ETHZ Computational Biology Seminar: Convergence and curvature of phylogenetic Markov chains

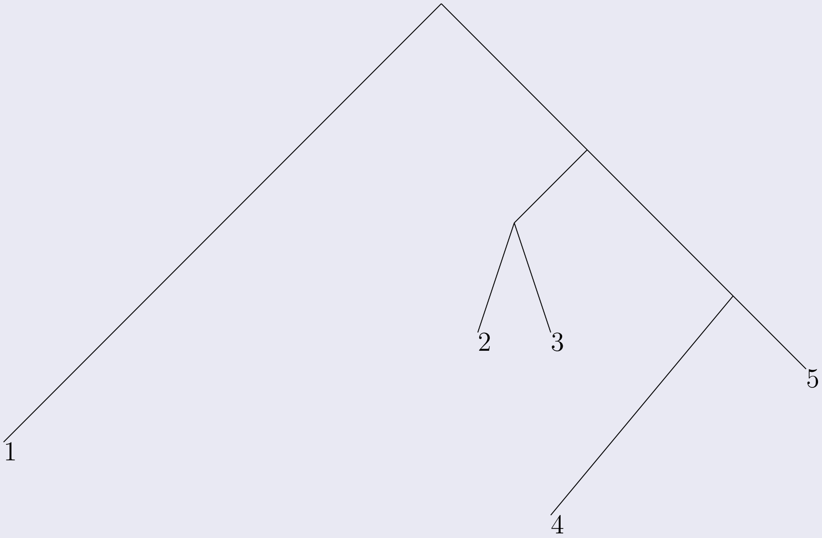
Alex Gavryushkin
(joint work with Alexei Drummond,
Chris Whidden, and Erick Matsen)

11th November 2015

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND

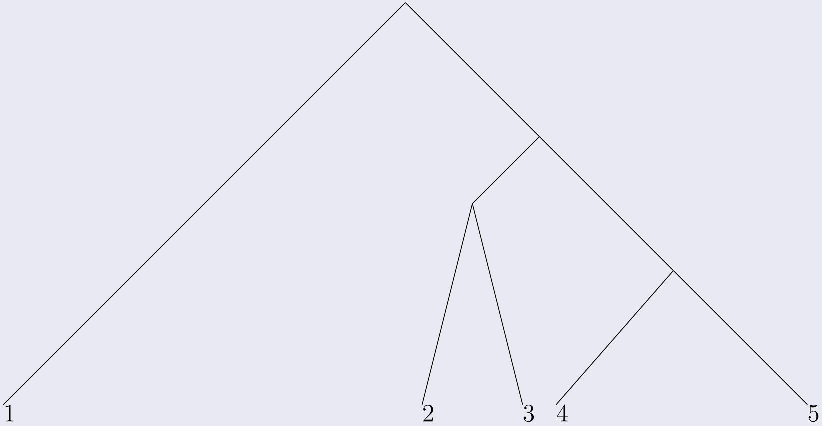
Rooted phylogenetic tree

Definition



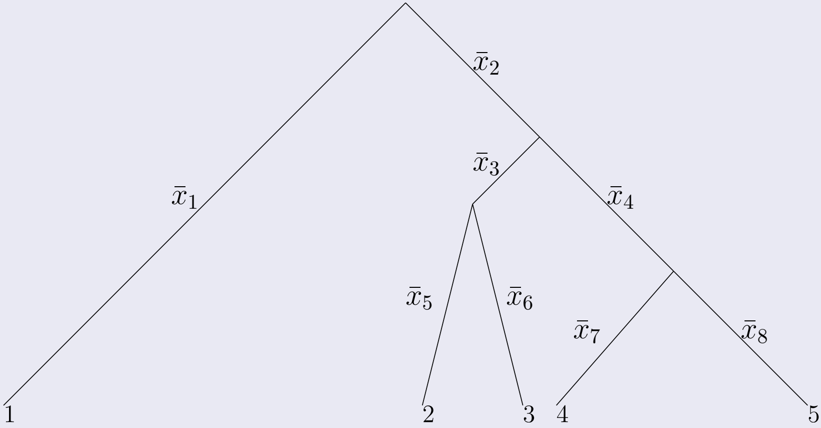
Equidistant (ultrametric) phylogenetic tree

Definition



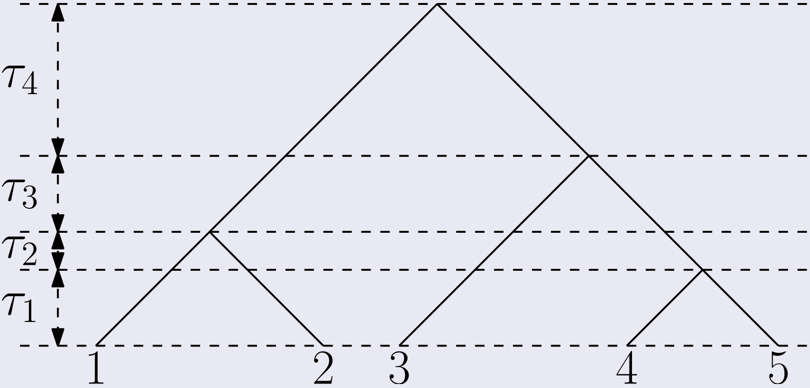
Equidistant phylogenetic tree with parameters

Definition

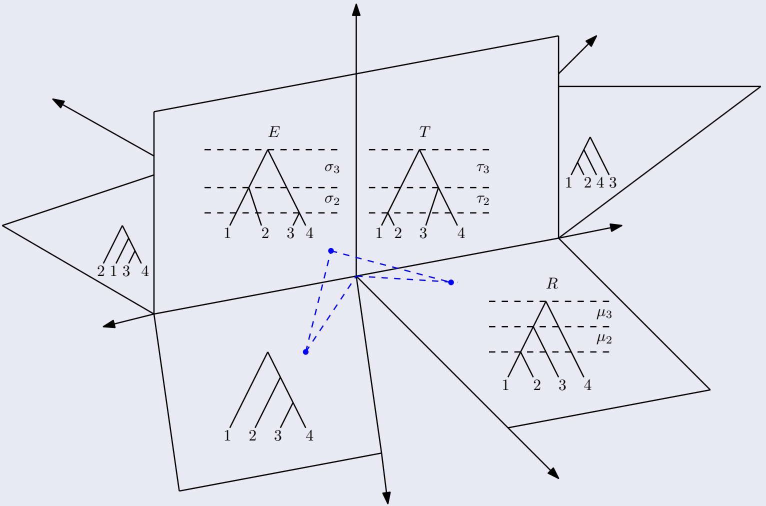


τ-parameterisation

Definition

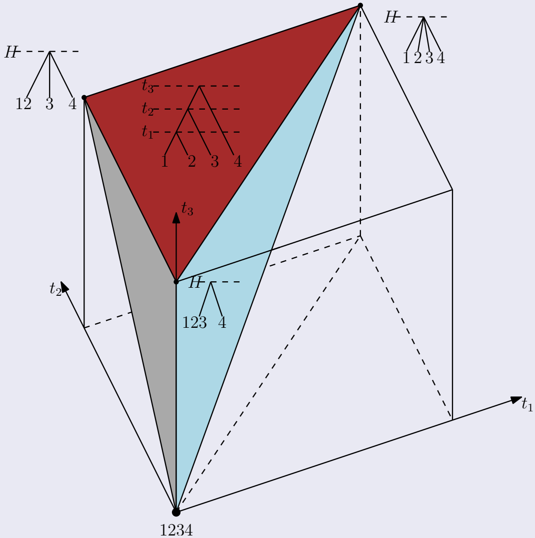


Definition



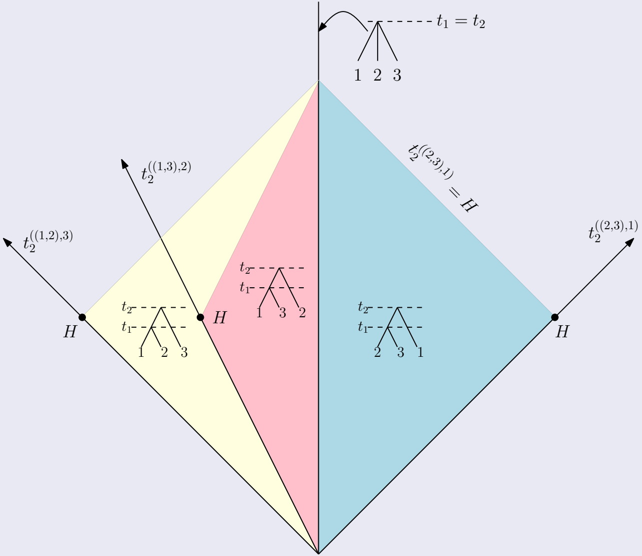
t-parameterisation

Definition



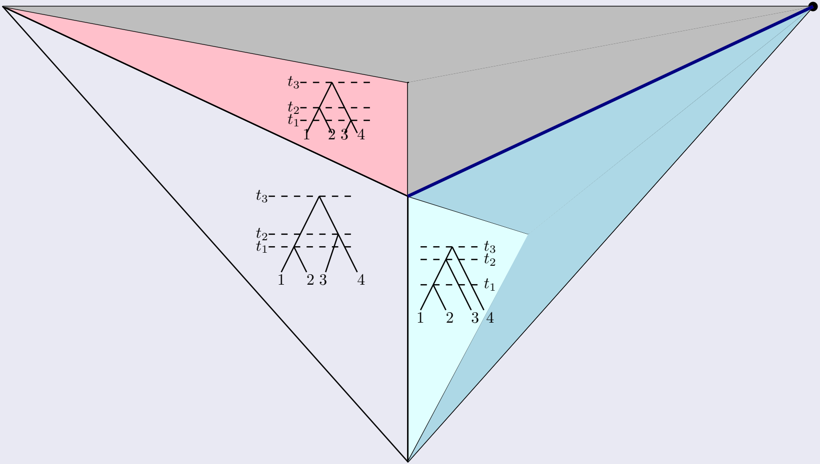
t-space

Definition



t-space

Definition



Motivation

1. Bayesian MCMC: Mixing rate, access time, efficient proposals.
2. Summarising posterior: No need to introduce several random variables on different probability spaces, no need to fit inconsistent data together.
3. Interesting algorithmic/data structures problems: How to solve NP-complete problems on real computers for real data (Chris and Erick can compute SPR-distance).
4. Interesting geometries: "Every new example of a non-trivial simplicial complex of non-positive curvature is a big deal."

Nice metric spaces

Definition

A metric space is called *nice* if most statisticians would like it.

Examples of nice metric spaces include real line, Euclidean space, and its nice subspaces.
Examples of not nice metric spaces include all non-measurable subsets of a Euclidean space, all nowhere dense subsets of a Euclidean space, and most importantly the spaces where it is hard to define a random variable.

Theorem (Billera, Holmes, and Vogtmann [2001])

The space of phylogenetic trees is a nice space.

Theorem (G and Drummond [2015])

The space of equidistant phylogenetic trees is a nice space.

Parameterisation matters!

Theorem (G and Drummond [2015])

t*-space is not so nice.*

Parameterisation matters!

Theorem (G and Drummond [2015])

t-*space is not so nice.*

More formally

Definition

A geodesic metric space is called *nice* if it is a convex path-connected subspace of a computable metric space with unique geodesics of the same dimension.

Theorem (G and Drummond [2015])

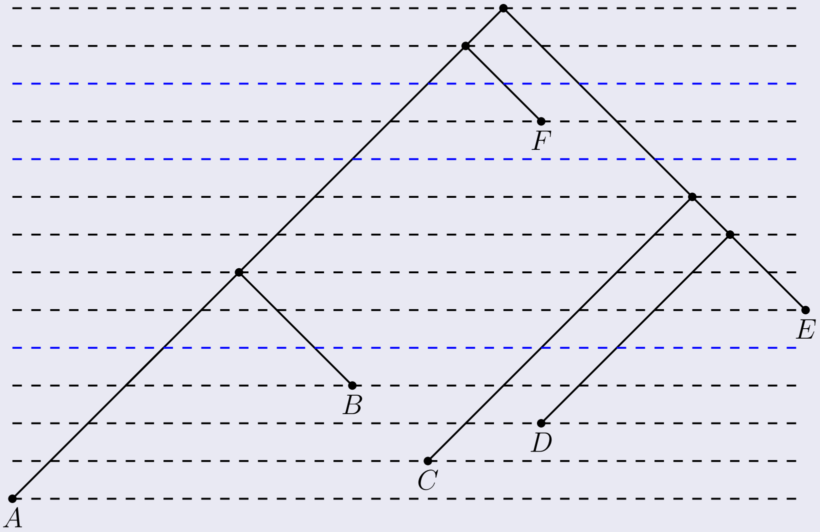
τ-*space is an efficiently computable cubical complex with unique geodesics.*

Theorem (G and Drummond [2015])

t-*space is a simplicial complex with unique geodesics.*

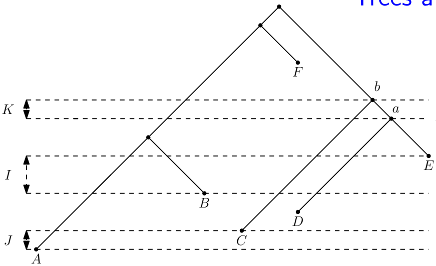
Discrete time-trees

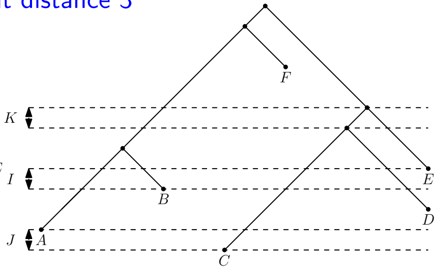
Definition (Discrete time-tree)



Discrete time-tree space

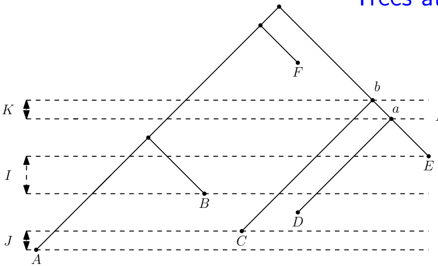
Trees at distance 3

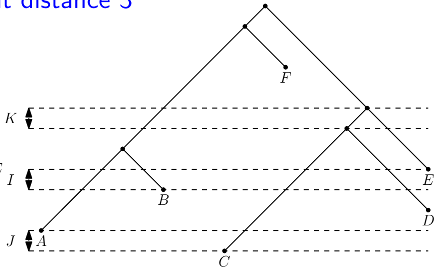




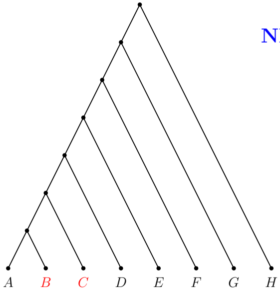
Discrete time-tree space

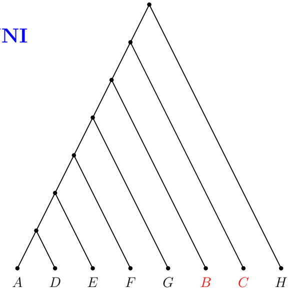
Trees at distance 3





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.

Curvature of Markov chains on graphs

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:

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

Random walks

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* rNNI*-graph, and the* DtT*-graph.*

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

Whidden and Matsen [2015]

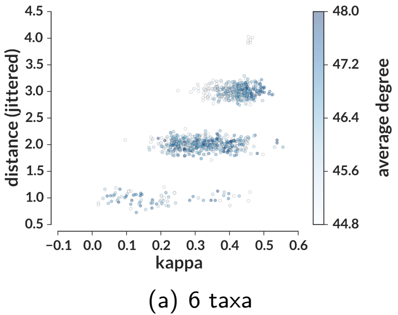


Figure: Scatter plot of $\kappa(\mathrm{MH}; T_1, T_2)$ values versus $d_{SPR}(T_1, T_2)$ for the rSPR graph. Colour displays the average degree of T_1 and T_2. Distance values randomly perturbed ("jittered") a small amount to avoid superimposed points.

Take-home message

$$1.001^{10000} = 21916.68\ldots$$

BUT

$$0.999^{10000} = 0.000045\ldots$$

Take-home message

$$1.001^{10000} = 21916.68\ldots$$

BUT

$$0.999^{10000} = 0.000045\ldots$$

- The curvature of basic random walks is normally positive.
- Although the spaces flatten out when the number of taxa n grows, there always are negatively curved pieces.
- Importantly, the number of those pieces grows with n.

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
- `https://github.com/gavruskin/tauGeodesic`
- `https://github.com/gavruskin/tTauCurvature`