Asymptotic Genealogies of sequential Monte Carlo algorithms

Jere Koskela, Paul Jenkins[1], **Adam M. Johansen**[2], and Dario Spanò, all at Department of Statistics, University of Warwick.

arxiv: 1804.01811v2

CIRM — 25.10.2018

Compiled: Wednesday 24th October, 2018: 22:05

[1]Also at Department of Computer Science, University of Warwick
[2]Also at The Alan Turing Institute

CoSInES Launch Day, 2/11/18 at Warwick

.

CoSInES a collaboration of researchers from Warwick, Bristol, Lancaster, Oxford and the Alan Turing Institute to tackle fundamental challenges in Computational and Bayesian Statistics.

Dates The project will run 1st October 2018 till 30th September 2023, is primarily funded by EPSRC.

Launch We'd like to invite anyone who would like to attend to register their interest by emailing Shital Desai (S.Desai.3@warwick.ac.uk).

Jobs! Soon 5 4-year PDRA positions associated with the project based at any of the 5 institutions involved in the grant. Informal enquiries to Gareth Roberts (gareth.o.roberts@warwick.ac.uk) are very welcome.

`http://www.cosines.org`

Outline

Importance sampling

- Intractable target:

$$\mathbb{E}^{\pi}[f(\mathbf{X})] := \int f(\mathbf{x})\pi(d\mathbf{x}).$$

- Monte Carlo: let $\mathbf{X}^{(1:N)} \sim \pi^{\otimes N}$. Then

$$\mathbb{E}^{\pi}[f(\mathbf{X})] \approx \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{X}^{(i)}).$$

- Change of measure:

$$\mathbb{E}^{\pi}[f(\mathbf{X})] = \mathbb{E}^{\mu}\left[\frac{d\pi}{d\mu}(\mathbf{X})f(\mathbf{X})\right] = \int \frac{d\pi}{d\mu}(\mathbf{x})f(\mathbf{x})\mu(d\mathbf{x}).$$

- Importance sampling: let $\mathbf{X}^{(1:N)} \sim \mu^{\otimes N}$. Then

$$\mathbb{E}^{\pi}[f(\mathbf{X})] \approx \frac{1}{N}\sum_{i=1}^{N} \frac{d\pi}{d\mu}(\mathbf{X}^{(i)})f(\mathbf{X}^{(i)}).$$

Sequential Monte Carlo / Interacting Particle System

1: **for** $i \in \{1, \ldots, N\}$ Sample $\mathbf{X}_0^{(i)} \sim \mu_0$.
2: **for** $i \in \{1, \ldots, N\}$ Set

$$w_0^{(i)} \leftarrow \frac{\pi_0(\mathbf{X}_0^{(i)})/\mu_0(\mathbf{X}_0^{(i)})}{\sum_{j=1}^N \pi_0(\mathbf{X}_0^{(j)})/\mu_0(\mathbf{X}_0^{(j)})}.$$

3: **for** $t \in \{1, \ldots, T\}$ **do**
4: Sample $(a_t^{(1)}, \ldots, a_t^{(N)}) \sim \text{Categorical}(w_{t-1}^{(1)}, \ldots, w_{t-1}^{(N)})$.
5: **for** $i \in \{1, \ldots, N\}$ Sample $\mathbf{X}_t^{(i)} \sim \mu_t(\cdot|\mathbf{X}_{t-1}^{(a_t^{(i)})})$.
6: **for** $i \in \{1, \ldots, N\}$ Set

$$w_t^{(i)} \leftarrow \frac{\pi_t(\mathbf{X}_t^{(i)}|\mathbf{X}_{t-1}^{(a_t^{(i)})})/\mu_t(\mathbf{X}_t^{(i)}|\mathbf{X}_{t-1}^{(a_t^{(i)})})}{\sum_{j=1}^N \pi_t(\mathbf{X}_t^{(j)}|\mathbf{X}_{t-1}^{(a_t^{(j)})})/\mu_t(\mathbf{X}_t^{(j)}|\mathbf{X}_{t-1}^{(a_t^{(j)})})}.$$

Example: Hidden Markov Model

- Let $\{\mathbf{X}_t\}_{t\geq 0}$ be a Markov process with transition density $p(\mathbf{x}, \mathbf{x}')$ and initial density $\pi(\mathbf{x})$.
- Suppose a noisy observation $\mathbf{Y}_t$ with density $g(\mathbf{y}|\mathbf{x})$ is made of each state $\mathbf{X}_t$.
- SMC algorithms with

$$\pi_0(\mathbf{x}_0) \propto \pi(\mathbf{x}_0) g(\mathbf{y}_0|\mathbf{x}_0),$$
$$\pi_t(\mathbf{x}_t|\mathbf{x}_{1:(t-1)}) \propto p(\mathbf{x}_{t-1}, \mathbf{x}_t) g(\mathbf{y}_t|\mathbf{x}_t)$$

  target $\mathbb{P}(\mathbf{X}_{0:T} \in d\mathbf{x}_{0:T} | Y_{0:T} = y_{0:T})$.
- E.g. the boostrap particle filter: $\mu_0 = \pi$, $\mu_t(\mathbf{x}_t|\mathbf{x}_{1:(t-1)}) = p(\mathbf{x}_{t-1}, \mathbf{x}_t)$, and $w_t \propto g(\mathbf{y}_t|\mathbf{x}_t)$.

Example: Bootstrap Particle Filter (Gordon et al., 1993)

1: **for** $i \in \{1, \ldots, N\}$ Sample $\mathbf{X}_0^{(i)} \sim \pi$.
2: **for** $i \in \{1, \ldots, N\}$ Set

$$w_0^{(i)} \leftarrow \frac{g(y_0|X_0^{(i)})}{\sum_{j=1}^N g(y_0|X_0^{(j)})}.$$

3: **for** $t \in \{1, \ldots, T\}$ **do**
4: Sample $(a_t^{(1)}, \ldots, a_t^{(N)}) \sim \text{Categorical}(w_{t-1}^{(1)}, \ldots, w_{t-1}^{(N)})$.
5: **for** $i \in \{1, \ldots, N\}$ Sample $\mathbf{X}_t^{(i)} \sim p(\mathbf{X}_{t-1}^{(a_t^{(i)})}, \cdot)$.
6: **for** $i \in \{1, \ldots, N\}$ Set

$$w_t^{(i)} \leftarrow \frac{g(y_t|X_t^{(i)})}{\sum_{j=1}^N g(y_t|X_t^{(j)})}.$$

Path degeneracy

- Suppose $T \gg 1$, and $f(\mathbf{X}_{0:T})$ depends on every time point.
- Mergers due to resampling mean that times $t \ll T$ are estimated from $m \ll N$ paths.
- $\Rightarrow$ High variance estimators.
- Loss of paths also means that fewer than $N \times T$ particles need to be stored, reducing memory cost.
- Aim: *a priori* estimates of:

$$\mathbb{E}[T_{MRCA}], \qquad \text{Var}(T_{MRCA}), \qquad \mathbb{P}(T_{MRCA} > t),$$

etc.

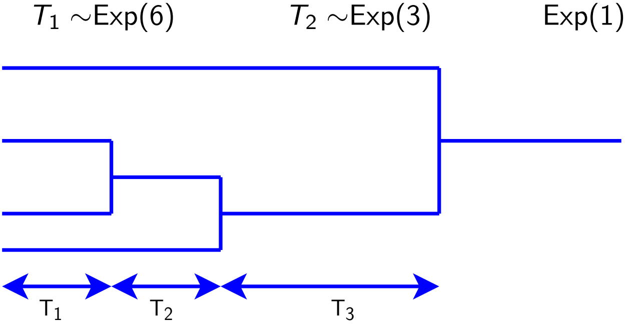
Reasons to Characterize Path Degeneracy include...

- Qualitative understanding of methods.
- Calibrating fixed-lag techniques, e.g. Doucet and Sénécal (2004).
- Relationship with estimator variance (Chan et al., 2013; Lee and Whiteley, 2015).
- Understanding storage requirements (Jacob et al., 2015).

The coalescent process (Kingman, 1982)

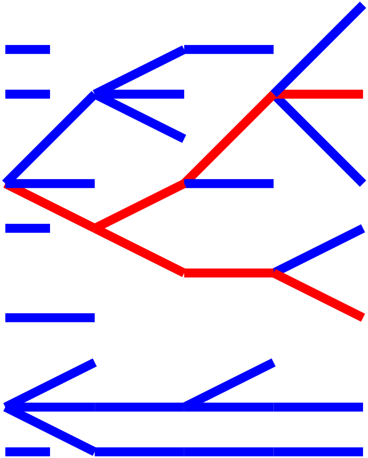
- Let $\{R_t^{(n)}\}_{t\geq 0}$ be a partition-valued process.
- $R_0^{(n)} = \{\{1\},\ldots,\{n\}\}$.
- Each pair of blocks $\{i\},\{j\}$ merge at rate 1.
- A "death" process of rate $\binom{k}{2}$ where k is the number of blocks.

Example: $n = 4$



The genealogical process

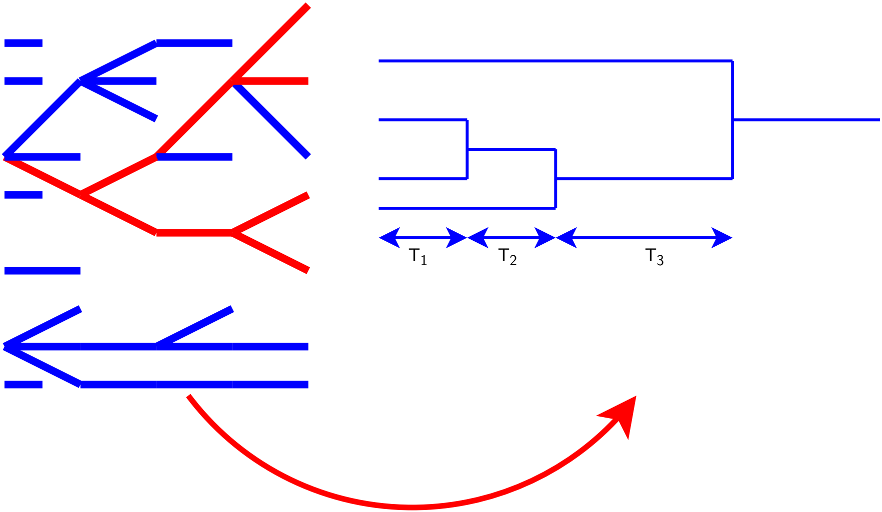
- It is convenient to reverse the direction of time...



- Let $\{G_t^{(n,N)}\}_{t\in\mathbb{N}_0}$ be the genealogy of $n \leq N$ particles sampled randomly from an N-particle SMC algorithm of interest.
- $G_0^{(n,N)} = \{\{1\},\ldots,\{n\}\}$.
- $i \sim j$ in $G_t^{(n,N)} \Rightarrow$ particles i and j have a common ancestor t generations ago.
- $G^{(2,7)}$ illustrated.

Objective: Establish conditions under which

As $N \to \infty$:



Rescaling time

- For $i \in \{1, \dots, N\}$ and $t \in \mathbb{N}$, let $\nu_t^{(i)}$ be the number of children of particle i, t generations ago.
- Define

$$
\begin{aligned}
c_N(t) &:= \frac{1}{(N)_2} \sum_{i=1}^{N} (\nu_t^{(i)})_2 \qquad \approx \mathbb{E}[\mathrm{ESS}(t)^{-1}], \\
\tau_N(t) &:= \inf \left\{ s \geq 1 : \sum_{r=1}^{s} c_N(r) \geq t \right\}, \\
D_N(t) &:= \frac{1}{N(N)_2} \sum_{i=1}^{N} (\nu_t^{(i)})_2 \left(\nu_t^{(i)} + \frac{1}{N} \sum_{j \neq i}^{N} (\nu_t^{(j)})^2 \right).
\end{aligned}
$$

Convergence theorem

Suppose that all assignments of offspring to parents are equally likely, and that

$$\lim_{N\to\infty} \mathbb{E}\left[\sum_{r=\tau_N(s)+1}^{\tau_N(t)} D_N(r)\right] = 0,$$

$$\lim_{N\to\infty} \mathbb{E}[c_N(t)] = 0,$$

$$\lim_{N\to\infty} \mathbb{E}\left[\sum_{r=\tau_N(s)+1}^{\tau_N(t)} c_N(r)^2\right] = 0,$$

$$\mathbb{E}[\tau_N(t) - \tau_N(s)] \leq C_{t,s} N.$$

Then $(G^{(n,N)}_{\tau_N(t)})_{t\geq 0}$ converges to the Kingman coalescent in the sense of finite dimensional distributions.

Proof outline

- Consider finite dimensional distributions.
- Apply straightforward, but intricate counting arguments,
- together with bounds on elementary transition probabilities,
- to upper and lower bound the elements of the FDDs.
- Compare these with those of the coalescent.

Sketch proof

- Let ξ and η be partitions of $\{1, \ldots, n\}$, with the block counts of η in terms of the blocks of ξ being $b_1, \ldots, b_{|\eta|}$, i.e. $b_1 + \ldots + b_{|\eta|} = |\xi|$.
- The conditional one-step transition probability of $G_t^{(n,N)}$ given family sizes is

$$p_{\xi\eta}(t) := \frac{1}{(N)_{|\xi|}} \sum_{i_1 \neq \ldots \neq i_{|\eta|}=1}^{N} (\nu_t^{(i_1)})_{b_1} \ldots (\nu_t^{(i_{|\eta|})})_{b_{|\eta|}}.$$

- FDDs:

$$\mathbb{P}(G_{\tau_N(t_1)}^{(n,N)} = \eta_1, \ldots, G_{\tau_N(t_k)}^{(n,N)} = \eta_k | G_{\tau_N(t_0)}^{(n,N)} = \eta_0)$$
$$= \mathbb{E}\left[\prod_{d=1}^{k} \left\{ \prod_{r=\tau_N(t_{d-1})+1}^{\tau_N(t_d)} P_N(r) \right\}_{\eta_{d-1}\eta_d} \right].$$

Sketch proof II

- For a single time interval

$$\left\{ \prod_{r=\tau_N(t_{d-1})+1}^{\tau_N(t_d)} P_N(r) \right\}_{\eta_{d-1}\eta_d} = \sum_{\boldsymbol{\xi}} \prod_{s=\tau_N(t_{d-1})+1}^{\tau_N(t_d)} p_{\xi_{s-1}\xi_s}(s),$$

  where $\boldsymbol{\xi} = (\eta_{d-1}, \xi_{\tau_N(t_{d-1})+1}, \ldots, \xi_{\tau_N(t_d)-1}, \eta_d)$.
- Each partition in $\boldsymbol{\xi}$ is either equal to its predecessor, or obtained from its predeceror by merging some subset(s) of blocks.

Sketch proof III

$$p_{\xi\xi}(t) \approx 1 - \binom{|\xi|}{2}\frac{1}{(N)_2} - \binom{|\xi|}{2}c_N(t).$$

If η is formed by merging two blocks of ξ,

$$c_N(t) - \binom{|\xi|-2}{2}D_N(t) \lesssim p_{\xi\eta}(t) \lesssim c_N(t).$$

If η is formed by merging more than two blocks of ξ,

$$p_{\xi\eta}(t) \lesssim \binom{|\xi|-2}{2}D_N(t).$$

Sketch proof IV

$$\sum_{\xi} \prod_{s=\tau_N(t_{d-1})+1}^{\tau_N(t_d)} p_{\xi_{s-1}\xi_s}(s) \approx \sum_{\alpha=1}^{|\eta_{d-1}|-|\eta_d|} \sum_{(\lambda,\mu)\in\Pi_2([\alpha])} C$$

$$\times \sum_{s_1<\ldots<s_\alpha=\tau_N(t_{d-1})+1}^{\tau_N(t_d)} \left\{\prod_{r\in\lambda} c_N(s_r)\right\} \left\{\prod_{r\in\mu} D_N(s_r)\right\},$$

$$D_N(t) \approx \frac{c_N(t)}{N},$$

$$\sum_{s<\ldots<s_\alpha=\tau_N(t_{d-1})+1}^{\tau_N(t_d)} \prod_{r=1}^{\alpha} c_N(s_r) \approx \frac{(t_d - t_{d-1})^\alpha}{\alpha!}.$$

Sketch proof V

When $\boldsymbol{\xi}$ consists of binary mergers only, i.e. $\alpha = |\eta_{d-1}| - |\eta_d|$,

$$\sum_{\boldsymbol{\xi}} \prod_{s=\tau_N(t_{d-1})+1}^{\tau_N(t_d)} p_{\xi_{s-1}\xi_s}(s)$$

$$\approx C' \sum_{s_1 < \ldots < s_\alpha = \tau_N(t_{d-1})+1}^{\tau_N(t_d)} \left\{ \prod_{r=1}^{\alpha} c_N(s_r) \right\} \prod_{r=\tau_N(t_{d-1})+1}^{\tau_N(t_d)} \{1 - C'' c_N(r)\}$$

$$\approx \sum_{\beta=0}^{\tau_N(t_d)-\tau_N(t_{d-1})-\alpha} C''' \sum_{s_1 < \ldots < s_{\alpha+\beta} = \tau_N(t_{d-1})+1}^{\tau_N(t_d)} \prod_{r=1}^{\alpha+\beta} c_N(s_r).$$

Sketch proof VI

It turns out that the constant C''' is *exactly* $(Q^{\alpha+\beta})_{\eta_{d-1}\eta_d}$, where Q is the Kingman coalescent generator!

$$
\begin{aligned}
&\sum_{\xi} \prod_{s=\tau_N(t_{d-1})+1}^{\tau_N(t_d)} p_{\xi_{s-1}\xi_s}(s) \\
&\approx \sum_{\beta=0}^{\tau_N(t_d)-\tau_N(t_{d-1})-\alpha} C''' \sum_{s_1<\ldots<s_{\alpha+\beta}=\tau_N(t_{d-1})+1}^{\tau_N(t_d)} \prod_{r=1}^{\alpha+\beta} c_N(s_r) \\
&\approx \sum_{\beta=0}^{\tau_N(t_d)-\tau_N(t_{d-1})-\alpha} (Q^{\alpha+\beta})_{\eta_{d-1}\eta_d} \frac{(t_d - t_{d-1})^{\alpha+\beta}}{(\alpha+\beta)!} \\
&\approx (e^{Q(t_d-t_{d-1})})_{\eta_{d-1}\eta_d}.
\end{aligned}
$$

Corollary 1

The genealogy of n particles sampled uniformly at random from an N-particle bootstrap particle filter with multinomial resampling converges to a Kingman coalescent under the time-scaling $\tau_N(t)$ whenever

$$\frac{1}{a} \leq g(y_t|x_t) \leq a,$$
$$\varepsilon h(x_t) \leq p(x_{t-1}, x_t) \leq \frac{1}{\varepsilon} h(x_t),$$

for some $0 < \varepsilon \leq 1 \leq a < \infty$, and some probability density $h(x)$ uniformly in time, space, and the observations.

Sketch proof

- Conditional on weights, the offspring counts have multinomial distributions with parameters $(N; w_t^{(1)}, \ldots, w_t^{(N)})$.
- Upper and lower bounds on observation densities imply

$$\frac{\varepsilon^2}{a^2 N} \leq w_t^{(i)} \leq \frac{a^2}{\varepsilon^2 N}.$$

- The required upper and lower bounds follow from these bounds, standard moment-calculations for multinomial distributions, and the local dependence structure of particle filters.

Corollary 2

Let T_n be the total Kingman coalescent tree height of n particles. Under the preceding assumptions,

$$\frac{2\varepsilon^4 N}{a^4}\left(1-\frac{1}{n}\right) \leq \mathbb{E}[\tau_N(T_n)] \leq \frac{2\varepsilon^4 N}{a^4}\left(1-\frac{1}{n}\right) + \frac{a^8}{\varepsilon^4},$$

$$\begin{aligned}
&\frac{N^2\varepsilon^8}{a^8}\left(\frac{4\pi^2}{3} - 12 + O(n^{-1})\right) \\
\leq &\,\mathrm{Var}(\tau_N(T_n)) \\
\leq &\frac{N^2 a^8}{\varepsilon^8}\left(\frac{4\pi^2}{3} - 12 + O(n^{-1})\right) + O(N).
\end{aligned}$$

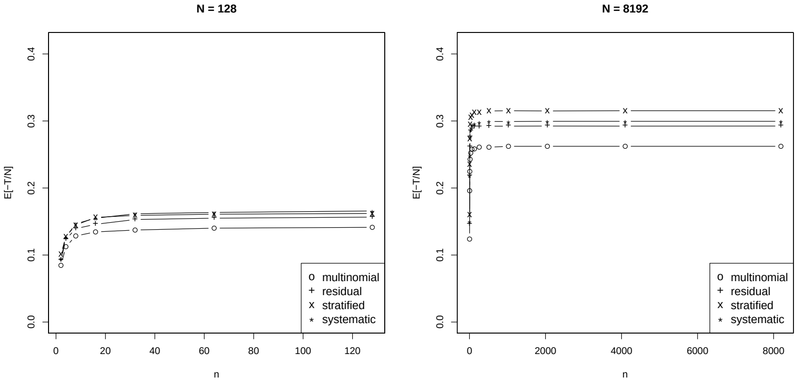
A numerical example

- Take the earlier HMM to be

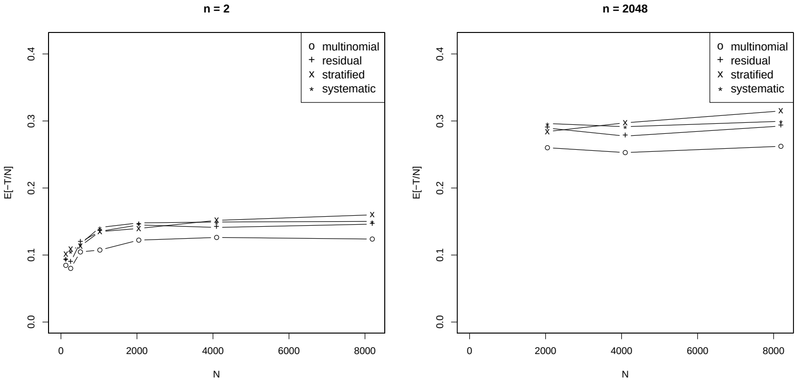
$$
\begin{aligned}
X_{t+1} &= (1-\Delta)X_t + \sqrt{\Delta}\xi_t \\
X_0 &\sim N(0,1), \\
Y_t | X_t &\sim N(X_t, \sigma^2),
\end{aligned}
$$

  where $\xi_t \sim N(0,1)$.
- This model violates the assumed lower bounds.
- Nevertheless, simulations using a boostrap particle filter show that the Kingman scalings hold, even when $n = N$.

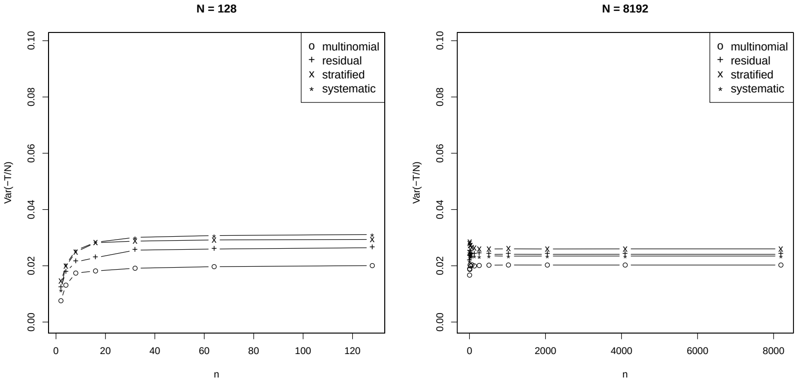
Mean tree height



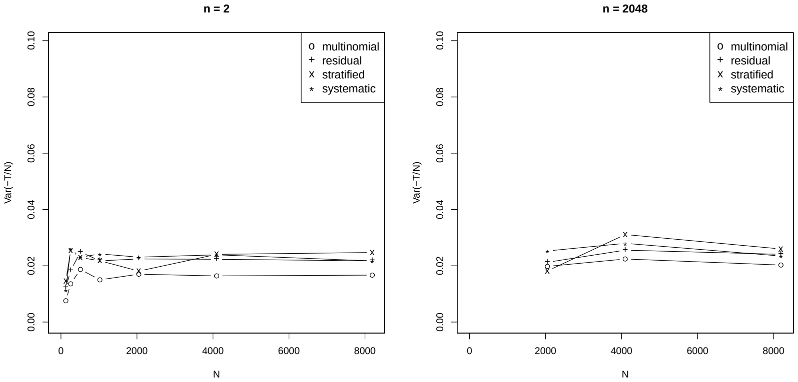
Mean tree height II



Tree height variance



Tree height variance II



Conclusions

- Genealogies of $n \ll N$ particles from N-particle SMC algorithms converge to the Kingman coalescent when time is measured in units of N, as $N \to \infty$.
- Strong technical assumptions (i.e. a compact state space) which do not seem necessary in practice.
- Predicted scalings observed in experiments for finite N, seem to hold even when $n \approx N$.
- Result holds for multinomial resampling, but other schemes agree with predictions empirically.
- This result also demonstrates that the domain of attraction of the Kingman coalescent includes certain non-Markovian genealogies.

Outlook

- Some areas in which genealogical results might be interesting:
  - Variance estimation from SMC output (Lee and Whiteley, 2015).
  - Smoothing and static parameter estimation.
  - Mixing in particle Gibbs/iterated cSMC.
- Room for improvement (selected topics...)
  - Relaxing strong assumptions.
  - Incorporating other resampling schemes.
  - Obtaining stronger modes of convergence.
  - (Formal analysis of $n \approx N$.)
  - Incorporating conditional SMC.

References

H. P. Chan, T. L. Lai, A general theory of particle filters in hidden Markov models and some applications. *Annals of Statistics*, 41 (6):2877–2904, 2013.

A. Doucet and Stéphane Sénécal. Fixed-lag sequential Monte Carlo. In *EURASIPCO*, 2004.

N. J. Gordon, S. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, 140(2):107–113, April 1993.

P. E. Jacob, L. Murray, and S. Rubenthaler. Path storage in the particle filter. *Statistics and Computing*, 25(2):487–496, 2015.

J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, 1982.

J. Koskela, P. Jenkins, A. M. Johansen, and D. Spanò. Asymptotic genealogies of interacting particle systems with an application to sequential Monte Carlo. ArXiv mathematics e-print 1804.01811, ArXiv Mathematics e-prints, April 2018.

A. Lee and N. Whiteley. Variance estimation and allocation in the particle filter. Technical Report 1509.00394, Arxiv, 2015.

Acknowledgments

- Paul Jenkins was supported in part by EPSRC grant EP/L018497/1.
- Adam Johansen was partially supported by funding from the Lloyd's Register Foundation — Alan Turing Institute Programme on Data-Centric Engineering.
- Jere Koskela was supported by EPSRC as part of the MASDOC DTC at the University of Warwick (Grant No. EP/HO23364/1), and by Deutsche Forschungsgemeinschaft (DFG) grant BL 1105/3-2.