6.S896 - Algorithmic Statistics

Lecture 7: MRF Distribution Learning (from finite samples)

- **Motivation:** estimating parameters/structure requires upper/lower bounds on the edge strengths, o.w. it cannot be done from a finite sample

  philosophical Question: if it takes a gazillion number of samples to distinguish whether or not an edge is present, in what applications does it really matter to find it?

- An alternative Goal: given finite # samples learn the target distribution in some distance
  ↳ TV, KL, ...

  ↳ could mean proper learning: return a distribution from the same family that is close to target
  e.g. proper learning an Ising model must return an Ising model

  or improper learning: return any distribution that is close to target

- From Parameter to Distribution Learning?

In settings where we can estimate parameters, we commonly can also do dist'n learning

e.g. suppose in Ising Model

$$P_\theta(x) \propto \exp\left(\sum_{i \neq j} \theta_{ij} x_i x_j + \sum_i \theta_i x_i \right)$$

we have estimated all θ_{ij}'s & θ_i's to high enough accuracy:

$\forall i \neq j$: $|\hat{\theta}_{ij} - \theta_{ij}| \leq \frac{\varepsilon^2}{n^2}$ ⟵ we can do this when $\lambda(\theta) = \max_i \left(\sum_{j \neq i} |\theta_{ij}| + |\theta_i| \right) < \infty$ using $\Omega\left(\frac{\lambda^2 \cdot e^{12\lambda} \cdot \mathrm{poly}(n)}{\varepsilon^8} \right)$ samples using main theorem of lecture 6

estimated interactions

$\forall i$: $|\hat{\theta}_i - \theta_i| \leq \frac{\varepsilon^2}{n}$

estimated external fields

can also be done from aforementioned number of samples using techniques of lecture 6 (small extension of lemma 4)

$$\boxed{\text{recall: } SKL(p,q) = KL(p\|q) + KL(q\|p)}$$

Now: claim 1: $SKL(P_{\hat{\theta}}, P_\theta) \leq 4\varepsilon^2$

$\Rightarrow TV(P_{\hat{\theta}}, P_\theta) \leq \varepsilon$

$$\boxed{\text{b.c. by Pinsker: } TV(p,q) \leq \sqrt{\tfrac{1}{2} KL(p\|q)}}$$

proof of claim1: from lecture 3/4

$$SKL(p_{\hat\theta}, p_\theta) = \sum_{i\neq j} (\theta_{ij} - \hat\theta_{ij}) \cdot \left(\mathbb{E}_\theta[X_iX_j] - \mathbb{E}_{\hat\theta}[X_iX_j]\right)$$
$$+ \sum_i (\theta_i - \hat\theta_i) \cdot \left(\mathbb{E}_\theta[X_i] - \mathbb{E}_{\hat\theta}[X_i]\right)$$

b.c.
$|\mathbb{E}_\theta X_iX_j - \mathbb{E}_{\hat\theta} X_iX_j| \le 2$
$|\mathbb{E}_\theta X_i - \mathbb{E}_{\hat\theta} X_i| \le 2$

$$\le n^2 \cdot \frac{\varepsilon^2}{n^2} \cdot 2 + n \cdot \frac{\varepsilon^2}{n} \cdot 2 = 4\varepsilon^2$$

- Can we do distribution learning without paying for low temperature?

  & more generally, without estimating parameters?

enters ... $\boxed{\text{Tournament-Based Approach}}$

↳ very general, quick & dirty method to understand sample complexity of distribution learning

↳ also provides algorithm to select among a set of hypotheses distributions one that is close to some target

- Tournament-Based Approach: 3 steps at a high level

  ↳ ① For the distance of interest d, the accuracy of interest ε, and the fami of dist'ns of interest $\mathcal{H}$ find a small ε-cover $\mathcal{H}_{(d,\varepsilon)} \subseteq \mathcal{H}$

  ↳ for all $p \in \mathcal{H}$ $\exists\, q \in \mathcal{H}_{(d,\varepsilon)}$ s.t. $d(p,q) < \varepsilon$

  ② Define a pairwise comparison procedure

  comp$(P; H_1, H_2)$: output 1) H_1 as winner
  2) H_2 as winner
  3) a tie

  want this to only use sample access to P which is unknown dist'n

  ③ Set-up a tournament among all distributions in $\mathcal{H}_{(d,\varepsilon)}$ which runs a sequence of pairwise comparisons of distributions in $\mathcal{H}_{(d,\varepsilon)}$ to select a winner

  WANT: winner of tournament is close to unknown P

- Today: tournament for TV distance

  [Devroye - Lugosi '96] : Scheffé Estimate

  [Yatracos '85], [Daskalakis et al '12], [Daskalakis-Kamath '14]

  [Acharya et al. '14]

  Can do other distances too [e.g. Feldman - O'Donnell - Servedio '06]

◦ **Main Primitive**: $\mathrm{Comp}(P; H_1, H_2)$ pairwise comparison procedure

(A) { **uses**: sample access to $P, H_1, H_2 \in \mathcal{H}$
PDF comparator: given x, compares $H_1(x)$ & $H_2(x)$
will assume PDFs exist

(B) { **behavior that I want**: Comp draws $m = m(\varepsilon, \delta)$ samples from P, H_1, H_2 and has the following properties:

For $(i,j) = (1,2)$ OR $(2,1)$:

1. If $d(P, H_i) \leq \varepsilon$ & $d(P, H_j) > 8\varepsilon$:

   H_i is declared winner, w.pr. $\geq 1-\delta$

2. If $d(P, H_i) \leq \varepsilon$ & $d(P, H_j) > 4\varepsilon$:

   H_i wins or it's a tie, w.pr. $\geq 1-\delta$

3. If $d(H_1, H_2) \leq 5\varepsilon$:

   it's a tie, w.pr. $\geq 1-\delta$

Lemma 1: For $d =$ TV distance & assuming (A) there exists a poly-time algorithm satisfying behavior (B) where:

$$m = \Theta\left(\frac{\log(1/\delta)}{\varepsilon^2}\right).$$

• Suppose we have the following

(A′) $\begin{cases} \text{- we have a comparator which assumes (A) and satisfies (B)} \\ \text{- we have a cover } \mathcal{H}_{(d,\varepsilon)} \text{ of our set of dist'ns } \mathcal{H} \\ \text{- we have sample access to some unknown } P \in \mathcal{H}. \end{cases}$

How do we set-up a tournament to select som $H \in \mathcal{H}$ that is close in d to P?

<u>Lemma 2</u>: Assume (A′). There exists an algorithm that:

- draws $m\left(\varepsilon, \frac{\delta}{2 \cdot N}\right)$ samples from P
- makes $\binom{N}{2}$ comparisons
- outputs H s.t. $d(P,H) < 8\varepsilon$ wpr $\geq 1-\delta$.

cardinality of $\mathcal{H}_{(d,\varepsilon)}$

Daskalakis-Kamath & Acharya et al improve this to $N \cdot \log N$

Lemma 1 + Lemma 2 $\Rightarrow$

Corollary: For $d = TV$ distance, suppose we have:

- a set of dist'ns $\mathcal{H}$ & ε-cover $\mathcal{H}_\varepsilon$ of cardinality N;
- sample access to unknown $P \in \mathcal{H}$;
- pdf comparator for all $H_1, H_2 \in \mathcal{H}_\varepsilon$;
- sample access to all $H \in \mathcal{H}_\varepsilon$;

Then there is an algorithm that

- draws $O\left(\frac{\log(N/\delta)}{\varepsilon^2}\right)$ samples from P (as well as all $H \in \mathcal{H}_\varepsilon$)
- runs in time $O\left(N^2 \cdot \frac{\log(N/\delta)}{\varepsilon^2}\right)$ $\rightsquigarrow$ can be improved to $N \cdot \log N \cdot \frac{\log(N/\delta)}{\varepsilon^2}$
- and outputs H s.t. $d(P, H) \leq 8\varepsilon$, w pr $\geq 1-\delta$.

Application of Corollary to learn Ising Models in TV

For $\lambda \geq 0$ define:

$$\mathcal{H}(\lambda) = \{\text{all Ising models s.t. } \max_i \Big(\sum_{i \neq j} |\theta_{ij}| + \sum_i |\theta_i|\Big) \leq \lambda\}$$

Theorem: given $\Omega\left(\frac{n^2 \log \frac{n\lambda}{\varepsilon \cdot \delta}}{\varepsilon^2}\right)$ samples from Ising model $p_\theta \in \mathcal{H}(\lambda)$, can find Ising model $p_{\theta'} \in \mathcal{H}(\lambda)$ such that $TV(p_\theta, p_{\theta'}) \leq \varepsilon$, wpr $\geq 1-\delta$

Proof: Recall: $SKL(p_\theta, p_{\theta'}) \leq \sum_{i \neq j} |\theta_{ij} - \theta'_{ij}| + \sum_i |\theta_i - \theta'_i|$

$$\Rightarrow \left.\begin{array}{l} SKL(p_\theta, p_{\theta'}) \leq \frac{\varepsilon^2}{16} \\ TV(p_\theta, p_{\theta'}) \leq \frac{\varepsilon}{8} \end{array}\right\} \text{ if } |\theta_{ij} - \theta'_{ij}| \leq \frac{\varepsilon^2}{32n^2}\ \forall ij \text{ and } |\theta_i - \theta'_i| \leq \varepsilon^2/32n, \forall i$$

$\Rightarrow$ gridding over θ_{ij}'s & θ_is exists $\frac{\varepsilon}{8}$-cover of $\mathcal{H}_\lambda$ in TV that has size $N \leq \left(\frac{32\lambda n^2}{\varepsilon^2}\right)^{n^2} \left(\frac{32\lambda n}{\varepsilon^2}\right)^n$ (using that all Ising models in $\mathcal{H}(\lambda)$ satisfy that $|\theta_{ij}| \leq \lambda$, $|\theta_i| \leq \lambda$)

$\xRightarrow{\text{tournament corollary}}$ $\frac{\log(N/\delta)}{\varepsilon^2}$ samples suffice to learn within ε in TV $\boxtimes$

Remark: Can strengthen Theorem to remove dependence on λ entirely! [see Theorem 9 of Bresler-Cai-Daskalakis '20]

Theorem 9 from Brustle-Cai-Daskalakis '20:

Theorem 9 (Learnability of MRFs in Total Variation and Prokhorov Distance). *Suppose we are given sample access to an MRF p, as in Definition 7, defined on an unknown graph with hyper-edges of size at most d.*

- ***Finite alphabet*** Σ: *Given $\frac{\mathrm{poly}\left(|V|^d,|\Sigma|^d,\log(\frac{1}{\varepsilon})\right)}{\varepsilon^2}$ samples from p we can learn some MRF q whose hyper-edges also have size at most d such that $\|p-q\|_{TV} \leq \varepsilon$. If the graph on which p is defined is known, then $\frac{\mathrm{poly}\left(|V|,|E|,|\Sigma|^d,\log(\frac{1}{\varepsilon})\right)}{\varepsilon^2}$-many samples suffice. Moreover, the polynomial dependence of the sample complexity on $|\Sigma|^d$ cannot be improved, and the dependence on ε is tight up to $\mathrm{poly}(\log\frac{1}{\varepsilon})$ factors.*
- ***Alphabet*** $\Sigma = [0, H]$: *If the log potentials $\phi_v(\cdot) \equiv \log(\psi_v(\cdot))$ and $\phi_e(\cdot) \equiv \log(\psi_e(\cdot))$ for every node v and every edge e are C-Lipschitz w.r.t. the ℓ_1-norm, then given $\mathrm{poly}\left(|V|^{d^2}, \left(\frac{H}{\varepsilon}\right)^d, C^d\right)$ samples from p we can learn some MRF q whose hyper-edges also have size at most d such that $\|p-q\|_P \leq \varepsilon$. If the graph on which p is defined is known, then $\mathrm{poly}\left(|V|, |E|^d, \left(\frac{H}{\varepsilon}\right)^d, C^d\right)$-many samples suffice.*

Prokhorov Distance

Definition 7. *A* Markov Random Field (MRF) *is a distribution defined by a hypergraph $G = (V, E)$. Associated with every vertex $v \in V$ is a random variable X_v taking values in some alphabet Σ, as well as a potential function $\psi_v : \Sigma \to [0,1]$. Associated with every hyperedge $e \subseteq V$ is a potential function $\psi_e : \Sigma^e \to [0,1]$. In terms of these potentials, we define a probability distribution p associating to each vector $x \in \Sigma^V$ probability $p(x)$ satisfying:*

$$p(x) = \frac{1}{Z} \prod_{v\in V} \psi_v(x_v) \prod_{e\in E} \psi_e(x_e), \tag{1}$$

where for a set of nodes e and a vector x we denote by x_e the restriction of x to the nodes in e, and Z is a normalization constant making sure that p, as defined above, is a distribution. In the degenerate case where the products on the RHS of (1) always evaluate to 0, we assume that p is the uniform distribution over Σ^V. In that case, we get the same distribution by assuming that all potential functions are identically 1. Hence, we can in fact assume that the products on the RHS of (1) cannot always evaluate to 0.

this is general no matter what d is and $comp(\cdot,\cdots)$ is as long as it satisfies desiderata stated above

Proof of Lemma 2:

- For all pairs $H_i, H_j \in \mathcal{H}_\varepsilon$ run

$$comp(P, H_i, H_j)$$

using $m(\varepsilon, \frac{\delta}{2N})$ samples

- Output any H_k that never lost (won or tied in every comparison it participated in)

If no such H_k exists output "failure"

- Claim 2.1: Suppose $H^* \in \mathcal{H}_\varepsilon$ satisfies $d(P, H^*) \leq \varepsilon$. (exists b.c. $\mathcal{H}_\varepsilon$ is a cover)

With prob. $\geq 1 - \frac{\delta}{2}$, H^* will never lose

thus the algorithm will not output "failure".

Proof: Take any $H' \in \mathcal{H}_\varepsilon$.

- If $d(P, H') > 4\varepsilon$, by property B.2 of the comparator, H^* either won or tied against H' wpr $\geq 1 - \frac{\delta}{2N}$

$d(P,H^*) \leq \varepsilon$

- If $d(P, H') \leq 4\varepsilon \implies d(H, H') \leq 5\varepsilon$

so by property B.3 of the comparator, H^* tied w/ H' wpr $\geq 1 - \delta/2N$

doing a union bound over all $H' \in \mathcal{H}_\varepsilon$ proves the claim $\boxtimes$

Claim 2.2: If $H \in \mathcal{H}_\varepsilon$ never lost, then $d(P,H) \le 8\varepsilon$,
w pr. $\geq 1 - \delta/2$

Proof: Suppose $d(P,H) > 8\varepsilon$, and take H^* s.t. $d(P,H^*) \le \varepsilon$.

By property B.1 of the comparator H loses to H^*
w.pr. $\geq 1 - \frac{\delta}{2N}$

By union bound, all $H \in \mathcal{H}_\varepsilon$ s.t. $d(P,H) > 8\varepsilon$,
lose to H^* with prob. $\geq 1 - \delta/2$.

$\boxtimes$

Claim 2.1 + Claim 2.2 $\Rightarrow$ w pr. $\geq 1-\delta$, algorithm does not fail & its output dist'n is 8ε-close to P

$\boxtimes$

↱ this is special for TV

Proof of Lemma 1: Set up a competition between H_1, H_2.

- Define set $\mathcal{W}_1 = \mathcal{W}(H_1 \| H_2) = \{ x \text{ s.t. } H_1(x) > H_2(x) \}$

  ↳ to determine for some x whether $x \in \mathcal{W}$ will use PDF comparator

- Define $p_1 = H_1(\mathcal{W}_1)$, $p_2 = H_2(\mathcal{W}_1)$

  clearly $p_1 > p_2$ and $TV(H_1, H_2) = p_1 - p_2$

- Here is how $Comp(P, H_1, H_2)$ works

  1a. draw $m = O\left(\frac{\log 1/\delta}{\varepsilon^2}\right)$ samples from P & call $\hat{\tau}$ the fraction of them that fall in $\mathcal{W}_1$

  1b. draw m samples from H_1, call $\hat{p}_1$ the fraction in $\mathcal{W}_1$

  1c. —//— —//— —//— H_2, —//— $\hat{p}_2$ —//—

  2. if $\hat{p}_1 - \hat{p}_2 \lesssim 6\varepsilon$, declare a tie

  3. If $\hat{\tau} > \hat{p}_1 - 2\varepsilon$, declare H_1 winner

  4. If $\hat{\tau} < \hat{p}_2 + 2\varepsilon$, declare H_2 winner

  5. declare a tie.

Claim: For $(i,j) = (1,2)$ OR $(2,1)$, suppose $TV(P, H_i) \le \varepsilon$.
Then:

1. If $TV(P, H_j) > 8\varepsilon$, then

$$H_i \text{ is declared winner w pr} \geq 1 - 6e^{-m\varepsilon^2/2} \quad \text{(B.1)}$$

2. If $TV(P, H_j) > 4\varepsilon$, then

$$H_i \text{ wins or it's a tie wpr} \geq 1 - 6e^{-m\varepsilon^2/2} \quad \text{(B.2)}$$

3. If $TV(H_1, H_2) \le 5\varepsilon$, then

$$\text{it's a tie wpr} \geq 1 - 6e^{-m\varepsilon^2/2}$$

Proof: • By Chernoff, w pr $\geq 1 - 6e^{-m\varepsilon^2/2}$ the following are true:

$$|p_1 - \hat{p}_1| < \varepsilon/2, \quad |p_2 - \hat{p}_2| < \varepsilon/2, \quad |\tau - \hat{\tau}| < \varepsilon/2 \quad (*)$$

$$\tau = P(W_1)$$

• Assuming $(*)$ we have:

<u>case 1</u>: $TV(P, H_i) \le \varepsilon$, $TV(P, H_j) > 8\varepsilon \overset{\text{triangle}}{\Rightarrow} \underbrace{TV(H_i, H_j)}_{\equiv p_1 - p_2} > 7\varepsilon$

$$\Rightarrow \hat{p}_1 - \hat{p}_2 \geq p_1 - p_2 - \varepsilon > 6\varepsilon$$

$$\Rightarrow \text{algorithm will go beyond step 2}$$

Now $TV(P, H_i) \le \varepsilon \Rightarrow |\tau - p_i| \le \varepsilon \Rightarrow |\hat{\tau} - \hat{p}_i| \le 2\varepsilon$
$\Rightarrow$ steps 3, 4 will choose H_i as winner
no matter if $i=1$ or $i=2$

<u>indeed</u>: suppose $i=1$; then b.c. $|\hat{\tau} - \hat{p}_1| \le 2\varepsilon \rightsquigarrow$ H_1 declared winner at step 3
If $i=2$, then $\hat{p}_1 > \hat{p}_2 + 6\varepsilon > \hat{\tau} + 4\varepsilon \rightsquigarrow$ algorithm goes to step 4
at step 4, b.c. $\hat{\tau} < \hat{p}_2 + 2\varepsilon \rightsquigarrow$ H_2 declared winner

<u>case 2</u>: $TV(P,H_i) \leq \varepsilon$, $TV(P,H_j) > 4\varepsilon$

<u>subcase 2.1</u>: If $\hat{p}_1 - \hat{p}_2 \leq 6\varepsilon$, then algorithm stops at step 2 declaring a tie

<u>subcase 2.2</u>: If $\hat{p}_1 - \hat{p}_2 > 6\varepsilon$, algorithm goes to step 3

- if $i=1$, then $TV(P,H_1) \leq \varepsilon \Rightarrow |\tau - p_1| \leq \varepsilon \Rightarrow$

  $\Rightarrow |\hat{\tau} - \hat{p}_1| \leq 2\varepsilon \rightsquigarrow$ algorithm declares H_1 as winner

- if $i=2$, then $TV(P,H_2) \leq \varepsilon \Rightarrow |\hat{\tau} - \hat{p}_2| < 2\varepsilon$

  thus $\hat{p}_1 > \hat{p}_2 + 6\varepsilon > \hat{\tau} + 4\varepsilon$
  so algorithm does not stop at step 3
  & at step 4, declares H_2 as winner

$\hookrightarrow$ so in this case H_i is declared the winner regardless of whether $i=1$ or 2

so in case 2, either H_i wins or it's a tie

<u>case 3</u>: $TV(H_1,H_2) \leq 5\varepsilon \Rightarrow \hat{p}_1 - \hat{p}_2 \leq 6\varepsilon \Rightarrow$ step 2 declares a tie

$\boxtimes$