Learning Graphical Models by Maximum Entropy Relaxation

Jason K. Johnson

(Joint work with V. Chandrasekaran and A.S. Willsky)

October 4, 2006

- Graphical Models and Chordal graphs

- Exponential Family Models
  - Gaussian model
  - Boltzmann machines (binary variables)
  - Information Projection and Max-Entropy

- Maximum Entropy Relaxation (MER)
  - Convex formulation
  - Primal-dual interior point method
  - Fisher information on chordal graphs
  - Incremental edge selection

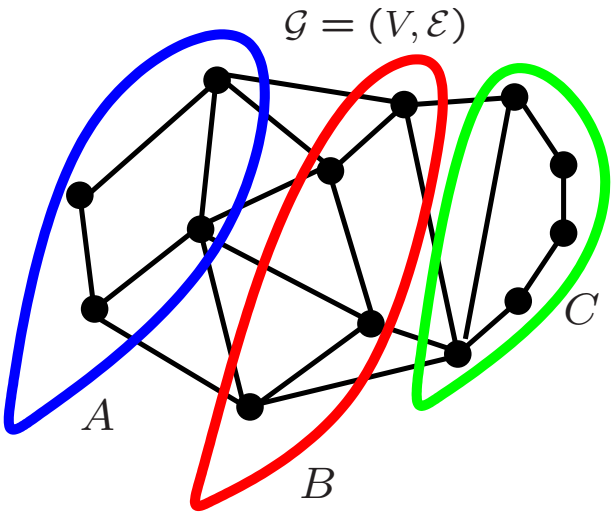
- Examples

Graphical Models

A probability distribution which *factors* over a graph (or hyper-graph) $\mathcal{G} = (V, \mathcal{E})$.

$$p(x_1, \ldots, x_n) \propto \prod_{E \in \mathcal{G}} \psi_E(x_E)$$

where ψ_E, defined on edge $E \subset V$, depends only on variables $x_E = (x_i, i \in E)$.

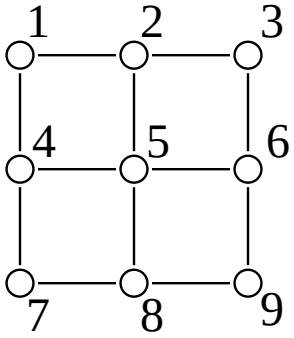


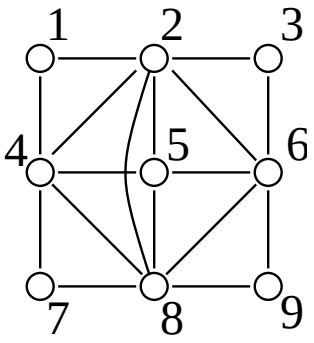
Markov Property: for every $A, B, C \subset V$ where B *separates* A and C in $\mathcal{G}$, x_A and x_C are independent given x_B:

$$p(x_A, x_C | x_B) = p(x_A | x_B) p(x_C | x_B)$$

Chordal Graphs

A graph is *chordal* if every cycle of four or more vertices is cut by a chord.

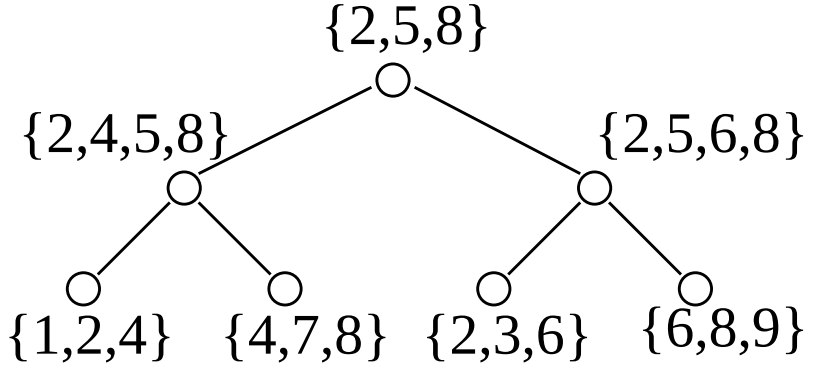




$\Leftrightarrow$ Exists perfect elimination order,

$$(1, 3, 7, 9, 4, 6, 2, 5, 8).$$

$\Leftrightarrow$ Exists a *junction tree* on cliques which satisfies running intersection property.



In thin chordal graphs, efficient algorithms to compute marginal distributions on cliques, and to recover ψ's from marginals.

Exponential Families

Parametric families of probability distributions:

$$p(x) = \exp\{\theta^T \phi(x) - \Phi(\theta)\}$$

based on a set of features $\phi : \mathcal{X} \to \mathbb{R}^d$, with parameter $\theta \in \mathbb{R}^d$.

The *cumulant-generating function* serves to normalize the distribution:

$$\Phi(\theta) = \log \int \exp \theta^T \phi(x) dx$$

and has the properties:

$$\begin{aligned} \nabla\Phi(\theta) &= \mathbb{E}_\theta\{\phi\} \triangleq \eta \\ \nabla^2\Phi(\theta) &= \mathrm{cov}_\theta\{\phi\} \triangleq G(\theta) \end{aligned}$$

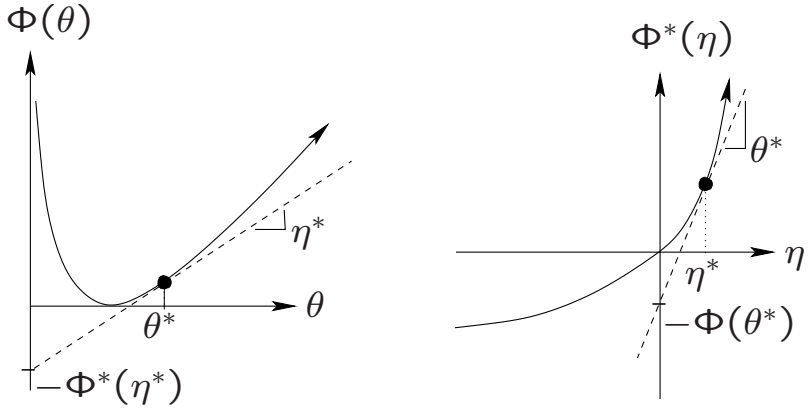
η are the *moment parameters* of the family. $G(\theta)$ is the *Fisher information* in θ.

Variational Principles

Convex duality The *convex conjugate* of Φ equals the *negative entropy* as a function of moments.

$$\Phi^*(\eta) \equiv \max_{\theta}\{\eta \cdot \theta - \Phi(\theta)\} = -h(\eta)$$

Due to convexity of Φ it holds that $(\Phi^*)^* = \Phi$



Learning Given a desired set of moments η^* the corresponding parameters θ^* minimize the convex function:

$$f(\theta) = \Phi(\theta) - \eta^* \cdot \theta$$

Inference Given θ^* the corresponding moments η^* minimize the convex function:

$$g(\eta) = \Phi^*(\eta) - \theta^* \cdot \eta$$

Boltzmann Machine

Let $x_v \in \{0, 1\}$ and

$$P(x) \propto \exp \sum_{E \in \mathcal{G}} \theta_E \phi_E(x_E)$$

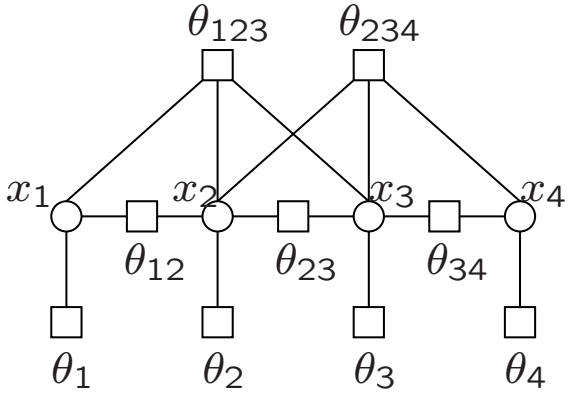
where $\phi_E(x_E) = \prod_{v \in E} x_v$. Moments,

$$\eta_E = P(\{x_i = 1, \text{for all } i \in E\}).$$

Example:

$$\begin{aligned} V &= \{1, 2, 3, 4\} \\ \mathcal{G} &= \{1, 2, 3, 4, 12, 23, 34, 123, 234\} \\ \theta(x) &= \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 \\ &\quad + \theta_{12} x_1 x_2 + \theta_{23} x_2 x_3 + \theta_{34} x_3 x_4 \\ &\quad + \theta_{123} x_1 x_2 x_3 + \theta_{234} x_2 x_3 x_4 \end{aligned}$$

Factor graph representation:



Mobius Transforms

Linear operator M_n^ω on vector space of functions $f : 2^{\{1,\ldots,n\}} \to \mathbb{R}$:

$$(M^\omega f)_A = \sum_{B \subset A} \omega^{|A-B|} f_B$$

Inverse transform given by $M_n^{-\omega}$. Using binary indexing of subsets, e.g. $\{1,3,5\} \equiv 10101$, we have:

$$M_n^\omega = \begin{pmatrix} M_{n-1}^\omega & 0 \\ \omega M_{n-1}^\omega & M_{n-1}^\omega \end{pmatrix}$$

Computed *recursively* with $\mathcal{O}(n2^n)$ complexity.

In the *complete* Boltzmann model, i.e. $\theta, \eta \in \mathbb{R}^{2^n}$,* we can relate θ and η by:

$$\begin{aligned} \eta &= M^T \exp(M\theta) \\ \theta &= M^{-1} \log(M^{-T}\eta) \end{aligned}$$

Which we will user later to perform projection to chordal graphs, and for Fisher information computations.

*With $\theta_\emptyset = -\Phi(\theta), \eta_\emptyset = 1$.

Gaussian Model

Information form zero-mean Gaussian:

$$p(x) \propto \exp\{-\frac{1}{2}x^T Jx\}$$

Moments given by the covariance matrix:

$$P = J^{-1}$$

Markov on $\mathcal{G}$ if $J_{ij} = 0$ for all $\{i, j\} \notin \mathcal{G}$.

Exponential Family Representation:

$$p(x) \propto \exp\{\theta^T \phi(x)\}$$

where

$$\begin{aligned} \phi(x) &= (x_i^2, \forall i) \cup (x_i x_j, \forall \{i, j\}) \\ \theta &= (-\tfrac{1}{2} J_{ii}, \forall i) \cup (-J_{ij}, \forall \{i, j\}) \\ \eta &= (P_{ii}, \forall i) \cup (P_{ij}, \forall \{i, j\}) \end{aligned}$$

Information Projection

Given p, minimize information divergence

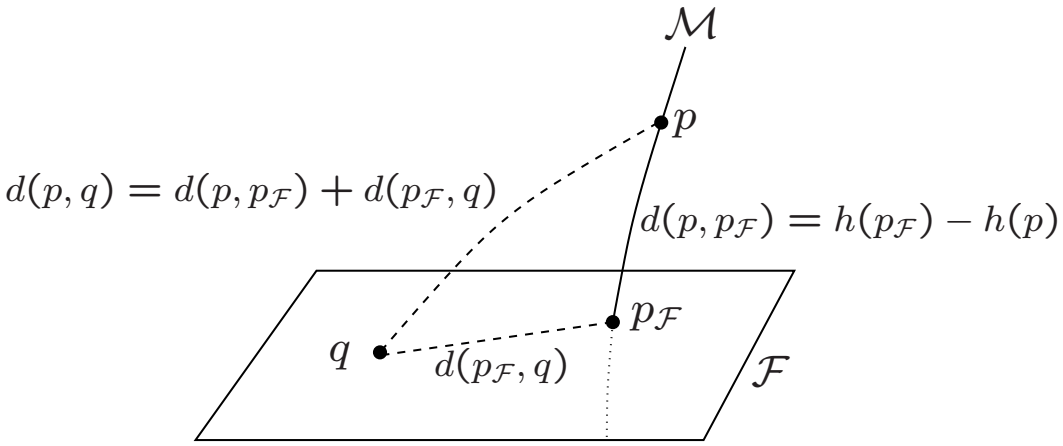
$$d(p,q) \triangleq \mathbb{E}_p\{\log \frac{p}{q}\}$$

over all $q \in \mathcal{F}$, an exponential family of Markov models on $\mathcal{G}$. Projection satisfies *Pythagorean property*.

Dual Problem: Maximize entropy

$$h(r) \triangleq -\mathbb{E}_r\{\log r\}$$

over all $r \in \mathcal{M}$ that satisfy linear moment constraints $\mathbb{E}_r\{\phi(x)\} = \eta$.



Maximum Entropy Relaxation

We pose graphical model selection as a convex optimization problem:

$$\begin{array}{ll} \max & h(\eta) \\ \text{s.t.} & d_E(\eta, \eta^*) \leq \delta_E, \; \forall E \in \mathcal{H} \end{array}$$

Maximize entropy subject to constraint that, for each subset $E \in \mathcal{H}$, the marginal probability distribution is close to that specified by η^*.

For example, we might impose all lower-order constraints $\mathcal{H} = \{E \subset V, |E| < k\}$, with tolerances

$$\delta_E = \gamma \times \begin{cases} |E| + \binom{|E|}{2}, & \text{Gaussian} \\ 2^{|E|}, & \text{Boltzmann} \end{cases}$$

Model Thinning Effect The MER solution $\widehat{\eta}$ is Markov with respect to $\widehat{\mathcal{G}}$ defined by active constraints. In other words, $\widehat{\theta}$ is sparse with respect to $\widehat{G}$.

Primal-Dual Interior Point Method*

Newton's method to solve for $\eta, \lambda > 0$ satisfying modified ($\epsilon > 0$) KKT conditions:

$$\begin{aligned} -\nabla h(\eta) + \sum_E \lambda_E \nabla d_E(\eta, \eta^*) &= 0 \\ \sum_E \lambda_E (\delta_E - d_E) &= \epsilon |\mathcal{H}| \end{aligned}$$

Eliminate $\Delta\lambda$, solve $H\Delta\eta = r$ where:

$$\begin{aligned} H &= G^* + \sum_E \lambda_E \left(G_E^* + \frac{(\theta_E^* - \theta_E)(\theta_E^* - \theta_E)^T}{\delta_E - d_E} \right) \\ r &= -\theta + \epsilon \sum_E \frac{\lambda_E}{\delta_E - d_E} (\theta_E^* - \theta_E) \end{aligned}$$

Back-substitution to compute $\Delta\lambda$.

Step-size chosen by back-tracking line search.

Decrease $\epsilon \to 0$ at each step.

*See pp.609–613, Boyd and Vandenbergh, Convex Optimization, '04.

Fisher Information in Gaussian Model

The relation between $J(\theta)$ and $P(\eta)$ is matrix inversion. Let $F(X) = X^{-1}$. We use the fact that $dF = -X^{-1}dXX^{-1}$.

Computing $\Delta\theta = G^*(\eta)\Delta\eta$ is equivalent to

$$\Delta J = -P^{-1}\Delta P P^{-1} \tag{1}$$

Similarly, computing $\Delta\eta = G(\theta)\Delta\theta$ is equivalent to

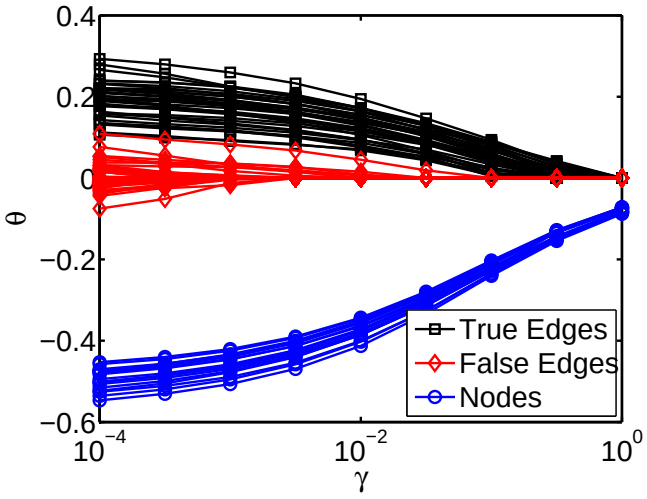
$$\Delta P = -J^{-1}\Delta J J^{-1} \tag{2}$$

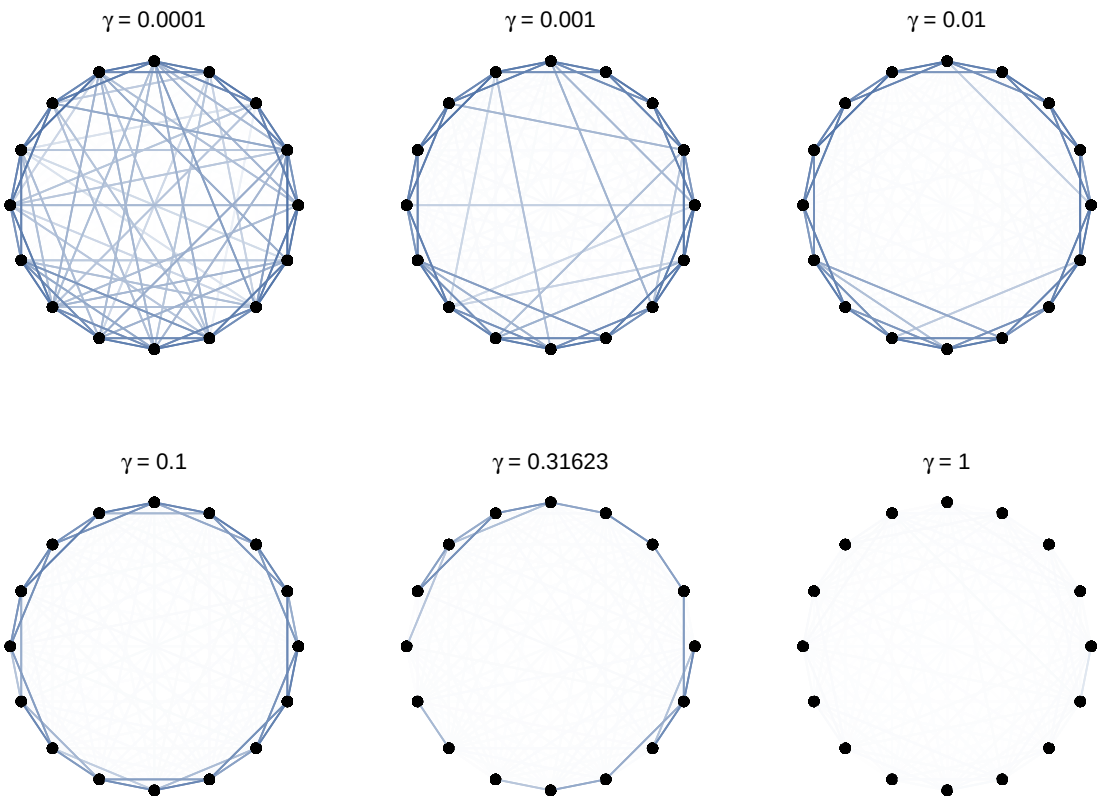
We can solve $H\Delta\eta = r$ using the *conjugate-gradient algorithm*, which requires multiplication by H. This can be efficiently implemented using (1).

Moreover, to accelerate convergence, we can use $G(\theta)$ as a preconditioner, implemented using (2).

Complexity per PCG iteration is $\mathcal{O}(n^3)$.

Gaussian Example





Fisher Information in Boltzmann Model

In the complete Boltzmann model, the Fisher information $G(\theta) = \frac{\partial \eta}{\partial \theta}$ is given by:

$$G(\theta) = M^T \text{Diag}(p_\theta) M$$

where $p_\theta = \exp(M\theta)$.

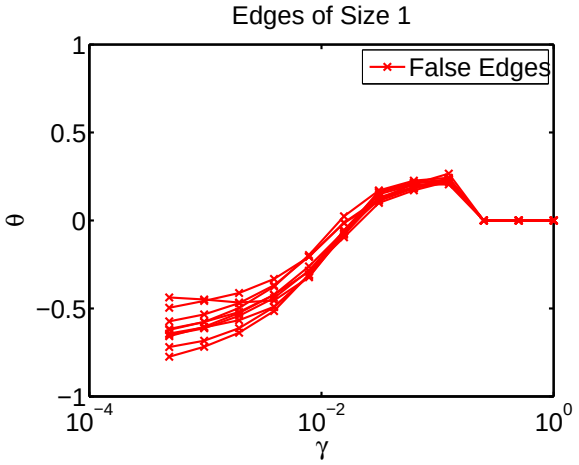
Its inverse is $G^*(\eta) = \frac{\partial \theta}{\partial \eta}$, given by:

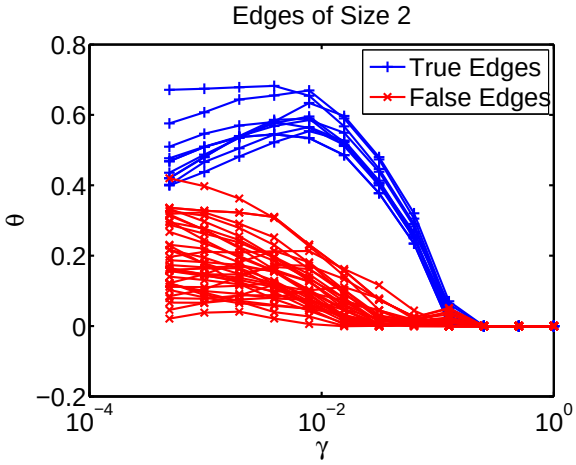
$$G^*(\eta) = M^{-1} \text{Diag}(1/p_\eta) M^{-T}$$

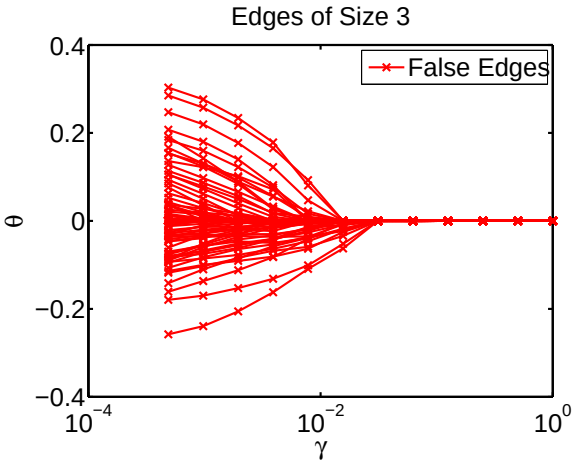
where $p_\eta = M^{-T}\eta$.

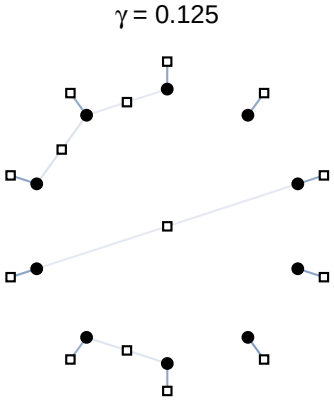
Hence, multiplication by either $G(\theta)$ or $G^*(\eta)$ requires $\mathcal{O}(n2^n)$ computation. We can use these computations to solve $H\Delta\eta = r$ using PCG, for $n < 16$.

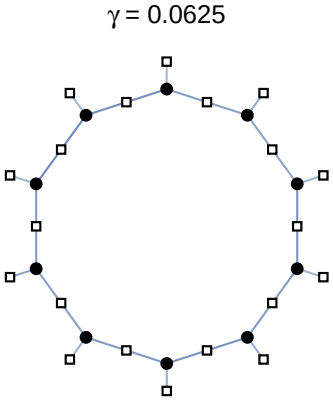
Boltzmann Example

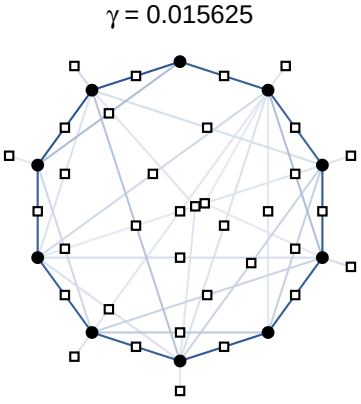












Fisher Information in Chordal Graphs

Let $\mathcal{G}$ be a *chordal* graph with maximal cliques $\mathcal{C}$, junction tree $T \subset \binom{\mathcal{C}}{2}$, and separators

$$\mathcal{S} = \{C_\alpha \cap C_\beta, (\alpha, \beta) \in T\}$$

Junction tree factorization:

$$p(x) = \prod_C p(x_C) / \prod_S p(x_S)$$

Entropy of a chordal model:

$$h(\eta) = \sum_C h_C(\eta) - \sum_S h_S(\eta).$$

Projection to chordal graph:

$$\Lambda^{-1}(\eta) = \sum_C \Lambda_C^{-1}(\eta) - \sum_S \Lambda_S^{-1}(\eta)$$

where $\Lambda : \theta \to \eta$.

Fisher information $G^*(\eta)$ of chordal model:

$$G^*(\eta) = \sum_C G_C^*(\eta) - \sum_S G_S^*(\eta).$$

So, $G^*(\eta)$ is *sparse*. Multiplication by $G^*(\eta)$ is $\mathcal{O}(nw^3)$ (Gaussian) or $\mathcal{O}(n2^w)$ (Boltzmann).

Incremental Graph Selection

Start with disconnected constraint graph $\mathcal{H}^{(0)}$, including only node constraints.

For $k = 0, 1, 2, \ldots$

1. Solve reduced MER with constraints $\mathcal{H}^{(k)}$, yielding relaxation $\eta^{(k)}$.

2. Evaluate $d_E(\eta^{(k)}, \eta^*)$ for all $E \in \mathcal{H} \setminus \mathcal{H}^{(k)}$.

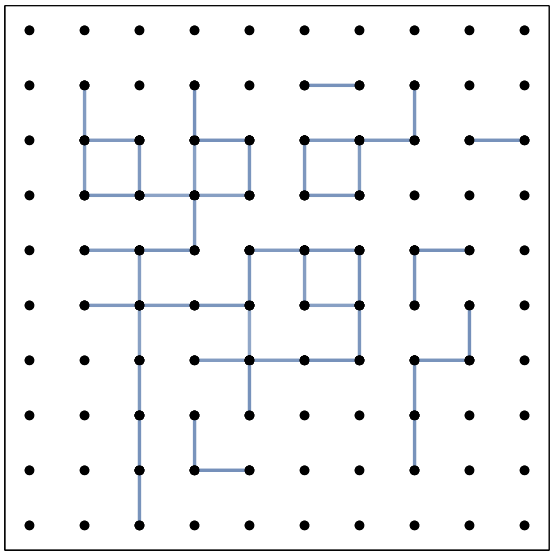
3. If $d_E \leq \delta_E$ for all E, STOP.

4. Else, build $\mathcal{H}^{(k+1)}$ by adding K largest violated constraints and continue.

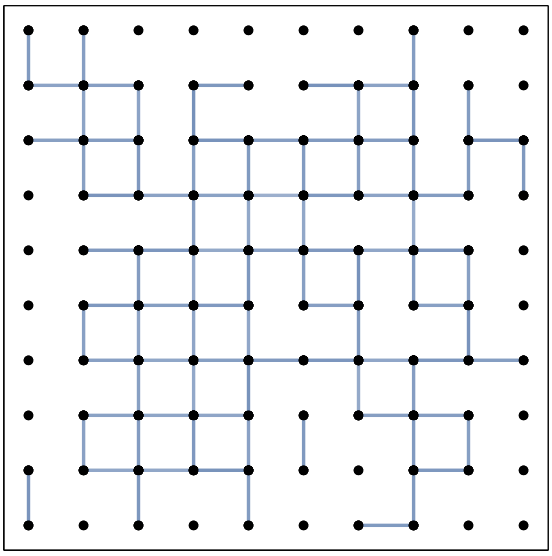
The solution to the full MER problem is obtained once the omitted constraints are all satisfied.

Gaussian Edge Selection

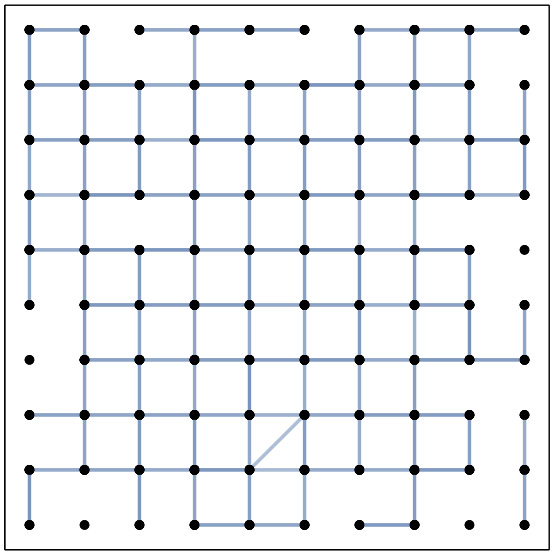
dim(G) = 150, dim(G_c) = 160



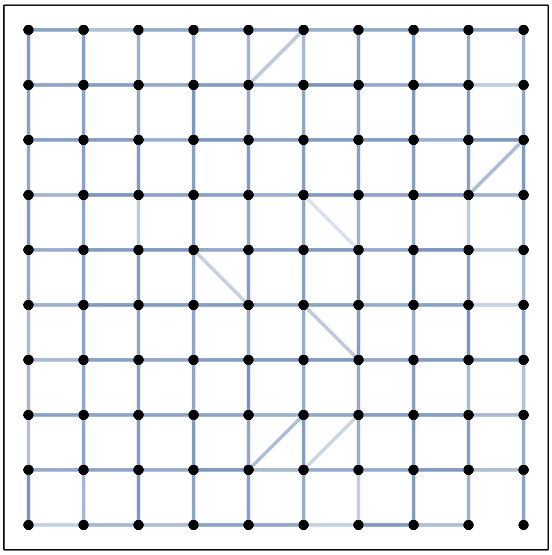
dim(G) = 200, dim(G_c) = 291



dim(G) = 250, dim(G_c) = 536



dim(G) = 287, dim(G_c) = 822



Summary

Relaxed maximum-entropy approach to learn graphical models, both structure and parameters are obtained through convex optimization.

Future work:

- Cross-validation to select γ
- Large-deviation analysis
- Bounds on generalization error
- Incomplete/inconsistent data
- Latent variables
- Tractable entropy approximations