Notes 14 : Steel's conjecture

MATH 833 - Fall 2012 *Lecturer: Sebastien Roch*

References: [Mos03], [Roc10].

1 Upper bound

Steel's conjecture asserts that ancestral reconstruction and phylogenetic tree reconstruction are closely related: when the ancestral reconstruction is solvable, it should as easy to build a deep tree as it is to build a shallow tree. A more quantitative form of the conjecture is that the dependence on the depth in tha asymptotic sample complexity results we discussed in a previous lecture should disappear when all branch lengths are below the critical threshold for ancestral recontruciton. The conjecure has been proved in many important cases, starting with the work of Mossel [Mos04].

An illustration of Steel's conjecture. Here we illustrate the conjecture on an example. We then discuss a lower bound in the next section.

EX 14.1 *Fix a rate matrix Q on C with stationary distribution π. Let $(\mathcal{T}, \{w_e\}_{e\in E})$ be a rooted binary phylogenetic tree such that the corresponding metric δ is ultrametric and all edge weights satisfy*

$$0 < f \leq w_e \leq g < g_*.$$

Suppose that we know the tree and branch lengths except for the top triplet: we are given three subtrees $\mathcal{T}_a$, $\mathcal{T}_b$, $\mathcal{T}_c$ rooted respectively at a, b, and c such that one of $X_t = \{a, b, c\}$ is a child of the root and the other two are grandchildren of the root. Our goal is to reconstruct the top triplet t from the partial information given and k samples $\{\Xi_X^i\}_{i=1}^k$ at the leaves $X = [n]$.

In a previous lecture, we gave an algorithm which (after being adapted for the setup above) reconstructs the top triplet with asymptotic sample complexity scaling exponentially in the depth of the tree, here $O(\log n)$ (where the constant depends on f and g). That is, the asymptotic sample complexity grows roughly as a polynomial of n.

A more accurate reconstruction algorithm. As before, let ν be the second right eigenvector of Q and let $\{\sigma_X^i\}_{i=1}^k$ be the samples mapped to ν. For $v \in V$, let $\mathcal{H}(v)$ be the weighted height of v, that is, the weighted distance between v and the leaves below it. For $u, v \in V$, let $\mathcal{A}(u, v)$ be the most recent common ancestor of u and v. To reconstruct the top triplet, it suffices to determine

$$(u, v) = \arg\min\{H(u, v) \equiv \mathcal{H}(\mathcal{A}(u, v)) \,:\, u \neq v \in X_t\}.$$

Let X_a, X_b, and X_c be the leaves below a, b, and c respectively. For any unit flows μ^u and μ^v on $\mathcal{T}_u$ and $\mathcal{T}_v$ respectively, the estimator

$$\hat{H}_{\mu^u,\mu^v}(u, v) = \sum_{x\in X_u, y\in X_v} \mu_x^u \mu_y^v \left(\frac{1}{k}\sum_{i=1}^k \sigma_x^i \sigma_y^i\right),$$

is unbiased for $e^{-2H(u,v)}$. So we seek the pair (u, v) in X_t which maximizes the latter estimator. To see the connection with ancestral reconstruction note that

$$\hat{H}_{\mu^u,\mu^v}(u, v) = e^{-\mathcal{H}(u)-\mathcal{H}(v)} \frac{1}{k}\sum_{i=1}^k \left(\sum_{x\in X_u} \frac{\mu_x^u \sigma_x^i}{e^{-\mathcal{H}(u)}}\right)\left(\sum_{y\in X_v} \frac{\mu_y^v \sigma_y^i}{e^{-\mathcal{H}(v)}}\right),$$

so that we are implicitly reconstructing ancestral sequences at u and v. We can compute the variance of $\hat{H}_{\mu^u,\mu^v}(u, v)$ in the case of the uniform flow. Note that, by independence,

$$\begin{aligned}
&\mathrm{Var}\left[e^{\mathcal{H}(u)+\mathcal{H}(v)} \hat{H}_{\mu^u,\mu^v}(u, v)\right] \\
&\quad = \frac{1}{k}\mathrm{Var}\left[\left(\sum_{x\in X_u} \frac{\mu_x^u \sigma_x^1}{e^{-\mathcal{H}(u)}}\right)\left(\sum_{y\in X_v} \frac{\mu_y^v \sigma_y^1}{e^{-\mathcal{H}(v)}}\right)\right] \\
&\quad \leq \frac{1}{k}\mathbb{E}\left[\left(\sum_{x\in X_u} \frac{\mu_x^u \sigma_x^1}{e^{-\mathcal{H}(u)}}\right)^2 \left(\sum_{y\in X_v} \frac{\mu_y^v \sigma_y^1}{e^{-\mathcal{H}(v)}}\right)^2\right] \\
&\quad \leq \frac{1}{k}\mathbb{E}\left[\mathbb{E}\left[\left(\sum_{x\in X_u} \frac{\mu_x^u \sigma_x^1}{e^{-\mathcal{H}(u)}}\right)^2 \left(\sum_{y\in X_v} \frac{\mu_y^v \sigma_y^1}{e^{-\mathcal{H}(v)}}\right)^2 \,\middle|\, \Xi_u^1, \Xi_v^1\right]\right] \\
&\quad \leq \frac{1}{k}(\pi_{\min}^{-1}\mathcal{V})^2,
\end{aligned}$$

where we used conditional independence and the fact that

$$\mathrm{Var}[Z_\mu] = \mathbb{E}[Z_\mu^2] = \sum_{\alpha\in C} \pi_\alpha \mathbb{E}[Z_\mu^2 \,|\, \Xi_\rho = \alpha],$$

so that $\mathbb{E}[Z_\mu^2 \,|\, \Xi_\rho] \leq \pi_{\min}^{-1}\mathcal{V}$ where $\pi_{\min} = \min_\alpha \pi_\alpha$. Recall Chebyshev's inequality:

LEM 14.2 (Chebyshev's Inequality) *Let X be a real random variable with finite second moment. Then for all $\alpha > 0$*

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq \psi] \leq \frac{\mathrm{Var}[X]}{\psi^2}.$$

Suppose we seek an error proability not exceeding $\varepsilon > 0$ (say, $\varepsilon = 0.01$) and that (a, b) is the minimizing pair. Then applying Chebyshev's inequality, we get

$$\begin{aligned}
&\mathbb{P}[\hat{H}_{\mu^a,\mu^b}(a,b) \leq e^{-2(H(a,b)+f/2)}] \\
&\quad \leq \mathbb{P}[|\hat{H}_{\mu^a,\mu^b}(a,b) - e^{-2H(a,b)}| \geq e^{-2H(a,b)}(1 - e^{-f/2})] \\
&\quad \leq \frac{\frac{1}{k}(\pi_{\min}^{-1}\mathcal{V})^2 e^{-2\mathcal{H}(a)-2\mathcal{H}(b)}}{e^{-4H(a,b)}(1-e^{-f/2})^2} \\
&\quad \leq \frac{\frac{1}{k}(\pi_{\min}^{-1}\mathcal{V})^2 e^{2\delta(a,b)}}{(1-e^{-f/2})^2} \\
&\quad \leq \frac{\frac{1}{k}(\pi_{\min}^{-1}\mathcal{V})^2 e^{4g}}{(1-e^{-f/2})^2} \\
&\quad \leq \frac{\varepsilon}{3},
\end{aligned}$$

if $k = \Omega_{f,g}(\varepsilon^{-1})$ and similarly for the other two pairs. Note that the latter does not depend on the depth of the tree.

2 Mossel's gedanken experiment

To show that the asymptotic sample complexity of any reconstruction algorithm must depend on the depth of the tree when edge lengths are such that the ancestral reconstruction problem is not solvable, we consider a simple thought experiment (again, in a special case). Consider again the setup of the previous section, but this time assume that Q is the CFN rate matrix, that $\mathcal{T}_a$, $\mathcal{T}_b$, and $\mathcal{T}_c$ are complete binary trees with edges lengths $g > g_*$ and depth H. Assume further that the top triplet is chosen uniformly between $t_1 = ab|c$ and $t_2 = ac|b$ and that the two closest leaves are at distance $2g$ from each other and from the root. From sequences at the leaves $\{\sigma_X^i\}_{i=1}^k$ we seek to infer whether t_1 or t_2 was used to generate the data.

We will use the mutual information.

DEF 14.3 *Let Y, Z be random variables with state space S_Y, S_Z. The mutual information between Y and Z is*

$$I(Y,Z) = \sum_{y \in S_Y, z \in S_Z} \mathbb{P}[Y=y, Z=z] \log \frac{\mathbb{P}[Y=y, Z=z]}{\mathbb{P}[Y=y]\mathbb{P}[Z=z]}.$$

The mutual information has the following useful properties. See e.g. [CT91].

LEM 14.4 *If W and Z are conditionally independent given Y, then*

$$I(W,Z) \leq I(Y,Z),$$

$$I((W,Y),Z) = I(Y,Z),$$

and

$$I((W,Z),Y) \leq I(W,Y) + I(Z,Y).$$

LEM 14.5 ([EKPS00]) *If Y is uniform in $\{1,2\}$ and μ_Z^1, μ_Z^2 are the conditional distributions of Z given Y, then*

$$\frac{1}{2}\|\mu_Z^1 - \mu_Z^2\|_1^2 \leq I(Y,Z) \leq \|\mu_Z^1 - \mu_Z^2\|_1.$$

We have already shown that

$$\|\mu_H^+ - \mu_H^-\|_1 \leq 2\sqrt{(2\theta^2)^H},$$

for $g = -\ln\theta$. (Recall that our assumption implies $2\theta^2 < 1$.) Hence, using Lemma 14.4, we get

$$\begin{aligned} I(t, \{\sigma_X^i\}_{i=1}^k) &\leq I(\{\sigma_{X_t}^i\}_{i=1}^k, \{\sigma_X^i\}_{i=1}^k) \\ &\leq \sum_{x \in \{a,b,c\}} I(\{\sigma_x^i\}_{i=1}^k, \{\sigma_X^i\}_{i=1}^k) \\ &\leq \sum_{x \in \{a,b,c\}} I(\{\sigma_x^i\}_{i=1}^k, \{\sigma_{X_x}^i\}_{i=1}^k) \\ &\leq 3k\|\mu_H^+ - \mu_H^-\|_1 \\ &\leq 6k\sqrt{(2\theta^2)^H}. \end{aligned}$$

Using Lemma 14.5 and denoting by $\mu_{\boldsymbol{\Xi}_X}^t$ the distribution of the data given t, we get

$$\|\mu_{\boldsymbol{\Xi}_X}^{t_1} - \mu_{\boldsymbol{\Xi}_X}^{t_2}\|_1 \leq \sqrt{12k}(2\theta^2)^{H/4}.$$

So for the probability of reconstruction to be close to 1, we need k to grow exponentially with H.

Further reading

See [DMR09] and [Roc10] for more details on the upper bound for general trees. The thought experiment is from [Mos03].

References

[CT91] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley Series in Telecommunications. John Wiley & Sons Inc., New York, 1991. A Wiley-Interscience Publication.

[DMR09] Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Evolutionary trees and the Ising model on the Bethe lattice: a proof of Steel's conjecture. *Probab Theor Relat Field (In Press)*. Available at http://arxiv.org/abs/math.PR/0509575, 2009.

[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.

[Mos03] E. Mossel. On the impossibility of reconstructing ancestral data and phylogenies. *J. Comput. Biol.*, 10(5):669–678, 2003.

[Mos04] E. Mossel. Phase transitions in phylogeny. *Trans. Amer. Math. Soc.*, 356(6):2379–2404, 2004.

[Roc10] Sebastien Roch. Toward Extracting All Phylogenetic Information from Matrices of Evolutionary Distances. *Science*, 327(5971):1376–1379, 2010.