Asymptotic Genealogies of Sequential Monte Carlo Algorithms

Suzie Brown

University of Warwick, U.K.
with Paul Jenkins, Adam Johansen & Jere Koskela

24 February 2021

Outline

1. Sequential Monte Carlo
2. Resampling & Genealogies
3. Results
4. Examples

Sequential Monte Carlo

- Monte Carlo = approximating probability distributions via sampling
- Sequence of probability distributions
- Dimension of distributions increases along the sequence
- Strong dependence between consecutive distributions

State space models

$$
\begin{tikzcd}
 & Y_t & Y_{t+1} & \\
\cdots \arrow[r] & X_t \arrow[u] \arrow[r] & X_{t+1} \arrow[u] \arrow[r] & \cdots
\end{tikzcd}
$$

$$X_0 \sim \mu(\cdot)$$
$$X_{t+1} \mid X_t$$
$$Y_t \mid X_t$$

State space models

$$
\begin{tikzcd}
 & Y_t & Y_{t+1} & \\
\cdots \arrow[r] & X_t \arrow[r] \arrow[u] & X_{t+1} \arrow[r] \arrow[u] & \cdots
\end{tikzcd}
$$

$$X_0 \sim \mu(\cdot)$$
$$X_{t+1} \mid X_t$$
$$Y_t \mid X_t$$

Inference problems ($s < t$):

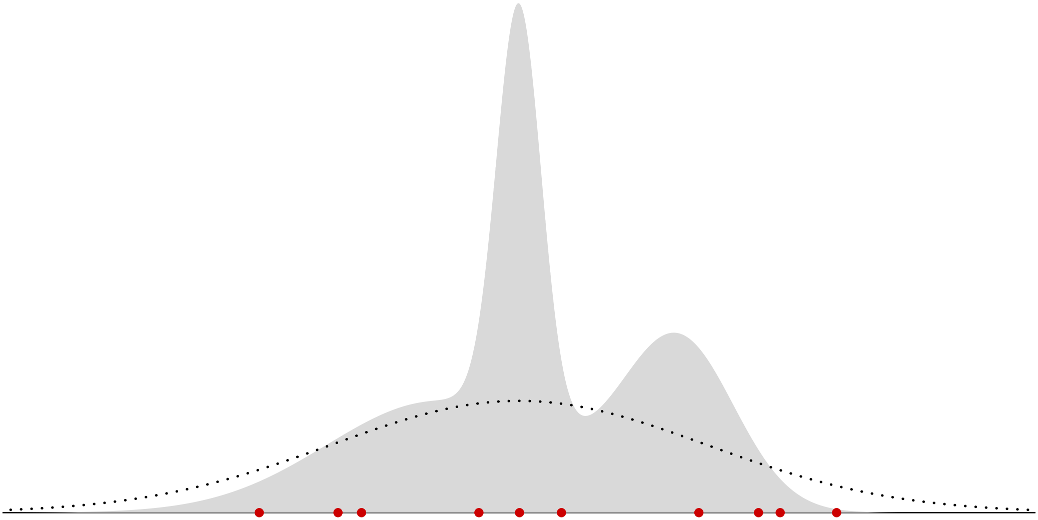
$p(x_t \mid y_{1:s})$	"prediction"
$p(x_t \mid y_{1:t})$	"filtering"
$p(x_s \mid y_{1:t})$	"smoothing"

State space models

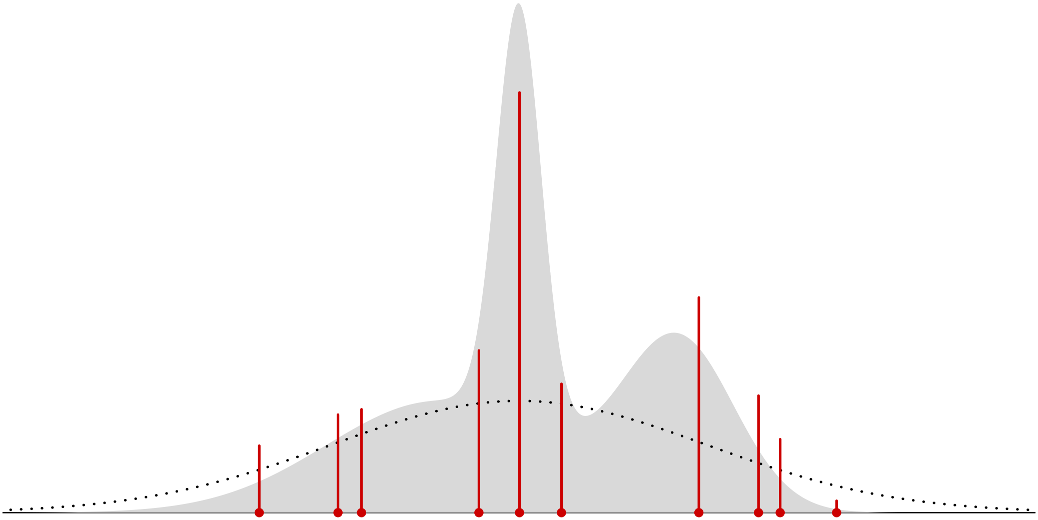
- **Kalman filter** and **Rauch-Tung-Striebel smoother** give exact solutions, but only for linear Gaussian models
- **Markov chain Monte Carlo** fails due to high dimension and high correlation between dimensions
- **Sequential Monte Carlo** has computational complexity that is linear in t, and can update posteriors on-line

Importance sampling

Importance sampling



Importance sampling

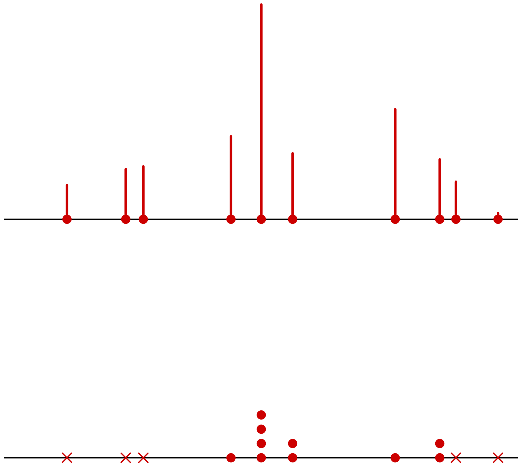


Sequential importance sampling

- Idea: use weighted samples from one time step to construct a proposal for the next step
- Multiplying weights over time causes *weight degeneracy*
- Can avoid this problem by resampling

Resampling

Stochastically map continuous weights $(w_t^{(1)}, \dots, w_t^{(N)})$ to discrete offspring counts $(\nu_t^{(1)}, \dots, \nu_t^{(N)})$

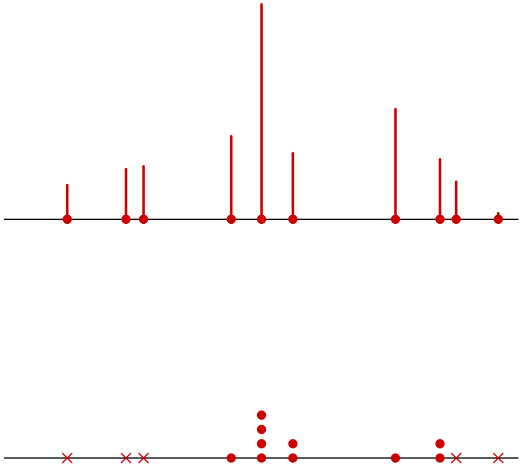


Resampling

Stochastically map continuous weights $(w_t^{(1)}, \dots, w_t^{(N)})$ to discrete offspring counts $(\nu_t^{(1)}, \dots, \nu_t^{(N)})$

Properties:

- Number of particles constant $\sum_{i=1}^N \nu_t^{(i)} = N$
- Equal weights after resampling $w_{t+}^{(i)} = 1/N$
- Unbiased $\mathbb{E}[\nu_t^{(i)} | w_t^{(1:N)}] = N w_t^{(i)}$



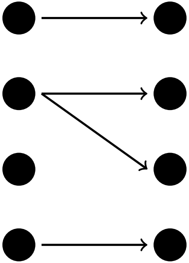
Sequential Monte Carlo

Initialise by sampling N particles from $\mu(\cdot)$.

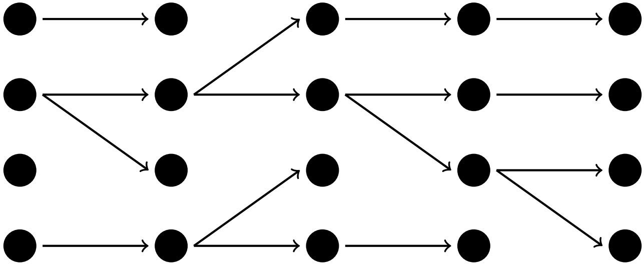
Iterate these steps:

1. **Mutate** particles via Markov transition kernel M_t
2. **Weight** particles by potential function g_t
3. **Resample** particles in proportion to their weights

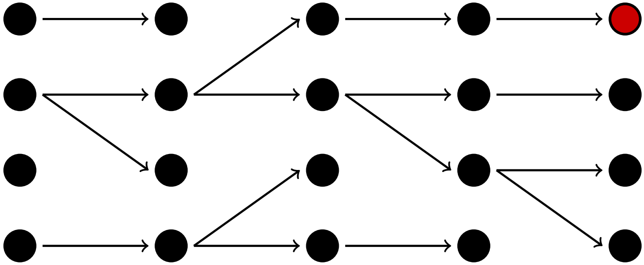
Resampling induces a genealogy



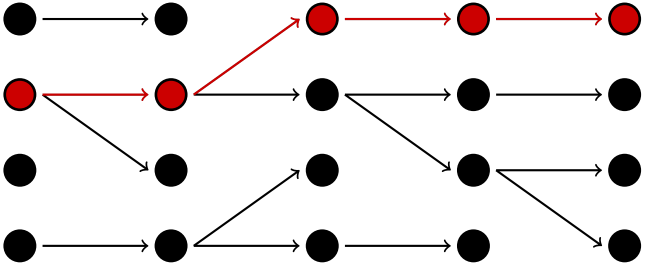
Resampling induces a genealogy



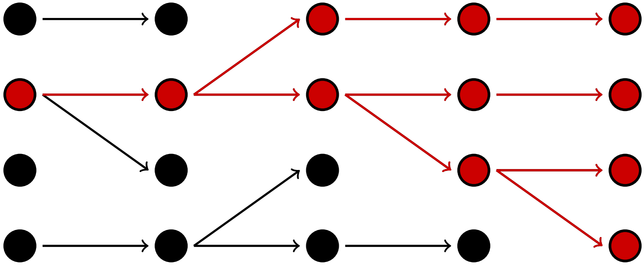
Resampling induces a genealogy



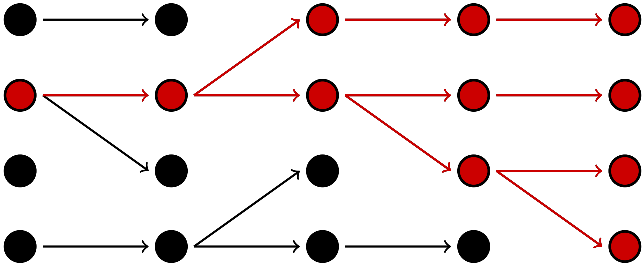
Resampling induces a genealogy



Resampling induces a genealogy



Resampling induces a genealogy

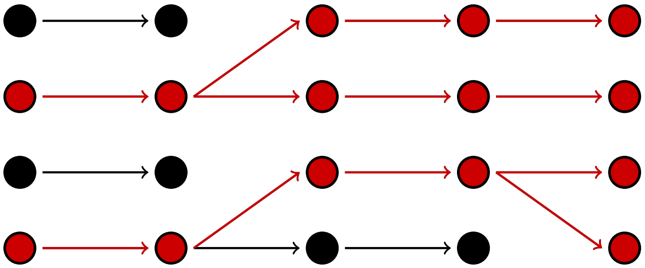


Ancestral degeneracy: for $t << T$, few distinct samples are available

Mitigating ancestral degeneracy

Resample less often?

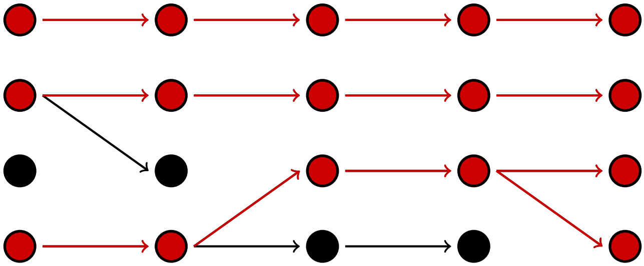
Adaptive resampling: only resample when effective sample size falls below some threshold.



Mitigating ancestral degeneracy

Resample more cleverly?

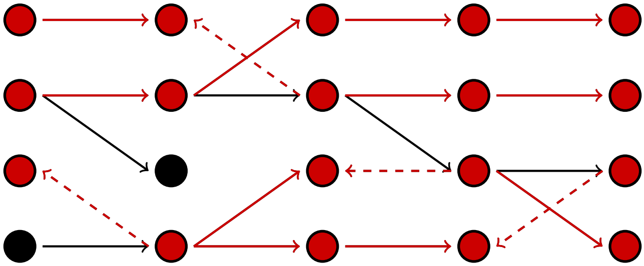
Low-variance resampling: resample in a way that reduces the extra randomness added by the resampling step.



Mitigating ancestral degeneracy

Make use of killed samples?

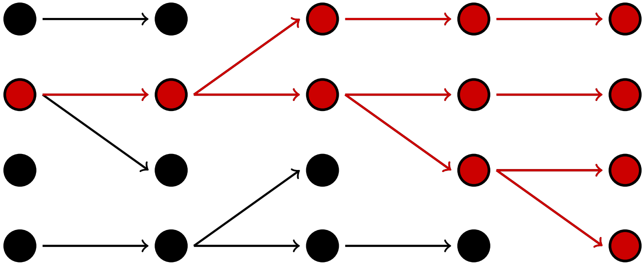
Backward simulation: use a backward pass to sample new ancestors for the terminal particles.



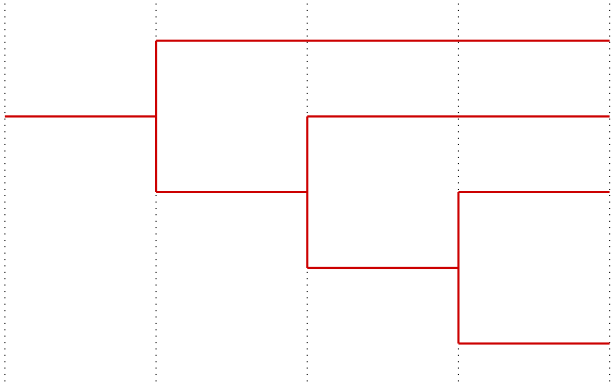
Analysing genealogies

- How many particles should I use to maintain k distinct trajectories across time horizon T?
- How big a lag can I use in fixed-lag smoothing?
- How reliable is my smoothing estimator?
- How do resampling schemes compare?

Encoding genealogies



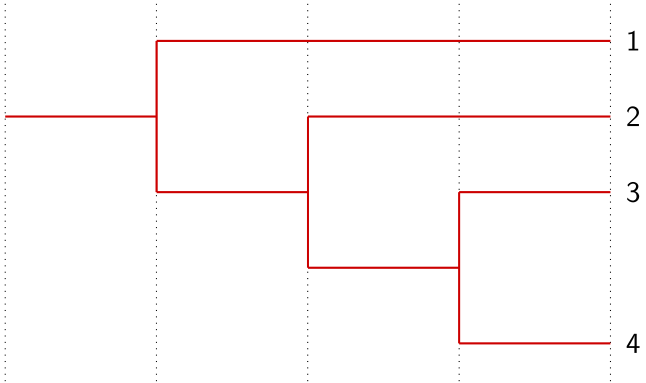
Encoding genealogies



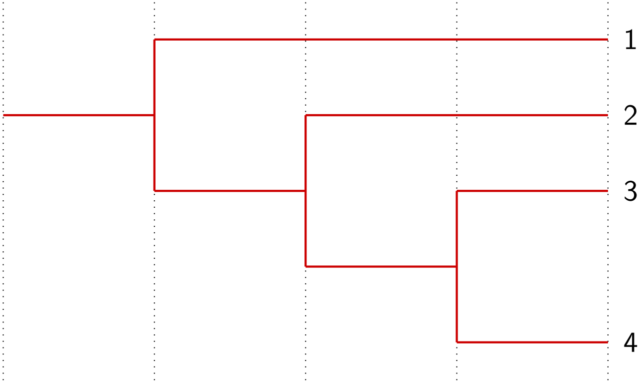
Encoding genealogies

- Label time in reverse
- Population of N particles; $N \to \infty$
- Sample $n \leq N$ terminal particles at random
- Describe genealogy by stochastic process $(G_t^{(n,N)})_{t \in \mathbb{N}_0}$ on space of partitions of $\{1, \ldots, n\}$
- Elements i, j are in the same block of the partition $G_t^{(n,N)}$ iff particles i and j share a common ancestor at time t
- Initially $G_0^{(n,N)} = \{\{1\}, \ldots, \{n\}\}$
- The only possible non-identity transitions are those that merge blocks
- The trivial partition $\{\{1, \ldots, n\}\}$ is an absorbing state

Encoding genealogies



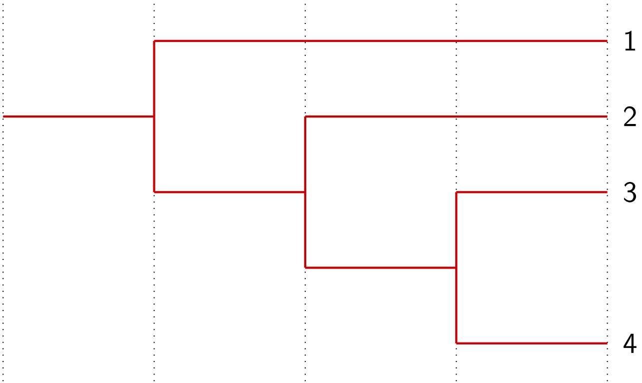
Encoding genealogies



$t = 0$

$\{1\}, \{2\}, \{3\}, \{4\}$

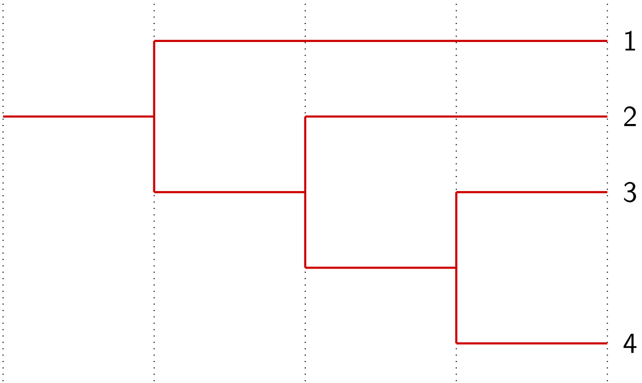
Encoding genealogies



$t = 1$

$\{1\}, \{2\}, \{3, 4\}$

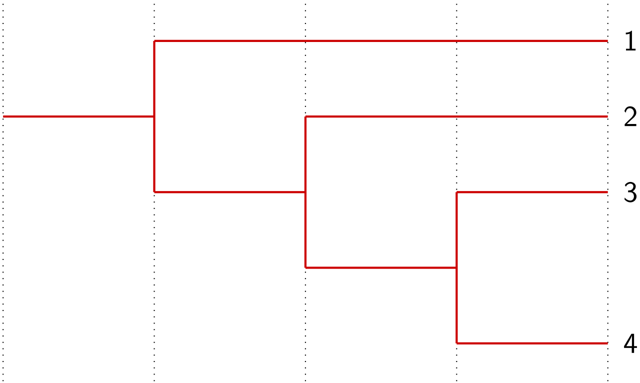
Encoding genealogies



$$t = 2$$

$$\{1\}, \{2, 3, 4\}$$

Encoding genealogies

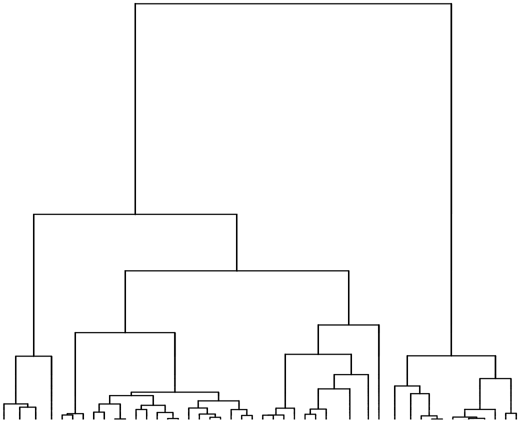


$t = 3$

$\{1, 2, 3, 4\}$

Kingman's n-coalescent[1]

- Continuous-time Markov chain on the space of partitions of $\{1, \ldots, n\}$
- Single pair mergers only
- Each pair merges independently at rate 1 (total merge rate $\binom{k}{2}$ while there are k distinct lineages)



[1] JFC Kingman, *Stochastic Processes & their Applications*, 1982.

Time scale

Pair merger probability conditional on $(\nu_t^{(1)}, \dots, \nu_t^{(N)})$:

$$c_N(t) = \frac{1}{(N)_2} \sum_{i=1}^{N} (\nu_t^{(i)})_2$$

Rescale time by inverse:

$$\tau_N(t) := \inf \left\{ s \geq 1 : \sum_{r=1}^{s} c_N(r) \geq t \right\}$$

Main theorem[2]

Conditions:

- Parent-offspring assignments are uniform given offspring counts
- Time scale does not explode (i.e. $\mathbb{P}[\tau_N(t) = \infty] = 0$ for all finite t)
- There exists a sequence (b_N) such that $\lim_{N\to\infty} b_N = 0$ and

$$\frac{1}{(N)_3}\sum_{i=1}^{N}\mathbb{E}_t[(\nu_t^{(i)})_3] \leq b_N \frac{1}{(N)_2}\sum_{i=1}^{N}\mathbb{E}_t[(\nu_t^{(i)})_2]$$

Then the time-rescaled genealogies $\left(G_{\tau_N(t)}^{(n,N)}\right)_{t\geq 0}$ converge weakly to Kingman's n-coalescent as $N \to \infty$.

[2] S Brown, PA Jenkins, AM Johansen, J Koskela, *Electronic Journal of Probability*, 2021.

Examples

- Multinomial resampling
- Stochastic rounding
- (Conditional SMC with multinomial resampling)

Multinomial resampling

Resample from a Categorical distribution, so offspring counts are Multinomial:

$$(\nu_t^{(1)}, \dots, \nu_t^{(N)}) \sim \text{Multinomial}\left(N, (w_t^{(1)}, \dots, w_t^{(N)})\right)$$

Suppose the transition kernels M_t admit densities m_t, and $\forall x, x'$,

$$\frac{1}{a} \leq g_t(x, x') \leq a \qquad \varepsilon h(x') \leq m_t(x, x') \leq \frac{1}{\varepsilon} h(x')$$

for constants $0 < \varepsilon \leq 1 \leq a < \infty$, and probability distribution $h(\cdot)$.

Then the rescaled genealogies converge to the n-coalescent.[3]

[3] J Koskela, PA Jenkins, AM Johansen, D Spanò. *Annals of Statistics*, 2020.

Stochastic rounding

$\mathbf{Y} : \mathbb{R}_+^N \to \mathbb{N}^N$ is a *stochastic rounding* of $\mathbf{X}$ if for $i = 1, \dots, N$

$$Y_i \mid X_i = \begin{cases} \lfloor X_i \rfloor & \text{with probability } 1 - X_i + \lfloor X_i \rfloor \\ \lfloor X_i \rfloor + 1 & \text{with probability } X_i - \lfloor X_i \rfloor \end{cases}$$

- Take $X_i = N w_t^{(i)}$ and $Y_i = \nu_t^{(i)}$
- By construction $\mathbb{E}[Y_i \mid \mathbf{X}] = X_i$
- Require further constraint $Y_1 + \cdots + Y_N = N$
- **Examples:** systematic resampling, residual-stratified resampling

Stochastic rounding

Resample using any stochastic rounding procedure.

Suppose the transition kernels M_t admit densities m_t, and $\forall x, x'$,

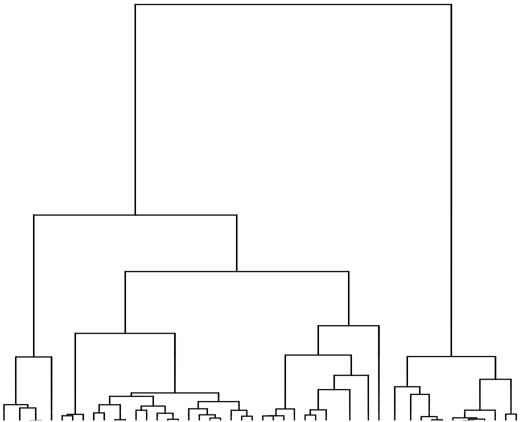
$$\frac{1}{a} \leq g_t(x, x') \leq a \qquad \varepsilon \leq m_t(x, x') \leq \frac{1}{\varepsilon}$$

for constants $0 < \varepsilon \leq 1 \leq a < \infty$.

Then the rescaled genealogies converge to the n-coalescent.

Quantities of interest

- Time to MRCA: first time when there is only one distinct lineage
- t_k: for how long are there exactly k distinct lineages
- Total branch length: storage cost



Comparing resampling schemes

- The expected time scale is the same for every stochastic rounding scheme
- Coalescence is faster for multinomial resampling than for stochastic rounding

Mitigating ancestral degeneracy, revisited

- **Adaptive resampling**: should slow down the time-scale on which the coalescent is recovered (it may have some other effect)
- **Low-variance resampling**: stochastic rounding schemes have minimal variance
- **Backward simulation**: the backward-in-time process is not a pure coalescent, and is not induced by resampling

In conclusion...

- Genealogies can help us to analyse performance of SMC algorithms which suffer ancestral degeneracy
- We have simple conditions under which these genealogies converge to Kingman's n-coalescent
- These conditions are verified for some important classes of SMC algorithms

Open questions

- Other resampling schemes (stratified, residual-multinomial, ...)
- Effect of adaptive resampling
- Estimating the time scale τ_N a priori (since it depends on observed offspring counts)
- Finite-N behaviour

Thank you!

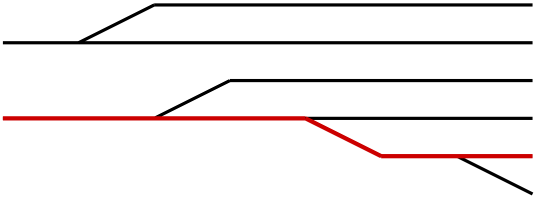
Conditional SMC

A particle Gibbs[4] scenario...

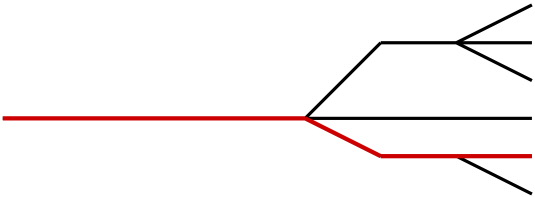
- Want to target $p(\theta, x_{0:T} \mid y_{0:T})$
- Gibbs sampler: alternate samples from
  $p(\theta \mid x_{0:T}, y_{0:T})$ (easy) and
  $p(x_{0:T} \mid \theta, y_{0:T})$ (using SMC)
- Standard SMC updates don't target the correct distribution
- Use SMC updates that are *conditioned* on the previous $X_{0:T}$ trajectory (states and ancestors)
- Resampling must deterministically propagate this "immortal lineage"

[4] C Andrieu, A Doucet, R Holenstein. *Journal of the Royal Statistical Society B*, 2010.

Conditional SMC



okay



not okay

Conditional SMC

Consider a conditional SMC algorithm with multinomial resampling.

Assume

$$\frac{1}{a} \leq g_t(x, x') \leq a \qquad \varepsilon h(x') \leq m_t(x, x') \leq \frac{1}{\varepsilon} h(x').$$

Then the rescaled genealogies converge to the n-coalescent.