Lecture 21

- Recall that the CFN model is a Markov Chain on a tree (T, P, μ_ρ), where T is a directed binary tree rooted at ρ and with leaf set $[n]$, whose edges have transition matrices

$$P_e = \begin{bmatrix} 1-p_e & p_e \\ p_e & 1-p_e \end{bmatrix} \quad \text{over the character set } \{0,1\},$$

and $\mu_\rho = (1/2 \ \ 1/2)$.

[p_e is the mutation probability of the edge]

- The tree reconstruction problem is the following:
  - given $\underset{\sim}{Z} = \left(Z^1_{[n]}, Z^2_{[n]}, \dots, Z^k_{[n]}\right)$, that is k independent samples from the CFN model at the leaves of T, the goal is to reconstruct the unrooted, undirected tree $T^{-\rho}$
  - the strong reconstruction problem also asks for a CFN model over the leaf set $[n]$ whose leaf character distn' is within ε total variation distance from the distn' of the actual model (which the samples $\underset{\approx}{Z}$ were sampled from)

- Our goal is to reconstruct $T^{-\rho}$ using a tree metric
- Our tree metric is inspired by the following decomposition:

$$P_e = (1-2p_e)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + 2p_e \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$$

interpretation: on an edge $e=(u,v)$: $\xi_v = \xi_u$ w/pr $\overbrace{1-2p_e}^{\theta_e \ \equiv}$

$\xi_v \perp \xi_u$ w/pr $2p_e$

→ independent

- consider now a path in tree $\text{Path}(a,b) = \{e_1, \dots, e_k\}$

$$\xi_b = \begin{cases} \xi_a & \text{w pr. } \prod_{i=1}^{k} \theta_{e_i} \\ 0/1 \text{ u.a.r.}, & \text{w pr. } 1 - \prod_{i=1}^{k} \theta_{e_i} \end{cases}$$

define: $\theta(a,b) := \prod_{i=1}^{k} \theta_{e_i}$ and $p(a,b) = \frac{1-\theta(a,b)}{2}$

- Since θ_e's multiply, reasonable to define

$$w_e = -\log \theta_e$$

then $$\sum_{i=1}^{k} w_{e_i} = -\log \prod \theta_{e_i} = -\log \theta(a,b)$$

- So CFN model defines tree metric $\delta(a,b) = -\log \theta(a,b)$
- If I could estimate $\delta(a,b)$ to within $\frac{1}{4}\overset{\min_e\{w_e\}}{w_*}$, I could solve the reconstruction problem (from last lecture)

- **Estimating $\delta(a,b)$ from Ξ**

define $\hat{P}_{ab} := \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}_{\{\Xi_a^i \neq \Xi_b^i\}}$; then $\mathbb{E}[\hat{P}_{ab}] = \frac{1-\theta(a,b)}{2} \equiv p(a,b)$

set $$\hat{\delta}(a,b) = \begin{cases} -\log[1 - 2\hat{p}^{ab}], & \text{if } \hat{p}^{ab} < 1/2 \\ \text{equal to } +\infty, & \text{if } \hat{p}^{ab} \geq \frac{1}{2} \end{cases}$$

Claim: Suppose that $\forall\, a,b \in [n]$

$$\left| p^{ab} - \hat{p}^{ab} \right| < \varepsilon \leq \frac{1}{2}\left(1 - e^{-w_*/4}\right)\left(1 - 2p^{ab}\right)$$

then

$$\max_{q=\{a,b,c,d\}} \left| \delta(q) - \hat{\delta}(q) \right| \leq 2 \max_{a,b} \left| \delta(a,b) - \hat{\delta}(a,b) \right| < \frac{1}{2} w_* .$$

where $\underset{\{a,b,c,d\}}{\delta(\underset{\shortparallel}{q})} = \frac{1}{2}\left[\delta(a,c) + \delta(b,d) - \delta(a,b) - \delta(c,d)\right]$.

Proof:

- Observation:

$$p^{ab} + \varepsilon < p^{ab} + \frac{1}{2} - p^{ab} < \frac{1}{2}$$

- hence $-\log\left(1 - 2(p^{ab} \pm \varepsilon)\right)$ well-defined (i.e. not $+\infty$)

- Now:

$$\left|\delta(a,b) - \hat{\delta}(a,b)\right| = \left| \log(1 - 2p^{ab}) - \log\left(1 - 2(p^{ab} \pm \varepsilon)\right) \right|$$

$$= \left| \log \frac{1 - 2p^{ab} \pm 2\varepsilon}{1 - 2p^{ab}} \right|$$

$$= \left| \log\left(1 \pm \frac{2\varepsilon}{1 - 2p^{ab}}\right) \right| \leq \frac{1}{4} w_*$$

$\hookrightarrow$ indeed:

$$\left| \log\left(1 \pm \frac{2\varepsilon}{1-2p^{ab}}\right) \right| \leq \max\left\{ \log\left(1 + \frac{2\varepsilon}{1-2p^{ab}}\right), -\log\left(1 - \frac{2\varepsilon}{1-2p^{ab}}\right) \right\}$$

$$\leq -\log\left(1 - \frac{2\varepsilon}{1-2p^{ab}}\right) \leq \frac{1}{4} w_*$$

$\boxtimes$

Theorem 1: Let $w_* = \min_e w(e)$ and $W_* = \max_{a,b} \delta(a,b)$.

Then $k = O\left(\frac{e^{2W_*}}{(1-e^{-w_*/4})^2} \cdot \log n \right)$ samples suffice

to get the correct tree T w/prob $1-o(1)$ as $n \to +\infty$.

Proof: • From Chernoff bounds:

$$\mathbb{P}r\left[\left| p_{ab} - \hat{p}_{ab} \right| \geq \varepsilon \right] \leq 2 \exp\left(-2\varepsilon^2 k \right) < \frac{1}{n^3} \qquad (*)$$

choosing $k = \Omega\left(\frac{1}{\varepsilon^2} \log n \right)$.

• Use $\varepsilon = \frac{1}{2} \left(1 - e^{-w_*/4} \right) \cdot e^{-W_*}$ $(**)$

• By $(*)$ and a union bound, it follows that w/ probability $\geq 1 - \frac{1}{n}$:

for all $a, b \in [n]$:

$$\left| p^{ab} - \hat{p}^{ab} \right| < \frac{1}{2} \left(1 - e^{-w_*/4} \right) \left(1 - 2p^{ab} \right).$$

• Hence by previous claim we'll get all quartets right w. pr. $\geq 1 - \frac{1}{n}$ and, therefore

the correct tree w/prob. $\geq 1 - \frac{1}{n}$.

• From $(*)$, $(**)$ it follows that $k = O\left(\frac{e^{2W_*}}{(1-e^{-w_*/4})^2} \log n \right)$ suffices

for this purpose.

$\boxtimes$

- So assume that mutation probabilities are bounded away from 0 and 1/2, i.e. $0 < c_1 \le p_e \le c_2 < 1/2$, for all edges e.

- Then $0 < \underbrace{-\log(1-2c_1)}_{= f} \le w_e \le \underbrace{-\log(1-2c_2)}_{= g} < +\infty$ (f, g are constants)

- In this case, $\min_e w_e \ge f$ and $W^* \le (n+1)\, g$

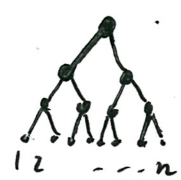
  so from previous theorem : $k = 2^{\Omega(n)}$ suffices.

- This is not tight; it turns out that Thm 1 holds if we replace W_* by the weighted depth of the tree.

  def (Weighted Depth): The depth of an edge is the length (under δ) of the shortest path between two leaves crossing e. The depth of a tree is the largest depth of an edge in the tree.

  e.g.

  caterpillar tree (suppose all edges have weight 1)
  the depth of all edges is constant



  full binary tree (suppose $w_e = 1$)
  the depth is at most $2\log_2 n$

- **Ex (0.5pt)**: If $w_e = 1, \forall e$, then the depth of an edge e in any binary tree is at most $2 \cdot \log_2 n + 2$.

- Hence, if $f \leq w_e \leq g$, $\forall e$, and we use the modified version of Thm 1, it follows that w/ $k = \text{poly}(n)$ samples from the CFN model we can reconstruct the tree w/ prob. $1-o(1)$ as $n \to +\infty$.

- Can we do better than poly(n) sequence length?
  (recall that our counting lower bound was just $\Omega(\log n)$)

Steel's Conjecture:

Let $\theta^* = 1/\sqrt{2}$. Then, if $\theta(e) \leq \theta^*$, $\forall e$, poly(n) samples are necessary for the tree reconstruction problem, while if $1 > \theta(e) > \theta^*$, $O(\log n)$ samples suffice.

More on Steel's Conjecture next time!