Improved Importance Sampling of Phylogenies

Mathias C. Cronjäger

Department of Statistics
University of Oxford

Presented at the LMS invited lecture series / CRISM summer school 2018 on computational statistics, University of Warwick, July 10th 2018

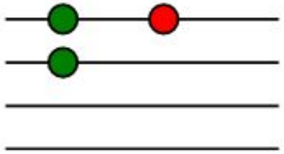
Outline

1. Our motivating problem
2. Importance sampling
    a. sequential proposal distributions
        i. existing proposals
3. Path density bias
    a. de-biasing by counting paths

Setup

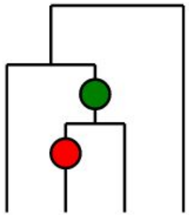
The data:

Single Nucleotide Polymorphisms (SNPs) in *n* aligned genetic sequences sampled in the present.



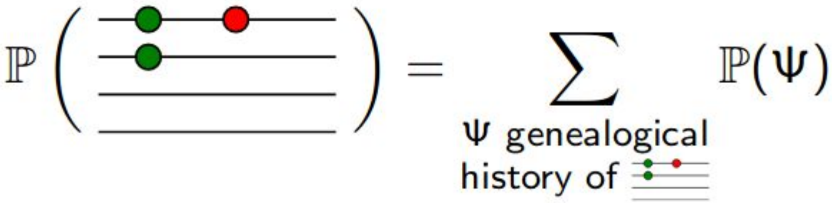
The model:

Individuals are leaves on an an unobserved random tree with the most recent common ancestor at the root (a coalescent). New segregating sites occur along branches at rate θ.



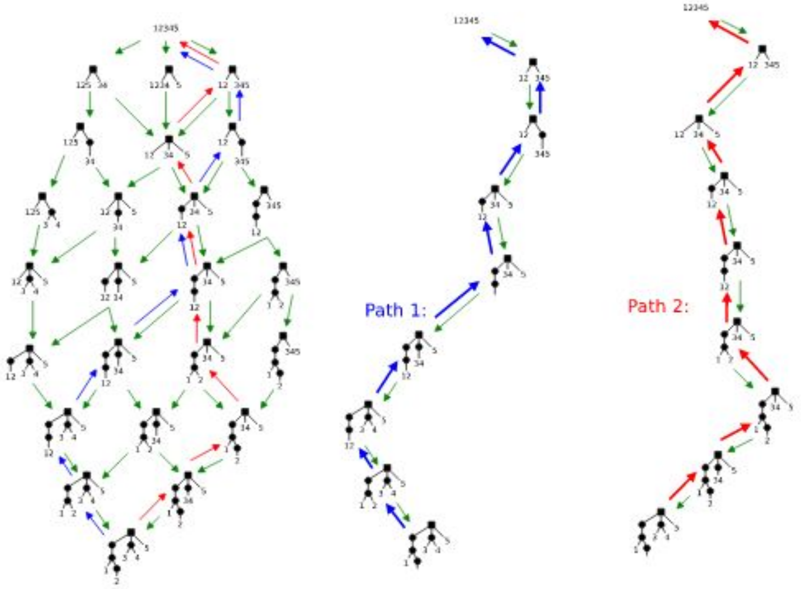
The challenge:

Computing likelihoods does not scale as *n* increases (there are too many possible trees to consider)



Importance sampling of ancestral histories

$$\mathbb{P}\left(\;\right) = \sum_{x \in \mathrm{H}(\;)} \frac{\mathbb{P}(x)}{\mathbb{Q}(x)} \mathbb{Q}(x) = \mathbb{E}_{X \sim \mathbb{Q}} \left[\frac{\mathbb{P}(X)}{\mathbb{Q}(X)} \right] \approx \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbb{P}(X_i)}{\mathbb{Q}(X_i)}$$



e.g. $\mathbb{P}\left(\;\right) \approx \frac{1}{2} \left(\frac{\mathbb{P}(\text{Path 1})}{\mathbb{Q}(\text{Path 1})} + \frac{\mathbb{P}(\text{Path 2})}{\mathbb{Q}(\text{Path 2})} \right)$

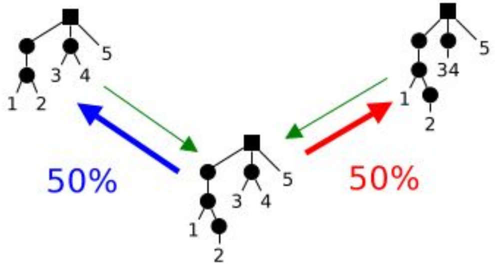
Ideally, the proposal distribution ***Q***, should

1. Approximate the target distribution ***P*** well
2. Sampling $X \sim$ ***Q*** should be fast
3. Computing weights $f_{\mathbf{Q}}(X)$ should be fast

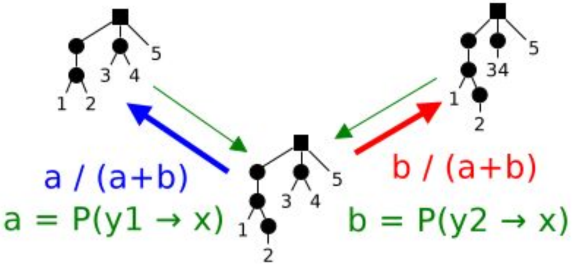
Conditions 2 & 3 allow us to pick large *N* (number of particles).
Condition 1 gives us better convergence in *N*.

For this problem: Known (feasible) proposals are **sequential**. Sampling from ***Q*** is done by constructing paths step-by-step, **from the bottom up**.

Existing proposals (a single step)



Stephens and Donnelly proposal



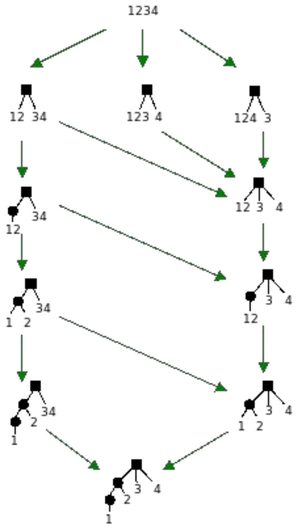
Griffiths and Taveré (G&T)

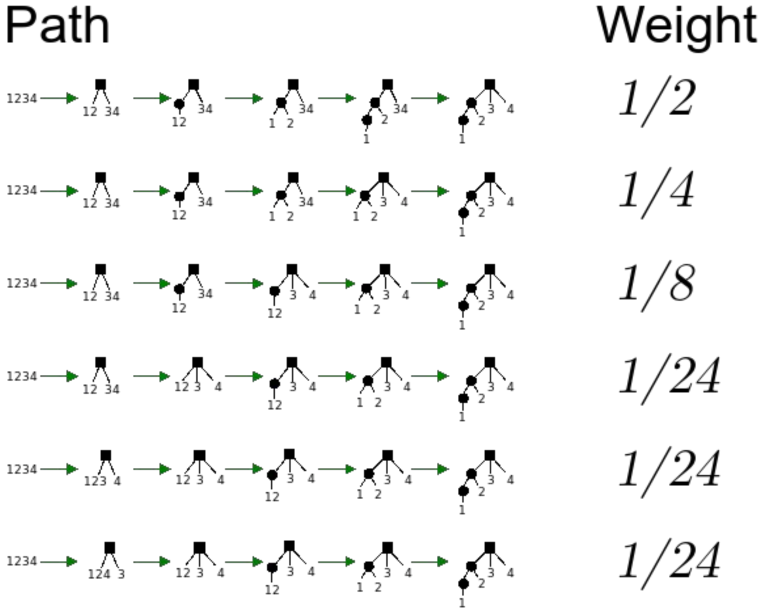
Sequential proposals and path density bias

Consider for example $\mathbf{Q}_{SD}$ which samples ancestral histories as follows:

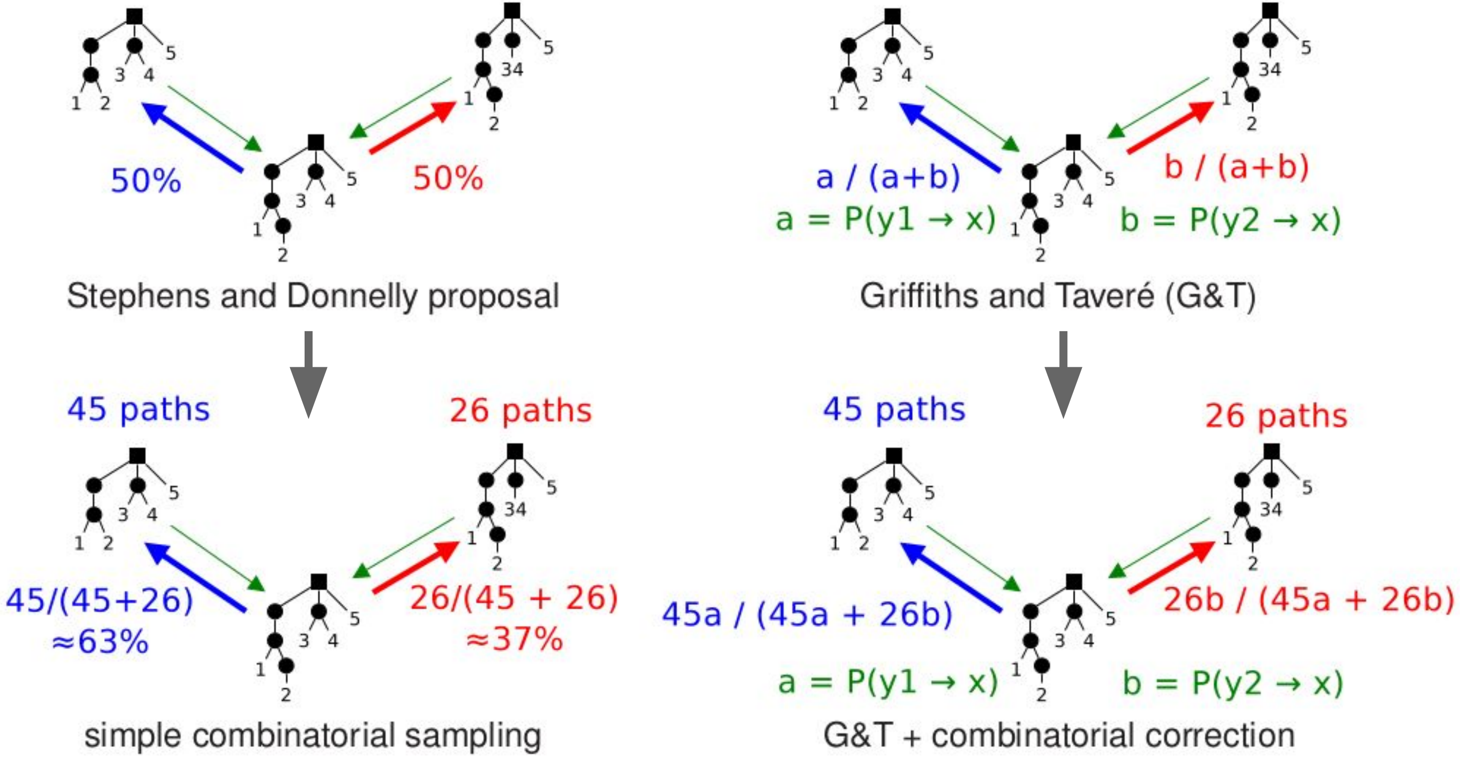
1. Start with state x_0 = Observed data
2. Given a partial path $p = [x_0, \dots, x_i]$, pick a state x_{i+1} uniformly at random from the predecessors of x_i, and update the partial path to $p' = [x_0, \dots, x_i, x_{i+1}]$
3. Terminate when hitting a state with no ancestors.

Major Issue: Oversampling of paths containing nodes of low in-degree.





Combinatorial importance sampling



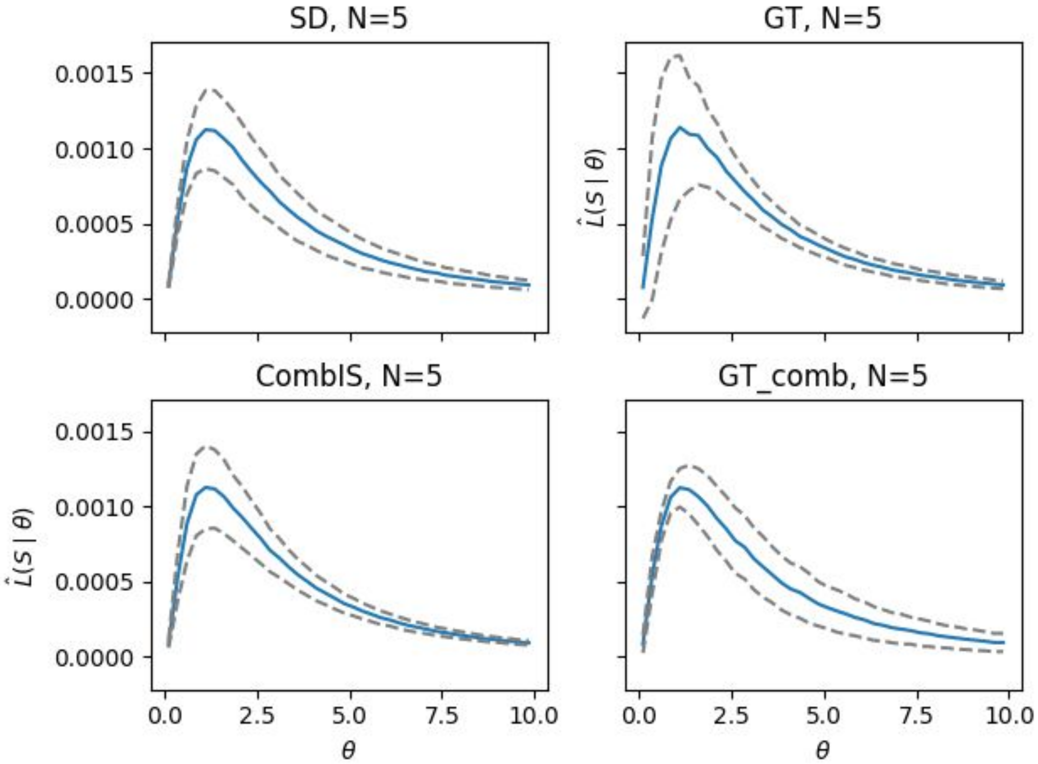
Stephens and Donnelly proposal

Griffiths and Taveré (G&T)

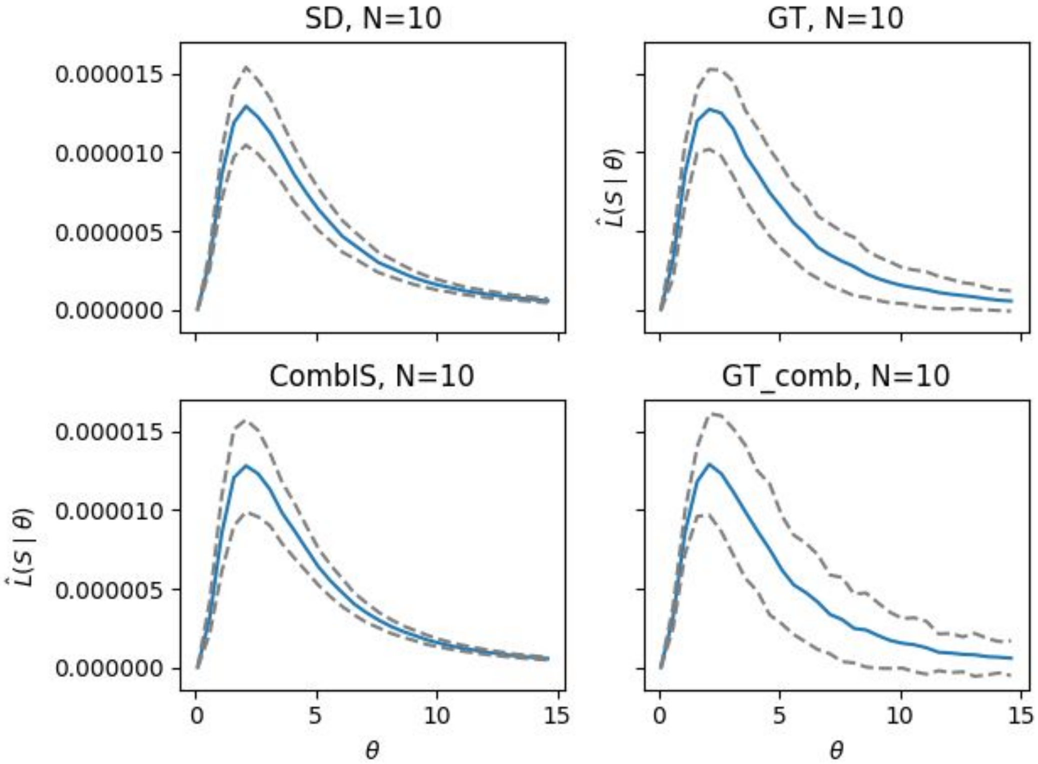
simple combinatorial sampling

G&T + combinatorial correction

Likelihoods (toy example 1: 4 seq, 2 sites)



Likelihoods (toy example 2: 5 seq; 4 sites)



Thank you.
Any Questions?

Collaborators (and supervisors)

Jotun Hein
(Oxford, Statistics)

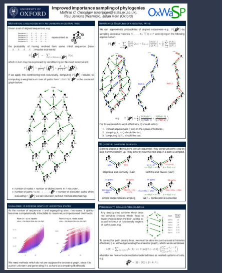
Paul Jenkins
(Warwick, Statistics & CS)

Code: https://github.com/Cronjaeger/combIS

Get in touch:

Email: cronjager@stats.ox.ac.uk
Twitter: @mcCronjaeger

I also have a poster →



UNIVERSITY OF OXFORD

Improved importance sampling of phylogenies
Mathias C. Cronjäger (cronjager@stats.ox.ac.uk),
Paul Jenkins (Warwick), Jotun Hein (Oxford)

OxWaSP

MOTIVATION: LIKELIHOODS WITH AN UNKNOWN ANCESTRAL TREE

Given a set of aligned sequences, e.g. represented as

the probability of having evolved from some initial sequence (here …A…A…A…A…) may be expressed:

which in turn may be expressed by conditioning on the most recent event:

If we apply this conditioning-trick recursively, computing $\mathbb{P}$ reduces to computing a weighted sum over all paths from "12345" to in the ancestral graph below.

- number of nodes = number of distinct terms in $\mathbb{P}$-recursion.
- number of paths "12345 → … →" = number of execution paths when evaluating $\mathbb{P}$ via tail-recursion (without memoization/tabling).

IMPORTANCE SAMPLING OF ANCESTRAL PATHS

We can approximate probabilities of aligned sequences–e.g. $\mathbb{P}$–by sampling ancestral histories $X_1, \dots, X_N \overset{iid}{\sim} \mathbb{Q} \ll \mathbb{P}$ and relying on the following approximation:

Path 1:

Path 2:

e.g. $\mathbb{P} \approx \frac{1}{2}\left(\frac{\mathbb{P}(\text{Path 1})}{\mathbb{Q}(\text{Path 1})} + \frac{\mathbb{P}(\text{Path 2})}{\mathbb{Q}(\text{Path 2})}\right)$

For this approach to work effectively, $\mathbb{Q}$ should satisfy:

1. $\mathbb{Q}$ must approximate $\mathbb{P}$ well on the space of histories;
2. sampling $X_i \sim \mathbb{Q}$ should be fast;
3. computing $\mathbb{Q}(X_i)$ should be fast.

SEQUENTIAL SAMPLING SCHEMES

Existing proposal distributions are all sequential: they construct paths step-by-step from the bottom up. They differ by how the next step in a path is sampled.

Stephens and Donnelly (S&D)

Griffiths and Tavaré (G&T)

simple combinatorial sampling

G&T + combinatorial correction

CHALLENGE: A GROWING GRAPH OF ANCESTRAL STATES

As the number of sequences n and segregating sites s increases, it quickly becomes computationally intractable to recursively compute exact likelihoods.

We need methods which do not pre-suppose the ancestral graph, since it is a priori unknown and generating it is as hard as computing likelihoods.

PATH DENSITY BIAS AND PATH COUNTING

Any step-by-step scheme which does not penalize choices which "lead to fewer choices down the line", will be biased in favour of low-density regions of path-space, e.g.

To correct for path density bias, we must be able to count ancestral histories effectively (i.e. without generating the ancestral graph), which we do as follows:

whereby we here encode rooted unordered trees as nested systems of sets, e.g.