Graphical models, exponential families, and variational methods

Martin Wainwright

Departments of Statistics, and

Electrical Engineering and Computer Science,

UC Berkeley

Email: `wainwrig@eecs.berkeley.edu`

Tutorial slides based on joint paper with Michael Jordan

Paper at: `www.eecs.berkeley.edu/~wainwrig/WaiJorVariational03.ps`

Introduction

- graphical models are used and studied in various applied statistical and computational fields:
  - machine learning and artificial intelligence
  - computational biology
  - statistical signal/image processing
  - communication and information theory
  - statistical physics
  - .....

- based on correspondences between graph theory and probability theory

- important but difficult problems:
  - computing likelihoods, marginal distributions, modes
  - estimating model parameters and structure from (noisy) data

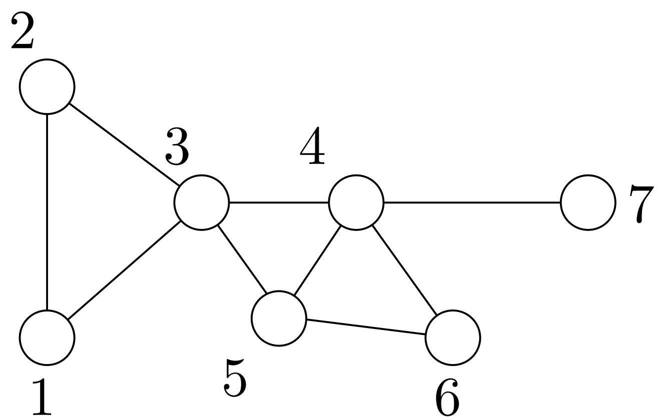
Outline

1. Introduction and motivation
    (a) Background on graphical models
    (b) Some applications and challenging problems
    (c) Illustrations of some variational methods
2. Exponential families and variational methods
    (a) What is a variational method (and why should I care)?
    (b) Graphical models as exponential families
    (c) The power of conjugate duality
3. Exact techniques as variational methods
    (a) Gaussian inference on arbitrary graphs
    (b) Belief-propagation/sum-product on trees (e.g., Kalman filter; α-β alg.)
    (c) Max-product on trees (e.g., Viterbi)
4. Approximate techniques as variational methods
    (a) Mean field and variants
    (b) Belief propagation and extensions on graphs with cycles
    (c) Semidefinite constraints and convex relaxations

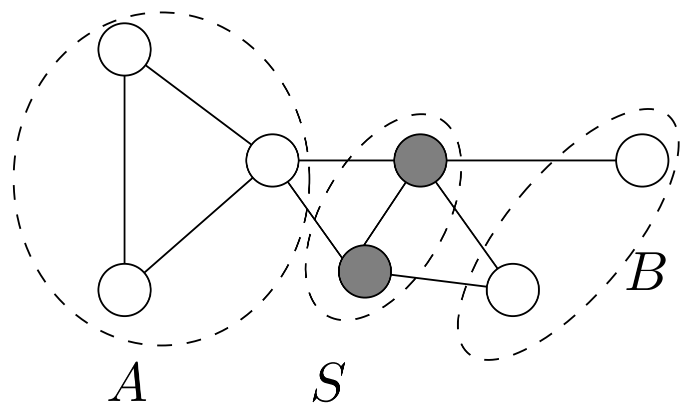
Undirected graphical models

Based on correspondences between graphs and random variables.

- given an undirected graph $G = (V, E)$, associate to each node s a random variable X_s
- for each subset $A \subseteq V$, define $X_A := \{x_s, s \in A\}$.



Maximal cliques $(123), (345), (456), (47)$



Vertex cutset S

- a *clique* $C \subseteq V$ is a subset of vertices all joined by edges
- a *vertex cutset* is a subset $S \subset V$ whose removal breaks the graph into two or more pieces

Factorization and Markov properties

The graph G can be used to impose constraints on the random vector $X = X_V$ (or on the distribution p) in different ways.

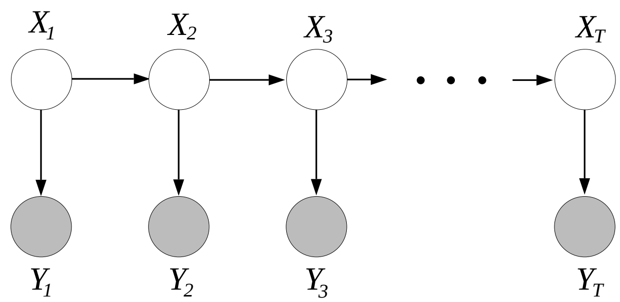
Markov property: X is *Markov w.r.t G* if X_A and X_B are conditionally indpt. given X_S whenever S separates A and B.

Factorization: The distribution p *factorizes according to G* if it can be expressed as a product over cliques:

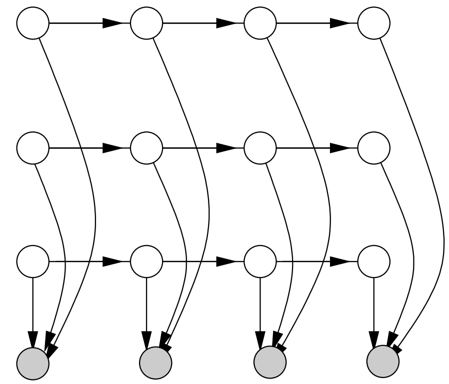
$$p(\mathbf{x}) \quad = \quad \frac{1}{Z} \prod_{C \in \mathcal{C}} \underbrace{\psi_C(x_C)}_{\text{compatibility function on clique } C}$$

Theorem: (Hammersley-Clifford) For strictly positive $p(\cdot)$, the Markov property and the Factorization property are equivalent.

Example: Hidden Markov models



(a) Hidden Markov model



(b) Coupled HMM

- HMMs are widely used in various applications

  discrete X_t: computational biology, speech processing, etc.

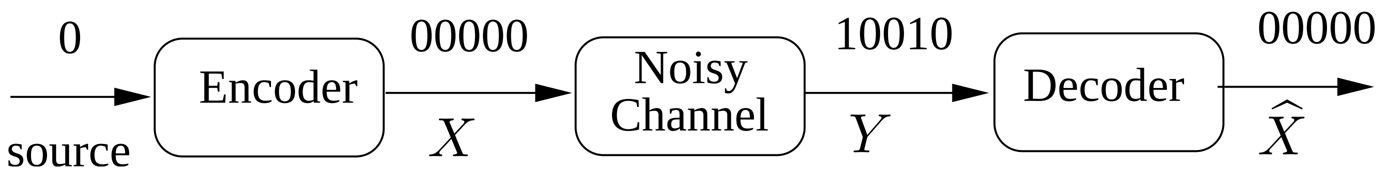
  Gaussian X_t: control theory, signal processing, etc.

- frequently wish to solve *smoothing* problem of computing $p(x_t \mid y_1, \ldots, y_T)$

- exact computation in HMMs is tractable, but coupled HMMs require algorithms for approximate computation (e.g., structured mean field)

Example: Graphical codes for communication

Goal: Achieve reliable communication over a noisy channel.



- wide variety of applications: satellite communication, sensor networks, computer memory, neural communication
- error-control codes based on careful addition of redundancy, with their fundamental limits determined by Shannon theory
- key implementational issues: *efficient* construction, encoding and decoding
- very active area of current research: *graphical codes* (e.g., turbo codes, low-density parity check codes) and iterative message-passing algorithms (belief propagation; max-product)

Graphical codes and decoding

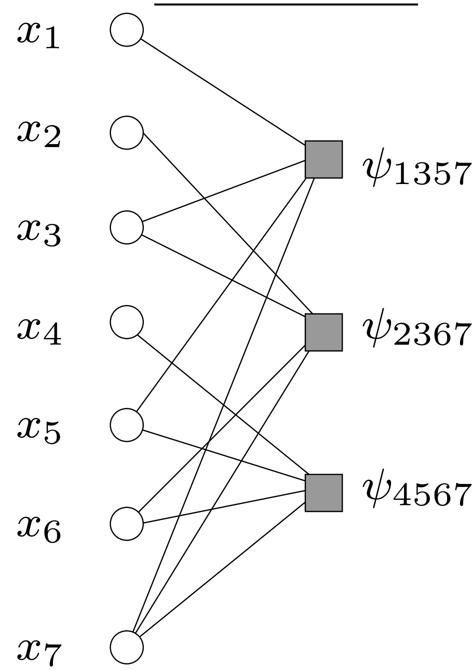
Parity check matrix

$$H = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Codeword: [0 1 0 1 0 1 0]

Non-codeword: [0 0 0 0 0 1 1]

Factor graph



- Decoding: requires finding maximum likelihood codeword:

$$\widehat{\mathbf{x}}_{ML} \;=\; \arg\max_{\mathbf{x}} p(\mathbf{y} \,|\, \mathbf{x}) \quad \text{s. t.} \quad H\mathbf{x} = 0 \pmod{2}.$$

- use of belief propagation as an approximate decoder has revolutionized the field of error-control coding

Challenging computational problems

Frequently, it is of interest to compute various quantities associated with an undirected graphical model:

(a) the log normalization constant $\log Z$

(b) local marginal distributions or other local statistics

(c) modes or most probable configurations

Relevant dimensions often grow rapidly in graph size $\Longrightarrow$ major computational challenges.

Example: Consider a naive approach to computing the normalization constant for binary random variables:

$$Z = \sum_{\mathbf{x} \in \{0,1\}^n} \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

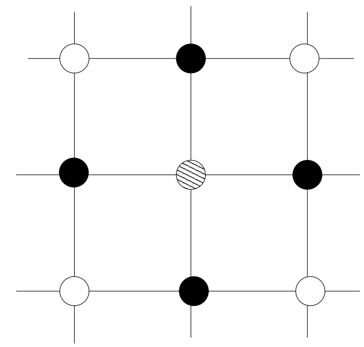
Complexity scales exponentially as 2^n.

Gibbs sampling in the Ising model

- binary variables on a graph $G = (V, E)$ with pairwise interactions:

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

Update $x_s^{(m+1)}$ stochastically based on values $x_{\mathcal{N}(s)}^{(m)}$ at neighbors:



1. Choose $s \in V$ at random.
2. Sample $u \sim \mathcal{U}(0,1)$ and update

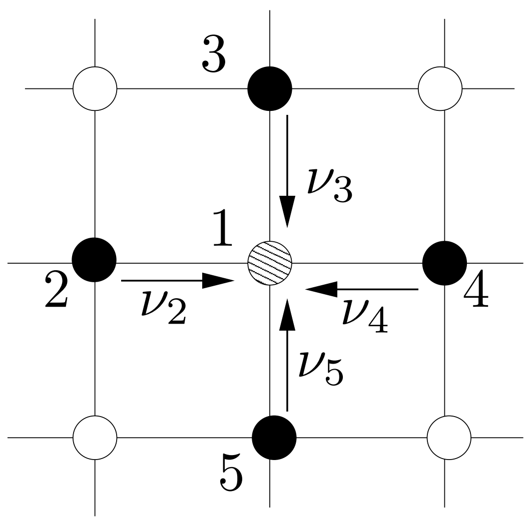
$$x_s^{(m+1)} = \begin{cases} 1 & \text{if } u \leq \{1 + \exp[-(\theta_s + \sum_{t\in\mathcal{N}(s)}\theta_{st}x_t^{(m)})]\}^{-1} \\ 0 & \text{otherwise} \end{cases},$$

Mean field updates in the Ising model

- binary variables on a graph $G = (V, E)$ with pairwise interactions:

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

- simple (deterministic) message-passing algorithm involving *variational parameters* $\nu_s \in (0, 1)$ at each node



1. Choose $s \in V$ at random.
2. Update ν_s based on neighbors $\{\nu_t,\ t \in \mathcal{N}(s)\}$:

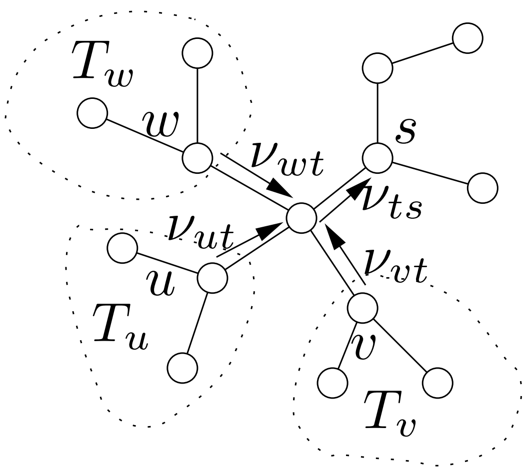
$$\nu_s \quad \longleftarrow \quad \Big\{1 + \exp[-(\theta_s + \sum_{t \in \mathcal{N}(s)} \theta_{st} \nu_t)]\Big\}^{-1}$$

<u>Questions:</u>

- principled derivation?
- convergence and accuracy?

Sum-product (belief-propagation) in the Ising model

- alternative set of message-passing updates (motivated by exactness for trees)



1. For each (direction of each) edge, update message:

$$\nu_{ts}(x_s) \leftarrow \sum_{x_t=0}^{1} \exp(\theta_t x_t + \theta_{st} x_s x_t) \prod_{u \in \mathcal{N}(t) \setminus s} \nu_{ut}(x_t)$$

2. Upon convergence, compute approx. to marginal:

$$p(x_s) \quad \propto \quad \exp\left(\theta_s x_s\right) \prod_{t \in \mathcal{N}(s)} \nu_{ts}(x_s).$$

- for any tree (i.e., no cycles), updates will converge (after a finite number of steps), and yield exact marginals (cf. Pearl, 1988)

- behavior for graphs with cycles?

Outline

1. Introduction and motivation
   (a) Background on graphical models
   (b) Some applications and challenging problems
   (c) Illustrations of some variational methods

2. Exponential families and variational methods
   (a) What is a variational method (and why should I care)?
   (b) Graphical models as exponential families
   (c) The power of conjugate duality

3. Exact techniques as variational methods
   (a) Gaussian inference on arbitrary graphs
   (b) Belief-propagation/sum-product on trees (e.g., Kalman filter; α-β alg.)
   (c) Max-product on trees (e.g., Viterbi)

4. Approximate techniques as variational methods
   (a) Mean field and variants
   (b) Belief propagation and extensions
   (c) Semidefinite constraints and convex relaxations

Variational methods

- *"variational":* umbrella term for optimization-based formulation of problems, and methods for their solution
- historical roots in the calculus of variations
- modern variational methods encompass a wider class of methods (e.g., dynamic programming; finite-element methods)

Variational principle: Representation of a quantity of interest $\widehat{\mathbf{u}}$ as the solution of an optimization problem.

1. allows the quantity $\widehat{\mathbf{u}}$ to be studied through the lens of the optimization problem
2. approximations to $\widehat{\mathbf{u}}$ can be obtained by approximating or relaxing the variational principle

Illustration: A simple variational principle

Goal: Given a vector $\mathbf{y} \in \mathbb{R}^n$ and a symmetric matrix $Q \succ 0$, solve the linear system $Q\mathbf{u} = \mathbf{y}$.

Unique solution $\widehat{\mathbf{u}}(\mathbf{y}) = Q^{-1}\mathbf{y}$ can be obtained by matrix inversion.

Variational formulation: Consider the function $J_{\mathbf{y}} : \mathbb{R}^n \to \mathbb{R}$ defined by

$$J_{\mathbf{y}}(\mathbf{u}) \quad := \quad \frac{1}{2}\mathbf{u}^T Q\mathbf{u} - \mathbf{y}^T\mathbf{u}.$$

It is strictly convex, and the minimum is uniquely attained:

$$\widehat{\mathbf{u}}(\mathbf{y}) \;=\; \arg\min_{\mathbf{u}\in\mathbb{R}^n} J_{\mathbf{y}}(\mathbf{u}) \;=\; Q^{-1}\mathbf{y}.$$

Various methods for solving linear systems (e.g., conjugate gradient) exploit this variational representation.

Useful variational principles for graphical models?

Consider an undirected graphical model:

$$p(\mathbf{x}) \quad = \quad \frac{1}{Z} \prod_{C \in \mathbf{C}} \psi_C(x_C)$$

Core problems that arise in many applications:

(a) computing the log normalization constant $\log Z$

(b) computing local marginal distributions (e.g., $p(x_s) = \sum_{x_t, t \neq s} p(\mathbf{x})$)

(c) computing modes or most likely configurations $\widehat{\mathbf{x}} \in \arg\max_{\mathbf{x}} p(\mathbf{x})$

Approach: Develop variational representations of all of these problems by exploiting ideas and results from:

(a) exponential families (e.g.,Brown, 1986)

(b) convex analysis (e.g., Rockafellar,1973)

Exponential families

$$
\begin{aligned}
\phi_\alpha : \mathcal{X}^n \to \mathbb{R} \quad &\equiv \quad \text{sufficient statistic} \\
\boldsymbol{\phi} = \{\phi_\alpha, \alpha \in \mathcal{I}\} \quad &\equiv \quad \text{vector of sufficient statistics} \\
\theta = \{\theta_\alpha, \alpha \in \mathcal{I}\} \quad &\equiv \quad \text{parameter vector} \\
\boldsymbol{\nu} \quad &\equiv \quad \text{base measure (e.g., Lebesgue, counting)}
\end{aligned}
$$

- parameterized family of densities (w.r.t. $\boldsymbol{\nu}$):

$$
p(\mathbf{x}; \theta) \quad = \quad \exp\Big\{\sum_\alpha \theta_\alpha \phi_\alpha(\mathbf{x}) \; - \; A(\theta)\Big\}
$$

- cumulant generating function (log normalization constant):

$$
A(\theta) \quad = \quad \log\Big(\int \exp\{\langle \theta, \, \boldsymbol{\phi}(\mathbf{x})\rangle\} \boldsymbol{\nu}(d\mathbf{x})\Big)
$$

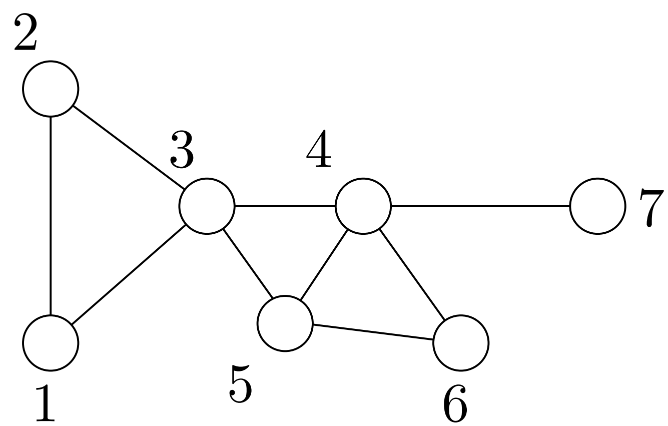
- set of valid parameters $\Theta := \{\theta \in \mathbb{R}^d \mid A(\theta) < +\infty\}$.
- will focus on *regular* families for which Θ is open.

Examples: Scalar exponential families

Family	$\mathcal{X}$	$\boldsymbol{\nu}$	$\log p(\mathbf{x};\theta)$	$A(\theta)$
Bernoulli	$\{0,1\}$	Counting	$\theta x - A(\theta)$	$\log[1+\exp(\theta)]$
Gaussian	$\mathbb{R}$	Lebesgue	$\theta_1 x + \theta_2 x^2 - A(\theta)$	$\frac{1}{2}[\theta_1 + \log\frac{2\pi e}{-\theta_2}]$
Exponential	$(0,+\infty)$	Lebesgue	$\theta\,(-x) - A(\theta)$	$-\log\theta$
Poisson	$\{0,1,2\ldots\}$	Counting $h(x)=1/x!$	$\theta x - A(\theta)$	$\exp(\theta)$

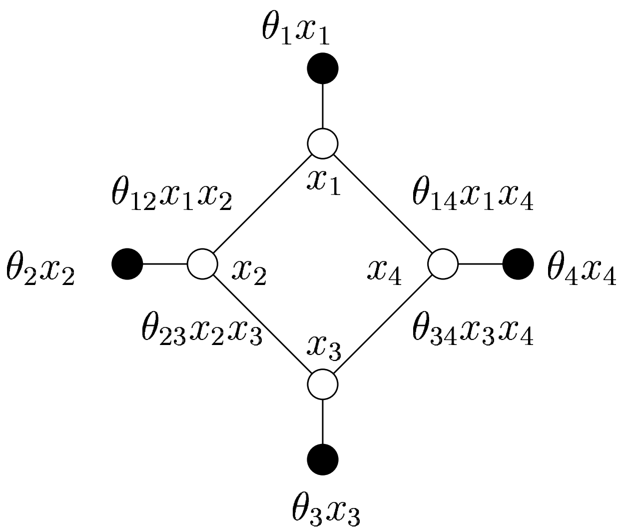
Graphical models as exponential families

- choose random variables X_s at each vertex $s \in V$ from an arbitrary exponential family (e.g., Bernoulli, Gaussian, Dirichlet etc.)
- exponential family can be the same at each node (e.g., multivariate Gaussian), or different (e.g., latent Dirichlet allocation model)



Key requirement: The collection $\boldsymbol{\phi}$ of sufficient statistics *must* respect the structure of G.

Example: Ising model



$$
\begin{aligned}
\phi &= \{\, x_s \,|\, s \in V \,\} \cup \{\, x_s x_t \,|\, (s,t) \in E \,\} \\
\mathcal{I} &= V \,\cup\, E \\
\mathcal{X}^n &= \{0,1\}^n
\end{aligned}
$$

Density (w.r.t. counting measure) of the form:

$$
p(\mathbf{x};\theta) \quad \propto \quad \exp\Big\{ \sum_{s=1}^{n} \theta_s x_s + \sum_{(s,t)\in E} \theta_{st} x_s x_t \Big\}
$$

Cumulant generating function (log normalization constant):

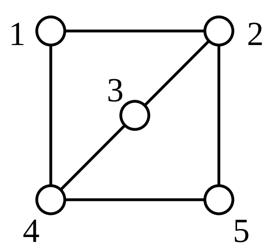
$$
A(\theta) \quad = \quad \log \sum_{\mathbf{x}\in\{0,1\}^n} \exp\Big\{ \sum_{s=1}^{n} \theta_s x_s + \sum_{(s,t)\in E} \theta_{st} x_s x_t \Big\}
$$

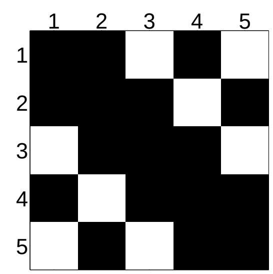
Example: Multivariate Gaussian

$U(\theta)$: Matrix of natural parameters $\qquad$ $\phi(\mathbf{x})$: Matrix of sufficient statistics

$$
\begin{bmatrix}
0 & \theta_1 & \theta_2 & \dots & \theta_n \\
\theta_1 & \theta_{11} & \theta_{12} & \dots & \theta_{1n} \\
\theta_2 & \theta_{21} & \theta_{22} & \dots & \theta_{2n} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\theta_n & \theta_{n1} & \theta_{n2} & \dots & \theta_{nn}
\end{bmatrix}
\qquad
\begin{bmatrix}
1 & x_1 & x_2 & \dots & x_n \\
x_1 & (x_1)^2 & x_1 x_2 & \dots & x_1 x_n \\
x_2 & x_2 x_1 & (x_2)^2 & \dots & x_2 x_n \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
x_n & x_n x_1 & x_n x_2 & \dots & (x_n)^2
\end{bmatrix}
$$

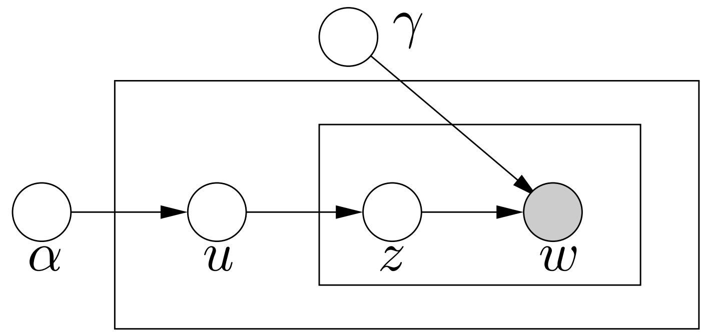
Edgewise natural parameters $\theta_{st} = \theta_{ts}$ must respect graph structure:





(a) Graph structure $\qquad$ (b) Structure of $[Z(\theta)]_{st} = \theta_{st}$.

Example: Latent Dirichlet Allocation model



Model components:

Dirichlet u	$\sim$	$\mathrm{Dir}(\alpha)$
Multinomial "topic" z	$\sim$	$\mathrm{Mult}(u)$
"Word" w	$\sim$	multinomial conditioned on z
		(with parameter γ)

With variables $\mathbf{x} := (u, z, w)$ and parameter $\theta := (\alpha, \gamma)$, density $p(u;\alpha)p(z;u)p(w \mid z, \gamma)$ is proportional to:

$$\exp\Big\{\sum_{i=1}^{n} \alpha_i \log u_i + \sum_{i=1}^{k} \mathbb{I}_i[z] \log u_i + \sum_{i=1}^{k}\sum_{j=1}^{l} \gamma_{ij} \mathbb{I}_i[z] \mathbb{I}_j[w]\Big\}.$$

The power of conjugate duality

Conjugate duality is a fertile source of variational principles.

(Rockafellar, 1973)

- any function f can be used to define another function f^* as follows:

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

- 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:

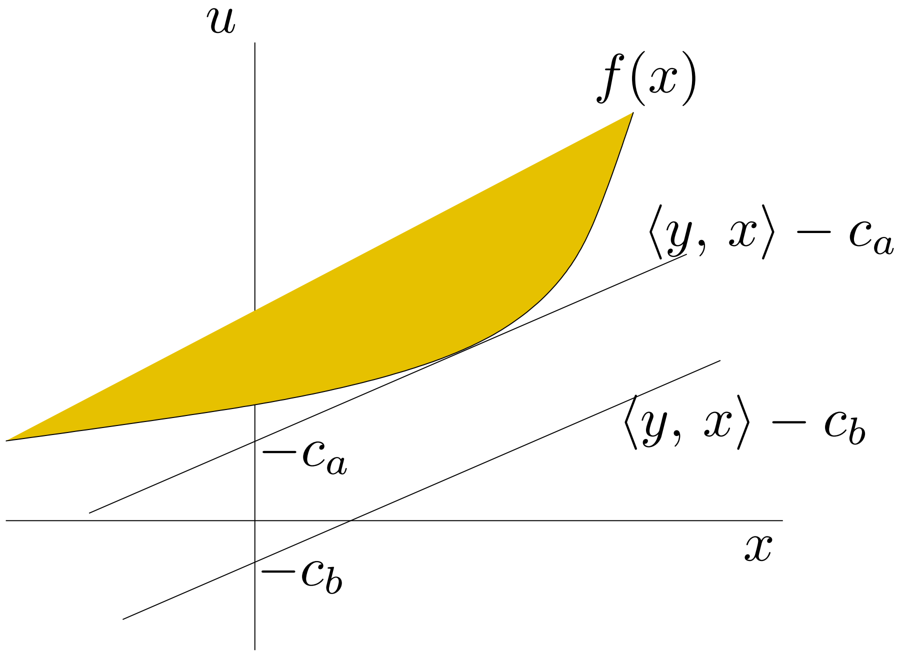
$$f(x) \quad = \quad \sup_{y \in \mathbb{R}^n} \big\{ \langle x,\, y \rangle - f^*(y) \big\}$$

Geometric view: Supporting hyperplanes

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

<u>Epigraph of f:</u>

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



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

$$\langle y,\, x\rangle - c \quad \le \quad f(x) \quad \text{for all} \quad x \in \mathbb{R}^n$$

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

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

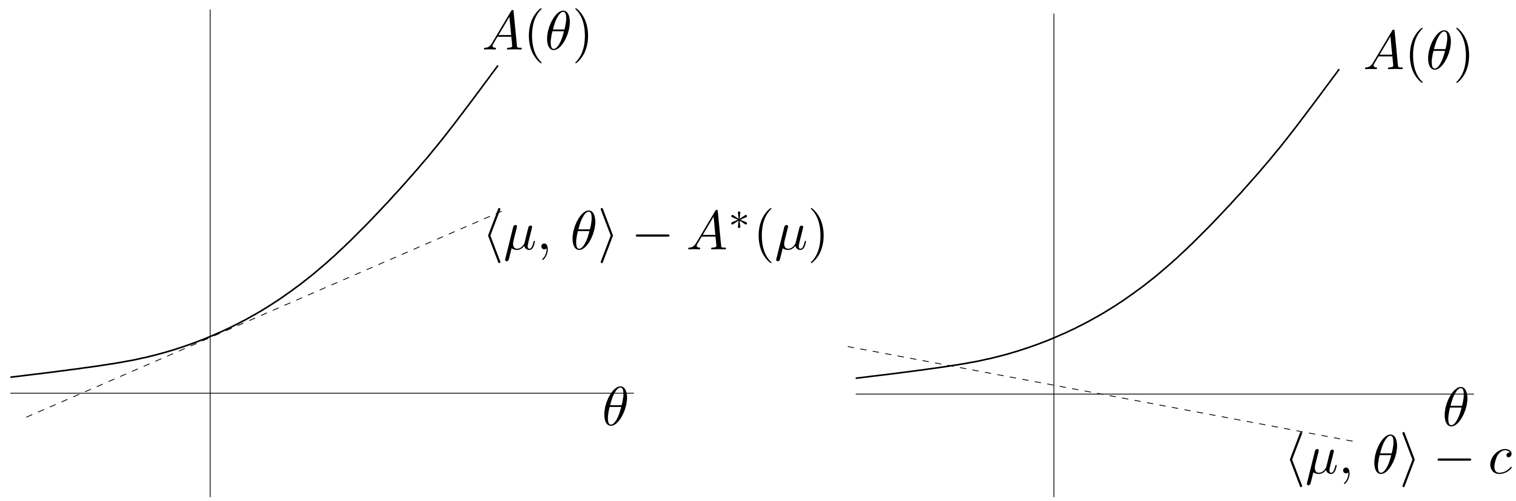
Example: Single Bernoulli

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

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

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

(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: $A(\theta) = \max_{\mu\in[0,1]}\left\{\mu\cdot\theta - A^*(\mu)\right\}$.

More general computation of the dual A^*

- consider the definition of the dual function:

$$A^*(\mu) \quad = \quad \sup_{\theta \in \mathbb{R}^d} \big\{\langle \mu,\, \theta \rangle - A(\theta)\big\}.$$

- taking derivatives w.r.t θ to find a stationary point yields:

$$\mu - \nabla A(\theta) \quad = \quad 0.$$

- Useful fact: Derivatives of A yield *mean parameters:*

$$\frac{\partial A}{\partial \theta_\alpha}(\theta) \quad = \quad \mathbb{E}_\theta[\phi_\alpha(\mathbf{x})] \;:=\; \int \phi_\alpha(\mathbf{x}) p(\mathbf{x};\theta) \boldsymbol{\nu}(\mathbf{x}).$$

Thus, stationary points satisfy the equation:

$$\mu \quad = \quad \mathbb{E}_\theta[\boldsymbol{\phi}(\mathbf{x})] \tag{1}$$

Computation of dual (continued)

- assume solution $\theta(\mu)$ to equation (1) exists
- strict concavity of objective guarantees that $\theta(\mu)$ attains global maximum with value

$$\begin{aligned} A^*(\mu) &= \langle \mu, \, \theta(\mu) \rangle - A(\theta(\mu)) \\ &= \mathbb{E}_{\theta(\mu)} \Big[\langle \theta(\mu), \, \boldsymbol{\phi}(\mathbf{x}) \rangle - A(\theta(\mu)) \Big] \\ &= \mathbb{E}_{\theta(\mu)} [\log p(\mathbf{x}; \theta(\mu))] \end{aligned}$$

- recall the definition of *entropy*:

$$H(p(\mathbf{x})) \quad := \quad -\int \big[\log p(\mathbf{x}) \big] p(\mathbf{x}) \boldsymbol{\nu}(d\mathbf{x})$$

- thus, we recognize that $A^*(\mu) = -H(p(\mathbf{x}; \theta(\mu)))$ when equation (1) has a solution

Question: For which $\mu \in \mathbb{R}^d$ does equation (1) have a solution $\theta(\mu)$?

Sets of realizable mean parameters

- for any distribution $p(\cdot)$, define a vector $\mu \in \mathbb{R}^d$ of *mean parameters*:

$$\mu_\alpha \quad := \quad \int \phi_\alpha(\mathbf{x}) p(\mathbf{x}) \boldsymbol{\nu}(d\mathbf{x})$$

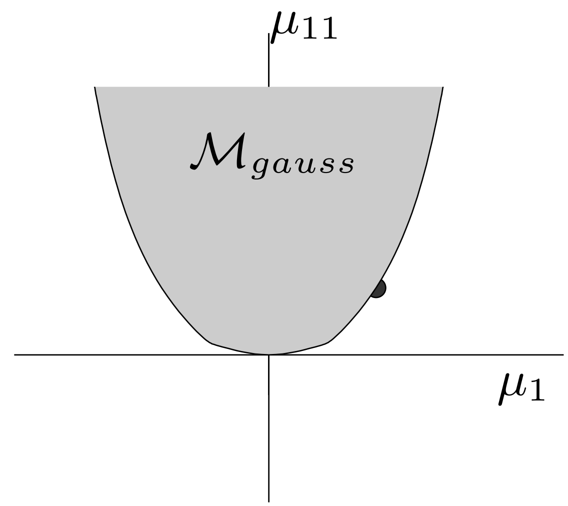
- now consider the set $\mathcal{M}(G; \boldsymbol{\phi})$ of all realizable mean parameters:

$$\mathcal{M}(G; \boldsymbol{\phi}) = \left\{ \mu \in \mathbb{R}^d \;\Big|\; \mu_\alpha = \int \phi_\alpha(\mathbf{x}) p(\mathbf{x}) \boldsymbol{\nu}(d\mathbf{x}) \quad \text{for } \textit{some } p(\cdot) \right\}$$

- for discrete families, we refer to this set as a *marginal polytope*, denoted by $\mathrm{MARG}(G; \boldsymbol{\phi})$

Examples of $\mathcal{M}$:

1. Gaussian MRF: Matrices of suff. statistics and mean parameters:



$$\phi(\mathbf{x}) = \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{x} \end{bmatrix}.$$

$$U(\mu) := \mathbb{E}\left\{ \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{x} \end{bmatrix} \right\}$$

Semidefinite set $\mathcal{M}_{Gauss} = \{\mu \mid U(\mu) \succeq 0\}$.

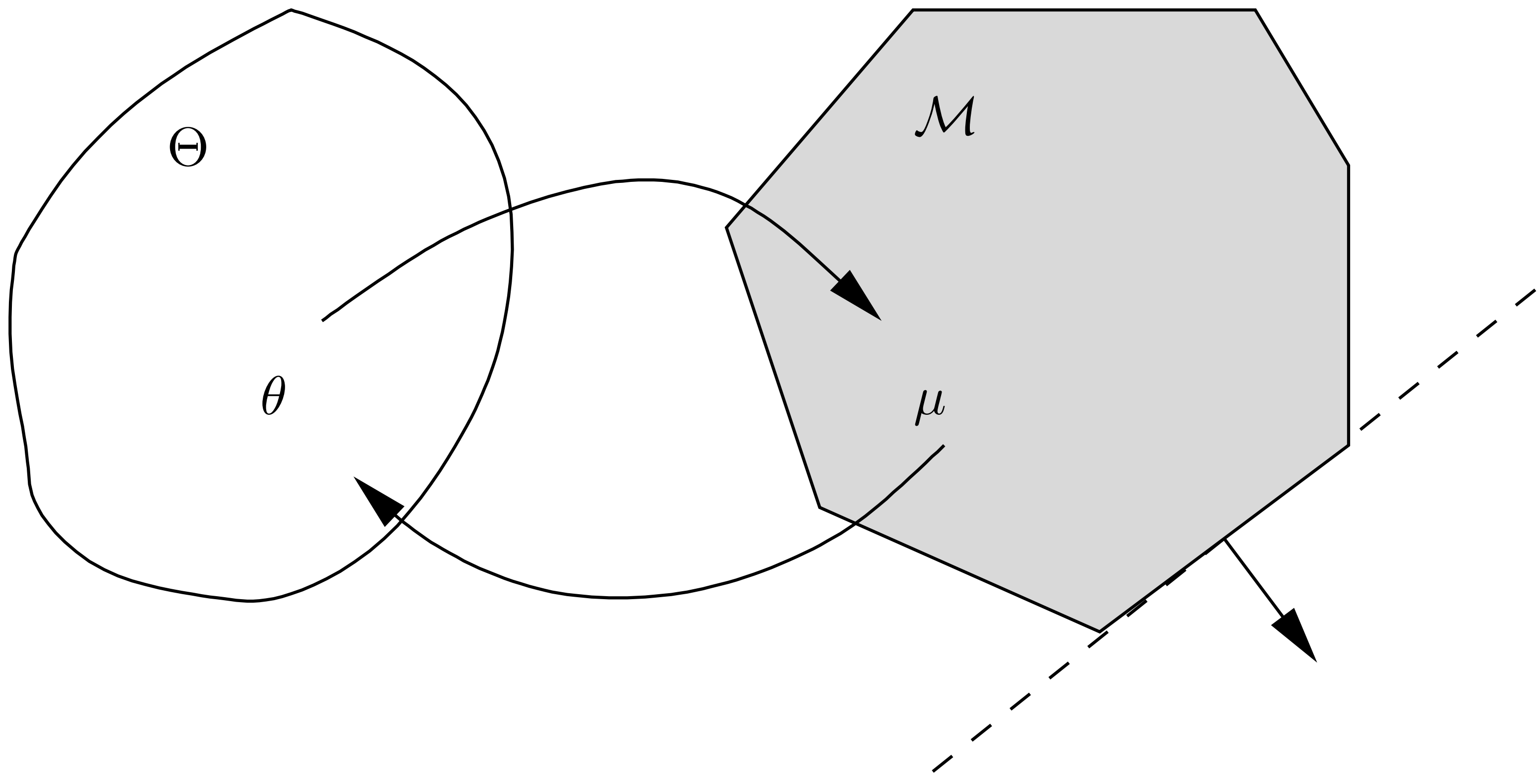
2. Ising model: Binary vector $X \in \{0,1\}^n$

Sufficient statistics: $\phi(\mathbf{x}) = \{x_s,\ s \in V\} \cup \{x_s x_t,\ (s,t) \in E\}$

$\mathcal{M}(G)$ is the *binary quadric polytope* of realizable singleton and pairwise marginal probabilities:

$$\mu_s = p(X_s = 1), \qquad \mu_{st} = p(X_s = 1, X_t = 1)$$

Geometry and moment mapping



Theorem: In a regular, minimal exponential family, the gradient map ∇A is one-to-one and onto the interior of the set $\mathcal{M}$.

(e.g., Barndorff-Nielsen, 1978; Brown, 1986; Efron, 1978)

Variational principles in terms of mean parameters

Theorem:

(a) The conjugate dual of A takes the form:

$$A^*(\mu) \quad = \quad \begin{cases} -H(p(\mathbf{x}; \theta(\mu))) & \text{if } \mu \in \operatorname{int} \mathcal{M}(G; \boldsymbol{\phi}) \\ +\infty & \text{if } \mu \notin \operatorname{cl} \mathcal{M}(G; \boldsymbol{\phi}). \end{cases}$$

Note: Boundary behavior by lower semi-continuity.

(b) The cumulant generating function A has the representation:

$$\underbrace{A(\theta)}_{\text{cumulant generating func.}} \quad = \quad \underbrace{\sup_{\mu \in \mathcal{M}(G; \boldsymbol{\phi})} \{\langle \theta, \, \mu \rangle \; - \; A^*(\mu)\}}_{\text{max. ent. problem over } \mathcal{M}},$$

with max. attained at mean parameters $\widehat{\mu}_\alpha = \mathbb{E}_\theta[\phi_\alpha(\mathbf{x})]$ (for all $\theta \in \Theta$).

(c) The problem of mode computation has the representation:

$$\sup_{\mathbf{x} \in \mathcal{X}^n} \log p(\mathbf{x}; \theta) + C \; = \; \sup_{\mathbf{x} \in \mathcal{X}^n} \langle \theta, \, \boldsymbol{\phi}(\mathbf{x}) \rangle \quad = \sup_{\mu \in \mathcal{M}(G; \boldsymbol{\phi})} \langle \theta, \, \mu \rangle.$$

Alternative view: Kullback-Leibler divergence

- Kullback-Leibler divergence defines "distance" between probability distributions:

$$D(p \,||\, q) \quad := \quad \int \Big[\log \frac{p(\mathbf{x})}{q(\mathbf{x})}\Big] p(\mathbf{x}) \boldsymbol{\nu}(d\mathbf{x})$$

- for two exponential family members $p(\mathbf{x}; \theta^1)$ and $p(\mathbf{x}; \theta^2)$, we have

$$D(p(\mathbf{x}; \theta^1) \,||\, p(\mathbf{x}; \theta^2)) \quad = \quad A(\theta^2) - A(\theta^1) - \langle \mu^1, \, \theta^2 - \theta^1 \rangle$$

- substituting $A(\theta^1) = \langle \theta^1, \, \mu^1 \rangle - A^*(\mu^1)$ yields a *mixed form*:

$$D(p(\mathbf{x}; \theta^1) \,||\, p(\mathbf{x}; \theta^2)) \quad \equiv \quad D(\mu^1 \,||\, \theta^2) \; = \; A(\theta^2) + A^*(\mu^1) - \langle \mu^1, \, \theta^2 \rangle$$

Hence, the following two assertions are equivalent:

$$A(\theta^2) \quad = \quad \sup_{\mu^1 \in \mathcal{M}(G; \boldsymbol{\phi})} \{\langle \theta^2, \, \mu^1 \rangle \; - \; A^*(\mu^1)\}$$

$$0 \quad = \quad \inf_{\mu^1 \in \mathcal{M}(G; \boldsymbol{\phi})} D(\mu^1 \,||\, \theta^2)$$

Challenges

1. In general, mean parameter spaces $\mathcal{M}$ can be very difficult to characterize (e.g., multidimensional moment problems).

2. Entropy $A^*(\mu)$ as a function of *only* the mean parameters μ typically lacks an explicit form.

Remarks:

1. Variational representation clarifies why certain models are tractable.

2. For intractable cases, one strategy is to solve an approximate form of the optimization problem.

Outline

1. Introduction and motivation
   (a) Background on graphical models
   (b) Some applications and challenging problems
   (c) Illustrations of some variational methods

2. Exponential families and variational methods
   (a) What is a variational method (and why should I care)?
   (b) Graphical models as exponential families
   (c) The power of conjugate duality

3. Exact techniques as variational methods
   (a) Gaussian inference on arbitrary graphs
   (b) Belief-propagation/sum-product on trees (e.g., Kalman filter; α-β alg.)
   (c) Max-product on trees (e.g., Viterbi)

4. Approximate techniques as variational methods
   (a) Mean field and variants
   (b) Belief propagation and extensions
   (c) Semidefinite constraints and convex relaxations

A(i): Multivariate Gaussian (fixed covariance)

Consider the set of all Gaussians with fixed *inverse* covariance $Q \succ 0$.

- potentials $\phi(\mathbf{x}) = \{x_1, \ldots, x_n\}$ and natural parameter $\theta \in \Theta = \mathbb{R}^n$.
- cumulant generating function:

$$A(\theta) \quad = \quad \log \int_{\mathbb{R}^n} \overbrace{\exp\Big\{\sum_{s=1}^{n} \theta_s x_s\Big\}}^{\text{density}} \quad \underbrace{\exp\Big\{-\frac{1}{2}\mathbf{x}^T Q \mathbf{x}\Big\} d\mathbf{x}}_{\text{base measure}}$$

- completing the square yields $A(\theta) = \frac{1}{2}\theta^T Q^{-1}\theta + \text{constant}$
- straightforward computation leads to the dual
$$A^*(\mu) = \tfrac{1}{2}\mu^T Q \mu - \text{constant}$$
- putting the pieces back together yields the variational principle

$$A(\theta) \quad = \quad \sup_{\mu \in \mathbb{R}^n} \Big\{\theta^T \mu - \frac{1}{2}\mu^T Q \mu\Big\} + \text{constant}$$

- optimum is uniquely obtained at the familiar Gaussian mean $\widehat{\mu} = Q^{-1}\theta$.

A(ii): Multivariate Gaussian (arbitrary covariance)

- matrices of sufficient statistics, natural parameters, and mean parameters:

$$\boldsymbol{\phi}(\mathbf{x}) = \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{x} \end{bmatrix}, \qquad U(\theta) := \begin{bmatrix} 0 & [\theta_s] \\ [\theta_s] & [\theta_{st}] \end{bmatrix} \qquad U(\mu) := \mathbb{E}\left\{ \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{x} \end{bmatrix} \right\}$$

- cumulant generating function:

$$A(\theta) \quad = \quad \log \int \exp\Big\{ \langle\!\langle U(\theta),\ \boldsymbol{\phi}(\mathbf{x}) \rangle\!\rangle \Big\} d\mathbf{x}$$

- computing the dual function:

$$A^*(\mu) \quad = \quad -\frac{1}{2} \log \det U(\mu) - \frac{n}{2} \log 2\pi e,$$

- exact variational principle is a *log-determinant problem*:

$$A(\theta) \quad = \quad \sup_{U(\mu) \succ 0,\ [U(\mu)]_{11}=1} \Big\{ \langle\!\langle U(\theta),\ U(\mu) \rangle\!\rangle + \frac{1}{2} \log \det U(\mu) \Big\} + \frac{n}{2} \log 2\pi e$$

- solution yields the *normal equations* for Gaussian mean and covariance.

B: Belief propagation/sum-product on trees

- multinomial variables $X_s \in \{0, 1, \ldots, m_s - 1\}$ on a *tree* $T = (V, E)$
- sufficient statistics: indicator functions for each node and edge

$$\begin{aligned} \mathbb{I}_j(x_s) \quad & \text{for} \quad s = 1, \ldots n, \quad j \in \mathcal{X}_s \\ \mathbb{I}_{jk}(x_s, x_t) \quad & \text{for} \quad (s,t) \in E, \quad (j,k) \in \mathcal{X}_s \times \mathcal{X}_t. \end{aligned}$$

- exponential representation of distribution:

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

where $\theta_s(x_s) := \sum_{j \in \mathcal{X}_s} \theta_{s;j} \mathbb{I}_j(x_s)$ (and similarly for $\theta_{st}(x_s, x_t)$)

- mean parameters are simply marginal probabilities, represented as:

$$\mu_s(x_s) \;:=\; \sum_{j \in \mathcal{X}_s} \mu_{s;j} \mathbb{I}_j(x_s), \qquad \mu_{st}(x_s, x_t) \;:=\; \sum_{(j,k) \in \mathcal{X}_s \times \mathcal{X}_t} \mu_{st;jk} \mathbb{I}_{jk}(x_s, x_t)$$

- the marginals must belong to the following *marginal polytope*:

$$\text{MARG}(T) \quad := \quad \{ \mu \geq 0 \mid \sum_{x_s} \mu_s(x_s) = 1, \; \sum_{x_t} \mu_{st}(x_s, x_t) = \mu_s(x_s) \},$$

Decomposition of entropy for trees

- by the junction tree theorem, any tree can be factorized in terms of its marginals $\mu \equiv \mu(\theta)$ as follows:

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

- taking logs and expectations leads to the following entropy decomposition:

$$H(p(\mathbf{x};\theta)) \;=\; -A^*(\mu(\theta)) \;=\; \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st})$$

where

$$H_s(\mu_s) \;:=\; -\sum_{x_s} \mu_s(x_s) \log \mu_s(x_s)$$

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

Exact variational principle on trees

- putting the pieces back together yields:

$$A(\theta) \quad = \quad \max_{\mu \in \mathrm{MARG}(T)} \Big\{ \langle \theta,\, \mu \rangle + \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E(T)} I_{st}(\mu_{st}) \Big\}.$$

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

$$\begin{aligned} \mathcal{L}(\mu; \lambda) = \langle \theta,\, \mu \rangle + \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E(T)} I_{st}(\mu_{st}) \\ + \sum_{(s,t) \in E} \Big[\sum_{x_t} \lambda_{st}(x_t) C_{st}(x_t) + \sum_{x_s} \lambda_{ts}(x_s) C_{ts}(x_s) \Big] \end{aligned}$$

Lagrangian derivation (continued)

- taking derivatives of the Lagrangian w.r.t μ_s and μ_{st} yields

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

- setting these partial derivatives to zero and simplifying:

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

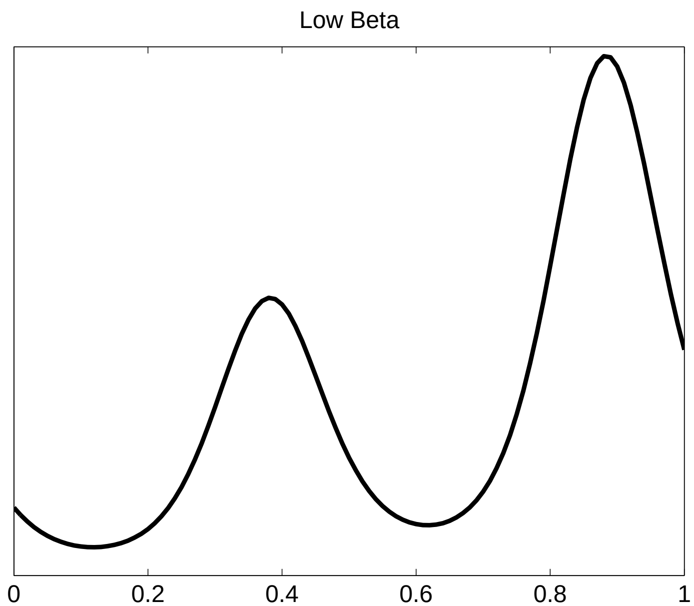
- enforcing the constraint $C_{ts}(x_s) = 0$ on these representations yields the familiar update rule for the *messages* $M_{ts}(x_s) = \exp(\lambda_{ts}(x_s))$:

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

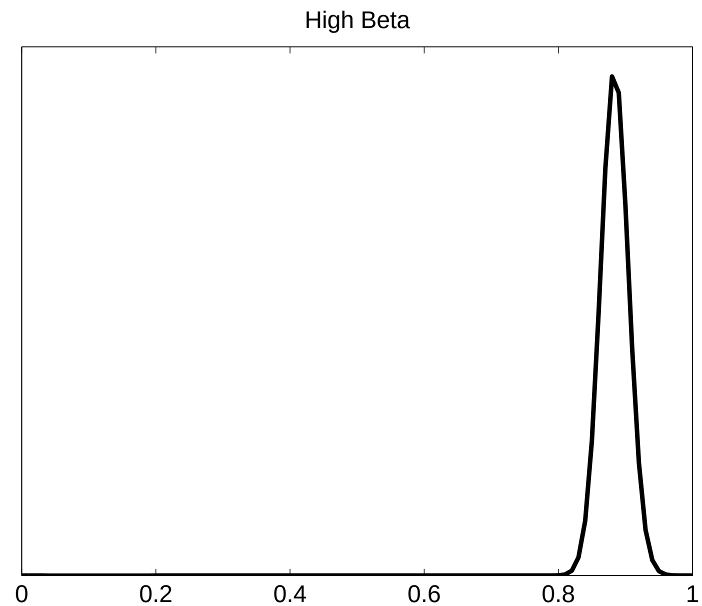
C: Max-product (belief revision) on trees

Question: What should be the form of a variational principle for computing modes?

Intuition: Consider behavior of the family $\{p(\mathbf{x}; \beta\theta) \mid \beta > 0\}$.



(a) Low β



(b) High β

Conclusion: Problem of computing modes should be related to limiting form $(\beta \to +\infty)$ of computing marginals.

Limiting form of variational principle (on trees)

- consider the tree-structured variational principle for $p(\mathbf{x}; \beta\theta)$:

$$\frac{1}{\beta} A(\beta\theta) \quad = \quad \frac{1}{\beta} \max_{\mu \in \mathrm{MARG}(T)} \Big\{ \langle \beta\theta,\, \mu \rangle + \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E(T)} I_{st}(\mu_{st}) \Big\}.$$

- taking limits as $\beta \to +\infty$ yields:

$$\underbrace{\max_{\mathbf{x} \in \mathcal{X}^N} \Big\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \Big\}}_{\text{computation of modes}} \quad = \quad \underbrace{\max_{\mu \in \mathrm{MARG}(T)} \big\{ \langle \theta,\, \mu \rangle \big\}}_{\text{linear program}}. \quad (2)$$

- recall the max-product (belief revision) updates:

$$M_{ts}(x_s) \quad \leftarrow \quad \max_{x_t} \exp\big\{ \theta_t(x_t) + \theta_{st}(x_s, x_t) \big\} \prod_{u \in \Gamma(t) \setminus s} M_{ut}(x_t)$$

- the LHS of equation (2) is a *linear program*: a similar Lagrangian formulation shows that max-product is an iterative method for solving it (details in Wainwright & Jordan, 2003)

Outline

1. Introduction and motivation
   (a) Background on graphical models
   (b) Some applications and challenging problems
   (c) Illustrations of some variational methods

2. Exponential families and variational methods
   (a) What is a variational method (and why should I care)?
   (b) Graphical models as exponential families
   (c) The power of conjugate duality

3. Exact techniques as variational methods
   (a) Gaussian inference on arbitrary graphs
   (b) Belief-propagation/sum-product on trees (e.g., Kalman filter; α-β alg.)
   (c) Max-product on trees (e.g., Viterbi)

4. Approximate techniques as variational methods
   (a) Mean field and variants
   (b) Belief propagation and extensions
   (c) Semidefinite constraints and convex relaxations

A: Mean field theory

Difficulty: (typically) no explicit form for $-A^*(\mu)$ (i.e., entropy as a function of mean parameters) $\Longrightarrow$ exact variational principle is intractable.

Idea: Restrict μ to a *subset* of distributions for which $-A^*(\mu)$ has a tractable form.

Examples:

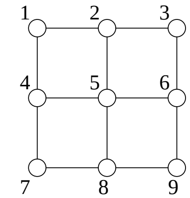
(a) For product distributions $p(\mathbf{x}) = \prod_{s \in V} \mu_s(x_s)$, entropy decomposes as $-A^*(\mu) = \sum_{s \in V} H_s(x_s)$.

(b) Similarly, for trees (more generally, decomposable graphs), the junction tree theorem yields an explicit form for $-A^*(\mu)$.

Definition: A subgraph H of G is *tractable* if the entropy has an explicit form for any distribution that respects H.

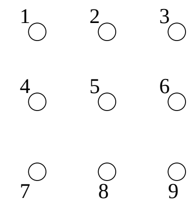
Geometry of mean field

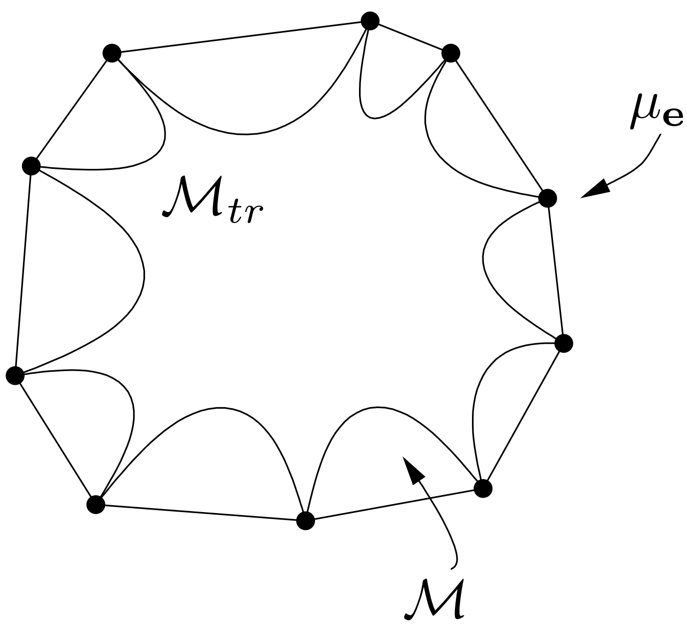
- let H represent a *tractable subgraph* (i.e., for which A^* has explicit form)



- let $\mathcal{M}_{tr}(G; H)$ represent tractable mean parameters:

$$\mathcal{M}_{tr}(G; H) := \{\mu \mid \mu = \mathbb{E}_\theta[\boldsymbol{\phi}(\mathbf{x})] \ \ \text{s. t. } \theta \ \text{ respects } \ H\}.$$





- under mild conditions, $\mathcal{M}_{tr}$ is a non-convex *inner approximation* to $\mathcal{M}$
- optimizing over $\mathcal{M}_{tr}$ (as opposed to $\mathcal{M}$) yields *lower bound*:

$$A(\theta) \ \geq \ \sup_{\widetilde{\mu} \in \mathcal{M}_{tr}} \big\{\langle \theta, \, \widetilde{\mu} \rangle - A^*(\widetilde{\mu})\big\}.$$

Alternative view: Minimizing KL divergence

- recall the *mixed form* of the KL divergence between $p(\mathbf{x};\theta)$ and $p(\mathbf{x};\widetilde{\theta})$:

$$D(\widetilde{\mu}\,||\,\theta) \quad = \quad A(\theta) + A^*(\widetilde{\mu}) - \langle \widetilde{\mu},\, \theta \rangle$$

- try to find the "best" approximation to $p(\mathbf{x};\theta)$ in the sense of KL divergence

- in analytical terms, the problem of interest is

$$\inf_{\widetilde{\mu} \in \mathcal{M}_{tr}} D(\widetilde{\mu}\,||\,\theta) \quad = \quad A(\theta) + \inf_{\widetilde{\mu} \in \mathcal{M}_{tr}} \Big\{ A^*(\widetilde{\mu}) - \langle \widetilde{\mu},\, \theta \rangle \Big\}$$

- hence, finding the tightest lower bound on $A(\theta)$ is equivalent to finding the best approximation to $p(\mathbf{x};\theta)$ from distributions with $\widetilde{\mu} \in \mathcal{M}_{tr}$

Example: Naive mean field algorithm for Ising model

- consider completely disconnected subgraph $H = (V, \emptyset)$
- permissible exponential parameters belong to subspace

$$\mathcal{E}(H) = \{\theta \in \mathbb{R}^d \mid \theta_{st} = 0 \ \forall \ (s,t) \in E\}$$

- allowed distributions take product form $p(\mathbf{x}; \theta) = \prod_{s \in V} p(x_s; \theta_s)$, and generate

$$\mathcal{M}_{tr}(G; H) = \{\mu \mid \mu_{st} = \mu_s \mu_t, \ \ \mu_s \in [0,1] \ \}.$$

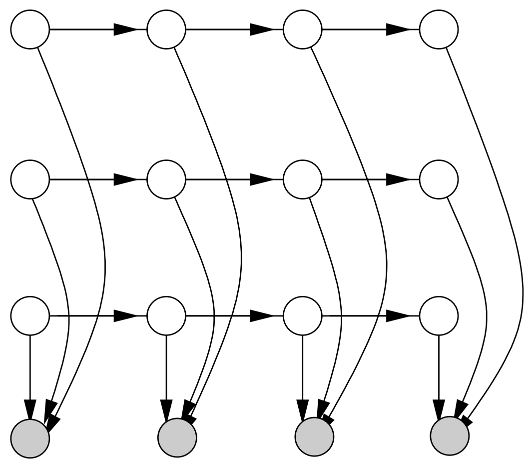
- approximate variational principle:

$$\max_{\mu_s \in [0,1]} \Big\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t - \big[\sum_{s \in V} \mu_s \log \mu_s + (1 - \mu_s) \log(1 - \mu_s) \big] \Big\}.$$

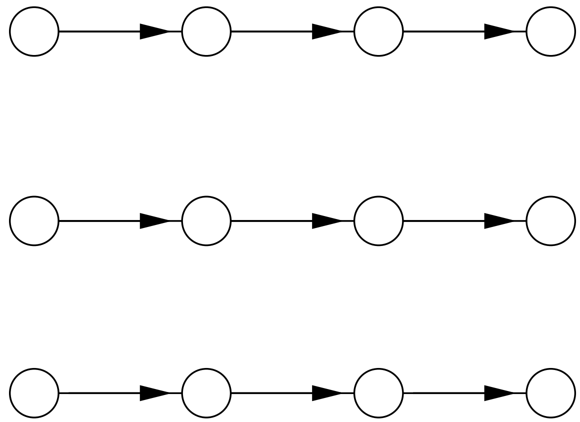
- **Co-ordinate ascent:** with all $\{\mu_t, t \neq s\}$ fixed, problem is strictly concave in μ_s and optimum is attained at

$$\mu_s \quad \longleftarrow \quad \Big\{ 1 + \exp[-(\theta_s + \sum_{t \in \mathcal{N}(s)} \theta_{st} \mu_t)] \Big\}^{-1}$$

Example: Structured mean field for coupled HMM



(a)



(b)

- entropy of distribution that respects H decouples into sum: one term for each chain.
- *structured mean field updates* are an iterative method for finding the tightest approximation (either in terms of KL or lower bound)

B: Belief propagation on arbitrary graphs

Two main ingredients:

1. Exact entropy $-A^*(\mu)$ is intractable, so let's approximate it.

   The *Bethe approximation* $A^*_{Bethe}(\mu) \approx A^*(\mu)$ is based on the exact expression for trees:

$$-A^*_{Bethe}(\mu) \quad = \quad \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st}).$$

2. The *marginal polytope* $\mathrm{MARG}(G)$ is also difficult to characterize, so let's use the following (tree-based) outer bound:

$$\mathrm{LOCAL}(G) \quad := \quad \{ \, \tau \geq 0 \mid \sum_{x_s} \tau_s(x_s) = 1, \; \sum_{x_t} \tau_{st}(x_s, x_t) = \tau_s(x_s) \, \},$$

Note: Use τ to distinguish these locally consistent *pseudomarginals* from globally consistent marginals.

Geometry of belief propagation

- combining these ingredients leads to the *Bethe variational principle:*

$$\max_{\tau \in \mathrm{LOCAL}(G)} \Big\{ \langle \theta, \tau \rangle + \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st}) \Big\}$$

- belief propagation can be derived as an iterative method for solving a Lagrangian formulation of the BVP (Yedidia et al., 2002)

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

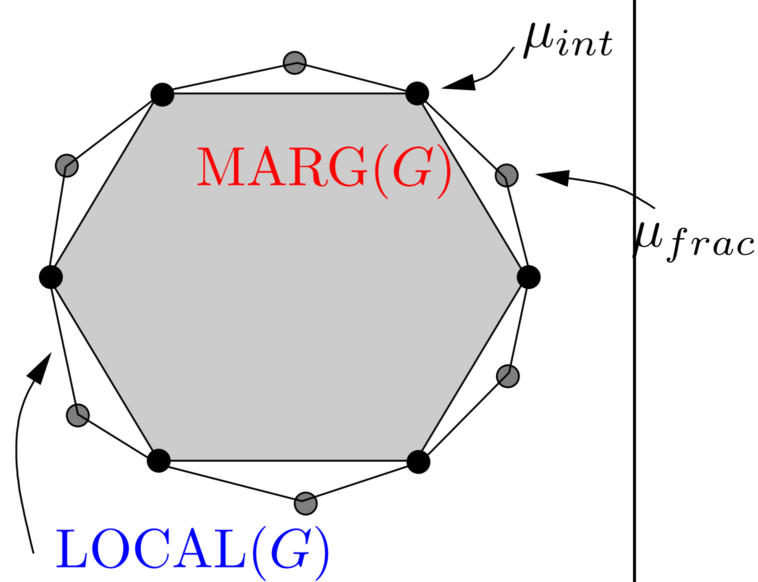
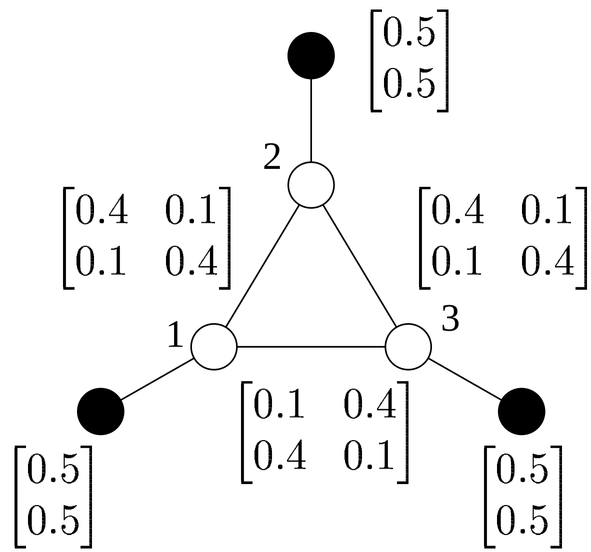


Illustration: Globally inconsistent BP fixed points

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

Locally consistent (pseudo)marginals



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

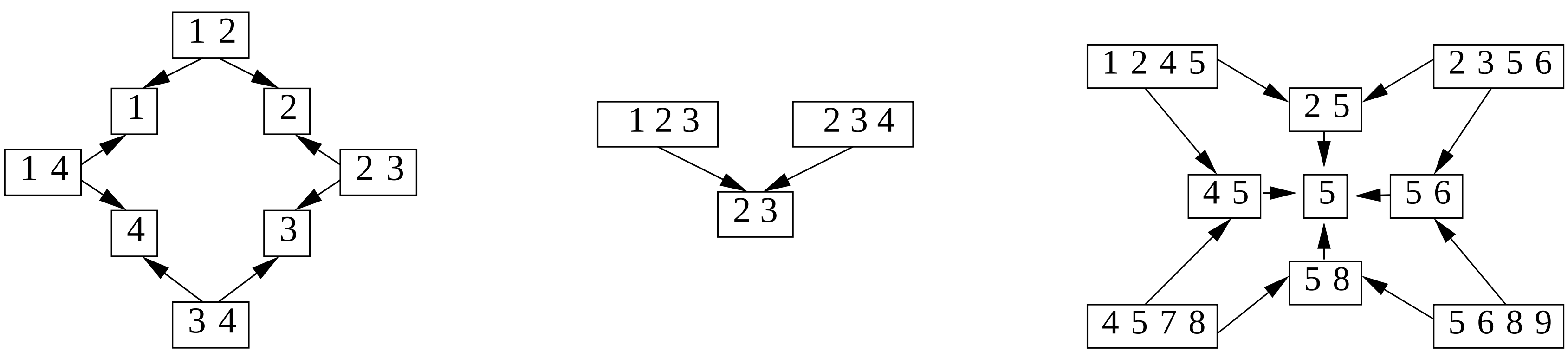
Note: More generally: for any τ in the interior of $\text{LOCAL}(G)$, can construct a distribution with τ as a BP fixed point.

High-level perspective

- 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 $\mathcal{M}$
  (b) various approximations to entropy function $-A^*(\mu)$
- <u>mean field:</u> non-convex inner bound and exact form of entropy
- <u>BP:</u> polyhedral outer bound and non-convex Bethe approximation
- <u>Kikuchi and variants:</u> tighter polyhedral outer bounds and better entropy approximations
  (e.g.,Yedidia et al., 2002)

Generalized belief propagation on hypergraphs

- a *hypergraph* is a natural generalization of a graph
- it consists of a set of vertices V and a set E of hyperedges, where each *hyperedge* is a subset of V
- convenient graphical representation in terms of *poset diagrams*



(a) Ordinary graph (b) Hypertree (width 2) (c) Hypergraph

- *descendants* and *ancestors* of a hyperedge h:

$$\mathcal{D}^+(h) := \{g \in E \mid g \subseteq h\}, \qquad \mathcal{A}^+(h) := \{g \in E \mid g \supseteq h\}.$$

Hypertree factorization and entropy

- hypertrees are an alternative way to describe junction trees

- associated with any poset is a Möbius function $\omega : E \times E \to \mathbb{Z}$

$$\omega(g, g) \;=\; 1, \quad \omega(g, h) \;=\; -\sum_{\{f \mid g \subseteq f \subset h\}} \omega(g, f)$$

  Example: For Boolean poset, $\omega(g, h) = (-1)^{|h| \backslash |g|}$.

- use the Möbius function to define a correspondence between the collection of marginals $\mu := \{\mu_h\}$ and new set of functions $\varphi := \{\varphi_h\}$:

$$\log \varphi_h(x_h) \;=\; \sum_{g \in \mathcal{D}^+(h)} \omega(g, h) \log \mu_g(x_g), \qquad \log \mu_h(x_h) \;=\; \sum_{g \in \mathcal{D}^+(h)} \log \varphi_g(x_g).$$

- any hypertree-structured distribution is guaranteed to factor as:

$$p(\mathbf{x}) \;=\; \prod_{h \in E} \varphi_h(x_h).$$

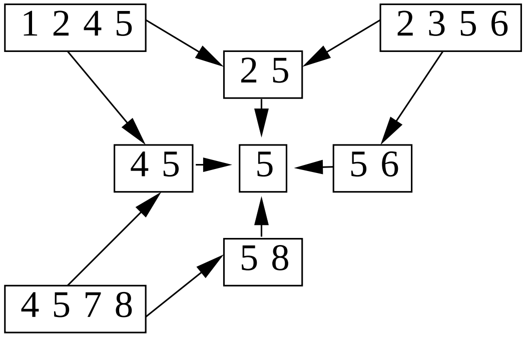
Examples: Hypertree factorization

1. **Ordinary tree:**

$$
\begin{aligned}
\varphi_s(x_s) &= \mu_s(x_s) && \text{for any vertex } s \\
\varphi_{st}(x_s, x_t) &= \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\ \mu_t(x_t)} && \text{for any edge } (s,t)
\end{aligned}
$$

2. **Hypertree:**

$$
\begin{aligned}
\varphi_{1245} &= \frac{\mu_{1245}}{\frac{\mu_{25}}{\mu_5}\ \frac{\mu_{45}}{\mu_5}\mu_5} \\
\varphi_{45} &= \frac{\mu_{45}}{\mu_5} \\
\varphi_5 &= \mu_5
\end{aligned}
$$

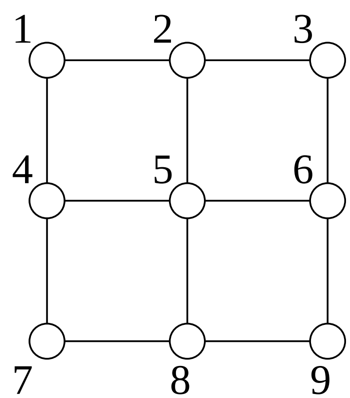


Combining the pieces:

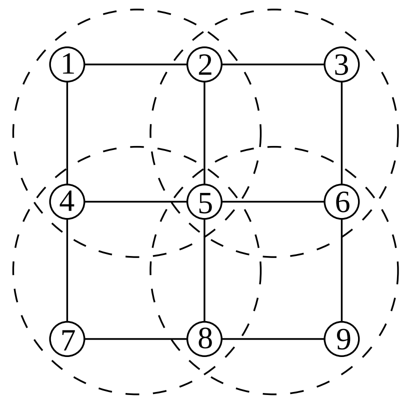
$$
p = \frac{\mu_{1245}}{\frac{\mu_{25}}{\mu_5}\ \frac{\mu_{45}}{\mu_5}\mu_5}\ \frac{\mu_{2356}}{\frac{\mu_{25}}{\mu_5}\ \frac{\mu_{56}}{\mu_5}\mu_5}\ \frac{\mu_{4578}}{\frac{\mu_{45}}{\mu_5}\ \frac{\mu_{58}}{\mu_5}\mu_5}\ \frac{\mu_{25}}{\mu_5}\ \frac{\mu_{45}}{\mu_5}\ \frac{\mu_{56}}{\mu_5}\ \frac{\mu_{58}}{\mu_5}\ \mu_5 = \frac{\mu_{1245}\ \mu_{2356}\ \mu_{4578}}{\mu_{25}\ \mu_{45}}
$$

Building augmented hypergraphs

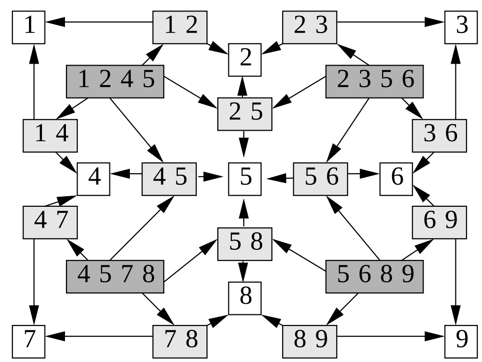
Better entropy approximations via augmented hypergraphs.



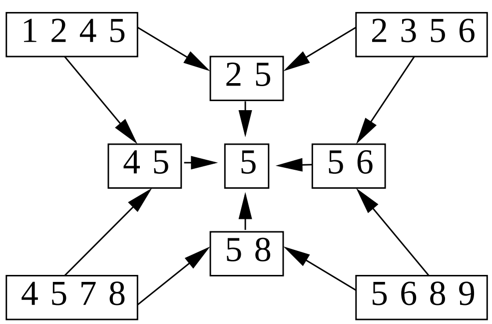
(a) Original



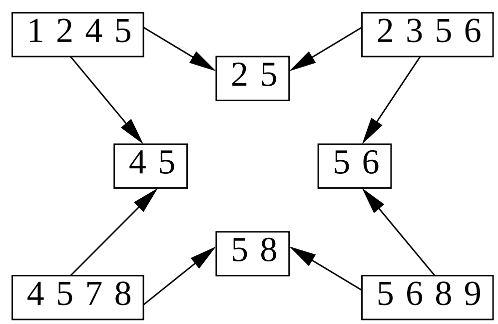
(b) Clustering



(c) Full covering



(d) Kikuchi



(e) Fails single counting

C. Convex relaxations

Possible concerns with the Bethe/Kikuchi problems and variations?

(a) lack of convexity $\Rightarrow$ multiple local optima, and substantial algorithmic complications

(b) failure to bound the log partition function

Goal: Techniques for approximate computation of marginals and parameter estimation based on:

(a) convex variational problems $\Rightarrow$ unique global optimum

(b) relaxations of exact problem $\Rightarrow$ upper bounds on $A(\theta)$

Usefulness of bounds:

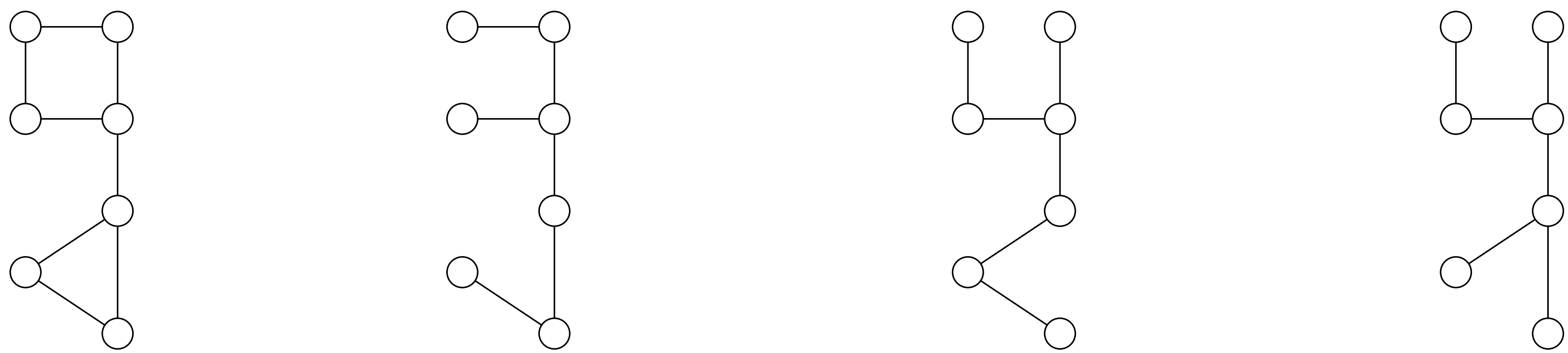
(a) interval estimates for marginals

(b) approximate parameter estimation

(c) large deviations (prob. of rare events)

Bounds from "convexified" Bethe/Kikuchi problems

Idea: Upper bound $-A^*(\mu)$ by convex combination of tree-structured entropies.



$$-A^*(\mu) \quad \leq \quad -\rho(T^1)A^*(\mu(T^1)) \quad - \quad \rho(T^2)A^*(\mu(T^2)) \quad - \quad \rho(T^3)A^*(\mu(T^3))$$

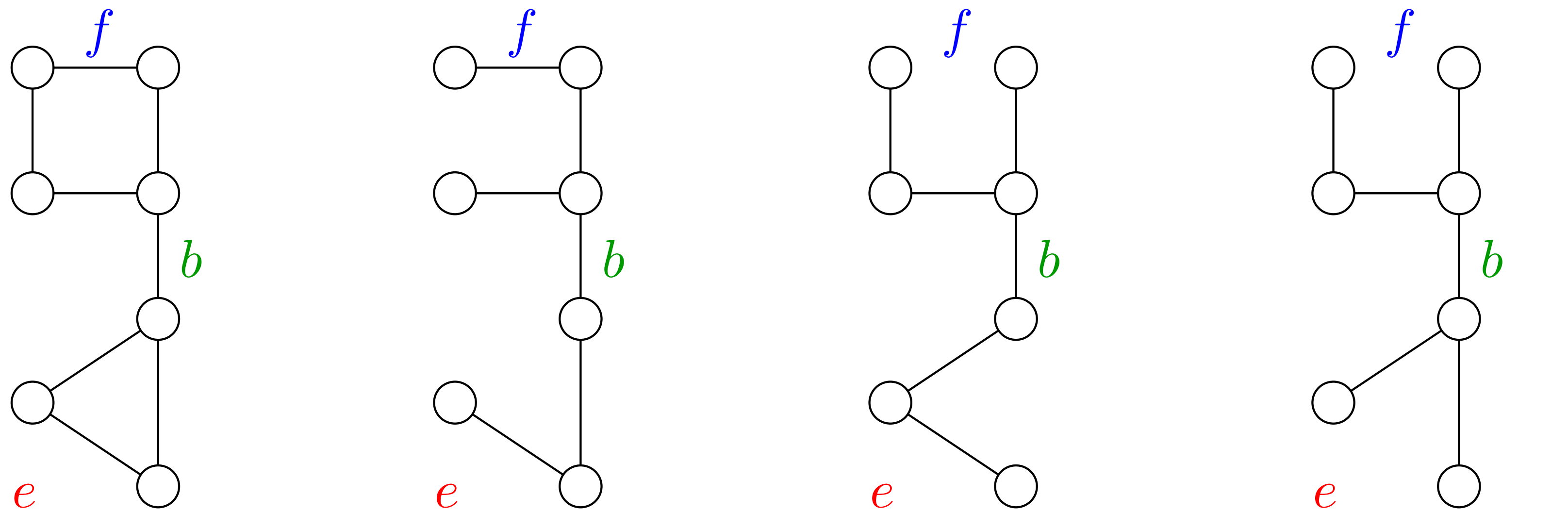
- given any spanning tree T, define the moment-matched tree distribution:

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

- use $-A^*(\mu(T))$ to denote the associated tree entropy
- let $\boldsymbol{\rho} = \{\rho(T)\}$ be a probability distribution over spanning trees

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)

Optimal bounds by tree-reweighted message-passing

Recall the constraint set of locally consistent marginal distributions:

$$\text{LOCAL}(G) \quad = \quad \{\, \tau \geq 0 \mid \underbrace{\sum_{x_s} \tau_s(x_s) = 1,}_{\text{normalization}} \; \underbrace{\sum_{x_s} \tau_{st}(x_s, x_t) = \tau_t(x_t)}_{\text{marginalization}} \,\}.$$

Theorem: (Wainwright, Jaakkola, & Willsky, UAI 2002)

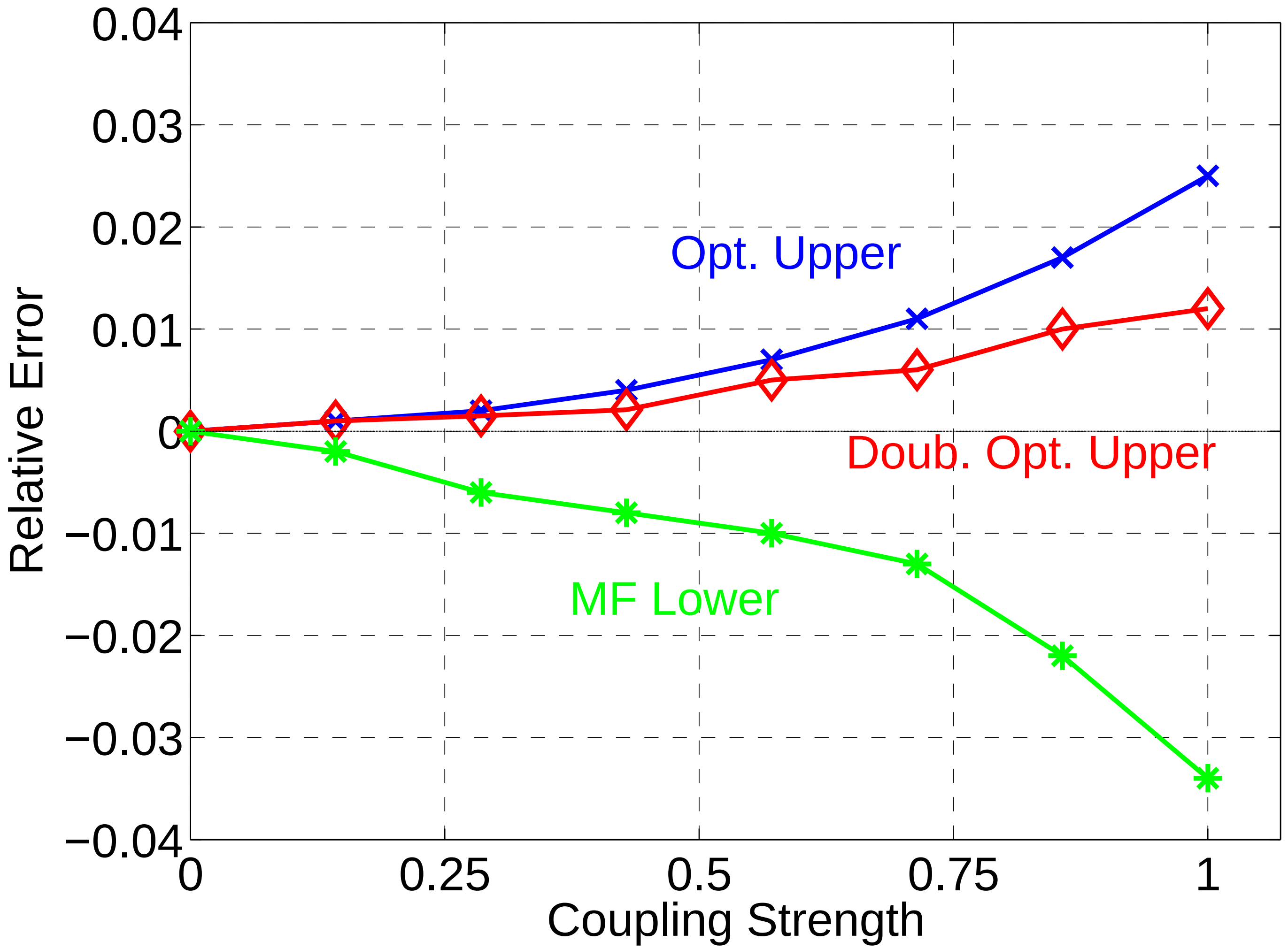
(a) For any given edge weights $\boldsymbol{\rho_e} = \{\rho_e\}$ in the spanning tree polytope, the optimal upper bound over *all* tree parameters is given by:

$$A(\theta) \;\leq\; \max_{\tau \in \text{LOCAL}(G)} \Big\{ \langle \theta, \, \tau \rangle + \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} \rho_{st} I_{st}(\tau_{st}) \,\Big\}.$$

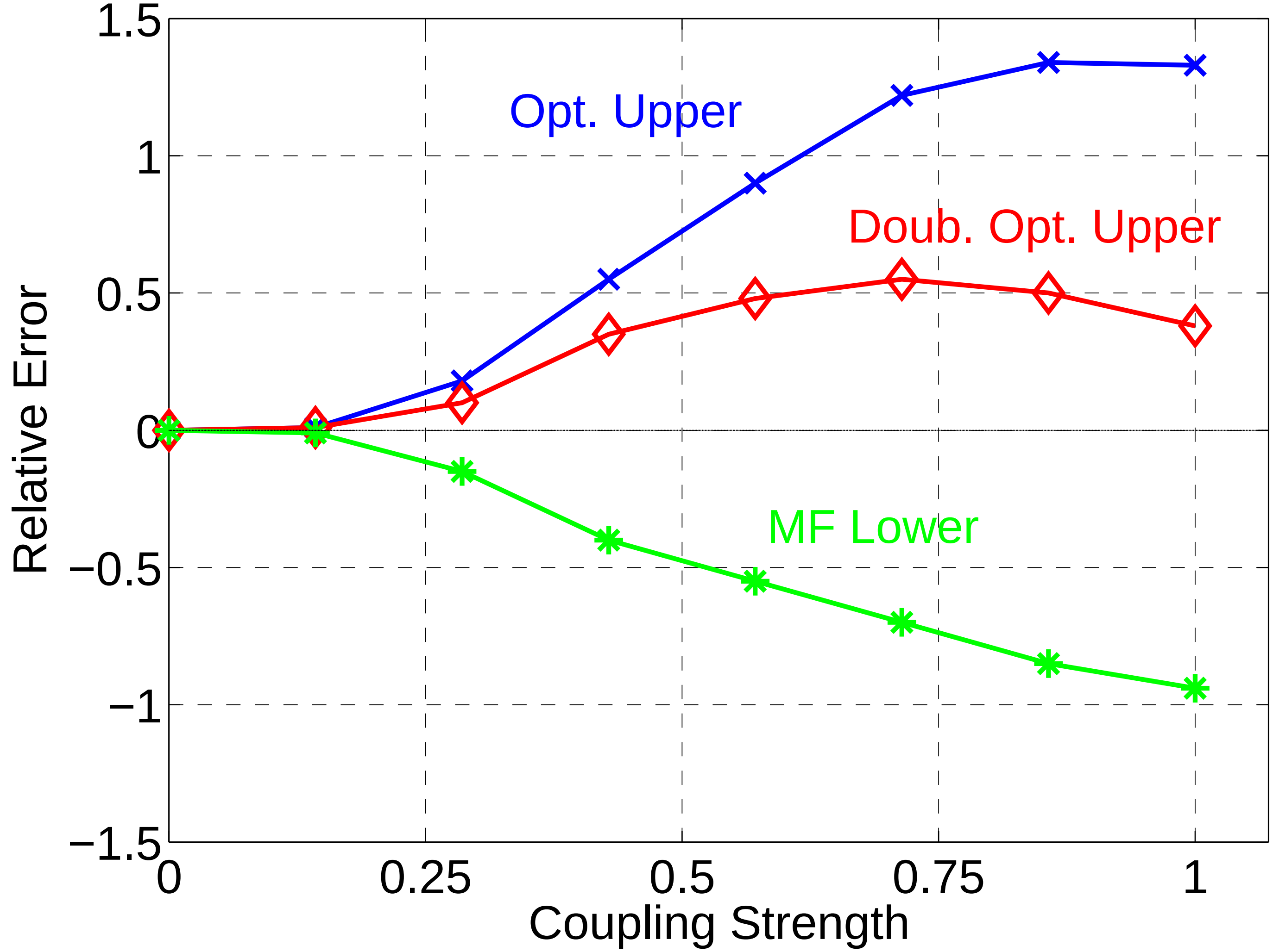
(b) This optimization problem is strictly convex, and its unique optimum is specified by the fixed point of $\boldsymbol{\rho_e}$-reweighted message passing:

$$M^*_{ts}(x_s) = \kappa \sum_{x'_t \in \mathcal{X}_t} \left\{ \exp\left[\frac{\theta_{st}(x_s, x'_t)}{\rho_{st}} + \theta_t(x'_t) \right] \frac{\prod_{v \in \Gamma(t) \backslash s} \big[M^*_{vt}(x_t) \big]^{\rho_{vt}}}{\big[M^*_{st}(x_t) \big]^{(1-\rho_{ts})}} \right\}.$$

Upper bounds on lattice model



Upper bounds on fully connected models



Semidefinite constraints in convex relaxations

Fact: Belief propagation and its hypergraph-based generalizations all involve polyhedral (i.e., *linear*) outer bounds on the marginal polytope.

Idea: Use *semidefinite* constraints to generate more global outer bounds.

Example: For the Ising model, relevant mean parameters are $\mu_s = p(X_s = 1)$ and $\mu_{st} = p(X_s = 1, X_t = 1)$.

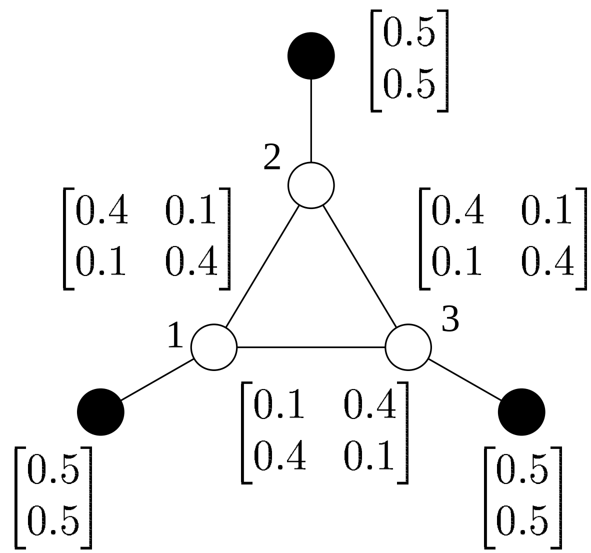
Define $\mathbf{y} = [1 \;\; \mathbf{x}]^T$, and consider the second-order moment matrix:

$$\mathbb{E}[\mathbf{y}\mathbf{y}^T] \;=\; \begin{bmatrix} 1 & \mu_1 & \mu_2 & \cdots & \mu_n \\ \mu_1 & \mu_1 & \mu_{12} & \cdots & \mu_{1n} \\ \mu_2 & \mu_{12} & \mu_2 & \cdots & \mu_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_n & \mu_{n1} & \mu_{n2} & \cdots & \mu_n \end{bmatrix}$$

It must be positive semidefinite, which imposes (an infinite number of) linear constraints on μ_s, μ_{st}.

Illustrative example

Locally consistent (pseudo)marginals



Second-order moment matrix

$$\begin{bmatrix} \mu_1 & \mu_{12} & \mu_{13} \\ \mu_{21} & \mu_2 & \mu_{23} \\ \mu_{31} & \mu_{32} & \mu_3 \end{bmatrix} = \begin{bmatrix} 0.5 & 0.4 & 0.1 \\ 0.4 & 0.5 & 0.4 \\ 0.1 & 0.4 & 0.5 \end{bmatrix}$$

Not positive-semidefinite!

Log-determinant relaxation

- based on optimizing over covariance matrices $M_1(\mu) \in \mathrm{SDEF}_1(K_n)$

Theorem: Consider an outer bound $\mathrm{OUT}(K_n)$ that satisfies:

$$\mathrm{MARG}(K_n) \quad \subseteq \quad \mathrm{OUT}(K_n) \;\subseteq\; \mathrm{SDEF}_1(K_n)$$

For any such outer bound, $A(\theta)$ is upper bounded by:

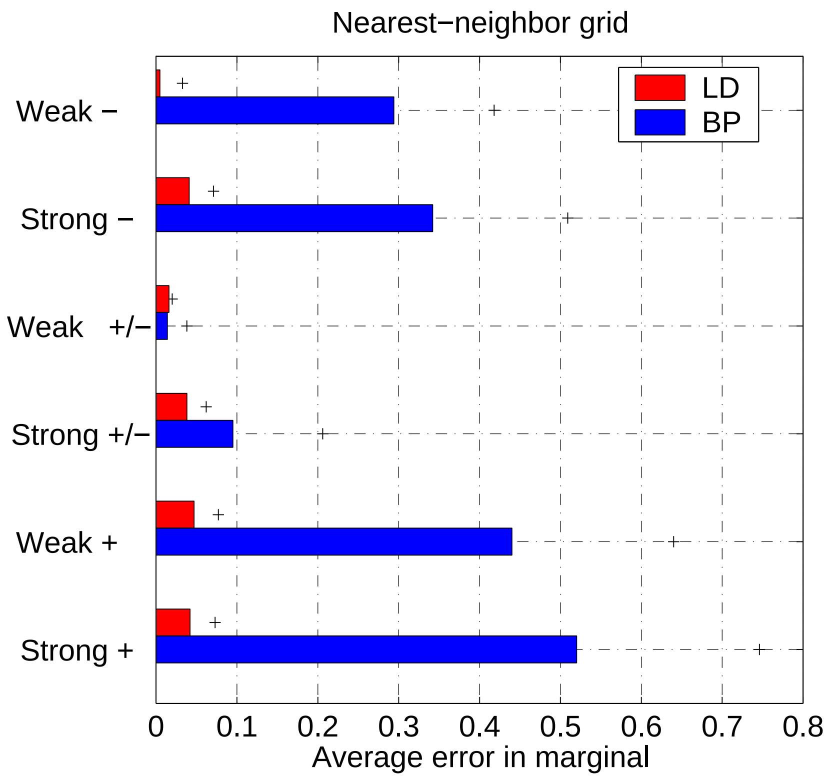
$$\max_{\mu \in \mathrm{OUT}(K_n)} \left\{ \langle \theta,\, \mu \rangle + \frac{1}{2} \log\det \left[M_1(\mu) + \frac{1}{3}\, \mathrm{blkdiag}[0, I_n] \right] \right\} + \frac{n}{2} \log(\frac{\pi e}{2})$$

Remarks:

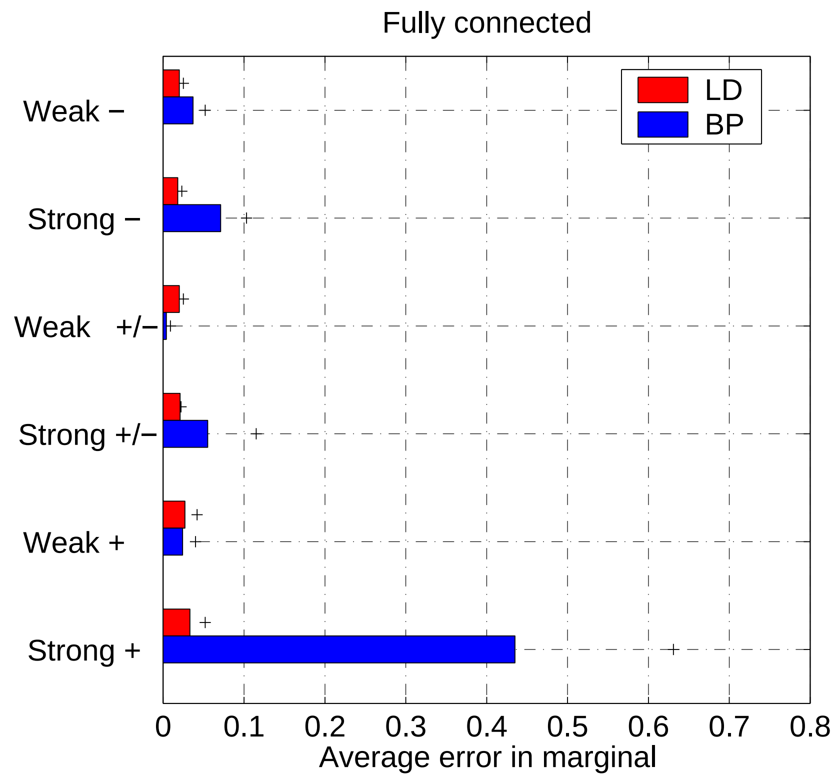
1. Log-det. problem can be solved efficiently by interior point methods.
2. Relevance for applications:
   (a) Upper bound on $A(\theta)$.
   (b) Method for computing approximate marginals.

(Wainwright & Jordan, 2003)

Results for approximating marginals



(a) Nearest-neighbor grid



(b) Fully connected

- average ℓ_1 error in approximate marginals over 100 trials
- coupling types: repulsive $(-)$, mixed $(+/-)$, attractive $(+)$

Summary and future directions

- variational methods are based on converting computational tasks to optimization problems:
  (a) complementary to sampling-based methods (e.g., MCMC)
  (b) a variety of new "relaxations" remain to be explored

- many open questions:
  (a) prior error bounds available only in special cases
  (b) extension to non-parametric settings?
  (c) hybrid techniques (variational and MCMC)
  (d) variational methods in parameter estimation
  (e) fast techniques for solving large-scale relaxations (e.g., SDPs, other convex programs)