Lecture 22

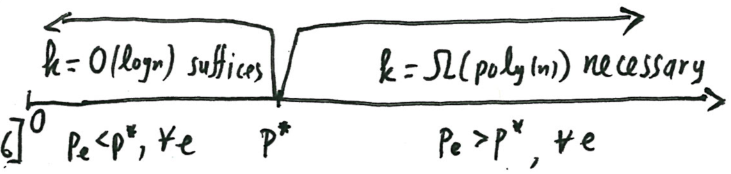
Last Time: • CFN model: Markov Chain on a tree (T, P, μ_ρ) where T is a directed binary tree rooted at ρ, w/ leaf set $[n]$, transition matrices $P_e = \begin{pmatrix} 1-p_e & p_e \\ p_e & 1-p_e \end{pmatrix}$ on every edge e, where $p_e \in (0, 1/2)$ is the mutation probability on edge e, and probability distn' $\mu_\rho = (1/2 \ \ 1/2)$ at the root, where all nodes of the tree have state space $\mathcal{C} = \{0, 1\}$.

- Tree Reconstruction thm: $k = O\left(\dfrac{e^{2W_*}}{(1-e^{-w_*/4})^2} \cdot \log n \right)$ samples from the CFN model at the leaves of the tree suffice to obtain $T^{-\rho}$ (w/ any desired probability), where $w_* = \min_e w_e = \min_e (-\log(1-2p_e))$ $= \min_e (-\log \theta_e)$ and W_* = is the longest leaf-to-leaf path under edge weights $w_e := -\log(1-2p_e)$

- We argued that W_* can be replaced by the weighted depth of the tree, which leads to $k = \text{poly}(n)$, if $0 < c_1 \le p_e \le c_2 < 1/2, \forall e$ where c_1, c_2 absolute constants.

- Can we go beyond poly(n) sequences?

- Steel's Conjecture: shown by [Mossel '03] [Daskalakis, Mossel, Roch '06]



- In this lecture, we want to give insight into the phase transition shown above; in particular, we want to understand why the reconstruction problem is easier when $p_e < p^*$. We start w/ analyzing how much information about the root DNA can be traced inside the leaf DNAs

- How much information about the root state is typically hidden in the boundary of a tree?

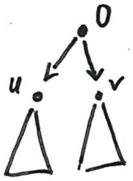
- simple setup: consider the CFN model on the complete binary tree T_h of height h; suppose mutation probability is p on all edges. where $p \in (0, 1/2)$; for notational simplicity work w/ character set $C = \{-1, +1\}$.

- T_h has 2^h leaves; label them $1, \dots, 2^h$; call the root 0.

- we are interested in guessing the state σ_0 at the root, by looking at the state $s_h = (\sigma_i)_{i \in [2^h]}$ at the leaves of the tree.

- Recall that

$$\begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} = \underbrace{(1-2p)}_{\theta} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + 2p \cdot \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$$



hence w/prob $(2p)^2$ the state σ_0 of the root and the states of its children are independent; under this event s_h and σ_0 are also independent and any guess is as good as a random coin flip.

- Intuitively, we are interested in getting an overall probability of correct guess that is better than a random coin flip, even as $h \to +\infty$.

- We claim that the following definition captures our desire.

Def (Ancestral Reconstruction Solvability):

- Let μ_h^+ be the distribution of s_h conditioned on the root state σ_0 being $+1$, and similarly define μ_h^-, in the above setup for a given $p \in (0, 1/2)$.
- We say that the ancestral reconstruction problem is solvable if

$$\liminf_h \| \mu_h^+ - \mu_h^- \|_{TV} > 0,$$ o.w. we call it unsolvable.

- Why is the definition capturing what we need?

- **Claim 1:** The best estimator is the maximum likelihood estimator, defined as:

$$\hat{\sigma}_0^{ML}(s_h) = \begin{cases} +1, & \mu_h^+(s_h) \geq \mu_h^-(s_h) \\ -1, & \text{o.w.} \end{cases}$$

Proof: • Consider any estimator $\hat{\sigma}_0(\cdot)$. Then:

$$\Pr\left[\hat{\sigma}_0(s_h) \neq \sigma_0\right] = \sum_{s_h} \mathbb{1}_{\{\hat{\sigma}_0(s_h)=+1\}} \cdot \overbrace{\frac{1}{2}\mu_h^-(s_h)}^{\Pr[s_h \wedge \sigma_0=-]}$$

$$+ \sum_{s_h} \mathbb{1}_{\{\hat{\sigma}_0(s_h)=-1\}} \cdot \underbrace{\frac{1}{2}\mu_h^+(s_h)}_{\Pr[s_h \wedge \sigma_0=+]}$$

• It is now obvious that the ML estimator minimizes $\Pr[\hat{\sigma}_0(s_h) \neq \sigma_0]$. ☒

• Now what's the performance of $\hat{\sigma}_0^{ML}(\cdot)$. Note that:

$$\Pr\left[\hat{\sigma}_0^{ML}(s_h) = \sigma_0\right] - \Pr\left[\hat{\sigma}_0^{ML}(s_h) \neq \sigma_0\right] =$$

$$= \frac{1}{2}\sum_{s_h} \mathbb{1}_{\{\hat{\sigma}_0^{ML}(s_h)=+1\}} \cdot \left(\mu_h^+(s_h) - \mu_h^-(s_h)\right)$$

$$+ \frac{1}{2}\sum_{s_h} \mathbb{1}_{\{\hat{\sigma}_0^{ML}(s_h)=-1\}} \cdot \left(\mu_h^-(s_h) - \mu_h^+(s_h)\right)$$

$$= \frac{1}{2}\sum_{s_h} \hat{\sigma}_0^{ML}(s_h) \cdot \left(\mu_h^+(s_h) - \mu_h^-(s_h)\right)$$

$$= \frac{1}{2}\sum_{s_h} \left|\mu_h^+(s_h) - \mu_h^-(s_h)\right| = \left\|\mu_h^+ - \mu_h^-\right\|_{TV}.$$

So $$\Pr\left[\hat{\sigma}_0^{ML}(s_h) = \sigma_0\right] = \frac{1}{2} + \underbrace{\frac{1}{2}\|\mu_h^+ - \mu_h^-\|_{TV}}$$

↝ whence the advantage over a random guess of the best estimator is captured by this TV distance

↝ our definition of "solvability" insists that the advantage remains non-trivial for all h.

Theorem: Let $\theta^* = 1 - 2p^* = \frac{1}{\sqrt{2}}$. The Ancestral Reconstruction Problem is solvable iff $p < p^*$.

- We'll prove only one side of the theorem. Namely that $p < p^*$ implies solvability.

- But first let us consider a potentially weaker estimator, via the majority function. The sign of this function could be used as an estimator of the spin at the root (in our simple setup, the sign of the majority is actually the ML estimator, but this is not the case for more complicated models, e.g. when edge probabilities are different.)

- **Def:** The majority at level h is defined as:

$$Z_h = \frac{1}{2^h \theta^h} \sum_{x \in [2^h]} \sigma_x ,$$

where $\theta = 1 - 2p$.

[note: the normalization by $2^h\theta^h$ makes it an unbiased estimator.

indeed $$\mathbb{E}\left[Z_h | \sigma_0\right] \underset{\substack{\text{linearity of expectation + symmetry of leaves}}}{=} \frac{1}{\theta^h}\mathbb{E}[\sigma_1 | \sigma_0] = \frac{1}{\theta^h}\cdot\left(\theta^h \sigma_0 + (1-\theta^h)\cdot 0\right) = \sigma_0 \quad]$$

Notation: $\mathbb{E}_h^+$ is the expectation under μ_h^+ and Var_h^+ is the variance under μ_h^+.

similarly, we define $\mathbb{E}_h^-$ and Var_h^- .

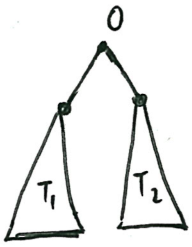
Theorem (Phase transition for majority):

$$\mathrm{Var}[Z_h] \xrightarrow[h\to+\infty]{} \begin{cases} \dfrac{1/2}{1-(2\theta^2)^{-1}} & , \ 2\theta^2>1 \\ +\infty & , \ 2\theta^2\leq 1 \end{cases}$$

Proof: ◦ We use the conditional variance formula:

$$\begin{aligned} \mathrm{Var}[Z_h] &= \mathrm{Var}\big[\mathbb{E}[Z_h|\sigma_0]\big] + \mathbb{E}\big[\mathrm{Var}[Z_h|\sigma_0]\big] \\ &= \mathrm{Var}[\sigma_0] + \tfrac{1}{2}\mathrm{Var}[Z_h|\sigma_0=+] + \tfrac{1}{2}\mathrm{Var}[Z_h|\sigma_0=-] \\ &= \mathrm{Var}[\sigma_0] + \mathrm{Var}_h^+[Z_h] \qquad \text{(using symmetry)} \\ &= 1 + \mathrm{Var}_h^+[Z_h]. \qquad (*) \end{aligned}$$

◦ Now let $Z_h = Z_h^{(1)} + Z_h^{(2)}$ (the decomposition corresponds to subtrees T_1, T_2 respectively)



we get $\mathrm{Var}[Z_h] = 1 + \mathrm{Var}_h^+[Z_h]$ (by the above)

$= 1 + 2\,\mathrm{Var}_h^+[Z_h^{(1)}]$ (since, conditioning on the root state, $Z_h^{(1)}$ and $Z_h^{(2)}$ are independent and are identically distributed)

$= 1 + 2\left(\mathbb{E}_h^+\left[\left(Z_h^{(1)}\right)^2\right] - \left(\mathbb{E}_h^+\left[Z_h^{(1)}\right]\right)^2\right).$

- Note that $\mathbb{E}_h^+\left[Z_h^{(1)}\right] = \frac{1}{2}$

- Moreover:

$$\mathbb{E}_h^+\left[\left(Z_h^{(1)}\right)^2\right] = (1-p)\frac{1}{(2\theta)^2}\mathbb{E}_{h-1}^+\left[Z_{h-1}^2\right] + p\frac{1}{(2\theta)^2}\mathbb{E}_{h-1}^-\left[Z_{h-1}^2\right]$$

$$= \frac{1}{(2\theta)^2}\mathbb{E}_{h-1}^+\left[Z_{h-1}^2\right].$$

using symmetry

⇝ since:
- w/prob (1-p): $2\theta \cdot Z_h^{(1)}$ has distribution μ_{h-1}^+
- w/prob p: $2\theta Z_h^{(1)}$ has distribution μ_{h-1}^-

- Therefore:

$$\mathrm{Var}\left[Z_h\right] = 1 + \frac{1}{2\theta^2}\mathbb{E}_{h-1}^+\left[Z_{h-1}^2\right] - \frac{1}{2}$$

$$= \frac{1}{2} + \frac{1}{2\theta^2}\left[\mathrm{Var}_{h-1}^+\left[Z_{h-1}\right] + \left(\mathbb{E}_{h-1}^+\left[Z_{h-1}\right]\right)^2\right]$$

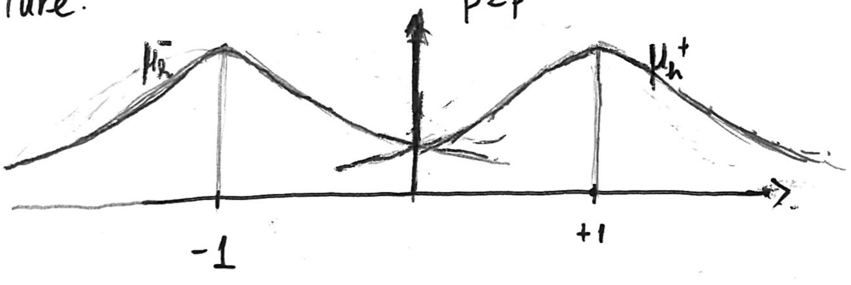
$$= \frac{1}{2} + \frac{1}{2\theta^2}\left[\mathrm{Var}_{h-1}^+\left[Z_{h-1}\right] + 1\right]$$

$$\overset{(*)}{=} \frac{1}{2} + \frac{1}{2\theta^2}\mathrm{Var}\left[Z_{h-1}\right].$$

Solving the recursion gives the result.

☒

Picture:



Proof of Reconstruction Theorem ($p<p^*$):

- Let $\hat{\mu}_h$ be the distn' of Z_h, $\hat{\mu}_h^+$ the distn' of Z_h conditioning on the root being a $+$ and $\hat{\mu}_h^-$ the distn' of Z_h conditioning in the root being a $-$.

- **Claim:**

$$\|\hat{\mu}_h^+ - \hat{\mu}_h^-\|_{TV} \leq \|\mu_h^+ - \mu_h^-|_{TV}$$

Proof: ex 0.5 pt

- So suffices to lower bound $\|\hat{\mu}_h^+ - \hat{\mu}_h^-\|_{TV}$, to show solvability for $p<p^*$.

- We have:

using: $\dfrac{|\hat{\mu}_h^+(z) - \hat{\mu}_h^-(z)|}{2\hat{\mu}_h(z)} \leq 1$, which follows from $\hat{\mu}_h(z) = \frac{1}{2}\hat{\mu}^+(z) + \frac{1}{2}\hat{\mu}^-(z)$

$$\sum_z |\hat{\mu}_h^+(z) - \hat{\mu}_h^-(z)| \geq 2 \sum_z \left(\frac{\hat{\mu}_h^+(z) - \hat{\mu}_h^-(z)}{2\hat{\mu}_h(z)}\right)^2 \hat{\mu}_h(z)$$

$$\geq 2 \frac{\left(\sum_z z\left(\frac{\hat{\mu}_h^+(z) - \hat{\mu}_h^-(z)}{2\hat{\mu}_h(z)}\right)\hat{\mu}_h(z)\right)^2}{\sum_z z^2 \hat{\mu}_h(z)}$$

[Cauchy-Schwartz]

$$= \frac{1}{2} \frac{\left(\mathbb{E}_h^+[Z_h] - \mathbb{E}_h^-[Z_h]\right)^2}{\mathrm{Var}[Z_h]}$$

$$\geq 4\left(1 - (2\theta^2)^{-1}\right) > 0$$

☒