Approximations for Inference in Undirected Graphical Models

Pradeep Ravikumar

Graphical Models, The History

Rennaissance Period

I want to model the world, and I like graphs...
The world was not ready, and so Da Vinci hid clues in architecture and paintings.

History

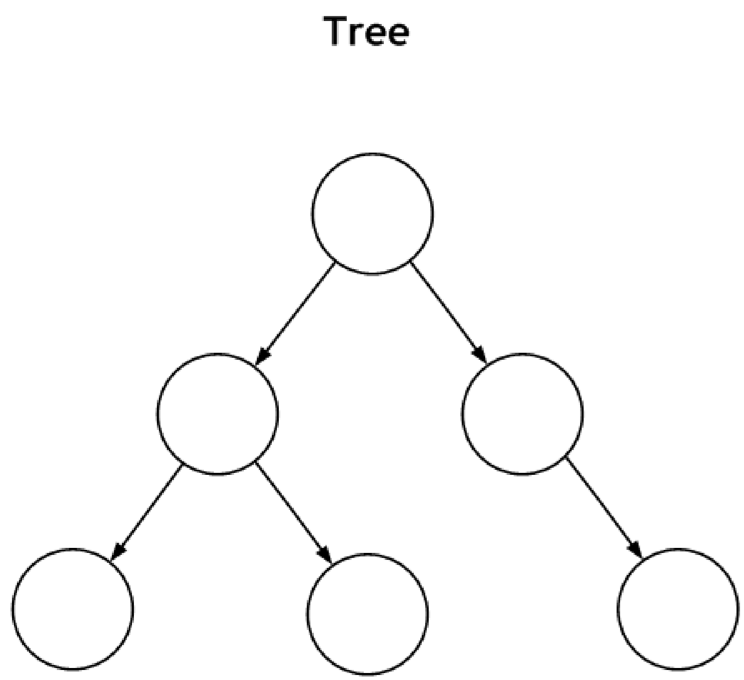
Mid to Late Twentieth Century

Pioneering work of Conspiracy Theorists

The System, it is all **connected**...

History

Late Twentieth Century: people realize that existing scientific literature (Statistics, The Da Vinci Code) offers a marriage between probability theory and graph theory – which can be used to model the world.



History

Common Misconception: Called graphical models after Grafici Modeles, a sculptor protege of Da Vinci.

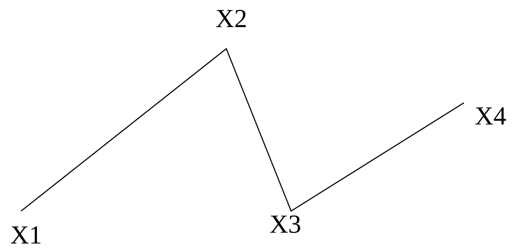
Called Graphical Models because it models stochastic systems using graphs.

History

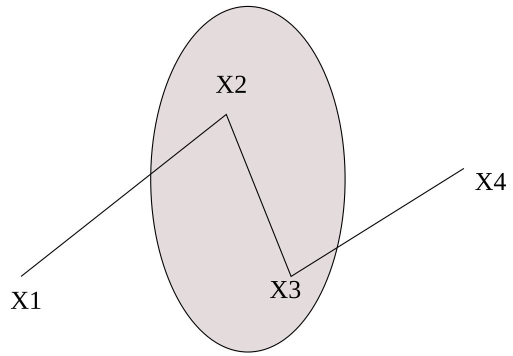
Common Misconception: Called graphical models after Grafici Modeles, a sculptor protege of Da Vinci.

Called Graphical Models because it models stochastic systems using graphs.

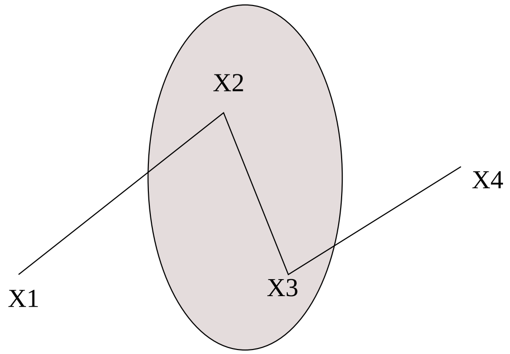
Graphical Models



Graphical Models

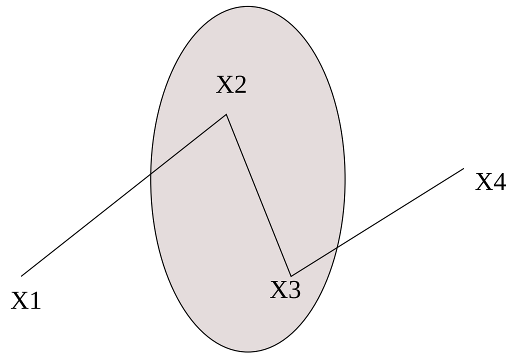


Graphical Models



$$X_1 \perp X_4 \mid X_2,\ X_3$$

Graphical Models



$$X_1 \perp X_4 \,|\, X_2,\, X_3$$

- Global Markov Property
- Graph Encodes Conditional Independencies

Hammersley and Clifford Theorem

$\mathcal{C}$ set of cliques in graph G.

Positive p over X is Markovian with respect to G iff p factorizes according to cliques in G,

$$p(X) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(X_C)$$

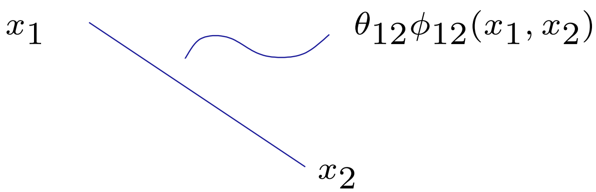
Exponential Family

$$
\begin{aligned}
p(X) &= \frac{1}{Z}\prod_{C\in\mathcal{C}}\psi_C(X_C) \\
&= \exp(\sum_{C\in\mathcal{C}}\log\psi_C(X_C) - \log Z)
\end{aligned}
$$

Exponential family: $p(X;\theta) \;=\; \exp\left(\sum_{\alpha\in\mathcal{C}} \theta_\alpha\phi_\alpha(X) - \Psi(\theta)\right)$

- $\{\phi_\alpha : \alpha \in \mathcal{C}\}$ features
- $\{\theta_\alpha : \alpha \in \mathcal{C}\}$ parameters
- $\{\Psi(\theta) :$ log partition function $\}$

Pairwise Graphical Models



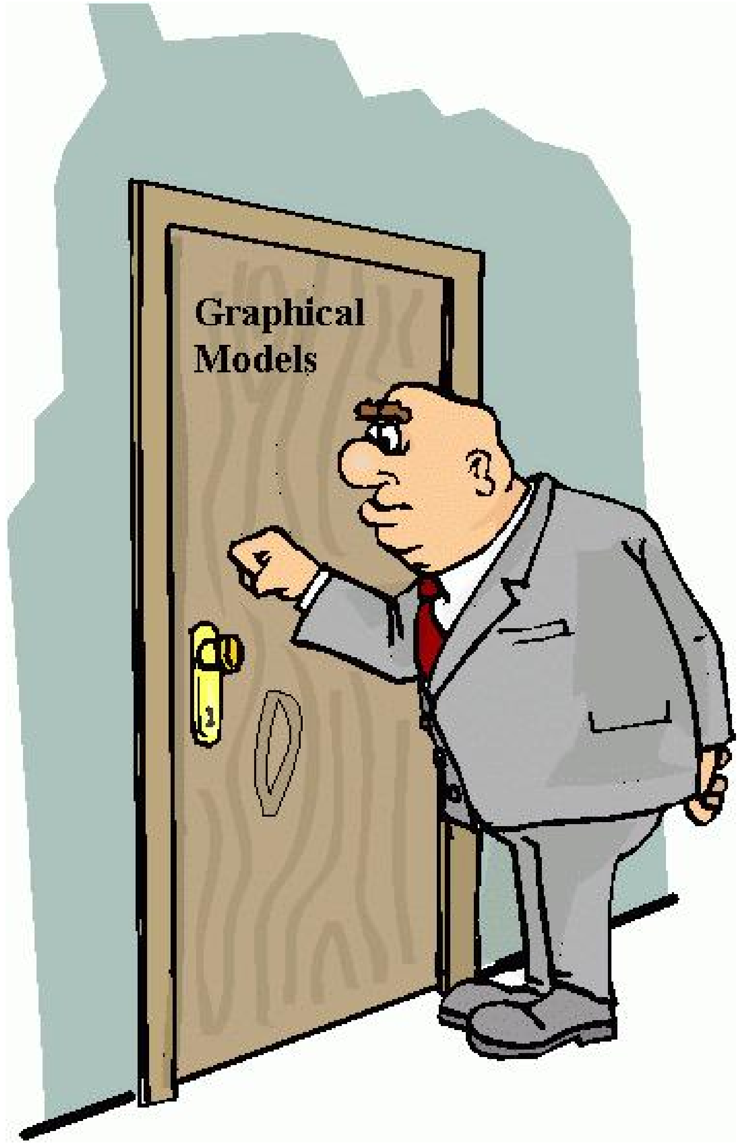
Overcomplete potentials:

$$\mathcal{I}_j(x_s) = \begin{cases} 1 & x_s = j \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{I}_{j,k}(x_s, x_t) = \begin{cases} 1 & x_s = j \text{ and } x_t = k \\ 0 & \text{otherwise}\,. \end{cases}$$

$$p(x|\theta) = \exp\left(\sum_{s,j} \theta_{s;j}\mathcal{I}_j(x_s) + \sum_{s,t;j,k} \theta_{s,j;t,k}\mathcal{I}_{j,k}(x_s, x_t) - \Psi(\theta) \right)$$

Inference



Inference

For undirected model $p(x;\theta) \; = \; \exp\left(\sum_{\alpha \in I} \theta_\alpha \phi_\alpha(x) - \Psi(\theta)\right)$ key inference problems are:

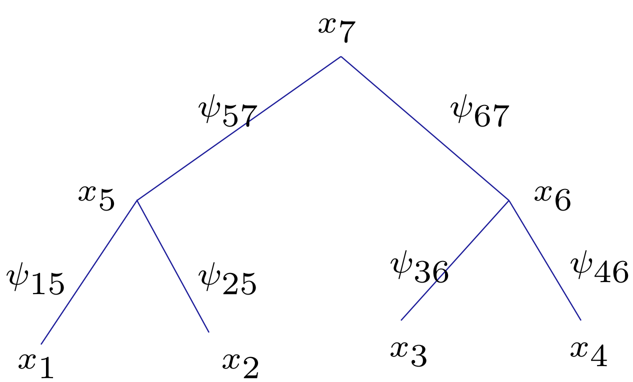
- compute log partition function (normalization constant) $\Psi(\theta)$
- marginals $p(x_A) = \sum_{x_v, v \notin A} p(x)$
- most probable configurations $x^* = \arg\max_x p(x \mid x_L)$

These problems are intractable in full generality.

Approximate inference techniques have focused on sampling and variational methods.

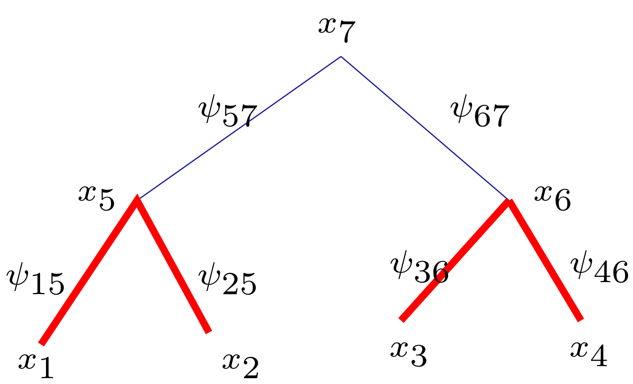
Log Partition Function

$$Z = \log \sum_{x} \prod_{\alpha \in I} \psi_\alpha(x_\alpha)$$



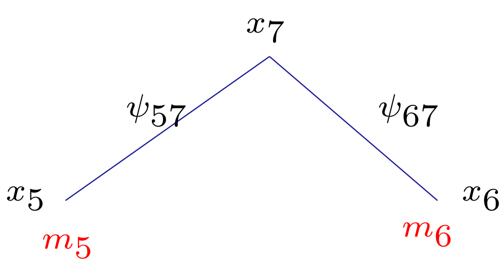
Log Partition Function

$$Z = \log \sum_x \prod_{\alpha \in I} \psi_\alpha(x_\alpha)$$



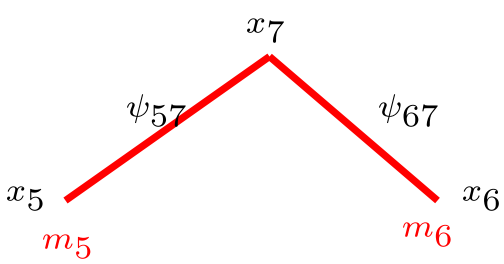
Log Partition Function

$$Z = \log \sum_x \prod_{\alpha \in I} \psi_\alpha(x_\alpha)$$



Log Partition Function

$$Z = \log \sum_{x} \prod_{\alpha \in I} \psi_{\alpha}(x_{\alpha})$$



Log Partition Function

$$Z = \log \sum_{x} \prod_{\alpha \in I} \psi_\alpha(x_\alpha)$$

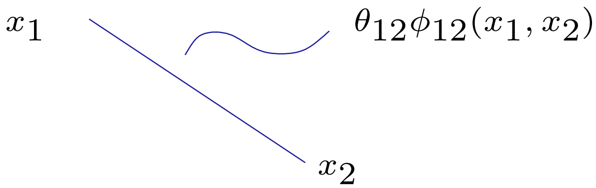
Exponential in tree-width.

Preconditioner Approximations

$$Z = \log \sum_{x} \prod_{\alpha \in I} \psi_\alpha(x_\alpha)$$

Recent scientific computing developments allow us to propose a preconditioner based approximation.

Pairwise Graphical Models



Matrix form:

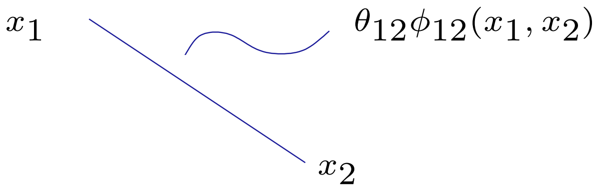
$$\Theta := \begin{pmatrix} \theta_{11} & \dots & \theta_{1n} \\ \vdots & \vdots & \vdots \\ \theta_{n1} & \dots & \theta_{nn} \end{pmatrix} \quad \Phi(x) := \begin{pmatrix} \phi_{11}(x_1, x_1) & \dots & \phi_{1n}(x_1, x_n) \\ \vdots & \vdots & \vdots \\ \phi_{n1}(x_n, x_1) & \dots & \phi_{nn}(x_n, x_n) \end{pmatrix}$$

Notation: $tr(AB) = \langle\langle A, B \rangle\rangle = \sum_{ij} A_{ij} B_{ij}$.

Energy $\sim \sum_{st} \theta_{st}\phi_{st}(x_s, x_t) = \langle\langle \Theta, \Phi(x) \rangle\rangle$

$$p(x;\theta) = \frac{\exp \langle\langle \Theta, \Phi(x) \rangle\rangle}{Z(\theta)} \qquad \mathcal{Z}(\theta) = \sum_{x \in \mathcal{X}} \exp \langle\langle \Theta, \Phi(x) \rangle\rangle$$

Pairwise Graphical Models



Matrix form:

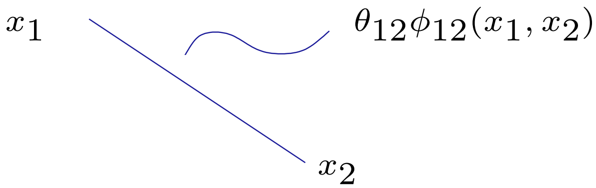
$$\Theta := \begin{pmatrix} \theta_{11} & \dots & \theta_{1n} \\ \vdots & \vdots & \vdots \\ \theta_{n1} & \dots & \theta_{nn} \end{pmatrix} \quad \Phi(x) := \begin{pmatrix} \phi_{11}(x_1, x_1) & \dots & \phi_{1n}(x_1, x_n) \\ \vdots & \vdots & \vdots \\ \phi_{n1}(x_n, x_1) & \dots & \phi_{nn}(x_n, x_n) \end{pmatrix}$$

Notation: $tr(AB) = \langle\langle A, B \rangle\rangle = \sum_{ij} A_{ij} B_{ij}$.

Energy $\sim \sum_{st} \theta_{st}\phi_{st}(x_s, x_t) = \langle\langle \Theta, \Phi(x) \rangle\rangle$

$$p(x;\theta) = \frac{\exp \langle\langle \Theta, \Phi(x) \rangle\rangle}{Z(\theta)} \qquad \mathcal{Z}(\theta) = \sum_{x \in \mathcal{X}} \exp \langle\langle \Theta, \Phi(x) \rangle\rangle$$

Pairwise Graphical Models



Matrix form:

$$\Theta := \begin{pmatrix} \theta_{11} & \dots & \theta_{1n} \\ \vdots & \vdots & \vdots \\ \theta_{n1} & \dots & \theta_{nn} \end{pmatrix} \quad \Phi(x) := \begin{pmatrix} \phi_{11}(x_1, x_1) & \dots & \phi_{1n}(x_1, x_n) \\ \vdots & \vdots & \vdots \\ \phi_{n1}(x_n, x_1) & \dots & \phi_{nn}(x_n, x_n) \end{pmatrix}$$

Notation: $tr(AB) = \langle\langle A, B \rangle\rangle = \sum_{ij} A_{ij} B_{ij}$.

Energy $\sim \sum_{st} \theta_{st}\phi_{st}(x_s, x_t) = \langle\langle \Theta, \Phi(x) \rangle\rangle$

$$p(x;\theta) \;=\; \frac{\exp \langle\langle \Theta, \Phi(x) \rangle\rangle}{Z(\theta)} \qquad \mathcal{Z}(\theta) \;=\; \sum_{x \in \mathcal{X}} \exp \langle\langle \Theta, \Phi(x) \rangle\rangle$$

Preconditioning Linear Systems

$A\,x = b$, A $n \times n$, m non-zero entries

Direct methods $\sim O(n^3)$

$B^{-1}Ax = B^{-1}b$

$B \sim A$, $B^{-1}A \sim I$

Simpler?

Preconditioning Linear Systems

$A\,x = b$, A $n \times n$, m non-zero entries

Direct methods $\sim O(n^3)$

$B^{-1}Ax = B^{-1}b$

$B \sim A$, $B^{-1}A \sim I$

Simpler?

Preconditioning Linear Systems

$B^{-1}Ax = B^{-1}b$

Computation required for ϵ-approximate solution:

$$T(A) = \sqrt{\kappa(A,B)}\,(m + T(B))\log\left(\frac{1}{\epsilon}\right)$$

$\kappa(A,B)$ is "quality of approximation" and $T(B)$ is time to solve $By = c$.

Objective: Find "sparse" matrix B, e.g., a spanning tree, with small condition number $\kappa(A,B)$

Preconditioning Linear Systems

$B^{-1}Ax = B^{-1}b$

Computation required for ϵ-approximate solution:

$$T(A) = \sqrt{\kappa(A,B)}\,(m + T(B)) \log\left(\frac{1}{\epsilon}\right)$$

$\kappa(A,B)$ is "quality of approximation" and $T(B)$ is time to solve $By = c$.

Objective: Find "sparse" matrix B, e.g., a spanning tree, with small condition number $\kappa(A,B)$

Recent Developments in Preconditioners

Nearly optimal linear system solvers: $O\left(m \log^{O(1)} n\right)$

- *Vaidya* (1990)
- *Gremban and Miller* (1996)
- *Boman and Hendrickson* (2002)
- *Spielman and Teng* (2003)
- *Elkin et al.* (2004)

Matrix A $\sim$ Laplacian of a graph

Methods require diagonally dominant matrices: $A_{ii} \geq \sum_{j \neq i} |A_{ij}|$

Basic Approach

Intuition: $\langle\langle \Theta, \Phi(x) \rangle\rangle \sim x^\top A x$.

If $B \sim A$, $\kappa(A, B)$ is small:

$$\kappa(A,B) = \max_x \frac{x^\top A x}{x^\top B x} \Bigg/ \min_x \frac{x^\top A x}{x^\top B x} = \lambda_{max}(A,B) \, / \lambda_{min}(A,B)$$

- B chosen to minimize condition number, rather than KL
  - ▹ Scale B appropriately
- If B is a tree, sums can be computed efficiently
- B can be used to approximate event probabilities

Basic Approach

Intuition: $\langle\langle \Theta, \Phi(x) \rangle\rangle \sim x^\top A x$.

If $B \sim A$, $\kappa(A, B)$ is small:

$$\kappa(A,B) = \max_x \frac{x^\top A x}{x^\top B x} \Bigg/ \min_x \frac{x^\top A x}{x^\top B x} = \lambda_{max}(A,B) \,/ \lambda_{min}(A,B)$$

- B chosen to minimize condition number, rather than KL
  - ▷ Scale B appropriately
- If B is a tree, sums can be computed efficiently
- B can be used to approximate event probabilities

Generalized Eigenvalues for Graphical Models

Bounds on energy: $\langle\langle C, \Phi(x)\rangle\rangle \;\leq\; \langle\langle \Theta, \Phi(x)\rangle\rangle \;\leq\; \langle\langle B, \Phi(x)\rangle\rangle$

Imply bounds on partition function and probabilities:

$$\frac{\exp\langle\langle C, \Phi(x)\rangle\rangle}{Z(B)} \;\leq\; p(x;\Theta) \;\leq\; \frac{\exp\langle\langle B, \Phi(x)\rangle\rangle}{Z(C)}$$

Want B to be as similar to Θ as possible. For upper bound, this leads to optimization problem

$$B^\star \;=\; \arg\max_B \;\min_x \quad \frac{\langle\langle \Theta, \Phi(x)\rangle\rangle}{\langle\langle B, \Phi(x)\rangle\rangle}$$

$$\text{such that} \quad \frac{\langle\langle \Theta, \Phi(x)\rangle\rangle}{\langle\langle B, \Phi(x)\rangle\rangle} \;\leq\; 1$$

Generalized Eigenvalues for Graphical Models

Bounds on energy: $\langle\langle C, \Phi(x) \rangle\rangle \leq \langle\langle \Theta, \Phi(x) \rangle\rangle \leq \langle\langle B, \Phi(x) \rangle\rangle$
Imply bounds on partition function and probabilities:

$$\frac{\exp \langle\langle C, \Phi(x) \rangle\rangle}{Z(B)} \leq p(x; \Theta) \leq \frac{\exp \langle\langle B, \Phi(x) \rangle\rangle}{Z(C)}$$

Want B to be as similar to Θ as possible. For upper bound, this leads to optimization problem

$$B^\star = \arg\max_B \min_x \frac{\langle\langle \Theta, \Phi(x) \rangle\rangle}{\langle\langle B, \Phi(x) \rangle\rangle}$$

$$\text{such that} \quad \frac{\langle\langle \Theta, \Phi(x) \rangle\rangle}{\langle\langle B, \Phi(x) \rangle\rangle} \leq 1$$

Generalized Eigenvalues for Graphical Models

Bounds on energy: $\langle\langle C, \Phi(x)\rangle\rangle \le \langle\langle \Theta, \Phi(x)\rangle\rangle \le \langle\langle B, \Phi(x)\rangle\rangle$
Imply bounds on partition function and probabilities:

$$\frac{\exp \langle\langle C, \Phi(x)\rangle\rangle}{Z(B)} \le p(x;\Theta) \le \frac{\exp \langle\langle B, \Phi(x)\rangle\rangle}{Z(C)}$$

Want B to be as similar to Θ as possible. For upper bound, this leads to optimization problem

$$B^\star = \arg\max_B \min_x \frac{\langle\langle \Theta, \Phi(x)\rangle\rangle}{\langle\langle B, \Phi(x)\rangle\rangle}$$

$$\text{such that} \quad \frac{\langle\langle \Theta, \Phi(x)\rangle\rangle}{\langle\langle B, \Phi(x)\rangle\rangle} \le 1$$

Reductions

Have to solve constrained minimax and maximin problems!

Generalized Eigenvalues for Graphical Models

For a graphical model with potential function matrix Φ and a pair of matrices (A, B), we define

Generalized graphical model eigenvalues:

$$\lambda^{\Phi}_{\mathsf{max}}(A,B) \;=\; \max_{x:\, \langle\langle B,\Phi(x)\rangle\rangle \neq 0} \frac{\langle\langle A,\Phi(x)\rangle\rangle}{\langle\langle B,\Phi(x)\rangle\rangle}$$

$$\lambda^{\Phi}_{\mathsf{min}}(A,B) \;=\; \min_{x:\, \langle\langle B,\Phi(x)\rangle\rangle \neq 0} \frac{\langle\langle A,\Phi(x)\rangle\rangle}{\langle\langle B,\Phi(x)\rangle\rangle} \;=\; \frac{1}{\lambda^{\Phi}_{\mathsf{max}}(B,A)}$$

Graphical model condition number:

$$\kappa^{\Phi}(A,B) = \frac{\lambda^{\Phi}_{\mathsf{max}}(A,B)}{\lambda^{\Phi}_{\mathsf{min}}(A,B)}$$

Condition Number Relaxation

Proposition 1. Let $C = \arg\min_B \kappa^{\Phi}(\Theta, B)$. Then the optimal upper bound matrix $B^U = \lambda_{\max}(\Theta, C)\, C$ and the optimal lower bound matrix $B^L = \lambda_{\min}(\Theta, C)\, C$.

Optimizing the constrained minimax and maximin problems can be reduced to minimizing (graphical model) condition numbers!

Condition Number Relaxation

Proposition 2. For a potential function matrix $\Phi(x) \succeq 0$ then $\lambda^{\Phi}_{\max}(\Theta, B) \leq \lambda_{\max}(\Theta, B)$. If $B \succeq 0$ then $\kappa^{\Phi}(\Theta, B) \leq \kappa(\Theta, B)$.

Optimizing graphical model CN can be reduced to minimizing classical CN!

Recipe for Graphical Model Preconditioner Approximations:

- $B^* = \arg\min_B \kappa(\Theta, B)$
- Upper bound matrix $\lambda_{\max}(\Theta, B^*)\ B^*$
- Lower bound matrix $\lambda_{\min}(\Theta, B^*)\ B^*$

Reduction to Ising Form

For the Ising potential function, $\phi_{ij}(x_i, x_j) = x_i x_j$

$$\kappa^{\Phi}(A, B) = \kappa(A, B)$$

Can reduce any potential function ϕ over discrete k-ary random variables X_i, to a binary Ising potential function:

$$\mathcal{E}(x) \;=\; \sum_{(i,j)\in E} \theta_{ij}\phi_{ij}(x_i, x_j) \;=\; \bar{x}^{\top} A(\theta, \phi)\bar{x}$$

where $\bar{x}_{(i,l)} = \delta(x_i, l)$ and

$$A(\theta, \phi) \;=\; \left[A_{(i,l),(j,m)}(\theta, \phi)\right] \;=\; [\theta_{ij}\phi_{ij}(l, m)]$$

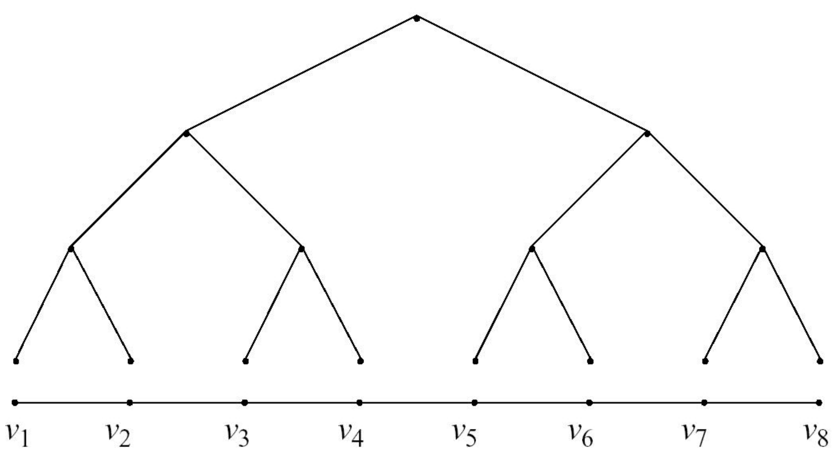
Solve for preconditioner bounds of the matrix $A(\theta, \phi)$

Simple Preconditioners

Vaidya's Spanning Tree Preconditioner

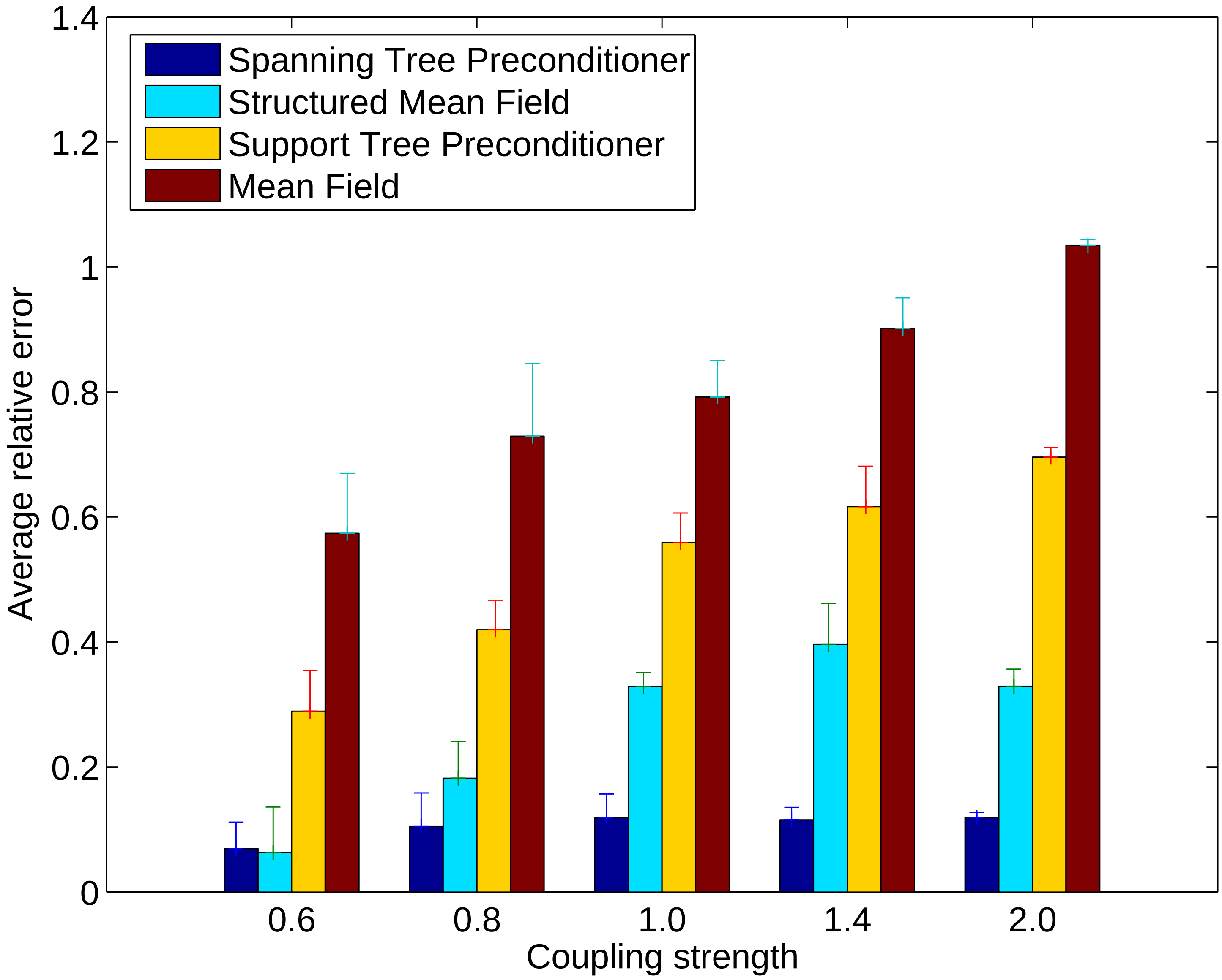
- Matrix corresponding to the maximum spanning tree of the given graph.
- Requires the parameter matrix to be Laplacian.

Gremban-Miller Support Tree Preconditioner

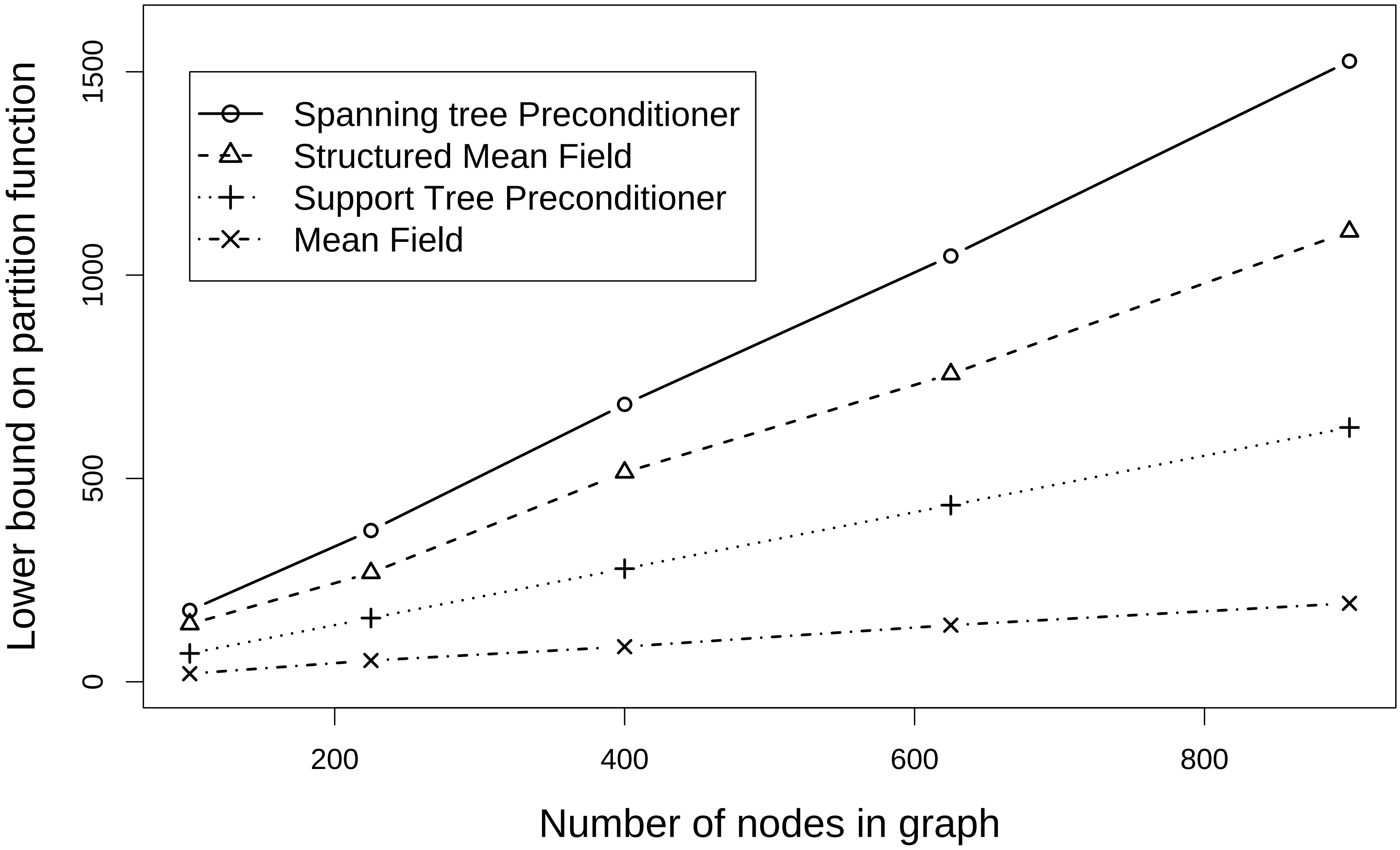


$$\begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix} \sim T$$

Experiments: $\log Z(\theta)$ Lower Bounds, 2D Grid



Experiments

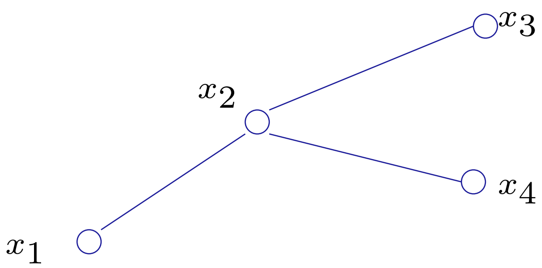


Preconditioner Approximations

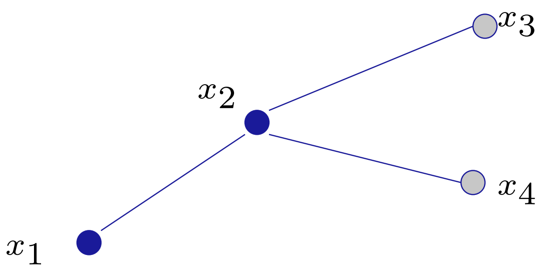
- The framework yields upper and lower bounds on energy, and consequently on the log-partition function, general event probabilities, and the MAP energy.

- The procedure has a low time complexity: both the construction of a sparse preconditioner and inference using a sparse (e.g. tree-based) preconditioner matrix are typically linear.

MAP Estimation

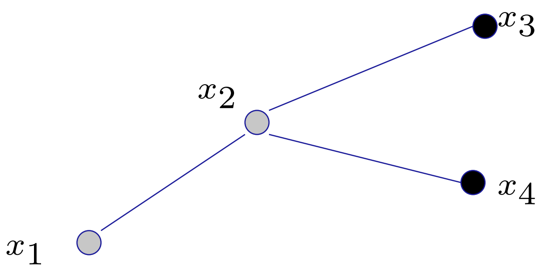


MAP Estimation



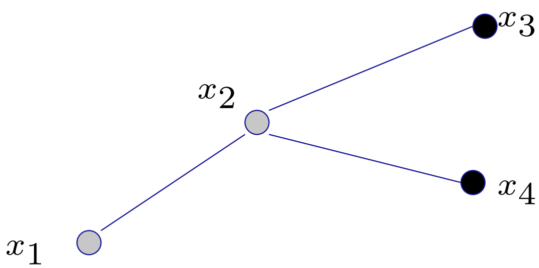
$$\mathrm{PROB} = 0.01$$

MAP Estimation



$$\text{PROB} = 0.2$$

MAP Estimation



$$\text{PROB} = 0.2$$

Most Probable Configuration?

MAP Estimation

$$p(x|\theta) \propto \exp\left(\sum_{s,j} \theta_{s;j} \mathcal{I}_j(x_s) + \sum_{s,t;j,k} \theta_{s,j;t,k} \mathcal{I}_{j,k}(x_s, x_t)\right).$$

$$x^* = \arg\max_x \sum_{s,j} \theta_{s;j} \mathcal{I}_j(x_s) + \sum_{s,t;j,k} \theta_{s,j;t,k} \mathcal{I}_{j,k}(x_s, x_t).$$

Integer Linear Program

$$I_j(x_s) \sim \mu_1(s;j)$$

$$I_{j,k}(x_s, x_t) \sim \mu_2(s,j;t,k)$$

$$\begin{aligned}
\text{max} \quad & \sum_{s;j} \theta_{s;j}\, \mu_1(s;j) + \sum_{s,t;j,k} \theta_{s,j;t,k}\, \mu_2(s,j;t,k) \\
\text{such that} \quad & \sum_{k} \mu_2(s,j;t,k) = \mu_1(s;j) \\
& \sum_{j} \mu_1(s;j) = 1 \\
& \mu_1(s;j) \in \{0,1\} \\
& \mu_2(s,j;t,k) \in \{0,1\}.
\end{aligned}$$

Linear Relaxation

$$\max \quad \sum_{s;j} \theta_{s;j}\, \mu_1(s;j) + \sum_{s,t;j,k} \theta_{s,j;t,k}\, \mu_2(s,j;t,k)$$

$$\text{such that} \quad \sum_{k} \mu_2(s,j;t,k) = \mu_1(s;j)$$

$$\sum_{j} \mu_1(s;j) = 1$$

$$0 \leq \mu_1(s;j) \leq 1$$

$$0 \leq \mu_2(s,j;t,k) \leq 1.$$

Linear Relaxation

LP $\sim$ Chekuri, Khanna, Naor, Zosin

Tree-reweighted Belief Propagation, Dual of LP $\sim$ Wainwright, Jaakkola, Willsky

Quadratic Relaxation

$$I_{j,k}(x_s, x_t) = I_j(x_s) I_k(x_t)$$

$$\mu_2(s, j; t, k) \sim \mu_1(x_s; j) \mu_1(x_t; k)$$

$O(|E|K^2) \rightarrow O(nK)$ variables!

Quadratic Integer Program

$$\max \quad \sum_{s;j} \theta_{s;j} \mu(s;j) + \sum_{s,t;j,k} \theta_{s,j;t,k}\, \mu(s;j)\, \mu(t;k)$$

$$\text{subject to} \quad \sum_{j} \mu(s;j) = 1$$

$$\mu(s;j) \in \{0,1\}$$

Quadratic Relaxation

$$\begin{aligned}
\text{max} \quad & \sum_{s;j} \theta_{s;j}\mu(s;j) + \sum_{s,t;j,k} \theta_{s,j;t,k}\,\mu(s;j)\,\mu(t;k) \\
\text{subject to} \quad & \sum_{j} \mu(s;j) = 1 \\
& 0 \leq \mu(s;j) \leq 1
\end{aligned}$$

Theorem: QP Relaxation is equivalent to the MAP problem.

- Relaxation is tight!
- MAP is in P when $\{\Theta_{s,j;t,k}\}$ is negative semi-definite

Quadratic Relaxation

$$\begin{aligned} \text{max} \quad & \sum_{s;j} \theta_{s;j} \mu(s;j) + \sum_{s,t;j,k} \theta_{s,j;t,k}\, \mu(s;j)\, \mu(t;k) \\ \text{subject to} \quad & \sum_j \mu(s;j) = 1 \\ & 0 \leq \mu(s;j) \leq 1 \end{aligned}$$

Theorem: QP Relaxation is equivalent to the MAP problem.

- Relaxation is tight!
- MAP is in P when $\{\Theta_{s,j;\,t,k}\}$ is negative semi-definite

Convex approximation

Want $-\Theta \sim$ Positive semi-definite

Convex approximation

Want $-\Theta \sim$ Diagonally Dominant

Convex approximation

Want $-\Theta \sim$ Diagonally Dominant

$-\Theta_{s,j;s,j} \geq \sum_{t,k} |\Theta_{s,j;\,t,k}|$

Solution: Substract from diagonal?

Convex approximation

Want $-\Theta \sim$ Diagonally Dominant

$-\Theta_{s,j;s,j} \geq \sum_{t,k} |\Theta_{s,j;\,t,k}|$

Solution: Substract from diagonal?

Diagonally Dominant

$$\mu(s;j) \in \{0,1\} \quad , \quad \mu(s;j)^2 = \mu(s;j)$$

$$\mu^\top \begin{bmatrix} -d_1 & & \\ & -d_2 & \\ & & \ddots \end{bmatrix} \mu + d^\top \mu = 0$$

Diagonally Dominant

$\Theta' = \Theta - \mathsf{diag}(d(s;j))$

$\theta'_{s;j} = \theta_{s;j} + d(s;j)$

$$\begin{aligned} \max_{\mu} \quad & \sum_{s;j} \theta'_{s;j}\, \mu(s;j) + \sum_{s,t;j,k} \theta'_{s,j;t,k}\, \mu(s;j)\, \mu(t;k) \\ \text{such that} \quad & \sum_{j} \mu(s;j) = 1 \\ & \mu(s;j) \in \{0,1\} \end{aligned}$$

Convex Approximation

$$\Theta' = \Theta - \text{diag}(d(s;j))$$

$$\theta'_{s;j} = \theta_{s;j} + d(s;j)$$

$$\begin{aligned} \max_{\mu} \quad & \sum_{s;j} \theta'_{s;j}\, \mu(s;j) + \sum_{s,t;j,k} \theta'_{s,j;t,k} \mu(s;j)\mu(t;k) \\ \text{such that} \quad & \sum_{j} \mu(s;j) = 1 \\ & \mu(s;j) \in [0,1] \end{aligned}$$

Convex Approximation

- Convex QP (with simple box constraints); polynomial time
- Not tight!
- Additive Guarantee: $E(y^{CVX}) \geq E^* - \frac{1}{4}\sum_{s,i} d(s;i)$

Iterative Procedure

$$\max \quad \sum_{s;j} \theta_{s;j}\mu(s;j) + \sum_{s,t;j,k} \theta_{s,j;t,k}\,\mu(s;j)\,\mu(t;k)$$

$$\text{subject to} \quad \sum_{j} \mu(s;j) = 1$$

$$0 \leq \mu(s;j) \leq 1$$

Co-ordinate Ascent: Optimize $\mu(s;.)$ for node s, fixing values of other nodes,

$$\mu(s;.) \; = \; \max_{\mu(s;.)} \sum_{j} \theta_{s;j}\mu(s;j) + \sum_{t;j,k} \theta_{s,j;t,k}\,\mu(s;j)\,\mu(t;k)$$

subject to $\sum_j \mu(s;j) = 1$.

Iterative Procedure

$$\begin{aligned} \text{max} \quad & \sum_{s;j} \theta_{s;j} \mu(s;j) + \sum_{s,t;j,k} \theta_{s,j;t,k}\, \mu(s;j)\, \mu(t;k) \\ \text{subject to} \quad & \sum_j \mu(s;j) = 1 \\ & 0 \le \mu(s;j) \le 1 \end{aligned}$$

Co-ordinate Ascent: Optimize $\mu(s;.)$ for node s, fixing values of other nodes,

$$\mu(s;.) \;=\; \max_{\mu(s;.)} \sum_j \theta_{s;j} \mu(s;j) + \sum_{t;j,k} \theta_{s,j;t,k}\, \mu(s;j)\, \mu(t;k)$$

subject to $\sum_j \mu(s;j) = 1$.

Iterative Procedure

$$\mu(s;.) \;=\; \max_{\mu(s;.)} \sum_j \theta_{s;j} \mu(s;j) + \sum_{t;j,k} \theta_{s,j;t,k}\, \mu(s;j)\, \mu(t;k)$$

Solution of fixed point equation,

$$j^*(s) = \arg\max_j \; \theta_{s;j} + \sum_{t;j,k} \theta_{s,j;t,k} \mu(t;k)$$

and setting $\mu(s,j) = \mathcal{I}_{j^*(s)}(j)$.

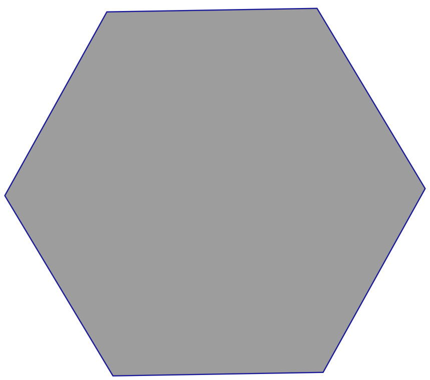
This is the Iterative Conditional Modes algorithm! (Besag 86)

Alternatively, use conjugate gradient for the convex approximation.

Inner Polytope Approximations

Consider Wainwright et al's polytope formulation of MAP,

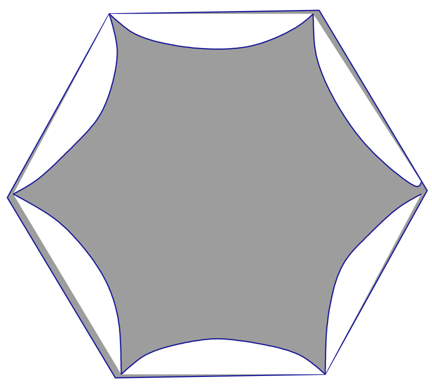
$$\mu^* = \max_x \langle \theta, \phi \rangle = \sup_{\mu \in \mathcal{M}} \langle \theta, \mu \rangle$$



Inner Polytope Approximations

If $M_I \subset \mathcal{M}$ is any subset of the marginal polytope that includes all of the vertices,

$$\mu^* = \max_x \langle \theta, \phi \rangle = \sup_{\mu \in M_I} \langle \theta, \mu \rangle$$

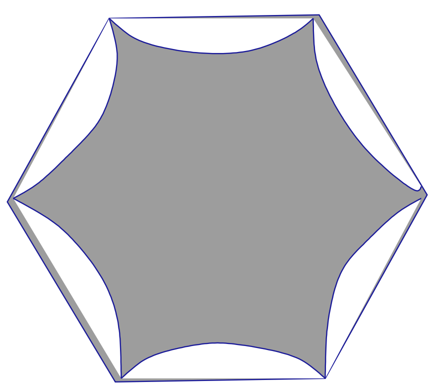


Inner Polytope Approximations

For the given graph G and a subgraph H, let

$$\mathcal{E}(H) = \{\theta' \mid \theta'_{st} = \theta_{st} \, 1_{(s,t) \in H}\}$$

$$\begin{aligned} \mathcal{M}(G; H) &= \{\mu \mid \mu = E_\theta[\phi(x)] \text{ for some } \theta \in \mathcal{E}(H)\}. \\ \mathcal{M}(G; H) &\subseteq \mathcal{M}(G) \end{aligned}$$



Inner Polytope Approximations

Mean Field parameters,

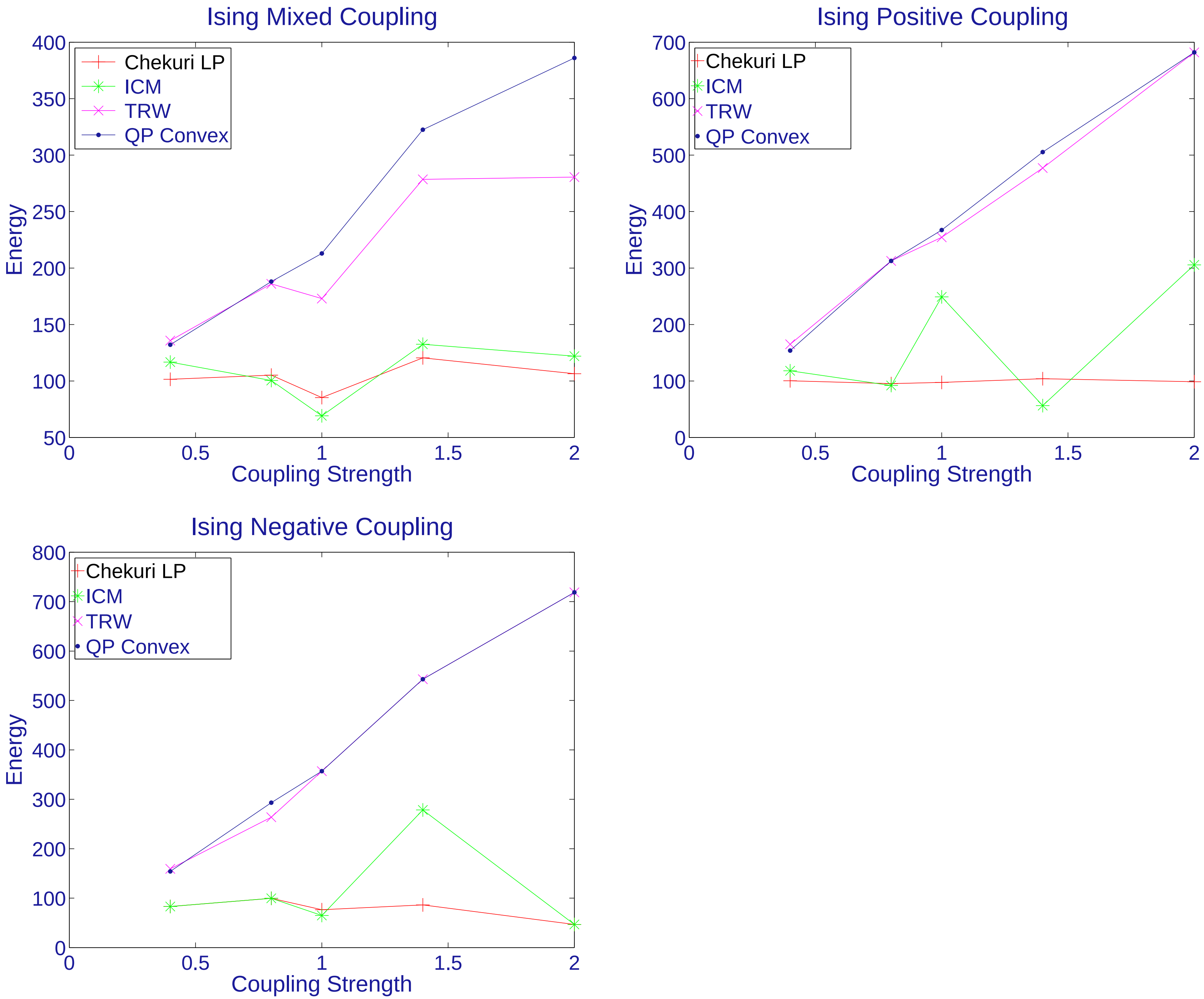
$$\mathcal{M}(G;H_0) \quad = \quad \{\mu(s;j), \mu(s,j;t,k) \,|\, 0 \leq \mu(s;j) \leq 1, \mu(s,j;t,k) = \mu(s;j)\mu(t;k)\}$$

Mean Field Relaxation,

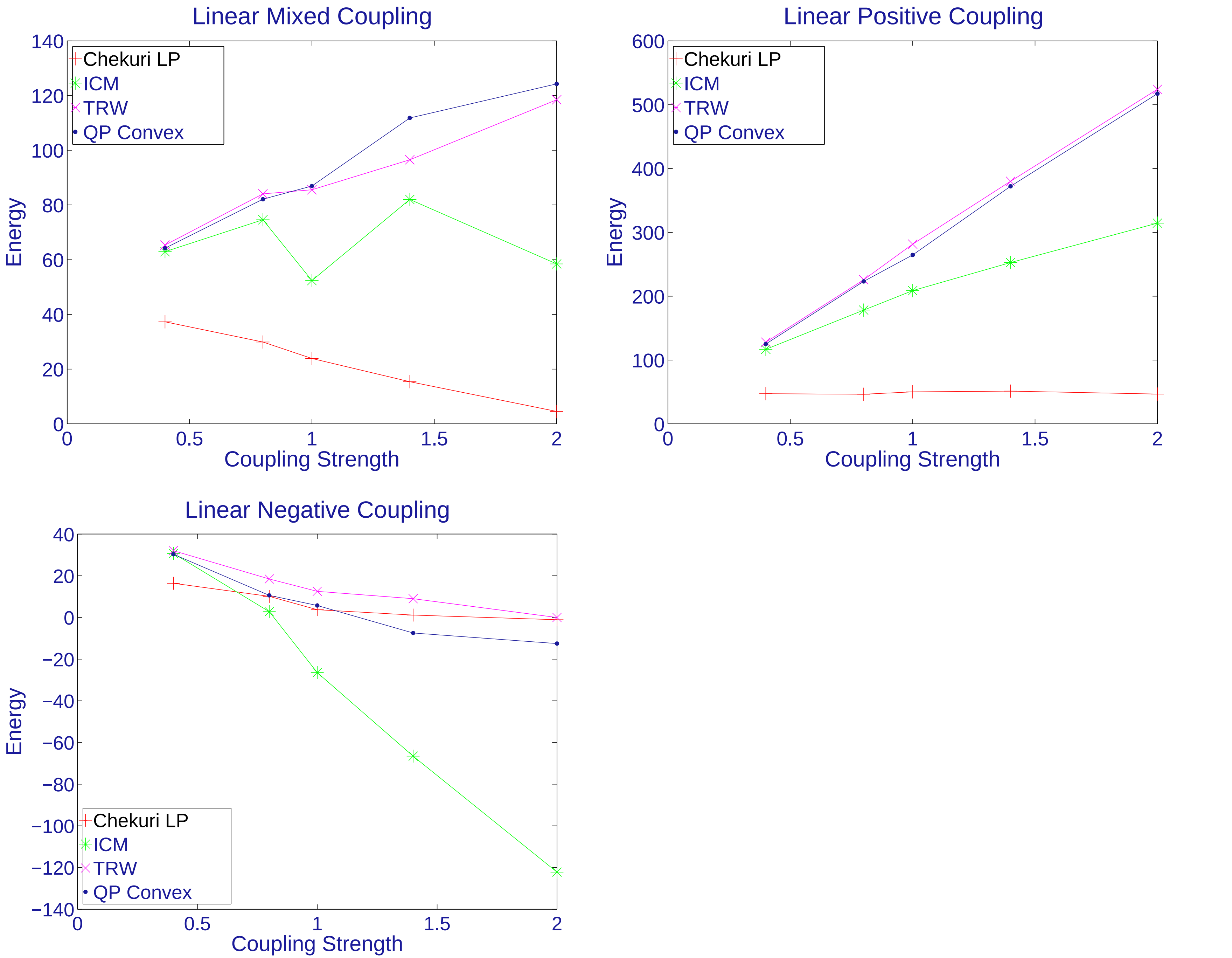
$$\begin{aligned}
& \sup_{\mu \in \mathcal{M}(G;H_0)} \langle \theta, \mu \rangle \\
& \quad = \sup_{\mu \in \mathcal{M}(G;H_0)} \sum_{s;j} \theta_{s;j} \mu(s;j) + \sum_{st;jk} \theta_{s,j;t,k} \mu(s,j;t,k) \\
& \quad = \sup_{\mu \in \mathcal{M}(G;H_0)} \sum_{s;j} \theta_{s;j} \mu(s;j) + \sum_{st;jk} \theta_{s,j;t,k} \mu(s;j)\mu(t;k)
\end{aligned}$$

Experiments

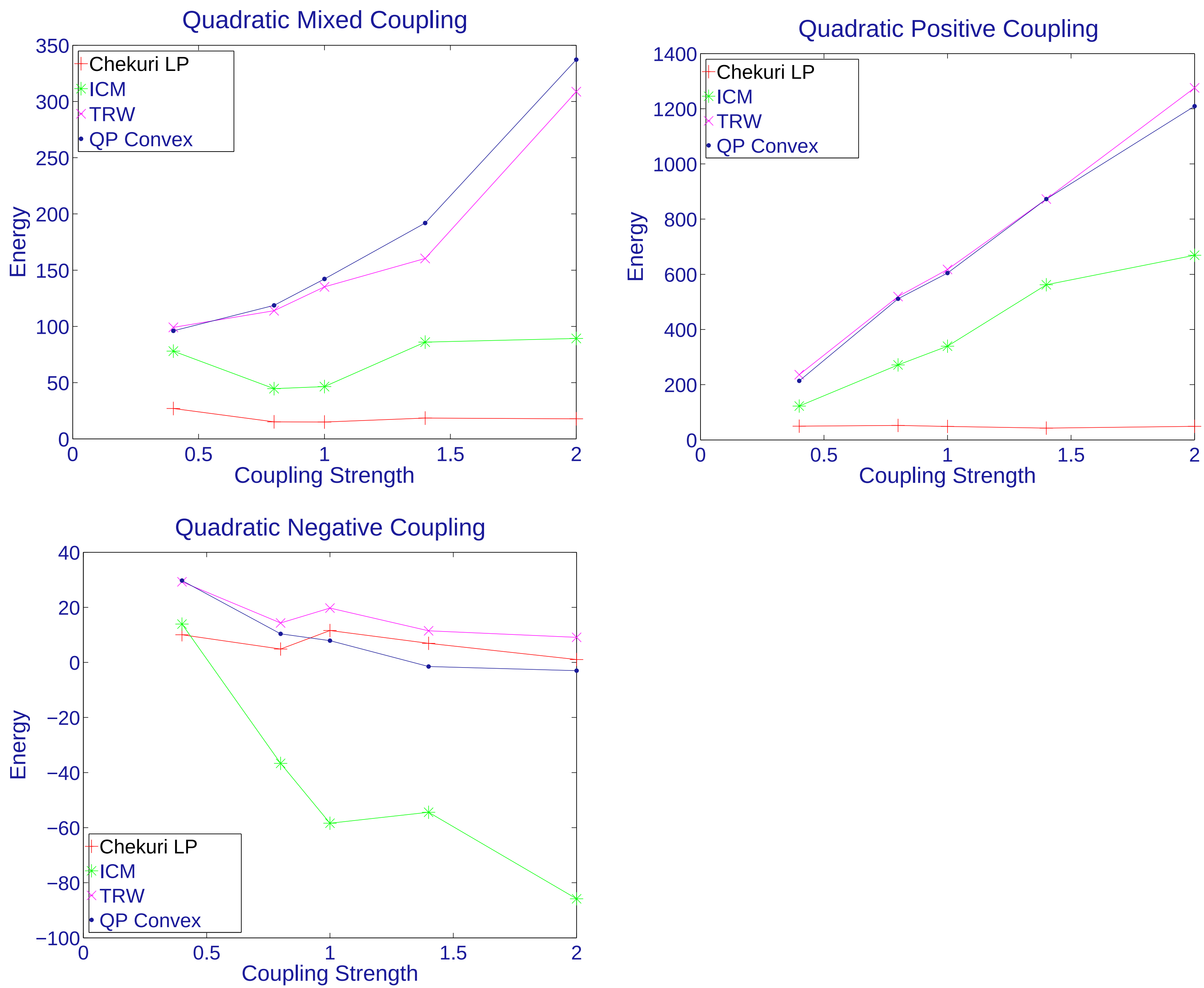
- 10×10 grid graphs; $n = 100$.
- Number of labels, $k = 4$.
- Potential Functions: Ising, Quadratic, Linear, Uniform.
- Methods Compared:
  - ▷ Iterative Conditional Modes
  - ▷ Chekuri LP
  - ▷ Tree-reweighted Max-Product
  - ▷ Convex Approximation to QP



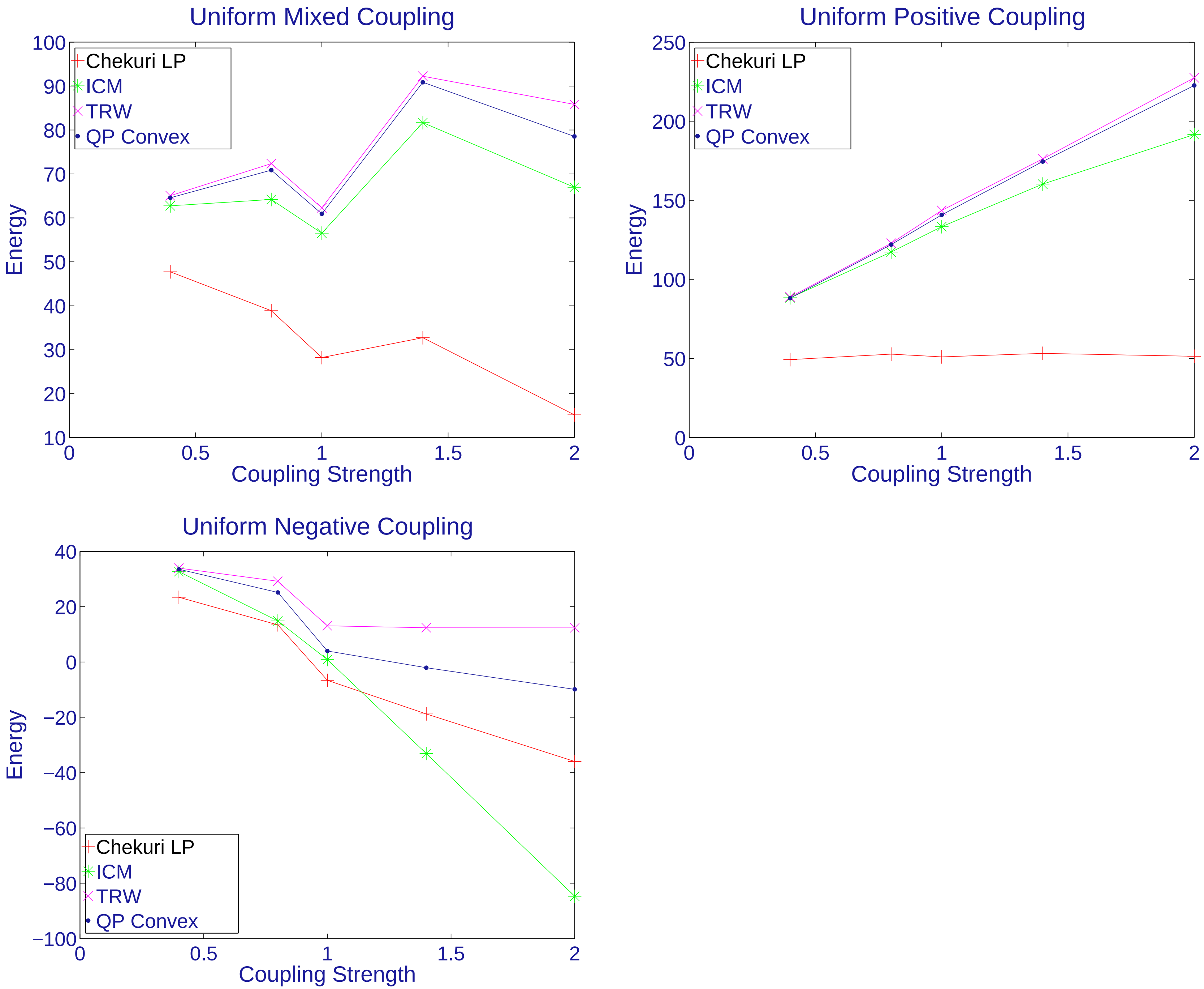
Comparison on 10×10 grid graphs using Ising potentials.



Comparison on 10×10 grid graphs using linear potentials.



Comparison on 10×10 grid graphs using quadratic potentials.



Comparison on 10×10 grid graphs using uniform potentials.

Experiments

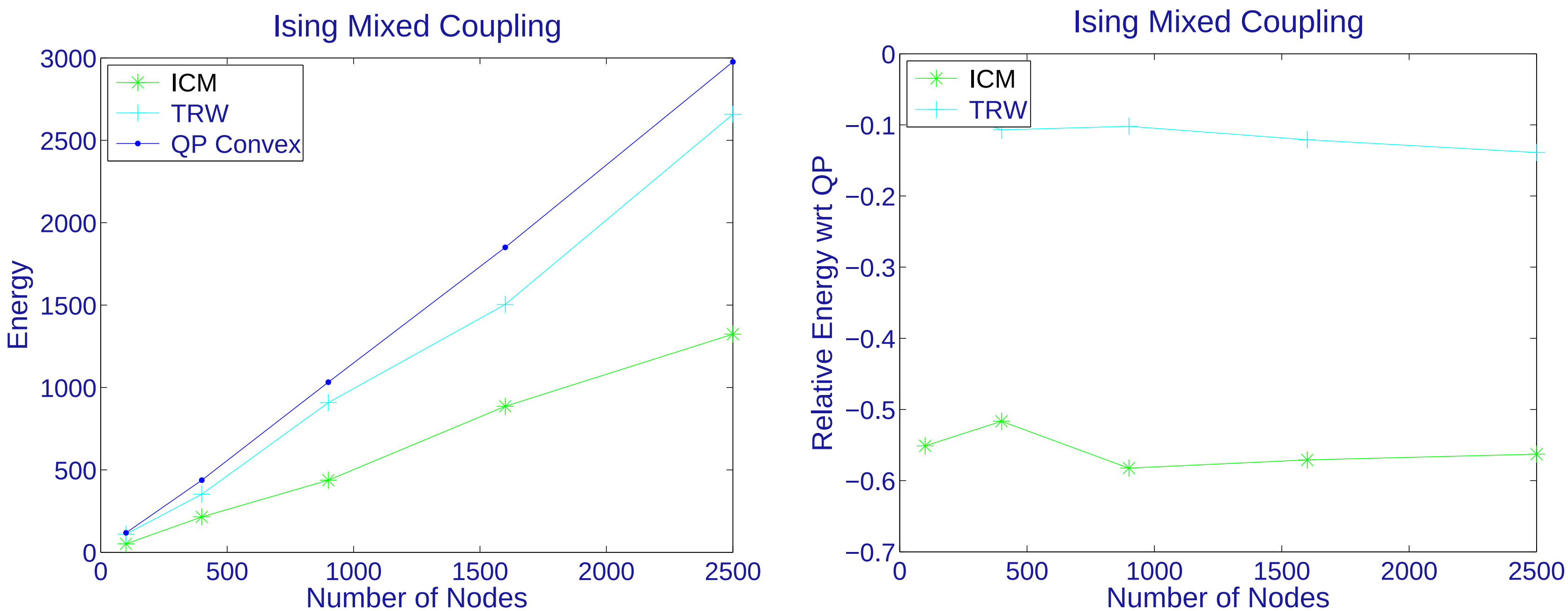


Figure 5: Comparison of ICM and TRW on larger graphs, using Ising potentials with mixed coupling. The right plot shows $(e_{\mathrm{ICM}} - e_{\mathrm{QP}})/e_{\mathrm{ICM}}$ and $(e_{\mathrm{TRW}} - e_{\mathrm{QP}})/e_{\mathrm{TRW}}$.

QP Relaxation

- The QP has $O(nk)$ variables in contrast to $O(|E|k^2)$ variables for linear relaxation.
- The QP more accurately represents objective function: Relaxation is tight!
- There exists a convex approximation with an additive guarantee.
- The QP can be extended to variational inner polytope relaxations.

Structure Selection