Semidefinite relaxations for approximate inference on graphs with cycles
Introduction

Graphical models are used and studied in various fields (e.g., machine learning; error-correcting coding; statistical physics; computer vision) and the following problems are important but difficult:

(a) computing marginal distributions
(b) estimating model parameters from data
(c) mean field methods (e.g., Jordan et al., 1999)
(b) Bethe/Kikuchi approximations and variations

Role of variational methods

Following problems are important but difficult:

Computer vision

Error-correcting coding; statistical physics;

Graphical models are used and studied in various fields.
Possible concerns with the Bethe/Kikuchi problems and variational algorithms:

(a) Lack of convexity
(b) Failure to bound the log partition function
(c) Multiple local optima, and substantial algorithmic complications

Overview
Overview

Possible concerns with the Bethe/Kikuchi problems and variations?

(a) lack of convexity
(b) failure to bound the Log partition function

Algorithmic complications

Goal: Techniques for approximate inference and parameter estimation based on:

(a) unique global optimum
(b) convex variational problems

Goal: Techniques for approximate inference and parameter estimation based on:

(a) unique global optimum
(b) convex variational problems

Upper bound on Log partition function

Possible concerns with the Bethe/Kikuchi problems and variations?
Variational approach

**Basic idea:** Represent a quantity of interest \( \phi(x) \) as the solution of an optimization problem:

(a) Study via the optimization problem.

(b) Approximate \( \phi(x) \) by approximating the optimization problem.

**Goal:** Obtain a variational representation for:

(a) The log partition function.

(b) The inference problem of computing

\[
\int_x \phi(x) \, d^x x = \eta \]

**Varational approach**
Classical form of convex duality

\[ 0 = (d \parallel b) D^{d \in b} \]

Equivalent to the assertion min

\[ + \left( x \right)^{b \in \log} \left( x \right)_{b} \sum_{x} =: (b)H \]

where is the usual (Boltzmann-Shannon) entropy

\[ \left\{ (b)H + \left[ \left( x \right)^{\phi \in \theta} \sum_{x} \right] \left( x \right)_{b} \sum_{x} \right\}^{d \in b} \max_{x} = \log D \]

Problem over \( D \)

Log partition function can be recovered as a maximum entropy

Let be the set of all possible distributions over \( x \)

Classical form of convex duality
Exponential representations

Parameterized family of distributions:

\[ p(x; \mu) = \exp^{-\mu x} \]

Log partition function:

\[ \log p(x; \mu) = \sum_{\mathcal{X} \subseteq \mathcal{X}} \exp \left( \sum_{\mathcal{X} \subseteq \mathcal{X}} \mu \mathcal{X} \phi \right) \]

Weights on potentials:

\[ \equiv \left\{ \mathcal{I} \ni \nu : \nu \theta \right\} = \theta \]

Collection of potential functions:

\[ \equiv \left\{ \mathcal{I} \ni \nu : \nu \phi \right\} = \phi \]

\[ \log \Phi = (\theta) \Phi \]

Log partition function:

\[ \log \Phi = (\theta) \Phi \]

Parameterized family of distributions:

\[ \exp(\theta; x) \]

Exponential representations
\[ \left\{ (\theta)\Phi - t^s x^s \sum_{s} + s^s x \sum_{s} \right\} dx \varepsilon = (\theta, x)d \]

\[ u \{ 1, 0 \} = u\mathcal{X} \]
\[ \mathcal{E} \cap \Lambda = \mathcal{I} \]
\[ \{ \mathcal{E} \in (t,s) | t^s x \} \cap \{ \Lambda \in s | s x \} = \phi \]

Binary variables on a graph with pairwise cliques

Special case: Ising model
An alternative view

Idea:
Think about optimization not in terms of distributions \( p \), but rather in terms of only the mean parameters:

\[
\{ ( \cdot ) d \text{ for some } x \in \mathbb{R}^d \mid \int_{\mathbb{R}^d} x d(\cdot) = \eta \} = (\phi; M(\mathcal{C}) )
\]

\[ M(\mathcal{C}) = \begin{cases} 
\text{marginals} \\
\text{A marginal polytope is a set of realizable or globally consistent marginals.} 
\end{cases} \]

Question: What is the relevant constraint set?

\[
( \mathbf{x} )^{ \odot \phi ( \mathbf{x} ) d} =: \odot \eta
\]

but rather in terms of only the mean parameters:

Idea: Think about optimization not in terms of distributions \( p \).
Potentials

\[ \left[ \lambda^s x \right] \theta \mathbb{E} = \lambda^s r \quad \left[ \gamma^s x \right] \theta \mathbb{E} = \gamma^s r \]

Relevant marginals

\( \{ \mathcal{E} \in (\gamma, s) \mid \lambda^s x \} \cap \{ \Lambda \in s \mid \gamma^s x \} = \emptyset \)

Associated constraint set is known as the correlation polytope or the binary quadratic polytope (e.g., Deza & Laurent, 1997).

Ising model example
Geometry and moment mapping
\[
[(\mathbf{x}, \phi)]_{\theta} = (\mathbf{x}, \phi(\theta; \mathbf{x}) d \mathbf{x} \in \mathcal{X}) = \phi_{\eta}
\]

Moreover, maximum is attained uniquely at desired marginals:

Log partition function

maximum entropy problem over marginal polytope

\[
\max_{(\eta) \in \text{MARG}(G; \mathcal{A})} \{ (\eta) \Phi - \langle \theta, \eta \rangle \} = (\theta) \Phi
\]

leads to a representation of \( \Phi \) in terms of \( \Phi \):

otherwise,

\[
(\phi; \mathcal{C}) \in \eta, \text{ if } \eta \in \text{MARG} \left( (\eta) \theta; \mathbf{x} \right) d \mathbf{x} \in \mathcal{X} \setminus \mathcal{C} \right\} = (\eta) \Phi
\]

the dual to has the form:

Variational principle in terms of marginals
Convex relaxations

Strategy:
Obtain upper bounds by relaxation of original problem.

Requirements:
(a) convex outer approximation to marginal polytope
(b) concave upper bound on entropy function $-\Phi'$.
(c) combination of semidefinite and hypertree methods
(q) semidefinite methods
(p) tree and hypertree approaches (Bethe/Kikuchi etc.)

Tools:

(a) tree and hypertree approaches (Bethe/Kikuchi etc.)
(b) semidefinite methods
(c) combination of semidefinite and hypertree methods

Results:
Obtain upper bounds by relaxation of original problem.
Semidefinite outer bounds on binary marginal polytope.

\begin{equation}
(\forall \theta, s, t) \quad [x^s x^t]^{\theta} = \theta^{st}
\end{equation}

Relevant marginals:

\begin{align*}
\eta_{1} &= E\left[ x^{s} | E\right] = E^{s} \\
\text{for all } s \\
\eta_{1} &= E\left[ x | E\right] = E
\end{align*}

Refer to the associated marginal polytope as \text{MARG}(K^u).

$q$) complete graph $K^u$ on $u$ nodes.

Focus on: (a) binary case with “spins” $x \in \{-1, +1\}$
First order: Optimizing over covariance matrices

The covariance matrix of $\mathbf{x}$ must be positive semidefinite:

$$0 \preceq [L \mathbf{x}] \mathbb{E} [\mathbf{x}] \mathbb{E} - [L \mathbf{x}] \mathbb{E} = (\mathbf{x}) \text{cov}$$

By Schur complement, equivalent to enforce PSD constraint on

$$\begin{bmatrix}
  I & \cdots & u_3 & u_2 & u_1 \\
  \vdots & \ddots & \vdots & \vdots & \vdots \\
  u_3 & \cdots & I & 1 & 1 \\
  u_2 & \cdots & 1 & I & 1 \\
  u_1 & \cdots & 1 & 1 & I \\
  \vdots & \vdots & \vdots & \vdots & \vdots \\
  u_n & \cdots & 1 & 1 & 1 \\
  \end{bmatrix} = \left\{ \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix} \right\} \mathbb{E}$$
Illustrative example

\[
\begin{bmatrix}
0.5 & 0.4 & 1.0 \\
0.4 & 0.5 & 0.4 \\
1.0 & 0.4 & 0.5
\end{bmatrix}
= \begin{bmatrix}
0.3 & 0.4 & 0.3 \\
0.4 & 0.3 & 0.3 \\
0.3 & 0.3 & 0.3
\end{bmatrix}
\]

Not positive-semidefinite!

\[
\begin{bmatrix}
0.5 \\
0.4 \\
1.0
\end{bmatrix}
\begin{bmatrix}
0.5 \\
0.4 \\
1.0
\end{bmatrix}
= \begin{bmatrix}
0.3 \\
0.4 \\
0.3
\end{bmatrix}
\]

Second-order moment matrix

\[
\begin{bmatrix}
3.0 & 2.0 & 1.0 \\
2.0 & 3.0 & 2.0 \\
1.0 & 2.0 & 1.0
\end{bmatrix}
\]

Illustrative example

Tree-consistent (pseudo)marginals
The differential entropy $h(x)$ can be expressed as:

\[
x p(x) d \log p(x) \int - =: (x) n
\]

**Lemma:** The differential entropy of any is upper-bounded.

By the covariance-matched Gaussian, it holds:

\[
\frac{1}{2} \log \text{det } \mathbf{C} + \frac{1}{2} \log \text{det } \mathbf{I} \geq (x)n
\]

**Challenge:** Recall that entropy function $\Phi$ in terms of only $n$ lacks an explicit form.

For the Ising model, we have second-order information:

\[
T \in (T', s) \land [^T x^s x] \mathbb{W} =: t s^n \quad \land \quad s \in \mathcal{L} \land [^s x] \mathbb{W} =: s^n
\]

Concave upper bound on entropy

\[
\mathcal{H} \in (T', s) \land [^T x^s x] \mathbb{W} =: t s^n \quad \land \quad s \in \mathcal{L} \land [^s x] \mathbb{W} =: s^n
\]
Log-determinant relaxation

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

\[ \text{MARG}(K_n) \leq \text{OUT}(K_n) \mu \leq S_{\text{DEF}}^{(K_n)} \]

Let $M_1^{(\theta)}$ be a covariance matrix. Note that constraints imply that $\text{MARG}(K_n) \leq \text{OUT}(K_n)$.

Log-determinant relaxation: For any such $\text{OUT}(K_n)$, $\mu$ is upper bounded by:

\[ \max_{\eta \in \text{OUT}(K_n)} \left\{ \frac{u}{\eta} \log \left( \frac{2}{\theta} \right) + \left\{ \left[ [u I, 0] + (\eta I) I \right] \right\} \right\} \]

Note: Such a log-det problem with LMIs constraints can be solved efficiently by an interior-point method. (Vandenberghe, Boyd, & Wu, 1998)

Note:

\[ \Phi \left( \text{OUT}(K_n) \mu \right) \]

Consider an outer bound $\text{OUT}(K_n)^{\mu}$ that satisfies:

Log-determinant relaxation
### Results for Fully Connected Graph

<table>
<thead>
<tr>
<th>Problem Type</th>
<th>Method</th>
<th>Range</th>
<th>Mean ± std</th>
<th>Range</th>
<th>Mean ± std</th>
<th>Problem Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weak</td>
<td>Log-determinant</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>Strong</td>
</tr>
<tr>
<td>Weak</td>
<td>Sum-product</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>Strong</td>
</tr>
<tr>
<td>Weak</td>
<td>Sum-product</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>Weak</td>
</tr>
<tr>
<td>Weak</td>
<td>Sum-product</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>Weak</td>
</tr>
<tr>
<td>Weak</td>
<td>Sum-product</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>Weak</td>
</tr>
<tr>
<td>Weak</td>
<td>Sum-product</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>Weak</td>
</tr>
<tr>
<td>Weak</td>
<td>Sum-product</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>Weak</td>
</tr>
<tr>
<td>Weak</td>
<td>Sum-product</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>Weak</td>
</tr>
<tr>
<td>Weak</td>
<td>Sum-product</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>[0.01, 0.06]</td>
<td>Weak</td>
</tr>
</tbody>
</table>

---

**Log-determinant Method**
<table>
<thead>
<tr>
<th>Problem Type</th>
<th>Method</th>
<th>Mean ± Std</th>
<th>Range</th>
<th>Mean ± Std</th>
<th>Range</th>
<th>Strong</th>
<th>Weak</th>
<th>±</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weak</td>
<td>0.01, 0.12</td>
<td>0.06 ± 0.031</td>
<td>[0.042, 0.042]</td>
<td>0.06 ± 0.030</td>
<td>[0.01, 0.11]</td>
<td>0.44 ± 0.200</td>
<td>0.40 ± 0.11</td>
<td>+</td>
</tr>
<tr>
<td>Weak</td>
<td>0.01, 0.13</td>
<td>0.09 ± 0.030</td>
<td>[0.047, 0.047]</td>
<td>0.06 ± 0.034</td>
<td>[0.01, 0.11]</td>
<td>0.09 ± 0.200</td>
<td>0.06 ± 0.24</td>
<td>+</td>
</tr>
<tr>
<td>Weak</td>
<td>0.01, 0.11</td>
<td>0.38 ± 0.024</td>
<td>[0.038, 0.038]</td>
<td>0.01 ± 0.024</td>
<td>[0.00, 0.020]</td>
<td>0.34 ± 0.200</td>
<td>0.30 ± 0.24</td>
<td>+/-</td>
</tr>
<tr>
<td>Weak</td>
<td>0.01, 0.12</td>
<td>0.01 ± 0.028</td>
<td>[0.047, 0.047]</td>
<td>0.04 ± 0.028</td>
<td>[0.01, 0.12]</td>
<td>0.02 ± 0.028</td>
<td>0.02 ± 0.030</td>
<td>+/-</td>
</tr>
<tr>
<td>Weak</td>
<td>0.01, 0.12</td>
<td>0.01 ± 0.028</td>
<td>[0.01, 0.12]</td>
<td>0.01 ± 0.028</td>
<td>[0.01, 0.12]</td>
<td>0.01 ± 0.028</td>
<td>0.01 ± 0.030</td>
<td>+/-</td>
</tr>
</tbody>
</table>

Results for nearest-neighbor grid
Sum-product versus log-determinant

(a) (b) (c) (d)
Summary

- Faster distributed techniques for solving relaxations
- Integer programming results (e.g., Goemans & Williamson, 1995)
- Performance guarantees for specific problem classes (link to integer programming results for specific problem classes: link to Goemans & Williamson, 1995)

Open questions:

- Relative roles of approximations to MAP and CRG
- Log-determinant relaxation for approximate inference
- Faster distributed techniques for solving relaxations
- Role of mean parameters and marginal polytopes in variational principle
Papers at: http://www.eecs.berkeley.edu/~wainwrig

Martin Wainwright

Contact Information
Supplementary material
Higher order extensions

1. Moment matrices involving higher-order multinomials.

Example:

\[
\begin{pmatrix}
1 & x^1 & x^2 & \ldots & x^m \\
1 & 1^2 & 1^2 & \ldots & 1^2 \\
1^2 & 1^2 & 1^2 & \ldots & 1^2 \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
1^m & 1^m & 1^m & \ldots & 1^m
\end{pmatrix}
= \begin{pmatrix}
(s) \\
0 \\
\vdots \\
0
\end{pmatrix}
\]

Consider correlations among vectors of multinomials:

\[
\{ \frac{s}{t} x^1, \ldots, \frac{s}{t} x^m \} = (s) d
\]

2. For more general discrete spaces \( \mathcal{X} \), consider correlations among vectors of monomials:

\[
\{ \text{monomials} \} = \mathcal{X}
\]

Higher order extensions

\[\text{Example:}\]

I. Moment matrices involving higher-order multinomials.
For fixed $\theta$, consider the 1-parameter family of distributions: $Z_{\text{zero temperature limit}}$
For strong coupling, behavior of log-det relaxation (for inference) approaches that of a SDP relaxation for integer programming.

\[
\langle \eta, \theta \rangle \Big( \max_{\nu \in \nu} \{ \eta \} \Big) \geq \langle (x) \phi, \theta \rangle \Big( \max_{\nu \in \nu} \{ x \} \Big)
\]

Result is a well-known SDP relaxation for integer programming:

\[
\max \{ \sum_{i} \log \det [I + \eta I] \} \geq \max \{ \sum_{i} \log \det [I + \eta I] \} \]

Taking limits as $\nu \to \infty$ corresponds to computing a recession function.

For all bounded by the following:

\[
\Phi(\eta) \leq \frac{C + \left\{ \left[ [u I, 0] \right], \right\} \frac{3}{2} + (\eta) \left[ \log \det [I + \eta] \right] \frac{7}{2} + \langle \eta, \theta \rangle \rangle \max_{\nu \in \nu} \{ \eta \} \frac{\theta}{\nu} \]

\[
\forall \theta, \eta > 0, \quad \Phi(\eta) \leq C + \left\{ \left[ [u I, 0] \right], \right\} \frac{3}{2} + (\eta) \left[ \log \det [I + \eta] \right] \frac{7}{2} + \langle \eta, \theta \rangle \rangle \max_{\nu \in \nu} \{ \eta \} \frac{\theta}{\nu}
\]

Link to SDP relaxation for integer programming.