Notes 10 : Asymptotic Sample Complexity

MATH 833 - Fall 2012 *Lecturer: Sebastien Roch*

References: [DMR09].

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 10.1 (Asymptotic Sample Complexity (ASC)) *Fix $\delta > 0$. Let*

$$\mathbf{\Xi} = \{\Xi_X^1, \ldots\},$$

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 10.2 (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 10.3 (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 10.4 *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 10.5 (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* $n \to \infty$ *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

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

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

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

■

Yule case. For random trees, one can obtain bounds on the diameter.

Branching processes are commonly used to model species phylogenies. In the continuous-time Yule process (or pure-birth process), one starts with two species (representing the two branches emanating from the root). At any given time, each species generates a new offspring at rate $0 < \nu < +\infty$. We stop the process when the number of species is exactly $n+1$ (and ignore the $n+1$st species). This process generates a species phylogeny with n leaves and branch lengths given by the inter-speciation times in the above process.

Let Z_i be the $(i-1)$-th inter-speciation time. As a minimum of i independent exponential distributions with mean $1/\nu$, Z_i is an exponential with mean $1/(i\nu)$. Moreover the Z_is are independent. Hence the height of the phylogeny in time units, that is, the total time until $n+1$ species are present (recall that we ignore the $(n+1)$-st species) is

$$\mathbf{Z} = \sum_{i=2}^{n+1} Z_i,$$

and we have

$$\mathbb{E}[\mathbf{Z}] = \sum_{i=2}^{n+1}\mathbb{E}[Z_i] = \sum_{i=2}^{n+1}\frac{1}{i\nu} = \Theta(\nu^{-1}\log n),$$

and

$$\mathrm{Var}[\mathbf{Z}] = \sum_{i=2}^{n+1}\mathrm{Var}[Z_i] = \sum_{i=2}^{n+1}\frac{1}{i^2\nu^2} = \Theta(\nu^{-2}).$$

By Chebyshev's inequality,

$$\mathbb{P}[\mathbf{Z} \geq C_1 \log n] \leq \frac{C_2}{C_3 \log^2 n} \to 0,$$

for appropriately chosen Cs not depending on n.

2 Depth v. Diameter

It turns out that Theorem 10.5 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 10.6 (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 $O(n)$ 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 $2\log_2 +3$ then the path to the closest leaf on one side of e would be at least $\log_2 +1$ 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.