Lecture 12 : Ancestral Reconstruction by Majority

MATH285K - Spring 2010 *Lecturer: Sebastien Roch*

References: [EKPS00, Mos01, MP03, BCMR06].

Previous class

DEF 12.1 (CFN Model) *A CFN model is an MCT $(\mathcal{T}, \mathcal{P}, \mu_\rho)$ on $C = \{0, 1\}$ with symmetric transition matrices with positive determinant and uniform μ_ρ. In particular, each transition matrix $P^e = \bar{P}^e$ is characterized by a single parameter $0 < p_e < 1/2$, the mutation probability along edge e.*

1 Ancestral reconstruction

For simplicity, we begin by considering a special case. Let $T^{(\infty)}$ be the infinite complete binary tree where the root is denoted by 0. For $h \geq 0$, let $\mathcal{T}^{(h)} = (T^{(h)}, \phi^{(h)})$ with $T^{(h)} = (V^{(h)}, E^{(h)})$ be the first h levels of $T^{(\infty)}$ starting from the root where the leaves are labeled by $[2^h]$ (say, from left to right in a natural planar embedding). In particular, the tree $\mathcal{T}^{(0)}$ is simply the root. For $0 < p < 1/2$, we denote by $(\mathcal{T}^{(h)}, p)$ the CFN model on $\mathcal{T}^{(h)}$ with state space $C = \{+1, -1\}$ where all edge mutation probabilities are fixed to p. We denote by $\boldsymbol{\sigma}_V = \{\sigma_v\}_{v \in V^{(h)}}$ the vector of states of a sample from $(\mathcal{T}^{(h)}, p)$. With a sligh abuse of notation, we let $\boldsymbol{\sigma}_h = \{\sigma_\ell\}_{\ell \in [2^h]}$ be the vector of states at the leaves and we denote by μ_h the distribution of $\boldsymbol{\sigma}_h$.

Recall that, under the CFN model, the root state σ_0 is assumed to be uniform in $\{+1, -1\}$. The ancestral reconstruction problem consists in trying to guess the value at the root σ_0 given the states $\boldsymbol{\sigma}_h$ at level h. We first note that in general we cannot expect an arbitrarily good estimator. Indeed, re-writing the transition matrix in its *random cluster* form

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

we see that the states $\boldsymbol{\sigma}_1$ at the first level are completely randomized (i.e., independent of σ_0) with probabiltity $(2p)^2$—in which case we cannot hope to reconstruct the root state better than a coin flip. Intuitively, the ancestral reconstruction problem is solvable if we can find an estimator of the root state which outperforms a random coin flip even as the tree grows to ∞.

Formally:

DEF 12.2 (Ancestral reconstruction solvability) *Let μ_h^+ be the distribution μ_h conditioned on the root state σ_0 being $+1$, and similarly for μ_h^-. We say that the ancestral reconstruction problem (under the CFN model) for $0 < p < 1/2$ is solvable if*

$$\liminf_h \|\mu_h^+ - \mu_h^-\|_1 > 0,$$

otherwise the problem is unsolvable. *Recall that*

$$\|\mu_h^+ - \mu_h^-\|_1 \equiv \sum_{\mathbf{s}_h \in \{+1,-1\}^h} |\mu_h^+(\mathbf{s}_h) - \mu_h^-(\mathbf{s}_h)|.$$

To see the connection with the description above, consider an arbitrary root estimator $\hat{\sigma}_0$. Then the probability of a mistake is

$$\begin{aligned}\mathbb{P}[\hat{\sigma}_0(\mathbf{s}_h) \neq \sigma_0] &= \frac{1}{2} \sum_{\mathbf{s}_h \in \{+1,-1\}^h} \mu_h^-(\mathbf{s}_h) \mathbb{1}\{\hat{\sigma}_0(\mathbf{s}_h) = +1\} \\ &\quad + \frac{1}{2} \sum_{\mathbf{s}_h \in \{+1,-1\}^h} \mu_h^+(\mathbf{s}_h) \mathbb{1}\{\hat{\sigma}_0(\mathbf{s}_h) = -1\}\end{aligned}$$

This expression is minimized by choosing

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

This is simply the ML estimator which we will denote by $\hat{\sigma}_0^{\mathrm{ML}}$.

Now note that

$$\begin{aligned}\mathbb{P}[\hat{\sigma}_0(\mathbf{s}_h) = \sigma_0] - \mathbb{P}[\hat{\sigma}_0(\mathbf{s}_h) \neq \sigma_0] &= \frac{1}{2} \sum_{\mathbf{s}_h \in \{+1,-1\}^h} \mu_h^+(\mathbf{s}_h) \hat{\sigma}_0^{\mathrm{ML}}(\mathbf{s}_h) \\ &\quad - \frac{1}{2} \sum_{\mathbf{s}_h \in \{+1,-1\}^h} \mu_h^-(\mathbf{s}_h) \hat{\sigma}_0^{\mathrm{ML}}(\mathbf{s}_h) \\ &= \frac{1}{2} \sum_{\mathbf{s}_h \in \{+1,-1\}^h} |\mu_h^+(\mathbf{s}_h) - \mu_h^-(\mathbf{s}_h)| \\ &= \frac{1}{2} \|\mu_h^+ - \mu_h^-\|_1,\end{aligned}$$

where the second line comes from

$$|a-b| = (a-b)\mathbb{1}\{a \geq b\} + (b-a)\mathbb{1}\{a < b\}.$$

2 Majority

It turns out that the accuracy of the ML estimator undergoes a phase transition at a critical p_* mutation probability.

THM 12.3 (Solvability) *Let $\theta_* = 1 - 2p_* = 1/\sqrt{2}$. Then when $p \leq p_*$ the ancestral reconstrution problem is solvable.*

Rather than analyzing maximum likelihood, we look at a simpler estimator first. We come back to the proof of Theorem 12.3 in the next section. The *majority* at level h is defined as

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

where

$$\theta = 1 - 2p.$$

The normalization in Z_h turns it into an unbiased estimator:

THM 12.4 (Unbiasedness) *Denoting by $\mathbb{E}_h^+$ the expectation operator under μ_h^+, and similarly for $\mathbb{E}_h^-$, we have*

$$\mathbb{E}_h^+[Z_h] = +1, \qquad \mathbb{E}_h^-[Z_h] = -1.$$

Proof: By applying the Markov transition matrix on the first level,

$$\begin{aligned} \mathbb{E}_h^+[\sigma_1] &= (1-p)\mathbb{E}_{h-1}^+[\sigma_1] + p\mathbb{E}_{h-1}^-[\sigma_1] \\ &= (1-2p)\mathbb{E}_{h-1}^+[\sigma_1], \end{aligned}$$

where the second line follows from the $+1/-1$ symmetry. By iteration,

$$\mathbb{E}_h^+[\sigma_1] = \theta^h,$$

from which the result follows by linearity. ■

To locate the phase transition, we compute the variance of Z_h.

THM 12.5 (Phase transition for majority) *We have*

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

Proof: By the conditional variance formula

$$\begin{aligned} \mathrm{Var}[Z_h] &= \mathrm{Var}[\mathbb{E}[Z_h \,|\, \sigma_0]] + \mathbb{E}[\mathrm{Var}[Z_h \,|\, \sigma_0]] \\ &= \mathrm{Var}[\sigma_0] + \mathbb{E}[\mathrm{Var}[Z_h \,|\, \sigma_0]] \\ &= 1 + \mathrm{Var}_h^+[Z_h], \end{aligned}$$

where the last line follows from symmetry with Var_h^+ being the conditional variance at level h given that the root is $+1$. Writing $Z_h = Z_h^{(1)} + Z_h^{(2)}$ as a sum over the two subtrees below the root and using the conditional independence of these two subtrees given the root state we get

$$\begin{aligned} \mathrm{Var}[Z_h] &= 1 + 2\mathrm{Var}_h^+[Z_h^{(1)}] \\ &= 1 + 2(\mathbb{E}_h^+[(Z_h^{(1)})^2] - (\mathbb{E}_h^+[Z_h^{(1)}])^2). \end{aligned}$$

Using $\mathbb{E}_h^+[Z_h^{(1)}] = 1/2$ and applying the Markov transition matrix on the first level and re-normalizing $Z_h^{(1)}$, we get

$$\begin{aligned} \mathrm{Var}[Z_h] &= 1 - 2(\mathbb{E}_h^+[Z_h^{(1)}])^2 + 2\mathbb{E}_h^+[(Z_h^{(1)})^2] \\ &= 1 - 1/2 + 2[(1-p)(2\theta)^{-2}\mathbb{E}_{h-1}^+[Z_{h-1}^2] + p(2\theta)^{-2}\mathbb{E}_{h-1}^-[Z_{h-1}^2]] \\ &= 1/2 + (2\theta^2)^{-1}\mathbb{E}_{h-1}^+[Z_{h-1}^2] \\ &= 1/2 + (2\theta^2)^{-1}\mathrm{Var}[Z_{h-1}], \end{aligned} \tag{1}$$

where we used that

$$\mathrm{Var}[Z_{h-1}] = \mathbb{E}[Z_{h-1}^2] = \mathbb{E}_{h-1}^+[Z_{h-1}^2] = \mathbb{E}_{h-1}^-[Z_{h-1}^2],$$

by symmetry and the fact that $\mathbb{E}[Z_{h-1}] = 0$. Solving the affine recursion (1) gives the result. ■

3 Solvability

In essence Theorem 12.5 says that majority is a useful root estimator when $2\theta^2 > 1$, that is, when $p < p_*$. (The proof below and a correlation inequality proved in [EKPS00, Theorem 1.4] gives a lower bound on the probability of reconstruction of majority. We leave the details to the reader.) We can now prove Theorem 12.3.

Proof:(of Theorem 12.3) Let $\bar{\mu}_h$ be the dsitribution of Z_h and define $\bar{\mu}_h^+$ and $\bar{\mu}_h^-$ similarly. We give a bound on $\|\mu_h^+ - \mu_h^-\|_1$ through a bound on $\|\bar{\mu}_h^+ - \bar{\mu}_h^-\|_1$.

Indeed, letting $\bar{\mathbf{s}}_h$ be the majority estimator applied to $\mathbf{s}_h \in \{+1,-1\}$,

$$\begin{aligned}\sum_z |\bar{\mu}_h^+(z) - \bar{\mu}_h^-(z)| &= \sum_z \left|\sum_{\mathbf{s}_h:\bar{\mathbf{s}}_h = z} (\mu_h^+(\mathbf{s}_h) - \mu_h^-(\mathbf{s}_h))\right| \\ &\leq \sum_z \sum_{\mathbf{s}_h:\bar{\mathbf{s}}_h = z} |\mu_h^+(\mathbf{s}_h) - \mu_h^-(\mathbf{s}_h)| \\ &= \sum_{\mathbf{s}_h} |\mu_h^+(\mathbf{s}_h) - \mu_h^-(\mathbf{s}_h)|.\end{aligned}$$

To lower bound $\|\bar{\mu}_h^+ - \bar{\mu}_h^-\|_1$, we apply Cauchy-Schwarz and use the variance bound in Theorem 12.5. Note that $\frac{1}{2}\bar{\mu}_h^+ + \frac{1}{2}\bar{\mu}_h^- = \bar{\mu}_h$ so that

$$\frac{|\bar{\mu}_h^+(z) - \bar{\mu}_h^-(z)|}{2\bar{\mu}_h(z)} \leq 1,$$

and we get

$$\begin{aligned}\sum_z |\bar{\mu}_h^+(z) - \bar{\mu}_h^-(z)| &\geq 2\sum_z \left(\frac{|\bar{\mu}_h^+(z) - \bar{\mu}_h^-(z)|}{2\bar{\mu}_h(z)}\right)^2 \bar{\mu}_h(z) \\ &\geq 2\frac{\left(\sum_z z\left(\frac{\bar{\mu}_h^+(z) - \bar{\mu}_h^-(z)}{2\bar{\mu}_h(z)}\right)\bar{\mu}_h(z)\right)^2}{\sum_z z^2 \bar{\mu}_h(z)} \\ &= \frac{1}{2}\frac{(\mathbb{E}_h^+[Z_h] - \mathbb{E}_h^-[Z_h])^2}{\mathrm{Var}[Z_h]} \\ &\geq 4(1 - (2\theta^2)^{-1}) \\ &> 0.\end{aligned}$$

■

Further reading

Most of the material discussed here (and much more) can be found in [EKPS00]. See also [Mos01, MP03, BCMR06] for further results.

References

[BCMR06] Christian Borgs, Jennifer T. Chayes, Elchanan Mossel, and Sébastien Roch. The Kesten-Stigum reconstruction bound is tight for roughly symmetric binary channels. In *FOCS*, pages 518–530, 2006.

[EKPS00] W. S. Evans, C. Kenyon, Y. Peres, and L. J. Schulman. Broadcasting on trees and the Ising model. *Ann. Appl. Probab.*, 10(2):410–433, 2000.

[Mos01] E. Mossel. Reconstruction on trees: beating the second eigenvalue. *Ann. Appl. Probab.*, 11(1):285–300, 2001.

[MP03] E. Mossel and Y. Peres. Information flow on trees. *Ann. Appl. Probab.*, 13(3):817–844, 2003.