6.S896 - Algorithmic Statistics

Lecture 6: MRF Structure Learning (from finite samples)

So far: Introduced MRFs, GRFs, Ising Models
└ capture cond. independence structure expressible via undirected graph

Showed that MRF is exploitable assumption to test hypotheses about high-dim dist'ns from polynomial in the dimension samples e.g. for uniformity testing (lectures 3,4)

In terms of learning:

- can identify structure of tree-structured MRF using infinitely-many samples (Chow-Liu)

- can learn tree-structured Ising model in total variation distance computationally efficiently using a tight $\Theta\left(\frac{n\log n}{\varepsilon^2}\right)$-samples
  ↳ upper bound: Chow-Liu
  lower bound: Fano

[open question: continuous tree-structured MRFs]

How about learning general MRFs?
└ structure learning: today
TV learning: Thursday

Structure Learning of MRFs $\boxed{\text{from finite samples}}$

Focus: Ising models

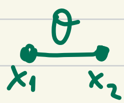
[Similar ideas: Gibbs Random Fields]

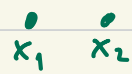
$$p(x) \propto \exp\Big(\sum_{i \neq j} \theta_{ij} x_i x_j + \sum_i \theta_i x_i \Big)$$

Infeasible goal: identify support of $(\theta_{ij})_{ij}$ matrix

i.e. all pairs (i,j) s.t. $\theta_{ij} \neq 0$

why infeasible? B.c. for any finite N number of samples exists small enough $\theta > 0$ can't distinguish w pr of error ≤ 0.49

between 

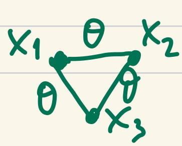
and 

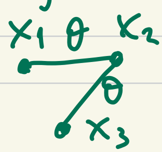
More Realistic goal: identify (i,j) pairs s.t. θ_{ij} "large enough"

$\hookrightarrow$ depends on #samples

Another caveat though: can still get confused if θ_{ij} are too large

for any finite N number of samples exists large enough $\theta > 0$ s.t. can't distinguish w pr of error ≤ 0.49

between 

and 

thus #samples should depend on the "edge strengths"

Theorem [Klivans-Meka'17, Rigollet&Hütter'17, Wu-Sanghavi-Dimakis'19]
improving on Bresler'15, Vuffray-Misra-Lokhov-Chertkov'16

Exist polynomial-time algorithm which,
given $N \gtrsim \frac{\lambda^2 \exp(12\cdot\lambda)\log(n/\delta)}{\varepsilon^4}$ samples

ignores constant factors

from an Ising model $p(x) \propto \exp\left(\sum_{i\neq j}\theta_{ij}x_ix_j + \sum_i \theta_i x_i\right)$
where $\lambda = \lambda(\theta) \triangleq \max_i \left\{\sum_{j\neq i}|\theta_{ij}| + |\theta_i|\right\}$, outputs
$(\hat{\theta}_{ij})_{ij}$ s.t. with prob. $\geq 1-\delta$ satisfy:

$$\forall i,j: \quad |\hat{\theta}_{ij} - \theta_{ij}| \leq \varepsilon$$

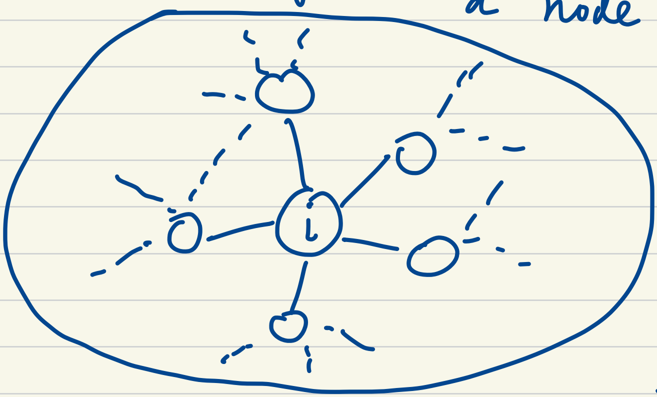
Corollary: If all $\theta_{ij} \neq 0$ satisfy $|\theta_{ij}| > \eta > 0$ we can identify support of $(\theta_{ij})_{ij}$ matrix using

$$N \gtrsim \frac{\lambda^2}{\eta^4}\exp(12\cdot\lambda)\cdot\log(n/\delta) \text{ samples.}$$

with success prob. $\geq 1-\delta$.

[Wainwright-Santhanam]: lower bound on $N \geq \frac{2^{\lambda/4}\log n}{\eta\, 2^{3\eta}}$

Idea for algorithm: look at the neighborhood of a node



$$\Pr(X_i = s \mid X_{-i}) = \frac{\exp\left(\sum_{j\neq i} \theta_{ij} x_j s + \theta_i s\right)}{\exp\left(\sum_{j\neq i} \theta_{ij} x_j s + \theta_i s\right) + \exp\left(-\sum_{j\neq i} \theta_{ij} x_j s - \theta_i s\right)}$$

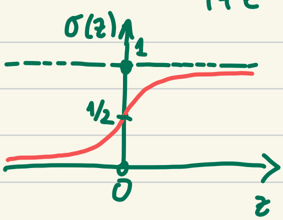
$$= \frac{1}{1 + \exp\left(-2 \cdot s \cdot \left(\sum_{j\neq i} \theta_{ij} x_j + \theta_i\right)\right)}$$

$$= \frac{1}{1 + \exp\left(-2 \cdot s \cdot \left(\langle \underbrace{\theta_{i\cdot}}_{\text{vector } (\theta_{ij})_{j\neq i}}, x_{-i}\rangle + \theta_i\right)\right)}$$

sigmoid function

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

$$= \sigma\left(2 \cdot s \cdot \left(\langle \theta_{i\cdot}, x_{-i}\rangle + \theta_i\right)\right)$$



Idea: think of X_i as $\{\pm 1\}$ - outcome: Y
& X_{-i} as feature vector: $\vec{z}$
in a linear logistic model

$$\Pr[Y \mid \vec{z}] = \sigma\left(s \cdot \left(2\langle \theta_{i\cdot}, \vec{z}\rangle + 2\theta_i\right)\right)$$

want to estimate

- Plan: estimate neighborhood-by-neighborhood
  do logistic regression to estimate $\theta_{i\cdot}$ and θ_i

- From now on focus on node i
  - given samples $X^{(1)}, X^{(2)}, \dots, X^{(N)}$ from p_θ
    create dataset for logistic regression

$$Y^{(\ell)} = X_i^{(\ell)} \; ; \quad Z^{(\ell)} = X_{-i}^{(\ell)} \;, \quad \ell = 1, \dots, N$$

  - do empirical MLE to estimate logistic model

suppose $(\hat{w}, \hat{c})$ is argmin $\rightsquigarrow$

$$\min_{\substack{\vec{w}, c \\ |w|_1 + c \le \lambda}} \underbrace{\frac{1}{N} \sum_{\ell=1}^{N} \log\left(1 + e^{-Y^{(\ell)}\left(\langle w, Z^{(\ell)} \rangle + c\right)}\right)}_{\hat{L}(w,c)\,:\ \text{empirical negative log likelihood}}$$

  - consider also population negative log likelihood
    ↳ means w/ infinite many samples

$$L(w,c) = \mathbb{E}_{\substack{(Y,Z) \sim p_\theta \\ = (X_i, X_{-i})}} \left[\log\left(1 + e^{-Y(\langle w, Z \rangle + c)}\right) \right]$$

known fact: $(w^*, c^*) = (\theta_{i\cdot}, \theta_i)$ is an optimal solution to $\min\limits_{\substack{w, c \\ |w|_1 + c \le \lambda}} L(w,c)$

goal: compare $(\hat{w}, \hat{c})$ with (w^*, c^*)

want to show they are close w.pr $\geq 1-\delta$

over the randomness in $X^{(1)}, \dots, X^{(N)} \sim p_\theta$

if N is large enough

step 1: If $N \geq \Omega(\lambda^2 \log\frac{n}{\delta} / \gamma^2)$ then:

(will show later)

$$L(\hat{w}, \hat{c}) - L(w^*, c^*) \leq \gamma \text{ , w pr } \geq 1-\delta$$

step 2: for any w, c:

(will show later)

$$L(w,c) - L(w^*, c^*) \geq 2 \cdot \mathbb{E}_{z \equiv X_{-i} \sim p_\theta}\left[\left(\sigma(\langle w, z\rangle + c) - \sigma(\langle w^*, z\rangle + c^*)\right)^2\right]$$

step 3:

(will show later)

$\forall\, w, c, w', c'$:

$$\mathbb{E}_{z \equiv X_{-i} \sim p_\theta}\left[\left(\sigma(\langle w, z\rangle + c) - \sigma(\langle w', z\rangle + c')\right)^2\right] \leq \gamma$$

$$\Rightarrow \|w - w'\|_\infty \leq e^{\|w\|_1 + |c| + \|w'\|_1 + |c'|} \cdot \sqrt{16\gamma \cdot e^{2\lambda}}$$

Putting everything together: choose $\gamma \leq O(\varepsilon^2 \cdot e^{-6\lambda})$

use step 1

\+

step 2 setting $(w,c) = (\hat{w}, \hat{c})$

\+

step 3 setting $(w,c) = (\hat{w}, \hat{c})$, $(w', c') = (w^*, c^*)$

$\Rightarrow N \geq \Omega(\lambda^2 e^{12\lambda} \log\frac{n}{\delta} / \varepsilon^4)$ as promised

$\Rightarrow \|\hat{w} - w^*\|_\infty \leq \varepsilon$ w pr $\geq 1-\delta$

Step 1

Lemma 1 (see eg. Shalev-Shwartz & Ben-David book) → understanding ML

Suppose $(Z, Y) \sim D$ s.t. wpr 1 under D: $|Z|_\infty \leq 1$

$Y \in \{\pm 1\}$, $Z \in \mathbb{R}^{q-1}$

Take $L(w,c) = \mathbb{E}_{(Z,Y)\sim D}\left[\log\left(1+e^{-Y\cdot(\langle w,Z\rangle + c)}\right)\right]$

$$\hat{L}(w,c) = \frac{1}{N}\sum_{\ell=1}^{N}\left[\log\left(1+e^{-Y_i^{(\ell)}(\langle w, Z_i^{(\ell)}\rangle + c)}\right)\right]$$

where $(Z^{(1)}, Y^{(1)}), \dots, (Z^{(N)}, Y^{(N)}) \overset{iid}{\sim} D$

Then wpr $\geq 1-\delta$, for all w, c s.t. $|w|_1 + c \leq \lambda$:

$$L(w,c) \leq \hat{L}(w,c) + 2\cdot\lambda\cdot\sqrt{\frac{2\log(2n)}{N}} + \lambda\sqrt{\frac{2\log(2/\delta)}{N}}$$

Proof: via Rademacher complexity analysis. ⊠

Lemma 1 $\overset{\text{if } N \geq \Omega(\lambda^2\log(n/\delta)/\gamma^2)}{\Longrightarrow}$ $L(\hat{w},\hat{c}) - L(w^*,c^*) \leq O(\gamma)$, wpr $\geq 1-\delta$

why? B.c.

$$\begin{aligned}
L(\hat{w},\hat{c}) &\leq \hat{L}(\hat{w},\hat{c}) + O(\gamma) && \text{(By lemma 1 \& choice of } N) \\
&\leq \hat{L}(w^*,c^*) + O(\gamma) && \text{(optimality of } (\hat{w},\hat{c}) \text{ for } \hat{L}) \\
&\leq L(w^*,c^*) + O(\gamma)
\end{aligned}$$

↑ by Chernoff: $\hat{L}(w^*,c^*) - L(w^*,c^*)$ are $\sqrt{\frac{\log(1/\delta)}{N}}$ close w pr $\geq 1-\delta$

$\boxed{\text{Step 2}}$

<u>Lemma</u> 2. In same setting as Lemma 1, suppose
$\Pr_D[Y=1 \mid z] = \sigma(\langle w^*, z\rangle + c^*)$ for some (w^*, c^*)

Then:

$$L(w,c) - L(w^*, c^*) \geq 2 \cdot \mathbb{E}_z\left[\left(\sigma(\langle w, z\rangle + c) - \sigma(\langle w^*, z\rangle + c^*)\right)^2\right]$$

<u>Proof:</u>

claim 1 (next page)

$$L(w,c) - L(w^*,c^*) = \mathbb{E}_{(z,Y)\sim D}\Big[-\frac{Y+1}{2}\log(\sigma(\langle w,z\rangle+c)) - \frac{1-Y}{2}\log(1-\sigma(\langle w,z\rangle+c))$$
$$+\frac{Y+1}{2}\log(\sigma(\langle w^*,z\rangle+c^*)) + \frac{1-Y}{2}\log(1-\sigma(\langle w^*,z\rangle+c^*))\Big]$$

$\Pr[Y=1 \mid z] = \sigma(\langle w^*, z\rangle + c^*)$

$$= \mathbb{E}_z \mathbb{E}_{Y|z}\left[\frac{Y+1}{2}\log\frac{\sigma(\langle w^*,z\rangle+c^*)}{\sigma(\langle w,z\rangle+c)} + \frac{1-Y}{2}\log\frac{1-\sigma(\langle w^*,z\rangle+c^*)}{1-\sigma(\langle w,z\rangle+c)}\right]$$

$$= \mathbb{E}_z\left[\sigma(\langle w^*,z\rangle+c^*)\log\frac{\sigma(\langle w^*,z\rangle+c^*)}{\sigma(\langle w,z\rangle+c)} + (1-\sigma(\langle w^*,z\rangle+c^*))\cdot\log\frac{1-\sigma(\langle w^*,z\rangle+c^*)}{1-\sigma(\langle w,z\rangle+c)}\right]$$

$$= \mathbb{E}_z\left[KL\left(\text{Bernoulli}(\sigma(\langle w^*,z\rangle+c^*)) \,\|\, \text{Bernoulli}(\sigma(\langle w,z\rangle+c))\right)\right]$$

$$\geq \mathbb{E}_z\left[2\cdot\left(\sigma(\langle w^*,z\rangle+c^*) - \sigma(\langle w,z\rangle+c)\right)^2\right]$$

$\frac{1}{2}KL(\text{Bernoulli}(p) \,\|\, \text{Bernoulli}(q)) \geq (p-q)^2$ by Pinsker

Claim 1: $L(w,c) = \underset{(z,Y)\sim D}{\mathbb{E}}\Big[-\frac{Y+1}{2}\log\big(\sigma(\langle w,z\rangle+c)\big)$
$\qquad - \frac{1-Y}{2}\log\big(1-\sigma(\langle w,z\rangle+c)\big)\Big]$

Proof: $L(w,c) = \underset{(z,Y)\sim D}{\mathbb{E}}\Big[-\log\big(\sigma(Y\cdot(\langle w,z\rangle+c))\big)\Big]$

$$= \underset{(z,Y)\sim D}{\mathbb{E}}\Big[-\frac{1+Y}{2}\cdot\log\big(\sigma(\langle w,z\rangle+c)\big) - \frac{1-Y}{2}\log\big(\underbrace{\sigma(-\langle w,z\rangle-c)}_{\overset{\text{III}}{1-\sigma(\langle w,z\rangle+c)}}\big)$$

$\boxtimes$

$\boxed{\text{Step 3}}$

Lemma 3: Suppose $\underline{X} \sim P_\theta$ ← Ising & $\lambda(\theta) = \max_i \left(\sum_j |\theta_{ij}| + \theta_i \right)$

Then $\min_i \min_{s \in \{\pm 1\}} \min_{s_{-i}} \mathbb{P}r[X_i = s \mid X_{-i} = s_{-i}] \geq \frac{1}{2} e^{-2\lambda(\theta)}$

Proof: $\mathbb{P}r[X_i = s \mid X_{-i} = s_{-i}] = \dfrac{1}{1 + \exp\left(-2 \cdot \left(\sum_j \theta_{ij} s_j + \theta_i\right) \cdot s\right)}$

$$\geq \frac{1}{1 + \exp\left(2\left(\sum_j |\theta_{ij}| + |\theta_i|\right)\right)}$$

$$\geq \frac{1}{1 + \exp(2\lambda(\theta))} \geq \frac{1}{2} e^{-2\lambda(\theta)} . \quad \boxtimes$$

Def: A dist'n D over boolean vectors X is $\mathcal{I}$-unbiased iff for all i, $\forall s \in \{\pm 1\}$, $\forall s_{-i}$: $\mathbb{P}r[X_i = s \mid X_{-i} = s_{-i}] \geq \mathcal{I}$.

Ising model is $\left(\frac{1}{2} e^{-2\lambda(\theta)}\right)$-unbiased.

Lemma 4: Suppose dist'n D over $\{\pm 1\}$-vectors is ξ-unbiased.

Then $\forall w, w', c, c'$:

$$\mathbb{E}_{z \sim D}\left[\left(\sigma(\langle w, z\rangle + c) - \sigma(\langle w', z\rangle + c')\right)^2\right] \leq \gamma$$

$$\Rightarrow \|w - w'\|_\infty \leq e^{\overbrace{\|w\|_1 + |c| + \|w'\|_1 + |c'|}^{e^{=\Lambda}}} \cdot \sqrt{\frac{8\gamma}{\xi}}$$

Proof: Pick arbitrary coordinate, say coordinate k:

$$\gamma \geq \mathbb{E}_{Z_{-k}}\left[\mathbb{E}_{Z_k}\left[\left(\sigma(\langle w, Z\rangle + c) - \sigma(\langle w', Z\rangle + c')\right)^2 \mid Z_{-i}\right]\right]$$

$$\geq \mathbb{E}_{Z_{-k}}\Big[\Pr[Z_k = 1 \mid Z_{-k}] \cdot \Big(\sigma(w_k + \underbrace{\langle w_{-k}, Z_{-k}\rangle}_{A(Z_{-k})} + c) - \sigma(w_k' + \underbrace{\langle w_{-k}', Z_{-k}\rangle}_{B(Z_{-k})} + c')\Big)^2$$

$$+ \Pr[Z_k = -1 \mid Z_{-k}] \cdot \left(\sigma(-w_k + A(Z_{-k})) - \sigma(-w_k' + B(Z_{-k}))\right)^2\Big]$$

claim 2 next page

$$\geq \mathbb{E}_{Z_{-k}}\Big[\frac{\xi \cdot e^{-2\Lambda}}{16} \cdot |w_k - w_k' + A(Z_{-k}) - B(Z_{-k})|^2 + $$

$$+ \frac{\xi \cdot e^{-2\Lambda}}{16} \cdot |w_k' - w_k + A(Z_{-k}) - B(Z_{-k})|^2\Big]$$

$$\geq \xi \cdot e^{-2\Lambda} \cdot \frac{1}{8} |w_k - w_k'|^2 \Rightarrow |w_k - w_k'| \leq e^{\Lambda} \sqrt{\frac{8\gamma}{\xi}} \quad \boxtimes$$

Claim 2: $\forall\, x, y \in \mathbb{R}: \ |\sigma(x) - \sigma(y)| \geq \frac{1}{4} e^{-|x|} \cdot e^{-|y|} \cdot |y-x|$

Proof: $|\sigma(x) - \sigma(y)| = \left| \frac{1}{1+e^{-x}} - \frac{1}{1+e^{-y}} \right|$

$$= \frac{|e^{-y} - e^{-x}|}{(1+e^{-x})(1+e^{-y})} \overset{\substack{\text{suppose} \\ \text{wlog } y \geq x}}{=} \frac{e^{-y}|1 - e^{y-x}|}{(1+e^{-x})(1+e^{-y})}$$

$$\geq \frac{|y-x|}{(1+e^{-x}) \cdot (1+e^{y})}$$

$$\geq \frac{e^{-|x|-|y|}}{4} \cdot |y-x|$$

$\boxtimes$