Nonreversible MCMC for latent phylogenetic trees

Jere Koskela

Inference for expensive systems in mathematical biology
University of Oxford, 23–24 March 2022

EPSRC
Engineering and Physical Sciences Research Council

Warwick Statistics

Outline

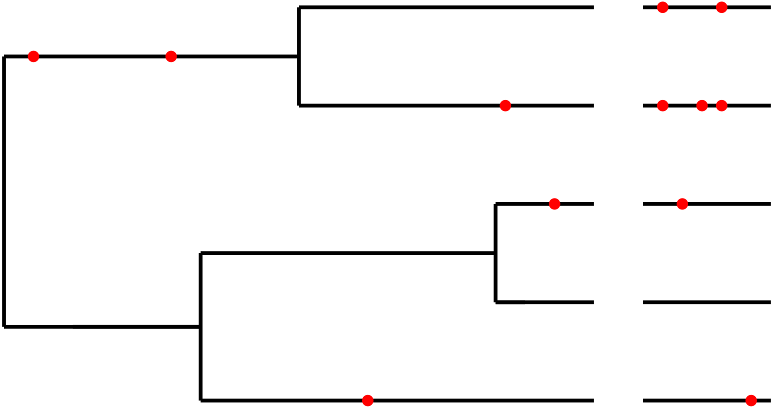
The coalescent

Metropolis-Hastings

The zig-zag process

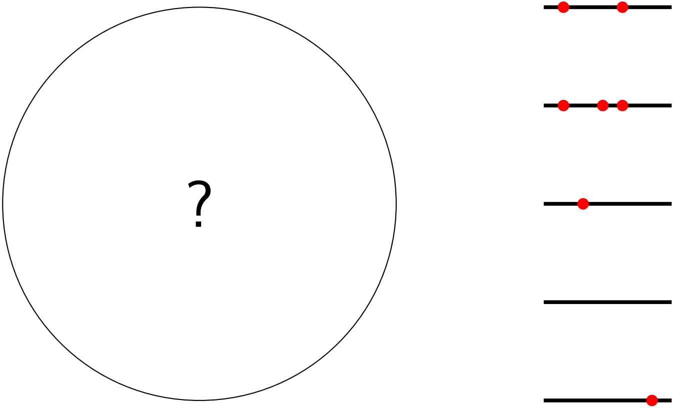
Zig-zag for the coalescent

The coalescent[1]



[1] J F C Kingman. The coalescent, Stoch Proc Appl 13(3):235–248, 1982.

The coalescent as missing data



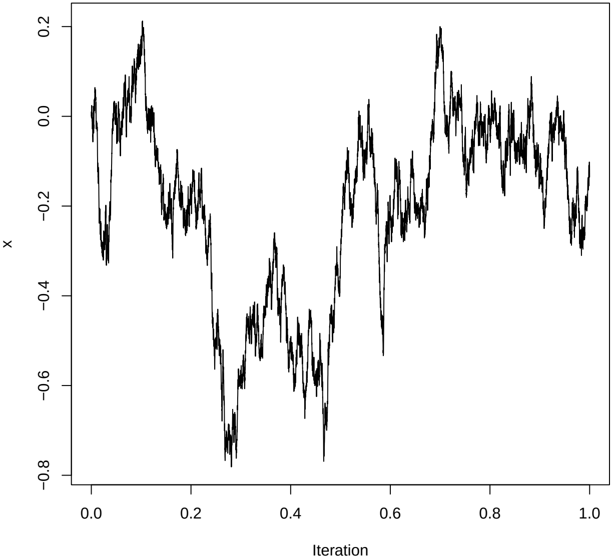
$$P(D|\theta) = \int_A P(D|A,\theta)P(A|\theta)dA$$

Metropolis–Hastings[2]

```
1. Set $X_1 \leftarrow x$.
2. For $i \in \{2, \dots, m\}$ do
    2.1 Sample $Y \sim q(X_{i-1}, \cdot)$.
    2.2 Sample $U \sim U(0, 1)$.
    2.3 If $U \leq \min\left\{1, \frac{\pi(Y) q(Y, X_{i-1})}{\pi(X_{i-1}) q(X_{i-1}, Y)}\right\}$, set $X_i \leftarrow Y$.
    2.4 Else set $X_i \leftarrow X_{i-1}$.
```

[2] W K Hastings. Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57(1):97–109, 1970.

Reversibility ⇒ diffusive behaviour



The zig-zag process[3]

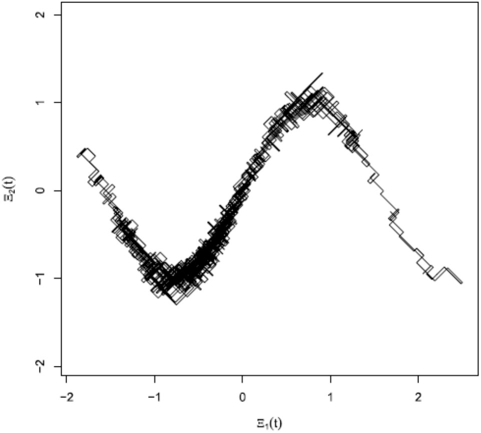
```
1. Set $(x, v) \in \mathbb{R}^n \times \{-1, 1\}^n$.
2. Set $t \leftarrow 0$.
3. While $t < T$ do
    3.1 Sample $Y \sim \mathrm{Exp}(1)$.
    3.2 Set $\rho$ such that $\sum_{i=1}^{n} \int_t^{t+\rho} \lambda_i(x + sv, v)\mathrm{d}s = Y$.
    3.3 Set $x \leftarrow x + \rho v$.
    3.4 Set $t \leftarrow t + \rho$.
    3.5 Sample $I \sim \mathrm{Categorical}(\lambda_1(x, v), \ldots, \lambda_n(x, v))$.
    3.6 Set $v_I \leftarrow -v_I$.
```

- Between switches, x moves with constant velocity v.
- Target density π is invariant if

$$\lambda_i(x, v) = v_i \partial_i \log \pi(x) \vee 0.$$

[3] J Bierkens, P Fearnhead and G Roberts. The zig-zag process and super-efficient sampling for Bayesian analysis of big data, Ann Stat 47(3):1288–1320, 2019.

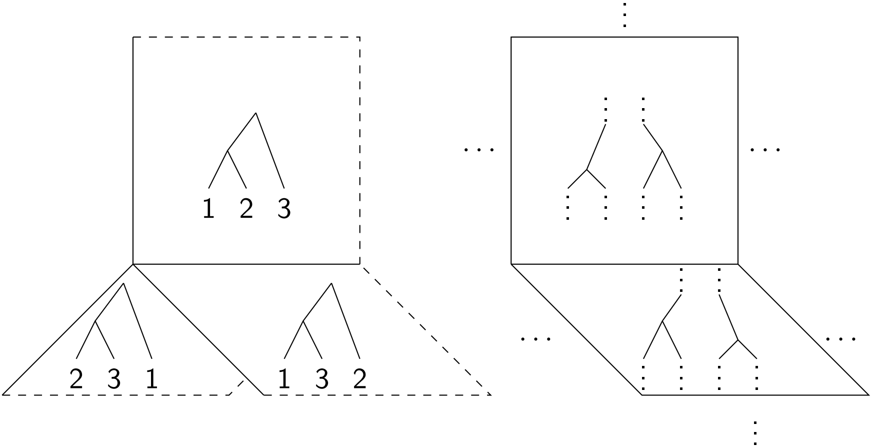
The zig-zag process



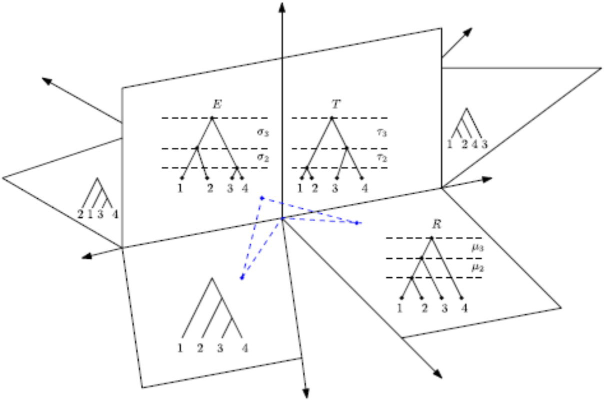
(d) 2D S-shaped density

[3] J Bierkens, P Fearnhead and G Roberts. The zig-zag process and super-efficient sampling for Bayesian analysis of big data, Ann Stat 47(3):1288–1320, 2019.

The three leaf τ-space



One third of the four leaf τ-space[4]



[4] A Gavryushkin and A J Drummond. The space of ultrametric phylogenetic trees, J Theor Biol 403:197–208, 2016.

Simulation study[5,6,7]

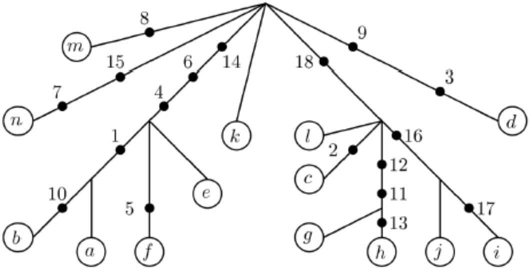


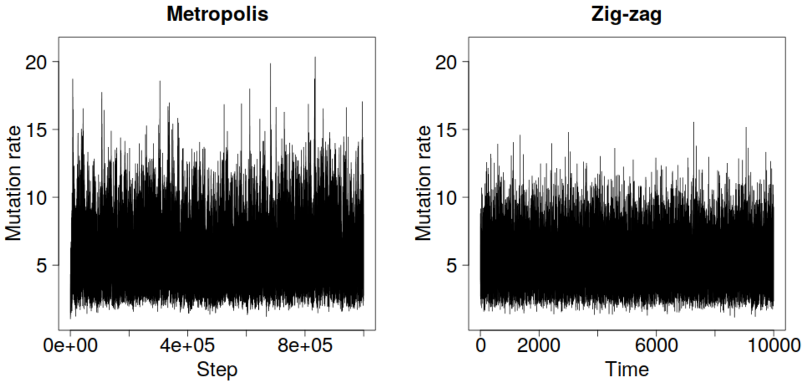
Figure 3.
Perfect phylogeny of the Griffiths and Tavaré (1994) data set.

[5] A Hobolth, M K Uyenoyama and C Wiuf. Importance sampling for the infinite sites model, Stat Appl Genet Mol Biol 7(1) Article 32, 2008.

[6] R C Griffiths and S Tavaré. Ancestral inference in population genetics, Stat Sci 9:307–319, 1994.

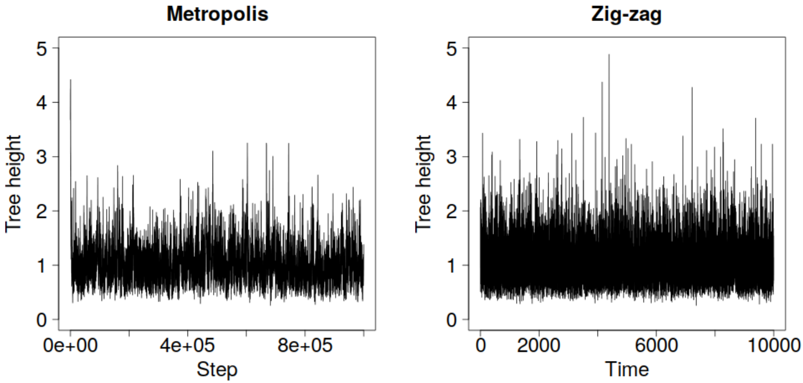
[7] R H Ward, B L Frazier, K Dew and S Pääbo. Extensive mitochondrial diversity within a single Amerindian tribe, Proc Natl Acad Sci USA 88:8720–8724, 1991.

Mutation rate (55 samples, 18 segregating sites)



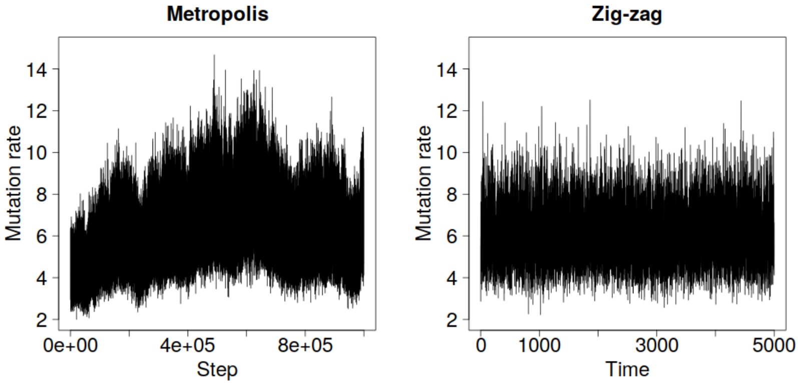
Run times: 3.5 min vs. 0.5 min.

Tree height (55 samples, 18 segregating sites)



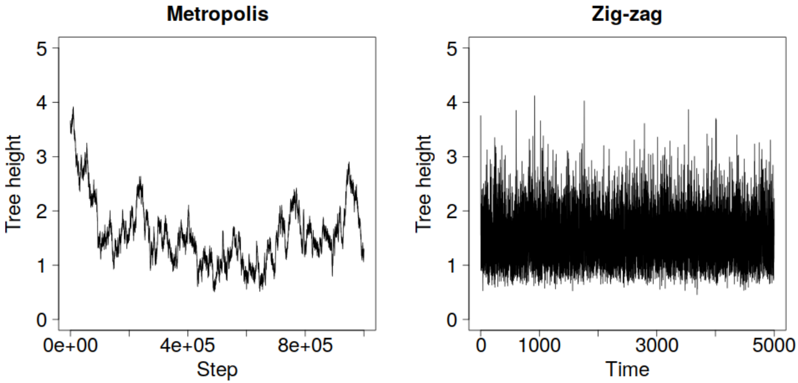
Run times: 3.5 min vs. 0.5 min.

Mutation rate (550 samples, 38 segregating sites)



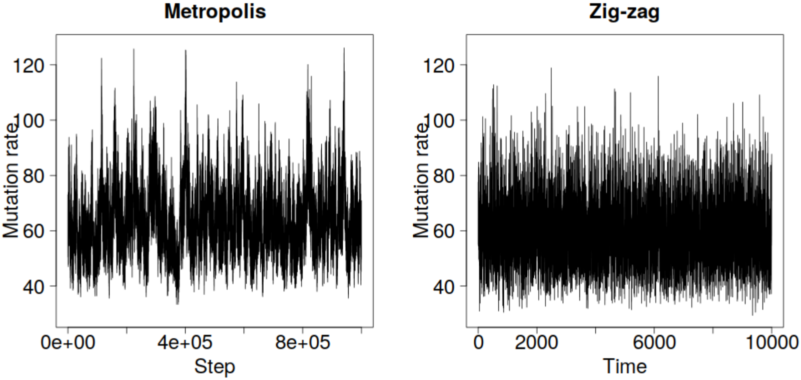
Run times: 70 min vs. 45 min.

Tree height (550 samples, 38 segregating sites)



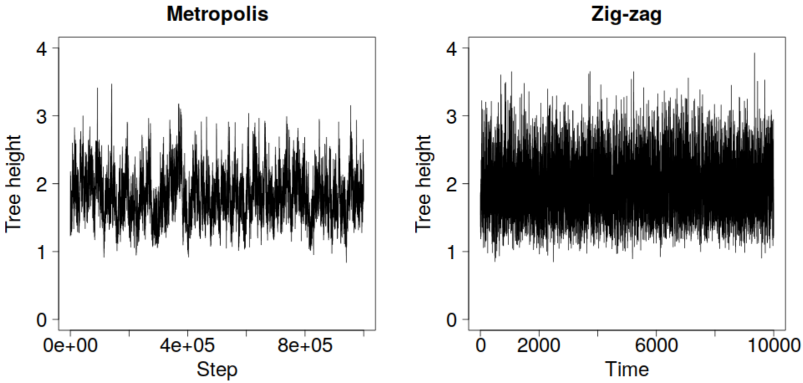
Run times: 70 min vs. 45 min.

Mutation rate (55 samples, 252 segregating sites)



Run times: 50 min vs. 1 min.

Tree height (55 samples, 252 segregating sites)



Run times: 50 min vs. 1 min.