

# Semidefinite relaxations for approximate inference and counting problems

Speaker: Martin Wainwright, EE & CS, UC Berkeley  
`martinw@eecs.berkeley.edu`

Joint work with: Michael Jordan, CS & Statistics, UC Berkeley  
`jordan@cs.berkeley.edu`

# Outline

## 1. Background

- (a) Exponential families and Markov random fields
- (b) Inference and counting problems

## 2. Role of convex optimization

- (a) Conjugate duality and Legendre mapping
- (b) Moment polytopes
- (c) Entropy functions

## 3. Semidefinite relaxations

- (a) Semidefinite bounds on moment polytopes
- (b) Log-determinant bound on entropy
- (c) Zero-temperature limit and integer programming

## 4. Summary and open questions

# Exponential families and graphical models

Given: A collection  $\mathbf{x} = \{x_s \mid s = 1, \dots, n\}$  of random variables.

Focus: A distribution  $p(\mathbf{x})$  defined by (relatively) local interactions.

Examples:

**statistical physics:** models of gases, magnets, crystals (e.g, Ising model; Potts model)

**statistics:** log-linear models; maximum entropy; Markov random fields (e.g., Hammersley & Clifford, 1973; Lauritzen, 1996)

**image processing and computer vision:** Markov image models; Gibbs sampler (e.g., Woods, 1978; Geman & Geman, 1984)

**error-correcting coding:** various graphical codes including LDPCs, turbo codes (e.g., Gallager, 1963; Luby et al. 1998, McEliece et al., 1998)

**machine learning:** computational biology; robotics; natural language modeling (e.g, Pearl, 1988; Jordan et al., 1999)

# Exponential families

## Notation:

$$\begin{aligned}\mathbf{x} = \{x_s \mid s = 1, \dots, n\} &\equiv \text{a collection of random variables} \\ \phi_\alpha : \mathcal{X}^n \rightarrow \mathbb{R} &\equiv \text{potential function} \\ \theta = \{ \theta_\alpha \mid \alpha \in \mathcal{I} \} &\equiv \text{collection of real-valued weights} \\ \langle \theta, \phi(\mathbf{x}) \rangle := \sum_{\alpha \in \mathcal{I}} \theta_\alpha \phi_\alpha(\mathbf{x}) &\equiv \text{Euclidean inner product in } \mathbb{R}^d\end{aligned}$$

The associated *exponential family* is the collection of densities:

$$p(\mathbf{x}; \theta) = \exp \{ \langle \theta, \phi(\mathbf{x}) \rangle - \Phi(\theta) \}$$

Log partition function:

$$\Phi(\theta) = \log \left( \sum_{\mathbf{x} \in \mathcal{X}^n} \exp \left\{ \sum_{\alpha} \theta_\alpha \phi_\alpha(\mathbf{x}) \right\} \right)$$

## Simple examples

1. Scalar Gaussian:  $\phi_a(x) = x, \quad \phi_b(x) = x^2$

$$p(x; \theta) = \exp\{\theta_a x + \theta_b x^2 - \Phi(\theta)\}$$

Here  $\theta$  is restricted to  $\{(\theta_a, \theta_b) \in \mathbb{R}^2 \mid \theta_b < 0\}$ .

2. Binary random variables:  $\phi(x_1, x_2) = \{x_1, x_2, x_1 x_2\}$

$$p(\mathbf{x}; \theta) = \exp\{\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2 - \Phi(\theta)\}$$

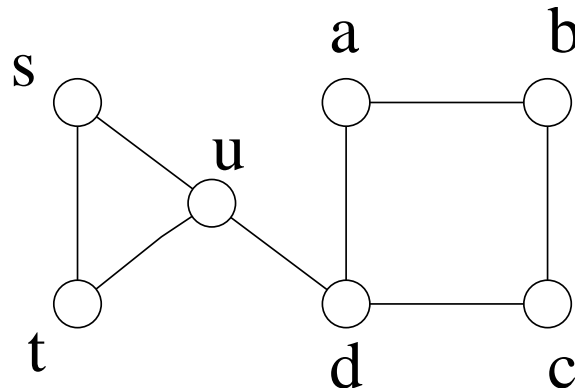
3. General  $m$ -state case:

Singleton terms:  $\{x_s, x_s^2, \dots, x_s^{m-1}\}$

Coupling terms:  $\{x_s^a x_t^b \mid a, b = 1, \dots, m-1\}$

# Markov random fields

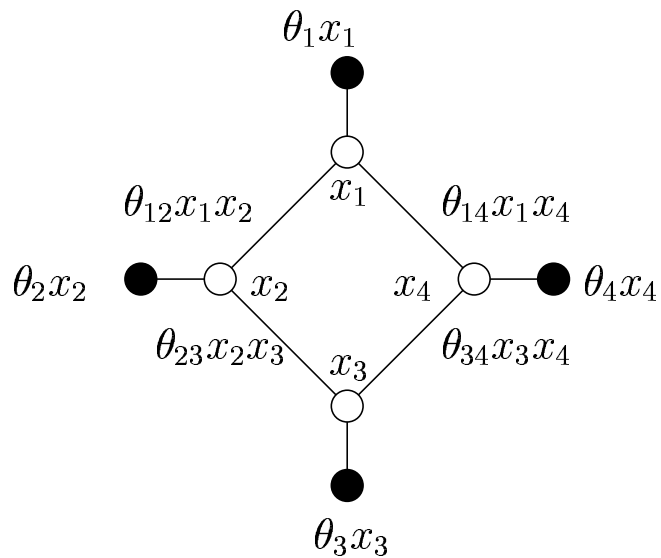
Based on an undirected (simple) graph  $G = (V, E)$ :



- vector  $\mathbf{x} = \{x_s \mid s \in V\}$  of random variables living at the vertices.
- graph clique: fully connected subset  $C_\alpha$  of the vertex set  $V$
- potential functions  $\phi_\alpha$  defined on cliques  
(I.e.,  $\phi_\alpha$  depends only on  $\mathbf{x}_\alpha = \{x_s \mid s \in C_\alpha\}$ ).

## Illustration: Ising model

Binary variables on a graph with maximal cliques of size two:



$$\phi = \{ x_s \mid s \in V \} \cup \{ x_s x_t \mid (s, t) \in E \}$$

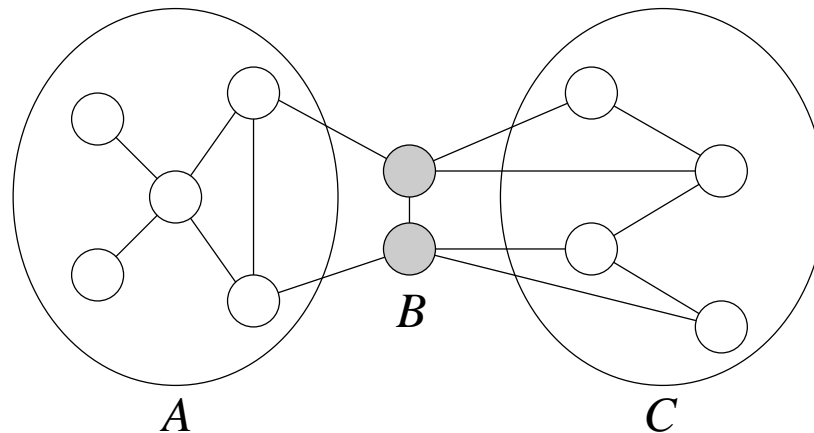
$$\mathcal{I} = V \cup E$$

$$\mathcal{X}^n = \{0, 1\}^n$$

$$p(\mathbf{x}; \theta) = \exp \left\{ \sum_{s \in V} \theta_s x_s + \sum_{(s, t) \in E} \theta_{st} x_s x_t - \Phi(\theta) \right\}$$

## Graph separation and Markov

- random vectors  $\mathbf{x}$  of interest are *Markov* with respect to the graph



**Markov property:**  $\mathbf{x}_{A|B} \perp \mathbf{x}_{C|B}$  if  $B$  separates  $A$  from  $C$ .

**Note:** The notation  $\mathbf{x}_{A|B} \perp \mathbf{x}_{C|B}$  means that  $\mathbf{x}_A$  is conditionally independent of  $\mathbf{x}_C$  given  $\mathbf{x}_B$ .

## Extensions to the basic model

Third order terms:

$$p(\mathbf{x}; \theta) = \exp \left\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t)} \theta_{st} x_s x_t + \sum_{(s,t,u)} \theta_{stu} x_s x_t x_u - \Phi(\theta) \right\}$$

More generally, add multinomials up to degree  $d \leq n$ .

Most general binary model (on the complete graph  $K_n$ ):

$$p(\mathbf{x}; \theta) = \exp \left\{ \sum_{s=1}^n \theta_s x_s + \sum_{(s,t)} \theta_{st} x_s x_t + \sum_{(s,t,u)} \theta_{stu} x_s x_t x_u + \dots \right. \\ \left. \dots + \theta_{1\dots n} \prod_{s=1}^n x_s - \Phi(\theta) \right\}$$

## Inference problems

- given observations  $\mathbf{y} = \{ y_s \mid s \in V \}$  specified by measurement distribution:

$$p(\mathbf{y} \mid \mathbf{x}) = \prod_{s \in V} p(y_s \mid x_s)$$

- by Bayes' rule:

$$p(\mathbf{x} \mid \mathbf{y}; \theta) \propto p(\mathbf{x}; \theta) p(\mathbf{y} \mid \mathbf{x})$$

- this conditional distribution is central to various inference problems:

(a) MAP estimate:  $\hat{\mathbf{x}}_{MAP} = \arg \max_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x} \mid \mathbf{y}; \theta)$

(b) marginal distributions:  $p(x_s \mid \mathbf{y}; \theta) = \sum_{\{ \mathbf{x}' \mid x'_s = x_s \}} p(\mathbf{x}' \mid \mathbf{y}; \theta)$

## Examples of inference problems

1. **Image processing:** vector  $\mathbf{x}$  is a representation of the image (e.g., pixels, wavelets, features); vector  $\mathbf{y}$  is a noise-corrupted version.
2. **Error-correcting coding:** vector  $\mathbf{y}$  represents bits received from noisy channel;  $\mathbf{x}$  represents the codeword.
  - (a)  $\hat{\mathbf{x}}_{MAP} = \arg \max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y})$  minimizes word error rate (WER).
  - (b)  $\hat{x}_s = \begin{cases} 1 & \text{if } p(x_s = 1 | \mathbf{y}) > 0.5 \\ 0 & \text{otherwise} \end{cases}$  minimizes bit error rate (BER).

## Role of the log partition function

Recall the log partition function:

$$\Phi(\theta) = \log \left[ \sum_{\mathbf{x} \in \mathcal{X}^n} \exp\{\langle \theta, \phi(\mathbf{x}) \rangle\} \right].$$

Plays an important role in various contexts:

- (a) bounds on marginal distributions
- (b) maximum likelihood (ML) parameter estimation
- (c) large deviations exponents
- (d) combinatorial enumeration

## Nature of problems

1. MAP estimation is an *integer program*.

**Example:** For the Ising model, MAP estimation is equivalent to a binary quadratic program:

$$\max_{\mathbf{x} \in \{0,1\}^n} \left\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t \right\}.$$

2. Computing marginal distributions and the log partition function are *counting problems*.

Both problems suffer from the exponential explosion in the number of configurations, which grows as  $m^n$  (for random variables in a  $m$ -state alphabet).

## Dependence on graph structure

- complexity of inference/estimation problems depends critically on graph structure
- for graphs without cycles, there exist efficient forms of non-serial dynamic-programming for trees (e.g., Gallager, 1963; Briochi et al., 1973; Pearl, 1988)
- algorithms generalize to graphs of low treewidth via hypertree decompositions (e.g., Lauritzen & Spiegelhalter, 1988)
- for general graphs with cycles, these same problems are intractable — hence, need for approximate methods

## §2. Role of convex optimization

# Variational formulation

**Basic idea:** Express a quantity of interest (say  $\hat{q}$ ) as the solution of an optimization problem:

- (a) study the behavior of  $\hat{q}$  via the optimization problem.
- (b) approximate  $\hat{q}$  by solving approximations to the optimization problem.

**Goal:** Obtain a variational formulation for:

- (a) computing the log partition function  $\Phi(\theta) = \log \sum_{\mathbf{x} \in \mathcal{X}^n} \exp\{\langle \theta, \phi(\mathbf{x}) \rangle\}$ .
- (b) computing mean parameters  $\mathbb{E}_{\theta}[\phi_{\alpha}(\mathbf{x})] := \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}; \theta) \phi(\mathbf{x})$

## Properties of log partition function

**Lemma:** The log partition function  $\Phi(\theta) = \log \sum_{\mathbf{x}} \exp \{ \langle \theta, \phi(\mathbf{x}) \rangle \}$  is a cumulant-generating function, and hence convex in terms of  $\theta$ .

(a) First derivatives:

$$\frac{\partial \Phi}{\partial \theta_{\alpha}}(\theta) = \mathbb{E}_{\theta}[\phi_{\alpha}(\mathbf{x})] := \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}; \theta) \phi_{\alpha}(\mathbf{x})$$

(b) Second derivatives:

$$\begin{aligned} \frac{\partial^2 \Phi}{\partial \theta_{\alpha} \partial \theta_{\beta}}(\theta) &= \text{cov}_{\theta} \{ \phi_{\alpha}(\mathbf{x}), \phi_{\beta}(\mathbf{x}) \} \\ &:= \mathbb{E}_{\theta}[\phi_{\alpha}(\mathbf{x}) \phi_{\beta}(\mathbf{x})] - \mathbb{E}_{\theta}[\phi_{\alpha}(\mathbf{x})] \mathbb{E}_{\theta}[\phi_{\beta}(\mathbf{x})] \end{aligned}$$

## Variational representation via duality

Exploit convexity to express  $\Phi$  in terms of its dual:

$$\Phi(\theta) = \sup_{\mu \in \text{dom } \Phi^*} \{ \langle \theta, \mu \rangle - \Phi^*(\mu) \}$$

We need to determine:

(a) Form of dual function:

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

(b) Effective domain:

$$\text{dom } \Phi^* := \{ \mu \in \mathbb{R}^d \mid \Phi^*(\mu) < +\infty \}.$$

## Classical example

- exponential family based on indicator functions for individual configurations  $\{\mathbf{x} = \mathbf{e}\}$ :

$$p(\mathbf{x}; \theta) = \exp \left\{ \sum_{\mathbf{e} \in \mathcal{X}^n} \theta_{\mathbf{e}} \delta(\mathbf{x} = \mathbf{e}) - \Phi(\theta) \right\}.$$

- log partition function has the form

$$\Phi(\theta) = \log \left[ \exp(\theta_{e_0}) + \cdots + \exp(\theta_{e_{|\mathcal{X}^n|}}) \right]$$

- supremum defining  $\Phi^*$  is  $+\infty$  unless  $\mu$  belongs to probability simplex  $\{ \mu \in \mathbb{R}^{|\mathcal{X}^n|} \mid \mu \geq 0, \sum_{\mathbf{e}} \mu_{\mathbf{e}} = 1 \}$
- on the simplex, dual function is equal to the negative of (Boltzmann-Shannon) entropy:

$$\Phi^*(\mu) = \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}) \log p(\mathbf{x}) \quad \text{where } \mu_{\mathbf{e}} = p(\mathbf{x} = \mathbf{e})$$

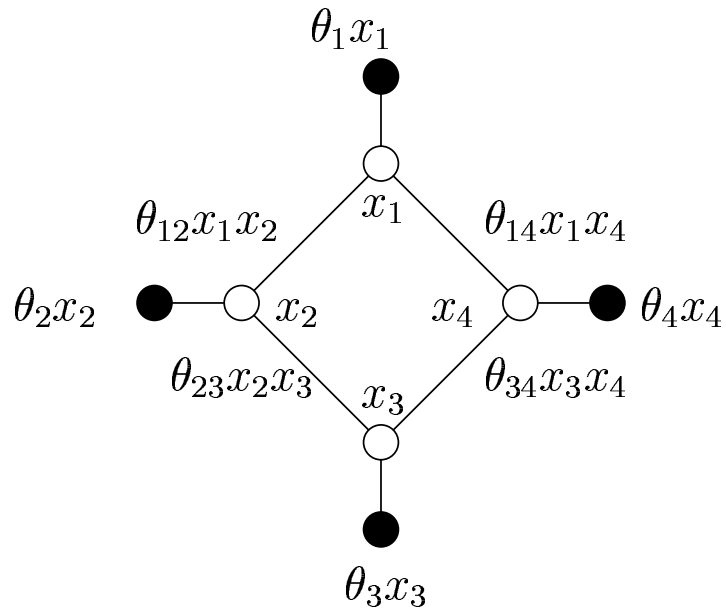
## Moment polytopes

- consider the convex hull of  $\{\phi(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}^n\}$ :

$$\text{MARG}(\phi) := \left\{ \mu \in \mathbb{R}^d \mid \mu = \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x})\phi(\mathbf{x}) \text{ for some } p(\cdot) \right\}$$

- explicitly,  $\text{MARG}(\phi)$  is the set of *mean parameters* of  $\phi(\mathbf{x})$  that are realizable
- as a convex hull of a finite number of vectors,  $\text{MARG}(\phi)$  can be described in terms of a finite number of inequalities
- however, a very large number of inequalities are required in general

# Ising model example



Potentials  $\phi = \{x_s \mid s \in V\} \cup \{x_s x_t \mid (s, t) \in E\}$

Relevant moments  $\mu_s = \mathbb{E}_\theta[x_s]$   $\mu_{st} = \mathbb{E}_\theta[x_s x_t]$

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

## Negative entropy as conjugate dual

**Proposition 1:** (a) Consider the conjugate dual:

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

For each  $\mu \in \text{ri MARG}(\phi)$ , the supremum is attained at a point  $\theta(\mu)$  such that the *moment matching* conditions hold:

$$\mu = \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}; \theta(\mu)) \phi(\mathbf{x}) = \mathbb{E}_{\theta(\mu)}[\phi(\mathbf{x})].$$

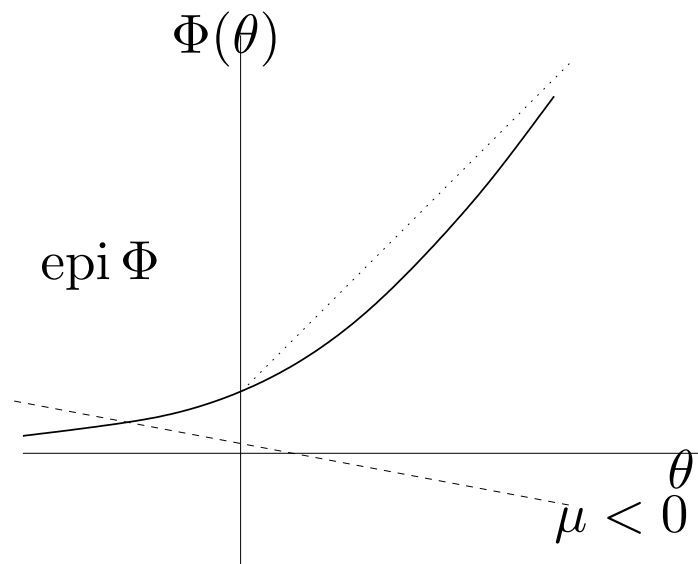
(b) Moreover, the dual function is given as follows:

$$\Phi^*(\mu) = \begin{cases} -H(p(\mathbf{x}; \theta(\mu))) & \text{if } \mu \in \text{MARG}(\phi) \\ +\infty & \text{otherwise.} \end{cases}$$

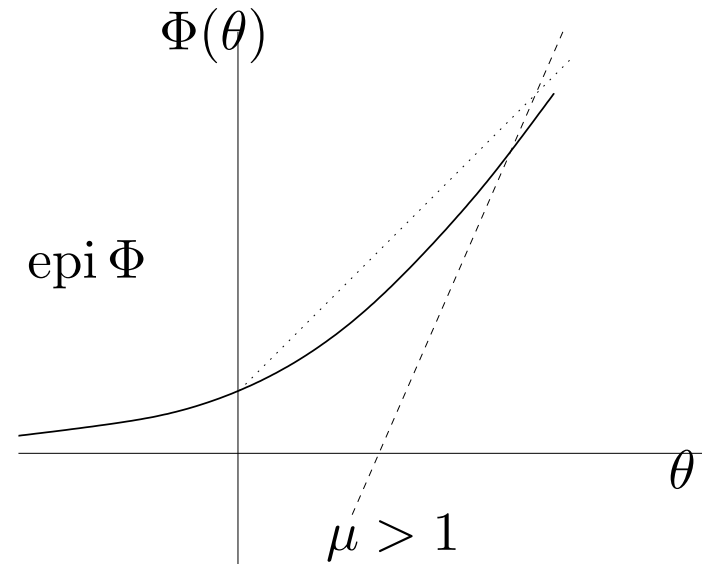
## Bernoulli example

Consider a scalar binary random variable  $x \in \{0, 1\}$ :

$$p(x; \theta) = \exp\{\theta x - \Phi(\theta)\}, \quad \Phi(\theta) = \log[1 + \exp(\theta)], \quad \mu := \mathbb{E}_\theta[x]$$



(a)



(b)

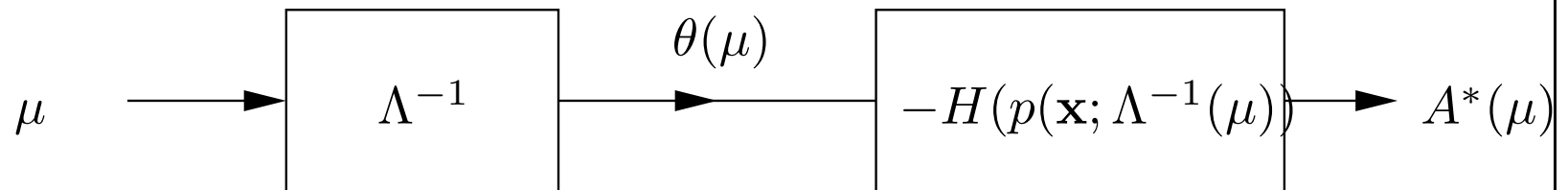
Dual function

$$\Phi^*(\mu) = \begin{cases} \mu \log \mu + (1 - \mu) \log(1 - \mu) & \text{if } \mu \in [0, 1] \\ +\infty & \text{otherwise} \end{cases}$$

## Nature of the dual function

The dual function  $\Phi^*$  is the composition of two functions:

- (a) the inverse Legendre mapping  $\Lambda^{-1} : \mu \mapsto \theta(\mu)$
- (b) the ordinary negative entropy  $-H(p(\mathbf{x}; \theta(\mu)))$  of the resulting distribution.



**Consequence:** It typically lacks a closed form expression in terms of  $\mu$ .

## Variational representation

### Proposition 2:

(a) The log partition function has the variational representation:

$$\underbrace{\Phi(\theta)}_{\text{log partition function}} = \underbrace{\sup_{\mu \in \text{MARG}(\phi)} \{ \langle \mu, \theta \rangle - \Phi^*(\mu) \}}_{\text{convex program over the marginal polytope}}$$

log partition function

convex program over  
the marginal polytope

(b) The supremum is attained uniquely at the mean parameters of  $p(\mathbf{x}; \theta)$ :

$$\mu = \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}; \theta) \phi(\mathbf{x}) = \mathbb{E}_{\theta}[\phi(\mathbf{x})]$$

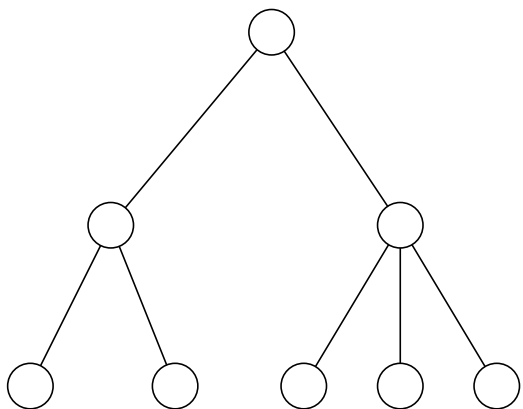
## Challenges

1. Moment polytopes  $\text{MARG}(\phi)$  are extremely difficult to characterize.
2. Entropy  $\Phi^*(\mu)$  as a function of *only* the mean parameters  $\mu$  is implicitly defined. It typically lacks an explicit form.

### Remarks:

- (a) Complexity of problem depends critically on graph structure associated with the potentials  $\phi$ .
- (b) Problem of characterizing moment polytopes also arises in integer programming (link via zero-temperature limit).

## Why are tree-structured problems easy?



Ising model on a tree  $T = (V, E(T))$ :

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

Relevant mean parameters or moments are:

$$\mu_s = \mathbb{E}_\theta[x_s], \quad \mu_{st} = \mathbb{E}_\theta[x_s x_t]$$

From a variational perspective, trees are special for two reasons:

- (a) Moment polytope  $\text{MARG}(T)$  is easy to characterize.
- (b) Negative entropy  $\Phi^*$  has an explicit form.

## Characterization of MARG( $T$ )

The mean parameters associated with edge  $(s, t)$

$$\mu_s = \mathbb{E}_\theta[x_s], \quad \mu_t = \mathbb{E}_\theta[x_t], \quad \mu_{st} = \mathbb{E}_\theta[x_s x_t].$$

determine a local marginal distribution as follows:

$$p(x_s, x_t; \mu) := \begin{bmatrix} (1 + \mu_{st} - \mu_s - \mu_t) & (\mu_t - \mu_{st}) \\ (\mu_s - \mu_{st}) & \mu_{st} \end{bmatrix} \geq 0$$

For a tree, this local consistency implies global consistency:

$$\text{LOCAL}(T) = \text{MARG}(T)$$

Special case of junction tree theorem (e.g., Cowell et al., 1997)

## Tree-structured factorization

Junction tree theorem also implies that  $p$  factorizes in terms of its marginals:

$$p(\mathbf{x}; \theta(\mu)) = \prod_{s \in V} p(x_s; \mu) \prod_{(s,t) \in E(T)} \frac{p(x_s, x_t; \mu)}{p(x_s; \mu)p(x_t; \mu)}$$

Hence for a tree  $T = (V, E(T))$ :

$$\Phi^*(\mu) = - \sum_{s \in V} H_s(\mu) + \sum_{(s,t) \in E(T)} I_{st}(\mu)$$

Single node entropy  $H_s(\mu) = - \sum_{x_s} p(x_s; \mu) \log p(x_s; \mu)$

Mutual information  $I_{st}(\mu) = \sum_{x_s, x_t} p(x_s, x_t; \mu) \log \frac{p(x_s, x_t; \mu)}{p(x_s; \mu)p(x_t; \mu)}$

## Variational problem for trees

For trees,  $\text{MARG}(T) \equiv \text{LOCAL}(T)$  is fully determined by  $\mathcal{O}(n)$  constraints:

$$\begin{aligned} 1 + \mu_{st} - \mu_s - \mu_t &\geq 0 \\ \mu_{st} &\geq 0 \\ \mu_s - \mu_{st} &\geq 0. \end{aligned}$$

Dual  $\Phi^*$  has an explicit form, and variational principle takes the form:

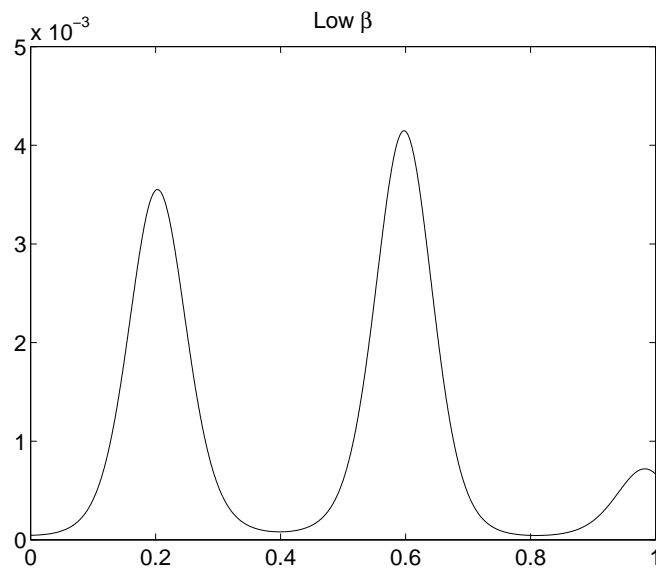
$$\Phi(\theta) = \max_{\mu \in \text{LOCAL}(T)} \left\{ \langle \theta, \mu \rangle + \sum_{s \in V} H_s(\mu) - \sum_{(s,t) \in E(T)} I_{st}(\mu) \right\}.$$

## Zero temperature limit

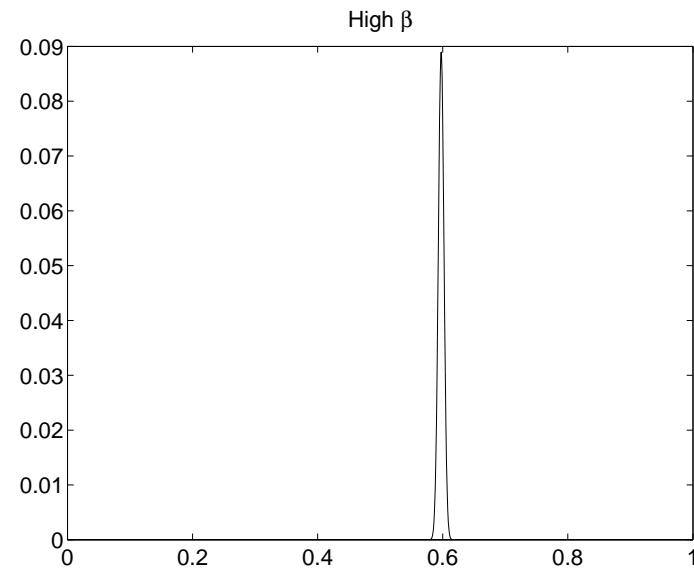
For fixed  $\theta$ , consider the 1-parameter family of distributions:

$$p(\mathbf{x}; \beta\theta) = \exp \{ \beta \langle \theta, \phi(\mathbf{x}) \rangle - \Phi(\beta\theta) \}$$

Here  $\beta$  should be viewed as inverse “temperature”.



(a) Low  $\beta$



(b) High  $\beta$

## Limiting form of exact variational principle

From Proposition 2, for each  $\beta > 0$ , we have:

$$\frac{1}{\beta} \Phi(\beta\theta) = \frac{1}{\beta} \sup_{\mu \in \text{MARG}(\phi)} \{ \langle \beta\theta, \mu \rangle - \Phi^*(\mu) \}.$$

Take the limit as  $\beta \rightarrow +\infty$ :

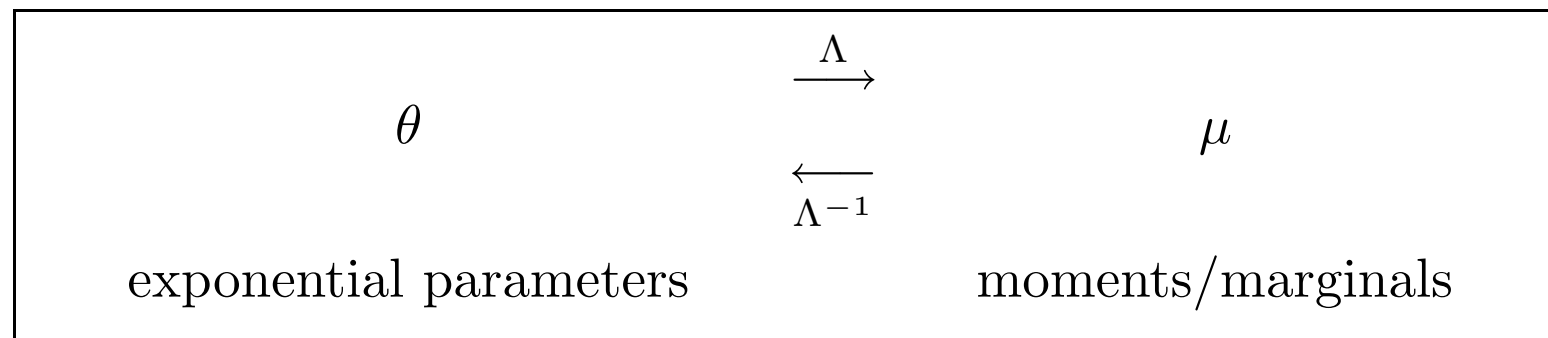
$$\underbrace{\max_{\mathbf{x} \in \mathcal{X}^n} \langle \theta, \phi(\mathbf{x}) \rangle}_{\text{integer program}} = \underbrace{\sup_{\mu \in \text{MARG}(\phi)} \{ \langle \theta, \mu \rangle \}}_{\text{linear program over the moment polytope}}$$

# Legendre mapping

The log partition function  $\Phi$  and negative entropy  $\Phi^*$  are conjugate duals.

This duality establishes a one-to-one Legendre mapping.

(Rockafellar, 1970)



1. Inference involves computing the *forward* Legendre mapping  $\mu = \Lambda(\theta)$ .
2. Parameter estimation involves computing the *backward* Legendre mapping  $\theta = \Lambda^{-1}(\mu)$ .

### **§3. Convex and semidefinite relaxations**

## Motivation for relaxation

- recall the exact variational principle:

$$\Phi(\theta) = \sup_{\mu \in \text{MARG}(\phi)} \{ \langle \mu, \theta \rangle - \Phi^*(\mu) \}$$

- since exact principle is intractable in general, seek a convex relaxation
- requirements for a (convex) relaxation:
  - (a) convex approximation to  $\text{MARG}(\phi)$
  - (b) convex approximation to  $\Phi^*$

# Semidefinite outer bounds on moment polytopes

Recall the *moment or marginal polytope*:

$$\text{MARG}(\phi) = \left\{ \mu \mid \mu_\alpha = \sum_{\mathbf{x}} p(\mathbf{x}) \phi_\alpha(\mathbf{x}) \text{ for some } p(\cdot) \right\}$$

Focus on:

- (a) binary case with “spins”  $\mathbf{x} \in \{-1, +1\}^n$ .
- (b) complete graph  $K_n$  on  $n$  nodes.

Refer to the associated moment polytope as  $\text{MARG}(K_n)$ .

Relevant moments:

$$\mu_s = \mathbb{E}_\theta[x_s] \quad \text{for all } s = 1, \dots, n$$
$$\mu_{st} = \mathbb{E}_\theta[x_s x_t] \quad \text{for all } (s, t)$$

Sequence of semidefinite relaxations on the binary moment polytope  $\text{MARG}(K_n)$  (e.g., Lasserre, 2001)

## Correlation matrices

Correlation matrix of the random variables  $\mathbf{x} = \{x_1, \dots, x_n\}$ :

$$\text{cor}(\mathbf{x}) = \mathbb{E}[\mathbf{x}\mathbf{x}^T] = \begin{bmatrix} 1 & \mu_{12} & \mu_{13} & \cdots & \mu_{1n} \\ \mu_{21} & 1 & \mu_{23} & \cdots & \mu_{2n} \\ \mu_{31} & \mu_{32} & 1 & \vdots & \mu_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_{n1} & \mu_{n2} & \cdots & \cdots & 1 \end{bmatrix} \succeq 0$$

### Remarks:

- (a) For  $\mathbf{x} \in \{-1, +1\}^n$ , we have  $x_s^2 = 1$ .
- (b) Correlation matrix does not involve  $\mu_s$ .

## Covariance matrix

Slight strengthening via covariance matrix:

$$\text{cov}(\mathbf{x}) = \mathbb{E}[\mathbf{x}\mathbf{x}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}^T] \succeq 0$$

By Schur complement, equivalent to enforce PSD constraint on:

$$\text{cor}(1, \mathbf{x}) = \begin{bmatrix} 1 & \mu_1 & \mu_2 & \mu_3 & \cdots & \mu_n \\ \mu_1 & 1 & \mu_{12} & \mu_{13} & \cdots & \mu_{1n} \\ \mu_2 & \mu_{21} & 1 & \mu_{23} & \cdots & \mu_{2n} \\ \mu_3 & \mu_{31} & \mu_{32} & 1 & \vdots & \mu_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_n & \mu_{n1} & \mu_{n2} & \cdots & \cdots & 1 \end{bmatrix} = \begin{bmatrix} 1 & \mu_s^T \\ \mu_s & \text{cor}(\mathbf{x}) \end{bmatrix}$$

## Higher order extensions

1. Correlations involving higher-order multinomials.

Example:

$$\text{cor} (1, x_1, x_2, x_1 x_2) = \begin{bmatrix} 1 & \mu_1 & \mu_2 & \mu_{12} \\ \mu_1 & 1 & \mu_{12} & \mu_2 \\ \mu_2 & \mu_{12} & 1 & \mu_1 \\ \mu_{12} & \mu_2 & \mu_1 & 1 \end{bmatrix} \propto 0$$

2. For more general discrete spaces  $\mathcal{X} = \{0, 1, \dots, m-1\}$ , consider correlations among vectors of monomials:

$$\mathcal{P}(s) = \{x_s, x_s^2, \dots, x_s^{m-1}\}$$

## Overview of outer bounds

Sequence of semidefinite relaxations of different orders:

$$\text{SDEF}_1 \supseteq \text{SDEF}_2 \supseteq \cdots \supseteq \text{SDEF}_n \equiv \text{MARG}(K_n).$$

(Lasserre, 2001)

Can combine with various linear constraints as well; e.g., hypertree marginal constraints:

$$\text{LOCAL}_1(n) \supseteq \text{LOCAL}_2(K_n) \supseteq \cdots \supseteq \text{LOCAL}_{n-1}(K_n) \equiv \text{MARG}(K_n)$$

Other linear constraints are possible. (e.g., Deza & Laurent, 1997)

Let  $\text{OUT}(K_n)$  be some convex outer bound on the marginal polytope  $\text{MARG}(K_n)$ .

## Gaussian as maximum entropy

Differential entropy for a continuous-valued random vector  $\tilde{\mathbf{x}}$ :

$$h(\tilde{\mathbf{x}}) = - \int p(\tilde{\mathbf{x}}) \log p(\tilde{\mathbf{x}}) d\tilde{\mathbf{x}}$$

**Lemma:** The differential entropy of any  $\tilde{\mathbf{x}}$  is upper-bounded as follows:

$$h(\tilde{\mathbf{x}}) \leq \frac{1}{2} \log \det \text{cov}(\tilde{\mathbf{x}}) + \frac{n}{2} \log(2\pi e)$$

(Cover & Thomas, 1990)

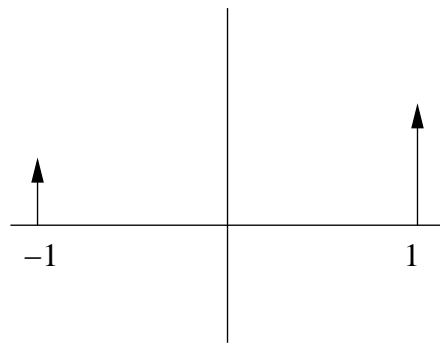
**Note:** The RHS is the entropy of a Gaussian with covariance  $\text{cov}(\tilde{\mathbf{x}})$ .

## From discrete to differential entropy

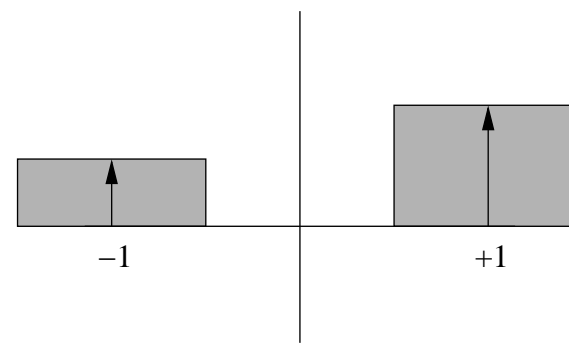
Need to relate the discrete entropy  $H(\mathbf{x})$  to the differential entropy

**Solution:** “Smoothing” by addition of independent randomness.

Let  $\mathbf{u} \sim \mathcal{U}[-\frac{1}{2}, \frac{1}{2}]$  be uniformly distributed.



Discrete  $\mathbf{x}$



Cts.  $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{u}$

**Lemma:** The differential entropy of the smoothed version  $\tilde{\mathbf{x}}$  is matched to the discrete entropy of  $\mathbf{x}$ . (I.e.,  $H(\mathbf{x}) = h(\tilde{\mathbf{x}})$ .)

## Log-determinant relaxation

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

$$\text{MARG}(K_n) \subseteq \text{OUT}(K_n) \subseteq \text{SDEF}_1$$

Let  $M_1(\mu) = \begin{bmatrix} 1 & \mu_s^T \\ \mu_s & \mu_{st} \end{bmatrix} =$  covariance matrix

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

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

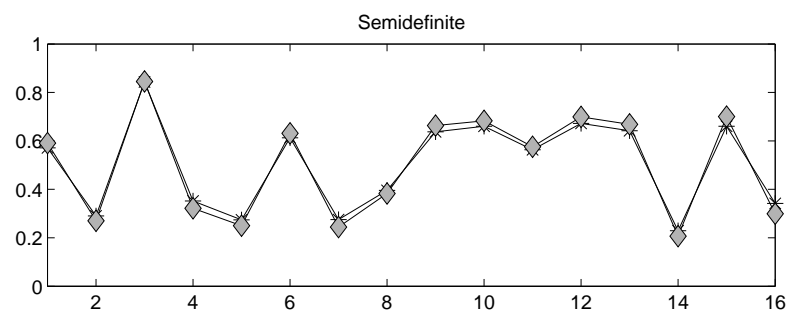
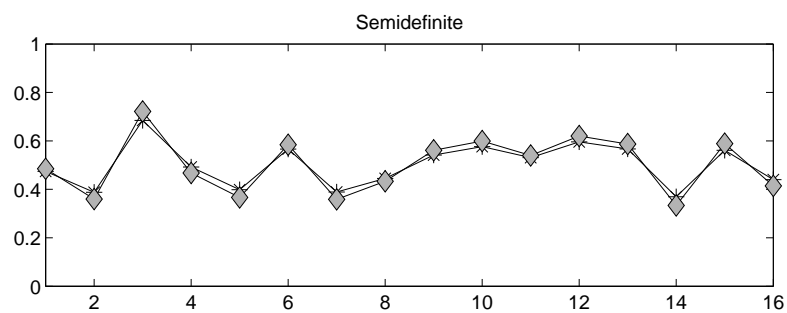
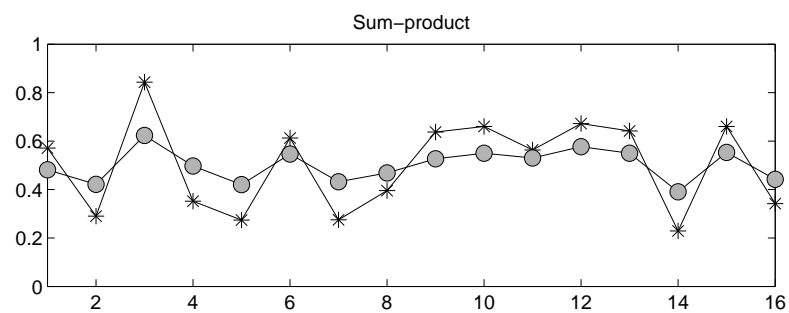
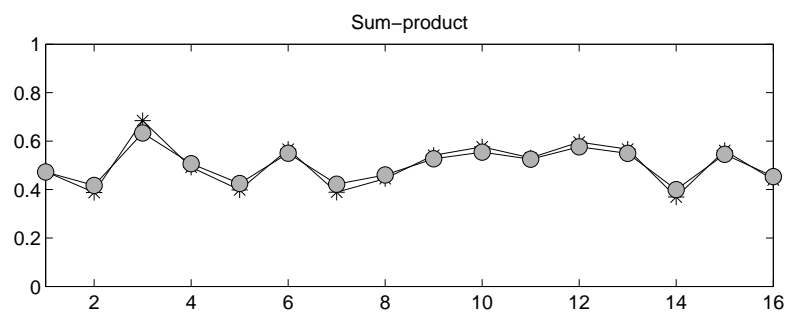
**Note:** Such a log-det problem with LMI constraints can be solved efficiently by an interior-point method. (Vandenberghe, Boyd, & Wu, 1998)

## Practical uses of log-determinant relaxation

1. Provides an efficiently computable upper bound on the log partition function.
2. Also useful in the context of inference:
  - recall that optimizing arguments of the exact variational principle are the mean parameters  $\mu = \mathbb{E}_\theta[\phi(\mathbf{x})]$
  - suggests solving log-det relaxation, and taking optimizing arguments as approximations to the mean parameters

# Simple illustration

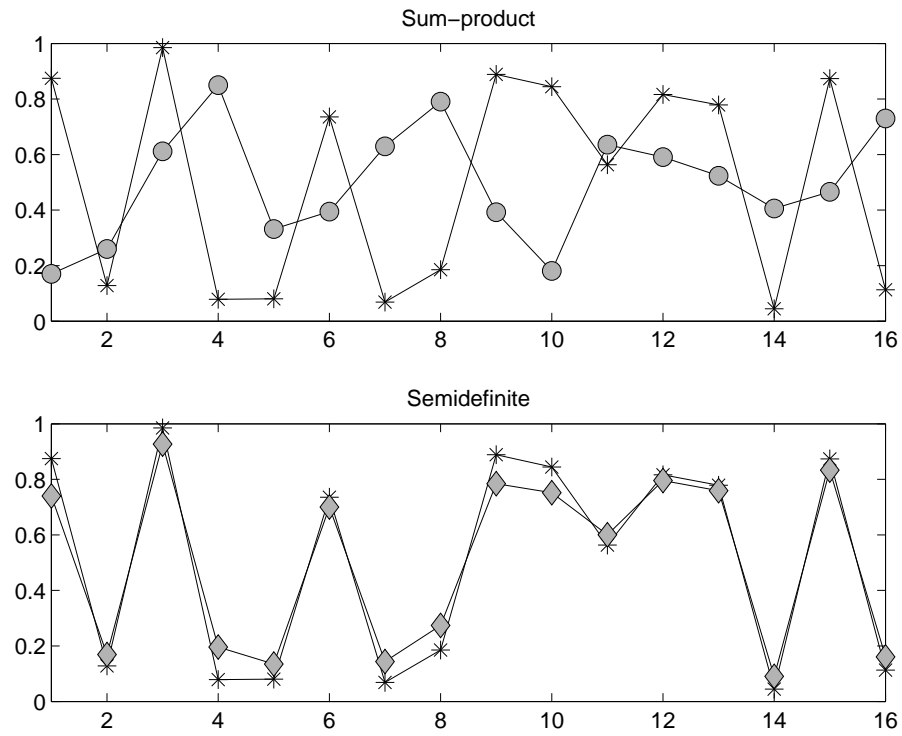
Binary vector  $\mathbf{x}$  on complete graph  $K_{16}$ .



(a) Weak

(b) Medium

# Strong couplings



(c) Strong

## Results for fully connected graph

Problem type		Method			
		Sum-product		Log-determinant	
Coup.	Str.	Mean $\pm$ std	Range	Mean $\pm$ std	Range
–	Weak	$0.037 \pm 0.015$	[0.01, 0.10]	$0.020 \pm 0.005$	[0.01, 0.03]
–	Strong	$0.071 \pm 0.032$	[0.03, 0.20]	$0.018 \pm 0.005$	[0.01, 0.04]
+/-	Weak	$0.004 \pm 0.005$	[0.00, 0.04]	$0.020 \pm 0.005$	[0.01, 0.03]
+/-	Strong	$0.055 \pm 0.060$	[0.01, 0.31]	$0.021 \pm 0.010$	[0.01, 0.06]
+	Weak	$0.024 \pm 0.016$	[0.00, 0.08]	$0.027 \pm 0.015$	[0.01, 0.06]
+	Strong	$0.435 \pm 0.196$	[0.08, 0.86]	$0.033 \pm 0.019$	[0.01, 0.09]

## Results for nearest-neighbor grid

Problem type		Method			
		Sum-product		Log-determinant	
Coup.	Str.	Mean $\pm$ std	Range	Mean $\pm$ std	Range
–	Weak	$0.294 \pm 0.124$	[0.04, 0.59]	$0.047 \pm 0.028$	[0.01, 0.12]
–	Strong	$0.342 \pm 0.167$	[0.04, 0.78]	$0.041 \pm 0.030$	[0.00, 0.12]
+/-	Weak	$0.014 \pm 0.024$	[0.00, 0.20]	$0.016 \pm 0.004$	[0.01, 0.02]
+/-	Strong	$0.095 \pm 0.111$	[0.01, 0.54]	$0.038 \pm 0.024$	[0.01, 0.11]
+	Weak	$0.440 \pm 0.200$	[0.06, 0.90]	$0.047 \pm 0.030$	[0.00, 0.13]
+	Strong	$0.520 \pm 0.226$	[0.06, 0.94]	$0.042 \pm 0.031$	[0.00, 0.12]

## Link to SDP relaxation for integer programming

For all  $\beta > 0$ ,  $\frac{1}{\beta}\Phi(\beta\theta)$  is upper bounded by the following:

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

Taking limits as  $\beta \rightarrow \infty$  corresponds to computing a recession function.  
(Rockafellar, 1970)

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

$$\max_{\mathbf{x} \in \mathcal{X}^n} \langle \theta, \phi(\mathbf{x}) \rangle \leq \max_{\mu \in \text{OUT}(K_n)} \langle \theta, \mu \rangle$$

For strong coupling, behavior of log-det relaxation (for inference) approaches that of a SDP relaxation for integer programming.

## §4. Summary and open questions

## Summary

- connections between exponential families and convex optimization:
  - (a) Legendre duality and mapping
  - (b) marginal/moment polytopes
  - (c) entropy functions
  
- convex and semidefinite methods lead to useful relaxations for:
  - (a) approximate inference
  - (b) upper bounds on partition functions

# Open questions

- Performance analysis
  - (a) analysis and bounds for different problem classes  
[link to results for integer programming (e.g., Goemans & Williamson, 1995; Nesterov, 1997)]
  - (b) tailoring relaxations to problems by the choice of:
    - (i) approximation to  $\text{MARG}(\phi)$
    - (ii) approximation to  $\Phi^*$
  
- Algorithmic questions
  - (a) (more) efficient algorithms for solving log-determinant relaxation
  - (b) how to exploit problem structure effectively?

## Contact information

Martin Wainwright

`martinw@eecs.berkeley.edu`

Papers at: <http://www.eecs.berkeley.edu/~martinw>