11. Learning graphical models

- Maximum likelihood
- Parameter learning
- Structural learning
- Learning partially observed graphical models

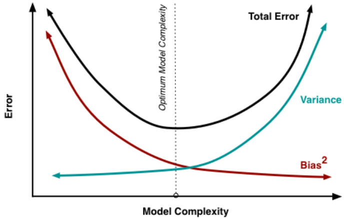
- statistical inference addresses how to overcome computational challenges given a graphical model
- learning addresses how to overcome both computational and statistical challenges to learn graphical models from data
- the ultimate goal of learning could be (a) to uncover structure; (b) solve particular problem such as estimation, classification, or decision; or (c) learn the joint distribution in order to make predictions
- and the right learning algorithm depends on the particular goal

- choosing the right model to learn is crucial, in the fundamental tradeoff between overeating and underfitting, often called **bias-variance tradeoff**
- consider the task of learning a function over $x \in \mathbb{R}^d$ where

$$Y = f(x) + Z$$

  with i.i.d noise Z
- the goal is to predict y from x by learning the function f, from "training data" $\{(x_1, y_1), \ldots, (x_n, y_n)\}$
- bias-variance tradeoff for $\text{Err}(x)$ for a given x is

$$\mathbb{E}[(Y - \hat{f}_{Y_1^n}(x))^2] = \underbrace{(\mathbb{E}[\hat{f}_{Y_1^n}(x)] - f(x))^2}_{\text{Bias}} + \underbrace{\mathbb{E}[(\hat{f}_{Y_1^n}(x) - \mathbb{E}[\hat{f}_{Y_1^n}(x)])^2]}_{\text{Variance}} + \underbrace{\mathbb{E}[(Y - f(x))^2]}_{\text{Irreducible error}}$$



- when the model is too complex, the model **overfits** the training data and does not generalize to new data, which is referred to as high "variance"
- when the model is too simple, the model **undercuts** the data and has high "bias"
- common strategy to combat this is to (a) constrain the class of models or (b) penalize the model complexity
- graphical models provide a hierarchy of model complexity to avoid overfitting/underfitting

- in increasing level of difficulty, learning tasks can be classified as follows
  - ⋆ we know the graph, or we impose a particular graph structure, and we need to learn the parameters
  - ⋆ we observe all the variables, and we need to infer the structure of the graph and the parameters
  - ⋆ the variables are partially observation, and we need to infer the structure

Maximum likelihood

$$\mu(x) \;=\; \frac{1}{Z}\prod_{a\in F}\psi_a(x_{\partial a})$$

- we want to learn the graphical model consisting of p nodes from n i.i.d. samples
$$x^{(1)},\ldots,x^{(n)} \in \mathcal{X}^p$$
- it is more convenient to parametrize the compatibility functions as $e^{\theta_a(x_{\partial a})} = \psi_a(x_{\partial a})$, and one approach is to use **maximum likelihood estimator**
$$\begin{aligned} L_n(\theta, G, \{x^{(\ell)}\}_{\ell\in[n]}) &= \frac{1}{n}\log\Big(\prod_{\ell=1}^{n}\mu(x^{(\ell)})\Big) \\ &= \frac{1}{n}\sum_{\ell\in[n]}\sum_{a\in F}\theta_a(x^{(\ell)}_{\partial a}) - \log Z(\theta, G) \end{aligned}$$
this is a strictly concave function over $\theta = [\theta_a(x_{\partial a})] \in \mathbb{R}^{\sum_{a\in F}|\mathcal{X}|^{|\partial a|}}$,
- however, evaluating this function requires computing the log partition function, which is in general #p-hard, even if G is given
- if inference is efficient then learning is efficient

In case of Ising models,

$$\mu(x) = \frac{1}{Z} \prod_{(i,j)\in E} e^{\theta_{ij} x_i x_j} \prod_{i \in V} e^{\theta_i x_i}$$

- the log likelihood function is

$$L_n(\theta, G, \{x^{(\ell)}\}_{\ell\in[n]}) = \underbrace{\sum_{(i,j)\in E} \theta_{ij} \Big\{ \frac{1}{n} \sum_{\ell=1}^{n} x_i^{(\ell)} x_j^{(\ell)} \Big\} + \sum_{i\in V} \theta_i \Big\{ \frac{1}{n} \sum_{\ell=1}^{n} x_i^{(\ell)} \Big\}}_{\triangleq \langle \widehat{M}, \theta \rangle} - \underbrace{\log Z(\theta, G)}_{\triangleq \phi(\theta)}$$

- $\widehat{M} = [\widehat{M}_{ij}, \ldots, \widehat{M}_i, \ldots] \in \mathbb{R}^{|E|+p}$ where $\widehat{M}_{ij} = \frac{1}{n} \sum_\ell x_i^{(\ell)} x_j^{(\ell)}$ and $\widehat{M}_i = \frac{1}{n} \sum_\ell x_i^{(\ell)}$
- this is a strictly concave function over $\theta = [\theta_{ij}, \ldots, \theta_i, \ldots] \in \mathbb{R}^{|E|+p}$

- this maximum likelihood estimator is **consistent**, i.e. let $\widehat{\theta} = \arg\max_\theta L_n(\theta, G, \{x^{(\ell)}\})$, and let θ^* denote the true parameter, then

$$\lim_{n\to\infty} \widehat{\theta} \;=\; \theta^*$$

- **Proof.** by optimality of $\widehat{\theta}$,

$$\langle \widehat{M}, \widehat{\theta} \rangle - \phi(\widehat{\theta}) \;\geq\; \langle \widehat{M}, \theta^* \rangle - \phi(\theta^*)$$

by strong convexity of $\phi(\cdot)$, there exists $\sigma > 0$ such that

$$\phi(\widehat{\theta}) \;\geq\; \phi(\theta^*) + \langle \mathsf{M}, (\widehat{\theta} - \theta^*) \rangle + \frac{\sigma}{2} \|\widehat{\theta} - \theta^*\|^2$$

which follows form the fact that $\nabla_\theta \phi(\theta^*) = \mathsf{M}$, where $\mathsf{M}_{ij} = \mathbb{E}_\mu[x_i x_j]$ and $\mathsf{M}_i = \mathbb{E}_\mu[x_i]$, and together we get

$$\langle \widehat{M}, (\widehat{\theta} - \theta^*) \rangle \;\geq\; \phi(\widehat{\theta}) - \phi(\theta^*) \;\geq\; \langle \mathsf{M}, (\widehat{\theta} - \theta^*) \rangle + \frac{\sigma}{2} \|\widehat{\theta} - \theta^*\|^2$$

it follows that

$$\|\widehat{M} - \mathsf{M}\| \, \|\widehat{\theta} - \theta^*\| \;\geq\; \langle (\widehat{M} - \mathsf{M}), (\widehat{\theta} - \theta^*) \rangle \;\geq\; \frac{\sigma}{2} \|\widehat{\theta} - \theta^*\|^2$$

- we have

$$\|\widehat{\theta} - \theta^*\| \;\leq\; \frac{2}{\sigma}\|\widehat{M} - \mathsf{M}\|$$

  which gives, with high probability,

$$\frac{1}{|E| + p}\|\widehat{\theta} - \theta^*\|^2 \;\leq\; \frac{4}{\sigma^2\, n}$$

  where $\sigma = \inf_\theta \sigma_{\min}(\nabla^2\phi(\theta))$
- computing the gradient of ϕ is straight forward:

$$\begin{aligned} \phi(\theta) &= \log\sum_x \prod_{(i,j)\in E} e^{\theta_{ij}x_ix_j} \prod_{i\in V} e^{\theta_i x_i} \\ \frac{\partial\phi(\theta)}{\partial\theta_{ij}} &= \frac{1}{Z(\theta, G)}\sum_x \prod_{(i,j)\in E} e^{\theta_{ij}x_ix_j} \prod_{i\in V} e^{\theta_i x_i} x_i x_j \\ &= \mathbb{E}_\mu[x_ix_j] \;=\; \mathsf{M}_{ij} \end{aligned}$$

- again, this is computationally challenging, in general, but if such inference is efficient, then learning can also be efficient

- recall that

$$\mu_\theta(x) \quad = \quad \frac{1}{Z} \prod_{a \in F} \psi_a(x_{\partial a})$$

Let

$$c(x) = \sum_{i=1}^{n} \mathbb{I}(x^{(i)} = x) \text{ , } and$$

$$c(x_{\partial a}) = \sum_{i=1}^{n} \mathbb{I}(x^{(i)}_{\partial a} = x_{\partial a})$$

then the log-likelihood is

$$\begin{aligned} L(\psi, x^{(1)}, \ldots, x^{(n)}) \quad &= \quad \frac{1}{n} \sum_{i \in [n]} \sum_{a \in F} \log(\psi_a(x^{(i)}_{\partial a})) - \log Z \\ &= \quad \sum_{a \in F} \sum_{x_{\partial a}} \underbrace{\frac{c(x_{\partial a})}{n}}_{\triangleq \widehat{\mu}(x_{\partial a})} \log(\psi_a(x^{(i)}_{\partial a})) - \log Z \end{aligned}$$

- taking the derivative

$$\frac{\partial L(\psi, x^{(1)}, \ldots, x^{(n)})}{\partial \psi_a(x_{\partial a})} = \frac{\widehat{\mu}(x_{\partial a})}{\psi_a(x_{\partial a})} - \frac{\mu(x_{\partial a})}{\psi_a(x_{\partial a})}$$

- if the graph is a tree, then one can find the ML estimate using the above equation
- in general, it is computationally challenging, and one resorts to approximate solutions
- **iterative proportional fitting (IPF)** updates the estimates iteratively using the above fixed point equation:

$$\psi_a^{(t+1)}(x_{\partial a}) \leftarrow \psi_a^{(t)}(x_{\partial a}) \frac{\widehat{\mu}(x_{\partial a})}{\mu^{(t)}(x_{\partial a})}$$

Parameter learning

- suppose G is given, and want to estimate the compatibility functions $\{\psi_a(x_{\partial a})\}_{a \in F}$ from i.i.d. samples $x^{(1)}, \ldots, x^{(n)} \in \mathcal{X}^p$ drawn from
$$\mu(x) \;=\; \frac{1}{Z} \prod_{a \in F} \psi_a(x_{\partial a})$$
- **Example.** a single node with $x \in \{0,1\}$, and let
$$\mu(x) = \begin{cases} \theta & \text{for } x = 1 \\ 1-\theta & \text{for } x = 0 \end{cases}$$
consider an estimator $\widehat{\theta} = \frac{1}{n}\sum_{\ell=1}^{n} x^{(\ell)}$,
- **Claim.** For any $\varepsilon, \delta > 0$, let $n^*(\varepsilon,\delta)$ denote the minimum number of samples required to ensure that $\mathbb{P}(|\widehat{\theta} - \theta| > \varepsilon) \le \delta$, then
$$n^*(\varepsilon,\delta) \le \frac{1}{2\varepsilon^2} \log \frac{2}{\delta}$$
- this follows directly from Hoeffding's inequality:
$$\mathbb{P}\Big(|\frac{1}{n}\sum_{\ell=1}^{n} x^{(\ell)} - \theta| > \varepsilon\Big) \le 2\exp\{-2n\varepsilon^2\}$$
but it does not depend on the value of θ

- often we want ε to be in the same range as θ, and letting $\varepsilon = \alpha\theta$ gives

$$n^*(\alpha\theta, \delta) \le \frac{1}{2\alpha^2\theta^2} \log \frac{2}{\delta}$$

but applying Chernoff's bound:

$$\begin{aligned} \mathbb{P}\Big(\frac{1}{n}\sum_{\ell=1}^{n} x^{(\ell)} - \theta > \alpha\theta\Big) &\le \exp\{-D_{\rm KL}(\theta(1+\alpha)\|\theta)\, n\}\,, \text{ and} \\ \mathbb{P}\Big(\frac{1}{n}\sum_{\ell=1}^{n} x^{(\ell)} - \theta < -\alpha\theta\Big) &\le \exp\{-D_{\rm KL}(\theta(1-\alpha)\|\theta)\, n\} \end{aligned}$$

and using the fact that $D_{\rm KL}((1+\alpha)\theta\|\theta) \ge (1/4)\alpha^2\theta$, for $|\alpha| < 1/2$, this can be tightened to

$$n^*(\alpha\theta, \delta) \le \frac{4}{\alpha^2\theta} \log \frac{2}{\delta}$$

- **Example.** consider now a joint distribution $\mu(x)$ (satisfying $\mu(x) > 0$ for all x) over $x \in \{0,1\}^p$ with a large p, and an estimator $\widehat{\mu}(x) = \frac{1}{n}\sum_{\ell=1}^{n} \mathbb{I}(x^{(\ell)} = x)$
- it follows that the minimum number of samples required to ensure that $\mathbb{P}(\exists x \text{ such that } |\widehat{\mu}(x) - \mu(x)| > \alpha\mu(x)) \leq \delta$ is

  $$n^*(\alpha, \delta) \leq \frac{4}{\alpha^2 \min_x\{\mu(x)\}} \log \frac{2^{p+1}}{\delta}$$

  which follows from the Chernoff's bound with union bound over $x \in \{0,1\}^p$
- since $\min_x\{\mu(x)\} \leq 2^{-p}$, the sample complexity is exponential in the dimension p
- this is unavoidable in general, but when $\mu(\cdot)$ factorizes according to a graphical model with a sparse graph, then the sample complexity significantly decreases

- we want to learn the compatibility functions of a graphical model using 'local' data, but compatibility functions are not uniquely defined,

$$\mu(x) = \frac{1}{Z} \prod_{a \in F} \psi_a(x_{\partial a})$$

- we first need to define canonical form of compatibility functions
- recall that a positive joint distribution that satisfies all conditional independencies according to a graph G has a factorization according to the graph G, which is formally shown in the following
- **Hammersely-Clifford Theorem.** Suppose a $\mu(x) > 0$ satisfies all independencies implied by a graphical model $G(V, F, E)$, then

$$\mu(x) = \mu(0) \prod_{a \in F} \widetilde{\psi}_a(x_{\partial a})$$

where

$$\begin{aligned} \widetilde{\psi}_a(x_{\partial a}) &\triangleq \prod_{U \in \partial a} \mu(x_U, 0_{V \setminus U})^{(-1)^{|\partial a \setminus U|}} \\ &= \prod_{U \in \partial a} \left(\frac{\mu(x_U, 0_{V \setminus U})}{\mu(0_U, 0_{V \setminus U})} \right)^{(-1)^{|\partial a \setminus U|}} \end{aligned}$$

- to estimate each term from samples, notice that

$$\frac{\mu(x_U, 0_{V\setminus U})}{\mu(0_U, 0_{V\setminus U})} = \frac{\mu(x_U|0_{\partial\partial U\setminus U})}{\mu(0_U|0_{\partial\partial U\setminus U})}$$

  if the degree is bounded by k, then this involves only k^3 variable nodes

Structural learning

- consider p random variables $x_1, \ldots, x_p$ represented as a graph with p nodes
- the number of edges is $\binom{p}{2} = p(p-1)/2$, and each is either present or not, resulting in $2^{p(p-1)/2}$ possible graphs
- given n i.i.d. observations $x^{(1)}, \ldots, x^{(n)} \in \mathcal{X}^p$, (assuming there is no hidden nodes and all variables are observed), we want to select the best graph among $2^{p(p-1)/2}$, i.e.
  - ⋆ how do we give scores to each model (i.e. graph) given data?
  - ⋆ how do we find the model with the highest score?

- in statistics there are two schools of thoughts: frequentist and Bayesian
  - frequentist assume the parameter θ of interest are deterministic but unknown, in particular there is no distribution over θ
  - maximum likelihood (ML) estimation finds θ that maximizes the score of a model parameter θ, with the score being the log likelihood $\log \mathbb{P}_\theta(x^{(1)}, \ldots, x^{(n)})$

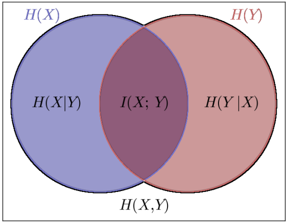
  - Baysians assume there exists a distribution over the parameter of interest $\mu(\theta)$
  - maximum a posteriori (MAP) estimation finds θ maximizing the posterior distribution evaluated on the given data, $\mathbb{P}(\theta | x^{(1)}, \ldots, x^{(n)})$

Frequentist's approach to graph learning

$$\arg\max_{G} \underbrace{\arg\max_{\theta_G} \frac{1}{n}\sum_{i=1}^{n} \log\left(\mu_{G,\theta_G}(x^{(i)})\right)}_{\mathcal{L}(G, x^{(1)}, \ldots, x^{(n)})}$$

- likelihood scores for graphical models are closely related to mutual information and entropy
- recall that

$$\begin{aligned} I(X;Y) &\triangleq \sum_{x,y} \mu(x,y) \log \frac{\mu(x,y)}{\mu(x)\mu(y)} \;, \\ H(X) &\triangleq -\sum_{x} \mu(x) \log \mu(x) \\ H(Y|X) &\triangleq -\sum_{x,y} \mu(x,y) \log \mu(y|x) \\ &= -I(X;Y) + H(Y)\;. \end{aligned}$$



- consider a directed acyclic graphical model (DAG)

$$\mu(x) = \prod_{i=1}^{p} \mu_{\theta_G^{(i)}}(x_i|x_{\pi(i)})$$

where $\theta_G^{(i)}$ represents all the parameters required to describe the conditional distribution (e.g. the entries of the table describing the conditional distribution)

$$\mu(x_i|x_{\pi(i)}) = [\theta_G^{(i)}]_{x_i,x_{\pi(i)}}$$

- in computing $\mathcal{L}(G, x^{(1)}, \ldots, x^{(n)})$,

$$\arg\max_G \underbrace{\arg\max_{\theta_G} \frac{1}{n}\sum_{i=1}^{n} \log\left(\mu_{G,\theta_G}(x^{(i)})\right)}_{\mathcal{L}(G,x^{(1)},\ldots,x^{(n)})}$$

the ML estimate $\widehat{\theta}_G$ for given data is the empirical distribution, i.e.

$$[\widehat{\theta}_G^{(i)}]_{x_i,x_{pi(i)}} = \underbrace{\widehat{\mu}(x_i|x_{\pi(i)})}_{\text{denotes empirical distribution}} = \frac{\widehat{\mu}(x_i, x_{\pi(i)})}{\widehat{\mu}(x_{\pi(i)})}$$

- it follows that the log likelihood is

$$\mathcal{L}(G, x^{(1)}, \ldots, x^{(n)}) = \sum_{i=1}^{p} \mathcal{L}_i(G, x^{(1)}, \ldots, x^{(n)})$$

where

$$\begin{aligned}\mathcal{L}_i(G, x^{(1)}, \ldots, x^{(n)}) &= \frac{1}{n} \sum_{\ell \in [n]} \log \widehat{\mu}(x_i^{(\ell)} | x_{\pi(i)}^{(\ell)}) \\ &= \sum_{x_i, x_{\pi(i)}} \widehat{\mu}(x_i, x_{\pi(i)}) \log \widehat{\mu}(x_i | x_{\pi(i)}) \\ &= -H(\hat{X}_i | \hat{X}_{\pi(i)}) \\ &= I(\hat{X}_i; \hat{X}_{\pi(i)}) - H(\hat{X}_i)\end{aligned}$$

together, it gives

$$\begin{aligned}\mathcal{L}(G, x^{(1)}, \ldots, x^{(n)}) &= \sum_{i=1}^{p} \mathcal{L}(G, x^{(1)}, \ldots, x^{(n)}) \\ &= \sum_{i=1}^{p} I(\hat{X}_i; \hat{X}_{\pi(i)}) - \underbrace{\sum_{i=1}^{p} H(\hat{X}_i)}_{\text{independent of } G}\end{aligned}$$

- given a graph G, this gives the likelihood, and all graphs are subtracted the same node entropies, and we only need to compare the sum of mutual information terms per graph
- **Chow-Liu 1968** addressed this for graphs restricted to trees, where each node only has one parent. In this case, the pairwise mutual information can be computed for all $\binom{p}{2}$ pairs and ML graphs amounts to being the maximum spanning tree on this weighted complete graph. This can be done efficiently using for example Kruskal's algorithm, where you sort the edges according to the mutual information, and you add edges iteratively starting from the largest one, adding an edge each time if it does not introduce a cycle. One needs to be careful with the direction in order to avoid a node having two parents.

- the ML score always favors more complex models, i.e. by adding more edges one can easily increase the likelihood (this follow from the fact that complex models include simpler models as special cases)
- without a penalty on complexity or a restriction to a class of models, ML will end up giving a complete graph
- one way to address this issue is to introduce prior, which we do not address in this lecture

Structural learning

$$n_{\mathsf{Alg}}(G,\theta) \;\equiv\; \inf\{n \in \mathbb{N} : \mathbb{P}_{n,G,\theta}\{\mathsf{Alg}(x^{(1)},\dots,x^{(n)}) = G\} \ge 1-\delta\}\,,$$

$$\chi_{\mathsf{Alg}}(G,\theta) \;\equiv\; \#\text{ operations of Alg when run on } n_{\mathsf{Alg}}(G,\theta) \text{ samples}$$

Typically, we assume G sparse

How would you modify maximum likelihod?

$$\begin{array}{rl} \text{minimize} & \mathcal{L}(\theta; \{x^{(\ell)}\}) \\ \text{subject to} & \|\theta\|_0 \leq m \end{array}$$

Intractable!

ℓ_1-regularized maximum likelihood

$$\begin{aligned} \widehat{\theta} &= \arg\min_{\theta} \; \mathcal{L}(\theta; \{x^{(\ell)}\}) + \lambda\|\theta\|_1 \\ &= -\langle \widehat{M}, \theta\rangle + \phi(\theta) + \lambda\|\theta\|_1 \end{aligned}$$

[cf. J.Friedman, T.Hastie, R.Tibshirani, Biostatistics, 2008]

Local independence test

Idea: For each $i \in V$, and for any candidate neighborood S,

test independence of x_i and $x_{V\setminus S_i}$, $S_i \equiv S \cup \{i\}$.

A possible implementation

```
Local Independence Test( samples $\{x^{(\ell)}\}$ )
For each $i \in V$;
    For each $S \subseteq V \setminus \{i\}$, $|S| \leq k$,
        Compute Score$(S, i) = \widehat{H}(X_i|X_S)$;
    Set $S^* = \arg\min_S$ Score$(S, i)$ and connect $i$ to all $j \in S^*$;
Prune the resulting graph.
```

[P.Abeel, D.Koller, A.Ng, 2006]

Another implementation

$$\begin{aligned}\mathsf{Score}(S,i) \equiv \min_{W\subseteq V\setminus S, j\in S} \max_{x_i,x_W,x_S,x_j} \\ \Big|\widehat{\mathbb{P}}_{n,G,\theta}\{X_i = x_i | X_W = x_W, X_S = x_S\} - \\ \widehat{\mathbb{P}}_{n,G,\theta}\{X_i = x_i | X_W = x_W, X_{S\setminus j} = x_{S\setminus j}, X_j = z_j\}\Big| \,.\end{aligned}$$

[G.Bresler, E.Mossel and A.Sly, APPROX 2008]

Another implementation

```
Local Independence Test( samples $\{x^{(\ell)}\}$, thresholds $(\varepsilon, \gamma)$ )
For each $i \in V$;
    For each $S \subseteq V \setminus \{i\}$, $|S| \le k$,
        Compute Score$(S, i)$;
    $S^* = \arg\max\{|S| : \text{Score}(S, i) > \varepsilon\}$ and connect $i$ to all $j \in S^*$;
```

Nobody would ever use these in practice!!

Why?

n^{k+1} operations!

For the sake of simplicity

$\theta_{ij} = \beta,\ \theta_i = 0$

$$\mu_{G,\beta}(x) = \frac{1}{Z_G(\beta)}\ \exp\Big\{\beta \sum_{(i,j)\in E} x_i x_j\Big\}$$

$$M = (M_{ij})_{1\le i,j\le n}\,, \quad M_{ij} = \mathbb{E}_{G,\beta}\{x_i x_j\}\,, \quad \widehat{M}_{ij} = \frac{1}{n}\sum_{\ell=1}^{n} x_i^{(\ell)} x_j^{(\ell)}$$

A very simple algorithm

```
Thresholding( samples $\{x^{(\ell)}\}$, threshold $\tau$ )
Compute the empirical correlations $\{\widehat{M}_{ij}\}_{(i,j)\in V\times V}$;
For each $(i,j) \in V \times V$
    If $\widehat{M}_{ij} \geq \tau$, set $(i,j) \in E$;
```

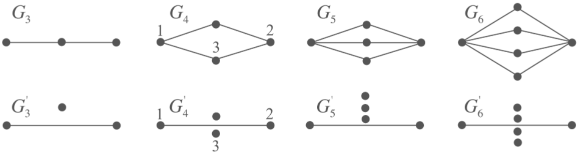
Basic intuition

Thresholding works if

$$\min_{(i,j)\in E} M_{ij} > \max_{(k,l)\notin E} M_{kl}$$

This is true at small β because...

Does not work always because



And its analysis

Theorem

If G is a tree, and $\tau(\beta) = (\tanh\beta + \tanh^2\beta)/2$, then

$$n_{\mathsf{Thr}(\tau)}(G,\theta) \le \frac{8}{(\tanh\beta - \tanh^2\beta)^2}\ \log\frac{2p}{\delta}\,.$$

Theorem

If G has maximum degree $k > 1$ and if $\beta < \operatorname{atanh}(1/(2k))$ then

$$n_{\mathsf{Thr}(\tau)}(G,\beta) \le \frac{8}{(\tanh\beta - \frac{1}{2k})^2}\ \log\frac{2p}{\delta}\,.$$

Theorem

If $k > 3$ and $\theta > C/k$, there are graphs such that for any τ, $n_{\mathsf{Thr}(\tau)} = \infty$.

[J.Bento and A.Montanari, NIPS 2009, and arXiv:1110.1769]

High temperature series

$$\tau = \tanh \beta \, .$$

Theorem (R.Griffiths, J. Math. Phys., 1967)

$$\mathbb{E}_{G,\beta}\{x_i x_j\} \le \sum_{\gamma \in \mathrm{SAW}(i \to j)} \tau^{|\gamma|}$$

This phenomenon is generic

Example: Regularized pseudo-likelihoods
[Meinshausen , Bühlmann, Ann.Stat. 2006]
[P.Ravikumar, M.Wainwright, J.Lafferty, Ann.Stat. 2010]

$\theta_{(i)} \equiv \{\theta_{i,j} : j \in [p] \setminus \{i\}\}$.

$$\text{minimize} \qquad -\frac{1}{n}\sum_{\ell=1}^{n} \log \mathbb{P}_\theta\{x_i^{(\ell)} | x_{\partial i}^{(\ell)}\} + \lambda \|\theta_{(i)}\|_1$$

The first therm only depends on $\theta_{(i)}$! Has explicit expression!