

High-dimensional graphical model selection: Practical and information-theoretic limits

Martin Wainwright

Departments of Statistics, and EECS

UC Berkeley, California, USA

Based on joint works with:

Pradeep Ravikumar (UC Berkeley), and John Lafferty (CMU)

ℓ_1 -regularized graph selection

Prasad Santhanam (UC Berkeley)

Info. theory of graph selection

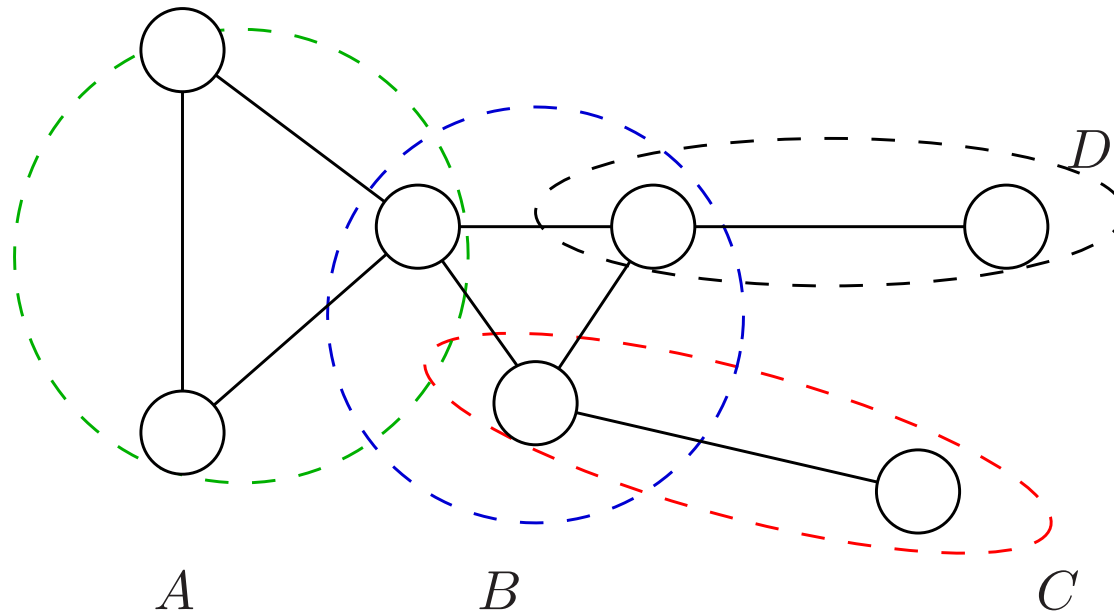
Supported by: NSF grants DMS-0605165 and CCF-0545862, and
a Sloan Foundation Fellowship

Introduction

- classical asymptotic theory of statistical inference:
 - number of observations $n \rightarrow +\infty$
 - model dimension p stays fixed
- not suitable for many modern applications:
 - { images, signals, systems, networks } frequently large ($p \approx 10^3 - 10^8$)...
 - interesting consequences: might have $p = \Theta(n)$ or even $p \gg n$
- curse of dimensionality: frequently impossible to obtain consistent procedures unless $p/n \rightarrow 0$
- can be saved by a lower *effective dimensionality*, due to some form of complexity constraint

Graphical models

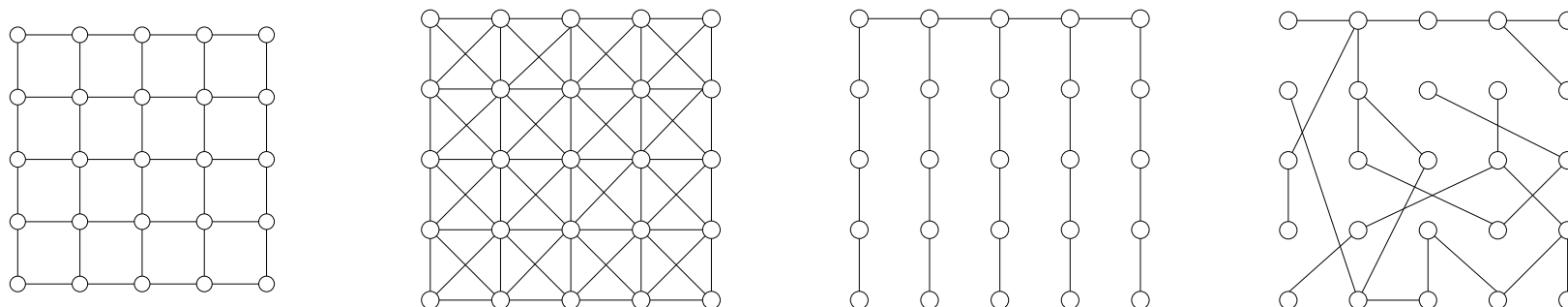
- probability \oplus graph theory = powerful formalism for capturing multivariate statistical dependencies
- Markov random field: random vector (X_1, \dots, X_p) with distribution factoring according to a graph $G = (V, E)$:



- (X_1, \dots, X_p) being Markov w.r.t G implies factorization:

$$\mathbb{P}(x_1, \dots, x_p) \propto \exp \left\{ \beta_A(x_A) + \beta_B(x_B) + \beta_C(x_C) + \beta_D(x_D) \right\}.$$

Problem of graphical model selection

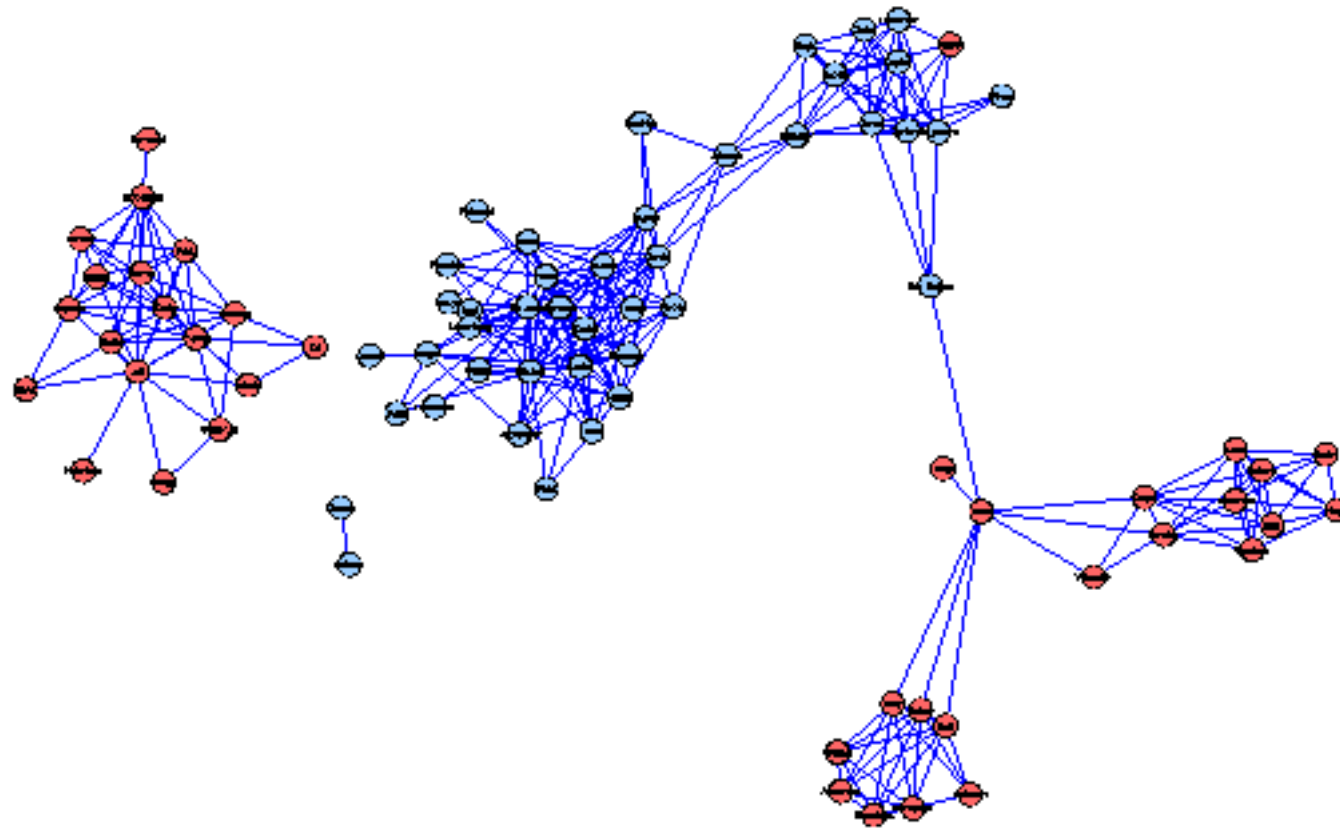


- consider p -dimensional random vector $X = (X_1, \dots, X_p)$:

$$\mathbb{P}(X_1, \dots, X_p; \beta) \propto \exp \left\{ \sum_{i \in V} \beta_i X_i + \sum_{(i,j) \in E} \beta_{ij} X_i X_j \right\}.$$

- given n independent and identically distributed (i.i.d.) samples of X , identify underlying graph $G = (V, E)$
- many applications: social network analysis, network tomography, computer vision, computational biology.....
- lower effective dimensionality:
 - # edges $k \ll \binom{p}{2}$ edges
 - Node degrees d slowly growing

Illustration: Social network analysis of US senators



Graphical model fit to voting records of US senators (Bannerjee et al., 2008)

Standard maximum likelihood is intractable

- graphical model as an exponential family:

$$\mathbb{P}(x_1, \dots, x_p; \beta) = \frac{1}{Z(\beta)} \exp \left\{ \sum_{i \in V} \beta_i x_i + \sum_{(i,j) \in E} \beta_{ij} x_i x_j \right\}$$

- given i.i.d. $X^{(1)}, \dots, X^{(n)}$ samples, might consider methods based on likelihood $\frac{1}{n} \sum_{i=1}^n \log \mathbb{P}(X^{(i)}; \beta)$
- exact likelihood involves partition function $Z(\beta)$: computationally intractable to compute in general
- possible solutions:
 - MCMC methods (e.g., Jerrum & Sinclair, 1993)
 - stochastic approximation (e.g., Potamianos & Goutsias, 1997)
 - approximation by “thinned” graphical models (Johnson et al., 2007)
 - variational approximations (WaiJor03, Lee et al., 2006)

Markov property and neighborhood structure

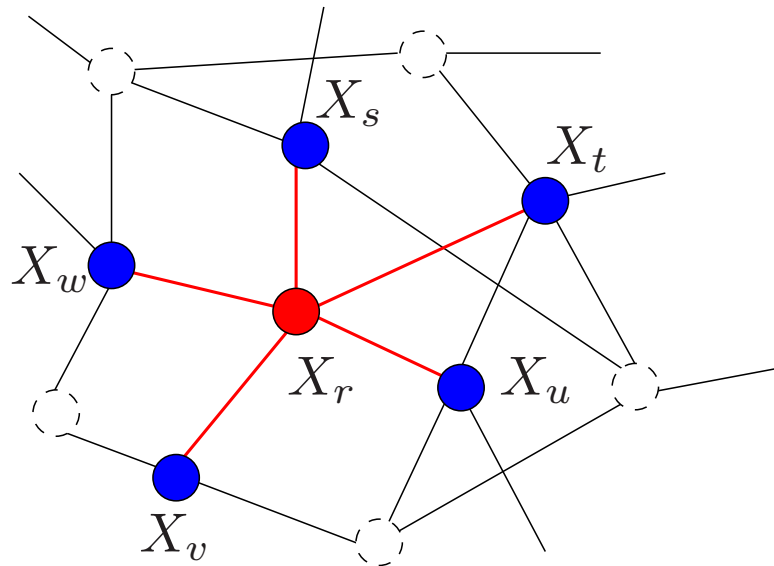
- Markov properties encode neighborhood structure:

$$\underbrace{(X_r \mid X_{V \setminus r})}_{\text{Condition on full graph}} \stackrel{d}{=} \underbrace{(X_r \mid X_{N(r)})}_{\text{Condition on Markov blanket}}$$

Condition on full graph

Condition on Markov blanket

$$N(r) = \{s, t, u, v, w\}$$



- basis of pseudolikelihood method (Besag, 1974, 1975, 1977)
- used for Gaussian model selection (Meinshausen & Buhlmann, 2006)

Method and notation

Observation: Recovering graph G equivalent to recovering neighborhood set $N(r)$ for all $r \in V$

Method: Given n i.i.d. samples $\{X^{(1)}, \dots, X^{(n)}\}$, perform logistic regression of each node X_s on $X_{\setminus r} := \{X_t, t \neq r\}$ to estimate neighborhood structure $\hat{N}(r)$.

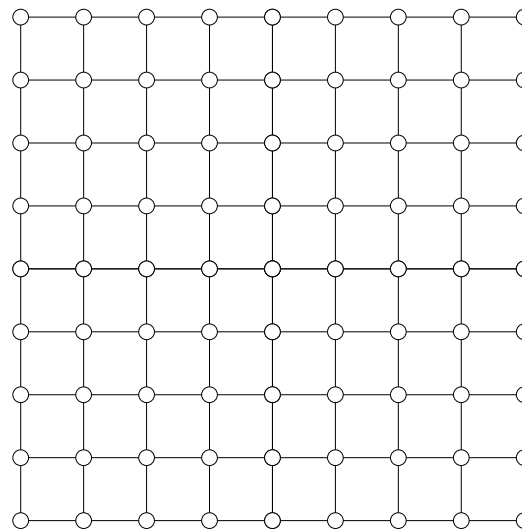
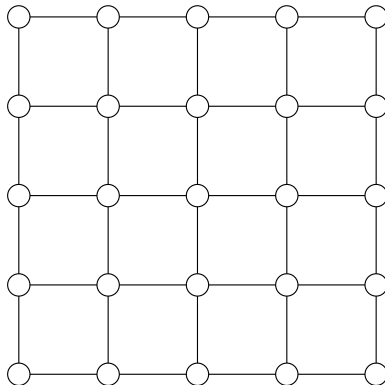
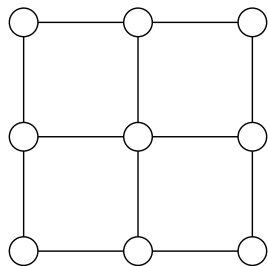
1. For each node $r \in V$, perform ℓ_1 regularized logistic regression of X_r on the remaining variables $X_{\setminus r}$:

$$\hat{\beta}[r] := \arg \min_{\beta \in \mathbb{R}^{p-1}} \left\{ \underbrace{\frac{1}{n} \sum_{i=1}^n f(\beta; X_{\setminus r}^{(i)})}_{\text{logistic likelihood}} + \underbrace{\rho_n \|\beta\|_1}_{\text{regularization}} \right\}$$

2. Estimate the local neighborhood $\hat{N}(r)$ as the support (non-negative entries) of the regression vector $\hat{\beta}[r]$.
3. Combine the neighborhood estimates in a consistent manner (AND, or OR rule).

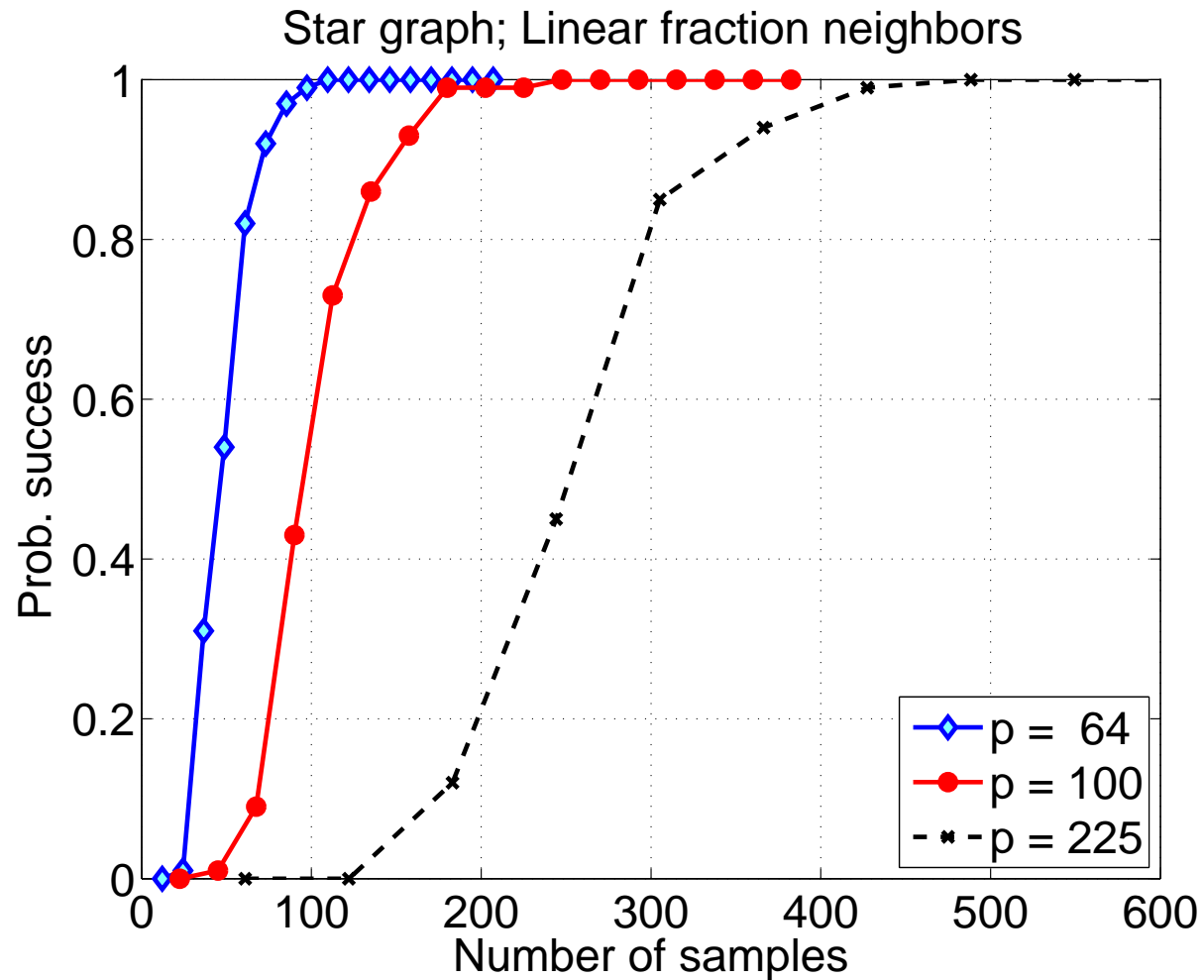
High-dimensional analysis

- classical analysis: dimension p fixed, sample size $n \rightarrow +\infty$
- high-dimensional analysis: allow both dimension p , sample size n , and maximum degree d to increase at arbitrary rates



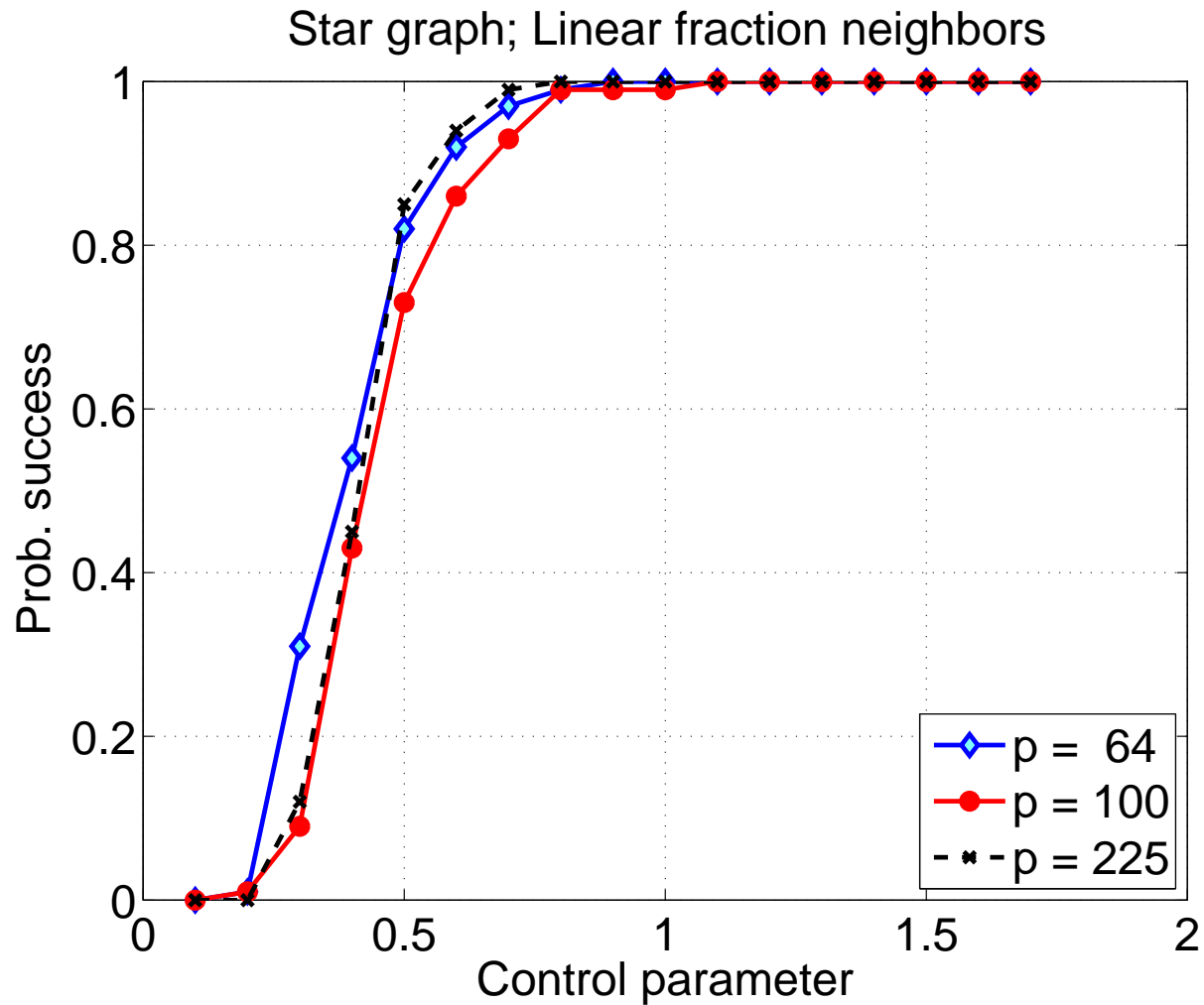
- take n i.i.d. samples from MRF defined by $G_{p,d}$
- study probability of success as a function of three parameters:
$$\text{Success}(n, p, d) = \mathbb{P}[\text{Method recovers graph } G_{p,d} \text{ from } n \text{ samples}]$$
- theory is non-asymptotic: explicit probabilities for finite (n, p, d)

Empirical behavior: Unrescaled plots



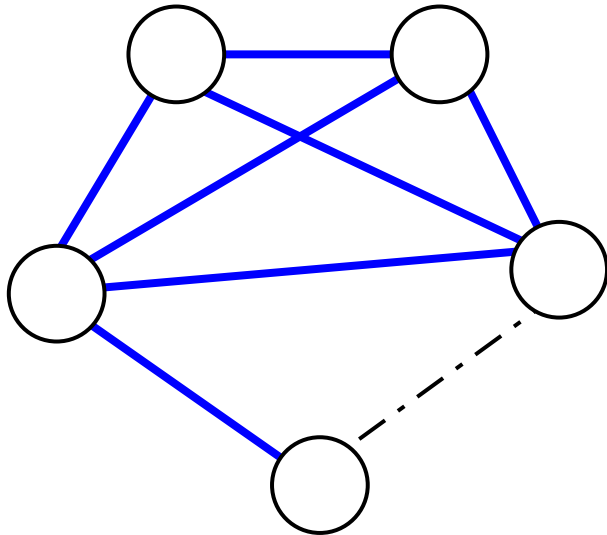
- star-shaped graphs (central node with $d = \alpha p$ neighbors)
- range of problem sizes: study probability of correct graph recovery

Empirical behavior: Appropriately rescaled

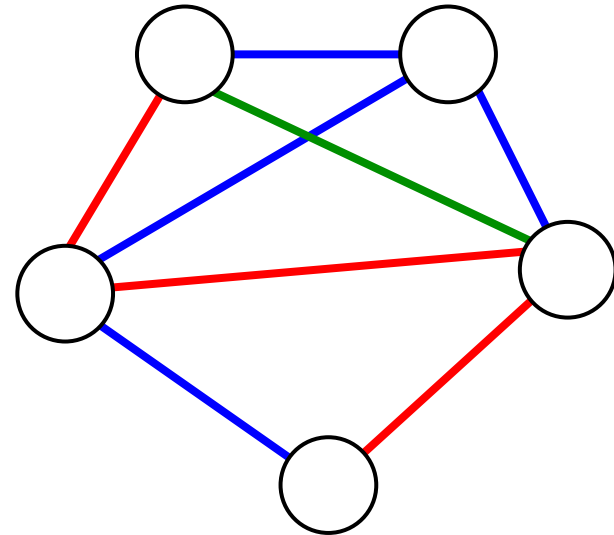


Plots of success probability versus suitable control parameter $C(n, p, d)$.

Some challenges in distinguishing graphs



Guilt by association



Hidden interactions

Conditions on Fisher information matrix $Q^* = \mathbb{E}[\nabla^2 f(\beta^*)]$

A1. Bounded eigenspectra: $\lambda(Q_{SS}^*) \in [C_{min}, C_{max}]$.

A2. Mutual incoherence There exists an $\nu \in (0, 1]$ such that

$$\|Q_{S^c S}^* (Q_{SS}^*)^{-1}\|_{\infty, \infty} \leq 1 - \nu.$$

where $\|A\|_{\infty, \infty} := \max_i \sum_j |A_{ij}|$.

Sufficient conditions for consistent model selection

- graph sequences $G_{p,d} = (V, E)$ with p vertices, and maximum degree d .
- drawn n i.i.d, samples, and analyze prob. success indexed by (n, p, d)

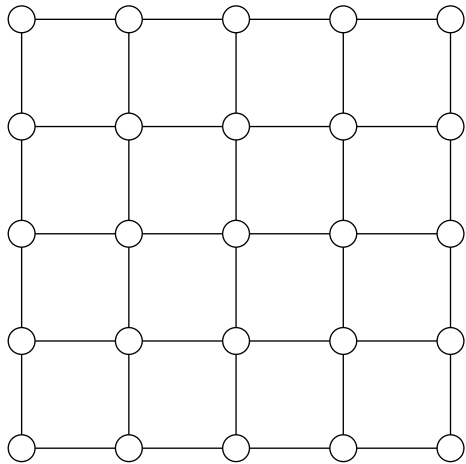
Theorem: For a rescaled sample size (RavWaiLaf06, RavWaiLaf08)

$$\theta_{\log}(n, p, d) := \frac{n}{d^3 \log p} > \theta_{\text{crit}}^*$$

and regularization parameter $\rho_n \geq c_1 \tau \sqrt{\frac{\log p}{n}}$, then with probability greater than $1 - 2 \exp(-c_2(\tau - 2) \log p) \rightarrow 1$:

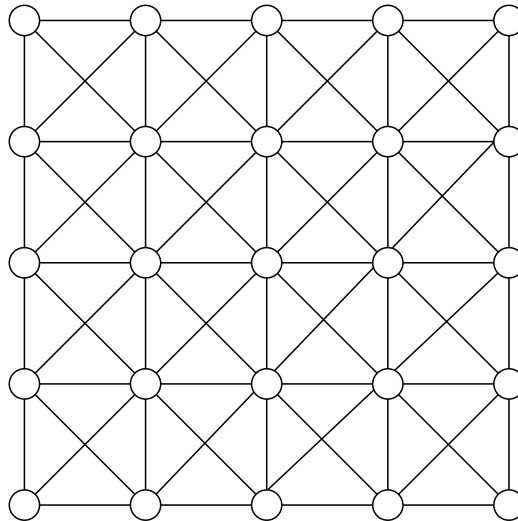
- For each node $r \in V$, the ℓ_1 -regularized logistic convex program has a unique solution. (Non-trivial since $p \gg n \implies$ not strictly convex).
- The estimated sign neighborhood $\widehat{\mathbb{N}}_{\pm}(r)$ correctly excludes all edges *not* in the true neighborhood.
- For $\beta_{\min} \geq c_3 \tau \sqrt{\frac{d^2 \log p}{n}}$, the method selects the correct signed neighborhood.

Example graphs



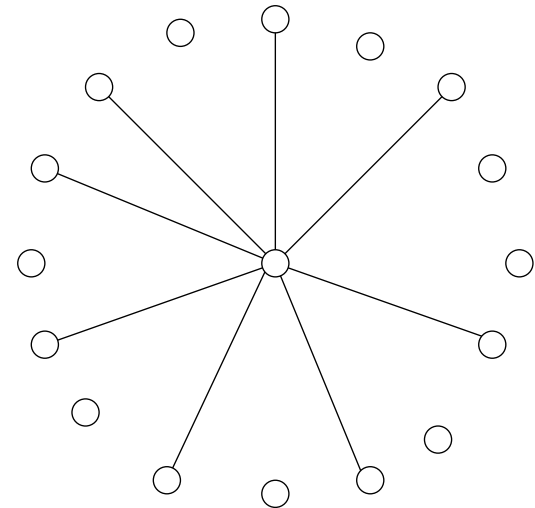
(a) 4-grid

$$d = 4$$



(b) 8-grid

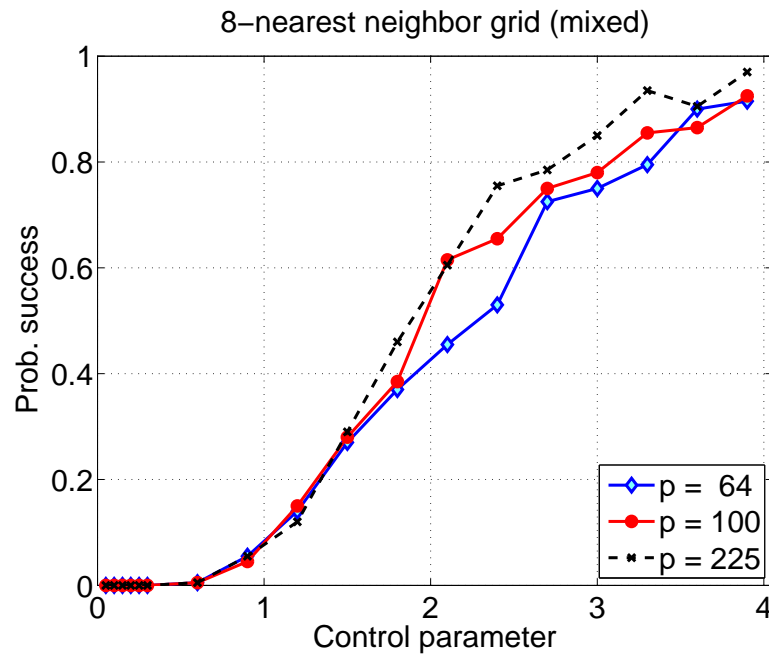
$$d = 8$$



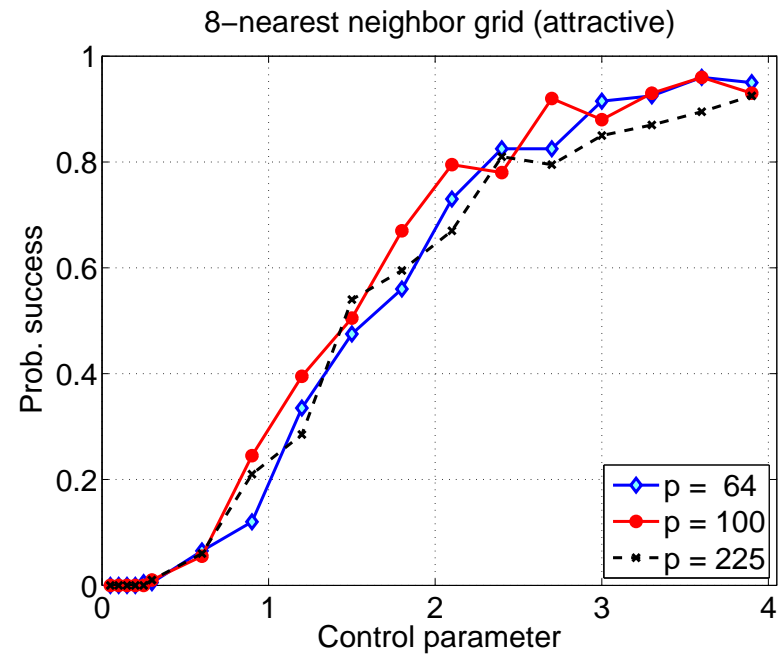
(c) Star

$$d \in \{\mathcal{O}(\log p), \alpha p\}$$

Results for 8-grid graphs



(a) Mixed interactions

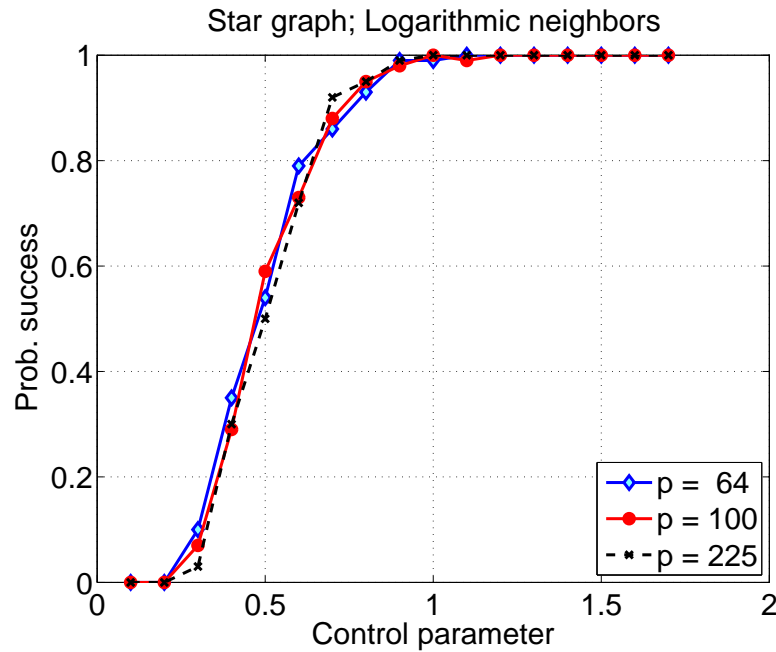


(b) Attractive interactions

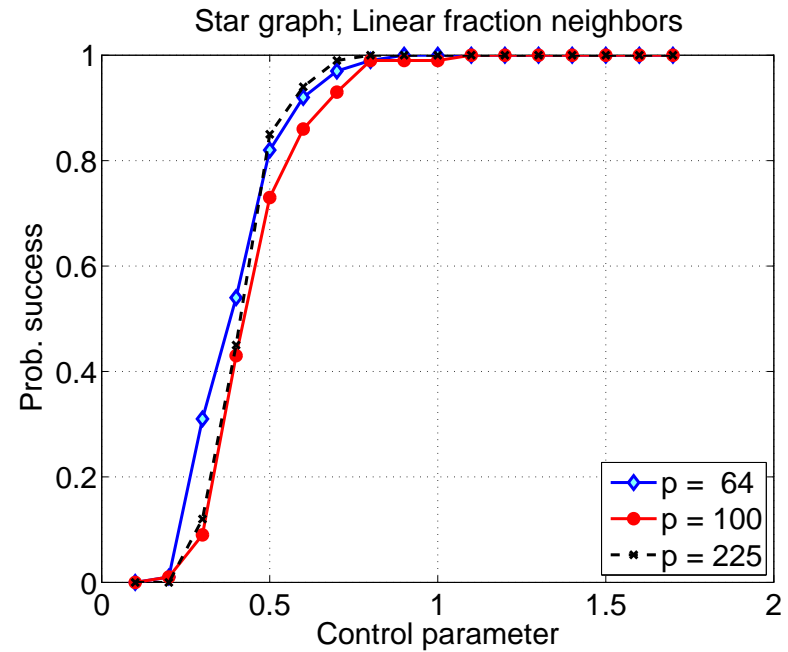
- Vertical axis: success probability $\mathbb{P}[\hat{N}(s) = N(s)]$
- Horizontal axis: control parameter

$$C(n, p, d) = \frac{n}{d \log(p - d)}.$$

Results for star graphs



(a) Degree $d = \mathcal{O}(\log p)$



(b) Degree $d = \alpha p$

- Vertical axis: success probability $\mathbb{P}[\hat{N}(s) = N(s)]$
- Horizontal axis: control parameter

$$C(n, p, d) = \frac{n}{d \log(p - d)}.$$

Proof sketch: Primal-dual certificate

- construct *candidate* primal-dual pair $(\tilde{\beta}, \tilde{z}) \in \mathbb{R}^{p-1} \times \mathbb{R}^{p-1}$.
- proof technique—not a practical algorithm!

(A) First, we solve the restricted program

$$\tilde{\beta} = \arg \min_{\beta \in \mathbb{R}^{p-1}, \beta_{S^c} = \mathbf{0}} \left\{ \frac{1}{n} \sum_{i=1}^n f(\beta; X_{\setminus r}^{(i)}) + \rho_n \|\beta\|_1 \right\},$$

thereby obtaining candidate solution $\tilde{\beta} = (\tilde{\beta}_S, \mathbf{0}_{S^c})$.

- (B) We choose $\tilde{z}_S \in \mathbb{R}^{|S|}$ as an element of the subdifferential $\partial \|\tilde{\beta}_S\|_1$.
- (C) Using optimality conditions from original convex program, verify *strict dual feasibility*

$$|\tilde{z}_j| < 1 \quad \text{for all } j \in S^c.$$

Lemma: Full convex program recovers neighborhood \iff primal-dual witness succeeds.

Intermediate step: Analysis of population version

- intermediate step: impose eigenvalue bounds and incoherence directly on sample Fisher matrix $\tilde{Q} = \frac{1}{n} \sum_{i=1}^n f(\beta^*; X^{(i)})$.

Proposition: If conditions imposed directly on sample Fisher matrix, then $n = \Omega(d^2 \log(p - d))$ samples are sufficient.

- construct primal-dual solution by first “setting” $\hat{\beta}_S^c = 0$ and $\hat{z}_S = \text{sign}(\beta_S^*)$, and applying Taylor expansion to gradient equation:

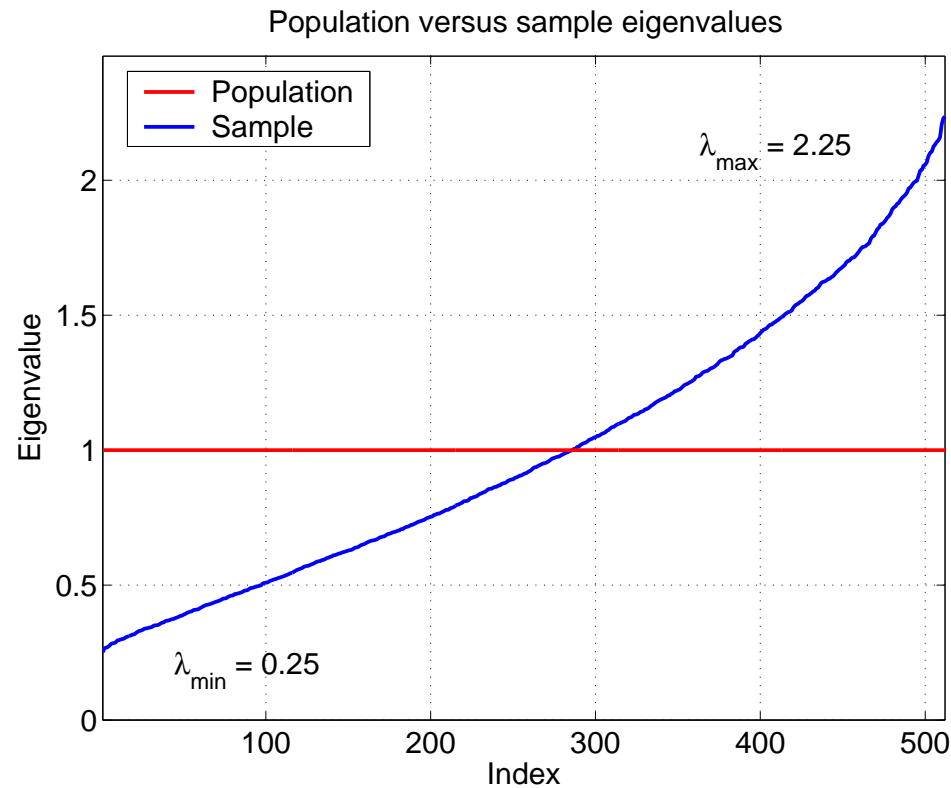
$$\begin{bmatrix} Q_{SS}^* & Q_{SS^c}^* \\ Q_{S^cS}^* & Q_{S^cS^c}^* \end{bmatrix} \begin{bmatrix} \hat{\beta}_S - \beta_S^* \\ 0 \end{bmatrix} + \text{”lots of noise”} = -\rho_n \begin{bmatrix} \text{sign}(\beta_S^*) \\ \hat{z}_S \end{bmatrix}$$

- solve for $\hat{\beta}_S$ and \hat{z}_S^c :

$$\hat{z}_S^c = Q_{S^cS}^* (Q_{SS}^*)^{-1} (\rho_n \text{sign}(\beta_S^*) + \text{”noise”}) + \text{”more noise”}.$$

Challenges with high-dimensional random matrices

- say $k = n$ for some $\alpha \in (0, 1)$, and consider $\hat{\Lambda} = \frac{1}{n}X^T X$, with $X_{ij} \sim N(0, 1)$



- sample eigenspectrum $\hat{\Lambda}$: $[\lambda_{\min}(\hat{\Lambda}), \lambda_{\max}(\hat{\Lambda})] \xrightarrow{p} (1 \pm \sqrt{\alpha})^2$
(Marcenko & Pastur, 1967; Geman, 1980)

Concentration bounds on random matrix norms

- say population matrices are *well-behaved* (i.e., satisfy eigenvalue bounds and incoherence assumptions)
- consider sample versions:

$$\tilde{Q} := \frac{1}{n} \sum_{i=1}^n \nabla^2 f(\beta^*; X^{(i)})$$

- for $n = \Omega(d^3 \log p)$, sample matrices are also well-behaved:
 - bounded eigenvalues:

$$\mathbb{P}[\lambda_{\min}(\tilde{Q}_{SS}) \leq C_{\min} - \delta] \leq 2 \exp\left(-A \frac{\delta^2 n}{d^2} + B \log(d)\right).$$

- mutual incoherence:

$$\mathbb{P}\left[\|\tilde{Q}_{S^c S}(\tilde{Q}_{SS})^{-1}\|_{\infty, \infty} \geq 1 - \frac{\nu}{2}\right] \leq \exp\left(-L \frac{n}{d^3} + \log(p)\right).$$

Information-theoretic limits on graph selection

(Santhanam & Wainwright, 2008)

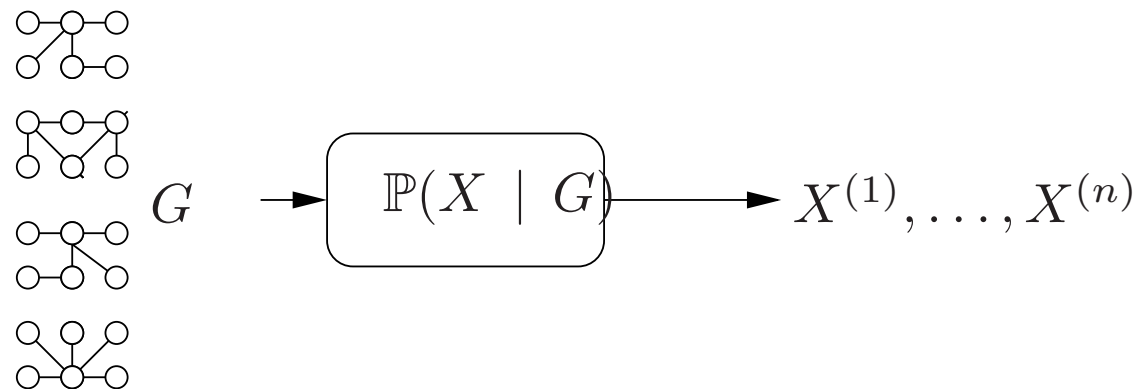
- thus far: have exhibited a particular polynomial-time method can recover structure if

$$n > \Omega(d^3 \log(p - d))$$

- but....is this a “good” result?
- are there polynomial-time methods that can do better?
- information theory can answer the question: is there an exponential-time method that can do better?

Graph selection as channel coding

- graphical model selection is an *unorthodox* channel coding problem:
- nature is sending one of $M(p, d)$ graphs of size p vertices and maximum degree d



- decoding problem: use observations $\{X^{(1)}, \dots, X^{(n)}\}$ to correctly distinguish the “codeword”
- channel capacity for graph decoding: balance between
 - log number of models: $\log |M(p, d)| = \Theta \left(pd \log \frac{p}{d} \right)$.
 - relative distinguishability of different models

Necessary conditions for graph recovery

Theorem: With the minimum edge weight $\beta_{min} = \min_{(s,t) \in E} |\beta_{st}^*|$, any algorithm requires at least

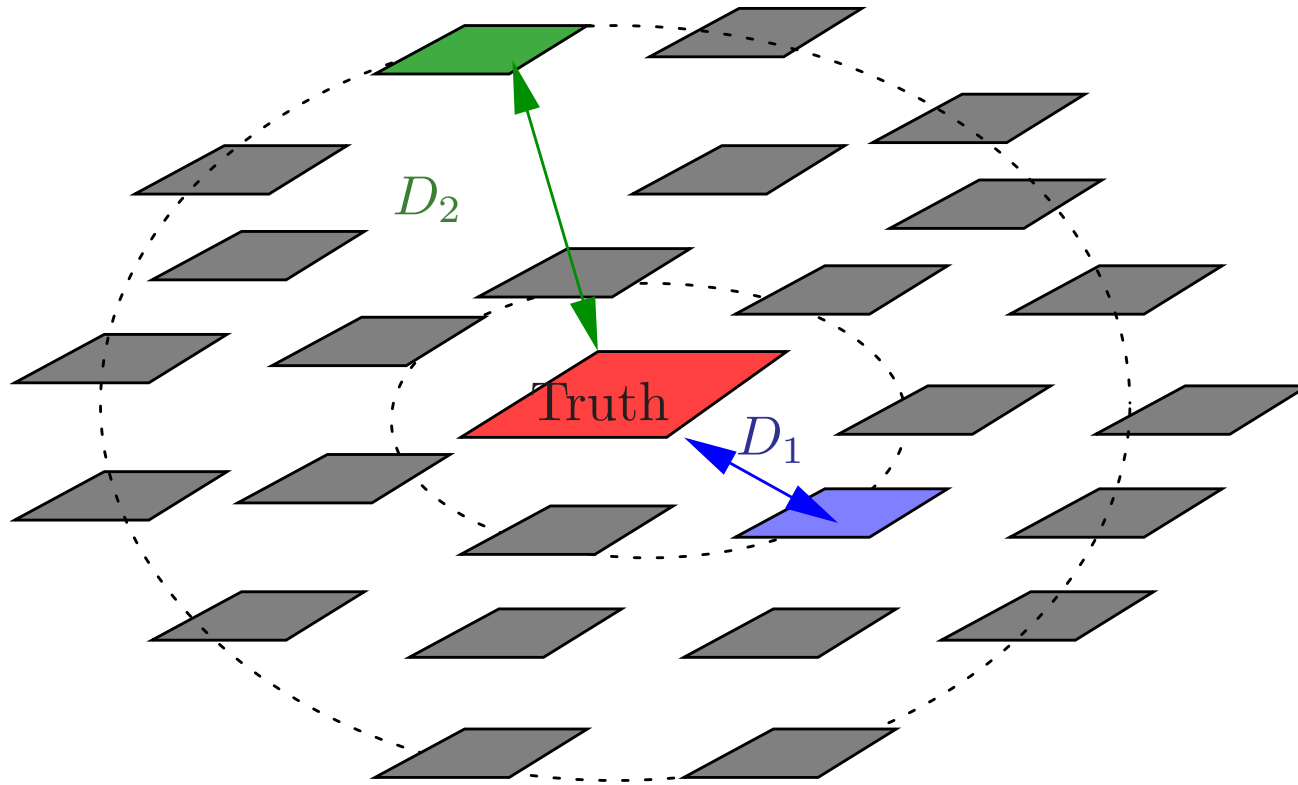
$$n > \max \left\{ c_1 d \log(p), \quad c_2 \frac{\log\left(\binom{p}{2} - pd/2\right)}{\beta_{min}^2} \right\}$$

observations to succeed.

(PraWai08)

- for constant degree graphs, ℓ_1 -regularized log. regression is order-optimal
- for d tending to infinity, gap between theory and practice
 - theoretical guarantees: $n > C d^3 \log p$
 - empirically: $n = \Theta(d \log p)$ appears to suffice, matching info.-theoretic limits

Geometric intuition



Error probability controlled by two competing quantities:

Model type	# Models	Distance scaling
Near-by	$\binom{p}{2} - pd/2$	$1/\beta_{min}^2$
Far-away	$\exp\left(pd \log \frac{p}{d}\right)$	$\Theta(d)$

Proof sketch: Main ideas

- consider two ensembles of models:
 - full model class $M_1(p, d)$: all degree d graphs
 - a smaller but “hard” sub-ensemble $M_2(p, d)$
- for both ensembles, by Fano’s inequality:

$$\mathbb{P}[\text{error}] \geq 1 - \frac{I(X; G)}{\log(M - 1)} - o(1),$$

where $I(X; G)$ is mutual info. between G and observations X

- by construction, have lower bounds:

$$\begin{aligned} \log(M_1 - 1) &\geq \Omega(pd \log \frac{p}{d}) \\ \log(M_2 - 1) &\geq \Omega\left(\log \left[\binom{p}{2} - pd/2 \right]\right) \end{aligned}$$

- require good upper bounds on mutual information

$$I(X^{(1)}, \dots, X^{(n)}; G) = H(X) - H(X | G).$$

Summary and open questions

- have analyzed graphical model selection:
 - high-dimensional analysis: sample size n , graph size p and maximum degree d allowed to diverge
 - ℓ_1 -regularized log. regression (polynomial-time) succeeds

$$n = \Omega(d^3 \log(p))$$

- even oracle decoders (exponential complexity) require

$$n = \Omega(d \log(p))$$

- more broadly:
 - optimal trade-offs between statistical and computational efficiency in inference? limits of bounded computation?
 - other types of “complexity”-theoretic regularization: efficient algorithms and high-dimensional guarantees?