AWC Anual Meeting: Convergence and curvature of phylogenetic Markov chains

Alex Gavryushkin
(joint work with Chris Whidden and Erick Matsen)

21st October 2015

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND

Outline

- Motivation
- Discrete time-trees
- Curvature

Motivation

Question

Why all Bayesian phylogeny inference software is (sometimes / so) slow?

Motivation

Question

Why all Bayesian phylogeny inference software is (sometimes / so) slow?

Question

Specifically: Why does it take (sometimes / so) much time to converge?

Motivation

Question

Why all Bayesian phylogeny inference software is (sometimes / so) slow?

Question

Specifically: Why does it take (sometimes / so) much time to converge?

Question

Even more specifically: Why no reliable convergence criterion is know?

Answer

Motivation

Question

Why all Bayesian phylogeny inference software is (sometimes / so) slow?

Question

Specifically: Why does it take (sometimes / so) much time to converge?

Question

Even more specifically: Why no reliable convergence criterion is know?

Answer

I don't know

Motivation

Question

Why all Bayesian phylogeny inference software is (sometimes / so) slow?

Question

Specifically: Why does it take (sometimes / so) much time to converge?

Question

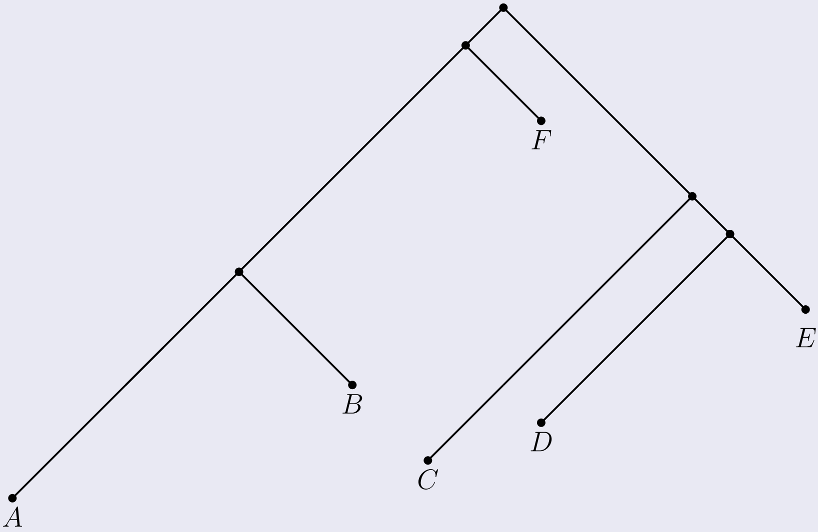
Even more specifically: Why no reliable convergence criterion is know?

Answer

I don't know geometry well.

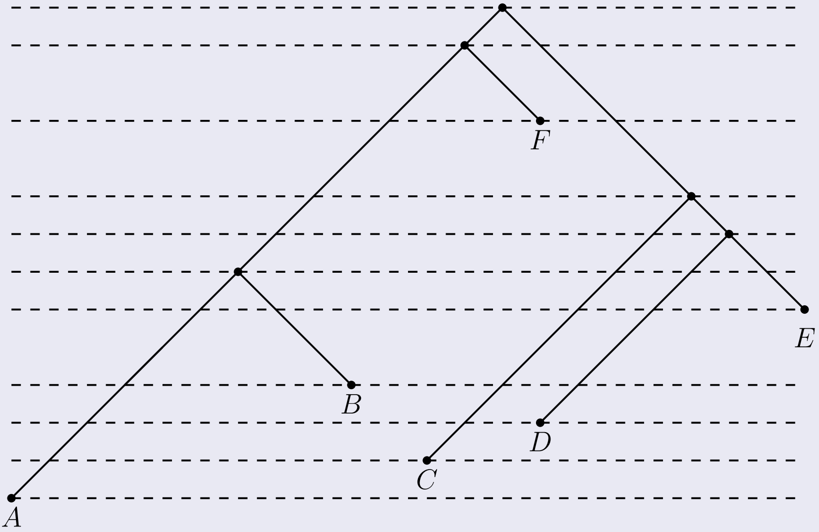
Discrete time-trees

Definition (Ranked tree topology)



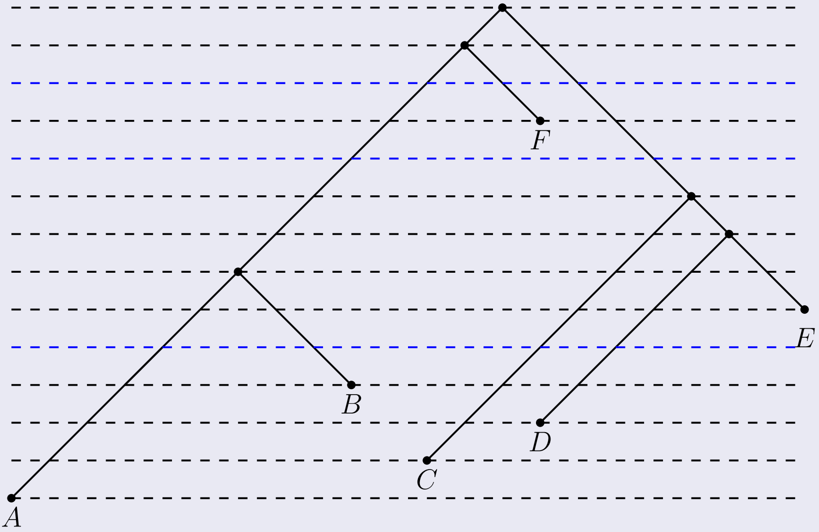
Discrete time-trees

Definition (Ranked tree topology)



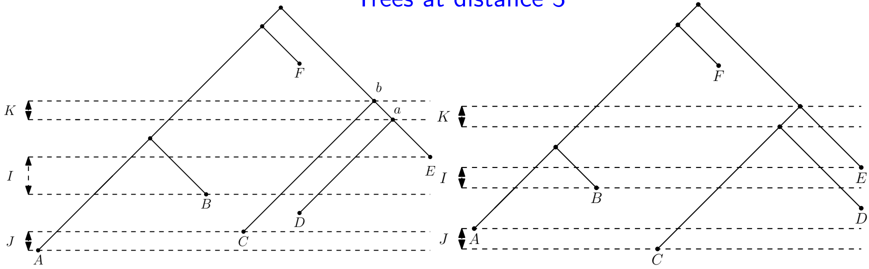
Discrete time-trees

Definition (Discrete time-tree)

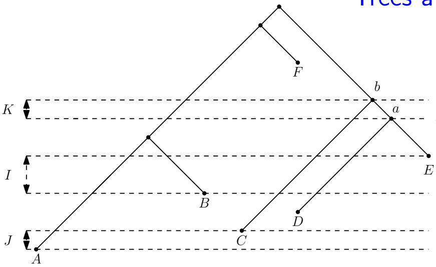


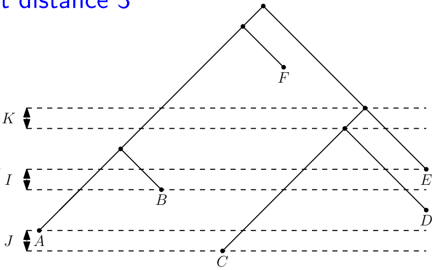
Discrete time-tree space

Trees at distance 3

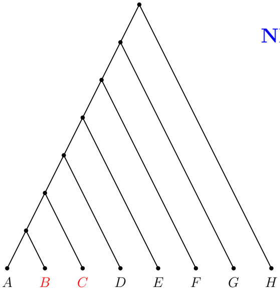


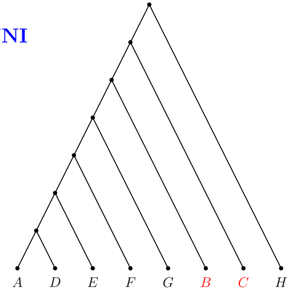
Trees at distance 3





NNI vs rNNI





Ricci-Ollivier curvature

Definition (Ollivier [2009])

Let $(\mathcal{T}, d)$ be a metric (tree) space with a random walk

$$m = (m_T)_{T \in \mathcal{T}}.$$

Let $T, R \in \mathcal{T}$ be two distinct points (trees). The Ricci-Ollivier curvature of $(\mathcal{T}, d, m)$ along $\overrightarrow{TR}$ is

$$\kappa_m(T, R) = 1 - \frac{W(m_T, m_R)}{d(T, R)},$$

where $W(\cdot\,,\cdot)$ is the earth mover's distance.

In a nutshell

Negative VS positive

$$\kappa_m(T, R) \leq 0 \iff W(m_T, m_R) \geq d(T, R)$$

In a nutshell

Negative VS positive

$$\kappa_m(T,R) \leq 0 \iff W(m_T, m_R) \geq d(T,R)$$

Take-home message

Negative curvature is bad.

Theorem (Ollivier [2009])

If $(\mathcal{T}, d)$ is a geodesic space then curvature is a local property.

Definition

Let $(\mathcal{T}, d)$ be a graph with a Markov chain m.Then the *curvature of the Markov chain m* on the graph $\mathcal{T}$ is the greatest number χ_m such that

$$\chi_m \leq \kappa_m(T, R) \text{ for adjacent } T \text{ and } R.$$

Trivial observation

Under a distance-one random walk, the following is true for any finite metric d and any pair of points T, R:

$$\frac{-2}{d(T,R)} \leq \kappa(T,R) \leq \frac{2}{d(T,R)}.$$

For now, we consider three simplest random walks on various phylogenetic tree spaces.

- Metropolis-Hastings random walk: Choose a tree from the one neighbourhood and accept it with probability $\min(1, \frac{|N_1(T_{old})|}{|N_1(T_{new})|})$.
- Uniform random walk.
- Uniform p-lazy random walk, where p is the laziness probability.

Lower bounds

Theorem (G, Whidden, Matsen [2015])

Let T and R be adjacent trees. Then both the asymptotic curvature of the space with p-lazy uniform random walk and the curvature of the space with uniform random walk are at least

$$\kappa(T,R) \geq \frac{-n^2+2n}{3.5n^2-15n+16} \geq -2/5 \qquad \textit{in } \mathrm{r}SPR \textit{ space,}$$

$$\kappa(T,R) \geq -\frac{4}{n-1} \qquad \textit{in } \mathrm{DtT} \textit{ space,}$$

$$\kappa(T,R) \geq -\frac{4}{n-2} \qquad \textit{in } \mathrm{NNI} \textit{ space,}$$

$$\kappa(T,R) \geq -\frac{8}{n-1} \qquad \textit{in } \mathrm{rNNI} \textit{ space.}$$

The bounds are tight.

Upper bounds

Theorem (G, Whidden, and Matsen [2015])

Let T and R be adjacent trees. Then the curvature of the following spaces with uniform random walk satisfy

$$\kappa(T,R) \leq \frac{6n-17}{3n^2-13n+14} \qquad \textit{in } \mathrm{rSPR} \textit{ space,}$$

$$\kappa(T,R) \leq \frac{1}{2(n-1)} \qquad \textit{in } \mathrm{DtT} \textit{ space,}$$

$$\kappa(T,R) \leq \frac{1}{2(n-2)} \qquad \textit{in } \mathrm{NNI} \textit{ space, and}$$

$$\kappa(T,R) \leq \frac{1}{n-1} \qquad \textit{in } \mathrm{rNNI} \textit{ space.}$$

The bounds are tight.

Life is good, at infinity

Theorem (G, Whidden, and Matsen [2015])

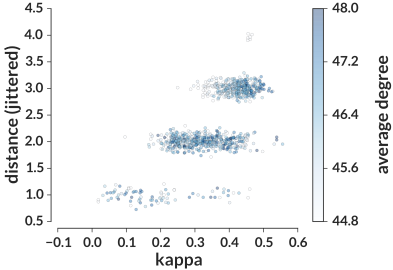
Let $\{T_n \mid n \in \mathbb{N}\}$ *and* $\{S_n \mid n \in \mathbb{N}\}$ *be two sequences of phylogenetic trees such that* $d(T_n, R_n) = 1$ *for all* n*. Then*

$$\lim_{n\to\infty} \kappa_n(T_n, S_n) = 0$$

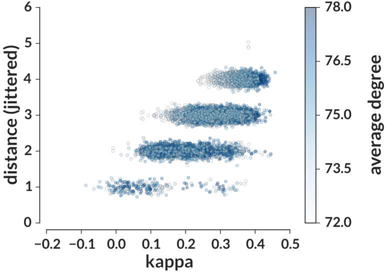
for the uniform random walk on the SPR graph, the NNI graph, the* τ*-graph, and the discrete* τ*-space.*

---

*For the SPR graph, we have to bound the size of the subtree which is getting moved.



(a) 6 taxa



(b) 7 taxa

Figure: Scatter plot of $\kappa(\mathrm{MH}; T_1, T_2)$ values versus $d_{SPR}(T_1, T_2)$ for the rSPR graph. Color displays the average degree of T_1 and T_2. Distance values randomly perturbed ("jittered") a small amount to avoid superimposed points.

Thank you for your attention!

Yann Ollivier
Ricci curvature of Markov chains on metric spaces
J. Functional Analysis, 256, 3, 810–864, 2009

Alex Gavryushkin and Alexei Drummond
The space of ultrametric phylogenetic trees
arXiv preprint arXiv:1410.3544, 2014

Chris Whidden and Frederick A. Matsen IV
Quantifying MCMC exploration of phylogenetic tree space
Systematic Biology, doi:10.1093/sysbio/syv006, 2015

Chris Whidden and Frederick A. Matsen IV
Ricci-Ollivier curvature of two random walks on rooted phylogenetic subtree-prune-regraft graph
To appear in the proceedings of the *Thirteenth Workshop on Analytic Algorithmics and Combinatorics*, 2015

Alex Gavryushkin, Chris Whidden, and Frederick A. Matsen IV
Random walks over discrete time-trees
To appear on the *arXiv*, 2015

Alex Gavryushkin
`https://github.com/gavruskin/tTauCurvature`