

Lecture 11 : Asymptotic Sample Complexity

MATH285K - Spring 2010

Lecturer: Sebastien Roch

References: [DMR09].

Previous class

THM 11.1 (Strong Quartet Evidence) *Let \mathcal{Q} be a collection of quartet trees on X such that for all S with $|S| = 4$ at most one quartet tree of \mathcal{Q} has label set S . Then there is a phylogenetic tree \mathcal{T} on X such that $\Sigma(\mathcal{Q}) \cup \Sigma_X^0 = \Sigma(\mathcal{T})$.*

Recall that the log-det distance is given by:

DEF 11.2 (Logdet Distance) *For $a, b \in X$, let P^{ab} be defined as follows*

$$\forall \alpha, \beta \in C, P_{\alpha, \beta}^{ab} = \mathbb{P}[\Xi_{\phi(b)} = \beta \mid \Xi_{\phi(a)} = \alpha].$$

The logdet distance between a and b is the dissimilarity map

$$\delta(a, b) = -\frac{1}{2} \log \det[P^{ab} P^{ba}].$$

In the case of the CFN model, we have

$$P^{ab} = \begin{pmatrix} 1 - p^{ab} & p^{ab} \\ p^{ab} & 1 - p^{ab} \end{pmatrix}$$

where p^{ab} is probability that the states at a and b differ—in particular, $p^{ab} = p^{ba}$. Hence

$$-\frac{1}{2} \log \det[P^{ab} P^{ba}] = -\frac{1}{2} \log[\det(P^{ab})^2] = -\log[1 - 2p^{ab}].$$

For $q = ab|cd$, define

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

Recall that, if δ is a tree metric (as is the case for the log-det distance), then among all 4-tuples over $X' = \{a, b, c, d\}$ $\delta(q)$ takes three possible values

$$\delta(q) \in \{w_{e_0}, 0, -w_{e_0}\}, \quad (1)$$

where e_0 is the middle edge of $\mathcal{T}|X'$.

Consider the following algorithm.

- For all $a, b \in X$ distinct, let

$$\hat{p}^{ab} = \frac{1}{k} \sum_{i=1}^k \mathbb{1}\{\Xi_a^i \neq \Xi_b^i\}.$$

and

$$\hat{\delta}(a, b) = -\log[1 - 2\hat{p}^{ab}].$$

(Make the last quantity $+\infty$ if the term inside the log is negative.)

- Set $\mathcal{Q} = \emptyset$.
- For all $a, b, c, d \in X$ distinct,
 - Setting $X' = \{a, b, c, d\}$, let

$$xy|wz = \arg \max\{\hat{\delta}(xy|wz) : x, y, w, z \in X' \text{ distinct}\},$$

where $\hat{\delta}(q)$ is defined similarly to $\delta(q)$ above.

- Add $xy|wz$ to \mathcal{Q} .

- Apply the Strong Quartet Evidence algorithm to \mathcal{Q} to recover \mathcal{T} .

Clearly, by (1), if we were to run the algorithm above with δ rather than $\hat{\delta}$, the correct tree \mathcal{T} would be reconstructed. However, $\hat{\delta}$ is only an approximation of δ . By the strong law of large numbers, this approximation gets arbitrarily better as $k \rightarrow \infty$ with probability 1. The consistency of the algorithm then follows from the following inequality:

$$\max_q |\delta(q) - \hat{\delta}(q)| \leq 2 \max_{a,b} |\delta(a, b) - \hat{\delta}(a, b)| < \frac{1}{2} \min_e w_e \equiv \frac{1}{2} w_*, \quad (2)$$

with probability 1 for all sufficiently large k .

1 Concentration in the CFN Model

Basic definition. Statistical consistency is a coarse property which does not allow to distinguish different inference methods very well. A more quantitative comparison between methods can be obtained from the following concept. For simplicity, we restrict ourselves to estimating the tree.

DEF 11.3 (Asymptotic Sample Complexity (ASC)) Fix $\delta > 0$. Let

$$\Xi = \{\Xi_X^1, \dots\},$$

be a sequence of i.i.d. samples generated by a CFN model $(\mathcal{T}, \mathcal{P})$. A sequence of estimators $\{\hat{\mathcal{T}}_k\}_{k \geq 0}$ of \mathcal{T} , where $\hat{\mathcal{T}}_k$ is based on k samples, has asymptotic sample complexity (ASC) at confidence level δ (at most) k_0 if for all $k \geq k_0$ the probability that $\hat{\mathcal{T}}_k = \mathcal{T}$ is at least δ .

Typically, the ASC is expressed as an asymptotic expression of structural parameters of the model such as the number of leaves n , the shortest branch length f , or the diameter of the tree.

Concentration bound. The law of large numbers itself is not enough to prove ASC results. Rather we need concentration inequalities such as Chernoff's bound. We give a proof for completeness.

THM 11.4 (Chernoff's bound) Let X be a binomial with parameters n and p . Then, for all $t > 0$

$$\mathbb{P}[|X - np| \geq t] \leq 2e^{-t^2/(2n)}.$$

Proof: Recall the following easy inequality.

LEM 11.5 (Markov's inequality) If $X \geq 0$ with finite mean then

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t},$$

for all $t > 0$.

Proof: Note that

$$\mathbb{E}[X] \geq \mathbb{E}[X \mathbb{1}\{X \geq t\}] \geq t\mathbb{P}[X \geq t].$$

■

Write $X - np$ as a centered iid sum

$$X - np = \sum_{i \leq n} Y_i,$$

in the obvious way. By Markov and independence,

$$\begin{aligned}
 \mathbb{P}[X - np \geq t] &= \mathbb{P}[\exp(h(X - np)) \geq \exp(ht)] \\
 &\leq \exp(-ht) \mathbb{E}[\exp(h(X - np))] \\
 &= \exp(-ht) \mathbb{E}[\exp(hY_n + h \sum_{i \leq n-1} Y_i)] \\
 &= \exp(-ht) \mathbb{E}[\exp(hY_n)] \mathbb{E}[\exp(h \sum_{i \leq n-1} Y_i)] \\
 &= \exp(-ht) \mathbb{E}[\exp(hY_1)]^n.
 \end{aligned}$$

LEM 11.6 Assume $\mathbb{E}[Y] = 0$ and $|Y| \leq 1$. Then, for all h

$$\mathbb{E}[\exp(hY)] \leq \exp(h^2/2).$$

Proof: By convexity

$$e^{hy} \leq \frac{1-y}{2} e^{-h} + \frac{1+y}{2} e^h,$$

for $|y| \leq 1$. By Taylor expansion (check!),

$$\mathbb{E}[\exp(hY)] \leq \frac{1}{2}(e^{-h} + e^h) \leq e^{h^2/2}.$$

Choose $h = t/n$ and apply the previous lemma to Y_1 . ■

ASC of Distance Methods. We apply the previous bound to the estimation of distances. For simplicity, we state the result for binary phylogenetic trees, although this is not necessary. ■

THM 11.7 (ASC of Distance Methods) Let

$$w_* = \min_e w(e)$$

and

$$W_* = \max_{a,b} \delta(a, b).$$

Then, the algorithm above recovers the correct tree with probability $1 - o(1)$ as $n \rightarrow \infty$ with

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

Proof: Assume that for all $a, b \in X$

$$|p^{ab} - \hat{p}^{ab}| < \varepsilon.$$

For (2) to hold for all pairs of leaves, it must be that

$$\begin{aligned} \frac{1}{4}w_* &> -\log(1 - 2(p^{ab} + \varepsilon)) + \log(1 - 2p^{ab}) \\ &= -\log\left(\frac{1 - 2(p^{ab} + \varepsilon)}{1 - 2p^{ab}}\right) \\ &= -\log\left(1 - \frac{2\varepsilon}{1 - 2p^{ab}}\right), \end{aligned}$$

and similarly for the other direction. Rewriting this equation, it is enough that

$$\begin{aligned} \varepsilon &< \frac{1}{2}(1 - e^{-w_*/4})e^{-W_*} \equiv \mathcal{W}_* \\ &\leq \frac{1}{2}(1 - e^{-w_*/4})(1 - 2p^{ab}). \end{aligned}$$

Plugging this expression into Chernoff's bound with $t = \mathcal{W}_*k$ gives a probability of failure

$$\leq 2 \exp\left(-\frac{\mathcal{W}_*^2 k}{2}\right),$$

which will be $\leq 1/n^3$ (so that we can apply a union bound over all pairs of leaves) if

$$k = O\left(\frac{1}{\mathcal{W}_*^2} \log n\right) = O\left(\frac{e^{2W_*}}{(1 - e^{-w_*/4})^2} \log n\right).$$

■

2 Depth v. Diameter

It turns out that Theorem 11.7 is not tight. In particular, the dependence of k in the *weighted diameter* W_* can be replaced by the *weighted depth* using a more sophisticated algorithm:

DEF 11.8 (Weighted Depth) *The depth of an edge e is the length (under δ) of the shortest path between two leaves crossing e . The depth of a tree is the maximum edge depth.*

In general, the depth can be much smaller than the diameter. Assume all branch lengths are 1 (that is, consider the graphical distance). Then on the caterpillar tree, the diameter is $O(n)$ while the depth is $O(1)$. In fact, under the graphical distance, the depth is always at most $2 \log_2 n + 2$. Indeed, if the depth of an edge e was $2 \log_2 n + 3$ then the path to the closest leaf on one side of e would be at least $\log_2 n + 1$ which would imply that the number of leaves on that side of e would exceed n —a contradiction.

For details, see e.g. [DMR09].

References

- [DMR09] Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Phylogenies without branch bounds: Contracting the short, pruning the deep. In *RECOMB*, pages 451–465, 2009.