "Divide and conquer": break global maximization into simpler sub-problems. (Lauritzen & Spiegelhalter, 1988; Dawid, 1992)

Max-product recursions

Decompose: $\max_{x_1, x_2, x_3, x_4, x_5} p(\mathbf{x}) = \max_{x_1} \Big[\exp(\theta_1(x_1)) \prod_{t \in N(2)} M_{t2}(x_2) \Big].$



Update messages:

$$M_{32}(x_3, x_2) = \max_{x_3} \left[\exp(\theta_3(x_3) + \theta_{23}(x_2, x_3)) \prod_{v \in N(3) \backslash 2} M_{v3}(x_3) \right]$$

Variational view: Max-product and linear programs

- MAP as integer program: $f^* = \max_{\mathbf{x} \in \mathcal{X}^N} \Big\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \Big\}$

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

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

- alternative formulation of MAP as linear program

$$g^* = \max_{(\mu_s, \mu_{st}) \in \mathbb{M}(T)} \Big\{ \sum_{s \in V} \mathbb{E}_{\mu_s}[\theta_s(x_s)] + \sum_{(s,t) \in E} \mathbb{E}_{\mu_{st}}[\theta_{st}(x_s, x_t)] \Big\}$$

Local expectations: $\mathbb{E}_{\mu_s}[\theta_s(x_s)] := \sum_{x_s} \mu_s(x_s) \theta_s(x_s).$

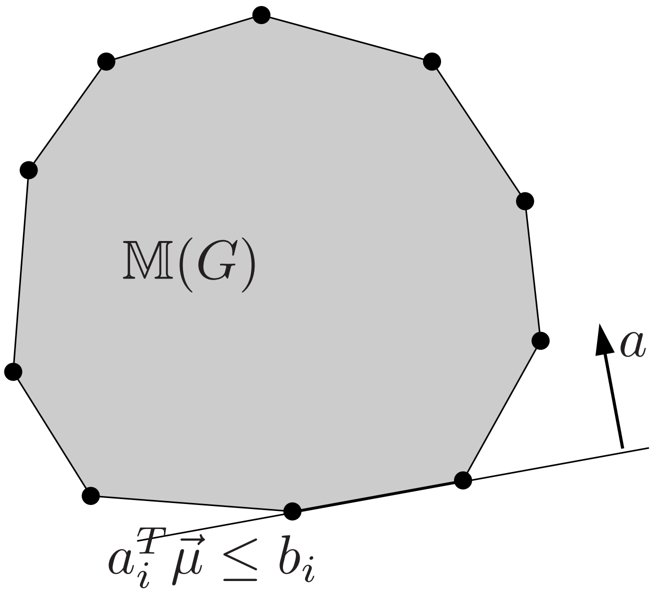
Key question: What constraints must local marginals $\{\mu_s, \mu_{st}\}$ satisfy?

Marginal polytopes for general undirected models

- $\mathbb{M}(G) \equiv$ set of all *globally realizable* marginals $\{\mu_s, \mu_{st}\}$:

$$\left\{ \vec{\mu} \in \mathbb{R}^{m^N} \;\middle|\; \mu_s(x_s) = \sum_{x_t, t \neq s} p_\mu(\mathbf{x}), \text{ and } \mu_{st}(x_s, x_t) = \sum_{x_u, u \neq s,t} p_\mu(\mathbf{x}) \right\}$$

for some $p_\mu(\cdot)$ over $(X_1, \ldots, X_N) \in \{0, 1, \ldots, m-1\}^N$.



- polytope in $m|V| + m^2|E|$ dimensions (m per vertex, m^2 per edge)
- with m^N vertices
- number of facets?

Marginal polytope for trees

- $\mathbb{M}(T) \equiv$ special case of marginal polytope for tree T
- local marginal distributions on nodes/edges (e.g., $m = 3$)

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

Consequence of junction tree theorem: If $\{\mu_s, \mu_{st}\}$ are non-negative and *locally consistent*:

$$\begin{aligned} \textbf{Normalization}: \quad \sum_{x_s} \mu_s(x_s) &= 1 \\ \textbf{Marginalization}: \quad \sum_{x'_t} \mu_{st}(x_s, x'_t) &= \mu_s(x_s), \end{aligned}$$

then on any tree-structured graph T, they are *globally consistent*.

(Lauritzen & Spiegelhalter, 1988)

Max-product on trees: Linear program solver

- MAP problem as a simple linear program:

$$f(\widehat{\mathbf{x}}) \quad = \quad \arg \max_{\vec{\mu} \in \mathbb{M}(T)} \left\{ \sum_{s \in V} \mathbb{E}_{\mu_s}[\theta_s(x_s)] + \sum_{(s,t) \in E} \mathbb{E}_{\mu_{st}}[\theta_{st}(x_s, x_t)] \right\}$$

subject to $\vec{\mu}$ in tree marginal polytope:

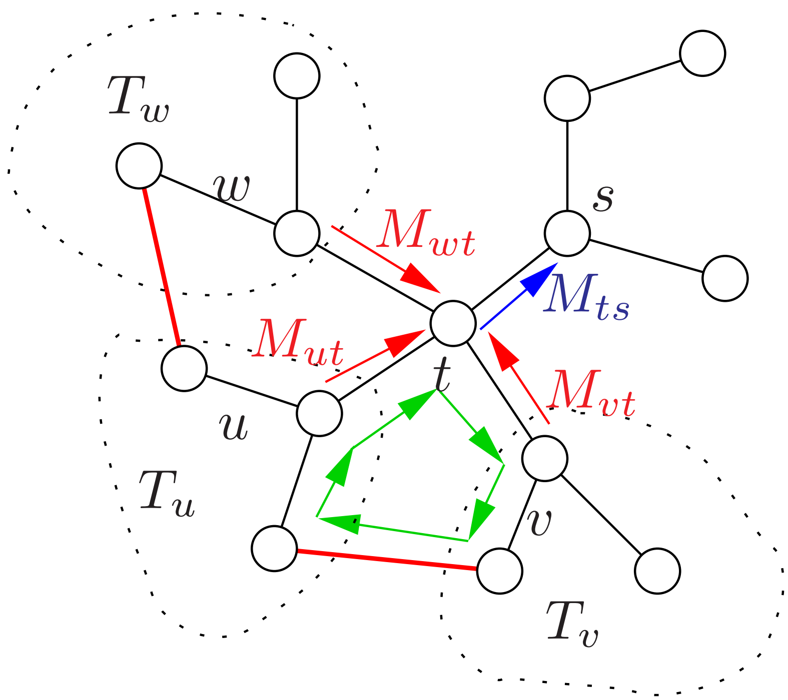
$$\mathbb{M}(T) \;=\; \left\{ \vec{\mu} \geq 0, \quad \sum_{x_s} \mu_s(x_s) \;=\; 1, \qquad \sum_{x'_t} \mu_{st}(x_s, x'_t) \;=\; \mu_s(x_s) \right\}.$$

Max-product and LP solving:

- on tree-structured graphs, max-product is a dual algorithm for solving the tree LP. (Wai. & Jordan, 2003)
- max-product message $M_{ts}(x_s) \equiv$ Lagrange multiplier for enforcing the constraint $\sum_{x'_t} \mu_{st}(x_s, x'_t) = \mu_s(x_s)$.

Standard message-passing algorithms: With cycles

Exact for trees, but approximate for graphs with cycles.



M_{ts}	$\equiv$	message from node t to s
$\mathcal{N}(t)$	$\equiv$	neighbors of node t

<u>Sum-product:</u> for marginals

<u>Max-product:</u> for modes

<u>Update:</u> $\mathbf{M_{ts}(x_s)} \leftarrow \max_{x'_t \in \mathcal{X}_t} \left\{ \exp\left[\theta_{st}(x_s, x'_t) + \theta_t(x'_t)\right] \prod_{v \in \mathcal{N}(t) \backslash s} \mathbf{M_{vt}(x_t)} \right\}$.

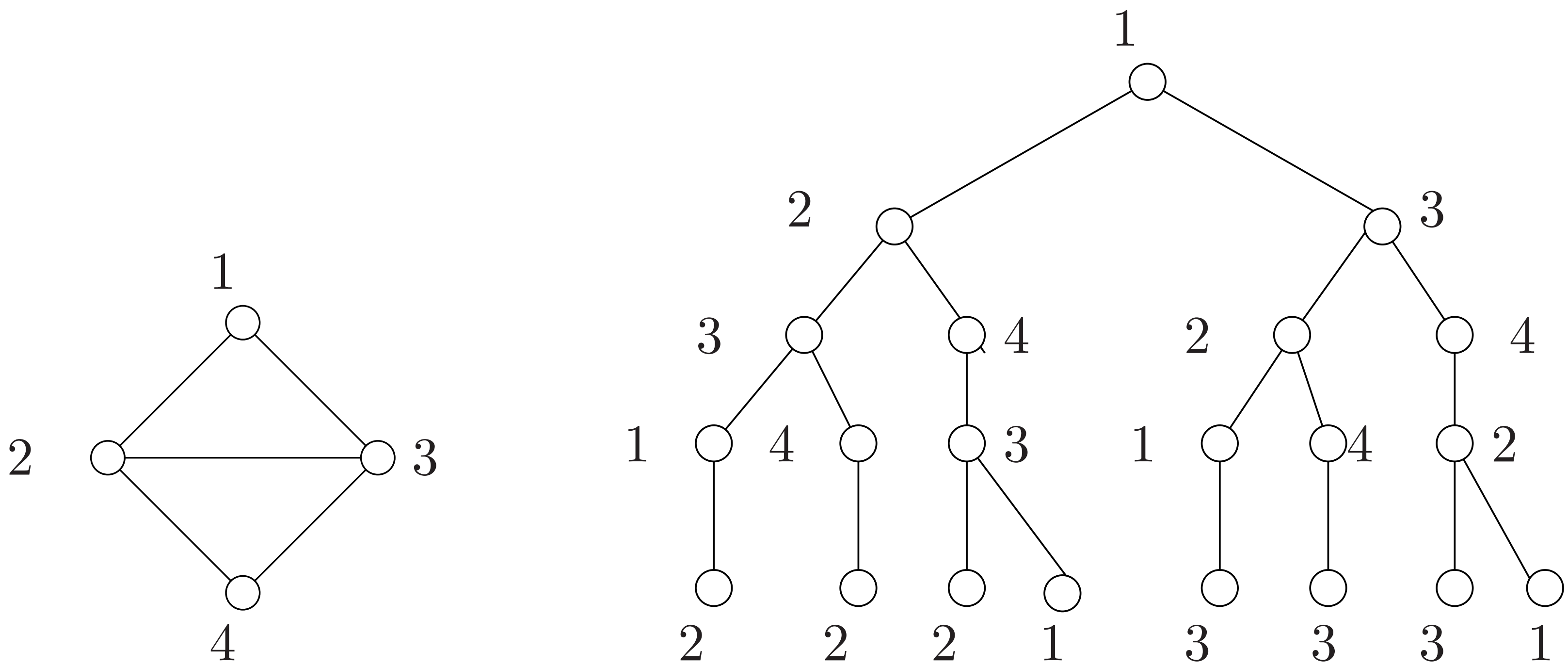
Question: What does max-product compute on a graph with cycles?

Some previous theory on ordinary max-product

- optimal for trees, and junction trees (Lauritzen & Spiegelhalter, 1988; Pearl, 1988; Dawid, 1992)
- analysis of graphs with large girth (Gallager, 1963; many others from 1990s onwards)
- single-cycle graphs (Aji & McEliece, 1998; Horn, 1999; Weiss, 1998)
- existence of fixed points for positive couplings (Wainwright et al., 2003)
- local optimality guarantees:
  - "tree-plus-loop" neighborhoods (Weiss & Freeman, 2001)
  - strengthened optimality results and computable error bounds (Wainwright et al., 2003)
- some exactness results for particular types of matching problems (Bayati et al., 2006, 2008; Jebara & Huang, 2007; Sanghavi, 2008)

Standard analysis via computation tree

- standard tool: computation tree of message-passing updates (Gallager, 1963)



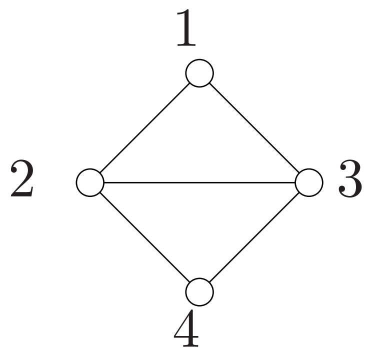
(a) Original graph (b) Computation tree (4 iterations)

- level t of tree: all nodes whose messages reach the root (node 1) after t iterations of message-passing

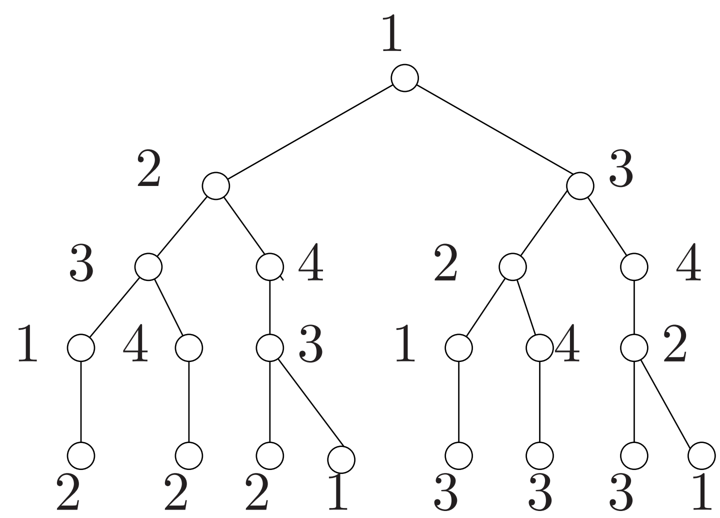
Illustration: Non-exactness of standard max-product

Intuition:

- max-product solves (exactly) modified problem on computation tree
- edge/nodes *not equally weighted* ⇒ incorrectness of max-product



(a) Diamond graph G_{dia}

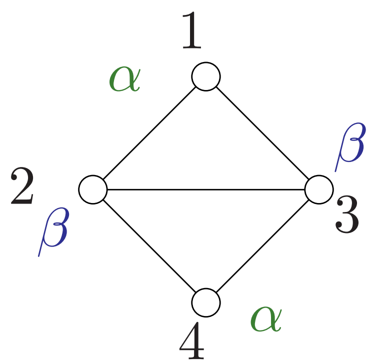


(b) Computation tree (4 iterations)

- for example: asymptotic node fractions in this computation tree:

$$\begin{bmatrix} f(1) & f(2) & f(3) & f(4) \end{bmatrix} = \begin{bmatrix} 0.2393 & 0.2607 & 0.2607 & 0.2393 \end{bmatrix}$$

A whole family of non-exact examples



$$\theta_s(x_s) \quad \begin{cases} \alpha x_s & \text{if } s = 1 \text{ or } s = 4 \\ \beta x_s & \text{if } s = 2 \text{ or } s = 3 \end{cases}$$

$$\theta_{st}(x_s, x_t) = \begin{cases} -\gamma & \text{if } x_s \neq x_t \\ 0 & \text{otherwise} \end{cases}$$

- for γ sufficiently large, optimal solution is always either $1^4 = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}$ or $(-1)^4 = \begin{bmatrix} (-1) & (-1) & (-1) & (-1) \end{bmatrix}$

- max-product and optimal decision based on *different* boundaries:

Optimal boundary: $\quad \widehat{\mathbf{x}} = \begin{cases} 1^4 & \text{if } 0.25\alpha + 0.25\beta \geq 0 \\ (-1)^4 & \text{otherwise} \end{cases}$

Max-product boundary: $\quad \widehat{\mathbf{x}} = \begin{cases} 1^4 & \text{if } 0.2393\alpha + 0.2607\beta \geq 0 \\ (-1)^4 & \text{otherwise} \end{cases}$

Tree-reweighted max-product algorithm

Message update from node t to node s:

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

Properties:

1. Modified updates remain *distributed* and *purely local* over the graph.

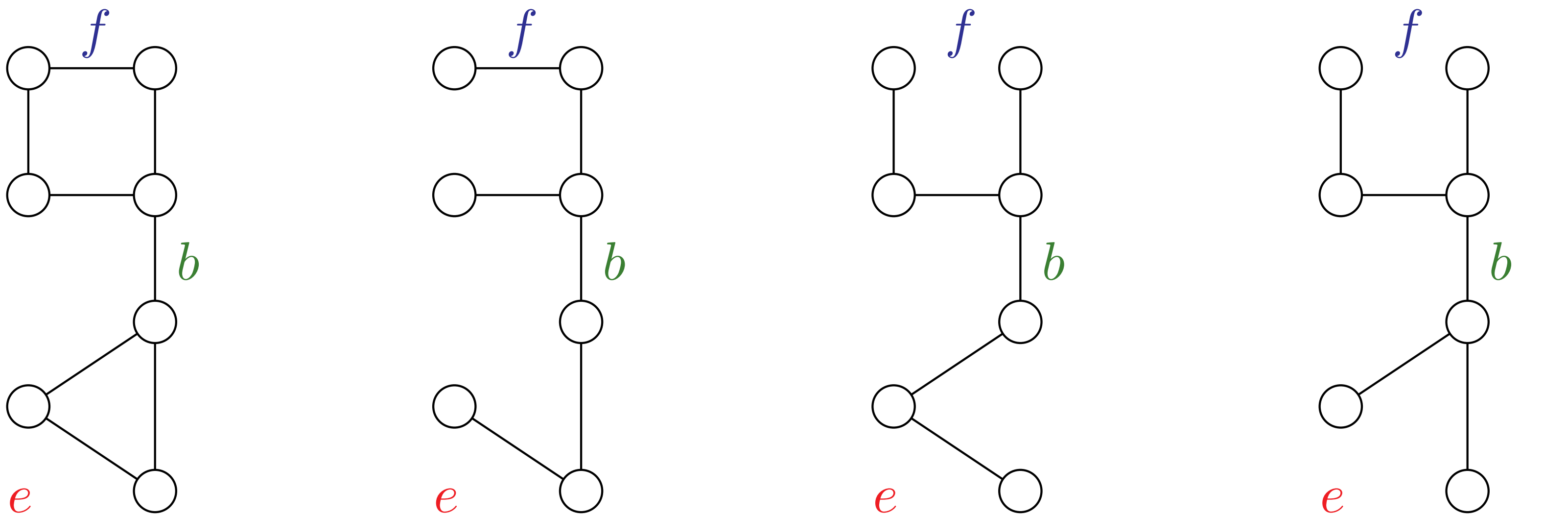
2. Key differences:
   - Messages are reweighted with $\rho_{st} \in [0, 1]$.
   - Potential on edge (s, t) is rescaled by $\rho_{st} \in [0, 1]$.
   - Update involves the reverse direction edge.

3. The choice $\rho_{st} = 1$ for all edges (s, t) recovers standard update.

(Wainwright, Jaakkola & Willsky, 2002)

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*, denoted $\mathbb{T}(G)$. (Edmonds, 1971)

TRW max-product does not lie

- from message fixed point M^*, compute *pseudo-max-marginals* associated with vertex s,

$$\nu_s(x_s) \quad = \quad \exp(\theta_s(x_s)) \prod_{t \in N(s)} [M^*_{ts}(x_s)]^{\rho_{ts}},$$

  and similar quantity for edge (s, t).

- say strong tree agreement holds if there exists a configuration $\mathbf{x}^*$ such that:

$$\begin{aligned} x^*_s &\in \arg\max_{x_s} \nu_s(x_s) \qquad \text{for all } s \in V \\ (x^*_s, x^*_t) &\in \arg\max_{x_s, x_t} \nu_{st}(x_s, x_t) \qquad \text{for all } (s,t) \in E. \end{aligned}$$

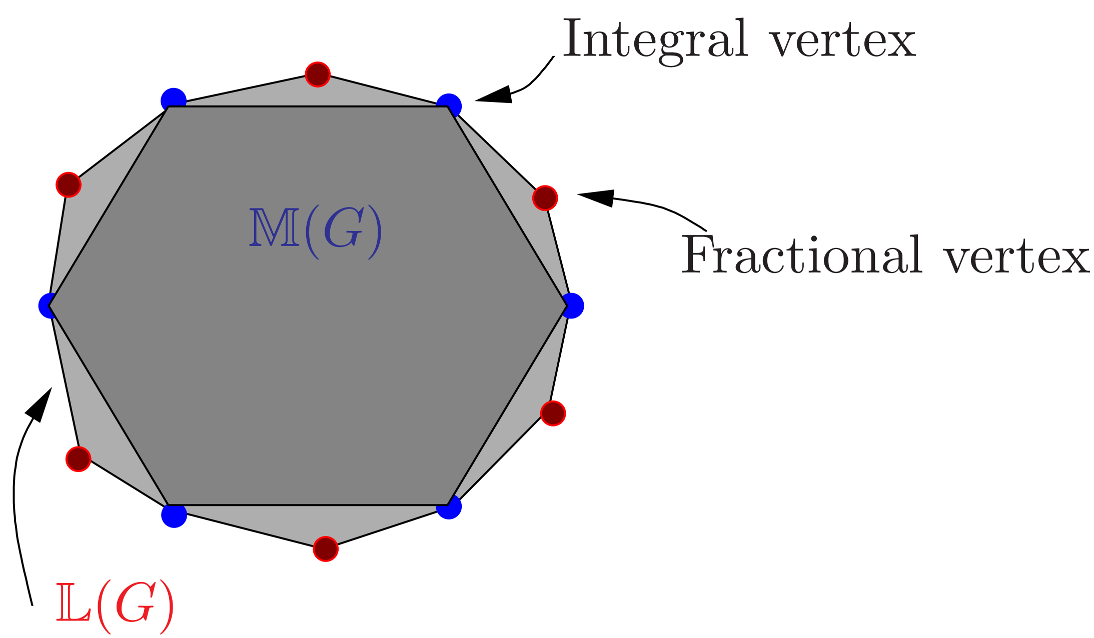
Theorem: For any fixed point M^* any STA configuration $\mathbf{x}^*$ is a mode (most probable configuration) on the full graph G.
(WaiJaaWil05)

- sharp contrast to ordinary max-product, which does lie

Tree-based relaxation for graphs with cycles

Set of *locally consistent pseudomarginals* for general graph G:

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



Key: For a general graph, $\mathbb{L}(G)$ is an outer bound on $\mathbb{M}(G)$, and yields a *linear-programming relaxation* of the MAP problem:

$$f(\widehat{\mathbf{x}}) = \max_{\vec{\mu} \in \mathbb{M}(G)} \theta^T \vec{\mu} \leq \max_{\vec{\tau} \in \mathbb{L}(G)} \theta^T \vec{\tau}.$$

TRW max-product and LP relaxation

First-order (tree-based) LP relaxation:

$$f(\widehat{\mathbf{x}}) \;\leq\; \max_{\vec{\tau} \in \mathbb{L}(G)} \left\{ \sum_{s \in V} \mathbb{E}_{\tau_s}[\theta_s(x_s)] + \sum_{(s,t) \in E} \mathbb{E}_{\tau_{st}}[\theta_{st}(x_s, x_t)] \right\}$$

Theorem: (WaiJaaWil05; Kolmogorov & Wainwright, 2005):

(a) **Strong tree agreement** Any TRW fixed-point that satisfies the strong tree agreement condition specifies an optimal LP solution.

(b) **LP solving:** For any binary pairwise problem, TRW max-product solves the first-order LP relaxation.

(c) **Persistence for binary problems:** Let $S \subseteq V$ be the subset of vertices for which there exists a single point $x_s^* \in \arg\max_{x_s} \nu_s^*(x_s)$. Then for *any optimal solution*, it holds that $y_s = x_s^*$.

Basic idea: convex combinations of trees

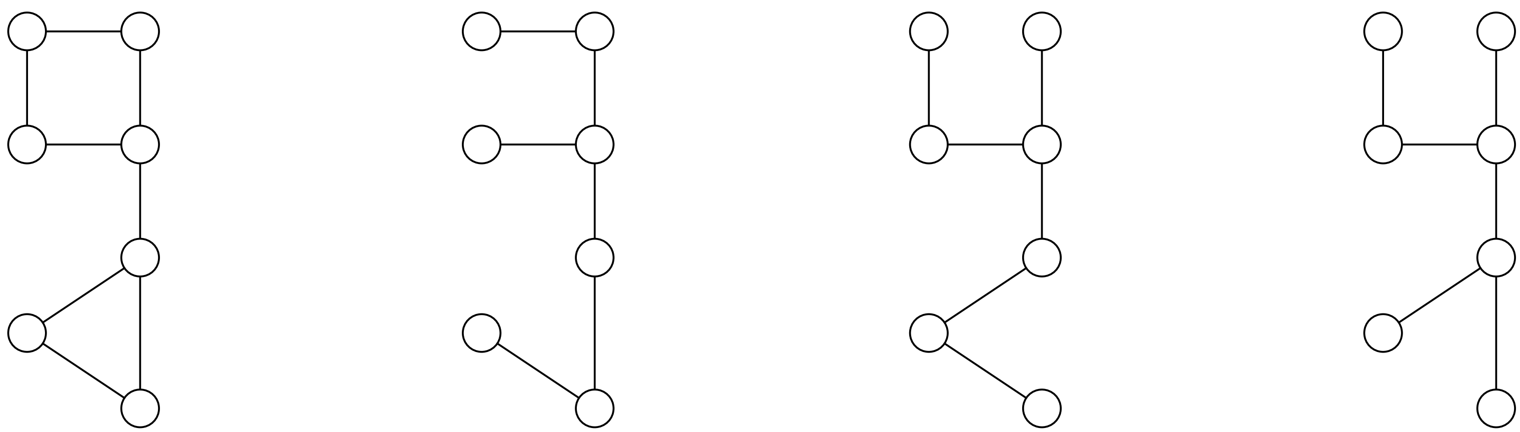
Observation: Easy to find its MAP-optimal configurations on trees:

$$\text{OPT}(\theta(T)) \quad := \quad \big\{\mathbf{x} \in \mathcal{X}^n \mid \mathbf{x} \text{ is MAP-optimal for } p(\mathbf{x}; \theta(T))\big\}.$$

Idea: Approximate original problem by a convex combination of trees.

$\boldsymbol{\rho} = \{\rho(T)\} \quad \equiv \quad$ probability distribution over spanning trees

$\theta(T) \quad \equiv \quad$ tree-structured parameter vector



$$* \quad \theta^* \quad = \quad \rho(T^1)\theta(T^1) \quad + \quad \rho(T^2)\theta(T^2) \quad + \quad \rho(T^3)\theta(T^3)$$

$$\dagger \quad \text{OPT}(\theta^*) \quad \supseteq \quad \text{OPT}(\theta(T^1)) \quad \cap \quad \text{OPT}(\theta(T^2)) \quad \cap \quad \text{OPT}(\theta(T^3)).$$

Dual perspective: linear programming relaxation

- Upper bound maintained by reweighted message-passing:

$$\max_{\mathbf{x} \in \mathcal{X}^N} \langle \theta^*, \, \phi(\mathbf{x}) \rangle \quad \leq \quad \sum_{T \in \mathfrak{T}} \rho(T) \max_{\mathbf{x} \in \mathcal{X}^N} \langle \theta(T), \, \phi(\mathbf{x}) \rangle$$

- Dual of finding optimal upper bound $\equiv$ tree-based LP relaxation:

$$\max_{\mathbf{x} \in \mathcal{X}^N} \langle \theta^*, \, \phi(\mathbf{x}) \rangle \quad \leq \quad \max_{\mu \in \mathrm{LOCAL}(G)} \langle \mu, \, \phi(\mathbf{x}) \rangle$$

- TRW-MP algorithm fixed points specify LP optimum:
  - whenever strong tree agreement holds (WaiJaaWil05)
  - for any binary problem (KolWai05)
  - ....but TRW-MP does not solve LP in general (Kol05)

Various connections and extensions

- max-sum diffusion framework (Schlesinger et al., 1960s, 70s; Werner, 2007)
- binary QPs and roof duality: equivalent to relaxation using $\mathbb{L}(G)$ (Hammer et al., 1984; Boros et al., 1990)
- hierarchy of LP relaxations based on treewidth:

$$\mathbb{M}(G) = \mathbb{L}_t(G) \subset \mathbb{L}_{t-1}(G) \subset \ldots \subset \mathbb{L}_1(G)$$

- treewidth hierarchy: equivalent to Boros et al. (1990) and Sherali-Adams (1990) hierarchies for binary problems (WaiJor04)
- other approaches with links to first-order $\mathbb{L}(G)$ LP relaxation:
  - sequential TRW and conv. guarantees (Kolmogorov, 2005)
  - convex free energies (Weiss et al., 2007)
  - sub-gradients (Feldman et al, 2003; Komodakis et al., 2007)
  - proximal projections (Ravikumar et al., 2008)

Extensions to computing/bounding likelihoods

- log normalization/likelihood for an undirected model:

$$A(\theta) \quad = \quad \log \sum_{\mathbf{x} \in \mathcal{X}^N} \exp \Big\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \Big\}$$

- variational reformulation as a convex optimization problem:

$$A(\theta) \quad = \quad \max_{\vec{\mu} \in \mathbb{M}(G)} \left\{ \theta^T \vec{\mu} + H(\vec{\mu}) \right\}.$$

  where

  - $H(\vec{\mu})$ is maximized entropy, over all distributions with mean parameters $\vec{\mu}$
  - marginal polytope $\mathbb{M}(G)$ of all globally realizable distributions

- both $H(\cdot)$ and $\mathbb{M}(G)$ pose significant challenges for general graphs

- as before hypertrees are easy, and inspire the same relaxation philosophy (Wainwright & Jordan, 2003)

Summary

- marginal polytope: fundamental object associated with any discrete graphical model
- connections between LP relaxation and message-passing algorithms on graphs
- marginal polytopes and relaxations: also relevant for approximating/bounding marginals and likelihoods
- many open questions/issues:
  - approximation guarantees for LP relaxations: role of graph structure
  - guarantees for marginal/likelihood approximations
  - extensions to mixed discrete/continuous graphs, non-parametric settings
  - hybrid variational and MCMC methods

Some papers

- Wainwright, M. J. & Jordan, M. (2003) *Graphical models, exponential families, and variational methods.* Department of Statistics, UC Berkeley, Technical Report 649. To appear in Foundation and Trends in Machine Learning.

- Wainwright, M. J., Jaakkola, T. S., and Willsky, A. S., (2005), *Exact MAP estimates via agreement on hypertrees: Message-passing and linear programming*. IEEE Trans. Information Theory, 51:3697–3717.

- Wainwright, M. J., Jaakkola, T. S. and Willsky, A. S. (2005). A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory*. July, 51:2313–2335.

- Daskalakis, C., Dimakis, A. D., Karp, R. and Wainwright, M. J. (2008). *Probabilistic analysis of linear programming decoding.* To appear in IEEE Trans. Info. Theory.

- Ravikumar, P., Agarwal, A. and Wainwright, M. J. (2008). *Message-passing for graph-structured linear programs: Proximal projections and convergence.* To appear in Int. Conference on Machine Learning, Helsinki, Finland.