Graphical models and variational methods:  
Message-passing, convex relaxations, and all that

Martin Wainwright  
Department of Statistics, and  
Department of Electrical Engineering and Computer Science,  
UC Berkeley, Berkeley, CA USA

Email: wainwrig@{stat,eecs}.berkeley.edu

For further information (tutorial slides, papers, course lectures), see:  
www.eecs.berkeley.edu/~wainwrig/
Introduction

- graphical model:
  * graph $G = (V, E)$ with $N$ vertices
  * random vector: $(X_1, X_2, \ldots, X_N)$

- useful in many statistical and computational fields:
  - machine learning, artificial intelligence
  - computational biology, bioinformatics
  - statistical signal/image processing, spatial statistics
  - statistical physics
  - communication and information theory
Graphs and random variables

- associate to each node $s \in V$ a random variable $X_s$
- for each subset $A \subseteq V$, random vector $X_A := \{X_s, s \in A\}$.

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:

$$p(x) = \frac{1}{Z} \prod_{C \in C} \text{exp} \left\{ \theta_C(x_C) \right\}$$

Normalization compatibility function on clique $C$

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

Given an undirected graphical model (Markov random field):

$$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \exp \{ \theta_C(x_C) \}$$

How to efficiently compute?

- the data likelihood or normalization constant

  **Summation/integration**

  $$Z = \sum_{x \in \mathcal{X}^N} \prod_{C \in \mathcal{C}} \exp \{ \theta_C(x_C) \}$$

- marginal distributions at single sites, or subsets:

  **Summation/integration**

  $$p(X_s = x_s) = \sum_{x_t, t \neq s} \prod_{C \in \mathcal{C}} \exp \{ \theta_C(x_C) \}.$$ 

- most probable configuration (MAP estimate):

  **Maximization**

  $$\hat{x} = \arg \max_{x \in \mathcal{X}^N} p(x) = \arg \max_{x \in \mathcal{X}^N} \prod_{C \in \mathcal{C}} \exp \{ \theta_C(x_C) \}.$$
Variational methods

• "variational": umbrella term for optimization-based formulations
• many modern algorithms are variational in nature:
  – dynamic programming, finite-element methods
  – max-product message-passing
  – sum-product message-passing: generalized belief propagation, convexified belief propagation, expectation-propagation
  – mean field algorithms

Classical example: Courant-Fischer for eigenvalues:

\[ \lambda_{\text{max}}(Q) = \max_{\|x\|_2 = 1} x^T Q x \]

Variational principle: Representation of interesting quantity \( u^* \) as the solution of an optimization problem.

1. \( u^* \) can be analyzed/bounded through “lens” of the optimization
2. approximate \( u^* \) by relaxing the variational principle
§1. Convex relaxations and message-passing for MAP

**Goal:** Compute most probable configuration (MAP estimate) on a tree:

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

Max-product strategy: “Divide and conquer”: break global maximization into simpler sub-problems. (Lauritzen & Spiegelhalter, 1988)
Max-product on trees

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

Update messages:

\[
M_{32}(x_2) = \max_{x_3} \left[ \exp(\theta_3(x_3)) + \theta_{23}(x_2, x_3) \prod_{v \in N(3) \setminus 2} M_{v3}(x_3) \right]
\]
Variational view: Max-product and linear programming

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

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

  $\mu_s(x_s) = \begin{bmatrix} \mu_s(0) \\ \mu_s(1) \\ \mu_s(2) \end{bmatrix}$

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

• alternative formulation of MAP as linear program?

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

  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

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

$$
\begin{aligned}
\{ \vec{\mu} \in \mathbb{R}^d \mid & \mu_s(x_s) = \sum_{x_t, t \neq s} p_\mu(x), \text{ and } \mu_{st}(x_s, x_t) = \sum_{x_u, u \neq s, t} p_\mu(x) \}
\end{aligned}
$$

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

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

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

\[
\mu_s(x_s) = \begin{bmatrix} \mu_s(0) \\ \mu_s(1) \\ \mu_s(2) \end{bmatrix} \quad \mu_{st}(x_s, x_t) = \begin{bmatrix} \mu_{st}(0, 0) & \mu_{st}(0, 1) & \mu_{st}(0, 2) \\ \mu_{st}(1, 0) & \mu_{st}(1, 1) & \mu_{st}(1, 2) \\ \mu_{st}(2, 0) & \mu_{st}(2, 1) & \mu_{st}(2, 2) \end{bmatrix}
\]

Deep fact about tree-structured models: If $\{\mu_s, \mu_{st}\}$ are non-negative and locally consistent:

Normalisation: \[ \sum_{x_s} \mu_s(x_s) = 1 \]

Marginalisation: \[ \sum_{x_t} \mu_{st}(x_s, x'_t) = \mu_s(x_s), \]

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

Follows from junction tree theorem (Lauritzen & Spiegelhalter, 1988).
Max-product on trees: Linear program solver

- MAP problem as a simple linear program:

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

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

\[
 \mathcal{M}(T) = \left\{ \vec{\mu} \geq 0, \sum_{x_s} \mu_s(x_s) = 1, \sum_{x'_t} \mu_{st}(x_s, x'_t) = \mu_s(x_s) \right\}.
\]

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 \(\sum_{x'_t} \mu_{st}(x_s, x'_t) = \mu_s(x_s)\).
Tree-based relaxation for graphs with cycles

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

$$L(G) = \left\{ \vec{\tau} \in \mathbb{R}^d \mid \vec{\tau} \geq 0, \sum_{x_s} \tau_s(x_s) = 1, \sum_{x_t} \tau_{st}(x_s, x_t') = \tau_s(x_s) \right\}.$$ 

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

$$f(\hat{x}) = \max_{\vec{\mu} \in M(G)} \theta^T \vec{\mu} \leq \max_{\vec{\tau} \in L(G)} \theta^T \vec{\tau}.$$
Max-product and graphs with cycles

Early and on-going work:

- single-cycle graphs and Gaussian models

- local optimality guarantees:
  - “tree-plus-loop” neighborhoods (Weiss & Freeman, 2001)
  - optimality on more general sub-graphs (Wainwright et al., 2003)


A natural “variational” conjecture:

- max-product on trees is a method for solving a linear program

- is max-product solving the first-order LP relaxation on graphs with cycles?
Standard analysis via computation tree

- standard tool: computation tree of message-passing updates
  (Gallager, 1963; Weiss, 2001; Richardson & Urbanke, 2001)

(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
Example: Standard max-product does not solve LP

(Wainwright et al., 2005)

Intuition:

- max-product solves (exactly) a modified problem on computation tree
- nodes *not equally weighted* in computation tree ⇒ max-product can output an incorrect configuration

(a) Diamond graph $G_{\text{dia}}$

(b) Computation tree (4 iterations)

- for example: asymptotic node fractions $\omega$ in this computation tree:

$$\begin{bmatrix} \omega(1) & \omega(2) & \omega(3) & \omega(4) \end{bmatrix} = \begin{bmatrix} 0.2393 & 0.2607 & 0.2607 & 0.2393 \end{bmatrix}$$
A whole family of non-exact examples

\[
\begin{align*}
\theta_s(x_s) &= \begin{cases} 
\alpha x_s & \text{if } s = 1 \text{ or } s = 4 \\
\beta x_s & \text{if } s = 2 \text{ or } s = 3 
\end{cases} \\
\theta_{st}(x_s, x_t) &= \begin{cases} 
-\gamma & \text{if } x_s \neq x_t \\
0 & \text{otherwise}
\end{cases}
\end{align*}
\]

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

- first-order LP relaxation always exact for this problem

- max-product and LP relaxation give different decision boundaries:

\[
\begin{align*}
\text{Optimal/LP boundary: } \hat{x} &= \begin{cases} 
1^4 & \text{if } 0.25\alpha + 0.25\beta \geq 0 \\
(-1)^4 & \text{otherwise}
\end{cases} \\
\text{Max-product boundary: } \hat{x} &= \begin{cases} 
1^4 & \text{if } 0.2393\alpha + 0.2607\beta \geq 0 \\
(-1)^4 & \text{otherwise}
\end{cases}
\end{align*}
\]
**Tree-reweighted max-product algorithm**

(Wainwright, Jaakkola & Willsky, 2002)

Message update from node $t$ to node $s$:

$$M_{ts}(x_s) \leftarrow \kappa \max_{x'_t \in X_t} \left\{ \exp \left[ \frac{\theta_{st}(x_s, x'_t)}{\rho_{st}} \right] + \theta_t(x'_t) \right\} \frac{\prod_{v \in \Gamma(t) \setminus s} \left[ M_{vt}(x_t) \right]^{\rho_{vt}}}{\left[ M_{st}(x_t) \right]^{(1 - \rho_{ts})}}.$$ 

Properties:

1. Modified updates remain distributed and purely local over the graph.
   - Messages are reweighted with $\rho_{st} \in [0, 1]$.
2. Key differences:
   - 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.
**Edge appearance probabilities**

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

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

The vector $\rho_e = \{ \rho_e \mid e \in E \}$ must belong to the *spanning tree polytope*.  

(Edmonds, 1971)
TRW max-product and LP relaxation

First-order (tree-based) LP relaxation:

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

Results: (Wainwright et al., 2003; 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 \( y_s = x_s^* \).
On-going work: Distributed methods for solving LPs

• tree-reweighted max-product solves first-order LP for any binary pairwise problem  
  (Kolmogorov & Wainwright, 2005)

• convergent dual ascent scheme; LP-optimal for binary pairwise problems  
  (Globerson & Jaakkola, 2007)

• convex free energies and zero-temperature limits  
  (Wainwright et al., 2005, Weiss et al., 2006; Johnson et al., 2007)

• coding problems: adaptive cutting-plane methods  
  (Taghavi & Siegel, 2006; Dimakis et al., 2006)

• arbitrary problems: proximal minimization and rounding schemes with correctness guarantees  
  (Ravikumar et al., ICML 2008)
Hierarchies of conic programming relaxations


- hierarchies of SDP relaxations for polynomial programming (Lasserre, 2001; Parrilo, 2002)

- intermediate between LP and SDP: second-order cone programming (SOCP) relaxations (Ravikumar & Lafferty, 2006; Pawan et al., 2008)

- all relaxations: particular outer bounds on the marginal polyope

Key questions:

- when are particular relaxations tight?

- when does more computation (e.g., LP $\rightarrow$ SDP) yield performance gains?
Variational principles for marginalization/summation

Undirected graphical model:

\[ p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \exp \{ \theta_C(x_C) \}. \]

Core computational challenges

(a) computing most probable configurations \( \hat{x} \in \arg \max_{x \in \mathcal{X}^N} p(x) \)

(b) computing normalization constant \( Z \)

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

Variational formulation of problems (b) and (c): not immediately obvious!

Approach: Develop variational representations using exponential families, and convex duality.
Maximum entropy formulation of graphical models

- suppose that we have measurements $\hat{\mu}$ of the average values of some (local) functions $\phi_\alpha : \mathcal{X}^n \to \mathbb{R}$

- in general, will be many distributions $p$ that satisfy the measurement constraints $\mathbb{E}_p[\phi_\alpha(x)] = \hat{\mu}$

- will consider finding the $p$ with maximum “uncertainty” subject to the observations, with uncertainty measured by entropy

$$H(p) = -\sum_x p(x) \log p(x).$$

Constrained maximum entropy problem: Find $\hat{p}$ to solve

$$\max_{p \in \mathcal{P}} H(p) \quad \text{such that} \quad \mathbb{E}_p[\phi_\alpha(x)] = \hat{\mu}$$

- elementary argument with Lagrange multipliers shows that solution belongs to exponential family

$$\hat{p}(x; \theta) \propto \exp \left\{ \sum_{\alpha \in \mathcal{I}} \theta_\alpha \phi_\alpha(x) \right\}. $$
### Examples: Scalar exponential families

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

- **Parameterized family of densities (w.r.t. some base measure)**

$$p(x; \theta) = \exp \left\{ \sum_{\alpha} \theta_\alpha \phi_\alpha(x) - A(\theta) \right\}$$

- **Cumulant generating function** (log normalization constant):

$$A(\theta) = \log \left( \int \exp\{\langle \theta, \phi(x) \rangle\} \nu(dx) \right)$$
Example: Discrete Markov random field

Indicators: \[ \mathbb{I}_j(x_s) = \begin{cases} 1 & \text{if } x_s = j \\ 0 & \text{otherwise} \end{cases} \]

Parameters: \[ \theta_s = \{\theta_{s;j}, j \in \mathcal{X}_s\} \]
\[ \theta_{st} = \{\theta_{st;jk}, (j, k) \in \mathcal{X}_s \times \mathcal{X}_t\} \]

Compact form: \[ \theta_s(x_s) := \sum_j \theta_{s;j}\mathbb{I}_j(x_s) \]
\[ \theta_{st}(x_s, x_t) := \sum_{j,k} \theta_{st;jk}\mathbb{I}_j(x_s)\mathbb{I}_k(x_t) \]

Probability mass function of form:
\[ p(x; \theta) \propto \exp \left\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \right\} \]

Cumulant generating function (log normalization constant):
\[ A(\theta) = \log \sum_{x \in \mathcal{X}^n} \exp \left\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \right\} \]
Special case: Hidden Markov model

- Markov chain \( \{X_1, X_2, \ldots \} \) evolving in time, with noisy observation \( Y_t \) at each time \( t \)

\[
\begin{align*}
\theta_{23}(x_2, x_3) &= \log p(x_3 | x_2) \\
\theta_5(x_5) &= \log p(y_5 | x_5)
\end{align*}
\]

- an HMM is a particular type of discrete MRF, representing the conditional \( p(x | y; \theta) \)

- exponential parameters have a concrete interpretation

\[
\begin{align*}
\theta_{23}(x_2, x_3) &= \log p(x_3 | x_2) \\
\theta_5(x_5) &= \log p(y_5 | x_5)
\end{align*}
\]

- the cumulant generating function \( A(\theta) \) is equal to the log likelihood \( \log p(y; \theta) \)
Example: Multivariate Gaussian

\( U(\theta) \): Matrix of natural parameters  \( \phi(x) \): Matrix of sufficient statistics

\[
\begin{bmatrix}
0 & \theta_1 & \theta_2 & \ldots & \theta_n \\
\theta_1 & \theta_{11} & \theta_{12} & \ldots & \theta_{1n} \\
\theta_2 & \theta_{21} & \theta_{22} & \ldots & \theta_{2n} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\theta_n & \theta_{n1} & \theta_{n2} & \ldots & \theta_{nn}
\end{bmatrix}
\]

\[
\begin{bmatrix}
1 & x_1 & x_2 & \ldots & x_n \\
x_1 & (x_1)^2 & x_1x_2 & \ldots & x_1x_n \\
x_2 & x_2x_1 & (x_2)^2 & \ldots & x_2x_n \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
x_n & x_nx_1 & x_nx_2 & \ldots & (x_n)^2
\end{bmatrix}
\]

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

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

1 2 3 4 5
1 2 3 4 5
4 5

28
Example: Mixture of Gaussians

- can form *mixture models* by combining different types of random variables

- let \( Y_s \) be conditionally Gaussian given the discrete variable \( X_s \) with parameters \( \gamma_{s;j} = (\mu_{s;j}, \sigma^2_{s;j}) \):
  \[
  X_s \quad \bigcirc \quad p(x_s; \theta_s)
  \]
  \[
  p(y_s \mid x_s; \gamma_s)
  \]
  \[
  Y_s \quad \bigcirc 
  \]
  \[X_s \equiv \text{mixture indicator}\]
  \[Y_s \equiv \text{mixture of Gaussian}\]

- couple the mixture indicators \( X = \{X_s, s \in V\} \) using a discrete MRF

- overall model has the exponential form
  \[
p(y, x; \theta, \gamma) \propto \prod_{s \in V} p(y_s \mid x_s; \gamma_s) \exp \left\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \right\}.
  \]
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:

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

- 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:

$$f(u) = \sup_{v \in \mathbb{R}^n} \{ \langle u, v \rangle - f^*(v) \}.$$
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 $\text{epi}(f)$?

**Epigraph of $f$:**

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

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

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

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

$$c^* = \sup_{u \in \mathbb{R}^n} \{\langle v, u \rangle - f(u)\}.$$
Example: Single Bernoulli

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

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

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

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

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:

$$A(\theta) = \max_{\mu \in [0,1]} \{ \mu \cdot \theta - A^*(\mu) \}.$$
More general computation of the dual $A^*$

- consider the definition of the dual function:
  \[ A^*(\mu) = \sup_{\theta \in \mathbb{R}^d} \{ \langle \mu, \theta \rangle - A(\theta) \}. \]

- taking derivatives w.r.t $\theta$ to find a stationary point yields:
  \[ \mu - \nabla A(\theta) = 0. \]

- **Useful fact:** Derivatives of $A$ yield *mean parameters:*
  \[ \frac{\partial A}{\partial \theta_\alpha}(\theta) = \mathbb{E}_\theta[\phi_\alpha(X)] := \int \phi_\alpha(x)p(x; \theta)\nu(x). \]

Thus, stationary points satisfy the equation:
\[ \mu = \mathbb{E}_\theta[\phi(X)] \] (1)
Computation of dual (continued)

• assume solution $\theta(\mu)$ to equation $\mu = \mathbb{E}_\theta[\phi(X)]$ (*)

• strict concavity of objective guarantees that $\theta(\mu)$ attains global maximum with value

$$A^*(\mu) = \langle \mu, \theta(\mu) \rangle - A(\theta(\mu))$$

$$= \mathbb{E}_{\theta(\mu)} \left[ \langle \theta(\mu), \phi(X) \rangle - A(\theta(\mu)) \right]$$

$$= \mathbb{E}_{\theta(\mu)} [\log p(X; \theta(\mu))]$$

• recall the definition of entropy:

$$H(p(x)) := -\int \left[ \log p(x) \right] p(x) \nu(dx)$$

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

**Question:** For which $\mu \in \mathbb{R}^d$ does equation (*) 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:
  \[
  \mu_\alpha := \int \phi_\alpha(x)p(x)\nu(dx)
  \]

- now consider the set $\mathcal{M}(G; \phi)$ of all realizable mean parameters:
  \[
  \mathcal{M}(G; \phi) = \{ \mu \in \mathbb{R}^d \mid \mu_\alpha = \int \phi_\alpha(x)p(x)\nu(dx) \text{ for some } p(\cdot) \}\n  \]

- for discrete families, we refer to this set as a marginal polytope (as discussed previously)
Examples of $\mathbb{M}$: Gaussian MRF

$\phi(x)$ Matrix of sufficient statistics

\[
\begin{bmatrix}
1 & x_1 & x_2 & \ldots & x_n \\
x_1 & (x_1)^2 & x_1x_2 & \ldots & x_1x_n \\
x_2 & x_2x_1 & (x_2)^2 & \ldots & x_2x_n \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
x_n & x_nx_1 & x_nx_2 & \ldots & (x_n)^2
\end{bmatrix}
\]

$U(\mu)$ Matrix of mean parameters

\[
\begin{bmatrix}
1 & \mu_1 & \mu_2 & \ldots & \mu_n \\
\mu_1 & \mu_{11} & \mu_{12} & \ldots & \mu_{1n} \\
\mu_2 & \mu_{21} & \mu_{22} & \ldots & \mu_{2n} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\mu_n & \mu_{n1} & \mu_{n2} & \ldots & \mu_{nn}
\end{bmatrix}
\]

- Gaussian mean parameters are specified by a single semidefinite constraint as $\mathbb{M}_{\text{Gauss}} = \{\mu \in \mathbb{R}^{n+\binom{n}{2}} | U(\mu) \succeq 0\}$.

Scalar case:

$U(\mu) = \begin{bmatrix} 1 & \mu_1 \\ \mu_1 & \mu_{11} \end{bmatrix}$
Examples of $\mathbb{M}$: Discrete MRF

- **sufficient statistics:**
  \[
  \mathbb{I}_j(x_s) \quad \text{for } s = 1, \ldots, n, \quad j \in \mathcal{X}_s \\
  \mathbb{I}_{jk}(x_s, x_t) \quad \text{for } (s, t) \in E, \quad (j, k) \in \mathcal{X}_s \times \mathcal{X}_t
  \]

- **mean parameters are simply marginal probabilities, represented as:**
  \[
  \mu_s(x_s) := \sum_{j \in \mathcal{X}_s} \mu_{s;j} \mathbb{I}_j(x_s), \quad \mu_{st}(x_s, x_t) := \sum_{(j, k) \in \mathcal{X}_s \times \mathcal{X}_t} \mu_{st;jk} \mathbb{I}_{jk}(x_s, x_t)
  \]

- denote the set of realizable $\mu_s$ and $\mu_{st}$ by $\mathbb{M}(G)$
- refer to it as the *marginal polytope*
- extremely difficult to characterize for general graphs
For suitable classes of graphical models in exponential form, the gradient map $\nabla A$ is a bijection between $\Theta$ and the interior of $M$.

(e.g., Brown, 1986; Efron, 1978)
Variational principle in terms of mean parameters

- The conjugate dual of $A$ takes the form:

$$A^*(\mu) = \begin{cases} -H(p(x; \theta(\mu))) & \text{if } \mu \in \text{int } \mathbb{M}(G; \phi) \\ +\infty & \text{if } \mu /\in \text{cl } \mathbb{M}(G; \phi) \end{cases}$$

*Interpretation:*
- $A^*(\mu)$ is finite (and equal to a certain negative entropy) for any $\mu$ that is globally realizable
- if $\mu /\in \text{cl } \mathbb{M}(G; \phi)$, then the max. entropy problem is *infeasible*

- The cumulant generating function $A$ has the representation:

$$A(\theta) = \sup_{\mu \in \mathbb{M}(G; \phi)} \left\{ \langle \theta, \mu \rangle - A^*(\mu) \right\}$$

  cumulant generating func.  \quad max. ent. problem over $\mathbb{M}$

- in contrast to the “free energy” approach, solving this problem provides both the value $A(\theta)$ and the exact mean parameters $\hat{\mu}_\alpha = \mathbb{E}_\theta [\phi_\alpha(x)]$
**Alternative view: Kullback-Leibler divergence**

- Kullback-Leibler divergence defines “distance” between probability distributions:

\[
D(p || q) := \int \left[ \log \frac{p(x)}{q(x)} \right] p(x) \nu(dx)
\]

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

\[
D(p(x; \theta^1) || p(x; \theta^2)) = A(\theta^2) - A(\theta^1) - \langle \mu^1, \theta^2 - \theta^1 \rangle
\]

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

\[
D(p(x; \theta^1) || p(x; \theta^2)) \equiv D(\mu^1 || \theta^2) = A(\theta^2) + A^*(\mu^1) - \langle \mu^1, \theta^2 \rangle
\]

Hence, the following two assertions are equivalent:

\[
A(\theta^2) = \sup_{\mu^1 \in M(G; \phi)} \left\{ \langle \theta^2, \mu^1 \rangle - A^*(\mu^1) \right\}
\]

\[
0 = \inf_{\mu^1 \in M(G; \phi)} D(\mu^1 || \theta^2)
\]
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 $\mu$ 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.
Example: Multivariate Gaussian (fixed covariance)

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

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

$$A(\theta) = \log \int_{\mathbb{R}^n} \exp \left\{ \sum_{s=1}^{n} \theta_s x_s \right\} \exp \left\{ -\frac{1}{2} x^T Q x \right\} dx$$

- completing the square yields $A(\theta) = \frac{1}{2} \theta^T Q^{-1} \theta + \text{constant}$
- straightforward computation leads to the dual

$$A^*(\mu) = \frac{1}{2} \mu^T Q \mu - \text{constant}$$
- putting the pieces back together yields the variational principle

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

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

42
Example: Multivariate Gaussian (arbitrary cov.)

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

\[
\phi(X) = \begin{bmatrix} 1 \\ X \end{bmatrix} \begin{bmatrix} 1 & X \end{bmatrix}, \quad U(\theta) := \begin{bmatrix} 0 & \theta_s \\ \theta_s & \theta_{st} \end{bmatrix} \quad U(\mu) := \mathbb{E}\left\{ \begin{bmatrix} 1 \\ X \end{bmatrix} \begin{bmatrix} 1 & X \end{bmatrix} \right\}
\]

- cumulant generating function:

\[
A(\theta) = \log \int \exp \left\{ \text{trace}(U(\theta) \phi(X)) \right\} dX
\]

- computing the dual function:

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

- exact variational principle is a log-determinant problem:

\[
A(\theta) = \sup_{U(\mu) > 0, [U(\mu)]_{11} = 1} \left\{ \text{trace}(U(\theta) U(\mu)) + \frac{1}{2} \log \det U(\mu) \right\} + C.
\]

- solution yields the normal equations for Gaussian mean and covariance.
Example: Belief propagation and Bethe principle

Problem set-up

• discrete variables $X_s \in \{0, 1, \ldots, m_s - 1\}$ on graph $G = (V, E)$

• sufficient statistics: indicator functions for each node and edge

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

• exponential representation of distribution:

\[ p(x; \theta) \propto \exp \left\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s, t) \in E} \theta_{st}(x_s, x_t) \right\} \]

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

Two main ingredients:

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

2. The marginal polytope $\mathcal{M}(G)$ is also difficult to characterize, so let’s use the tree-based outer bound $\mathcal{L}(G)$. 
Bethe entropy approximation

- mean parameters are simply marginal probabilities, represented as:
  \[ \mu_s(x_s) := \sum_{j \in \mathcal{X}_s} \mu_{s;j}(x_s), \quad \mu_{st}(x_s, x_t) := \sum_{(j,k) \in \mathcal{X}_s \times \mathcal{X}_t} \mu_{st;jk}(x_s, x_t) \]

- Bethe entropy approximation
  \[ -A_{Bethe}^*(\mu) = \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st}), \]
  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 for trees, using the factorization:
  \[ p(x; \theta) = \prod_{s \in V} \mu_s(x_s) \prod_{(s,t) \in E} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s) \mu_t(x_t)} \]
Bethe variational principle

- Bethe entropy approximation, and outer bound $\mathbb{L}(G)$:

$$\mathbb{L}(G) = \left\{ \bar{\tau} \mid \sum_{x_s} \tau_s(x_s) = 1, \sum_{x'_t} \tau_{st}(x_s, x'_t) = \tau_s(x_s) \right\}.$$ 

- Combining these ingredients leads to the Bethe variational problem (BVP):

$$\max_{\tau \in \mathbb{L}(G)} \left\{ \langle \theta, \tau \rangle + \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st}) \right\}.$$ 

**Key fact:** Belief propagation can be derived as an iterative method for solving a Lagrangian formulation of the BVP (Yedidia et al., 2002)
Lagrangian derivation of belief propagation

• 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:

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

- taking derivatives of the Lagrangian w.r.t $\tau_s$ and $\tau_{st}$ yields

\[
\frac{\partial L}{\partial \tau_s(x_s)} = \theta_s(x_s) - \log \tau_s(x_s) + \sum_{t\in\mathcal{N}(s)} \lambda_{ts}(x_s) + C
\]

\[
\frac{\partial L}{\partial \tau_{st}(x_s, x_t)} = \theta_{st}(x_s, x_t) - \log \frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s)\tau_t(x_t)} - \lambda_{ts}(x_s) - \lambda_{st}(x_t) + C'
\]

- setting these partial derivatives to zero and simplifying:

\[
\tau_s(x_s) \propto \exp \left\{ \theta_s(x_s) \right\} \prod_{t\in\mathcal{N}(s)} \exp \left\{ \lambda_{ts}(x_s) \right\}
\]

\[
\tau_s(x_s, x_t) \propto \exp \left\{ \theta_s(x_s) + \theta_t(x_t) + \theta_{st}(x_s, x_t) \right\} \times \prod_{u\in\mathcal{N}(s)\setminus t} \exp \left\{ \lambda_{us}(x_s) \right\} \prod_{v\in\mathcal{N}(t)\setminus s} \exp \left\{ \lambda_{vt}(x_t) \right\}
\]

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

\[
M_{ts}(x_s) \leftarrow \sum_{x_t} \exp \{ \theta_t(x_t) + \theta_{st}(x_s, x_t) \} \prod_{u\in\mathcal{N}(t)\setminus s} M_{ut}(x_t)
\]
Geometry of Bethe variational problem

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

**Natural question:** Do BP fixed points ever fall outside of the marginal polytope $\mathcal{M}(G)$?
Illustration: Globally inconsistent BP fixed points

Consider the following assignment of pseudomarginals $\tau_s, \tau_{st}$:

\begin{align*}
\tau \in \mathbb{L}(G^t), \text{ and that }\tau \text{ is a fixed point of belief propagation (with all constant messages)}
\end{align*}

• however, $\tau$ is globally inconsistent

Note: More generally: for any $\tau$ in the interior of $\mathbb{L}(G^t)$, can construct a distribution with $\tau$ 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)
Generalized belief propagation on hypergraphs

(Yedidia et al., 2002)

- 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 $V$

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

- ancestor/descendant relationships:
  - $g \subset h$ if $g$ is contained within hyperedge $h$
  - $g \supset h$ for opposite relationship
Hypertree factorization

- for each hyperedge: $\log \varphi_h(x_h) := \sum_{g \subseteq h} (-1)^{|h\setminus g|} [\log \tau_g(x_g)]$.

- any hypertree-structured distribution is guaranteed to factor as:
  $$p(x) = \prod_{h \in E} \varphi_h(x_h).$$

- **Ordinary tree:**
  $$\varphi_s(x_s) = \mu_s(x_s) \quad \text{for any vertex } s$$
  $$\varphi_{st}(x_s, x_t) = \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s) \mu_t(x_t)} \quad \text{for edge } (s, t).$$

- **Hypertree:**

  $\varphi_{1245} = \frac{\mu_{1245}}{\mu_{25} \mu_{45} \mu_5}$
  $\varphi_{45} = \frac{\mu_{45}}{\mu_5}$
  $\varphi_5 = \mu_5$
Building augmented hypergraphs

Better entropy approximations via augmented hypergraphs.

(a) Original  (b) Clustering  (c) Full covering

(d) Kikuchi  (e) Fails single counting
Expectation-propagation (EP)

- originally derived in terms of assumed density filtering (Minka, 2002)
- another instance of a relaxed variational principle:
  - “Bethe-like” (termwise) approximation to entropy
  - local consistency constraints on marginals
- distribution with tractable/intractable decomposition:

\[
f(x, \gamma, \Gamma) \propto \exp(\langle \gamma, \phi(x) \rangle) \prod_{i=1}^{k} T_i(x)\]

\[\begin{array}{ll}
\text{Tractable} & \text{Intractable}
\end{array}\]

- auxiliary parameters \(\theta\), and term-by-term entropy approx.: 

\[
H(f) \approx H(q_{base}(x; \theta, \gamma)) + \sum_{i=1}^{k} \left[ H(q_{aug}^{i}(x; \theta, \gamma, T_i)) - H(q_{base}(x; \theta, \gamma)) \right]
\]

\[\begin{array}{ll}
\text{Base entropy} & \text{Term approximations}
\end{array}\]
EP updates for Gaussian mixtures

- distribution formed by tractable/intractable combination:

\[ f(x, \Sigma) \propto \exp \left( -\frac{1}{2} x^T \Sigma^{-1} x \right) \prod_{i=1}^{n} f(y^i \mid X = x) \]

- Gaussian mixture likelihoods

\[ f(y^i \mid X = x) = \alpha \mathcal{N}(y^i; 0, \sigma_0^2) + (1 - \alpha) \mathcal{N}(y^i; x, \sigma_1^2) \]

- base/augmented distributions take form:

Base: \[ q_{base}(x; \Sigma, \theta, \Theta) \propto \exp \left( \langle \gamma, x \rangle - \frac{1}{2} \text{trace}(\Theta + \Sigma^{-1} xx^T) \right) \]

Augmented: \[ q_{aug}^i(x; \Sigma, \theta, \Theta, T_i) \propto q(x; \Sigma, \theta, \Theta) T_i(x). \]

- variational problem: maximize term-by-term entropy approximation, subject to marginalization constraints:

\[ \mathbb{E}_{q_{base}}[X] = \mathbb{E}_{q_{aug}^i}[X] \]

\[ \mathbb{E}_{q_{base}}[XX^T] = \mathbb{E}_{q_{aug}^i}[XX^T]. \]
Convex relaxations and upper bounds

Possible concerns with Bethe/Kikuchi, expectation-propagation etc.?

(a) lack of convexity $\Rightarrow$ multiple local optima, and 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.

\[-A^*(\mu) \leq -\rho(T^1)A^*(\mu(T^1)) - \rho(T^2)A^*(\mu(T^2)) - \rho(T^3)A^*(\mu(T^3))\]

- given any spanning tree \(T\), define the moment-matched tree distribution:
  \[p(x; \mu(T)) := \prod_{s \in V} \mu_s(x_s) \prod_{(s,t) \in E} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s) \mu_t(x_t)}\]

- use \(-A^*(\mu(T))\) to denote the associated tree entropy
- let \(\rho = \{\rho(T)\}\) be a probability distribution over spanning trees
Optimal bounds by tree-reweighted message-passing

Recall the constraint set of locally consistent marginal distributions:
\[
\mathbb{L}(G) = \{ \tau \geq 0 \mid \sum_{x_s} \tau_s(x_s) = 1, \sum_{x_s} \tau_{st}(x_s, x_t) = \tau_t(x_t) \}.
\]

**Theorem:**  
(Wainwright et al., UAI-02)

(a) For any given edge weights \( \rho_e = \{ \rho_e \} \) in the spanning tree polytope, the optimal upper bound over all tree parameters is given by:
\[
A(\theta) \leq \max_{\tau \in \mathbb{L}(G)} \{ \langle \theta, \tau \rangle + \sum_{s \in V} H_s(\tau_s) - \sum_{(s, t) \in E} \rho_{st} I_{st}(\tau_{st}) \}.
\]

(b) This optimization problem is strictly convex, and its unique optimum is specified by the fixed point of \( \rho_e \)-reweighted sum-product:
\[
M_{ts}^*(x_s) = \kappa \sum_{x_t' \in \mathcal{X}_t} \left\{ \exp \left[ \frac{\theta_{st}(x_s, x_t')}{\rho_{st}} + \theta_t(x_t') \right] \frac{\prod_{v \in \Gamma(t) \setminus s} [M_{vt}^*(x_t)]^{\rho_{vt}}}{[M_{st}^*(x_t)]^{(1-\rho_{ts})}} \right\}.
\]
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:** *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:

$$
\mathbb{E}[\mathbf{Y}\mathbf{Y}^T] = 
\begin{bmatrix}
1 & \mu_1 & \mu_2 & \cdots & \mu_n \\
\mu_1 & \mu_1 & \mu_{12} & \cdots & \mu_{1n} \\
\mu_2 & \mu_{12} & \mu_2 & \cdots & \mu_{2n} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\mu_n & \mu_{n1} & \mu_{n2} & \cdots & \mu_n
\end{bmatrix} = M_1[\mu].
$$

- since it must be positive semidefinite, this (an infinite number of) linear constraints on $\mu_s, \mu_{st}$.
- defines the *first-order semidefinite relaxation of* $\mathbb{M}(G)$:

$$
\mathbb{S}(G) = \left\{ \mu \in \mathbb{R}^d \mid M_1[\mu] \succeq 0 \right\}.
$$
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 $M_1(\mu) \in \mathbb{S}_1(K_n)$

**Theorem:** Consider an outer bound $\mathbb{O}(K_n)$ that satisfies:

$$\mathbb{M}(K_n) \subseteq \mathbb{O}(K_n) \subseteq \mathbb{S}_1(K_n)$$

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

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

**Remarks:**

1. Log-det. problem can be solved efficiently by interior point methods.
2. Relevance for applications (e.g., Banerjee et al., 2008)
   (a) Upper bound on $A(\theta)$.
   (b) Method for computing approximate marginals.

(Wainwright & Jordan, 2003)
Mean field theory

Recap: All variational methods discussed until now are based on:

- *outer bounding* the set of valid mean parameters.
- approximating the entropy (negative dual function $-A^*(\mu)$)

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

Examples:

(a) For product distributions $p(x) = \prod_{s \in V} \mu_s(x_s)$, 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 $M_{tr}(G; H)$ represent tractable mean parameters:
  
  $$M_{tr}(G; H) := \{\mu | \mu = \mathbb{E}_\theta[\phi(x)] \text{ s.t. } \theta \text{ respects } H\}.$$

- under mild conditions, $M_{tr}$ is a non-convex *inner approximation* to $M$

- optimizing over $M_{tr}$ (as opposed to $M$) yields lower bound:

  $$A(\theta) \geq \sup_{\tilde{\mu} \in M_{tr}} \left\{ \langle \theta, \tilde{\mu} \rangle - A^*(\tilde{\mu}) \right\}.$$
Alternative view: Minimizing KL divergence

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

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

• try to find the “best” approximation to $p(x; \theta)$ in the sense of KL divergence

• in analytical terms, the problem of interest is

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

• hence, finding the tightest lower bound on $A(\theta)$ is equivalent to finding the best approximation to $p(x; \theta)$ from distributions with $\tilde{\mu} \in \mathcal{M}_{tr}$
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 $p(x; \theta) = \prod_{s \in V} p(x_s; \theta_s)$, and generate
  
  $$\mathbb{M}_{tr}(G; H) = \{ \mu \mid \mu_{st} = \mu_s \mu_t, \ \mu_s \in [0, 1] \}.$$

- approximate variational principle:

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

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

  $$\mu_s \leftarrow \left\{ 1 + \exp\left[ - \left( \theta_s + \sum_{t \in \mathcal{N}(s)} \theta_{st} \mu_t \right) \right] \right\}^{-1}$$
Example: Structured mean field for coupled HMM

- 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)
Summary and future directions

• variational methods: statistical/computational tasks converted to optimization problems:
  (a) complementary to sampling-based methods (e.g., MCMC)
  (b) require entropy approximations, and characterization of marginal polytopes (sets of valid mean parameters)
  (c) a variety of new “relaxations” remain to be explored

• many open questions:
  (a) strong performance guarantees? (only for special cases thus far...)
  (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)