Resampling and genealogies in sequential Monte Carlo algorithms

Susanna Elizabeth Brown

A thesis submitted for the degree of
Doctor of Philosophy in Statistics

Department of Statistics
University of Warwick

August 2021

Abstract

This thesis attempts to quantify the problem of ancestral degeneracy of sequential Monte Carlo samples, which is known to have a critical effect on the performance of the resulting estimators. To facilitate comparisons between different algorithms, the induced genealogical processes are analysed under an asymptotic regime in which the number of particles tends to infinity. Simple conditions are derived under which these genealogical processes converge weakly to Kingman's well-studied n-coalescent, with a certain time change. These sufficient conditions are verified for the many of the most popular sequential Monte Carlo algorithms, giving a novel insight into the large-sample behaviour of the associated estimators. The asymptotic regime serves to unify these different algorithms in one framework, the genealogical differences between the algorithms then being fully captured by the respective time-change functions. The results also have implications in theoretical population genetics, where the processes studied may be seen as population models incorporating selection. Our main result then comprises a novel weak convergence theorem for genealogies arising from non-neutral populations.

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

Parts of this thesis have been published by the author. Some results of Chapters 3 and 5 appear in condensed form in the article
Suzie Brown et al. (2021). "Simple Conditions for Convergence of Sequential Monte Carlo Genealogies with Applications". In: *Electronic Journal of Probability* 26.1, pp. 1–22. ISSN: 1083-6489. DOI: 10.1214/20-EJP561

The work presented (including data generated and data analysis) was carried out by the author except in the cases outlined below:
The final part of Section 3.3.1, comprising the modification to the proof of Koskela et al. (2018), was lifted from the aforementioned article (Brown et al. 2021) where the calculations were primarily carried out by collaborators, and subsequently edited by myself to improve readability and consistency of presentation.

Acknowledgements

Firstly I would like to thank my supervisors, Dr. Paul Jenkins, Dr. Jere Koskela and Prof. Adam Johansen, who for the past three years have guided me through swamps of calculations and done your best to keep my focus on things that are important. Your combined expertise in all things Monte Carlo and population genetics have been invaluable, and your humour and general cynicism have enlightened many a dreary week. Thank you also to those people sitting near me in the office, who have been called upon many times to stupid-check my calculations — Francesca Crucinio, James Hodgson and Marco Palma — your enduring patience is gratefully received. I am also very grateful to Dr. Dario Spanò and Prof. Martin Möhle for agreeing to examine this thesis — perhaps you didn't know what you were getting into!

I also owe my thanks to the EPSRC for providing funding (under grant EP/L016710/1), without which I would not have been able to embark upon this PhD, and to the organisers of the OxWaSP CDT, which equipped me with a broad knowledge of modern statistics and gave me the confidence to complete this research. Thanks also to Prof. Oliver Johnson and Prof. Jonty Rougier for encouraging me to continue my academic studies and to apply for this PhD programme.

Thank you to all of the colleagues who helped to create such a fun and stimulating research environment at Warwick: fellow OxWaSP students, office-mates and the wider young researchers community in the department. Special thanks to Ana Ignatieva and Jaro Sant for providing the bastion of continuity that is the population genetics reading group. A very special thank you must go to the bridge club — Jack Carter, Francesca Crucinio, Giulio Morina, Marco Palma and William Thomas — for your best efforts in preventing me from finishing my thesis. Perhaps it is for the best that a pandemic put an end to our excessively long lunch breaks.

I am grateful to James Brixey and Katie Farnes for your prayers and support during the writing-up phase, for keeping me sane during the lockdowns, and for many interesting discussions on unrelated topics! A heartfelt thank you to my church family for your hospitality, friendship and prayers throughout the "year of plenty" and beyond, especially to Charissa Brain and the other Sunday Zoom regulars, as well as the Griffithses, Kibbles, Murphies and many others who have made Coventry feel like home. I've learnt at least as much from you lovely people as from my research.

To the other friends I've found during these four years — Francesca Panero, Claire Atkins, Ann Cordery, Claire Silvester and Rachel G-H — I'm very glad to have met you, and I'm sure that our friendship will continue. And of course, a heartfelt thank you to those friends and family who have been with me since before the start of this chapter and continue to stand by me: Mum, Dad, Nick, Sheriff, Jamie, Rhys, Julia and the Wiggles, to name a few.

I am so grateful to have had the opportunity to spend four years doing something I love, surrounded by brilliant people, and to reach the end having accomplished all that I hoped. I have learnt so much and grown in so many ways over this time, and I can't wait for the next chapter!

S.D.G.

Suzie Brown
29 July 2021

Contents

List of Figures

List of Tables

List of Algorithms

Abbreviations

CDF	cumulative distribution function
i.i.d.	independent and identically distributed
MRCA	most recent common ancestor
MVB	minimal variance branching
PRNG	pseudo-random number generator
SMC	sequential Monte Carlo
SSP	Srinivasan sampling process

Notation and Conventions

$\mathbb{N}$	the natural numbers starting from one, $\{1, 2, \dots\}$
$\mathbb{N}_0$	the natural numbers starting from zero, $\{0, 1, 2, \dots\}$
$\mathcal{P}_n$	the set of partitions of $\{1, \dots, n\}$
$[a]$	the set $\{1, 2, \dots, a\}$ where $a \in \mathbb{N}$, or the empty set if $a = 0$
$a : b$	the set $\{a, a+1, \dots, b\}$ where $a \leq b \in \mathbb{N}$, defined to be the empty set when $a > b$
$\mathcal{S}_k$	the k-dimensional unit simplex $\{x_{1:k+1} \geq 0 : \sum_{i=1}^{k+1} x_i = 1\}$
x_A	the subvector consisting of the elements of x with index in set $A \subseteq \mathbb{N}$
x_{-a}	the subvector x_A where $A = \{1, 2, \dots, a-1, a+1, \dots n\}$, $a \in \{1, \dots, n\}$, and n is the length of x which should be clear from context
$(a)_b$	the falling factorial $a(a-1)\cdots(a-b+1)$ where $a \in \mathbb{N}_0, b \in \mathbb{N}$, and define $(a)_0 = 1$
$\binom{a}{b}$	binomial coefficient where $a, b \in \mathbb{N}_0$, defined to be 0 when $a < b$
$a \wedge b$	the minimum of a and b
$\prod_\emptyset$	the empty product is taken to be 1
$\sum_\emptyset$	the empty sum is taken to be 0, while the sum over an index vector of length zero is the identity operator
$\mathcal{F}_t$	the (backward) filtration generated by offspring counts up to time t
$\mathbb{E}$	expectation
$\mathbb{E}_t$	filtered expectation $\mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$
$\mathbb{E}_\mathbb{P}$	expectation with respect to a specific probability measure $\mathbb{P}$
Var	variance
Cov	covariance
$\overset{d}{=}$	equal in distribution
$\sim^{iid}$	sampled i.i.d. from

A^c	the complement of set A
$\|A\|$	the cardinality of set A
$O(\cdot)$	standard asymptotic notation: $f(x) = O(g(x))$ if there exist $M \in [0, \infty), x_0 \in \mathbb{R}$ such that $f(x) \leq Mg(x)$ for all $x \geq x_0$
$o(\cdot)$	standard asymptotic notation: $f(x) = o(g(x))$ if for all $\epsilon > 0$ there exists x_0 such that for all $x \geq x_0$, $f(x) \leq \epsilon g(x)$
$\Omega(\cdot)$	standard asymptotic notation: $f(x) = \Omega(g(x))$ if and only if $g(x) = o(f(x))$
1_N	asymptotic notation for a sequence that converges to 1 as $N \to \infty$

1 Introduction

I wonder why. I wonder why.
I wonder why I wonder.
I wonder *why* I wonder why
I wonder why I wonder!

Richard P. Feynman

Since their introduction in the 1990s, sequential Monte Carlo (SMC) methods, sometimes known as particle filters, have found applications in virtually every branch of science. This is due to the ubiquity of the types of problems in which SMC is most powerful. As more and more data are collected and scientific models made ever more complex, practitioners are frequently reaching for numerical methods to solve problems. SMC is a likely candidate whenever the aim is to make inferences from sequentially-observed data. Moreover, SMC is used as a tool to speed up other numerical methods, by artificially introducing some sequential structure, for instance: tempering to enable Monte Carlo sampling from multimodal distributions; constructing nested sequences of events to enable rare event simulation; or sequentially decreasing the tolerance level in approximate Bayesian computation.

Almost three decades of study have produced a menagerie of variations on the standard "bootstrap" SMC algorithm, along with a deeper understanding of their theoretical underpinnings. Even so, the problems to which SMC is applied are inherently hard, so there are still problems to overcome. One such unresolved issue, which is the primary concern of the current work, is that of ancestral or path degeneracy, which is described in Section 2.3.1. Although this problem was noted in the original article on SMC by Gordon, Salmond, and Smith (1993), it still has not been adequately solved.

The current work makes no attempt to provide solutions to the problem of ancestral degeneracy. The focus is instead on analysing and quantifying it, using a combination of techniques from the SMC and population genetics literatures. The hope is that, equipped with more information about this phenomenon, the practitioner will be able to make better judgements about their choice of algorithm and tuning parameters, and how much trust they should put in the resulting estimates.

The bulk of the thesis is divided into four chapters. Chapter 2 provides the relevant background on sequential Monte Carlo and coalescent theory, and explains in more detail the relevance of genealogies to the study of SMC algorithms. It also includes a detailed comparison of the most important "resampling schemes" in the SMC literature, in terms of various properties of interest. Most of the results included are well-known, but Section 2.4.3 provides a more complete summary than can be found elsewhere in the literature.

Chapter 3 sets up the framework for the asymptotic analysis of genealogies, and presents the first result (Theorem 3.6), a sufficient condition for convergence of finite-dimensional distributions to those of Kingman's n-coalescent (Section 2.2.1). The proof of the theorem builds on a related result of Koskela et al. (2018), which is reviewed in Section 3.2.

In Chapter 4 it is shown that under the same sufficient conditions, the processes under consideration also converge *weakly* to the n-coalescent (Theorem 4.1), the first weak convergence result for SMC genealogies. This is a stronger result than that of Chapter 3, additionally requiring tightness of the processes.

Chapter 5 consists of a series of corollaries, each of which verifies the theorem conditions for a particular class of SMC algorithms. This includes the majority of SMC algorithms commonly used by practitioners.

2 Background

> Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin.
>
> — John von Neumann

2.1 Sequential Monte Carlo

The idea of Monte Carlo is to use (pseudo-)random numbers to approximate expectations under an intractable probability distribution of interest. Sequential Monte Carlo (SMC) is a class of Monte Carlo algorithms which are implemented sequentially, allowing efficient sampling from sequences of distributions. SMC was developed for inference in intractable state space models (details in Section 2.1.1) and introduced to the statistics community by Gordon, Salmond, and Smith (1993). The basic idea behind SMC is that of sequential importance sampling, whereby the importance samples from one target distribution are used to generate proposals for the next. A full derivation of the SMC recursions is beyond the scope of this work, but the reader is referred to e.g. Chopin and Papaspiliopoulos (2020) and Doucet and Johansen (2011) for more background. Here it suffices to provide a motivation in the context of state space models (Section 2.1.1) and the formalism of Feynman-Kac models (Section 2.1.3).

2.1.1 State space models

State space models (sometimes called hidden Markov models) are a flexible class of statistical models which are suitable in all sorts of applications where observations appear sequentially. The general model has two components: a Markov process $(X_t)_{t\in\mathbb{N}_0}$ representing the (unobservable) underlying state of the system, and a sequence $(Y_t)_{t\in\mathbb{N}_0}$ of observations containing information about the underlying state. The model is characterised by its conditional independence structure (Figure 2.1) along with an initial distribution μ, the Markov *transition* kernels $(K_t)_{t\in\mathbb{N}}$ and the *emission* distributions $(g_t)_{t\in\mathbb{N}_0}$. Written

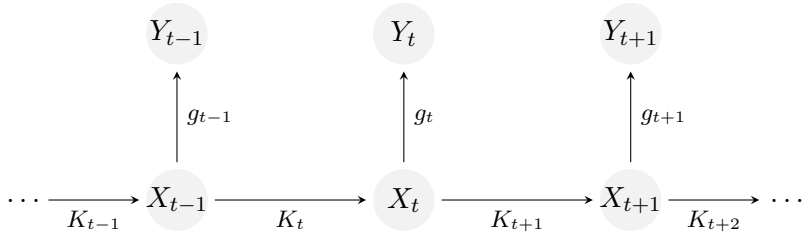


Figure 2.1: Conditional independence graph for a general state space model. (X_t) is a Markov process with transition kernels (K_t) representing the underlying state of the system. Y_t is a noisy observation of X_t for each t.

as a hierarchical model,

$$\begin{aligned} X_0 &\sim \mu(\cdot) \\ X_{t+1} \mid X_t &\sim K_{t+1}(\cdot|X_t) && \text{for } t = 0, 1, \ldots \\ Y_t \mid X_t &\sim g_t(\cdot|X_t) && \text{for } t = 0, 1, \ldots \end{aligned} \tag{2.1}$$

The index t will frequently be referred to as time, since in many applications the sequence is indeed a time series, but it need not be.

Here X and/or Y may be multivariate and observation times need not be equally spaced. Straightforward generalisations of the stated model can allow for situations in which observations are not available as often as the state is updated (up to and including the extreme where the state is a continuous-time Markov process but the observations are available only at discrete times) or on the other hand where observations are made more frequently than the state is updated.

Applications include target tracking, where X is the true position of some object and Y encodes some measurements from sensors e.g. radar; stochastic volatility models, where X is the volatility and Y is the observed value e.g. the price of a stock; change-point detection; and many other situations in which there is an observed time series from which one would like to do inference or prediction.

The principal inferences of interest in state space models are:

filtering $p(x_t \mid y_{0:t})$: inferring the current state x_t from the observations up to now $y_{0:t}$

prediction $p(x_{t+h} \mid y_{0:t})$: inferring a future state x_{t+h} from the observations up to now $y_{0:t}$

(complete) smoothing $p(x_{0:t} \mid y_{0:t})$: inferring the sequence of states up to now $x_{0:t}$ from the observations up to now $y_{0:t}$

fixed-lag smoothing $p(x_{t-h:t} \mid y_{0:t})$: inferring the last h states $x_{t-h:t}$ from the observations up to now $y_{0:t}$

If the dynamics of the state space model are parametrised by some θ, i.e. g_t and/or K_t depend on θ, we may also be interested in parameter inference or computing the likelihood

$p(y_{0:t})$ of the observed data for particular values of θ. Such a model is considered in Section 2.5.1.

In certain cases, these inference problems may be solved analytically (Section 2.1.2), but this is not typically the case. For intractable models we must resort to numerical methods such as Monte Carlo. However, state space inference is problematic even with Monte Carlo. The main difficulties are that the dimension of the target distributions may increase along the sequence, and that there is strong dependence between consecutive distributions. Markov chain Monte Carlo (MCMC), for instance, is known to struggle with highly correlated targetsand its performance drops drastically as dimension increases, despite convergence rates that are independent of dimension.

As we will see in Section 2.1.4, sequential Monte Carlo somewhat overcomes these problems, turning the problematic properties of the target distribution to its benefit. Correlation between consecutive targets is exploited for sequential updating, which is able to handle the incrementing dimensionality. The resulting linear-in-t computational complexity also allows inference to be performed on-line, that is, updating the posterior distribution(s) as observations arrive.

2.1.2 Inference in state space models

If the state space model has linear dynamics with Gaussian noise, the posterior distributions of interest are also Gaussian. The posterior mean and covariance satisfy certain recursions, implemented by the Kalman filter (Kalman 1960) and Rauch-Tung-Striebel smoother (Rauch, Striebel, and Tung 1965). Recursions are also available for some other conjugate models: see for example Kon Kam King, Papaspiliopoulos, and Ruggiero (2021) and Vidoni (1999). Another analytic case occurs if the state space $\mathcal{X}$ is finite, in which case any integrals become finite sums, and the forward-backward algorithm (Baum et al. 1970) yields the exact posteriors. However, if the state space becomes large, albeit finite, exact computation becomes infeasible.

If the model is Gaussian but non-linear, the posterior filtering distributions can be estimated using the *extended Kalman filter* (see for example Jazwinski (2007)), which applies a first-order approximation in order to make use of the Kalman filter. This method performs well on models that are "almost linear". The resulting predictor is only *optimal* when the model is actually linear, in which case the extended Kalman filter coincides with the Kalman filter.

For models that are high-dimensional or highly non-linear or for which gradients are not readily available, the exact Kalman filter updates can be replaced by sample approximations. The *ensemble Kalman filter* (Evensen 1994) uses a Monte Carlo sample from the current time, propagates these points through the transition dynamics, and uses the sample covariance as an estimator of the updated covariance matrix. The means, which are cheaper to evaluate and more stable than the covariances, are still updated using the Kalman filter recursion, based on the estimated covariance. The *unscented Kalman filter*

(Wan and Merwe 2000) uses a deterministic sample chosen via the *unscented transformation*, which is then propagated through the non-linear transition kernel to obtain a characterisation of the distribution at the next time step. The sample consists of $2d + 1$ points, where d is the dimension of the state space, and defines a Gaussian approximation to the updated distribution. If the model is really linear-Gaussian then the sample points are sufficient to recover the correct distribution.

In complex or high-dimensional models, exact inference is often infeasible, and we turn instead to Monte Carlo methods. Markov chain Monte Carlo performs woefully on state space models due to the high dimension of the parameter space and high correlation between dimensions. But we can exploit the sequential nature of the underlying dynamics to decompose the problem into a sequence of inferences of fixed dimension. This is the motivation behind sequential Monte Carlo (SMC).

2.1.3 Feynman-Kac models

State space models are very natural and intuitive applications, but they do not do justice to the scope of SMC algorithms, which is much wider. On the other hand, every SMC algorithm is a Monte Carlo approximation of some *Feynman-Kac* model. Before formally introducing SMC let us therefore define a generic Feynman-Kac model. For a more in-depth study, the reader is directed to the exhaustive books by Del Moral (2004, 2013) or the more accessible Chopin and Papaspiliopoulos (2020, Chapter 5).

Define a state space $\mathcal{X}$, which in this presentation we assume to be common for all times: this is often not the case in practice, but the generalisation to a sequence of state spaces is straightforward. The basic components of the Feynman-Kac model are a Markov law, defined by an initial distribution $\mathbb{M}_0$ on $\mathcal{X}$ and transition kernels $M_t : \mathcal{X} \mapsto \mathcal{X}$ for $t \in \mathbb{N}$; and a sequence of *potential* functions $G_0 : \mathcal{X} \mapsto [0, \infty)$ and $G_t : \mathcal{X}^2 \mapsto [0, \infty)$ for $t \in \mathbb{N}$. From these we can construct, for any time horizon T, a sequence of Feynman-Kac measures $(\mathbb{Q}_t)_{t=0:T}$ defined by the changes of measure

$$\mathbb{Q}_t(dx_{0:T}) = \frac{1}{L_t} G_0(x_0)\mathbb{M}_0(dx_0) \left\{ \prod_{s=1}^{t} G_s(x_{s-1}, x_s) \right\} \left\{ \prod_{s=1}^{T} M_s(x_{s-1}, dx_s) \right\}, \tag{2.2}$$

where L_t is the normalising constant required to make $\mathbb{Q}_t$ a probability measure. Other quantities such as $\mathbb{Q}_t(dx_{0:t})$ can be obtained as marginals of (2.2), allowing us to treat all of the inference problems described in Section 2.1.1 by approximating $\mathbb{Q}_t$ and then possibly marginalising.

The generic state space model defined in (2.1) may be described by a Feynman-Kac

model where:

$$\begin{aligned} \mathbb{M}_0 &:= \mu \\ M_t(x_{t-1}, dx_t) &:= K_t(dx_t \mid x_{t-1}) && \text{for } t = 1, 2, \ldots \\ G_0(x_0) &:= g_0(y_0 \mid x_0) \\ G_t(x_{t-1}, x_t) &:= g_t(y_t \mid x_t) && \text{for } t = 1, 2, \ldots. \end{aligned} \tag{2.3}$$

This is not the only Feynman-Kac model for (2.1); this corresponds to the *bootstrap* SMC algorithm, which is the simplest implementation. Abusing notation, g_t now denotes the density of the corresponding emission distribution; state space models in which these densities do not exist can still be expressed as Feynman-Kac models, the but not this bootstrap model. In practice the bootstrap SMC algorithm may be significantly outperformed by more involved algorithms such as *auxiliary particle filters* (Carpenter, Clifford, and Fearnhead 1999; Pitt and Shephard 1999) and those using *locally optimal proposals* (e.g. Doucet, Godsill, and Andrieu 2000) or *lookahead methods* (Lin, Chen, and Liu 2013). Feynman-Kac formalisms for some of these variants are presented for example in Chopin and Papaspiliopoulos (2020, Section 5.1.2).

It remains to demonstrate that the measures $\mathbb{Q}_t$ arising from (2.3) are sufficient for all the usual inference problems in the corresponding state space model (2.1). By construction, the complete smoothing distribution is precisely

$$\begin{aligned} \mathbb{Q}_t(dx_{0:t}) &= \frac{1}{L_t} G_0(x_0)\mathbb{M}_0(dx_0) \prod_{s=1}^{t} G_s(x_{s-1}, x_s) M_s(x_{s-1}, dx_s) \\ &= g_0(y_0 \mid x_0)\mu(dx_0) \prod_{s=1}^{t} g_s(y_s \mid x_s) K_s(dx_s \mid x_{s-1}) \\ &= p(dx_{0:t} \mid y_{0:t}). \end{aligned}$$

The filtering, prediction and fixed-lag smoothing distributions are all also marginals of some $\mathbb{Q}_t(dx_{0:T})$:

$$\begin{aligned} p(dx_t \mid y_{0:t}) &= \mathbb{Q}_t(dx_t) \\ p(dx_{t+h} \mid y_{0:t}) &= \mathbb{Q}_t(dx_{t+h}) \\ p(dx_{t-h:t} \mid y_{0:t}) &= \mathbb{Q}_t(dx_{t-h:t}), \end{aligned} \tag{2.4}$$

while the likelihood $p(y_{0:t}) = L_t$. This means that Monte Carlo approximation of $\mathbb{Q}_t(dx_{0:T})$ is sufficient for inference on any of these distributions, since marginalisation of Monte Carlo samples is trivial. The likelihood, on the other hand, is not obtained by marginalisation; nevertheless, we will see that likelihood estimates can also be obtained "for free". The next section describes how we may obtain Monte Carlo samples from $\mathbb{Q}_t(dx_{0:T})$.

2.1.4 Sequential Monte Carlo for Feynman-Kac models

In order to implement the SMC algorithm corresponding to a given Feynman-Kac model, we need to be able to sample from $\mathbb{M}_0$ and from $M_t(x, \cdot)$ for all x, t; and evaluate $G_t(x, y)$ pointwise for each x, y, t. Under these conditions we may implement Algorithm 2.1, which describes a generic SMC algorithm. The only free choices are the parameter N, which dictates the number of *particles* used, and the RESAMPLE procedure. However, remember that given a particular state space model there is also a choice of possible Feynman-Kac descriptions, and this choice can strongly affect performance.

Input: $T, N, \mathbb{M}_0, (M_t)_{t=1}^T, (G_t)_{t=0}^T$
for $i \in \{1, \ldots, N\}$ **do** Sample $X_0^{(i)} \sim \mathbb{M}_0(\cdot)$
for $i \in \{1, \ldots, N\}$ **do** $w_0^{(i)} \leftarrow \left\{\sum_{j=1}^N G_0(X_0^{(j)})\right\}^{-1} G_0(X_0^{(i)})$
for $t \in \{1, \ldots, T\}$ **do**
  Sample $a_{t-1}^{(1:N)} \sim \text{RESAMPLE}(\{1, \ldots, N\}, w_{t-1}^{(1:N)})$
  for $i \in \{1, \ldots, N\}$ **do** Sample $X_t^{(i)} \sim M_t(X_{t-1}^{(a_{t-1}^{(i)})}, \cdot)$
  for $i \in \{1, \ldots, N\}$ **do** $w_t^{(i)} \leftarrow \left\{\sum_{j=1}^N G_t(X_{t-1}^{(a_{t-1}^{(j)})}, X_t^{(j)})\right\}^{-1} G_t(X_{t-1}^{(a_{t-1}^{(i)})}, X_t^{(i)})$
end

Algorithm 2.1: Sequential Monte Carlo for a generic Feynman-Kac model

The choice of RESAMPLE procedure can also have a profound effect on performance and is discussed in detail in Section 2.4. Resampling is not necessary for the algorithm to be valid, but it is important to ensure good performance. Its purpose is to periodically "reset" the weights $w_t^{(1:N)}$, preventing *weight degeneracy*. This is the phenomenon that multiplying importance weights over time causes the variance of the unnormalised weights to increase exponentially, until after some iterations practically all of the weight is concentrated on one particle, with the rest having weights very close to zero. This means that the Monte Carlo sample (of size N) essentially consists of just one sample, so that the Monte Carlo approximations have very high variance.

The idea of resampling is to make multiple copies of particles with high weights and eliminate particles with low weights, then reset the weights to $w_t^{(1:N)} = (1, \ldots, 1)/N$. Done correctly, this procedure does not introduce any bias and, although it increases variance in the short term by adding extra randomness, it improves stability in the long term. It also prevents computational budget being "wasted" on simulating low-weight particles which do not contribute much to the approximation.

For each $i = 1, \ldots, N$, the RESAMPLE procedure selects a *parent*, indexed by $a_{t-1}^{(i)} \in \{1, \ldots, N\}$, which copies its state to the ith particle in the next iteration. These copies are then mutated independently and so the algorithm goes on. We define the *offspring counts* for each i, t as

$$\nu_{t-1}^{(i)} := |\{j : a_{t-1}^{(j)} = i\}|,$$

the number of copies in generation t of particle i from generation $t-1$ appearing by resampling. Notice that $\nu_{t-1}^{(1:N)}$ is expressed as a non-injective function of $a_{t-1}^{(1:N)}$, and as such carries less information. To ensure Algorithm 2.1 is valid, we assume that the resampling procedure is *unbiased*, that is, the expectation of $\nu_{t-1}^{(i)}$ is proportional to $w_{t-1}^{(i)}$ for each i. This requirement is formalised in Definition 2.2.

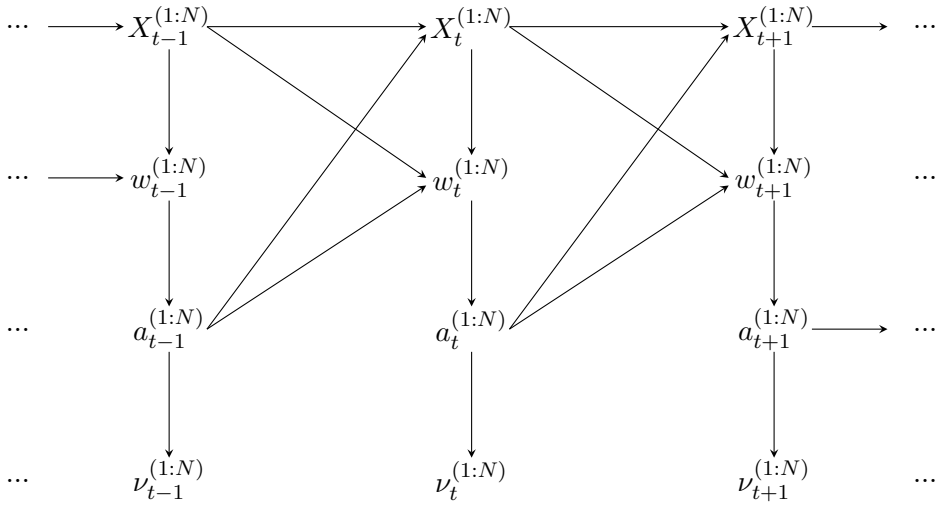


Figure 2.2: Part of the conditional dependence graph implied by Algorithm 2.1. The direction of time is from left to right.

Figure 2.2 shows a section of the conditional dependence graph implied by Algorithm 2.1. Because the algorithm proceeds sequentially, its computational cost is linear in the time horizon T, assuming that the cost of evaluating G_t is $O(1)$. Furthermore, the bootstrap algorithm, where the Feynman-Kac model is (2.3), processes the data $y_{0:T}$ one observation at a time via $G_t(x_{t-1}, x_t) = g_t(y_t \mid x_t)$, which means that it can be run on-line, incorporating each observation as it becomes available. This is in stark contrast to a standard MCMC approach, for example, which would have to process all of the data at once up to a fixed time horizon. Adding one more observation would require running the MCMC algorithm from scratch on the extended target, making the computational cost at least linear *per time point* and rendering on-line inference infeasible.

The output of Algorithm 2.1 is, for $i = 1, \ldots, N$ and $t = 0, \ldots, T$, the *states* $X_t^{(i)} \in \mathcal{X}$, the *weights* $w_t^{(i)} \in [0, 1]$ and, for $i = 1, \ldots, N$ and $t = 0, \ldots, T-1$, the *parental indices* $a_t^{(i)} \in \{1, \ldots, N\}$. Depending on the application, one may want to retain only a subset of this output in order to reduce memory usage.

The output can be used to construct discrete approximations of the various probability measures of interest, with which one may estimate integrals against test functions, i.e. expectations. The measure $\mathbb{Q}_t(dx_t)$, corresponding to a filtering distribution in the state space model example, is approximated by the empirical measure

$$\sum_{i=1}^{N} w_t^{(i)} \delta_{X_t^{(i)}}, \tag{2.5}$$

where δ_x denotes a unit mass at x. Expectations of appropriate test functions $\varphi : \mathcal{X} \mapsto \mathbb{R}$ are then approximated by their expectations with respect to the empirical measure,

$$\mathbb{E}_{\mathbb{Q}_t}[\varphi(dx_t)] \simeq \sum_{i=1}^{N} w_t^{(i)} \varphi(X_t^{(i)}).$$

The precise meaning of approximation (or $\simeq$) is clarified in Section 2.1.5. To approximate $\mathbb{Q}_t(dx_{0:t})$, we first define the *trajectories* $X_{t,0:t}^{(i)}$ (for each $i \in \{1, \ldots, N\}$) by setting $X_{t,t}^{(i)} := X_t^{(i)}$ and tracing back through the ancestors via the recursion $X_{t,s}(i) = X_{t,s+1}^{(a_t^{(i)})}$ for each $s \in \{0, \ldots, t\}$. We can then construct the approximation

$$\sum_{i=1}^{N} w_t^{(i)} \delta_{X_{t,0:t}^{(i)}}$$

of $\mathbb{Q}_t(dx_{0:t})$, corresponding to a smoothing distribution in a state space model, with which we can calculate expectations as above. Similar approximations can be constructed for the other measures in (2.4). We can also approximate the normalising constants, which correspond to marginal likelihoods in a state space model, using the *unnormalised* weights:

$$L_t \simeq \frac{1}{N} \sum_{i=1}^{N} G_0(X_0^{(i)}) \prod_{t=1}^{T} \frac{1}{N} \sum_{i=1}^{N} G_t(X_{t-1}^{a_{t-1}^{(i)}}, X_t^{(i)}). \tag{2.6}$$

The unnormalised weights could be output directly from Algorithm 2.1, or re-calculated from the states as shown here.

2.1.5 Theoretical justification

It can be shown that SMC approximations of expectations of test functions possess various desirable properties. For instance, it is quite easy to show that the approximations (2.5) satisfy a law of large numbers:

$$\sum_{i=1}^{N} w_t^{(i)} \varphi(X_t^{(i)}) \longrightarrow \mathbb{Q}_t(\varphi),$$

almost surely and in the L_2 sense, as $N \to \infty$, under some conditions (Crisan and Doucet 2002). Moreover, they satisfy a central limit theorem:

$$\sqrt{N} \left(\sum_{i=1}^{N} w_t^{(i)} \varphi(X_t^{(i)}) - \mathbb{Q}_t(\varphi) \right) \longrightarrow \text{Normal}(0, \sigma_t(\varphi))$$

in distribution, as $N \to \infty$ (Chopin 2004; Del Moral and Guionnet 1999). The $\sqrt{N}$ scaling agrees with the standard convergence rate for Monte Carlo approximations. Under additional conditions, the asymptotic variances $\sigma_t(\varphi)$ are stable over t (e.g. Chopin and Papaspiliopoulos 2020, Proposition 11.13), justifying the use of SMC filtering on-line. It

can also easily be shown that the likelihood estimates (2.6) are unbiased (see for example Chopin and Papaspiliopoulos 2020, Proposition 16.3).

There are many other results concerning convergence, stability and error bounds for SMC algorithms. A full exposition of these results and their conditions is beyond the scope of this work, but Del Moral (2004, 2013) provides an exhaustive treatment, and some of the key ideas and results are also developed in Chopin and Papaspiliopoulos (2020, Chapter 11). Suffice it to say that SMC algorithms enjoy enough theoretical properties to be useful in practice.

2.2 Coalescent theory

The current work draws on the literature around coalescent theory, primarily from population genetics. This section summarises the relevant parts of that literature. We will see in Section 2.3 how it applies to SMC.

2.2.1 Kingman's coalescent

The Kingman coalescent (Kingman 1982a,b,c) is a continuous-time Markov process on the space of partitions of $\mathbb{N}$. For our purposes we need only consider its restriction to $\{1, \dots, n\}$, termed the n-coalescent (Definition 2.1), since we only ever consider finite samples from a population.

Definition 2.1. Let $\mathcal{P}_n$ denote the set of partitions of $\{1, \dots, n\}$. The *n-coalescent* is the homogeneous continuous-time Markov process on $\mathcal{P}_n$ with infinitesimal generator Q having entries

$$q_{\xi,\eta} = \begin{cases} 1 & \xi \prec \eta \\ -|\xi|(|\xi| - 1)/2 & \xi = \eta \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

for every $\xi, \eta \in \mathcal{P}_n$, where $|\xi|$ denotes the number of blocks in ξ, and $\xi \prec \eta$ means that η is obtained from ξ by merging exactly one pair of blocks.

A particularly attractive feature of the n-coalescent is its tractability; its distribution and those of many statistics of interest are available in closed form (Section 2.2.2). It turns out also to be extremely useful as a limiting distribution in population genetics, including in its domain of attraction the genealogies of a wide range of population models (Section 2.2.3).

2.2.2 Properties of Kingman's coalescent

The simplicity of Q allows various properties of the n-coalescent to be studied analytically. Starting with n blocks, exactly $n-1$ coalescences are required to reach the absorbing state where all blocks have coalesced, known in population genetics as the *most recent common ancestor* (MRCA).

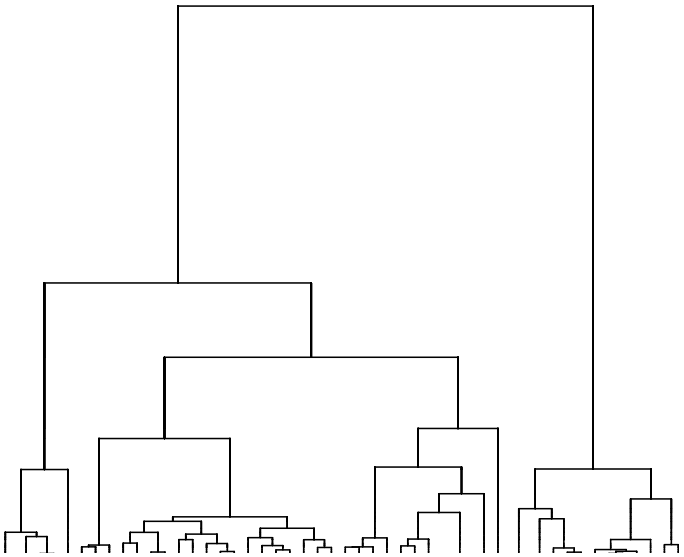


Figure 2.3: A realisation of the n-coalescent with $n = 50$.

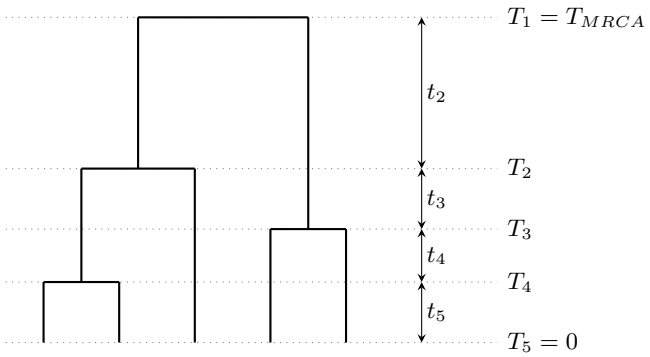


Figure 2.4: Definitions of t_i, T_i in the n-coalescent.

Denote by $t_2, \ldots, t_n$ the waiting times between coalescent events, where t_i is the amount of time for which the coalescent has exactly i distinct lineages (see Figure 2.4). A consequence of Definition 2.1 is that these waiting times are independent and have distributions

$$t_i \sim \text{Exp}\left(\binom{i}{2}\right).$$

The partial sum $T_k := \sum_{i=k+1}^{n} t_i$ gives the total time up to the $(n-k)^{\text{th}}$ coalescence event, that is, the first time at which there are only k lineages remaining out of the initial n (see Figure 2.4). The partial sums, being sums of independent Exponential random variables, have HypoExponential distributions.

Another important property of the n-coalescent is *exchangeability*. That is, its law is invariant under permutations of the branches. This can be seen from (2.7) since the merge rate is equal for every pair of partitions $\xi \prec \eta$.

Time to MRCA

Of particular interest is the tree height or time to the most recent common ancestor, $T_{MRCA} := T_1$. With some algebra we find, for instance,

$$\mathbb{E}[T_{MRCA}] = \sum_{i=2}^{n} \mathbb{E}[t_i] = \sum_{i=2}^{n} \frac{2}{i(i-1)} = 2\sum_{i=2}^{n} \left\{ \frac{1}{i-1} - \frac{1}{i} \right\} = 2\left(1 - \frac{1}{n}\right)$$

and

$$\mathrm{Var}[T_{MRCA}] = \sum_{i=2}^{n} \mathrm{Var}[t_i] = \sum_{i=2}^{n} \left(\frac{2}{i(i-1)} \right)^2.$$

The expected tree height converges to 2 as $n \to \infty$, and the variance converges to $4(\pi^2 - 9)/3 \simeq 1.16$. The somewhat surprising fact that the tree height does not diverge with n is a result of the very high rate of coalescence close to the bottom of the tree. This rate is large enough that the full Kingman coalescent (on $\mathbb{N}$) *comes down from infinity*, that is, despite starting with infinitely many blocks, after any positive amount of time these have coalesced into finitely many blocks.

Total branch length

Another quantity of interest is the total branch length, $L := \sum_{i=2}^{n} i t_i$. For instance

$$\mathbb{E}[L] = \sum_{i=2}^{n} i\mathbb{E}[t_i] = \sum_{i=2}^{n} \frac{2}{i-1} = \sum_{i=1}^{n-1} \frac{2}{i} \simeq 2\ln(n-1)$$

and

$$\mathrm{Var}[L] = \sum_{i=2}^{n} i^2 \,\mathrm{Var}[t_i] = \sum_{i=2}^{n} \frac{4}{(i-1)^2} = \sum_{i=1}^{n-1} \frac{4}{i^2}.$$

Note that although the mean total branch length diverges with n, the variance converges to a constant, $4\pi^2/6 \simeq 6.58$.

Probability that sample MRCA is population MRCA

One other interesting quantity is the probability that the MRCA of k random lineages coincides with the population MRCA (e.g. Durrett 2008, Theorem 1.7). Consider a random subsample of size k among n lineages distributed according to the n-coalescent. Denote by $S_{k,n}$ the event that these k lineages have the same MRCA as all n lineages. The probability of this event is calculated in Saunders, Tavaré, and Watterson (1984, Example 1) and again in Spouge (2014, Equation (3)), in both cases arising as a special case of more general results. A direct proof is given below.

Consider the two subtrees produced by cutting the full population tree just below the population MRCA. The k sampled lineages coalesce before the full-sample MRCA if and only if all k sampled leaves lie in just one of these two subtrees. Let X be the number of leaves in the left subtree, so $X \in \{1, \ldots, n-1\}$, and a consequence of the exchangeability

of the n-coalescent is that X is uniformly distributed on that set. Conditional on X we have

$$\mathbb{P}[S^c_{k,n} \mid X = x] = \left[\binom{x}{k} + \binom{n-x}{k}\right]\binom{n}{k}^{-1}.$$

Integrating against the distribution of X gives

$$\begin{aligned}\mathbb{P}[S_{k,n}] &= 1 - \frac{1}{n-1}\binom{n}{k}^{-1}\sum_{x=1}^{n-1}\left[\binom{x}{k} + \binom{n-x}{k}\right] \\ &= 1 - \frac{1}{n-1}\binom{n}{k}^{-1}\left[\binom{n}{k+1} + \binom{n}{k+1}\right] \\ &= \frac{k-1}{k+1}\frac{n+1}{n-1}\end{aligned}$$

using binomial identities and some algebra. In particular, when $k = 2$ we have

$$\mathbb{P}[S_{2,n}] = \frac{n+1}{3(n-1)}$$

as the probability that a randomly chosen pair of lineages does not coalesce until the MRCA of all n lineages.

2.2.3 Models in population genetics

The Kingman coalescent is the limiting coalescent process (in the large population limit) for a surprisingly wide range of population models. Some important examples of models in this domain of attraction are introduced in this section. Common to all of these models are the following assumptions:

- The population has constant size N
- Reproduction happens in discrete generations
- The mechanism for assigning offspring to parents is identical at each generation, and independent between generations
- The offspring distribution is exchangeable.

As in Section 2.1.4, we define offspring counts in terms of parental indices as $\nu_j := |\{i : a_i = j\}|$. Since the assignment of offsrping to parents is i.i.d. across generations, there is no dependence on t. Also, under the assumption of exchangeability, it is sufficient to consider only the offspring counts, rather than the parental indices (which generally carry more information). These models are all *neutral*, that is exhibiting no natural selection, because the offspring counts at each generation are independent, so there can be no preferential propagation of certain "fitter" lineages.

Cannings model

The neutral Cannings model (Cannings 1974, 1975) is a general class which encompasses some other important models as special cases.

The Cannings model does not specify a particular distribution for the offspring counts; it just requires that the distribution is exchangeable, i.i.d. between generations, and preserves the population size. In particular, the probability of observing offspring counts $(v_1, \ldots, v_N)$ must be invariant under permutations of this vector.

Rescaled genealogies of the neutral Cannings model converge to the Kingman coalescent as $N \to \infty$, under some conditions on the moments of the offspring distribution. For example, one may apply the sufficient conditions of Kingman (1982b): if $\mathrm{Var}[\nu_1] \to \sigma^2 \in (0, \infty)$ and $\mathbb{E}[\nu_1^k]$ is bounded for all $k \in \mathbb{N}$ then, under the time scaling $N\sigma^{-2}$, the genealogies of the neutral Cannings model converge to the Kingman coalescent.

Wright-Fisher model

The neutral Wright-Fisher model (Fisher 1923, 1930; Wright 1931) is one of the most studied models in population genetics. At each generation the existing population dies and is replaced by N offspring. The offspring descend from parents $(a_1, \ldots, a_N)$ which are selected according to

$$a_i \overset{iid}{\sim} \text{Categorical}(\{1, \ldots, N\}, (1/N, \ldots, 1/N)).$$

The joint distribution of the offspring counts is therefore

$$(v_1, \ldots, v_N) \sim \text{Multinomial}(N, (1/N, \ldots, 1/N)).$$

Since the Multinomial distribution is exchangeable, the Wright-Fisher model is a special case of the Cannings model.

Kingman (1982b) showed that the Wright-Fisher model satisfies his sufficient conditions, and thus the resulting genealogies, appropriately rescaled, converge to the Kingman coalescent as $N \to \infty$. The correct time scale in this instance is N, since

$$\mathrm{Var}[\nu_1] = N\frac{1}{N}\left(1 - \frac{1}{N}\right) = \frac{N-1}{N} \to 1 =: \sigma^2,$$

so $N\sigma^{-2} = N$.

Moran model

The neutral Moran model (Moran 1958), while perhaps less biologically relevant, is mathematically appealing because its simple dynamics make it particularly tractable.

At each generation, an ordered pair of individuals is selected uniformly at random. The first individual in this pair dies (i.e. leaves no offspring in the next generation), while the

other reproduces (leaving two offspring). All of the other individuals leave exactly one offspring. This is another special case of the neutral Cannings model, where the offspring distribution is now uniform over all permutations of $(0, 2, 1, 1, \ldots, 1)$.

Under a suitable time-scaling, its genealogies converge to the Kingman coalescent, although the sufficient conditions of Kingman (1982b) do not apply:

$$\begin{aligned}\mathrm{Var}[\nu_1] = \mathbb{E}[\nu_1^2] - \mathbb{E}[\nu_1]^2 &= (0 + 1 \cdot \frac{N-2}{N} + 4 \cdot \frac{1}{N}) - (0 + 1 \cdot \frac{N-2}{N} + 2 \cdot \frac{1}{N})^2 \\ &= \frac{N+2}{N} - 1^2 = \frac{2}{N} \to 0,\end{aligned}$$

violating the condition that $\sigma^2 > 0$. That condition turns out not to be necessary, and $\mathrm{Var}[\nu_1]$ gives us the correct time scale on which to recover the Kingman coalescent: $N(\mathrm{Var}[\nu_1])^{-1} = N^2/2$. It is not surprising that the time scale is an order bigger than in the Wright-Fisher model, because the Moran model has a reproduction rate $O(N)$ times lower than in the Wright-Fisher model: at each generation 2 individuals are involved in reproduction, as opposed to N in the Wright-Fisher model.

2.2.4 Other convergence results

The original work of Kingman (1982b) provides sufficient conditions for the finite-dimensional distributions of genealogies of Cannings models to converge to those of the Kingman coalescent. Möhle (1998) provides another set of sufficient conditions which apply to the wider class of models in which the population size may vary deterministically, the offspring distributions are independent but not identical across generations, and exchangeability is replaced by the weaker *random assignment* condition. For that class of models under the same conditions, Möhle (1999) proves that the genealogies converge weakly as well as in the sense of finite-dimensional distributions. Möhle (2000) gives a simpler condition which is necessary and sufficient for convergence of Cannings genealogies to the Kingman coalescent.

Meanwhile many similar results were established for models in which the limiting process is not the Kingman coalescent. Relaxing the conditions to allow multiple mergers in the limit admits Λ-coalescents as limiting processes, with the Kingman coalescent as a special case (Möhle and Sagitov 1998; Pitman 1999; Sagitov 1999). If simultaneous mergers are also allowed, the limiting process belongs to the even more general class of Ξ-coalescents (Möhle and Sagitov 2001), and this class encompasses all possible limiting genealogies of Cannings models. On the other hand, Del Moral, Miclo, et al. (2009) do not consider asymptotics in the population size, but instead prove some properties of neutral models with fixed population N, particularly concerning the MRCA.

The focus of the current work is on systems that admit Kingman genealogies in the limit, among a wider class of models where, like Möhle (1998, 1999), exchangeability is relaxed to random assignment, and we also do not require independence between generations, so our models are not neutral. Our main condition for convergence to the Kingman

coalescent, which is introduced in Chapter 3, can be considered a non-exchangeable non-neutral analogue of the condition presented in Möhle (2000).

2.2.5 Particle populations

Much of the population genetics framework transfers readily to the case of SMC. The population is now a population of particles, with each iteration of the SMC algorithm corresponding to a generation, and resampling playing the part of reproduction. In fact, SMC populations are in some ways more suited to these population models than actual biological populations. For example, the assumptions that the population has constant size N and that reproduction occurs only at discrete generations are satisfied by construction.

However, we cannot assume independence between generations: as seen in Figure 2.2, the offspring counts at subsequent iterations of an SMC algorithm are not independent without some conditioning. This means that SMC populations are not neutral. In fact, after marginalising out the information about the positions of the particles, the genealogical process is not even Markovian.

2.3 Sequential Monte Carlo genealogies

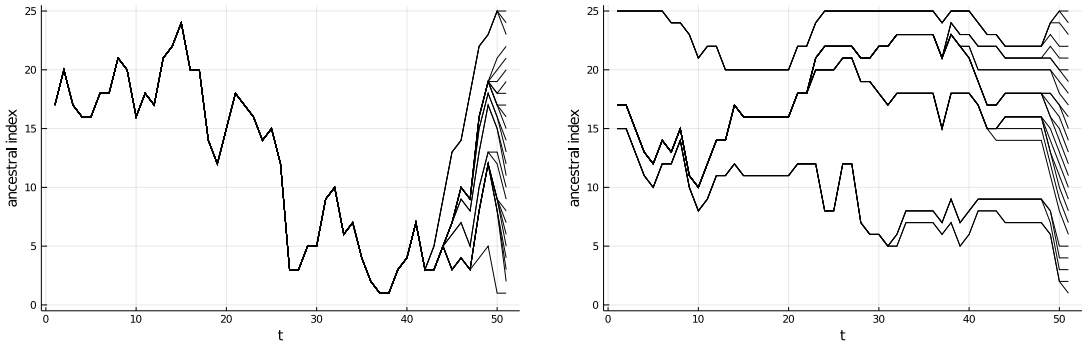
We have seen that genetic terminology applies quite naturally to SMC. The resampling step induces parent-offspring relationships, each duplicate of particle i after resampling being considered one of its offspring. Then follows the notion of offspring counts (also known as family sizes), that is, the number of offspring assigned to each parent. Viewed backwards in time, the parent-offspring relationships also imply a genealogy, obtained by tracing the lineages from each terminal particle through its ancestor in each generation. We will see in this section that these genealogies, induced by resampling, are not a mere curiosity but in fact have important implications for the performance of SMC algorithms.

2.3.1 Ancestral degeneracy

Suppose we were using SMC to sample from the smoothing distribution of some state space model. As described in Section 2.1.4, we run our chosen SMC algorithm forwards, then output the N sampled trajectories $X_{t,0:t}^{(i)}$ (for each $i \in \{1, \ldots, N\}$). Each trajectory was obtained by tracing back through the parent at each generation, starting from one of the terminal particles. This means that if two terminal particles i and j share a common ancestor at some generation s, then $X_{t,0:s}^{(i)}$ will be exactly equal to $X_{t,0:s}^{(j)}$, because their ancestries coincide from time s to 0.

At every resampling step, some parents may be assigned more than one offspring each, so the further back in time you look, the more of the ancestries of the terminal particles will have coalesced (see Figure 2.5a). The effect of this is that, instead of obtaining N separate sampled trajectories, we actually obtain N sampled trajectories that coalesce backwards in time, which means that the further back in time we look, the fewer distinct samples

we have from the corresponding component of the target distribution. Particularly if we are interested in smoothing over a long time horizon, the variance of the SMC estimator is going to blow up.



(a) Multinomial resampling (b) Adaptive minimum-variance resampling

Figure 2.5: Illustration of ancestral degeneracy and the mitigating effect of low-variance and adpative resampling. Each line is the lineage of one of the terminal particles, indicating the index of its ancestor in each generation: (a) with multinomial resampling; (b) the same system with adaptive systematic resampling.

On the other hand, ancestral degeneracy actually improves the memory efficiency of SMC. We do not need to store all of the particles generated at each time (at memory cost $O(NT)$), only those that are included in the resulting genealogy. Jacob, Murray, and Rubenthaler (2015) provide an algorithm for efficient storage of the genealogy, reducing the asymptotic memory cost to $O(N \log N + T)$. However, it is certainly still worth trying to reduce ancestral degeneracy because to achieve a given level of error with a highly degenerate system will require such a large N that any such memory gains are cancelled out.

Mitigating ancestral degeneracy

There are a few possible approaches to mitigating ancestral degeneracy. Firstly, we could try to limit the number of offspring assigned to any one parent during each resampling step. We can only go so far, because we need the resampling procedure to remain unbiased (as discussed in Section 2.1.4), but we can try to reduce the variance inherent in the resampling procedure. This idea, known as *low-variance resampling* is explored in detail in Section 2.4.

Another idea is to resample less often. Recall that the reason for resampling is to prevent weight degeneracy (that is, one of the weights tending to one while the others tend to zero). Now we see that, while solving one type of degeneracy, resampling creates another. The effect of ancestral degeneracy is essentially the same as that of weight degeneracy: both drastically increase the variance of the resulting SMC estimators. We can therefore consider a trade-off between the two, which is the idea behind *adaptive resampling* (Liu and Chen 1995, Section 4). The trick is to apply the resampling step only at iterations

in which a certain criterion is met. The most commonly-used criterion, suggested by Liu and Chen (1995, Equation (14)), is based on the *effective sample size*

$$ESS(t) := \left\{ \sum_{i=1}^{N} (w_t^{(i)})^2 \right\}^{-1},$$

which decreases as the weights degenerate. The resampling step is then applied only at iterations t such that $ESS(t)$ is less than some pre-specified threshold, typically $N/2$.

If adaptive resampling is used, some trivial changes are required to the calculation of the weights in Algorithm 2.1, to allow for the importance weights to accumulate sequentially until the particles are resampled. See e.g. Chopin and Papaspiliopoulos (2020, Section 10.2) for details.

As well as mitigating ancestral degeneracy, adaptive resampling has the virtue of saving some computation (although the overall asymptotic complexity of the SMC algorithm does not change). How effective adaptive resampling will be depends on the particular application and choice of SMC algorithm. If the proposals (i.e. transition kernels) are not very close to their targets then the weights will degenerate rapidly and the effective sample size criterion (or similar) will not reduce the frequency of resampling very much.

Low-variance resampling is also less effective under poor proposals: the resulting high-variance weights lead to high-variance offspring counts, even under minimum-variance resmapling schemes, because the resampling is required to be unbiased.

Adaptive resampling and low-variance resampling can be combined, and this is widely considered to be the best practice when implementing SMC. Figure 2.5 compares ancestral degeneration under multinomial resampling (a relatively high variance scheme) to the same under adaptive resampling with a minimum-variance resampling scheme. It is clear that the degeneration is much more severe in the former case.

There is one technique that completely solves the problem of ancestral degeneracy, namely *backward simulation* (Godsill, Doucet, and West 2004). This involves running an SMC algorithm as usual (the *forward pass*), and then sampling new ancestors for each particle during an additional *backward pass*. The backward-simulated parents in each generation are chosen among all N particles, making use of particles that were not included in the forward-sampled trajectories. The effect on genealogies is striking: the lineages are now sampled independently, so the coalescences caused by resampling do not feature at all in the output genealogies.

Since this work concerns genealogies induced by resampling, we will not say much more about backward simulation. There are many situations in which it is impossible to implement and therefore the study of SMC genealogies is still of interest. Firstly, backward simulation inherently requires a forward and backward pass through all of the data, so it cannot be implemented on-line. Secondly, calculating the backward-simulation probabilities requires the Markov kernels M_t of the corresponding Feynman-Kac model to admit densities that can be evaluated pointwise. This is much stronger than the ability

to simulate from M_t, the requirement for applying standard SMC algorithms.

2.3.2 Asymptotic genealogies

If we had access to information about the behaviour of SMC genealogies a priori (i.e. without having run the algorithm), we would be in a position to answer many questions of interest. These include practical questions about tuning, for example:

- How many particles should I use in order to maintain (with high enough probability) a given level of error over a time horizon T?
- With N particles, what is the largest lag over which fixed-lag smoothing produces reasonable estimates?
- How many particles should I use within particle Gibbs to ensure that (with high enough probability) at least two distinct trajectories survive each iteration?

This last question touches on a critical aspect of the performance of particle Gibbs algorithms, which is discussed in Section 2.5. We could also consider theoretical questions, such as:

- For a given class of models and algorithms, what is the effect of ancestral degeneracy on how the estimators behave over time?
- Which resampling schemes lead to the smallest amount of ancestral degeneracy?
- What is the effect on genealogies of adaptive resampling?

Many of these questions have already been partially addressed, without any explicit analysis of genealogies, by way of variance calculations and simulation experiments. But since these are all genealogical questions by nature, it seems sensible to work directly with the genealogies, if possible. The problem is that the genealogy of particles is a complex object, it is random, and it can depend strongly on the particular choice of Feynman-Kac model and SMC implementation.

It turns out that these problems can be somewhat overcome by considering the genealogies in an asymptotic regime where the number of particles N tends to infinity. In this regime, many different particle systems exhibit genealogies of a common form, namely Kingman's n-coalescent under suitable time-scalings. The genealogical differences between various algorithms is then encoded by their respective time-scale functions. This is still a random object but is less complicated than the genealogy itself; namely a càdlàg function as opposed to a labelled weighted tree.

In the context of SMC, these asymptotic genealogies were first analysed by Koskela et al. (2018). The simulations therein suggest that such asymptotic results also transfer to finite systems, making them practically useful. One of the contributions of the current work is to demonstrate that Kingman-type genealogies arise from a wide variety of SMC

algorithms, including those most commonly used in practice. In principle this means, for instance, that genealogies of different SMC algorithms can be compared by examining the corresponding time-scale functions.

2.4 Resampling

As we have seen, resampling is necessary within SMC to reset the weights in order to prevent weight degeneracy. Resampling is itself a Monte Carlo procedure: the discrete offspring counts can be viewed as stochastic estimates of the continuous weights. In order to obtain a valid SMC algorithm, these Monte Carlo samples must be unbiased; this and other desirable properties are formalised in Definition 2.2. There is a huge range of resampling procedures satisfying these properties, some of which perform better than others. Some of the most popular resampling schemes are introduced in Section 2.4.2 and their properties are explored in Section 2.4.3.

2.4.1 Definition

Definition 2.2. For our purposes, a valid resampling scheme is a stochastic function mapping weights $w_t^{(1:N)} \in \mathcal{S}_{N-1}$ to offspring counts $\nu_t^{(1:N)} \in \{0, \ldots, N\}^N$ that satisfies the following conditions:

1. the population size is conserved: $\sum_{i=1}^{N} \nu_t^{(i)} = N$
2. the weights are equal after resampling: $w_{t+}^{(i)} = 1/N$ for all i
3. the resampling is unbiased: $\mathbb{E}[\nu_t^{(i)} \mid w_t^{(i)}] = N w_t^{(i)}$ for all i.

It is possible to design resampling schemes that violate these properties. Resampling different numbers of particles in different iterations (violating condition 1) is of course possible (see for example Crisan, Del Moral, and Lyons 1999), but we typically have a fixed limit on computational resources, so in most cases it makes sense to simulate the maximum feasible number of particles N at every iteration. There may be circumstances under which it is beneficial to allow the number of particles to vary adaptively (Chau et al. 2012; Fox 2003; Lee and Whiteley 2018) or at random, but this is not commonly done in practice.

Condition 2 is sensible because the whole point of resampling is to reset the weights. However, there are several examples in the literature where the weights are only partially reset, so that the variance of weights after resampling is not zero, but is lower than before resampling. For example, a scheme of Liu and Chen (1998) uses the square roots of the weights for resampling, then corrects by setting unequal weights after resampling (violating conditions 2 and 3). Liu, Chen, and Logvinenko (2001, Section 3.1) generalise this further, and suggest setting the resampling weights adaptively as a function of the true weights $w_t^{(1:N)}$. Fearnhead and Clifford (2003) present an optimal resampling scheme

in the case that the state space is discrete, which similarly uses weights other than $w_t^{(1:N)}$ for resampling and corrects by giving the particles unequal weights after resampling. The *chopthin* resampling algorithm of Gandy and Lau (2016) also violates condition 2.

Deterministic resampling schemes (which cannot generally be unbiased, violating condition 3) have been used by some authors. One such scheme was proposed by Kitagawa (1996) but is now generally implemented only in its randomised form (see *systematic resampling* below). More recent examples include schemes based on optimal transport (Corenflos et al. 2021; Myers et al. 2021; Reich 2013) and the *importance support points* resampling of Huang, Joseph, and Mak (2020).

The mutation and weighting steps of SMC are embarrassingly parallel, but resampling is not easy to parallelise, presenting a bottleneck to running SMC on parallel or distributed computer architectures. Whiteley, Lee, and Heine (2016) show that it is possible to construct resampling schemes that perform well whilst only requiring interaction of a few particles at a time, suggesting that parallel resampling is possible, and further details concerning implementation are provided in Lee and Whiteley (2016). The *Metropolis resampler* of Murray, Lee, and Jacob (2016) resamples in parallel via a Metropolis MCMC algorithm, but this introduces bias and thus violates condition 3. Murray, Lee, and Jacob (2016) also propose *rejection resampling*, which is unbiased. This constitutes an alternative method for *multinomial resampling* (see below) which offers speed-ups when computing in parallel but requires that an upper bound on the weights is known. A variant that only requires an "approximate" upper bound on the weights is also presented, but this does not use the true weights for resampling, and so violates conditions 2 and 3.

The majority of resampling schemes in the literature fit within Definition 2.2, and it is not usually advantageous to violate the properties 1–3. Definition 2.2 still allows a great deal of flexibility in the choice of resampling scheme, and many such schemes have been proposed, some performing better than others. Some important resampling schemes are reviewed in Section 2.4.2, and their performance is discussed in detail in Section 2.4.3 and summarised in Table 2.3.

2.4.2 Examples

Multinomial resampling

Multinomial resampling (Efron and Tibshirani 1994; Gordon, Salmond, and Smith 1993; Rubin 1987) is one of the simplest resampling schemes. The parental indices are chosen independently from $\{1, \ldots, N\}$, each with probability given by the weight of the corresponding particle $w_t^{(i)}$. That is,

$$a_t^{(1:N)} \sim^{iid} \text{Categorical}(\{1, \ldots, N\}, w_t^{(1:N)}).$$

This implies that the joint distribution of the offspring counts is

$$\nu_t^{(1:N)} \overset{d}{=} \text{Multinomial}(N, w_t^{(1:N)}).$$

It follows from properties of the Multinomial distribution that this resampling scheme is unbiased.

A simple way to sample the parental indices is to use inversion sampling: partition the unit interval into N subintervals each of which will correspond to a certain index i and has length equal to the weight $w_t^{(i)}$; then draw N samples $U_i \sim \text{Uniform}[0,1]$ and classify them according to which of these subintervals they fall in. Explicitly, the parental index assigned to child i is the index a_i satisfying

$$\sum_{j=1}^{a_i-1} w_t^{(j)} \leq U_i \leq \sum_{j=1}^{a_i} w_t^{(j)}. \tag{2.8}$$

This is illustrated in Figure 2.6.

Fast implementations of multinomial resampling rely on $U_1, \ldots, U_N$ being pre-sorted, which speeds up the search step (2.8). Sorting N numbers requires $O(N \log N)$ computation, but this is not necessary since we can sample directly the order statistics of a Uniform$[0,1]$ distribution, at $O(N)$ cost. This can be done either by sampling $X_i \sim \text{Exp}(1)$ independently for $i = 1, \ldots, N+1$ and outputting the normalised sums

$$U_k := \frac{\sum_{i=1}^{k} X_i}{\sum_{i=1}^{N+1} X_i}$$

for $k = 1, \ldots, N$, or by sampling $X_i \sim \text{Uniform}[0,1]$ independently for $i = 1, \ldots, N$ and computing recursively

$$U_N := X_N^{1/N}, \qquad U_k := X_k^{1/k} U_{k+1};$$

see Devroye (1986, Chapter 5, Section 3.1). This allows multinomial resampling to be implemented at $O(N)$ cost. A side-effect is that the sampled ancestral indices will be ordered and therefore cannot be Categorically distributed, although the offspring counts still have the correct Multinomial distribution. For the purposes of resampling this isn't usually a problem, but the Categorical distribution can anyway be restored at $O(N)$ cost by applying a random permutation to the offspring indices.

Residual resampling

Residual resampling is described in Liu and Chen (1998) and also in Whitley (1994) where it is called remainder stochastic sampling.

Each particle $X_t^{(i)}$ is deterministically assigned $\lfloor N w_t^{(i)} \rfloor$ offspring, and the remaining

$$R := \sum_{i=1}^{N} (N w_t^{(i)} - \lfloor N w_t^{(i)} \rfloor) = N - \sum_{i=1}^{N} \lfloor N w_t^{(i)} \rfloor$$

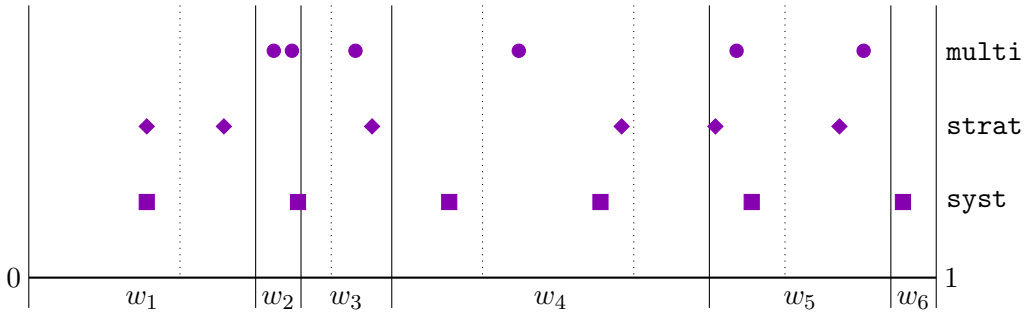


Figure 2.6: Inversion sampling interpretation of multinomial, stratified and systematic resampling. In this example, $N = 6$, $w^{(1:6)} = (0.25, 0.05, 0.1, 0.35, 0.2, 0.05)$ and the uniform random variables input to the resampling schemes are $u_{1:6} = (0.78, 0.29, 0.27, 0.92, 0.54, 0.36)$. The solid vertical lines show the partition of $[0, 1]$ into subintervals with lengths $w^{(1:6)}$. The dotted vertical lines show the partition of $[0, 1]$ into subintervals of length $1/N$, used for stratified and systematic resampling.
Top row (circles): in multinomial resampling, $u_{1:6}$ are fed directly into the inversion sampler. Which subinterval u_i falls into determines the parent of offspring i. The resulting offspring counts in this example are $\nu^{(1:6)} = (0, 2, 1, 1, 2, 0)$.
Middle row (diamonds): in stratified resampling, $u_{1:6}$ are transformed so that one point lies in each subinterval of length $1/N$. The resulting offspring counts are $\nu^{(1:6)} = (2, 0, 1, 1, 2, 0)$.
Bottom row (squares): in systematic resampling, only u_1 is used, being transformed to equally spaced points. The resulting offspring counts are $\nu^{(1:6)} = (1, 1, 0, 2, 1, 1)$.

offspring are assigned stochastically according to the residual weights

$$r^{(i)} := (N w_t^{(i)} - \lfloor N w_t^{(i)} \rfloor)/R.$$

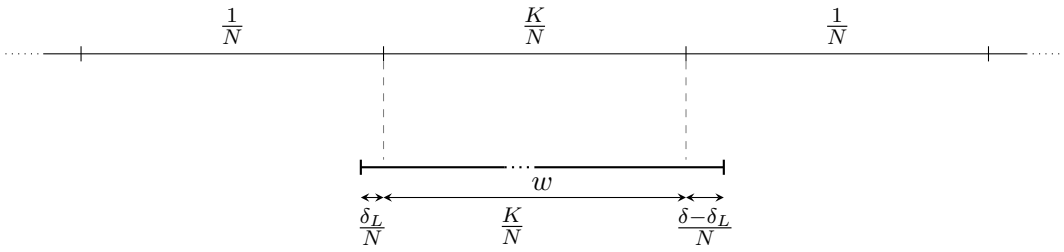
Notice that each $r^{(i)}$ lies in the interval $[0, 1/R)$, and $R \in \{0, \ldots, N-1\}$ with $R = 0$ only if all weights are multiples of $1/N$ in which case all residual weights are zero.

The stochastic part can be implemented using any of the other basic resampling schemes (e.g. multinomial, stratified, systematic). Most presentations focus on the case where multinomial resampling is used for the residuals, which is by no means the most sensible choice. We will explore several different options in what follows.

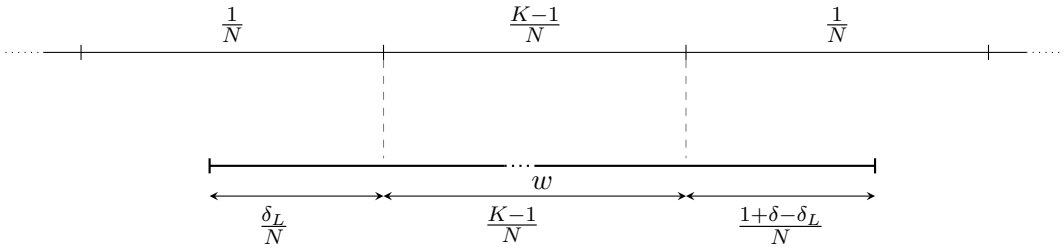
Stratified resampling

Stratified resampling was introduced by Kitagawa (1996). As in multinomial resampling, inversion sampling is used sample the parental indices. However, the samples used for inversion sampling are no longer i.i.d. Uniform$[0, 1]$ samples. Instead, one number is sampled independently from each subinterval of length $1/N$; that is,

$$U_i \sim \text{Uniform}\left[\frac{i-1}{N}, \frac{i}{N}\right].$$



(a) The overhang is less than $1/N$ and $\delta_L \in [0, \delta]$. The parent under consideration is automatically assigned K offspring, plus up to two more.



(b) The overhang is greater than $1/N$ (this case can only occur when $K \geq 1$) and $\delta_L \in (\delta, 1)$. The parent under consideration is automatically assigned $K - 1$ offspring, plus up to two more.

Figure 2.7: Cases for stratified resampling with a fixed weight $w = (K + \delta)/N$.

Alternatively, one may think of standard Uniform samples $u_1, \ldots, u_N \sim^{iid} \text{Uniform}[0, 1]$ with the transformation

$$U_i = \frac{u_i + i - 1}{N}.$$

The parents are then assigned as in (2.8), illustrated in Figure 2.6. The offspring distribution is no longer Multinomial, since parental indices are not identically distributed. Stratified resampling ensures that the samples are "well spread out", which reduces the probability of randomly losing high-weight particles or duplicating low-weight particles.

It will be useful later on to have a better idea about the marginal distributions of $\nu_t^{(i)}$ that are induced by stratified resampling. There are complex dependencies between the offspring counts, but we can still find some constraints on the distribution of each count conditional on the corresponding weight. Write the i^{th} weight in the form $w_t^{(i)} = (K + \delta)/N$, where $\delta \in [0, 1)$ and $K \in \{0, \ldots, N\}$. Considering the illustration Figure 2.6, the distribution of $\nu_t^{(i)}$ depends not only on $w_t^{(i)}$ but also on where the i^{th} weight interval falls with respect to the length-$(1/N)$ intervals. Denote the *left overhang* by δ_L. There are two cases to consider, which are illustrated in Figure 2.7. In Case (a) the total overhang is less than $1/N$ and $\delta_L \in [0, \delta]$. In Case (b) the total overhang is greater than $1/N$ and $\delta_L \in (\delta, 1)$. Arrangements such that one or both ends have no overhang are special cases of Case (a) where $\delta_L \in \{0, \delta\}$. Note that Case (b) cannot occur if $K = 0$.

In any case $\nu_t^{(i)} \in \{K - 1, K, K + 1, K + 2\}$ almost surely. To define a probability distribution over these four values, we introduce the notation

$$p_j := \mathbb{P}[\nu_t^{(i)} = \lfloor N w_t^{(i)} \rfloor + j \mid w_t^{(i)}],$$

for $j = -1, 0, 1, 2$. Since the sample within each interval of length $1/N$ is uniform over that interval, we find the probabilities given in Table 2.1, in terms of δ and δ_L. The probabilities do not depend on K, but of course the corresponding values of $\nu_t^{(i)}$ do.

	Case (a)	Case (b)	L.B.	U.B.
p_{-1}	0	$\delta_L(1+\delta-\delta_L)-\delta$	0	$1/4$
p_0	$1-\delta+\delta_L(\delta-\delta_L)$	$1+\delta-2\delta_L(1+\delta-\delta_L)$	$(1-\delta)/2$	$1-3\delta/4$
p_1	$\delta-2\delta_L(\delta-\delta_L)$	$\delta_L(1+\delta-\delta_L)$	$\delta/2$	$(1+\delta)/2$
p_2	$\delta_L(\delta-\delta_L)$	0	0	$1/4$

Table 2.1: Marginal probability distribution of $\nu_t^{(i)}$ conditional on $w_t^{(i)} = (K+\delta)/N$, in terms of δ and the left overhang δ_L, along with upper and lower bounds on these in terms of δ only, which hold in both cases.

Systematic resampling

Systematic resampling is described in Carpenter, Clifford, and Fearnhead (1999) and also in Whitley (1994) where it is called stochastic universal sampling.

Like stratified resampling, it uses the inversion sampler of multinomial resampling but starts with a more regular set of points in $[0, 1]$. In this scheme, only one standard Uniform sample is drawn, $u \sim \text{Uniform}[0, 1]$, from which the N samples are generated by via the transformation

$$U_i = \frac{u + i - 1}{N}$$

for $i = 1, \ldots, N$. The parental indices are again selected according to (2.8), as illustrated in Figure 2.6.

Kitagawa (1996) suggests a deterministic scheme in which the random u is replaced by a fixed $\alpha \in [0, 1]$; but, being deterministic, this scheme does not satisfy the unbiasedness condition (Property 1 in Definition 2.2). Whitley (1994) employs a different description of systematic resampling, where the interval $[0, 1]$ is joined up into a circle, and the systematic samples are evenly spaced pointers on an outer ring, which is spun around like a roulette wheel (Figure 2.8). This comprises adding a random phase to each U_i, modulo one, and is an exactly equivalent description of systematic resampling.

Like stratified resampling, systematic resampling ensures the random numbers are "well spread out"; the resulting samples are even more constrained than with stratified resampling. Systematic resampling also has the advantage of being extremely easy to implement and computationally efficient, requiring only one sample from a pseudo-random number generator (PRNG) followed by $O(N)$ elementary operations.

However, the systematic scheme is known to exhibit pathological behaviour in some cases because its performance depends on the ordering of the weights. A simple example of this phenomenon is presented in Douc, Cappé, and Moulines (2005). Such behaviour

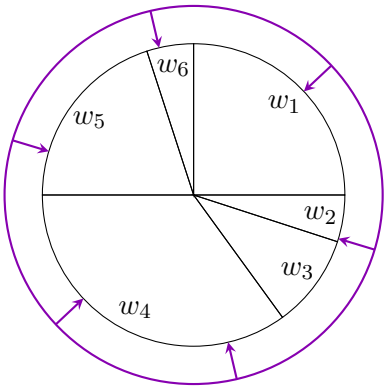


Figure 2.8: Whitley's "roulette wheel" description of systematic resampling. The inner circle stays fixed while the outer circle with its equally-spaced pointers is "spun" by a random amount. The weights and phase pictured are the same as in Figure 2.6, with $N = 6$. For systematic resampling, the two descriptions are equivalent.

can be avoided by randomly permuting the weights before resampling, and this is the recommended practice.

Star resampling

For the sake of comparison, we also construct a resampling scheme which is in some sense the worst possible. Sample

$$a \sim \text{Categorical}(\{1, \dots, N\}, w_t^{(1:N)})$$

and set $a_t^{(i)} = a$ for all i. The resulting offspring counts are all equal to zero except for $\nu_t^{(a)}$, which is equal to N. This resampling scheme is indeed unbiased, since each offspring count has marginal distribution

$$\nu_t^{(i)} \mid w_t^{(1:N)} = \begin{cases} 0 & \text{w.p. } 1 - w_t^{(i)} \\ N & \text{w.p. } w_t^{(i)}. \end{cases}$$

These offspring counts have the highest possible marginal variance subject to $\mathbb{E}[\nu_t^{(i)} \mid w_t^{(i)}] = N w_t^{(i)}$ and $\nu_t^{(i)} \in \{0, \dots, N\}$.

I call this scheme *star resampling* because the parent-offspring relationships at each iteration form a star graph.

Minimal variance branching

The minimal variance branching (MVB) algorithm of Crisan and Lyons (1999) provides a framework for resampling that enforces the minimal variance. It requires that each

offspring count $\nu_t^{(i)}$, conditionally on $w_t^{(i)}$, has marginal distribution

$$\nu_t^{(i)} \mid w_t^{(i)} \overset{d}{=} \lfloor N w_t^{(i)} \rfloor + \text{Bernoulli}(N w_t^{(i)} - \lfloor N w_t^{(i)} \rfloor). \tag{2.9}$$

We will see later on that this is exactly the framework of *stochastic rounding*.

The set-up of Crisan and Lyons (1999) does not require the number of particles to remain constant from one generation to the next (Property 1 in Definition 2.2), so the MVB algorithm could be implemented for instance by sampling each $\nu_t^{(i)}$ independently from (2.9). The authors remark that enforcing strictly negative correlation between the offspring counts can improve the rate of convergence, but they do not specify how this might be achieved. Since we are not given a specific implementation, MVB is not discussed much in the remainder of the chapter, and is for instance omitted from Table 2.3 since many of the properties included there are not well-defined for MVB.

Srinivasan sampling procedure

Gerber, Chopin, and Whiteley (2019) build on the work of Crisan and Lyons (1999) in that they construct a resampling scheme for which the marginal offspring counts are distributed as (2.9), but the number of particles is held constant and non-negative correlation of offspring counts is enforced. The resulting scheme is termed *Srinivasan sampling procedure (SSP) resampling* after Srinivasan (2001).

The implementation is somewhat complicated compared to the other schemes we have seen (for full details see Gerber, Chopin, and Whiteley 2019, Algorithm 1) but a brief description is given here. The offspring counts are initialised at $N w_t^{(i)}$, then we iterate through pairs of counts, rounding one of the pair up and the other down by an amount such that at least one of the pair ends up an integer. After at most N such adjustments, all of the counts are integers and can be returned. Each iteration adds and subtracts the same amount so that the sum of the counts is preserved, ensuring that the number of particles remains constant. Which of the selected pair is increased/decreased in each iteration is chosen at random with probabilities that guarantee the resampling is unbiased.

As well as proposing this resampling scheme, Gerber, Chopin, and Whiteley (2019) make several other contributions to the SMC resampling literature, some of which will be discussed later.

2.4.3 Properties

In this section we consider some important properties of resampling schemes, and see how the example schemes of Section 2.4.2 compare in terms of these. The findings are summarised in Table 2.3, with the exception of a few properties which depend on details of the implementation or are applicable only to a subset of the resampling schemes considered.

Abbreviation	Description
`multi`	multinomial resampling
`star`	star resampling
`strat`	stratified resampling
`syst`	systematic resampling
`res-multi`	residual resampling with multinomial residuals
`res-star`	residual resampling with star residuals
`res-strat`	residual resampling with stratified residuals
`res-syst`	residual resampling with systematic residuals
`ssp`	Srinivasan sampling procedure resampling

Table 2.2: Abbreviations for resampling schemes

Support of offspring numbers

Recall that the weights give an indication of how useful each particle is for the approximation. Killing a high-weight particle is likely to increase the variance of the SMC estimates, while duplicating a low-weight particle wastes computational resources on propagating particles that will not contribute much to reducing that variance. One way to assess the performance of a given resampling scheme, then, is to consider the support of the marginal offspring distributions, conditional on the weights. This tells us how many duplicates it is possible to obtain from a particle with a given weight, and is therefore an indication of performance, albeit a rather crude one.

Suppose that $w_t^{(i)} \in [K/N, (K+1)/N]$. The value of K roughly determines how useful particle i is. Conditional on K, we will determine the range of possible values $\nu_t^{(i)}$ can take, under each of the resampling schemes described in Section 2.4.2.

Under multinomial resampling, it is possible for $\nu_t^{(i)}$ to take any value from 0 to N (although some values are of course more likely than others). Thus it is possible for a high-weight particle to have zero offspring, or a low-weight particle to have many offspring, simply by chance.

Residual resampling ensures that every particle with above-average (i.e. $> 1/N$) weight has at least one offspring, avoiding the loss of high-weight particles. If the residuals are sampled using multinomial resampling then the duplication of low-weight particles is not avoided, $\nu_t^{(i)} \in \{K, \ldots, K+R\} \subseteq \{K, \ldots, N\}$, but this can be addressed by using a lower-variance scheme for the residual offspring. Various choices are included in Table 2.3.

Stratified resampling is more restrictive, $\nu_t^{(i)} \in \{K-1, K, K+1, K+2\}$, but allows the possibility of a particle with above-average weight having no offspring. This is not quite as good as the erroneous claim of Douc, Cappé, and Moulines (2005) that $|\nu_t^{(i)} - N w_t^{(i)}| \leq 1$ for stratified resampling. Systematic resampling has the smallest support, $\nu_t^{(i)} \in \{K, K+1\}$, that is possible whilst maintaining unbiasedness, as do SSP and MVB resampling.

Another way to quantify this property is by considering the maximum possible difference between the offspring count $\nu_t^{(i)}$ and its expected value $Nw_t^{(i)}$. This is also presented in Table 2.3.

Degeneracy under equal weights

In the case where all of the weights are multiples of $1/N$, low-variance schemes such as residual and systematic resampling become fully deterministic. Since $\lfloor Nw_t^{(i)} \rfloor = Nw_t^{(i)}$ for each i, residual resampling will have $R = 0$, leaving no remainder to be assigned stochastically. In systematic resampling exactly $\lfloor Nw_t^{(i)} \rfloor = Nw_t^{(i)}$ samples will fall in the i^{th} interval. In particular, if $w_t^{(1:N)} = (1, \ldots, 1)/N$ then each parent is assigned exactly one offspring deterministcially, so there is effectively no resampling.

The same phenomenon occurs with stratified resampling, although not if one uses Whitley's roulette wheel description (Figure 2.8). The random phase shift introduced by "spinning the wheel" prevents the inversion sampling intervals from lining up exactly with the weight intervals, so the resampled offspring counts may vary from their means by one either side. Whitley (1994) does not describe stratified resampling, but we see that unlike with systematic resampling, in the case of startified resampling the roulette wheel description is not equivalent to the standard inversion sampling description. The roulette wheel adds some unnecessary extra randomness, so the straightforward inversion sampler is preferred.

When the state space is continuous, it is often the case that the event that all weights are multiples of $1/N$ has zero measure. Even so, with non-zero probability we may get arbitrarily close to this regime in which resampling becomes deterministic.

Marginal variance of offspring counts

Another indication of the performance of resampling is the variance of the resampled offspring counts. For instance we might ask what is the marginal variance of $\nu_t^{(i)}$, conditional on the corresponding weight $w_t^{(i)}$. We would like to keep this variance small, limiting the additional randomness introduced to our Monte Carlo estimates by the resampling step.

In multinomial resampling, the marginal distributions are

$$\nu_t^{(i)} \mid w_t^{(i)} \stackrel{d}{=} \text{Binomial}(N, w_t^{(i)})$$

so the variance is

$$\text{Var}[\nu_t^{(i)} \mid w_t^{(i)}] = Nw_t^{(i)}(1 - w_t^{(i)}).$$

Compare this to star resampling, where the marginal offspring counts

$$\nu_t^{(i)} \mid w_t^{(i)} \stackrel{d}{=} N \,\text{Bernoulli}(w_t^{(i)})$$

having variance

$$\mathrm{Var}[\nu_t^{(i)} \mid w_t^{(i)}] = N^2 w_t^{(i)}(1 - w_t^{(i)}),$$

N times larger than in the multinomial case.

As pointed out in Crisan and Lyons (1999, p.557), their MVB process yields offspring variance

$$\mathrm{Var}[\nu_t^{(i)} \mid w_t^{(i)}] = (Nw_t^{(i)} - \lfloor Nw_t^{(i)} \rfloor)(1 - Nw_t^{(i)} + \lfloor Nw_t^{(i)} \rfloor) \leq \frac{1}{4},$$

since the stochastic part of $\nu_t^{(i)}$ is a Bernoulli$(Nw_t^{(i)} - \lfloor Nw_t^{(i)} \rfloor)$ random variable (as seen in (2.9)). The same marginal variance appears from systematic, residual-systematic and SSP resampling, since these all share the same marginal offspring distributions. We will see in Section 2.4.4 that all of these schemes fall within the *stochastic rounding* class, and marginal offspring variance is a property shared by all stochastic roundings.

The marginal variance is harder to calculate for other schemes such as residual-multinomial and stratified resampling because these were not defined in terms of marginal distributions, nor are the offspring counts independent conditional on the weights. However, it is possible in some cases to find upper bounds on the variance, and some such bounds are derived below.

In residual-multinomial resampling, $\nu_t^{(i)}$ depends on all of the other weights as well as $w_t^{(i)}$, but only through the statistic $R := \sum(Nw_t^{(i)} - \lfloor Nw_t^{(i)} \rfloor)$. We have

$$\nu_t^{(i)} \mid w_t^{(i)}, R \stackrel{d}{=} \lfloor Nw_t^{(i)} \rfloor + \mathrm{Binomial}\left(R, \frac{Nw_t^{(i)} - \lfloor Nw_t^{(i)} \rfloor}{R}\right).$$

Using the law of total variance,

$$\begin{aligned}
\mathrm{Var}[\nu_t^{(i)} \mid w_t^{(i)}] &= \mathbb{E}\left[\mathrm{Var}[\nu_t^{(i)} \mid w_t^{(i)}, R] \,\middle|\, w_t^{(i)}\right] + \mathrm{Var}\left[\mathbb{E}[\nu_t^{(i)} \mid w_t^{(i)}, R] \,\middle|\, w_t^{(i)}\right] \\
&= \mathbb{E}\left[(Nw_t^{(i)} - \lfloor Nw_t^{(i)} \rfloor)\left(1 - \frac{Nw_t^{(i)} - \lfloor Nw_t^{(i)} \rfloor}{R}\right) \,\middle|\, w_t^{(i)}\right] \\
&\quad + \mathrm{Var}\left[Nw_t^{(i)} \,\middle|\, w_t^{(i)}\right] \\
&= Nw_t^{(i)} - \lfloor Nw_t^{(i)} \rfloor - (Nw_t^{(i)} - \lfloor Nw_t^{(i)} \rfloor)^2 \,\mathbb{E}[R^{-1} \mid w_t^{(i)}] \\
&\leq Nw_t^{(i)} - \lfloor Nw_t^{(i)} \rfloor.
\end{aligned}$$

Here we have excluded the case $R = 0$, in which the variance is zero. Similarly, for residual resampling with star residuals,

$$\nu_t^{(i)} \mid w_t^{(i)}, R \stackrel{d}{=} \lfloor Nw_t^{(i)} \rfloor + R\,\mathrm{Bernoulli}\left(\frac{Nw_t^{(i)} - \lfloor Nw_t^{(i)} \rfloor}{R}\right).$$

and we find

$$\begin{aligned}
\mathrm{Var}[\nu_t^{(i)} \mid w_t^{(i)}] &= \mathbb{E}\left[\mathrm{Var}[\nu_t^{(i)} \mid w_t^{(i)}, R] \,\middle|\, w_t^{(i)}\right] + \mathrm{Var}\left[\mathbb{E}[\nu_t^{(i)} \mid w_t^{(i)}, R] \,\middle|\, w_t^{(i)}\right] \\
&= \mathbb{E}\left[R(Nw_t^{(i)} - \lfloor Nw_t^{(i)} \rfloor)\left(1 - \frac{Nw_t^{(i)} - \lfloor Nw_t^{(i)} \rfloor}{R}\right) \,\middle|\, w_t^{(i)}\right] \\
&\quad + \mathrm{Var}\left[Nw_t^{(i)} \,\middle|\, w_t^{(i)}\right] \\
&= \mathbb{E}\left[R(Nw_t^{(i)} - \lfloor Nw_t^{(i)} \rfloor)\left(1 - \frac{Nw_t^{(i)} - \lfloor Nw_t^{(i)} \rfloor}{R}\right) \,\middle|\, w_t^{(i)}\right] \\
&= (Nw_t^{(i)} - \lfloor Nw_t^{(i)} \rfloor)\mathbb{E}[R \mid w_t^{(i)}] - (Nw_t^{(i)} - \lfloor Nw_t^{(i)} \rfloor)^2 \\
&\leq N(Nw_t^{(i)} - \lfloor Nw_t^{(i)} \rfloor).
\end{aligned}$$

Again, if $R = 0$ then the variance is zero.

For stratified resampling, we can use the constraints on the marginal offspring distribution that were derived in Section 2.4.2. Recall that, conditional on $w_t^{(i)}$, $\nu_t^{(i)} = \lfloor Nw_t^{(i)} \rfloor + j$ with probability p_j for $j = -1, 0, 1, 2$. We can use the expressions for p_{-1}, p_0, p_1, p_2 in the two cases of Figure 2.7, as summarised in Table 2.1, to bound the variance. First write

$$\begin{aligned}
\mathrm{Var}[\nu_t^{(i)} \mid w_t^{(i)}] &= \mathrm{Var}\left[\nu_t^{(i)} - \lfloor Nw_t^{(i)} \rfloor \,\middle|\, w_t^{(i)}\right] \\
&= \mathbb{E}\left[(\nu_t^{(i)} - \lfloor Nw_t^{(i)} \rfloor)^2 \,\middle|\, w_t^{(i)}\right] - \mathbb{E}\left[\nu_t^{(i)} - \lfloor Nw_t^{(i)} \rfloor \,\middle|\, w_t^{(i)}\right]^2 \\
&= p_{-1} + p_1 + 4p_2 - (-p_{-1} + p_1 + 2p_2)^2. \qquad (2.10)
\end{aligned}$$

Using the upper and lower bounds in Table 2.1 and then optimising over δ, we obtain the bound

$$\mathrm{Var}[\nu_t^{(i)} \mid w_t^{(i)}] \leq \frac{1}{4} + \frac{1+\delta}{2} + 1 - (0 + \frac{\delta}{2} + 0)^2 = \frac{1}{4}(7 + 2\delta - \delta^2) \leq 2.$$

Optimising the exact expressions in each case (first two columns in Table 2.1) does not improve this overall bound.

Residual-stratified resampling has the further constraint that $p_{-1} = 0$ (i.e. Figure 2.7b doesn't occur) since the residual weights are between 0 and $1/R$. Now the bounds in Table 2.1 are too loose, so we bound the variance by using the exact expressions from Table 2.1 in each case and optimising over δ_L, δ. Setting $p_{-1} = 0$ in (2.10), substituting the expressions for Case (a) from Table 2.1, and maximising over δ_L and then δ yields

$$\begin{aligned}
\mathrm{Var}[\nu_t^{(i)} \mid w_t^{(i)}] &= p_1 + 4p_2 - (p_1 + 2p_2)^2 = \delta - 2\delta_L(\delta - \delta_L) + 4\delta_L(\delta - \delta_L) - \delta^2 \\
&= \delta - \delta^2 + 2\delta\delta_L - 2\delta_L^2 \leq \delta - \frac{1}{2}\delta^2 \leq \frac{1}{2},
\end{aligned}$$

since the maximum is achieved at $\delta_L = \delta/2$ and then at $\delta = 1$.

Table 2.3 includes upper bounds on $\mathrm{Var}[\nu_t^{(i)}]$ for various resampling schemes, indepen-

dent of $w_t^{(i)}$. Those general bounds are derived from the results of this section, bounded above independently of the weights. Some of the bounds may not be tight. We could also try to bound this variance below, but for every resampling scheme the only lower bound valid for all $w_t^{(i)}$ is zero (consider the case $w_t^{(i)} = 0$) so this does not provide any more information.

Contribution to the Monte Carlo variance

While the variance of the offspring counts goes some way towards providing a comparison between resampling schemes, a more relevant property is the contribution of the resampling step to the Monte Carlo variance. This quantifies directly the effect of a certain choice of resampling scheme on the variance of the resulting Monte Carlo estimators.

Let $(\mathcal{G}_t)_{t\geq 0}$ be the filtration generated by the particle positions and weights up to and including time t, so $\mathcal{G}_t$ is the σ-algebra generated by $(X_{0:t}^{(1:N)}, w_{0:t}^{(1:N)})$. Consider the position of the ith particle in generation $t+1$ just after resampling but before mutating, that is $X_t^{(a_t^{(i)})}$. Define the one-step Monte Carlo variance induced by resampling as

$$\sigma(\varphi) := \mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{N}\varphi(X_t^{(a_t^{(i)})}) \,\middle|\, \mathcal{G}_t\right] \tag{2.11}$$

where φ is an arbitrary test function.

Some results comparing this variance across different resampling schemes are presented in Douc, Cappé, and Moulines (2005). Their results, plus some additional ones, are presented in Proposition 2.3. It may be possible to derive similar results regarding residual-stratified and SSP resampling, but such results are hard to obtain due to the strong dependence between parental indices induced by these resampling schemes. This remains an interesting open problem.

In the case of systematic (but not necessarily residual-systematic) resampling, no such variance comparison can be made. Systematic resampling generally yields low variance in practice, but it is possible to construct pathological cases in which it yields higher variance than multinomial resampling (Douc, Cappé, and Moulines 2005, Section 3.4) and it lacks theoretical support more generally (e.g. Gerber, Chopin, and Whiteley 2019, Section 3.3).

Proposition 2.3 (Variance of resampling schemes)**.** *Let σ_{multi} etc. denote the variance* (2.11) *under the various resampling schemes, as abbreviated in Table 2.2. For any square-integrable function φ,*

(a) $\sigma_{multi}(\varphi) \geq \sigma_{res\text{-}multi}(\varphi)$

(b) $\sigma_{multi}(\varphi) \geq \sigma_{strat}(\varphi)$

(c) $\sigma_{star}(\varphi) = N\sigma_{multi}(\varphi)$

(d) $\sigma_{res\text{-}star}(\varphi) \geq \sigma_{res\text{-}multi}(\varphi) \geq \sigma_{res\text{-}strat}(\varphi)$

(e) $\sigma_{star}(\varphi) \geq \sigma_{res\text{-}star}(\varphi)$

The partial ordering suggested by Proposition 2.3 is depicted graphically in Figure 2.9.

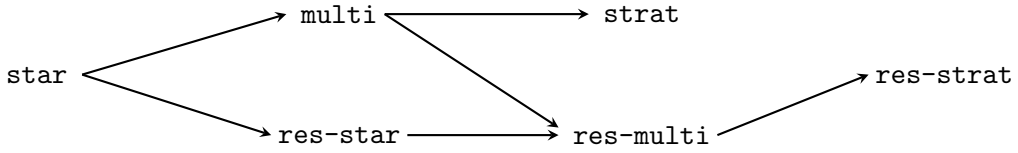


Figure 2.9: Graphical depiction of the conditional variance inequalities stated in Proposition 2.3. Conditional variance (2.11) is non-increasing along arrows.

Proof. **(a)** See Douc, Cappé, and Moulines (2005, Section 3).
(b) See Douc, Cappé, and Moulines (2005, Section 3).
(c) The following expression is derived in Douc, Cappé, and Moulines (2005, Equation (6)):

$$\sigma_{\texttt{multi}}(\varphi) = \frac{1}{N}\sum_{j=1}^{N} \varphi^2(X_t^{(j)}) w_t^{(j)} - \frac{1}{N}\left\{\sum_{j=1}^{N} \varphi(X_t^{(j)}) w_t^{(j)}\right\}^2 .$$

Under star resampling, all of the resampled indices are equal, say $X_t^{(a_t^{(1)})} = \cdots = X_t^{(a_t^{(N)})} = X_t^\star$, so

$$\begin{aligned}
\sigma_{\texttt{star}}(\varphi) &= \mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{N} \varphi(X_t^{(a_t^{(i)})}) \,\middle|\, \mathcal{G}_t\right] = \mathrm{Var}\left[\varphi(X_t^\star) \mid \mathcal{G}_t\right] \\
&= \mathbb{E}\left[\varphi^2(X_t^\star) \mid \mathcal{G}_t\right] - \mathbb{E}\left[\varphi(X_t^\star) \mid \mathcal{G}_t\right]^2 \\
&= \sum_{j=1}^{N} \varphi^2(X_t^{(j)}) \mathbb{P}[X_t^\star = X_t^{(j)} \mid \mathcal{G}_t] - \left\{\sum_{j=1}^{N} \varphi(X_t^{(j)}) \mathbb{P}[X_t^\star = X_t^{(j)} \mid \mathcal{G}_t]\right\}^2 \\
&= \sum_{j=1}^{N} \varphi^2(X_t^{(j)}) w_t^{(j)} - \left\{\sum_{j=1}^{N} \varphi(X_t^{(j)}) w_t^{(j)}\right\}^2 \qquad (2.12) \\
&= N\sigma_{\texttt{multi}}(\varphi),
\end{aligned}$$

as required.

(d) The second inequality follows from (b) and is stated in Gerber, Chopin, and Whiteley (2019, p.9). For the first inequality, we use the following expression which is a slight modification of Douc, Cappé, and Moulines (2005, Equation (8)):

$$\sigma_{\texttt{res-multi}}(\varphi) = \frac{R}{N^2}\sum_{j=1}^{N}\varphi^2(X_t^{(j)})r^{(j)} - \frac{R}{N^2}\left(\sum_{j=1}^{N}\varphi(X_t^{(j)})r^{(j)}\right)^2.$$

A derivation similar to theirs can also be used for residual-star resampling. First notice that, conditional on $\mathcal{G}_t$, the Monte Carlo estimate in (2.11) can be decomposed into a sum of conditionally deterministic terms plus a sum of stochastic terms:

$$\frac{1}{N}\sum_{i=1}^{N}\varphi(X_t^{(a_t^{(i)})}) = \frac{1}{N}\sum_{j=1}^{N}\lfloor Nw_t^{(j)}\rfloor\varphi(X_t^{(j)}) + \frac{1}{N}\sum_{i=1}^{R}\varphi(\hat{X}_t^{(i)}),$$

where the terms in the second sum are all equal, say $\hat{X}_t^{(1)} = \cdots = \hat{X}_t^{(R)} = X_t^\star$. The first sum is conditionally deterministic and hence does not contribute to the Monte Carlo variance (2.11). We have

$$\begin{aligned}
\sigma_{\texttt{res-star}}(\varphi) &= \operatorname{Var}\left[\frac{1}{N}\sum_{i=1}^{R}\varphi(\hat{X}_t^{(i)})\,\middle|\,\mathcal{G}_t\right] = \frac{R^2}{N^2}\operatorname{Var}\left[\varphi(X_t^\star)\,|\,\mathcal{G}_t\right] \\
&= \frac{R^2}{N^2}\mathbb{E}\left[\varphi^2(X_t^\star)\mid\mathcal{G}_t\right] - \frac{R^2}{N^2}\mathbb{E}\left[\varphi(X_t^\star)\mid\mathcal{G}_t\right]^2 \\
&= \frac{R^2}{N^2}\sum_{j=1}^{N}\varphi^2(X_t^{(j)})\mathbb{P}[X_t^\star = X_t^{(j)}\mid\mathcal{G}_t] - \frac{R^2}{N^2}\left\{\sum_{j=1}^{N}\varphi(X_t^{(j)})\mathbb{P}[X_t^\star = X_t^{(j)}\mid\mathcal{G}_t]\right\}^2 \\
&= \frac{R^2}{N^2}\sum_{j=1}^{N}\varphi^2(X_t^{(j)})r^{(j)} - \frac{R^2}{N^2}\left\{\sum_{j=1}^{N}\varphi(X_t^{(j)})r^{(j)}\right\}^2 \qquad (2.13) \\
&= R\sigma_{\texttt{res-multi}}(\varphi) \\
&\geq \sigma_{\texttt{res-multi}}(\varphi)
\end{aligned}$$

whenever $R \geq 1$. If $R = 0$ then all residual schemes have zero variance and (d) holds trivially.

(e) We have from (2.12)

$$\sigma_{\texttt{star}}(\varphi) = \sum_{j=1}^{N}\varphi^2(X_t^{(j)})w_t^{(j)} - \left\{\sum_{j=1}^{N}\varphi(X_t^{(j)})w_t^{(j)}\right\}^2$$

and from (2.13), noting that $r^{(j)} := (Nw_t^{(j)} - \lfloor Nw_t^{(j)} \rfloor)/R \leq Nw_t^{(j)}/R$,

$$\begin{aligned}
\sigma_{\texttt{res-star}}(\varphi) &= \frac{R^2}{N^2} \sum_{j=1}^{N} \varphi^2(X_t^{(j)}) r^{(j)} - \frac{R^2}{N^2} \left\{ \sum_{j=1}^{N} \varphi(X_t^{(j)}) r^{(j)} \right\}^2 \\
&\leq \frac{R^2}{N^2} \sum_{j=1}^{N} \varphi^2(X_t^{(j)}) \frac{Nw_t^{(j)}}{R} - \frac{R^2}{N^2} \left\{ \sum_{j=1}^{N} \varphi(X_t^{(j)}) \frac{Nw_t^{(j)}}{R} \right\}^2 \\
&= \frac{R}{N} \sum_{j=1}^{N} \varphi^2(X_t^{(j)}) w_t^{(j)} - \left\{ \sum_{j=1}^{N} \varphi(X_t^{(j)}) w_t^{(j)} \right\}^2 \\
&\leq \sigma_{\texttt{star}}(\varphi),
\end{aligned}$$

since $R \leq N - 1$. ■

Exchangeability of offspring

We say that a resampling scheme leaves the offspring exchangeable if the resulting distribution of parental indices is invariant under permutations of the offspring. To put it another way, each child chooses its parent from the same marginal distribution.

It is clear that true multinomial resampling satisfies this property since the parental indices are independent and distributed according to the same Categorical distribution. The same goes for star resampling. However, as mentioned earlier, the efficient implementation of multinomial resampling that takes sorted inputs does not leave the offspring exchangeable. Stratified and systematic resampling do not either since their inversion sampling points are sorted: for instance, child 1 is more likely to choose parent 1 than child N is. Residual resampling schemes are also typically implemented in such a way that the offspring are not exchangeable.

Whichever resampling scheme is used, exchangeability of offspring can easily be reintroduced, at $O(N)$ cost, by applying a random permutation to the vector of parental indices after sampling.

Operations in SMC that depend on the ancestral indices are typically independent of ordering, so sampling ancestral indices from a non-exchangeable distribution is not expected to cause any problem. (A notable exception is conditional SMC, which is why some care is needed when implementing conditional versions of non-exchangeable resampling schemes.) However, the results of Chapters 3 and 4 rely on the random assignment assumption (A1) which amounts to exchangeability of offspring, so to be sure that the current genealogical study applies, a permutation should be appended to any non-exchangeable resampling procedure.

Permutation sensitivity and sorting

Some resampling schemes are sensitive to the order in which the weights are input. That is, permuting the weight vector before resampling can affect the distribution of the resulting offspring counts. Note that this is different to the permutations of offspring discussed in the previous section; here it is the weights, i.e. the parents, that are permuted.

To give a concrete example, consider resampling schemes based on inversion sampling (multinomial, stratified, systematic). Figure 2.10 shows two partitions of $[0, 1]$ each constructed from a permutation of the weight vector $w^{(1:6)} = (0.25, 0.05, 0.1, 0.35, 0.2, 0.05)$. Under multinomial resampling this does not affect the distribution of the offspring counts, although it will affect the distribution of the parental indices if the fast implementation is used.

On the other hand, under stratified or systematic resampling the distribution of offspring counts is different for the two partitions. To see this, consider parents 2 and 6. When the weights are sorted, the probability that both of these parents are assigned a non-zero number of offspring is zero, because both of their subintervals lie within the same subinterval of length $1/N$, which gets exactly one inversion sampling point. When the weights are in their natural order, as in the top row of Figure 2.10, it is possible under stratified and systematic resampling for both parents 2 and 6 to be assigned one offspring. Clearly, then, the distribution of offspring counts under these resampling schemes differs between the two orderings of the weight vector pictured. This property is also pointed out in Douc, Cappé, and Moulines (2005, p.66). Table 2.3 includes a summary of which resampling schemes are permutation-sensitive or not.

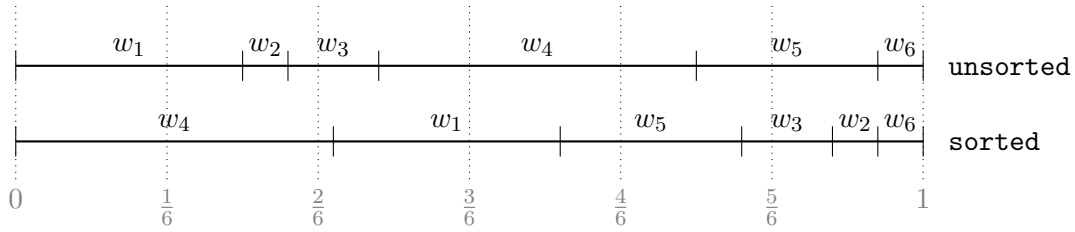


Figure 2.10: An example in which permuting the weights can affect the conditional distribution of offspring counts under certain resampling schemes. As in Figure 2.6, $N = 6$ and $w^{(1:6)} = (0.25, 0.05, 0.1, 0.35, 0.2, 0.05)$. The top row shows the weighted subintervals in the natural order, as in Figure 2.6. The bottom row shows the partition corresponding to the same weights, but sorted in decreasing order. The dotted lines are spaced $1/N$ apart. Under permutation-sensitive resampling schemes, the distribution of offspring counts differs depending on which partition is used.

Related to this phenomenon, Gerber, Chopin, and Whiteley (2019) prove some striking theoretical results concerning the effects of pre-sorting the particles. They show that sorting the particles in order of their states prior to resampling improves the rate of decay of resampling error from the usual $O(N^{-1})$ to $O(N^{-1-\frac{1}{d}})$, where d is the dimension of the state space. In dimension $d = 1$, this supports the numerical results of Kitagawa (1996),

who observed empirically that sorting improved the convergence rate from $O(N^{-1})$ to $O(N^{-2})$ when working in one dimension.

In dimension $d \geq 2$ things are more complicated because there is no full ordering of the state space. Gerber, Chopin, and Whiteley (2019) get around this by mapping the state space onto $[0, 1]^d$ and sorting by the Hilbert curve. The variance reduction from sorting the particles diminishes as the dimension increases, so in practice this has to be weighed up against the $O(N \log N)$ cost of sorting.

Another remarkable result of Gerber, Chopin, and Whiteley (2019) is that, when the particles are sorted by their states, systematic resampling admits some theoretical support that was lacking in the unsorted case. Recall that the possibly pathological behaviour of systematic resampling was related to "bad" orderings of the weight intervals; sorting the particles evidently prevents this.

The intuition behind these results is that sorting particles by their states ensures that the stratified and systematic resampling schemes select parents from a good range of locations in state space. The sorting step prevents the sampled parents being concentrated in one small part of the state space purely by a chance ordering of the weight intervals. Another explanation (Li et al. 2020; Webber 2019) is based on the observation that, under stratified or systematic resampling, the possible parents of a given offspring are always consecutive in the order in which the weights are input. Sorting these weights in order of the particle states ensures that these potential parents are "close" in state space, so that the state after resampling does not differ drastically depending on which parent is selected.

These results are only relevant to resampling schemes based on inversion sampling with fairly evenly-spaced points. Notably, multinomial resampling is not affected by sorting, since it is invariant under permutations of the weight vector.

Computational complexity

All of the resampling algorithms discussed in Section 2.4.2 can be implemented in $O(N)$ operations. Considering the complexity of each operation, Hol, Schön, and Gustafsson (2006) suggest that systematic resampling is fastest because it only requires one pseudo-random number generation, and multinomial resampling is slower than stratified resampling because of the transformations required (although this may depend on which method is used to sample the Uniform order statistics). Residual resampling is hard to compare directly because a random fraction of the operations are deterministic, so the number of pseudo-random numbers required is a random number between 0 and $N - 1$, but the authors' simulation experiments place it between multinomial and stratified resampling.

However, the analysis of per-particle cost is sensitive to the particular implementation of each resampling scheme, the system implementation of pseudo-random number generation and arithmetic operations, and the hardware used, so it is not clear how robust such comparisons are.

Negative association

Following Gerber, Chopin, and Whiteley (2019), we use the definition of negative association from Joag-Dev and Proschan (1983).

Definition 2.4. Let $(Z_1, \ldots, Z_n)$ be a collection of random variables. $Z_{1:n}$ are said to be *negatively associated* if, for every disjoint pair of subsets $I, J \subseteq \{1, \ldots, n\}$, for all real-valued coordinatewise non-decreasing functions φ, ψ for which the covariance is well defined,

$$\operatorname{Cov}\left[\varphi(Z_I), \psi(Z_J)\right] \leq 0.$$

Gerber, Chopin, and Whiteley (2019) show that negative association of offspring counts is a desirable property which may be used, along with some other machinery, to establish certain weak convergence results for the resampled measures.

Multinomial counts are negatively associated (Joag-Dev and Proschan 1983, Section 3.1), which implies that residual-multinomial resampling also satisfies this property. Gerber, Chopin, and Whiteley (2019) construct a counter-example to demonstrate that systematic resampling violates the negative association property. For residual-systematic resampling, we can cook up a counterexample in the same spirit by taking $\varphi(x) = \psi(x) = \mathbb{1}_{\{x=1\}}$, $I = \{1\}$, $J = \{3\}$ and considering a weight vector say $w^{(1:4)} = \frac{1}{8}(1, 1, 1, 5)$ for $N = 4$. Then the residual weights are $r^{(1:4)} = \frac{1}{4}(1, 1, 1, 1)$ with $R = 2$, so

$$\begin{aligned}
\operatorname{Cov}\left[\varphi(Z_I), \psi(Z_J)\right] &= \mathbb{E}[\varphi(Z_I)\psi(Z_J)] - \mathbb{E}[\varphi(Z_I)]\mathbb{E}[\psi(Z_J)] \\
&= \mathbb{P}[\nu^{(1)} = 1, \nu^{(3)} = 1] - \mathbb{P}[\nu^{(1)} = 1]\mathbb{P}[\nu^{(3)} = 1] \\
&= \frac{1}{2} - \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} > 0,
\end{aligned}$$

since the residual weight intervals corresponding to parents 1 and 3 both occupy the first half of a length-$(1/R)$ interval, hence $\{\nu^{(1)} = 1\}$ and $\{\nu^{(3)} = 1\}$ each have probability $1/2$ and $\nu^{(3)} = 1$ if and only if $\nu^{(1)} = 1$. So residual-systematic resampling also violates the negative association property.

Gerber, Chopin, and Whiteley (2019) also mention some resampling schemes that do result in negatively associated counts: stratified resampling, and by implication residual-stratified resampling; star resampling (see the remark at the end of Gerber, Chopin, and Whiteley (2019, Section 3.2)), and by implication residual-star resampling. The authors go on to introduce the SSP resampling scheme, which yields negatively associated offspring counts by construction.

These results are summarised in Table 2.3. The MVB algorithm does not enforce negative association, so this property depends on the particular implementation, and as such is left blank in Table 2.3.

Star discrepancy

The *star discrepancy* is a measure of the regularity of a given set of points $u_{1:N}$ in the unit hypercube. For our purposes it is sufficient to define the star discrepancy in one dimension, as in Kuipers and Niederreiter (1974, Definition 1.2):

$$D^\star(u_1, \ldots, u_N) := \sup_{u \in [0,1]} |d(u)| := \sup_{u \in [0,1]} \left| u - \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{u_i \leq u\}} \right|. \tag{2.14}$$

The quantity inside the supremum is the difference between the empirical CDF of the observed points $u_{1:N}$ and the CDF of the Uniform distribution on $[0, 1]$. Thus $D^\star$ measures, in a certain sense, how far the points are from being uniformly spaced.

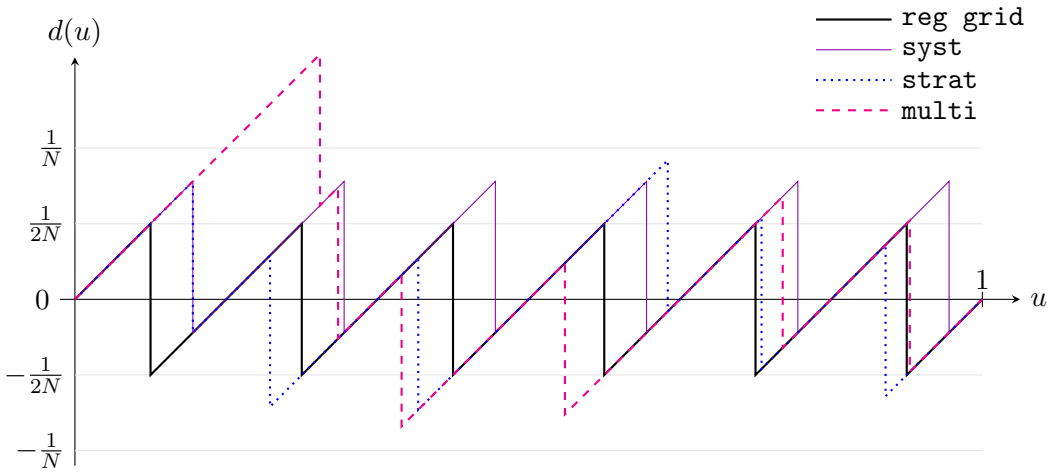


Figure 2.11: Plot of the function inside the absolute value in (2.14), for four different point sets. The points $u_{1:6}$ used are the same as in Figure 2.6.
The solid black line corresponds to the regular grid, which achieves the minimal discrepancy $1/(2N)$, but cannot be used for resampling. The star discrepancy of stratified and systematic points varies between $1/(2N)$ and $1/N$ depending on the realisation. In this example, the star discrepancy of the systematic points is $0.78/N$ and of the stratified points is $0.92/N$. The star discrepancy of standard multinomial resampling (that is, i.i.d. Uniform points) can be arbitrarily close to 1 for "bad" realisations; in this example it is $1.62/N$.

Star discrepancy is used in quasi-Monte Carlo, where *low-discrepancy* points are used in place of Uniform random numbers to decrease the variance of Monte Carlo estimates. We have noted already that resampling can itself be viewed as a Monte Carlo procedure. From this point-of-view, stratified and systematic resampling are quasi-Monte Carlo implementations of multinomial resampling, since they provide "more regular" point sets to be used in inversion sampling.

In one dimension, the lowest-discrepancy point set is the regular grid $\frac{1}{2N}(1, 3, \ldots, 2N - 1)$, which has star discrepancy $\frac{1}{2N}$ (see for example Kuipers and Niederreiter 1974, Corollary 1.2). However, resampling based on a deterministic point set cannot be unbiased since the resulting parental indices are conditionally deterministic given the weights. Systematic

resampling amounts to a randomisation of the regular grid, shifting each grid point by a random amount $u \sim \text{Uniform}[0, 1/N]$, which corresponds to a randomised quasi-Monte Carlo procedure. This yields star discrepancy $D^\star = \max\{u, \frac{1}{N} - u\}$, which is between $1/(2N)$ and $1/N$ almost surely. The point sets generated in stratified resampling also have star discrepancy between $1/(2N)$ and $1/N$, where the exact value depends on the realisation. This certainly seems to improve on independent uniform points which can have star discrepancy arbitrarily close to 1, the maximum possible value, albeit with diminishing probability as N increases. Figure 2.11 illustrates how the star discrepancy is computed, and how it compares between these sampling methods.

Matrix resampling

Some resampling schemes render the parental indices $a_t^{(1:N)}$ conditionally independent over i given the weights $w_t^{(1:N)}$, and such schemes admit a matrix representation conditional on the weights. These resampling matrices are of particular interest in distributed SMC as they can be used to characterise the communication between particles required in the resampling step. The matrices have been defined differently by different authors (cf. Webber 2019; Whiteley, Lee, and Heine 2016), but the general idea is the same. Following most closely to the presentation of Li et al. (2020), we use the following definition.

Definition 2.5. A *resampling matrix* for weights $w_t^{(1:N)}$ is a $N \times N$ matrix with entries in $[0, 1]$ such that:

1. each row sums to 1, and
2. the i^{th} column sums to $Nw_t^{(i)}$, for each $i \in \{1, \ldots, N\}$.

The ij^{th} element of the resampling matrix represents the conditional probability of offspring i being resampled from parent j. Some resampling schemes that are expressible as matrices are:

- no resampling, for which the resampling matrix is the identity matrix;
- multinomial resampling, for which each row of the resampling matrix is the vector of weights;
- stratified resampling;
- residual-multinomial resampling;
- residual-stratified resampling.

Illustrations of the structure of the resampling matrix resulting from each of these schemes can be found in Li et al. (2020, Figure 2), for example, although the residual resampling matrices may differ depending on the implementation. All of the other resampling schemes

we have encountered have some conditional dependence between the parental indices (as summarised in Table 2.3) and thus are not representable by matrices.

Within the matrix resampling class, Li et al. (2020) show that, in one dimension, stratified resampling on the sorted particles is optimal in terms of the conditional variance (2.11), where sorting is based on the test function φ applied to the states. They also prove that this scheme is optimal in other senses. Webber (2019) proves a generalisation to multiple dimensions: stratified resampling with the particles sorted by a certain functional of the states, based on φ, minimises the corresponding conditional variance. The optimal functional by which to sort the particles cannot typically be computed, but the author suggests alternative sorting rules that also significantly improve performance, including the Hilbert curve sorting proposed by Gerber, Chopin, and Whiteley (2019). Webber (2019) also uses the matrix representations to construct alternative proofs of several of the results of Proposition 2.3.

For our purposes, however, this class is too restrictive, as it excludes several resampling schemes that are prevalent in the literature and which perform comparably to, if not better than, conditionally independent resampling schemes.

	support of $\nu_t^{(i)}$ given $\frac{K}{N} \leq w_t^{(i)} < \frac{K+1}{N}$	worst case $\lvert\nu_t^{(i)} - N w_t^{(i)}\rvert$	degenerate if $w_t^{(1:N)} = \frac{1}{N}(1, \ldots, 1)$?	upper bound on $\mathrm{Var}[\nu_t^{(i)}]$	sensitive to permutations of weights?	PRNG calls	neg. assoc.?	cond. indep.?	stochastic rounding?
`multi`	$\{0, \ldots, N\}$	N	×	$N/4$	×	N	✓	✓	×
`star`	$\{0, N\}$	N	×	$N^2/4$	×	1	✓	×	×
`strat`	$\{K-1, K, K+1, K+2\}$	2	✓	2	✓	N	✓	✓	×
`syst`	$\{K, K+1\}$	1	✓	1/4	✓	1	×	×	✓
`res-multi`	$\{K, \ldots, N\}$	$N-1$	✓	1	×	$\leq N-1$	✓	✓	×
`res-star`	$\{K, N\}$	$N-1$	✓	N	×	1	✓	×	×
`res-strat`	$\{K, K+1, K+2\}$	2	✓	1/2	✓	$\leq N-1$	✓	✓	×
`res-syst`	$\{K, K+1\}$	1	✓	1/4	✓	1	×	×	✓
`ssp`	$\{K, K+1\}$	1	✓	1/4	×	$\leq N$	✓	×	✓

Table 2.3: Summary of some of the properties of resampling schemes explored in Section 2.4.3. Columns appear in order of their explanations in the text. The abbreviated names for the resampling schemes are explained in Table 2.2.

2.4.4 Stochastic rounding

Some of the resampling schemes we have met can be classified as *stochastic roundings*. This will be useful later on, as we will see (Section 5.3) that all members of this class admit some common convergence results. The stochastic roundings class is a subclass of MVB resampling (Section 2.4.2) additionally constrained to satisfy condition 1 of Definition 2.2.

Definition 2.6. Let $X = (X_1, \dots, X_N)$ be a $\mathbb{R}_+^N$-valued random variable. Then $Y = (Y_1, \dots, Y_N) \in \mathbb{N}^N$ is a *stochastic rounding* of X if each element Y_i takes values

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

By construction, $\mathbb{E}(Y_i) = X_i$ for each i. Taking X to be N times the vector of particle weights, we can therefore use stochastic rounding to construct a valid resampling scheme, under the further constraint that $Y_1 + \cdots + Y_N = N$. Several ways to enforce this constraint on the joint distribution have been proposed, including systematic resampling, residual resampling with systematic residuals, and SSP resampling.

Explicitly, the offspring counts are marginally distributed according to

$$\nu_t^{(i)} \mid w_t^{(i)} \overset{d}{=} \lfloor N w_t^{(i)} \rfloor + \text{Bernoulli}(N w_t^{(i)} - \lfloor N w_t^{(i)} \rfloor).$$

Some of the properties discussed earlier are common to every stochastic rounding scheme. Since all such schemes give offspring counts with the same marginal distributions, properties such as the marginal offspring variance are common to all stochastic roundings. Indeed it is easy to see that the marginal variance of the offspring counts, $\text{Var}[\nu_t^{(i)} \mid w_t^{(i)}]$ is as small as possible under the constraint of unbiasedness, and as such this is sometimes referred to as minimal-variance resampling. By definition, the support of an offspring count $\nu_t^{(i)}$ given that the associated weight lies in the interval $K/N \leq w_t^{(i)} < (K+1)/N$ is $\{K, K+1\}$. All stochastic roundings are also degenerate when the weights are all equal, i.e. $w_t^{(1:N)} = (1, \dots, 1)/N$ implies $\nu_t^{(1:N)} = (1, \dots, 1)$ almost surely.

2.5 Conditional SMC

Andrieu, Doucet, and Holenstein (2010) propose a number of *particle MCMC* algorithms, which combine SMC with MCMC in order to improve performance in certain situations. One of their algorithms, the *particle Gibbs* sampler (Andrieu, Doucet, and Holenstein 2010, Section 2.4.3), is of particular interest in the current work. For one thing, genealogies are particularly critical to its performance, and for another, the particle update uses a variant SMC algorithm which alters the distribution of genealogies.

In this section, we first introduce the particle Gibbs algorithm and the conditional SMC update, then discuss how ancestral degeneracy impacts the performance of particle Gibbs

and how ancestor sampling mitigates this.

2.5.1 Particle Gibbs

To motivate the particle Gibbs algorithm, we introduce a parametrised state space model and explain how combining SMC updates with MCMC sampling allows us to tackle the related inferences effectively. The particle Gibbs algorithm can be applied much more broadly, but this application is particularly intuitive and exhibits all the features of interest to our genealogical study.

Consider a parametrised state space model of the form

$$\begin{aligned} \theta &\sim p(\cdot) \\ X_0 &\sim \mu^\theta(\cdot) \\ X_{t+1} \mid X_t &\sim K^\theta_{t+1}(\cdot \mid X_t) && \text{for } t = 0, \dots, T-1 \\ Y_t \mid X_t &\sim g^\theta_t(\cdot \mid X_t) && \text{for } t = 0, \dots, T \end{aligned}$$

exactly like (2.1) except that the specification is now parametrised by θ (which may be multi-dimensional), and we place a prior distribution on θ. As usual, p, μ^θ, (K^θ_t) and (g^θ_t) are part of the model and are assumed to be known but not necessarily tractable.

Suppose that, given some data $y_{0:T}$, we wish to generate Monte Carlo samples from the joint posterior distribution of $X_{0:T}$ and θ. (Even if we are only interested in inferring θ, for instance, it is often more practical to target the joint posterior and then marginalise.) Notice that we are now working with a finite time horizon $T \in \mathbb{N}$. The inference of interest here is not inherently sequential; we are building an MCMC algorithm to sample from a single target distribution which happens to include some sequentially correlated components.

The conditional dependence structure of the model invites the use of a Gibbs sampler, sampling alternately from the conditional distributions $p(\theta \mid x_{0:T}, y_{0:T})$ and $p(x_{0:T} \mid \theta, y_{0:T})$. The θ update,

$$p(d\theta \mid x_{0:T}, y_{0:T}) \propto p(d\theta) p(x_{0:T}, y_{0:T} \mid \theta),$$

is often quite straightforward, if not analytically then by employing a Metropolis-Hastings step based on the current sampled values of θ and $x_{0:T}$. The X update, meanwhile, is high-dimensional with strong sequential correlations: exactly the situation in which one might use SMC. For the X update, we need a sample from

$$p(dx_{0:T} \mid \theta, y_{0:T}) \propto \mu^\theta(dx_0) g^\theta_0(y_0 \mid x_0) \prod_{s=1}^{T} K^\theta_s(dx_s \mid x_{s-1}) g^\theta_s(y_s \mid x_s), \tag{2.15}$$

which can be approximately obtained by running an SMC smoother then sampling one trajectory from its output in proportion to the associated weight.

However, the Markov chain associated to the procedure just described does not admit $p(x_{0:T}, \theta \mid y_{0:T})$ as an invariant distribution. It *approximately* targets this distribution, with some bias. A Gibbs sampler targeting $p(x_{0:T}, \theta \mid y_{0:T})$ exactly can be constructed by replacing the SMC step with a *conditional SMC* step, which takes into account the value of $x_{0:T}$ sampled at the previous iteration, as well as the observations and the current value of θ.

A conditional SMC algorithm for this scenario is presented in Algorithm 2.2. In contrast to Algorithm 2.1, the input now includes $x^\star_{0:T}$ and $a^\star_{0:T}$, which encode the states and parental indices, respectively, of the *immortal trajectory* (so called because it "survives" the SMC run with probability one). Within a particle Gibbs algorithm, the immortal trajectory is set to the trajectory sampled at the previous iteration. The resampling step now assigns the immortal offspring to the immortal parent deterministically, and the state of the immortal particle is also updated deterministically rather than via the Markov kernel. As in standard SMC, there is a choice of RESAMPLE procedures, but some care is needed to ensure the correct treatment of the immortal particle (for details see e.g. Lee, Murray, and Johansen 2019). In the case of multinomial resampling, exchangeability of the offspring means that conditioning on $a^\star_{t-1}$ has no effect on the resampling of the non-immortal particles.

Input: $T, N, \mu^\theta, (K^\theta_t), (g^\theta_t), y_{0:T}, x^\star_{0:T}, a^\star_{0:T}$
Set $X_0^{(a^\star_0)} \leftarrow x^\star_0$
for $i \in \{1, \ldots, N\} \setminus a^\star_0$ **do** Sample $X_0^{(i)} \sim \mu(\cdot)$
for $i \in \{1, \ldots, N\}$ **do** $w_0^{(i)} \leftarrow \left\{\sum_{j=1}^N g^\theta_0(y_0 \mid X_0^{(j)})\right\}^{-1} g^\theta_0(y_0 \mid X_0^{(i)})$
for $t \in \{1, \ldots, T\}$ **do**
  Set $a_{t-1}^{(a^\star_t)} \leftarrow a^\star_{t-1}$, $X_t^{(a^\star_t)} \leftarrow x^\star_t$
  Sample $a_{t-1}^{(1:N)} \setminus a^\star_{t-1} \sim \text{RESAMPLE}(\{1, \ldots, N\}, w_{t-1}^{(1:N)} \mid a^\star_{t-1})$
  for $i \in \{1, \ldots, N\} \setminus a^\star_t$ **do** Sample $X_t^{(i)} \sim K^\theta_t(\cdot \mid X_{t-1}^{(a_{t-1}^{(i)})})$
  for $i \in \{1, \ldots, N\}$ **do** $w_t^{(i)} \leftarrow \left\{\sum_{j=1}^N g^\theta_t(y_t \mid X_t^{(j)})\right\}^{-1} g^\theta_t(y_t \mid X_t^{(i)})$
end

Algorithm 2.2: Conditional sequential Monte Carlo for a parametrised state space model. The immortal particle at each generation has its new state and parental index set deterministically according to the values of $x^\star_{0:T}$ and $a^\star_{0:T}$ given as input.

The complete particle Gibbs algorithm for this example then consists of alternately sampling from the full conditional distribution of θ (e.g. using a Metropolis-Hastings update) and sampling a trajectory $(x_{0:T}, a_{0:T})$ using conditional SMC. See Andrieu, Doucet, and Holenstein (2010, Section 2.4.3) for more details.

2.5.2 Ancestral degeneracy in particle Gibbs

We have seen in Section 2.3 that the phenomenon of ancestral degeneracy can severely affect the performance of SMC algorithms, particularly in smoothing applications. The SMC update of particle Gibbs is a smoothing problem, however it requires only one sampled trajectory from the smoothing distribution, so one might imagine that we are safe from the curse of ancestral degeneracy. In fact, the loss of ancestors causes a different problem for particle Gibbs: it prevents some components of the Markov chain being refreshed, so that the chain mixes slowly.

To see this, consider the illustration in Figure 2.12, which shows the smoothing trajectories generated by a conditional SMC update at some iteration r. The thick black line is the immortal trajectory given as input, that is, the trajectory sampled by the conditional SMC update at iteration $r-1$. Backwards in time, the sampled trajectories quickly coalesce until at time 20 all of the trajectories have coalesced. The common trajectory from time 0 to 20 must necessarily be part of the immortal trajectory. A new trajectory (highlighted in purple) is then sampled among the N generated trajectories. Whichever trajectory we sample, it will certainly overlap with the previously sampled trajectory at least from time 0 to 20.

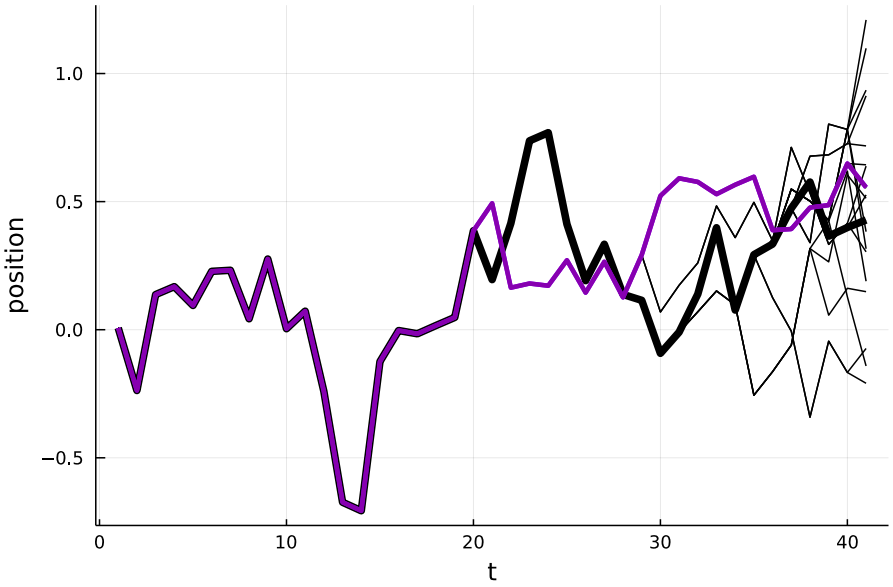


Figure 2.12: Illustration of how ancestral degeneracy causes particle Gibbs to mix slowly on some components. The thick black line is the immortal trajectory, i.e. the sampled trajectory from the previous iteration. Other lines are all of the trajectories generated by conditional SMC. One of these (highlighted in purple) is the sampled trajectory at the current iteration. Due to ancestral degeneracy, the current sample (purple) coincides with the previous sample (thick black) up to time 20, so the components $x_{0:20}$ are not updated in this iteration.

At the next iteration the newly sampled trajectories will again coalesce onto the im-

mortal trajectory, and this behaviour is repeated. If T is too large with respect to N, the early part of the trajectory will very rarely be updated, so the corresponding states will mix very slowly. For further intuition on this phenomenon the reader is directed to Lindsten and Schön (2013, Section 5.4).

The meaning of T "too large" here depends on the model and the type of SMC update used, but typically T is determined by the application and N is limited by computational resources, so we may not be able to control their relative size. The other brute-force approach would be to increase the number of iterations of the MCMC algorithm, but this too is infeasible on a limited computational budget. It is therefore worth investing some effort to find an alternative solution to the problem of ancestral degeneracy within particle Gibbs.

2.5.3 Ancestor sampling

An effective solution (where it is possible to implement it) was proposed by Whiteley (2010) and is known as *ancestor sampling*. It consists of a simple modification to the resampling step within the conditional SMC algorithm. In the basic algorithm with multinomial resampling, at each time step the non-immortal particles are resampled by multinomial resampling according to their weights, while the immortal offspring is deterministically assigned to the immortal parent. That is, at time t, for each $i \in \{1, \ldots, N\}$,

$$\mathbb{P}\left[a_t^{(j)} = i \,\middle|\, X_{0:t}^{(1:N)}, x_{0:T}^\star, a_{0:T}^\star\right] \propto \begin{cases} w_t^{(i)} & j \text{ non-immortal} \\ \mathbb{1}_{\{i=a_t^\star\}} & j \text{ immortal.} \end{cases}$$

Ancestor sampling combines the resampling step with a backward simulation step for the immortal particle. Instead of deterministically inheriting the immortal parent, the immortal particle samples its parent among all N possible parents. This is justified in the same way as backward simulation in general (Section 2.3.1), provided the ancestor sampling probabilities are chosen correctly, although we now apply the backward simulation step to the immortal trajectory only. Ancestor sampling can also be implemented for other choices of RESAMPLE, using the same backward simulation probabilities (but of course the resampling probabilities for non-immortal particles will be different, and there may be some additional dependence between parental indices). For simplicity we here restrict ourselves to multinomial resampling.

Assume that the smoothing distributions admit densities, that is $\mu^\theta(\cdot)$ and $K_t^\theta(\cdot \mid x)$ admit densities for all x, t. Denote the density of K_t^θ by q_t^θ. Define the trajectories $X_{t,0:t}^{(i)}$ (for any t, i) as in Section 2.1.4, starting from $X_{t,t}^{(i)} := X_t^{(i)}$ and tracing back the states of the parents via $X_{t,s}(i) = X_{t,s+1}^{(a_t^{(i)})}$. Then the correct resampling probabilities are, for each i,

$$\mathbb{P}\left[a_t^{(j)} = i \,\middle|\, X_{0:t}^{(1:N)}, x_{0:T}^\star, a_{0:T}^\star\right] \propto \begin{cases} w_t^{(i)} & j \text{ non-immortal} \\ w_t^{(i)} \frac{p((X_{t,0:t}^{(i)}, x_{t+1:T}^\star) \mid \theta, y_{0:T})}{p(X_{t,0:t}^{(i)} \mid \theta, y_{0:t})} & j \text{ immortal.} \end{cases} \tag{2.16}$$

Input: $T, N, \mu^\theta, (K_t^\theta), (q_t^\theta), (g_t^\theta), y_{0:T}, x_{0:T}^\star, a_{0:T}^\star$
Set $X_0^{(a_0^\star)} \leftarrow x_0^\star$
for $i \in \{1, \ldots, N\} \setminus a_0^\star$ **do** Sample $X_0^{(i)} \sim \mu^\theta(\cdot)$
for $i \in \{1, \ldots, N\}$ **do** $w_0^{(i)} \leftarrow \left\{\sum_{j=1}^N g_0^\theta(y_0 \mid X_0^{(j)})\right\}^{-1} g_0^\theta(y_0 \mid X_0^{(i)})$
for $t \in \{1, \ldots, T\}$ **do**
  Set $X_t^{(a_t^\star)} \leftarrow x_t^\star$
  Sample $a_{t-1}^{(a_t^\star)} \sim \text{Categorical}\left(\{1, \ldots, N\}, w_{t-1}^{(1:N)} q_t^\theta(x_t^\star \mid X_{t-1}^{(1:N)})\right)$
  Sample $a_{t-1}^{(1:N)} \setminus a_{t-1}^\star \sim \text{RESAMPLE}(\{1, \ldots, N\}, w_{t-1}^{(1:N)} \mid a_{t-1}^\star)$
  for $i \in \{1, \ldots, N\} \setminus a_t^\star$ **do** Sample $X_t^{(i)} \sim K_t^\theta(\cdot \mid X_{t-1}^{(a_{t-1}^{(i)})})$
  for $i \in \{1, \ldots, N\}$ **do** $w_t^{(i)} \leftarrow \left\{\sum_{j=1}^N g_t^\theta(y_t \mid X_t^{(j)})\right\}^{-1} g_t^\theta(y_t \mid X_t^{(i)})$
end

Algorithm 2.3: Conditional sequential Monte Carlo with ancestor sampling for a parametrised state space model. The parent of the immortal particle is updated at each iteration via an on-line backward simulation step. The transition kernels K_t^θ are assumed to admit densities q_t^θ. The second parameter of the Categorical variable should be interpreted element-wise, and is given up to a normalisation constant.

The ratio of densities can be interpreted as the conditional probability that the whole trajectory is the concatenation of $X_{t,0:t}^{(i)}$ with $x_{t+1:T}^\star$, given that its first $t+1$ states are $X_{t,0:t}^{(i)}$. To simplify the ratio, use (2.15) to write

$$p(X_{t,0:t}^{(i)} \mid \theta, y_{0:t}) \propto \mu^\theta(X_{t,0}^{(i)}) g_0^\theta(y_0 \mid X_{t,0}^{(i)}) \prod_{s=1}^t q_s^\theta(X_{t,s}^{(i)} \mid X_{t,s-1}^{(i)}) g_s^\theta(y_s \mid X_{t,s}^{(i)})$$

and

$$\begin{aligned} &p((X_{t,0:t}^{(i)}, x_{t+1:T}^\star) \mid \theta, y_{0:T}) \\ &\quad \propto \mu^\theta(X_{t,0}^{(i)}) g_0^\theta(y_0 \mid X_{t,0}^{(i)}) \left\{\prod_{s=1}^t q_s^\theta(X_{t,s}^{(i)} \mid X_{t,s-1}^{(i)}) g_s^\theta(y_s \mid X_{t,s}^{(i)})\right\} \\ &\qquad \times q_{t+1}^\theta(x_{t+1}^\star \mid X_{t,t}^{(i)}) g_{t+1}^\theta(y_{t+1} \mid x_{t+1}^\star) \left\{\prod_{s=t+2}^T q_s^\theta(x_s^\star \mid x_{s-1}^\star) g_s^\theta(y_s \mid x_s^\star)\right\}. \end{aligned}$$

The ratio then becomes

$$\begin{aligned} &\frac{p((X_{t,0:t}^{(i)}, x_{t+1:T}^\star) \mid \theta, y_{0:T})}{p(X_{t,0:t}^{(i)} \mid \theta, y_{0:t})} \\ &\qquad \propto q_{t+1}^\theta(x_{t+1}^\star \mid X_{t,t}^{(i)}) g_{t+1}^\theta(y_{t+1} \mid x_{t+1}^\star) \prod_{s=t+2}^T q_s^\theta(x_s^\star \mid x_{s-1}^\star) g_s^\theta(y_s \mid x_s^\star) \\ &\qquad \propto q_{t+1}^\theta(x_{t+1}^\star \mid X_{t,t}^{(i)}) = q_{t+1}^\theta(x_{t+1}^\star \mid X_t^{(i)}). \end{aligned}$$

The probabilities in (2.16) become

$$\mathbb{P}\left[a_t^{(j)} = i \,\middle|\, X_{0:t}^{(1:N)}, x_{0:T}^\star, a_{0:T}^\star\right] \propto \begin{cases} w_t^{(i)} & j \text{ non-immortal} \\ w_t^{(i)} q_{t+1}^\theta(x_{t+1}^\star \mid X_t^{(i)}) & j \text{ immortal.} \end{cases} \tag{2.17}$$

The conditional SMC algorithm with this adaptation is presented in Algorithm 2.3.

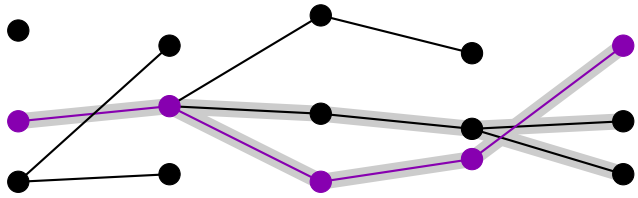
We see that, in order to do ancestor sampling, we need a stronger assumption on the Markov kernels than was required to simply run the conditional SMC algorithm: we now require that, for each t, K_t^θ admits a density q_t^θ and that $q_t^\theta(\cdot \mid x)$ can be evaluated pointwise for any x, whereas previously we only needed to draw samples from $K_t^\theta(\cdot \mid x)$ for any x. This additional requirement rules out ancestor sampling in some applications, for instance when the transitions are discretisations of some stochastic differential equation.

Recall that the usual backward simulation procedure requires a full forward pass to calculate the future states before the backward simulation probabilities can be computed. Ancestor sampling, on the other hand, does not require a forward pass because it only computes backward simulation probabilities for the immortal trajectory, for which all the future states are known in advance. This means that the additional computational cost of implementing ancestor sampling is negligible.

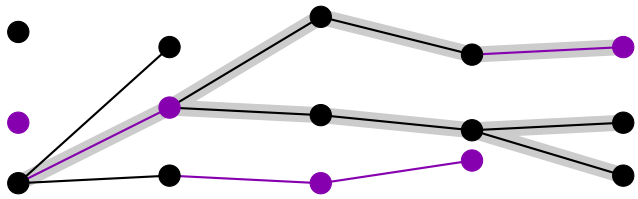
Why ancestor sampling works

We know that complete backward simulation eradicates ancestral degeneracy by sampling each lineage independently (Section 2.3.1). But here we are only backward-simulating one of the N particles, leaving the other $N-1$ lineages to coalesce as usual. So how does this help?

Recall that in particle Gibbs ancestral degeneracy is not itself a problem, because we only require a single sample from the smoothing distribution. The problem is that the consecutive samples are highly correlated, because of the repeated coalescence onto the immortal lineage. The contribution of ancestor sampling is to break up the immortal trajectory so that it no longer appears among the lineages; see Figure 2.13. While the non-immortal trajectories may still coalesce, they no longer preferentially coalesce onto the immortal trajectory. In turn, the sampled trajectory that is output does not overlap unduly with the immortal trajectory that was the previous output, and this completely solves the problem of the slow-mixing particle Gibbs chain.



(a) without ancestor sampling



(b) with ancestor sampling

Figure 2.13: Illustration of how ancestor sampling prevents coalescence onto the immortal trajectory. Immortal particles are highlighted in purple, along with their parent-offspring edges (the given ones in (a) and the ancestor-sampled ones in (b)). The resulting lineages of the terminal particles are highlighted in grey. In (a), the lineages of the terminal particles coalesce onto the immortal trajectory. Imagine time stretching further back: the lineages would continue to coincide with the immortal trajectory forever. In (b), the lineages still coalesce, but not onto the immortal trajectory. The immortal trajectory no longer exists as a lineage.

3 Convergence of Finite-Dimensional Distributions

To see a World in a Grain of Sand
And a Heaven in a Wild Flower,
Hold Infinity in the palm of your hand
And Eternity in an hour.

William Blake

In this chapter we derive conditions under which genealogies induced by SMC algorithms converge to the n-coalescent as the number of particles tends to infinity. Here we prove only the convergence of finite-dimensional distributions; weak convergence is proved under the same conditions in Chapter 4.

3.1 The genealogical process

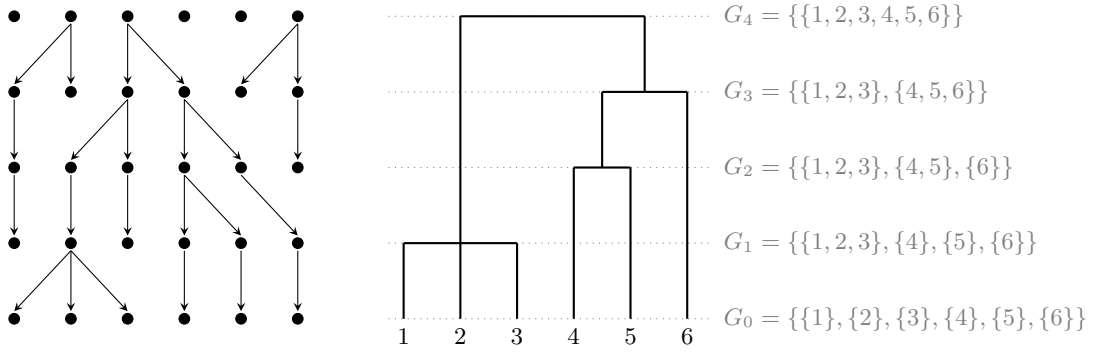
Before we can analyse genealogies, we need a way to encode them. The encoding will only include the information relevant to the sample genealogy, namely which lineages coalesce at which times. Information about particle positions and "killed" particles is ignored.

Let $\mathcal{P}_n$ be the space of partitions on $\{1, \ldots, n\}$. For convenience, we now label time in reverse, so the terminal particles are at time 0, their parents are at time 1, and so on. Consider a randomly chosen (uniformly, without replacement) sample of size n among the N terminal particles, and label the sampled particles $1, \ldots, n$. The *genealogical process* $(G_t^{(n,N)})_{t \in \mathbb{N}_0}$ for this sample is the $\mathcal{P}_n$-valued stochastic process such that labels i and j are in the same block of the partition $G_t^{(n,N)}$ if and only if terminal particles i and j have a common ancestor at time t (i.e. t generations back).

A formulation where $G_t^{(n,N)}$ takes values in the space of equivalence relations from $[n]$ to $[n]$ is sometimes used (e.g. Möhle 1999); interpreting partition blocks as equivalence classes, this formulation is equivalent to ours.

The initial value of the process is the partition of singletons $G_0^{(n,N)} = \{\{1\}, \ldots, \{n\}\}$, since all of the terminal particles are in separate lineages. The only possible non-identity transitions are those that merge some blocks of the partition, encoding the coalescence

of the corresponding lineages. The trivial partition $\{\{1, \ldots, n\}\}$ is therefore an absorbing state, corresponding to all lineages in the sample having coalesced, that is, the MRCA has been reached. The construction of the genealogical process from the resampling relationships, encoded by the vector of parental indices at each generation, is illustrated in Figure 3.1.



(a) Resampling relationships (b) Genealogy of terminal particles

Figure 3.1: Illustration of how the sample genealogy is encoded. (a) Relationships induced by resampling in a sample of $n = 6$ particles over four iterations. For clarity the pictured example has $n = N$, but in general $n \ll N$. (b) The genealogy of these six particles, labelled with the value of the genealogical process G_t at each time.

Under the assumption (A1) stated below, it is sufficient for our purposes to consider only offspring counts $\nu_t^{(1:N)} = (\nu_t^{(1)}, \ldots, \nu_t^{(N)})$, where $\nu_t^{(i)} = |\{j : a_t^{(j)} = i\}|$, rather than the parental indices $a_t^{(1:N)}$ which are generally more informative.

(A1) The conditional distribution of parental indices $a_t^{(1:N)}$ given offspring counts $\nu_t^{(1:N)}$ is uniform over all assignments such that $|\{j : a_t^{(j)} = i\}| = \nu_t^{(i)}$ for all i.

As we saw in Section 2.2, the n-coalescent is *exchangeable*, so for instance the pair of lineages merging at each event is chosen uniformly. Sometimes called the *random assignment condition*, (A1) is a weaker condition than exchangeability of the particles within a generation which is sufficient to admit an exchangeable process in the limit. Although the resampling in SMC is not generally exchangeable, (A1) can easily be enforced upon any SMC algorithm by applying a random permutation to the offspring indices immediately after resampling.

3.1.1 Time scale

In order to have a well-defined limit for the genealogical process as $N \to \infty$, we must scale time by a suitable function $\tau_N(\cdot)$. In the population genetics literature the time scale function is typically deterministic (Section 2.2.3), but in our case τ_N depends on the offspring counts and is therefore random. To define the time scale we first define the pair

merger rate

$$c_N(t) := \frac{1}{(N)_2} \sum_{i=1}^{N} (\nu_t^{(i)})_2. \tag{3.1}$$

This is the probability, conditional on $\nu_t^{(1:N)}$, that a randomly chosen pair of lineages in generation t merges exactly one generation back. The given expression for $c_N(t)$ is justified by assumption (A1), as are the expressions for $\tau_N(t)$ and $D_N(t)$ below. To achieve a limiting pair merger rate of 1, as in the n-coalescent, we rescale time by the generalised inverse

$$\tau_N(t) := \inf \left\{ s \in \mathbb{N} : \sum_{r=1}^{s} c_N(r) \geq t \right\}. \tag{3.2}$$

The function τ_N maps continuous to discrete time, providing the link between the discrete-time SMC dynamics and the continuous-time limit. We will also need the following quantity, which is an upper bound on the conditional probability of a multiple merger (three or more lineages merging, or two or more simultaneous pairwise mergers):

$$D_N(t) := \frac{1}{N(N)_2} \sum_{i=1}^{N} (\nu_t^{(i)})_2 \left\{ \nu_t^{(i)} + \frac{1}{N} \sum_{j \neq i} (\nu_t^{(j)})^2 \right\}. \tag{3.3}$$

This will be used to control the rate of multiple mergers, which must be dominated by the pair-merger rate as $N \to \infty$ if we are to recover a Kingman limit (in which almost surely the only non-identity transitions are pair mergers). Some basic properties of c_N, D_N and τ_N are stated in Proposition 3.1.

Proposition 3.1. *For all $t \in \mathbb{N}$, $t' > s' > 0$,*

(a) $c_N(t), D_N(t) \in [0, 1]$

(b) $D_N(t) \leq c_N(t)$

(c) $c_N(t)^2 \leq c_N(t)$

(d) $t' \leq \sum_{r=1}^{\tau_N(t')} c_N(r) \leq t' + 1.$

(e) $t' - s' - 1 \leq \sum_{r=\tau_N(s')+1}^{\tau_N(t')} c_N(r) \leq t' - s' + 1.$

(f) $\tau_N(t') \geq t'.$

Proof. **(a)** $c_N(t)$ and $D_N(t)$ are clearly non-negative. Both are maximised when one of the offspring counts is equal to N and the rest are zero, in which case $c_N(t) = D_N(t) = 1$.

(b) As outlined in Koskela et al. (2018, p.10),

$$\begin{aligned} D_N(t) &:= \frac{1}{(N)_2}\sum_{i=1}^{N}(\nu_t^{(i)})_2\frac{1}{N}\left\{\nu_t^{(i)} + \frac{1}{N}\sum_{j\neq i}^{N}(\nu_t^{(j)})^2\right\} \\ &\leq \frac{1}{(N)_2}\sum_{i=1}^{N}(\nu_t^{(i)})_2\frac{1}{N}\left\{\nu_t^{(i)} + \frac{1}{N}\sum_{j\neq i}^{N}N\nu_t^{(j)}\right\} \\ &= \frac{1}{(N)_2}\sum_{i=1}^{N}(\nu_t^{(i)})_2\frac{1}{N}\left\{\sum_{j=1}^{N}\nu_t^{(j)}\right\} = \frac{1}{(N)_2}\sum_{i=1}^{N}(\nu_t^{(i)})_2 = c_N(t). \end{aligned}$$

(c) is immediate given (a).

(d) follows directly from the definition of τ_N in (3.2).

(e) Writing

$$\sum_{r=\tau_N(s')+1}^{\tau_N(t')} c_N(r) = \sum_{r=1}^{\tau_N(t')} c_N(r) - \sum_{r=1}^{\tau_N(s')} c_N(r),$$

the result follows by applying (d) to both sums.

(f) follows from (a) and the definition of τ_N in (3.2). ■

Another useful property is the following, based on Koskela et al. (2018, Lemma 2). There the special case $f(\nu_r^{(1:N)}) \equiv c_N(r)$ is proved, but the authors remark that the result also holds for other choices of f. Here we state a more general version of the result.

Lemma 3.2. *Fix $t > 0$. Let $(\mathcal{F}_r)$ be the backwards-in-time filtration generated by the offspring counts $\nu_r^{(1:N)}$ at each generation r. Let $f : [N]^N \mapsto \mathbb{R}$ be any deterministic function such that for all $\nu^{(1:N)}$ there exists $B < \infty$ for which $0 \leq f(\nu^{(1:N)}) \leq B$. Then*

$$\mathbb{E}\left[\sum_{r=1}^{\tau_N(t)} f(\nu_r^{(1:N)})\right] = \mathbb{E}\left[\sum_{r=1}^{\tau_N(t)} \mathbb{E}[f(\nu_r^{(1:N)}) \mid \mathcal{F}_{r-1}]\right].$$

The lemma holds more generally for any bounded function f (that does not need to be non-negative) by decomposing into the positive and negative parts. However, the simplified statement here is sufficient for our purposes since we will only apply the lemma to non-negative functions.

Proof. Define

$$M_s := \sum_{r=1}^{s}\left\{f(\nu_r^{(1:N)}) - \mathbb{E}[f(\nu_r^{(1:N)}) \mid \mathcal{F}_{r-1}]\right\}.$$

It is easy to establish that (M_s) is a martingale with respect to $(\mathcal{F}_s)$, and $M_0 = 0$. Now fix $K \geq 1$ and note that $\tau_N(t) \wedge K$ is a bounded $\mathcal{F}_t$-stopping time. Hence we can apply

the optional stopping theorem:

$$\begin{aligned}\mathbb{E}[M_{\tau_N(t)\wedge K}] &= \mathbb{E}\left[\sum_{r=1}^{\tau_N(t)\wedge K}\left\{f(\nu_r^{(1:N)}) - \mathbb{E}[f(\nu_r^{(1:N)}) \mid \mathcal{F}_{r-1}]\right\}\right] \\ &= \mathbb{E}\left[\sum_{r=1}^{\tau_N(t)\wedge K} f(\nu_r^{(1:N)})\right] - \mathbb{E}\left[\sum_{r=1}^{\tau_N(t)\wedge K} \mathbb{E}[f(\nu_r^{(1:N)}) \mid \mathcal{F}_{r-1}]\right] = 0.\end{aligned}$$

Since this holds for all $K \geq 1$,

$$\lim_{K\to\infty} \mathbb{E}\left[\sum_{r=1}^{\tau_N(t)\wedge K} f(\nu_r^{(1:N)})\right] = \lim_{K\to\infty} \mathbb{E}\left[\sum_{r=1}^{\tau_N(t)\wedge K} \mathbb{E}[f(\nu_r^{(1:N)}) \mid \mathcal{F}_{r-1}]\right].$$

The monotone convergence theorem allows the limit to pass inside the expectation on each side (since increasing K can only increase each sum, by possibly adding some non-negative terms). Hence

$$\begin{aligned}\mathbb{E}\left[\sum_{r=1}^{\tau_N(t)} f(\nu_r^{(1:N)})\right] &= \mathbb{E}\left[\lim_{K\to\infty}\sum_{r=1}^{\tau_N(t)\wedge K} f(\nu_r^{(1:N)})\right] = \mathbb{E}\left[\lim_{K\to\infty}\sum_{r=1}^{\tau_N(t)\wedge K} \mathbb{E}[f(\nu_r^{(1:N)}) \mid \mathcal{F}_{r-1}]\right] \\ &= \mathbb{E}\left[\sum_{r=1}^{\tau_N(t)} \mathbb{E}[f(\nu_r^{(1:N)}) \mid \mathcal{F}_{r-1}]\right],\end{aligned}$$

which concludes the proof. ■

3.1.2 Transition probabilities

Recall that $\mathcal{P}_n$ denotes the space of partitions of $\{1,\dots,n\}$. For any $\xi,\eta \in \mathcal{P}_n$ and $t \in \mathbb{N}$, let $p_{\xi\eta}(t)$ denote the conditional transition probabilities of the genealogical process given $\nu_t^{(1:N)}$. The transition probability $p_{\xi\eta}(t)$ can only be non-zero when η is obtained from ξ by merging some blocks of ξ (i.e. some lineages coalescing). Ordering the blocks by their least element, denote by b_i the number of blocks of ξ that merge to form block i in η, for each $i \in \{1,\dots,|\eta|\}$. Hence $b_1 + \cdots + b_{|\eta|} = |\xi|$. Then the transition probability is given by

$$p_{\xi\eta}(t) := \frac{1}{(N)_{|\xi|}} \sum_{\substack{i_1,\dots,i_{|\eta|}=1 \\ \text{all distinct}}}^{N} (\nu_t^{(i_1)})_{b_1} \cdots (\nu_t^{(i_{|\eta|})})_{b_{|\eta|}}. \tag{3.4}$$

This expression is again justified by (A1). We will only need to work directly with the identity transition probabilities $p_{\xi\xi}(t)$. Upper and lower bounds on these probabilities are presented in Propositions 3.3 and 3.4.

Proposition 3.3. *Let $\xi \in \mathcal{P}_n$, $N > 2$. Then*

$$p_{\xi\xi}(t) \geq 1 - \binom{|\xi|}{2} \frac{N^{|\xi|-2}}{(N-2)_{|\xi|-2}} \left[c_N(t) + B_{|\xi|} D_N(t)\right]$$

where

$$B_{|\xi|} = K(|\xi|-1)!(|\xi|-2)\exp(2\sqrt{2(|\xi|-2)})$$

for some $K > 0$ that does not depend on $|\xi|$.

Proof. We have the following expression for $p_{\xi\xi}(t)$, by subtracting all possible non-identity transitions. The sum over k counts all transitions from ξ to η such that $k = |\eta| \leq |\xi| - 1$; the omitted $k = |\xi|$ term would count identity transitions.

$$p_{\xi\xi}(t) = 1 - \frac{1}{(N)_{|\xi|}} \sum_{k=1}^{|\xi|-1} \sum_{\substack{b_1 \geq \ldots \geq b_k \geq 1 \\ b_1 + \ldots + b_k = |\xi|}} \frac{|\xi|!}{\prod_{j=1}^{|\xi|} (j!)^{\kappa_j} \kappa_j!} \sum_{\substack{i_1,\ldots,i_k = 1 \\ \text{all distinct}}}^{N} (\nu_t^{(i_1)})_{b_1} \ldots (\nu_t^{(i_k)})_{b_k},$$

where $\kappa_i = |\{j : b_j = i\}|$ is the multiplicity of mergers of size i (κ_1 counts non-merger events, and we have the identity $\kappa_1 + 2\kappa_2 + \cdots + |\xi|\kappa_{|\xi|} = |\xi|$). The combinatorial factor is the number of partitions of a sequence of length $|\xi|$ having κ_j subsequences of length j for each j (Fu 2006, Equation (11)).

We separate the $k = |\xi| - 1$ term (which counts single pair mergers), for which $(b_1, b_2, \ldots, b_{|\xi|-1}) = (2, 1, \ldots, 1)$ and

$$\frac{|\xi|!}{\prod_{j=1}^{|\xi|} (j!)^{\kappa_j} \kappa_j!} = \binom{|\xi|}{2}.$$

For the remaining terms we use

$$\frac{|\xi|!}{\prod_{j=1}^{|\xi|} (j!)^{\kappa_j} \kappa_j!} \leq |\xi|!.$$

Thus

$$\begin{aligned} p_{\xi\xi}(t) \geq 1 &- \frac{1}{(N)_{|\xi|}} \binom{|\xi|}{2} \sum_{\substack{i_1,\ldots,i_{|\xi|-1} = 1 \\ \text{all distinct}}}^{N} (\nu_t^{(i_1)}) {}_2 \nu_t^{(i_2)} \ldots \nu_t^{(i_{|\xi|-1})} \\ &- \frac{1}{(N)_{|\xi|}} \sum_{k=1}^{|\xi|-2} \sum_{\substack{b_1 \geq \ldots \geq b_k \geq 1 \\ b_1 + \ldots + b_k = |\xi|}} |\xi|! \sum_{\substack{i_1,\ldots,i_k = 1 \\ \text{all distinct}}}^{N} (\nu_t^{(i_1)})_{b_1} \ldots (\nu_t^{(i_k)})_{b_k}. \end{aligned} \tag{3.5}$$

Now, for the $k = |\xi| - 1$ term we use the bound

$$\sum_{\substack{i_1,\dots,i_{|\xi|-1}=1 \\ \text{all distinct}}}^{N} (\nu_t^{(i_1)})_2 \nu_t^{(i_2)} \dots \nu_t^{(i_{|\xi|-1})} \leq N^{|\xi|-2} \sum_{i=1}^{N} (\nu_t^{(i)})_2$$

while for the other terms we have

$$\begin{aligned}
&\sum_{\substack{i_1,\dots,i_k=1 \\ \text{all distinct}}}^{N} (\nu_t^{(i_1)})_{b_1} \dots (\nu_t^{(i_k)})_{b_k} \\
&\qquad \leq \sum_{i=1}^{N} (\nu_t^{(i)})_2 \left\{ \sum_{j_1,\dots,j_{|\xi|-2}=1}^{N} \nu_t^{(j_1)} \dots \nu_t^{(j_{|\xi|-2})} - \sum_{\substack{j_1,\dots,j_{|\xi|-2}=1 \\ \text{all distinct and } \neq i}}^{N} \nu_t^{(j_1)} \dots \nu_t^{(j_{|\xi|-2})} \right\} \\
&\qquad = \sum_{i=1}^{N} (\nu_t^{(i)})_2 \left\{ \left(\sum_{j=1}^{N} \nu_t^{(j)} \right)^{|\xi|-2} - \sum_{\substack{j_1,\dots,j_{|\xi|-2}=1 \\ \text{all distinct and } \neq i}}^{N} \nu_t^{(j_1)} \dots \nu_t^{(j_{|\xi|-2})} \right\} \\
&\qquad = \sum_{i=1}^{N} (\nu_t^{(i)})_2 \left\{ N^{|\xi|-2} - \sum_{\substack{j_1,\dots,j_{|\xi|-2}=1 \\ \text{all distinct and } \neq i}}^{N} \nu_t^{(j_1)} \dots \nu_t^{(j_{|\xi|-2})} \right\}
\end{aligned}$$

where we have subtracted all the terms except those which either have one of the indices equal to i (in which case the largest merger consists of more than two lineages) or have two of the indices equal to each other (in which case there is a simultaneous merger). This leaves only those terms where $k \leq |\xi| - 2$, that is where there are multiple or simultaneous mergers. It is an inequality rather than an equality because some of the cases are double-counted.

The expression is further bounded by a reverse multinomial expansion, as in the proof of Koskela et al. (2018, Lemma 1 Case 3), which is then simplified using that $(N - x)^b \geq N^b - bxN^{b-1}$ for $x \leq N$, $b \geq 0$, an application of the Bernoulli inequality:

$$\begin{aligned}
&\sum_{i=1}^{N} (\nu_t^{(i)})_2 \left\{ N^{|\xi|-2} - (N - \nu_t^{(i)})^{|\xi|-2} + \binom{|\xi|-2}{2} \sum_{j \neq i} (\nu_t^{(j)})^2 \left(\sum_{k \neq i} \nu_t^{(k)} \right)^{|\xi|-4} \right\} \\
&\leq \sum_{i=1}^{N} (\nu_t^{(i)})_2 \left\{ (|\xi| - 2) \nu_t^{(i)} N^{|\xi|-3} + \binom{|\xi|-2}{2} \sum_{j \neq i} (\nu_t^{(j)})^2 N^{|\xi|-4} \right\}.
\end{aligned}$$

Hence

$$\begin{aligned}
p_{\xi\xi}(t) \geq 1 &- \frac{1}{(N)_{|\xi|}} \binom{|\xi|}{2} N^{|\xi|-2} \sum_{i=1}^{N} (\nu_t^{(i)})_2 \\
&- \frac{N^{|\xi|-3}}{(N)_{|\xi|}} |\xi|! \sum_{k=1}^{|\xi|-2} \sum_{\substack{b_1 \geq \dots \geq b_k \geq 1 \\ b_1 + \dots + b_k = |\xi|}} \sum_{i=1}^{N} (\nu_t^{(i)})_2 \left\{ (|\xi| - 2) \nu_t^{(i)} + \binom{|\xi|-2}{2} \frac{1}{N} \sum_{j \neq i} (\nu_t^{(j)})^2 \right\}.
\end{aligned}$$

The summands in the last line are independent of $k, b_1, \ldots, b_k$, and the number of terms in the sums over k and $b_1, \ldots, b_k$ is bounded by $\gamma_{|\xi|-2}(|\xi|-2)$, where γ_n is the number of integer partitions of n. By Hardy and Ramanujan (1918, Section 2), $\gamma_n < Ke^{2\sqrt{2n}}/n$ for a constant $K > 0$ independent of n. Thus, for $|\xi| > 2$,

$$\begin{aligned} p_{\xi\xi}(t) \geq 1 &- \frac{N^{|\xi|-2}}{(N-2)_{|\xi|-2}} \binom{|\xi|}{2} c_N(t) \\ &- \frac{N^{|\xi|-2}}{(N-2)_{|\xi|-2}} K \exp(2\sqrt{2(|\xi|-2)})|\xi|! \frac{1}{N(N)_2} \\ &\times \sum_{i=1}^{N} (\nu_t^{(i)})_2 \left\{ (|\xi|-2)\nu_t^{(i)} + \binom{|\xi|-2}{2} \frac{1}{N} \sum_{j \neq i} (\nu_t^{(j)})^2 \right\}. \end{aligned}$$

Notice that when $|\xi| > 2$, both $(|\xi|-2)$ and $\binom{|\xi|-2}{2}$ are less than or equal to $\binom{|\xi|-1}{2}$. Thus by definition of $D_N(t)$,

$$\begin{aligned} p_{\xi\xi}(t) &\geq 1 - \frac{N^{|\xi|-2}}{(N-2)_{|\xi|-2}} \binom{|\xi|}{2} c_N(t) \\ &\quad - \frac{N^{|\xi|-2}}{(N-2)_{|\xi|-2}} K \exp(2\sqrt{2(|\xi|-2)})|\xi|! \binom{|\xi|-1}{2} D_N(t) \\ &\geq 1 - \frac{N^{|\xi|-2}}{(N-2)_{|\xi|-2}} \binom{|\xi|}{2} \left[c_N(t) + B_{|\xi|} D_N(t) \right] \end{aligned}$$

where

$$\begin{aligned} B_{|\xi|} &= \binom{|\xi|}{2}^{-1} K \exp(2\sqrt{2(|\xi|-2)})|\xi|! \binom{|\xi|-1}{2} \\ &= K(|\xi|-1)!(|\xi|-2)\exp(2\sqrt{2(|\xi|-2)}). \end{aligned}$$

When $|\xi| = 2$, (3.5) becomes

$$p_{\xi\xi}(t) \geq 1 - c_N(t)$$

and when $|\xi| = 1$, (3.5) becomes

$$p_{\xi\xi}(t) \geq 1;$$

in both cases the result is immediate. ■

Proposition 3.4. *Let $\xi \in \mathcal{P}_n$. Then, for N sufficiently large,*

$$p_{\xi\xi}(t) \leq 1 - \binom{|\xi|}{2} \{1 + O(N^{-1})\} \left[c_N(t) - B'_{|\xi|} D_N(t) \right]$$

where $B'_{|\xi|} = \binom{|\xi|-1}{2}$.

A proof is given in Koskela et al. (2018, Lemma 1 Case 1).

3.2 An existing limit theorem

Koskela et al. (2018) proved the following theorem which gives sufficient conditions under which sampled genealogies of non-neutral interacting particle systems converge to the n-coalescent as $N \to \infty$. Such a result can only be expected to hold for genealogies of finite samples ($n \ll N$), and not for the entire genealogy of the N particles. For instance the genealogies arising in SMC algorithms are not restricted to single pair mergers only, although within a sparse sample we may, under mild conditions, see only single pair mergers. That is to say, there is not an extension of this result whereby the whole-population genealogy converges to the Kingman coalescent as $N \to \infty$, unless very restrictive conditions are imposed.

Theorem 3.5 (Koskela et al. 2018)**.** *Fix $n \leq N$ as the observed number of particles from the output of an interacting particle system with N particles which satisfies* (A1)*. Suppose that for any $0 \leq s < t < \infty$, we have*

$$\lim_{N\to\infty} \mathbb{E}\left[\sum_{r=\tau_N(s)+1}^{\tau_N(t)} D_N(r)\right] = 0, \tag{3.6}$$

$$\lim_{N\to\infty} \mathbb{E}[c_N(t)] = 0, \tag{3.7}$$

$$\lim_{N\to\infty} \mathbb{E}\left[\sum_{r=\tau_N(s)+1}^{\tau_N(t)} c_N(r)^2\right] = 0, \tag{3.8}$$

$$\textit{and} \qquad \mathbb{E}[\tau_N(t) - \tau_N(s)] \leq C_{t,s}N, \tag{3.9}$$

for some constant $C_{t,s} > 0$ that is independent of N. Then the finite-dimensional distributions of $(G^{(n,N)}_{\tau_N(t)})_{t\geq 0}$ converge to those of the n-coalescent as $N \to \infty$.

To ensure samples of size n have Kingman genealogies in the limit, with pair mergers only, we require that multiple mergers (that is, where more than two lineages merge into one, or where two or more mergers happen simultaneously) occur on a slower time scale than pair mergers. This is the role of condition (3.6).

Conditions (3.7) and (3.8) ensure that the limiting process is continuous and has the required unit pair merger rate. For (3.7) to fail to hold, the expected number of mergers at some generation would have to be $\Omega(N^2)$. This can only happen if the resampling scheme is very bad (e.g. star resampling) or the weights are particularly badly-behaved. The latter is ruled out in the corollaries of Chapter 5 by imposing bounds on the potential functions; this is discussed further in Section 5.1.

Condition (3.9) specifies that the time scale must be $O(N)$. As we saw in Section 2.2.3, this is the correct time scale for the Wright-Fisher model, but for instance the Moran model has time scale $O(N^2)$ and hence violates this condition. Since we know that the neutral Moran model also has Kingman genealogies in the limit, condition (3.9) cannot be

necessary. The simplified statement in Theorem 3.6 does not impose any such condition on the time scale.

The proof of Koskela et al. (2018) does not explicitly use (3.7) but rather the similar condition

$$\lim_{N\to\infty} \mathbb{E}[c_N(\tau_N(t))] = 0. \tag{3.10}$$

However, as we will see in the next section (Lemmata 3.8 and 3.9), both (3.7) and (3.10) are implied by (3.6), so the theorem is correct. Such redundancies in the statement of Theorem 3.5 are removed in Theorem 3.6.

The proof of Theorem 3.5 (i.e. Koskela et al. 2018, Theorem 1) proceeds in three parts. The first is a vanishing upper bound on finite-dimensional distributions of the genealogical process when the path of the process involves any multiple mergers. The second is showing that, when the path of the genealogy consists of only single pair mergers, the finite-dimensional distributions of the n-coalescent upper bound those of the genealogical process in the limit $N \to \infty$. The final piece is a similar lower bound, which together with the upper bound establishes convergence of the finite-dimensional distributions.

3.3 A new limit theorem

We now present a related theorem, having the same conclusion but with conditions that are more tractable and remove some redundancies in the statement of Theorem 3.5.

Theorem 3.6. *Let $\nu_t^{(1:N)}$ denote the offspring numbers in an interacting particle system satisfying* (A1) *such that, for any N sufficiently large, $\mathbb{P}[\tau_N(t) = \infty] = 0$ for all finite t. Suppose that there exists a deterministic sequence $(b_N)_{N\geq 1}$ such that $\lim_{N\to\infty} b_N = 0$ and*

$$\frac{1}{(N)_3} \sum_{i=1}^{N} \mathbb{E}_t \left[(\nu_t^{(i)})_3 \right] \leq b_N \frac{1}{(N)_2} \sum_{i=1}^{N} \mathbb{E}_t \left[(\nu_t^{(i)})_2 \right] \tag{3.11}$$

for all N, uniformly in $t \in \mathbb{N}$. Fix $n \leq N$ and consider a randomly chosen sample of n terminal particles. Then the finite-dimensional distributions of the resulting rescaled genealogical process $(G^{(n,N)}_{\tau_N(t)})_{t\geq 0}$ converge to those of the n-coalescent as $N \to \infty$.

On the right-hand side of (3.11) is the filtered expectation of $c_N(t)$, i.e. the expected pair merger rate, and the left-hand side is the corresponding rate of triple mergers. Intuitively, (3.11) says that pair mergers dominate triple mergers, the expected rate of which vanishes as $N \to \infty$. As we will see, this implies that pair mergers also dominate all other larger mergers, such as simultaneous pair mergers. The condition (3.11) is a non-exchangeable, non-neutral adaptation of the well-known necessary and sufficient condition for genealogies

of Cannings models to converge to the n-coalescent, namely

$$\lim_{N\to\infty} \frac{\mathbb{E}[(\nu^{(1)})_3]}{N\mathbb{E}[(\nu^{(1)})_2]} = 0,$$

found for example in Möhle (2000, Equation (16)). Since Cannings models assume exchangeability, the expectations are expressed, without loss of generality, for individual 1, and since offspring distributions are i.i.d. across generations there is no dependence on t and no conditioning.

Our result improves on Theorem 3.5 by eliminating the restrictive condition (3.9), which we know is unnecessary. This allows our result to apply to some models not previously included; for example the neutral Moran model. Although we do not prove that Theorem 3.6 is a true generalisation of Theorem 3.5, we know that in exchangeable neutral models the analogue of (3.11) is both necessary and sufficient, suggesting that in general this condition is not significantly stronger than (3.6)–(3.8) combined.

Our conditions are also significantly easier to verify than those of Theorem 3.5. Not only are four conditions replaced with one, but the condition (3.11) only involves marginal moments of the offspring counts, whereas (3.6) and (3.8) involve mixed moments. As we will see in Chapter 4, once we move beyond conditionally independent resampling schemes such as multinomial resampling, the joint distributions of offspring counts become complex and it may only be feasible to calculate their moments marginally. As such, we are able to verify the conditions of Theorem 3.6 in several cases, including for resampling schemes that induce strong correlations between offspring counts, whereas Koskela et al. (2018) apply their theorem only to multinomial resampling.

Our condition on the time scale, $\mathbb{P}[\tau_N(t) = \infty] = 0$, is not very restrictive. Essentially, it rules out systems in which coalescences occur at only finitely many generations. This condition is not actually necessary for Theorem 3.6 to hold, as such, but if it is violated then the limiting object is an n-coalescent under an infinite time-scaling, so that after some finite time the process is frozen forever and there are no more coalescences. This would constitute a qualitatively different result and one that is of little interest for SMC, so we follow Möhle (1998) and others in excluding it.

3.3.1 Proof of theorem

First we prove that (3.10) and the assumptions (3.6)–(3.8) of Theorem 3.5 all follow from (3.11). Figure 3.2 illustrates how the following Lemmata 3.7–3.10 fit together. The argument differs slightly from that presented in Brown et al. (2021) in that we will here show (3.11) $\Rightarrow$ (3.6) $\Rightarrow$ (3.7) rather than (3.11) $\Rightarrow$ (3.6) and (3.11) $\Rightarrow$ (3.7). This highlights the redundancy in Theorem 3.5, where condition (3.6) directly implies two of the other stated conditions.

The second step in the proof is to show that condition (3.9) is not necessary. In particular, the parts of the proof of Koskela et al. (2018) which relied on (3.9) are rewritten

using Proposition 3.3 instead. Proposition 3.3 is a lower bound on the probability of an identity transition, which holds in general without the need for further conditions, so we really are removing condition (3.9) and not replacing it with a different condition.

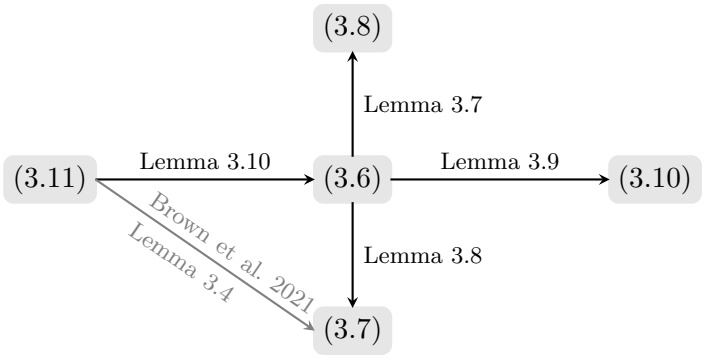


Figure 3.2: Dependencies between conditions of Theorems 3.5 and 3.6. Arrows represent logical implication; labels on arrows indicate the lemma in which the implication is stated. In Brown et al. (2021, Lemma 3.4) the direct implication (3.11) ⇒ (3.7) was proved, but here we will instead show that (3.6) ⇒ (3.7).

Lemma 3.7. *If for all* $0 \leq s < t < \infty$

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

then for all $0 \leq s < t < \infty$

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

Proof. We have

$$\begin{aligned}
c_N(t)^2 &= \frac{1}{N(N-1)(N)_2}\sum_{i=1}^{N}(\nu_t^{(i)})_2\left\{\nu_t^{(i)}(\nu_t^{(i)}-1)+\sum_{\substack{j=1\\j\neq i}}^{N}(\nu_t^{(j)})_2\right\}\\
&= \frac{1}{N(N)_2}\sum_{i=1}^{N}(\nu_t^{(i)})_2\left\{\frac{\nu_t^{(i)}(\nu_t^{(i)}-1)}{N-1}+\frac{1}{N-1}\sum_{\substack{j=1\\j\neq i}}^{N}(\nu_t^{(j)})_2\right\}\\
&\leq \frac{1}{N(N)_2}\sum_{i=1}^{N}(\nu_t^{(i)})_2\left\{\nu_t^{(i)}+\frac{1}{N-1}\sum_{\substack{j=1\\j\neq i}}^{N}(\nu_t^{(j)})_2\right\}\\
&\leq \frac{1}{N(N)_2}\sum_{i=1}^{N}(\nu_t^{(i)})_2\left\{\nu_t^{(i)}+\frac{N/(N-1)}{N}\sum_{\substack{j=1\\j\neq i}}^{N}(\nu_t^{(j)})^2\right\}\\
&\leq \frac{N/(N-1)}{N(N)_2}\sum_{i=1}^{N}(\nu_t^{(i)})_2\left\{\nu_t^{(i)}+\frac{1}{N}\sum_{\substack{j=1\\j\neq i}}^{N}(\nu_t^{(j)})^2\right\} = \frac{N}{N-1}D_N(t)
\end{aligned}$$

which is sufficient for the result. ■

Lemma 3.8. *If for all* $0 \leq s < t < \infty$

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

then for all $t \in \mathbb{N}$

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

Proof. Fix $\epsilon > 0$, and assume $N > 2/\epsilon$. Following Möhle and Sagitov (2003), define the events

$$A_i(t) := \{\nu_t^{(i)} \leq N\epsilon\}. \tag{3.12}$$

Then

$$\begin{aligned}
c_N(t) &= \frac{1}{(N)_2}\sum_{i=1}^{N}(\nu_t^{(i)})_2[\mathbb{1}_{A_i(t)}+\mathbb{1}_{A_i(t)^c}]\\
&\leq \frac{N\epsilon}{(N)_2}\sum_{i=1}^{N}\nu_t^{(i)}+\sum_{i=1}^{N}\mathbb{1}_{A_i(t)^c}\\
&= \frac{N\epsilon}{N-1}+\sum_{i=1}^{N}\mathbb{1}_{A_i(t)^c}.
\end{aligned}$$

Taking expectations and applying the generalised Markov inequality,

$$\begin{aligned}
\mathbb{E}[c_N(t)] &\leq \epsilon 1_N + \sum_{i=1}^{N} \mathbb{P}[\nu_t^{(i)} > N\epsilon] \\
&\leq \epsilon 1_N + \sum_{i=1}^{N} \frac{\mathbb{E}[(\nu_t^{(i)})_3]}{(N\epsilon)_3} \\
&\leq \epsilon 1_N + \frac{N(N)_2}{(N\epsilon)_3} \mathbb{E}[D_N(t)] \\
&= \epsilon 1_N + \epsilon^{-3} 1_N \mathbb{E}[D_N(t)] \\
&\leq \epsilon 1_N + \epsilon^{-3} 1_N \mathbb{E}\left[\sum_{r=1}^{t} D_N(r)\right] \\
&\leq \epsilon 1_N + \epsilon^{-3} 1_N \mathbb{E}\left[\sum_{r=\tau_N(0)+1}^{\tau_N(t)} D_N(r)\right],
\end{aligned}$$

where 1_N is used as an asymptotic shorthand for a sequence that converges to 1 as $N \to \infty$; for instance 1_N can be thought of as $1 + O(N^{-1})$. The last inequality is a consequence of $\tau_N(0) = 0$ and $\tau_N(t) \geq t$ (Proposition 3.1(f)). Taking limits,

$$\lim_{N\to\infty} \mathbb{E}[c_N(t)] \leq \epsilon.$$

Since ϵ was arbitrary this concludes the proof. ■

Lemma 3.9. *If for all* $0 \leq s < t < \infty$

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

then for all $0 < t < \infty$

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

Proof. Analogously to the proof of Lemma 3.8, we find

$$\begin{aligned}
\mathbb{E}[c_N(\tau_N(t))] &\leq \epsilon 1_N + \sum_{i=1}^{N} \mathbb{P}[\nu_{\tau_N(t)}^{(i)} > N\epsilon] \\
&\leq \epsilon 1_N + \epsilon^{-3} 1_N \mathbb{E}\left[D_N(\tau_N(t))\right] \\
&\leq \epsilon 1_N + \epsilon^{-3} 1_N \mathbb{E}\left[\sum_{r=\tau_N(0)+1}^{\tau_N(t)} D_N(r)\right] \\
&\underset{N\to\infty}{\longrightarrow} \epsilon
\end{aligned}$$

which concludes the proof since ϵ was arbitrary. ■

Lemma 3.10. *If there exists a deterministic sequence* $(b_N)_{N\geq 1}$ *such that* $\lim_{N\to\infty} b_N = 0$ *and*

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

for all N, *uniformly in* $t \in \mathbb{N}$, *then for all* $0 \leq s < t < \infty$

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

Proof. We decompose $D_N(t)$ as the sum of two terms and consider their filtered expectations. The first is

$$\begin{aligned}\frac{1}{N(N)_2}\sum_{i=1}^{N}\mathbb{E}_t[(\nu_t^{(i)})_2\nu_t^{(i)}] &= \frac{1}{N(N)_2}\sum_{i=1}^{N}\mathbb{E}_t[2(\nu_t^{(i)})_2 + (\nu_t^{(i)})_3] \\ &\leq \frac{2}{N}\mathbb{E}_t[c_N(t)] + \frac{1}{(N)_3}\sum_{i=1}^{N}\mathbb{E}_t[(\nu_t^{(i)})_3] \\ &\leq \left(\frac{2}{N} + b_N\right)\mathbb{E}_t[c_N(t)]. \end{aligned} \tag{3.13}$$

The second is

$$\begin{aligned}\frac{1}{N^2(N)_2}\sum_{j=1}^{N}\sum_{i\neq j}\mathbb{E}_t[(\nu_t^{(i)})_2(\nu_t^{(j)})^2] &= \frac{1}{N^2(N)_2}\sum_{j=1}^{N}\sum_{i\neq j}\mathbb{E}_t[(\nu_t^{(i)})_2(\nu_t^{(j)})_2 + (\nu_t^{(i)})_2\nu_t^{(j)}] \\ &\leq \frac{1}{N^2(N)_2}\sum_{j=1}^{N}\sum_{i\neq j}\mathbb{E}_t[(\nu_t^{(i)})_2(\nu_t^{(j)})_2] + \frac{\mathbb{E}_t[c_N(t)]}{N}. \end{aligned} \tag{3.14}$$

Now, with the events $A_i(t)$ defined as in (3.12),

$$\begin{aligned}\sum_{j=1}^{N}\sum_{i\neq j}\mathbb{E}_t\{(\nu_t^{(i)})_2(\nu_t^{(j)})_2\} &= \sum_{j=1}^{N}\sum_{i\neq j}\left\{\mathbb{E}_t[(\nu_t^{(i)})_2(\nu_t^{(j)})_2\mathbb{1}_{A_i(t)}] + \mathbb{E}_t[(\nu_t^{(i)})_2(\nu_t^{(j)})_2\mathbb{1}_{A_i(t)^c}]\right\} \\ &\leq N\epsilon\sum_{j=1}^{N}\sum_{i\neq j}\mathbb{E}_t[\nu_t^{(i)}(\nu_t^{(j)})_2\mathbb{1}_{A_i(t)}] + N^3\sum_{j=1}^{N}\sum_{i\neq j}\mathbb{E}_t[\nu_t^{(j)}\mathbb{1}_{A_i(t)^c}] \\ &\leq N^2(N)_2\epsilon\mathbb{E}_t[c_N(t)] + N^4\sum_{i=1}^{N}\mathbb{P}[\nu_t^{(i)} > N\epsilon \mid \mathcal{F}_{t-1}]. \end{aligned} \tag{3.15}$$

For $N \geq 3/\epsilon$, by the generalised Markov inequality,

$$\sum_{i=1}^{N} \mathbb{P}[\nu_t^{(i)} > N\epsilon \mid \mathcal{F}_{t-1}] \leq \frac{1}{(N\epsilon)_3} \sum_{i=1}^{N} \mathbb{E}_t[(\nu_t^{(i)})_3] = \frac{1_N}{\epsilon^3 (N)_3} \sum_{i=1}^{N} \mathbb{E}_t[(\nu_t^{(i)})_3]$$
$$\leq 1_N \frac{b_N}{\epsilon^3} \mathbb{E}_t[c_N(t)]. \tag{3.16}$$

Substituting (3.16) into (3.15) gives

$$\sum_{j=1}^{N} \sum_{i \neq j} \mathbb{E}_t[(\nu_t^{(i)})_2 (\nu_t^{(j)})_2] \leq N^4 1_N \left(\epsilon + \frac{b_N}{\epsilon^3} \right) \mathbb{E}_t[c_N(t)] \tag{3.17}$$

and substituting (3.17) into (3.14) gives

$$\frac{1}{N^2 (N)_2} \sum_{j=1}^{N} \sum_{i \neq j} \mathbb{E}_t[(\nu_t^{(i)})_2 (\nu_t^{(j)})^2] \leq \left[1_N \left(\epsilon + \frac{b_N}{\epsilon^3} \right) + \frac{1}{N} \right] \mathbb{E}_t[c_N(t)]. \tag{3.18}$$

Combining (3.13) and (3.18), we have that

$$\mathbb{E}_t[D_N(t)] = \left[1_N \left(\epsilon + \frac{b_N}{\epsilon^3} \right) + \frac{3}{N} + b_N \right] \mathbb{E}_t[c_N(t)].$$

Finally, invoking Lemma 3.2 twice gives

$$\mathbb{E}\left[\sum_{r=\tau_N(s)+1}^{\tau_N(t)} D_N(r) \right] = \mathbb{E}\left[\sum_{r=\tau_N(s)+1}^{\tau_N(t)} \mathbb{E}_r[D_N(r)] \right]$$
$$\leq \left\{ 1_N \left(\epsilon + \frac{b_N}{\epsilon^3} \right) + \frac{3}{N} + b_N \right\} \mathbb{E}\left[\sum_{r=\tau_N(s)+1}^{\tau_N(t)} c_N(r) \right]$$
$$\leq \left\{ 1_N \left(\epsilon + \frac{b_N}{\epsilon^3} \right) + \frac{3}{N} + b_N \right\} (t - s + 1)$$
$$\underset{N \to \infty}{\longrightarrow} \epsilon(t - s + 1),$$

and recalling that $\epsilon > 0$ was arbitrary concludes the proof. ■

To complete the proof of Theorem 3.6 it remains to show that condition (3.9) is unnecessary. We will show that Proposition 3.3 can be used instead of (3.9) to obtain the same result. The only part of Koskela et al. (2018, Proof of Theorem 1) making use of condition (3.9) is the lower bound on finite-dimensional distributions of the genealogical process for paths involving single pair mergers only (starting on p.15 therein). A slight modification of the argument allows a similar lower bound to be obtained via Proposition 3.3 such that as $N \to \infty$ the bound coincides with the corresponding finite-dimensional distributions of the n-coalescent, as required. The modified section of the proof is presented below, using the notation of Koskela et al. (2018) for ease of comparison.

Proof. Let $\chi_d^\star$ be the conditional transition probability of a transition from state η_{d-1} to state η_d at times $\tau_N(t_{d-1})$ and $\tau_N(t_d)$ respectively, conditional on the offspring counts between those times $\nu^{(1:N)}_{\tau_N(t_{d-1})+1}, \ldots, \nu^{(1:N)}_{\tau_N(t_d)}$. This transition can happen via any valid path of merger events, but we restrict to paths involving binary mergers only, and denote by χ_d the conditional transition probability subject to this restriction. Compared to Koskela et al. (2018, Proof of Theorem 1), the derivation of an upper bound on χ_d holds without modification, while the first step in the derivation of a lower bound (Koskela et al. 2018, bottom of p.15) involves the application of Koskela et al. (2018, Lemma 1 Case 1) to bound χ_d from below and the subsequent application of (3.9). Instead, we apply Proposition 3.3 to obtain, for sufficiently large N,

$$\begin{aligned}\chi_d \geq \sum_{\substack{s_1<\ldots<s_\alpha \\ =\tau_N(t_{d-1})+1}}^{\tau_N(t_d)} (\tilde{Q}^\alpha)_{\eta_{d-1}\eta_d} \left(\prod_{r=1}^{\alpha} \mathbb{1}_{\{c_N(s_r)>\binom{n-2}{2} D_N(s_r)\}} \left[c_N(s_r) - \binom{n-2}{2} 1_N D_N(s_r) \right] \right) \\ \times \prod_{\substack{r=\tau_N(t_{d-1})+1 \\ r\notin\{s_1,\ldots,s_\alpha\}}}^{\tau_N(t_d)} \left[1 - \tilde{B}_n 1_N D_N(r) - \binom{|\eta_{d-1}| - |\{i : s_i < r\}|}{2} 1_N c_N(r) \right] \\ \times \mathbb{1}_{\{c_N(r)<(\tilde{B}_n+\binom{n}{2})^{-1}\}}.\end{aligned}$$

Here $\tilde{Q}$ is the matrix obtained from the generator Q of Kingman's n-coalescent (see Definition 2.1) by setting the diagonal entries to 0. The number of pair-merger steps required to transition from η_{d-1} to η_d is $\alpha = |\eta_{d-1}| - |\eta_d|$. The sequences $s_1, \ldots, s_\alpha$ denote the times at which these pair-mergers happen. At the remaining times r the partition is unchanged, and the bound of Proposition 3.3 has been applied to the one-step transition probabilities corresponding to these identity transitions. The constant is $\tilde{B}_n := B_n\binom{n}{2}$ where B_n is the constant defined in Proposition 3.3, and we have replaced $|\eta_d|$ by its upper bound n.

The rest of the proof proceeds as in Koskela et al. (2018, pp.16–18), albeit from this modified initial lower bound. A multinomial expansion of the product on the second line, noting that $(1_N)^a = 1_N$ for any $a \in \mathbb{R}$, yields

$$\begin{aligned}\chi_d \geq \left(\prod_{r=\tau_N(t_{d-1})+1}^{\tau_N(t_d)} \mathbb{1}_{\{c_N(r)>\binom{n-2}{2} D_N(r)\}} \mathbb{1}_{\{c_N(r)<(\tilde{B}_n+\binom{n}{2})^{-1}\}} \right) \\ \times \sum_{\beta=0}^{\tau_N(t_d)-\tau_N(t_{d-1})-\alpha} (\tilde{Q}^\alpha)_{\eta_{d-1}\eta_d} \sum_{\substack{(\lambda,\mu)\in\Pi_2([\alpha+\beta]): \\ |\lambda|=\alpha}} 1_N \\ \times \sum_{\substack{s_1<\ldots<s_{\alpha+\beta} \\ =\tau_N(t_{d-1})+1}}^{\tau_N(t_d)} \left(\prod_{r\in\lambda} \left[c_N(s_r) - \binom{n-2}{2} 1_N D_N(s_r) \right] \right) \\ \times \prod_{r\in\mu} \left\{ - \binom{|\eta_{d-1}| - |\{i\in\lambda : i < r\}|}{2} c_N(s_r) - \tilde{B}_n D_N(s_r) \right\}\end{aligned}$$

where $\Pi_i([n])$ denotes the set of partitions of $\{1,\dots,n\}$ into exactly i blocks. Expanding the product over λ gives

$$\begin{aligned}\chi_d \geq \Bigg(\prod_{r=\tau_N(t_{d-1})+1}^{\tau_N(t_d)} \mathbb{1}_{\{c_N(r)>\binom{n-2}{2}D_N(r)\}} \mathbb{1}_{\{c_N(r)<(\tilde{B}_n+\binom{n}{2})^{-1}\}} \Bigg) \\ \times \sum_{\beta=0}^{\tau_N(t_d)-\tau_N(t_{d-1})-\alpha} (\tilde{Q}^\alpha)_{\eta_{d-1}\eta_d} \sum_{\substack{(\lambda,\mu,\pi)\in\Pi_3([\alpha+\beta]):\\ |\mu|=\beta}} \binom{n-2}{2}^{|\pi|} (-1)^{|\pi|} 1_N \\ \times \sum_{\substack{s_1<\dots<s_{\alpha+\beta}\\ =\tau_N(t_{d-1})+1}}^{\tau_N(t_d)} \Bigg\{ \prod_{r\in\lambda} c_N(s_r) \Bigg\} \Bigg\{ \prod_{r\in\pi} D_N(s_r) \Bigg\} \\ \times \prod_{r\in\mu} \Bigg\{ -\binom{|\eta_{d-1}| - |\{i\in\lambda\cup\pi : i<r\}|}{2} c_N(s_r) - \tilde{B}_n D_N(s_r) \Bigg\}\end{aligned}$$

and expanding the product over μ results in

$$\begin{aligned}\chi_d \geq \Bigg(\prod_{r=\tau_N(t_{d-1})+1}^{\tau_N(t_d)} \mathbb{1}_{\{c_N(r)>\binom{n-2}{2}D_N(r)\}} \mathbb{1}_{\{c_N(r)<(\tilde{B}_n+\binom{n}{2})^{-1}\}} \Bigg) \\ \times \sum_{\beta=0}^{\tau_N(t_d)-\tau_N(t_{d-1})-\alpha} (\tilde{Q}^\alpha)_{\eta_{d-1}\eta_d} \sum_{\substack{(\lambda,\mu,\pi,\sigma)\in\Pi_4([\alpha+\beta]):\\ |\mu|+|\sigma|=\beta}} \tilde{B}_n^{|\sigma|} \binom{n-2}{2}^{|\pi|} (-1)^{|\pi|+|\sigma|} \\ \times 1_N \Bigg\{ \prod_{r\in\mu} -\binom{|\eta_{d-1}| - |\{i\in\lambda\cup\pi : i<r\}|}{2} \Bigg\} \\ \times \sum_{\substack{s_1<\dots<s_{\alpha+\beta}\\ =\tau_N(t_{d-1})+1}}^{\tau_N(t_d)} \Bigg\{ \prod_{r\in\lambda\cup\mu} c_N(s_r) \Bigg\} \prod_{r\in\pi\cup\sigma} D_N(s_r).\end{aligned}$$

Via a further multinomial expansion, the lower bound for the k-step transition probability

can be written as

$$\begin{aligned}
\lim_{N\to\infty}\mathbb{E}\left[\prod_{d=1}^{k}\chi_d\right] \geq \lim_{N\to\infty}\mathbb{E}\Bigg[&\left(\prod_{r=\tau_N(t_0)+1}^{\tau_N(t_k)}\mathbb{1}_{\{c_N(r)>\binom{n-2}{2}D_N(r)\}}\mathbb{1}_{\{c_N(r)<(\tilde{B}_n+\binom{n}{2})^{-1}\}}\right)\\
&\times\sum_{\beta_1=0}^{\infty}\cdots\sum_{\beta_k=0}^{\infty}\sum_{\substack{(\lambda_1,\mu_1,\pi_1,\sigma_1)\in\Pi_4([\alpha_1+\beta_1]):\\ |\mu_1|+|\sigma_1|=\beta_1}}\cdots\sum_{\substack{(\lambda_k,\mu_k,\pi_k,\sigma_k)\in\Pi_4([\alpha_k+\beta_k]):\\ |\mu_k|+|\sigma_k|=\beta_k}}\\
&\tilde{B}_n^{\sum_{d=1}^k|\sigma_d|}\binom{n-2}{2}^{\sum_{d=1}^k|\pi_d|}(-1)^{\sum_{d=1}^k|\pi_d|+|\sigma_d|}\,1_N\\
&\times\left\{\prod_{d=1}^{k}(\tilde{Q}^{\alpha_d})_{\eta_{d-1}\eta_d}\prod_{r\in\mu_d}-\binom{|\eta_{d-1}|-|\{i\in\lambda_d\cup\pi_d:i<r\}|}{2}\right\}\\
&\times\sum_{\substack{s_1^{(1)}<\ldots<s_{\alpha_1+\beta_1}^{(1)}\\ =\tau_N(t_0)+1}}^{\tau_N(t_1)}\cdots\sum_{\substack{s_1^{(k)}<\ldots<s_{\alpha_k+\beta_k}^{(k)}\\ =\tau_N(t_{k-1})+1}}^{\tau_N(t_k)}\\
&\prod_{d=1}^{k}\mathbb{1}_{\{\tau_N(t_d)-\tau_N(t_{d-1})\geq\alpha_d+\beta_d\}}\left\{\prod_{r\in\lambda_d\cup\mu_d}c_N(s_r^{(d)})\right\}\prod_{r\in\pi_d\cup\sigma_d}D_N(s_r^{(d)})\Bigg].
\end{aligned}$$

An argument completely analogous to that in Koskela et al. (2018, p.24) shows that passing the expectation and the limit through the infinite sums is justified, by Fubini's theorem and dominated convergence, whereupon the contribution of terms with $\sum_{d=1}^{k}(|\pi_d|+|\sigma_d|)>0$ vanishes. To see why, follow the argument used to show that the contribution of multiple merger trajectories vanishes in the corresponding upper bound in Koskela et al. (2018, p.10). That leaves

$$\begin{aligned}
\lim_{N\to\infty}\mathbb{E}\left[\prod_{d=1}^{k}\chi_d\right] \geq &\sum_{\beta_1=0}^{\infty}\cdots\sum_{\beta_k=0}^{\infty}\sum_{\substack{(\lambda_1,\mu_1)\in\Pi_2([\alpha_1+\beta_1]):\\ |\mu_1|=\beta_1}}\cdots\sum_{\substack{(\lambda_k,\mu_k)\in\Pi_2([\alpha_k+\beta_k]):\\ |\mu_k|=\beta_k}}\\
&\left\{\prod_{d=1}^{k}(\tilde{Q}^{\alpha_d})_{\eta_{d-1}\eta_d}\prod_{r\in\mu_d}-\binom{|\eta_{d-1}|-|\{i\in\lambda_d\cup\pi_d:i<r\}|}{2}\right\}\\
&\times\lim_{N\to\infty}\mathbb{E}\Bigg[\left(\prod_{r=\tau_N(t_{d-1})+1}^{\tau_N(t_d)}\mathbb{1}_{\{c_N(r)>\binom{n-2}{2}D_N(r)\}}\mathbb{1}_{\{c_N(r)<(\tilde{B}_n+\binom{n}{2})^{-1}\}}\right)\\
&\times\sum_{\substack{s_1^{(1)}<\ldots<s_{\alpha_1+\beta_1}^{(1)}\\ =\tau_N(t_0)+1}}^{\tau_N(t_1)}\cdots\sum_{\substack{s_1^{(k)}<\ldots<s_{\alpha_k+\beta_k}^{(k)}\\ =\tau_N(t_{k-1})+1}}^{\tau_N(t_k)}\\
&\prod_{d=1}^{k}\mathbb{1}_{\{\tau_N(t_d)-\tau_N(t_{d-1})\geq\alpha_d+\beta_d\}}\left\{\prod_{r\in\lambda_d\cup\mu_d}c_N(s_r^{(d)})\right\}\Bigg].
\end{aligned}\tag{3.19}$$

We have, from Koskela et al. (2018, Equation (11)),

$$\sum_{\substack{(\lambda,\mu)\in\Pi_2([\alpha+\beta]):\\ |\mu|=\beta}} (\tilde{Q}^{\alpha})_{\eta_{d-1}\eta_d} \prod_{r\in\mu} -\binom{|\eta_{d-1}| - |\{i\in\lambda\cup\pi : i<r\}|}{2} = (Q^{\alpha+\beta})_{\eta_{d-1}\eta_d}.$$

Applying this k times in (3.19) yields

$$\begin{aligned}
\lim_{N\to\infty}\mathbb{E}\left[\prod_{d=1}^{k}\chi_d\right] \geq \sum_{\beta_1=0}^{\infty}\cdots\sum_{\beta_k=0}^{\infty}\left\{\prod_{d=1}^{k}(Q^{\alpha_d+\beta_d})_{\eta_{d-1}\eta_d}\right\} \\
\times \lim_{N\to\infty}\mathbb{E}\Bigg\{\left(\prod_{r=\tau_N(t_{d-1})+1}^{\tau_N(t_d)} \mathbb{1}_{\{c_N(r)>\binom{n-2}{2}D_N(r)\}}\mathbb{1}_{\{c_N(r)<(\tilde{B}_n+\binom{n}{2})^{-1}\}}\right) \\
\times\left(\prod_{d=1}^{k}\mathbb{1}_{\{\tau_N(t_d)-\tau_N(t_{d-1})\geq\alpha_d+\beta_d\}}\right) \\
\times \sum_{\substack{s_1^{(1)}<\ldots<s_{\alpha_1+\beta_1}^{(1)}\\ =\tau_N(t_0)+1}}^{\tau_N(t_1)}\cdots\sum_{\substack{s_1^{(k)}<\ldots<s_{\alpha_k+\beta_k}^{(k)}\\ =\tau_N(t_{k-1})+1}}^{\tau_N(t_k)}\prod_{d=1}^{k}\prod_{r\in\lambda_d\cup\mu_d} c_N(s_r^{(d)})\Bigg\}.
\end{aligned}$$

We now apply equations (14) and (15), respectively, of Koskela et al. (2018), to those terms with a negative ($|\boldsymbol{\beta}|$ odd) and positive ($|\boldsymbol{\beta}|$ even) sign, respectively, to obtain

$$\begin{aligned}
\lim_{N\to\infty}\mathbb{E}\left[\prod_{d=1}^{k}\chi_d\right] &\geq \sum_{\beta_1=0}^{\infty}\cdots\sum_{\beta_k=0}^{\infty}\left\{\prod_{d=1}^{k}(Q^{\alpha_d+\beta_d})_{\eta_{d-1}\eta_d}\frac{(t_d-t_{d-1})^{\alpha_d+\beta_d}}{(\alpha_d+\beta_d)!}\right\} \\
&\quad\times\lim_{N\to\infty}\mathbb{E}\left[\left(\prod_{r=\tau_N(t_{d-1})+1}^{\tau_N(t_d)} \mathbb{1}_{\{c_N(r)>\binom{n-2}{2}D_N(r)\}}\mathbb{1}_{\{c_N(r)<(\tilde{B}_n+\binom{n}{2})^{-1}\}}\right)\right. \\
&\qquad\left.\times\left(\prod_{d=1}^{k}\mathbb{1}_{\{\tau_N(t_d)-\tau_N(t_{d-1})\geq\alpha_d+\beta_d\}}\right)\right] \\
&\geq \sum_{\beta_1=0}^{\infty}\cdots\sum_{\beta_k=0}^{\infty}\left\{\prod_{d=1}^{k}(Q^{\alpha_d+\beta_d})_{\eta_{d-1}\eta_d}\frac{(t_d-t_{d-1})^{\alpha_d+\beta_d}}{(\alpha_d+\beta_d)!}\right\}
\end{aligned}$$

where the expectation of the indicators converges to 1 by a trivial modification of Koskela et al. (2018, Lemma 4). ■

4 Weak Convergence

> At the age of twenty-one he wrote a treatise upon the Binomial Theorem, and had, to all appearances, a most brilliant career before him.
>
> Sherlock Holmes

In this chapter we present a weak convergence result which is identical to Theorem 3.6 except that the mode of convergence is strengthened from convergence of the finite-dimensional distributions to weak convergence. Weak convergence is desirable because it implies convergence of a strictly larger class of functions of genealogies, granting access to the distributions of statistics such as the time to the sample MRCA, the total branch length, and the probability that the MRCA of a subsample is equal to the sample MRCA.

The extension from Theorem 3.6 to weak convergence requires an additional tightness argument. The proof is rather long-winded since we do not have the strong assumptions on the dynamics of the interacting particle system that are exploited for example in Möhle (1999). The proof is broken down into a series of technical results which culminate in Theorem 4.1. The overall structure of the proof is depicted graphically in Figure 4.1.

We start by defining a suitable metric space. Let $\mathcal{P}_n$ be the space of partitions of $\{1, \ldots, n\}$. Denote by $\mathcal{D}$ the set of all functions mapping $[0, \infty)$ to $\mathcal{P}_n$ that are right-continuous with left limits. Our rescaled genealogical process $(\mathcal{G}^{(n,N)}_{\tau_N(t)})_{t \geq 0}$ and our encoding of the n-coalescent are piecewise-constant functions mapping time $t \in [0, \infty)$ to partitions, and thus live in the space $\mathcal{D}$. Finally, equip the space $\mathcal{P}_n$ with the discrete metric,

$$\rho(\xi, \eta) = 1 - \delta_{\xi\eta} := \begin{cases} 0 & \text{if } \xi = \eta \\ 1 & \text{otherwise} \end{cases}$$

for any $\xi, \eta \in \mathcal{P}_n$.

Theorem 4.1. *Let $\nu_t^{(1:N)}$ denote the offspring numbers in an interacting particle system satisfying* (A1) *and such that, for any N sufficiently large, for all finite t, $\mathbb{P}[\tau_N(t) = \infty] = 0$. Suppose that there exists a deterministic sequence $(b_N)_{N\in\mathbb{N}}$ such that $\lim_{N\to\infty} b_N = 0$ and*

$$\frac{1}{(N)_3}\sum_{i=1}^{N}\mathbb{E}_t\left[(\nu_t^{(i)})_3\right] \leq b_N \frac{1}{(N)_2}\sum_{i=1}^{N}\mathbb{E}_t\left[(\nu_t^{(i)})_2\right] \tag{4.1}$$

almost surely for all N, uniformly in $t \geq 1$. Then the rescaled genealogical process $(G^{(n,N)}_{\tau_N(t)})_{t\geq 0}$ converges weakly in $(\mathcal{D}, \rho)$ to Kingman's n-coalescent as $N \to \infty$.

Proof of Theorem 4.1. The structure of the proof follows Möhle (1999), albeit with considerable technical complication due to the dependence between generations (non-neutrality) in our model. To make it digestible, the proof is broken down into a number of results which are organised into sections; the relationships between these are shown in Figure 4.1.

Since we already have convergence of the finite-dimensional distributions (Theorem 3.6), strengthening this to weak convergence requires relative compactness of the sequence of processes $\{(G^{(n,N)}_{\tau_N(t)})_{t\geq 0}\}_{N\in\mathbb{N}}$.

Ethier and Kurtz (1986, Chapter 3, Corollary 7.4) provide a necessary and sufficient condition for relative compactness: $\mathcal{P}_n$ is finite and therefore complete and separable, and the sample paths of $(G^{(n,N)}_{\tau_N(t)})_{t\geq 0}$ live in $\mathcal{D}$, so the conditions of their corollary are satisfied. The corollary states that the sequence of processes $\{(G^{(n,N)}_{\tau_N(t)})_{t\geq 0}\}_{N\in\mathbb{N}}$ is relatively compact if and only if the following two conditions hold:

1. For every $\epsilon > 0$, $t \geq 0$ there exists a compact set $\Gamma \subseteq \mathcal{P}_n$ such that

$$\liminf_{N\to\infty} \mathbb{P}\left[G^{(n,N)}_{\tau_N(t)} \in \Gamma\right] \geq 1 - \epsilon$$

2. For every $\epsilon > 0$, $t > 0$ there exists $\delta > 0$ such that

$$\liminf_{N\to\infty} \mathbb{P}\left[\omega\left(G^{(n,N)}_{\tau_N(\cdot)}, \delta, t\right) < \epsilon\right] \geq 1 - \epsilon$$

   where ω is the modified modulus of continuity:

$$\omega\left(G^{(n,N)}_{\tau_N(\cdot)}, \delta, t\right) := \inf \max_{i\in[K]} \sup_{u,v\in[T_{i-1},T_i)} \rho\left(G^{(n,N)}_{\tau_N(u)}, G^{(n,N)}_{\tau_N(v)}\right)$$

   with the infimum taken over all partitions of the form $0 = T_0 < T_1 < \cdots < T_{K-1} < t \leq T_K$ (for any K) such that $\min_{i\in[K]}(T_i - T_{i-1}) > \delta$.

In our case, Condition 1 is satisfied automatically with $\Gamma = \mathcal{P}_n$, since $\mathcal{P}_n$ is finite and hence compact. Intuitively, Condition 2 ensures that the jumps of the process are well-separated. In our case where ρ is the discrete metric, we see that $\rho(G^{(n,N)}_{\tau_N(u)}, G^{(n,N)}_{\tau_N(v)})$ is equal

to 1 if there is a jump between times u and v, and 0 otherwise. Taking the supremum and maximum then indicates whether there is a jump inside any of the intervals of the given partition; this can only be equal to zero if all of the jumps up to time t occur exactly at the times $T_0, \ldots, T_K$. The infimum over all allowed partitions, then, can only be equal to zero if no two jumps occur less than δ (unscaled) time apart, because of the restriction we placed on these partitions.

The proof is concentrated on proving Condition 2. To do this, we use a coupling with another process that contains all of the jumps of the genealogical process, with the addition of some extra jumps. This process is constructed in such a way that it can be shown to satisfy Condition 2, and hence so does the genealogical process.

Define $p_t := \max_{\xi \in \mathcal{P}_n}\{1 - p_{\xi\xi}(t)\} = 1 - p_{\Delta\Delta}(t)$, where Δ denotes the trivial partition of singletons $\{\{1\}, \ldots, \{n\}\}$. For a proof that the maximum is attained at $\xi = \Delta$, see Lemma 4.2. Following Möhle (1999), we now construct the two-dimensional Markov process $(Z_t, S_t)_{t \in \mathbb{N}_0}$ on $\mathbb{N}_0 \times \mathcal{P}_n$ with transition probabilities

$$\begin{aligned}
&\mathbb{P}\left[Z_t = j, S_t = \eta \mid Z_{t-1} = i, S_{t-1} = \xi, \mathcal{F}_\infty\right] \\
&\qquad = \begin{cases} 1 - p_t & \text{if } j = i \text{ and } \eta = \xi \\ p_{\xi\xi}(t) + p_t - 1 & \text{if } j = i+1 \text{ and } \eta = \xi \\ p_{\xi\eta}(t) & \text{if } j = i+1 \text{ and } \eta \neq \xi \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{4.2}$$

and initial state $Z_0 = 0$, $S_0 = \Delta$. Unlike the corresponding process in Möhle (1999), in our case the transition probabilities depend on offspring counts, thus the process is only Markovian conditional on $\mathcal{F}_\infty$. It can be thought of as a time-inhomogeneous Markov process in a random environment.

The construction is such that the marginal (S_t) has the same distribution as the genealogical process of interest, and (Z_t) has jumps at all the times (S_t) does plus some extra jumps. The definition of p_t ensures that the probability in the second case of (4.2) is non-negative, attaining the value zero when $\xi = \Delta$. Furthermore, the transition probabilities (and hence jump times) of (Z_t) do not depend on the current state.

Denote by $0 = T_0^{(N)} < T_1^{(N)} < \ldots$ the jump times of the rescaled process $(Z_{\tau_N(t)})_{t \geq 0}$, and by $\varpi_i^{(N)} := T_i^{(N)} - T_{i-1}^{(N)}$ the corresponding holding times.

Suppose that for some fixed $\varpi_1^{(N)}, \varpi_2^{(N)}, \ldots$ and $t > 0$, there exists $m \in \mathbb{N}$ and $\delta > 0$ such that $\varpi_i^{(N)} > \delta$ for all $i \in \{1, \ldots, m\}$, and $T_m^{(N)} \geq t$. Then $K_N := \min\{i : T_i^{(N)} \geq t\}$ is well-defined with $1 \leq K_N \leq m$, and $T_1^{(N)}, \ldots, T_{K_N}^{(N)}$ form a partition of the form required for Condition 2. Indeed $(Z_{\tau_N(\cdot)})$ is constant on every interval $[T_{i-1}^{(N)}, T_i^{(N)})$ by construction, so $\omega((Z_{\tau_N(\cdot)}), \delta, t) = 0$. We therefore have that for each $m \in \mathbb{N}$ and $\delta > 0$,

$$\mathbb{P}\left[\omega\left((Z_{\tau_N(\cdot)}), \delta, t\right) < \epsilon\right] \geq \mathbb{P}\left[T_m^{(N)} \geq t, \varpi_i^{(N)} > \delta \, \forall i \in \{1, \ldots, m\}\right].$$

Thus a sufficient condition for Condition 2 is: for any $\epsilon > 0$, $t > 0$, there exist $m \in \mathbb{N}$, $\delta > 0$ such that

$$\liminf_{N\to\infty} \mathbb{P}\left[T_m^{(N)} \geq t, \varpi_i^{(N)} > \delta \,\forall i \in \{1,\ldots,m\}\right] \geq 1-\epsilon. \tag{4.3}$$

Due to Lemma 4.3, the limiting distributions of $\varpi_i^{(N)}$ are i.i.d. Exp(α_n), where $\alpha_n := n(n-1)/2$, so

$$\limsup_{N\to\infty} \mathbb{P}\left[\varpi_i^{(N)} \leq \delta\right] = 1 - \liminf_{N\to\infty} \mathbb{P}\left[\varpi_i^{(N)} > \delta\right] = 1 - e^{-\alpha_n \delta}$$

for each i, and

$$\begin{aligned}
\limsup_{N\to\infty} \mathbb{P}\left[T_m^{(N)} < t\right] &= 1 - \liminf_{N\to\infty} \mathbb{P}\left[T_m^{(N)} \geq t\right] \\
&= 1 - \liminf_{N\to\infty} \mathbb{P}\left[\varpi_1^{(N)} + \cdots + \varpi_m^{(N)} \geq t\right] \\
&= 1 - e^{-\alpha_n t} \sum_{i=0}^{m-1} \frac{(\alpha_n t)^i}{i!}.
\end{aligned}$$

using the series expansion for the Erlang CDF (see for example Forbes et al. 2011, Chapter 15). Now

$$\begin{aligned}
&\liminf_{N\to\infty} \mathbb{P}\left[T_m^{(N)} \geq t, \varpi_i^{(N)} > \delta \,\forall i \in \{1,\ldots,m\}\right] \\
&\qquad = 1 - \limsup_{N\to\infty} \mathbb{P}\left[\{T_m^{(N)} < t\} \cup \bigcup_{i=1}^{m} \{\varpi_i^{(N)} \leq \delta\}\right] \\
&\qquad \geq 1 - \limsup_{N\to\infty} \mathbb{P}\left[T_m^{(N)} < t\right] - \sum_{i=1}^{m} \limsup_{N\to\infty} \mathbb{P}\left[\varpi_i^{(N)} \leq \delta\right] \\
&\qquad = 1 - \left(1 - e^{-\alpha_n t} \sum_{i=0}^{m-1} \frac{(\alpha_n t)^i}{i!}\right) - m(1 - e^{-\alpha_n \delta}),
\end{aligned}$$

which can be made $\geq 1-\epsilon$ by taking m sufficiently large and δ sufficiently small. Since this argument applies for any ϵ and t, (4.3) and hence Condition 2 is satisfied, and the proof is complete. ■

Lemma 4.2. $\max_{\xi\in\mathcal{P}_n}\{1 - p_{\xi\xi}(t)\} = 1 - p_{\Delta\Delta}(t)$.

Proof. Consider any $\xi \in E$ consisting of k blocks ($1 \leq k \leq n-1$), and any $\xi' \in E$ consisting of $k+1$ blocks. Setting $\eta = \xi$ in (3.4),

$$p_{\xi\xi}(t) = \frac{1}{(N)_k} \sum_{\substack{i_1,\ldots,i_k=1 \\ \text{all distinct}}}^{N} \nu_t^{(i_1)} \cdots \nu_t^{(i_k)}.$$

Similarly,

$$p_{\xi'\xi'}(t) = \frac{1}{(N)_{k+1}} \sum_{\substack{i_1,\dots,i_{k+1}=1 \\ \text{all distinct}}}^{N} \nu_t^{(i_1)} \cdots \nu_t^{(i_k)} \nu_t^{(i_{k+1})}$$
$$= \frac{1}{(N)_k (N-k)} \sum_{\substack{i_1,\dots,i_k=1 \\ \text{all distinct}}}^{N} \left\{ \nu_t^{(i_1)} \cdots \nu_t^{(i_k)} \sum_{\substack{i_{k+1}=1 \\ \notin\{i_1,\dots,i_k\}}}^{N} \nu_t^{(i_{k+1})} \right\}.$$

Discarding the zero summands,

$$p_{\xi'\xi'}(t) = \frac{1}{(N)_k (N-k)} \sum_{\substack{i_1,\dots,i_k=1 \\ \text{all distinct:} \\ \nu_t^{(i_1)},\dots,\nu_t^{(i_k)}>0}}^{N} \left\{ \nu_t^{(i_1)} \cdots \nu_t^{(i_k)} \sum_{\substack{i_{k+1}=1 \\ \notin\{i_1,\dots,i_k\}}}^{N} \nu_t^{(i_{k+1})} \right\}.$$

The inner sum is

$$\sum_{\substack{i_{k+1}=1 \\ \notin\{i_1,\dots,i_k\}}}^{N} \nu_t^{(i_{k+1})} = \left\{ \sum_{i=1}^{N} \nu_t^{(i)} - \sum_{i \in \{i_1,\dots,i_k\}} \nu_t^{(i)} \right\} \leq N - k,$$

since $\nu_t^{(i_1)}, \dots, \nu_t^{(i_k)}$ are all at least 1. Hence

$$p_{\xi'\xi'}(t) \leq \frac{N-k}{(N)_k (N-k)} \sum_{\substack{i_1,\dots,i_k=1 \\ \text{all distinct:} \\ \nu_t^{(i_1)},\dots,\nu_t^{(i_k)}>0}}^{N} \nu_t^{(i_1)} \cdots \nu_t^{(i_k)} = p_{\xi\xi}(t).$$

Thus $p_{\xi\xi}(t)$ is decreasing in the number of blocks of ξ, and is therefore minimised by taking $\xi = \Delta$, which uniquely achieves the maximum n blocks. This choice in turn maximises $1 - p_{\xi\xi}(t)$, as required. ■

Lemma 4.3. *The finite-dimensional distributions of $\varpi_1^{(N)}, \varpi_2^{(N)}, \dots$ converge as $N \to \infty$ to those of $\varpi_1, \varpi_2, \dots$, where the ϖ_i are independent* $\mathrm{Exp}(\alpha_n)$*-distributed random variables.*

Proof. There is a continuous bijection between the jump times $T_1^{(N)}, T_2^{(N)}, \dots$ and the holding times $\varpi_1^{(N)}, \varpi_2^{(N)}, \dots$, so convergence of the holding times to $\varpi_1, \varpi_2, \dots$ is equivalent to convergence of the jump times to $T_1, T_2, \dots$, where $T_i := \varpi_1 + \cdots + \varpi_i$. We will work with the jump times, following the structure of Möhle (1999, Lemma 3.2).

The idea is to prove by induction that, for any $k \in \mathbb{N}$ and $t_1, \dots, t_k > 0$,

$$\lim_{N\to\infty} \mathbb{P}\left[T_1^{(N)} \leq t_1, \dots, T_k^{(N)} \leq t_k\right] = \mathbb{P}[T_1 \leq t_1, \dots, T_k \leq t_k]. \tag{4.4}$$

Take the basis case $k = 1$, for which

$$\mathbb{P}[T_1 \leq t] = \mathbb{P}[\varpi_1 \leq t] = 1 - e^{-\alpha_n t}$$

and $T_1^{(N)} > t$ if and only if Z has no jumps up to time t:

$$\mathbb{P}\left[T_1^{(N)} > t\right] = \mathbb{E}\left[\mathbb{P}[T_1^{(N)} > t \mid \mathcal{F}_\infty]\right] = \mathbb{E}\left[\prod_{r=1}^{\tau_N(t)} (1 - p_r)\right].$$

Lemma 4.7 shows that this probability converges to $e^{-\alpha_n t}$ as required.

For the induction step, assume that (4.4) holds for some k. We have the following decomposition:

$$\begin{aligned}\mathbb{P}\left[T_1^{(N)} \leq t_1, \ldots, T_{k+1}^{(N)} \leq t_{k+1}\right] &= \mathbb{P}\left[T_1^{(N)} \leq t_1, \ldots, T_k^{(N)} \leq t_k\right] \\ &\quad - \mathbb{P}\left[T_1^{(N)} \leq t_1, \ldots, T_k^{(N)} \leq t_k, T_{k+1}^{(N)} > t_{k+1}\right].\end{aligned}$$

The first term on the right-hand side converges to $\mathbb{P}[T_1 \leq t_1, \ldots, T_k \leq t_k]$ by the induction hypothesis, and it remains to show that

$$\lim_{N\to\infty} \mathbb{P}\left[T_1^{(N)} \leq t_1, \ldots, T_k^{(N)} \leq t_k, T_{k+1}^{(N)} > t_{k+1}\right] = \mathbb{P}[T_1 \leq t_1, \ldots, T_k \leq t_k, T_{k+1} > t_{k+1}].$$

As shown in Möhle (1999),

$$\mathbb{P}[T_1 \leq t_1, \ldots, T_k \leq t_k, T_{k+1} > t_{k+1}] = \alpha_n^k e^{-\alpha_n t} \sum_{\substack{i_1 \leq \cdots \leq i_{k-1} \\ \in \{0,\ldots,k\}: \\ i_j \geq j \forall j}} \prod_{j=1}^{k} \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!},$$

while the probability on the left-hand side can be written

$$\begin{aligned}&\mathbb{P}\left[T_1^{(N)} \leq t_1, \ldots, T_k^{(N)} \leq t_k, T_{k+1}^{(N)} > t_{k+1}\right] \\ &\qquad = \mathbb{E}\left[\mathbb{P}[T_1^{(N)} \leq t_1, \ldots, T_k^{(N)} \leq t_k, T_{k+1}^{(N)} > t_{k+1} \mid \mathcal{F}_\infty]\right] \\ &\qquad = \mathbb{E}\left[\sum_{\substack{r_1 < \cdots < r_k: \\ r_i \leq \tau_N(t_i) \forall i}} \left(\prod_{i=1}^{k} p_{r_i}\right) \left(\prod_{\substack{r=1 \\ \notin \{r_1,\ldots,r_k\}}}^{\tau_N(t)} (1 - p_r)\right)\right].\end{aligned}$$

That is, there are jumps at some times $r_1, \ldots, r_k$ and identity transitions at all other times. A similar expression is derived in Möhle (1999), but here we have an additional outer expectation because the probabilities p_r depend on the offspring counts which are random. Lemmata 4.8 and 4.9 show that this probability converges to the correct limit. This completes the induction. ■

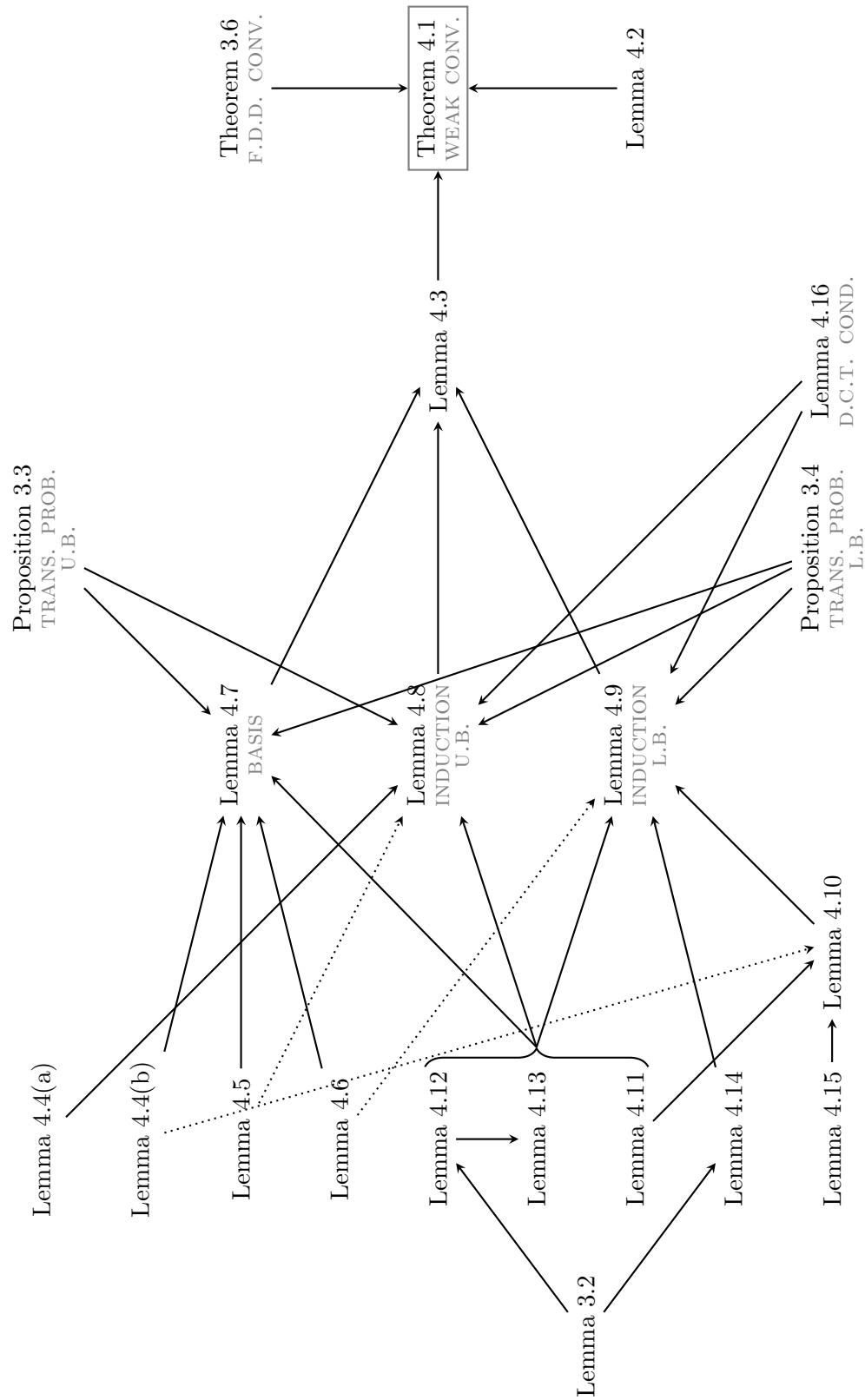


Figure 4.1: Graph showing dependencies between the lemmata used to prove weak convergence. Dotted arrows indicate dependence via a slight modification of the preceding lemma.

4.1 Bounds on sum-products

We start by proving some upper and lower bounds on sums of products of various quantities, which appear from our bounds on p_r (Propositions 3.3 and 3.4). These sum-product bounds will be applied multiple times in the lemmata of this chapter.

Lemma 4.4. *Fix $t > 0$, $l \in \mathbb{N}$.*

(a) $$\sum_{\substack{s_1,\dots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} c_N(s_j) \leq (t+1)^l$$

(b) $$t^l - \left(\sum_{s=1}^{\tau_N(t)} c_N(s)^2\right)\binom{l}{2}(t+1)^{l-2} \leq \sum_{\substack{s_1,\dots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} c_N(s_j) \leq t^l + c_N(\tau_N(t))(t+1)^l$$

Proof. **(a)** Firstly, we have the inequality

$$\sum_{\substack{s_1,\dots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} c_N(s_j) \leq \left(\sum_{s=0}^{\tau_N(t)} c_N(s)\right)^l,$$

as can be seen by considering the multinomial expansion of the right-hand side. Applying Proposition 3.1(d),

$$\sum_{\substack{s_1,\dots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} c_N(s_j) \leq (t+1)^l.$$

(b) As pointed out in Koskela et al. (2018, Equation (8)),

$$\sum_{\substack{s_1,\dots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} c_N(s_j) \geq \left(\sum_{s=0}^{\tau_N(t)} c_N(s)\right)^l - \binom{l}{2}\left(\sum_{s=0}^{\tau_N(t)} c_N(s)^2\right)\left(\sum_{s=0}^{\tau_N(t)} c_N(s)\right)^{l-2}. \tag{4.5}$$

Applying Proposition 3.1(d) on the right-hand side of (4.5) yields the lower bound.

For the upper bound we have

$$\sum_{\substack{s_1,\dots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} c_N(s_j) \leq \left(\sum_{s=0}^{\tau_N(t)} c_N(s)\right)^l \leq \left(\sum_{s=0}^{\tau_N(t)-1} c_N(s) + c_N(\tau_N(t))\right)^l \leq \left[t + c_N(\tau_N(t))\right]^l,$$

using the definition of τ_N. A binomial expansion yields

$$\left[t + c_N(\tau_N(t))\right]^l = t^l + \sum_{i=0}^{l-1}\binom{l}{i} t^i c_N(\tau_N(t))^{l-i} = t^l + c_N(\tau_N(t))\sum_{i=0}^{l-1}\binom{l}{i} t^i c_N(\tau_N(t))^{l-1-i},$$

then by Proposition 3.1(a),

$$\sum_{i=0}^{l-1}\binom{l}{i}t^i c_N(\tau_N(t))^{l-1-i} \le \sum_{i=0}^{l-1}\binom{l}{i}t^i \le (t+1)^l.$$

Putting this together yields the upper bound. ■

Lemma 4.5. *Fix $t > 0$, $l \in \mathbb{N}$. Then, for any constant $B > 0$,*

$$\sum_{\substack{s_1,\dots,s_l=1\\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} [c_N(s_j) + BD_N(s_j)] \le \sum_{\substack{s_1,\dots,s_l=1\\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} c_N(s_j) + \left(\sum_{s=1}^{\tau_N(t)} D_N(s)\right)(t+1)^{l-1}(1+B)^l.$$

Proof. We start with a binomial expansion:

$$\begin{aligned}\sum_{\substack{s_1,\dots,s_l=1\\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} [c_N(s_j) + BD_N(s_j)] &= \sum_{\substack{s_1,\dots,s_l=1\\ \text{all distinct}}}^{\tau_N(t)} \sum_{\mathcal{I}\subseteq[l]} B^{l-|\mathcal{I}|}\left(\prod_{i\in\mathcal{I}} c_N(s_i)\right)\left(\prod_{j\notin\mathcal{I}} D_N(s_j)\right)\\ &= \sum_{\mathcal{I}\subseteq[l]} B^{l-|\mathcal{I}|} \sum_{\substack{s_1,\dots,s_l=1\\ \text{all distinct}}}^{\tau_N(t)} \left(\prod_{i\in\mathcal{I}} c_N(s_i)\right)\left(\prod_{j\notin\mathcal{I}} D_N(s_j)\right)\end{aligned} \tag{4.6}$$

where $[l] := \{1,\dots,l\}$. Since we are summing over all permutations of $s_1,\dots,s_l$, the inner sum depends on $\mathcal{I}$ only through $I := |\mathcal{I}|$. We may therefore replace the sum over $\mathcal{I} \subseteq \{1,\dots,l\}$ with a sum over the size I of the subset and a binomial coefficient counting the number of terms in which the subset is of size I:

$$\begin{aligned}&\sum_{\mathcal{I}\subseteq[l]} B^{l-|\mathcal{I}|} \sum_{\substack{s_1,\dots,s_l=1\\ \text{all distinct}}}^{\tau_N(t)} \left(\prod_{i\in\mathcal{I}} c_N(s_i)\right)\left(\prod_{j\notin\mathcal{I}} D_N(s_j)\right)\\ &\qquad = \sum_{I=0}^{l}\binom{l}{I} B^{l-I} \sum_{\substack{s_1,\dots,s_l=1\\ \text{all distinct}}}^{\tau_N(t)} \left(\prod_{i=1}^{I} c_N(s_i)\right)\left(\prod_{j=I+1}^{l} D_N(s_j)\right).\end{aligned}$$

Separating the term $I = l$,

$$\sum_{I=0}^{l} \binom{l}{I} B^{l-I} \sum_{\substack{s_1,\dots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \left(\prod_{i=1}^{I} c_N(s_i) \right) \left(\prod_{j=I+1}^{l} D_N(s_j) \right)$$
$$= \sum_{\substack{s_1,\dots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} c_N(s_j) + \sum_{I=0}^{l-1} \binom{l}{I} B^{l-I} \sum_{\substack{s_1,\dots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \left(\prod_{i=1}^{I} c_N(s_i) \right) \left(\prod_{j=I+1}^{l} D_N(s_j) \right). \tag{4.7}$$

In the second term on the right-hand side, there is always at least one D_N term, so using that $c_N(s) \geq D_N(s)$ (Proposition 3.1(b)) we can write

$$\begin{aligned}
&\sum_{I=0}^{l-1} \binom{l}{I} B^{l-I} \sum_{\substack{s_1,\dots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \left(\prod_{i=1}^{I} c_N(s_i) \right) \left(\prod_{j=I+1}^{l} D_N(s_j) \right) \\
&\qquad \leq \sum_{I=0}^{l-1} \binom{l}{I} B^{l-I} \sum_{\substack{s_1,\dots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \left(\prod_{i=1}^{l-1} c_N(s_i) \right) D_N(s_l) \\
&\qquad \leq \sum_{I=0}^{l-1} \binom{l}{I} B^{l-I} \left(\sum_{\substack{s_1,\dots,s_{l-1}=1 \\ \text{all distinct}}}^{\tau_N(t)} \prod_{i=1}^{l-1} c_N(s_i) \right) \sum_{s_l=1}^{\tau_N(t)} D_N(s_l) \\
&\qquad \leq \sum_{I=0}^{l-1} \binom{l}{I} B^{l-I} (t+1)^{l-1} \sum_{s=1}^{\tau_N(t)} D_N(s)
\end{aligned} \tag{4.8}$$

using Lemma 4.4(a). Finally, by the Binomial Theorem,

$$\sum_{I=0}^{l-1} \binom{l}{I} B^{l-I} (t+1)^{l-1} \sum_{s=1}^{\tau_N(t)} D_N(s) \leq \left(\sum_{s=1}^{\tau_N(t)} D_N(s) \right) (t+1)^{l-1} (1+B)^l, \tag{4.9}$$

which, together with (4.7), concludes the proof. ■

Lemma 4.6. *Fix $t > 0$, $l \in \mathbb{N}$. Then, for any constant $B > 0$,*

$$\sum_{\substack{s_1,\dots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} [c_N(s_j) - B D_N(s_j)]$$
$$\geq \sum_{\substack{s_1,\dots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} c_N(s_j) - \left(\sum_{s=1}^{\tau_N(t)} D_N(s) \right) (t+1)^{l-1} (1+B)^l.$$

Proof. A binomial expansion and subsequent manipulation as in (4.6)–(4.7) gives

$$\begin{aligned}
&\sum_{\substack{s_1,\ldots,s_l=1\\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} [c_N(s_j) - BD_N(s_j)] \\
&\quad = \sum_{\mathcal{I}\subseteq[l]} (-B)^{l-|\mathcal{I}|} \sum_{\substack{s_1,\ldots,s_l=1\\ \text{all distinct}}}^{\tau_N(t)} \left(\prod_{i\in\mathcal{I}} c_N(s_i)\right)\left(\prod_{j\notin\mathcal{I}} D_N(s_j)\right) \\
&\quad = \sum_{I=0}^{l} \binom{l}{I} (-B)^{l-I} \sum_{\substack{s_1,\ldots,s_l=1\\ \text{all distinct}}}^{\tau_N(t)} \left(\prod_{i=1}^{I} c_N(s_i)\right)\left(\prod_{j=I+1}^{l} D_N(s_j)\right) \\
&\quad = \sum_{\substack{s_1,\ldots,s_l=1\\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} c_N(s_j) + \sum_{I=0}^{l-1} \binom{l}{I} (-B)^{l-I} \sum_{\substack{s_1,\ldots,s_l=1\\ \text{all distinct}}}^{\tau_N(t)} \left(\prod_{i=1}^{I} c_N(s_i)\right)\left(\prod_{j=I+1}^{l} D_N(s_j)\right) \\
&\quad \geq \sum_{\substack{s_1,\ldots,s_l=1\\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} c_N(s_j) - \sum_{I=0}^{l-1} \binom{l}{I} B^{l-I} \sum_{\substack{s_1,\ldots,s_l=1\\ \text{all distinct}}}^{\tau_N(t)} \left(\prod_{i=1}^{I} c_N(s_i)\right)\left(\prod_{j=I+1}^{l} D_N(s_j)\right)
\end{aligned}$$

where the last inequality just multiplies some positive terms by -1. Then (4.8)–(4.9) can be applied directly (noting that an upper bound on negative terms gives a lower bound overall):

$$\begin{aligned}
&-\sum_{I=0}^{l-1} \binom{l}{I} B^{l-I} \sum_{\substack{s_1,\ldots,s_l=1\\ \text{all distinct}}}^{\tau_N(t)} \left(\prod_{i=1}^{I} c_N(s_i)\right)\left(\prod_{j=I+1}^{l} D_N(s_j)\right) \\
&\qquad \geq -\left(\sum_{s=1}^{\tau_N(t)} D_N(s)\right)(t+1)^{l-1}(1+B)^l
\end{aligned}$$

which concludes the proof. ∎

4.2 Main components of induction argument

This section contains the technical aspects of the proof of Lemma 4.3, which establishes the limiting distributions of holding times of the coupled process, via an induction argument. This section is split into four lemmata: the first (Lemma 4.7) is used in the basis step and the others in the induction step. The induction step is established by combining upper and lower bounds, proved in Lemmata 4.8 and 4.9 respectively. Lemma 4.10 is a technical result which is common to both the upper and lower bounds, determining the limit as $N \to \infty$ of a certain expectation that arises in both bounds.

Recall that the following conditions are all consequences of (4.1): for all $t > s > 0$,

$$\mathbb{E}\left[c_N(\tau_N(t))\right] \to 0 \tag{4.10}$$

$$\mathbb{E}\left[\sum_{r=\tau_N(s)+1}^{\tau_N(t)} c_N(r)^2\right] \to 0 \tag{4.11}$$

$$\mathbb{E}\left[\sum_{r=\tau_N(s)+1}^{\tau_N(t)} D_N(r)\right] \to 0 \tag{4.12}$$

as $N \to \infty$. (See Lemmata 3.7, 3.8 and 3.10 for proofs.)

Lemma 4.7 (Basis step)**.** *Assume* (4.1) *holds. Then for any* $0 < t < \infty$*,*

$$\lim_{N\to\infty} \mathbb{E}\left[\prod_{r=1}^{\tau_N(t)} (1 - p_r)\right] = e^{-\alpha_n t}$$

where $\alpha_n := n(n-1)/2$*.*

Proof. We start by showing that $\lim_{N\to\infty} \mathbb{E}\left[\prod_{r=1}^{\tau_N(t)}(1-p_r)\right] \le e^{-\alpha_n t}$.
Setting $\xi = \Delta$ in Proposition 3.4, we have for each r and for sufficiently large N

$$1 - p_r = p_{\Delta\Delta}(r) \le 1 - \alpha_n 1_N \left[c_N(r) - B_n' D_N(r)\right]. \tag{4.13}$$

Recall that 1_N is asymptotic notation for a function that converges to 1 as $N \to \infty$. Since we will eventually take $N \to \infty$, it is sufficient to have bounds that hold for large enough N. However, some of the following manipulations require that these bounds are non-negative. For this reason we introduce some indicator functions (which will be almost surely equal to 1 in the limit) to keep the bounds non-negative. These indicators will later be dropped from certain terms that are clearly non-negative without them. The indicators introduced at this point are such that if their conditions do not hold then the bound becomes the trivial $1 - p_r \le 1$.

When $N \ge 3$, a sufficient condition to ensure that the expression on the right-hand side of (4.13) is non-negative is that the event

$$E_N^1(r) := \left\{c_N(r) < \alpha_n^{-1} A_N\right\} \tag{4.14}$$

occurs, where $A_N = 1_N$ as $N \to \infty$ and is independent of r but will not be specified explicitly. We will also need to control the sign of $c_N(r) - B_n' D_N(r)$, for which we define the event

$$E_N^2(r) := \left\{c_N(r) \ge B_n' D_N(r)\right\}, \tag{4.15}$$

and we define $E_N^1 := \bigcap_{r=1}^{\tau_N(t)} E_N^1(r)$ and $E_N^2 := \bigcap_{r=1}^{\tau_N(t)} E_N^2(r)$. Then

$$1 - p_r = p_{\Delta\Delta}(r) \leq 1 - \alpha_n 1_N \left[c_N(r) - B_n' D_N(r)\right] \mathbb{1}_{E_N^1 \cap E_N^2}.$$

Applying a multinomial expansion and then separating the positive and negative terms,

$$\begin{aligned}
\prod_{r=1}^{\tau_N(t)} (1 - p_r) &\leq 1 + \sum_{l=1}^{\tau_N(t)} (-\alpha_n)^l 1_N \frac{1}{l!} \sum_{\substack{s_1,\ldots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} \left[c_N(s_j) - B_n' D_N(s_j)\right] \mathbb{1}_{E_N^1 \cap E_N^2} \\
&= 1 + \sum_{\substack{l=2 \\ \text{even}}}^{\tau_N(t)} \alpha_n^l 1_N \frac{1}{l!} \sum_{\substack{s_1,\ldots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} \left[c_N(s_j) - B_n' D_N(s_j)\right] \mathbb{1}_{E_N^1 \cap E_N^2} \\
&\quad - \sum_{\substack{l=1 \\ \text{odd}}}^{\tau_N(t)} \alpha_n^l 1_N \frac{1}{l!} \sum_{\substack{s_1,\ldots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} \left[c_N(s_j) - B_n' D_N(s_j)\right] \mathbb{1}_{E_N^1 \cap E_N^2}. \qquad (4.16)
\end{aligned}$$

This is further bounded by applying Lemma 4.6 and then both bounds of Lemma 4.4(b):

$$\begin{aligned}
&\prod_{r=1}^{\tau_N(t)} (1 - p_r) \\
&\quad \leq 1 + \mathbb{1}_{E_N^1 \cap E_N^2} \left\{ \sum_{\substack{l=2 \\ \text{even}}}^{\tau_N(t)} \alpha_n^l 1_N \frac{1}{l!} \sum_{\substack{s_1,\ldots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} c_N(s_j) \right. \\
&\qquad \left. - \sum_{\substack{l=1 \\ \text{odd}}}^{\tau_N(t)} \alpha_n^l 1_N \frac{1}{l!} \left[\sum_{\substack{s_1,\ldots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} c_N(s_j) - \left(\sum_{s=1}^{\tau_N(t)} D_N(s) \right) (t+1)^{l-1} (1 + B_n')^l \right] \right\} \\
&\quad \leq 1 + \left\{ \sum_{\substack{l=2 \\ \text{even}}}^{\tau_N(t)} \alpha_n^l 1_N \frac{1}{l!} \left\{ t^l + c_N(\tau_N(t))(t+1)^l \right\} \right. \\
&\qquad - \sum_{\substack{l=1 \\ \text{odd}}}^{\tau_N(t)} \alpha_n^l 1_N \frac{1}{l!} \left[t^l - \left(\sum_{s=1}^{\tau_N(t)} c_N(s)^2 \right) \binom{l}{2} (t+1)^{l-2} \right] \\
&\qquad \left. - \left(\sum_{s=1}^{\tau_N(t)} D_N(s) \right) (t+1)^{l-1} (1 + B_n')^l \right\} \mathbb{1}_{E_N^1 \cap E_N^2}.
\end{aligned}$$

Collecting some terms,

$$\begin{aligned}
\prod_{r=1}^{\tau_N(t)}(1-p_r) &\le 1+\sum_{l=1}^{\tau_N(t)}(-\alpha_n)^l 1_N \frac{1}{l!}t^l \mathbb{1}_{E_N^1\cap E_N^2} + c_N(\tau_N(t))\sum_{\substack{l=2\\ \text{even}}}^{\tau_N(t)} \alpha_n^l 1_N \frac{1}{l!}(t+1)^l \\
&\quad + \left(\sum_{s=1}^{\tau_N(t)} c_N(s)^2\right)\sum_{\substack{l=1\\ \text{odd}}}^{\tau_N(t)} \alpha_n^l 1_N \frac{1}{l!}\binom{l}{2}(t+1)^{l-2} \\
&\quad + \left(\sum_{s=1}^{\tau_N(t)} D_N(s)\right)\sum_{\substack{l=1\\ \text{odd}}}^{\tau_N(t)} \alpha_n^l 1_N \frac{1}{l!}(t+1)^{l-1}(1+B_n')^l \\
&\le 1+\sum_{l=1}^{\infty}(-\alpha_n)^l 1_N \frac{1}{l!}t^l \mathbb{1}_{\{\tau_N(t)\ge l\}}\mathbb{1}_{E_N^1\cap E_N^2} + c_N(\tau_N(t))\exp[\alpha_n 1_N(t+1)] \\
&\quad + \left(\sum_{s=1}^{\tau_N(t)} c_N(s)^2\right)\frac{1}{2}\alpha_n^2\exp[\alpha_n 1_N(t+1)] \\
&\quad + \left(\sum_{s=1}^{\tau_N(t)} D_N(s)\right)\exp[\alpha_n 1_N(t+1)(1+B_n')]. \qquad (4.17)
\end{aligned}$$

The requirement $\tau_N(t) \ge l$ has been dropped in all but the first term, which constitutes adding some positive terms, giving an upper bound. Now, taking the expectation and limit, then applying (4.10)–(4.12), and using Lemmata 4.12, 4.13 and 4.14 to show that $\lim_{N\to\infty}\mathbb{P}\left[\{\tau_N(t)\ge l\}\cap E_N^1\cap E_N^2\right]=1$,

$$\begin{aligned}
\lim_{N\to\infty}\mathbb{E}\left[\prod_{r=1}^{\tau_N(t)}(1-p_r)\right] &\le 1+\sum_{l=1}^{\infty}(-\alpha_n)^l\frac{1}{l!}t^l \lim_{N\to\infty}\mathbb{P}\left[\{\tau_N(t)\ge l\}\cap E_N^1\cap E_N^2\right] \\
&\quad + \lim_{N\to\infty}\mathbb{E}\left[c_N(\tau_N(t))\right]\exp[\alpha_n(t+1)] \\
&\quad + \lim_{N\to\infty}\mathbb{E}\left[\sum_{s=1}^{\tau_N(t)} c_N(s)^2\right]\frac{1}{2}\alpha_n^2\exp[\alpha_n(t+1)] \\
&\quad + \lim_{N\to\infty}\mathbb{E}\left[\sum_{s=1}^{\tau_N(t)} D_N(s)\right]\exp[\alpha_n(t+1)(1+B_n')] \\
&= 1+\sum_{l=1}^{\infty}(-\alpha_n)^l\frac{1}{l!}t^l = e^{-\alpha_n t}. \qquad (4.18)
\end{aligned}$$

Passing the limit and expectation inside the infinite sum is justified by dominated convergence and Fubini.

It remains to show the corresponding lower bound

$$\lim_{N\to\infty}\mathbb{E}\left[\prod_{r=1}^{\tau_N(t)}(1-p_r)\right] \ge e^{-\alpha_n t}.$$

Setting $\xi = \Delta$ in Proposition 3.3, we have

$$1 - p_t = p_{\Delta\Delta}(t) \geq 1 - \frac{N^{n-2}}{(N-2)_{n-2}} \alpha_n [c_N(t) + B_n D_N(t)] \tag{4.19}$$

where $B_n > 0$. Due to Proposition 3.1((b)), a sufficient condition for this bound to be non-negative is

$$E_N^3(r) := \left\{ c_N(r) \leq \frac{(N-2)_{n-2}}{N^{n-2}} \alpha_n^{-1} (1 + B_n)^{-1} \right\}, \tag{4.20}$$

and we define $E_N^3 := \bigcap_{r=1}^{\tau_N(t)} E_N^3(r)$. Then

$$1 - p_t \geq \left\{ 1 - \frac{N^{n-2}}{(N-2)_{n-2}} \alpha_n [c_N(t) + B_n D_N(t)] \right\} \mathbb{1}_{E_N^3(t)}$$

is also a valid lower bound since if $E_N^3(t)$ does not occur then this collapses to the trivial lower bound $1 - p_t \geq 0$. We now apply a multinomial expansion to the product, and split into positive and negative terms:

$$\begin{aligned}
\prod_{r=1}^{\tau_N(t)} (1 - p_r) &\geq \left\{ 1 + \sum_{l=1}^{\tau_N(t)} (-\alpha_n)^l 1_N \frac{1}{l!} \sum_{\substack{s_1,\ldots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} [c_N(s_j) + B_n D_N(s_j)] \right\} \mathbb{1}_{E_N^3} \\
&= \left\{ 1 + \sum_{\substack{l=2 \\ \text{even}}}^{\tau_N(t)} \alpha_n^l 1_N \frac{1}{l!} \sum_{\substack{s_1,\ldots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} [c_N(s_j) + B_n D_N(s_j)] \right. \\
&\qquad \left. - \sum_{\substack{l=1 \\ \text{odd}}}^{\tau_N(t)} \alpha_n^l 1_N \frac{1}{l!} \sum_{\substack{s_1,\ldots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} [c_N(s_j) + B_n D_N(s_j)] \right\} \mathbb{1}_{E_N^3}
\end{aligned}$$

This is further bounded by applying Lemma 4.5 and both bounds in Lemma 4.4(b):

$$\begin{aligned}
&\prod_{r=1}^{\tau_N(t)}(1-p_r)\\
&\quad\geq \mathbb{1}_{E_N^3}\Bigg\{1+\sum_{\substack{l=2\\ \text{even}}}^{\tau_N(t)}\alpha_n^l 1_N\frac{1}{l!}\sum_{\substack{s_1,\dots,s_l=1\\ \text{all distinct}}}^{\tau_N(t)}\prod_{j=1}^{l}c_N(s_j)\\
&\qquad-\sum_{\substack{l=1\\ \text{odd}}}^{\tau_N(t)}\alpha_n^l 1_N\frac{1}{l!}\left[\sum_{\substack{s_1,\dots,s_l=1\\ \text{all distinct}}}^{\tau_N(t)}\prod_{j=1}^{l}c_N(s_j)+\left(\sum_{s=1}^{\tau_N(t)}D_N(s)\right)(t+1)^{l-1}(1+B_n)^l\right]\Bigg\}\\
&\quad\geq \mathbb{1}_{E_N^3}\Bigg\{1+\sum_{\substack{l=2\\ \text{even}}}^{\tau_N(t)}\alpha_n^l 1_N\frac{1}{l!}\left[t^l-\left(\sum_{s=1}^{\tau_N(t)}c_N(s)^2\right)\binom{l}{2}(t+1)^{l-2}\right]\\
&\qquad-\sum_{\substack{l=1\\ \text{odd}}}^{\tau_N(t)}\alpha_n^l 1_N\frac{1}{l!}\left[t^l+c_N(\tau_N(t))(t+1)^l+\left(\sum_{s=1}^{\tau_N(t)}D_N(s)\right)(t+1)^{l-1}(1+B_n)^l\right]\Bigg\}.
\end{aligned}$$

Collecting terms and dropping indicators from some non-positive terms,

$$\begin{aligned}
\prod_{r=1}^{\tau_N(t)}(1-p_r)&\geq \sum_{l=0}^{\tau_N(t)}(-\alpha_n)^l 1_N\frac{1}{l!}t^l\mathbb{1}_{E_N^3}-\left(\sum_{s=1}^{\tau_N(t)}c_N(s)^2\right)\sum_{\substack{l=2\\ \text{even}}}^{\tau_N(t)}\alpha_n^l 1_N\frac{1}{l!}\binom{l}{2}(t+1)^{l-2}\\
&\quad-c_N(\tau_N(t))\sum_{\substack{l=1\\ \text{odd}}}^{\tau_N(t)}\alpha_n^l 1_N\frac{1}{l!}(t+1)^l\\
&\quad-\left(\sum_{s=1}^{\tau_N(t)}D_N(s)\right)\sum_{\substack{l=1\\ \text{odd}}}^{\tau_N(t)}\alpha_n^l 1_N\frac{1}{l!}(t+1)^{l-1}(1+B_n)^l\\
&\geq \sum_{l=0}^{\infty}(-\alpha_n)^l 1_N\frac{1}{l!}t^l\mathbb{1}_{E_N^3}\mathbb{1}_{\{\tau_N(t)\geq l\}}-\left(\sum_{s=1}^{\tau_N(t)}c_N(s)^2\right)\frac{1}{2}\alpha_n^2\exp[\alpha_n 1_N(t+1)]\\
&\quad-c_N(\tau_N(t))\exp[\alpha_n 1_N(t+1)]\\
&\quad-\left(\sum_{s=1}^{\tau_N(t)}D_N(s)\right)\exp[\alpha_n 1_N(t+1)(1+B_n)]. \qquad (4.21)
\end{aligned}$$

Now, taking the expectation and limit, and applying (4.10)–(4.12) to show that all but the first sum vanish, and Lemmata 4.12 and 4.13 to show that $\lim_{N\to\infty}\mathbb{P}[\{\tau_N(t)\geq l\}\cap E_N^3]=$

1,

$$\begin{aligned}\lim_{N\to\infty}\mathbb{E}\left[\prod_{r=1}^{\tau_N(t)}(1-p_r)\right] &\geq \sum_{l=0}^{\infty}(-\alpha_n)^l 1_N \frac{1}{l!}t^l \lim_{N\to\infty}\mathbb{P}\left[\{\tau_N(t)\geq l\}\cap E_N^3\right] \\ &\quad - \lim_{N\to\infty}\mathbb{E}\left[\sum_{s=1}^{\tau_N(t)} c_N(s)^2\right]\frac{1}{2}\alpha_n^2\exp[\alpha_n(t+1)] \\ &\quad - \lim_{N\to\infty}\mathbb{E}\left[c_N(\tau_N(t))\right]\exp[\alpha_n(t+1)] \\ &\quad - \lim_{N\to\infty}\mathbb{E}\left[\sum_{s=1}^{\tau_N(t)} D_N(s)\right]\exp[\alpha_n(t+1)(1+B_n)] \\ &= \sum_{l=0}^{\infty}(-\alpha_n)^l\frac{1}{l!}t^l = e^{-\alpha_n t}. \end{aligned} \tag{4.22}$$

Again, passing the limit and expectation inside the infinite sum is justified by dominated convergence and Fubini. Combining the upper and lower bounds in (4.18) and (4.22) respectively concludes the proof. ■

Lemma 4.8 (Induction step upper bound)**.** *Assume* (4.1) *holds. Fix* $k\in\mathbb{N}$, $i_0 := 0$, $i_k := k$. *For any sequence of times* $0 = t_0 \leq t_1 \leq \cdots \leq t_k \leq t$,

$$\begin{aligned}\lim_{N\to\infty}\mathbb{E}\left[\sum_{\substack{r_1<\cdots<r_k:\\ r_i\leq\tau_N(t_i)\forall i}}\left(\prod_{i=1}^{k}p_{r_i}\right)\left(\prod_{\substack{r=1\\ \notin\{r_1,\ldots,r_k\}}}^{\tau_N(t)}(1-p_r)\right)\right] \\ \leq \alpha_n^k e^{-\alpha_n t}\sum_{\substack{i_1\leq\cdots\leq i_{k-1}\\ \in\{0,\ldots,k\}:\\ i_j\geq j\forall j}}\prod_{j=1}^{k}\frac{(t_j-t_{j-1})^{i_j-i_{j-1}}}{(i_j-i_{j-1})!}.\end{aligned}$$

Proof. We use the bound on $(1-p_r)$ from (4.13), which holds for sufficiently large N, and apply a multinomial expansion. Define as in (4.14) and (4.15) respectively the sequences

of events E_N^1 and E_N^2 which ensure that the following manipulations make sense:

$$\begin{aligned}
\prod_{\substack{r=1\\ \notin\{r_1,\dots,r_k\}}}^{\tau_N(t)} (1-p_r) &\le \prod_{\substack{r=1\\ \notin\{r_1,\dots,r_k\}}}^{\tau_N(t)} \left\{1-\alpha_n 1_N [c_N(r) - B_n' D_N(r)] \mathbb{1}_{E_N^1 \cap E_N^2}\right\} \\
&= 1 + \sum_{l=1}^{\tau_N(t)-k} (-\alpha_n)^l 1_N \frac{1}{l!} \sum_{\substack{s_1,\dots,s_l=1\\ \notin\{r_1,\dots,r_k\}\\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} [c_N(s_j) - B_n' D_N(s_j)] \mathbb{1}_{E_N^1 \cap E_N^2} \\
&= 1 + \sum_{l=1}^{\tau_N(t)-k} (-\alpha_n)^l 1_N \frac{1}{l!} \sum_{\substack{s_1,\dots,s_l=1\\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} [c_N(s_j) - B_n' D_N(s_j)] \mathbb{1}_{E_N^1 \cap E_N^2} \\
&\quad - \sum_{l=1}^{\tau_N(t)-k} (-\alpha_n)^l 1_N \frac{1}{l!} \sum_{\substack{s_1,\dots,s_l=1\\ \text{all distinct:}\\ \exists i,i': s_i = r_{i'}}}^{\tau_N(t)} \prod_{j=1}^{l} [c_N(s_j) - B_n' D_N(s_j)] \mathbb{1}_{E_N^1 \cap E_N^2}.
\end{aligned} \tag{4.23}$$

The penultimate line above is exactly the expansion we had in the basis step (4.16), except for the limit on l, and as such following the same arguments gives a bound analogous to that in (4.17):

$$\begin{aligned}
1 + \sum_{l=1}^{\tau_N(t)-k} (-\alpha_n)^l 1_N \frac{1}{l!} &\sum_{\substack{s_1,\dots,s_l=1\\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l} [c_N(s_j) - B_n' D_N(s_j)] \mathbb{1}_{E_N^1 \cap E_N^2} \\
&\le 1 + \sum_{l=1}^{\tau_N(t)-k} (-\alpha_n)^l 1_N \frac{1}{l!} t^l \mathbb{1}_{E_N^1 \cap E_N^2} + c_N(\tau_N(t)) \exp[\alpha_n 1_N (t+1)] \\
&\quad + \left(\sum_{s=1}^{\tau_N(t)} c_N(s)^2 \right) \frac{1}{2} \alpha_n^2 \exp[\alpha_n 1_N (t+1)] \\
&\quad + \left(\sum_{s=1}^{\tau_N(t)} D_N(s) \right) \exp[\alpha_n 1_N (t+1)(1+B_n')].
\end{aligned}$$

For the last line of (4.23), recalling that $D_N(t) \le c_N(t)$ (Proposition 3.1(b)),

$$\begin{aligned}
-\sum_{l=1}^{\tau_N(t)-k} (-\alpha_n)^l 1_N \frac{1}{l!} \sum_{\substack{s_1,\dots,s_l=1 \\ \text{all distinct:} \\ \exists i,i': s_i = r_{i'}}}^{\tau_N(t)} \prod_{j=1}^{l} \{c_N(s_j) - B'_n D_N(s_j)\} \mathbb{1}_{E_N^1 \cap E_N^2}
&\le \sum_{l=1}^{\tau_N(t)-k} \alpha_n^l 1_N \frac{1}{l!} \sum_{\substack{s_1,\dots,s_l=1 \\ \text{all distinct:} \\ \exists i,i': s_i = r_{i'}}}^{\tau_N(t)} \prod_{j=1}^{l} \{c_N(s_j) + B'_n D_N(s_j)\} \\
&\le \sum_{l=1}^{\tau_N(t)-k} \alpha_n^l 1_N \frac{1}{l!} \sum_{\substack{s_1,\dots,s_l=1 \\ \text{all distinct:} \\ \exists i,i': s_i = r_{i'}}}^{\tau_N(t)} (1+B'_n)^l \prod_{j=1}^{l} c_N(s_j) \\
&\le \sum_{l=1}^{\tau_N(t)-k} \alpha_n^l 1_N \frac{1}{(l-1)!} \sum_{s_1 \in \{r_1,\dots,r_k\}} \sum_{\substack{s_2,\dots,s_l=1 \\ \text{all distinct}}}^{\tau_N(t)} (1+B'_n)^l \prod_{j=1}^{l} c_N(s_j) \\
&= \sum_{s \in \{r_1,\dots,r_k\}} c_N(s) \sum_{l=1}^{\tau_N(t)-k} \alpha_n^l 1_N \frac{1}{(l-1)!} (1+B'_n)^l \sum_{\substack{s_1,\dots,s_{l-1}=1 \\ \text{all distinct}}}^{\tau_N(t)} \prod_{j=1}^{l-1} c_N(s_j) \\
&\le \sum_{j=1}^{k} c_N(r_j) \sum_{l=1}^{\tau_N(t)-k} \alpha_n^l 1_N \frac{1}{(l-1)!} (1+B'_n)^l (t+1)^{l-1} \\
&\le \left(\sum_{j=1}^{k} c_N(r_j) \right) \alpha_n (1+B'_n) \exp[\alpha_n 1_N (1+B'_n)(t+1)],
\end{aligned}$$

where the penultimate inequality uses Lemma 4.4(a). Putting these together, we have

$$\begin{aligned}
\prod_{\substack{r=1 \\ \notin \{r_1,\dots,r_k\}}}^{\tau_N(t)} (1-p_r) \le 1 &+ \sum_{l=1}^{\tau_N(t)-k} (-\alpha_n)^l 1_N \frac{1}{l!} t^l \mathbb{1}_{E_N^1 \cap E_N^2} + c_N(\tau_N(t)) \exp[\alpha_n 1_N (t+1)] \\
&+ \left(\sum_{s=1}^{\tau_N(t)} c_N(s)^2 \right) \frac{1}{2} \alpha_n^2 \exp[\alpha_n 1_N (t+1)] \\
&+ \left(\sum_{s=1}^{\tau_N(t)} D_N(s) \right) \exp[\alpha_n 1_N (t+1)(1+B'_n)] \\
&+ \left(\sum_{j=1}^{k} c_N(r_j) \right) \alpha_n (1+B'_n) \exp[\alpha_n 1_N (1+B'_n)(t+1)]. \qquad (4.24)
\end{aligned}$$

Meanwhile, using the bound on p_r from (4.19) then applying a modification of Lemma 4.5

where the sum is over ordered indices rather than distinct indices,

$$\begin{aligned}\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\prod_{i=1}^{k}p_{r_i} &\le \alpha_n^k 1_N \sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\prod_{i=1}^{k}[c_N(r_i)+B_nD_N(r_i)]\\ &\le \alpha_n^k 1_N \sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\prod_{i=1}^{k}c_N(r_i)+\left(\sum_{s=1}^{\tau_N(t)}D_N(s)\right)\alpha_n^k 1_N(t+1)^{k-1}(1+B_n)^k.\end{aligned} \tag{4.25}$$

A more liberal but simpler bound can be arrived at thus:

$$\begin{aligned}\prod_{i=1}^{k}p_{r_i} &\le \alpha_n^k 1_N \prod_{i=1}^{k}[c_N(r_i)+B_nD_N(r_i)]\\ &\le \alpha_n^k 1_N \prod_{i=1}^{k}c_N(r_i)(1+B_n)\\ &\le \alpha_n^k 1_N (1+B_n)^k \prod_{i=1}^{k}c_N(r_i)\end{aligned} \tag{4.26}$$

which, using Lemma 4.4(a), also leads to the deterministic bound

$$\begin{aligned}\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\prod_{i=1}^{k}p_{r_i} &\le \alpha_n^k 1_N (1+B_n)^k \sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\prod_{i=1}^{k}c_N(r_i)\\ &\le \alpha_n^k 1_N (1+B_n)^k \frac{1}{k!}\sum_{r_1\ne\cdots\ne r_k}^{\tau_N(t)}\prod_{i=1}^{k}c_N(r_i)\\ &\le \alpha_n^k 1_N (1+B_n)^k \frac{1}{k!}(t+1)^k.\end{aligned} \tag{4.27}$$

Combining this sum-product with (4.24), the expression inside the expectation in Lemma 4.8

is bounded above by

$$\begin{aligned}
&\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\left(\prod_{i=1}^{k}p_{r_i}\right)\left(\prod_{\substack{r=1\\ \notin\{r_1,\ldots,r_k\}}}^{\tau_N(t)}(1-p_r)\right)\\
&\qquad\le\left\{1+\sum_{l=1}^{\tau_N(t)-k}(-\alpha_n)^l 1_N\frac{1}{l!}t^l\mathbb{1}_{E_N^1\cap E_N^2}\right\}\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\prod_{i=1}^{k}p_{r_i}\\
&\qquad\quad+\left\{c_N(\tau_N(t))\exp[\alpha_n 1_N(t+1)]+\left(\sum_{s=1}^{\tau_N(t)}c_N(s)^2\right)\frac{1}{2}\alpha_n^2\exp[\alpha_n 1_N(t+1)]\right.\\
&\qquad\qquad\left.+\left(\sum_{s=1}^{\tau_N(t)}D_N(s)\right)\exp[\alpha_n 1_N(t+1)(1+B_n')]\right\}\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\prod_{i=1}^{k}p_{r_i}\\
&\qquad\quad+\exp[\alpha_n 1_N(1+B_n')(t+1)]\alpha_n(1+B_n')\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\sum_{j=1}^{k}c_N(r_j)\prod_{i=1}^{k}p_{r_i}.
\end{aligned}$$

Applying (4.25) to the first term, (4.27) to the second term and (4.26) to the third term, we have

$$\begin{aligned}
&\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\left(\prod_{i=1}^{k}p_{r_i}\right)\left(\prod_{\substack{r=1\\ \notin\{r_1,\ldots,r_k\}}}^{\tau_N(t)}(1-p_r)\right)\\
&\quad\le\alpha_n^k 1_N\left\{1+\sum_{l=1}^{\tau_N(t)-k}(-\alpha_n)^l 1_N\frac{1}{l!}t^l\mathbb{1}_{E_N^1\cap E_N^2}\right\}\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\prod_{i=1}^{k}c_N(r_i)\\
&\qquad+\left(\sum_{s=1}^{\tau_N(t)}D_N(s)\right)\alpha_n^k 1_N(t+1)^{k-1}(1+B_n)^k\sum_{l=0}^{\tau_N(t)}(\alpha_n)^l 1_N\frac{1}{l!}t^l\\
&\qquad+\left\{c_N(\tau_N(t))\exp[\alpha_n 1_N(t+1)]+\left(\sum_{s=1}^{\tau_N(t)}c_N(s)^2\right)\frac{1}{2}\alpha_n^2\exp[\alpha_n 1_N(t+1)]\right.\\
&\qquad\qquad\left.+\left(\sum_{s=1}^{\tau_N(t)}D_N(s)\right)\exp[\alpha_n 1_N(t+1)(1+B_n')]\right\}\alpha_n^k 1_N(1+B_n)^k\frac{1}{k!}(t+1)^k\\
&\qquad+\exp[\alpha_n(1+B_n')(t+1)]\alpha_n(1+B_n')\alpha_n^k 1_N(1+B_n)^k\\
&\qquad\qquad\times\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\sum_{j=1}^{k}c_N(r_j)\prod_{i=1}^{k}c_N(r_i).
\end{aligned}$$

Upon taking the expectation and limit, we have

$$\begin{aligned}
&\lim_{N\to\infty}\mathbb{E}\left[\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\left(\prod_{i=1}^{k}p_{r_i}\right)\left(\prod_{\substack{r=1\\ \notin\{r_1,\ldots,r_k\}}}^{\tau_N(t)}(1-p_r)\right)\right]\\
&\quad\le\alpha_n^k\lim_{N\to\infty}\mathbb{E}\left[\left(1+\sum_{l=1}^{\tau_N(t)-k}(-\alpha_n)^l\frac{1}{l!}t^l\mathbb{1}_{E_N^1\cap E_N^2}\right)\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\prod_{i=1}^{k}c_N(r_i)\right]\\
&\qquad+\lim_{N\to\infty}\mathbb{E}\left[\sum_{s=1}^{\tau_N(t)}D_N(s)\right]\alpha_n^k(t+1)^{k-1}(1+B_n)^k\exp[\alpha_n t]\\
&\qquad+\Bigg\{\lim_{N\to\infty}\mathbb{E}\left[c_N(\tau_N(t))\right]\exp[\alpha_n(t+1)]\\
&\qquad\qquad+\lim_{N\to\infty}\mathbb{E}\left[\sum_{s=1}^{\tau_N(t)}c_N(s)^2\right]\frac{1}{2}\alpha_n^2\exp[\alpha_n(t+1)]\\
&\qquad\qquad+\lim_{N\to\infty}\mathbb{E}\left[\sum_{s=1}^{\tau_N(t)}D_N(s)\right]\exp[\alpha_n(t+1)(1+B_n')]\Bigg\}\alpha_n^k(1+B_n)^k\frac{1}{k!}(t+1)^k\\
&\qquad+\exp[\alpha_n(1+B_n')(t+1)]\alpha_n^{k+1}(1+B_n')(1+B_n)^k\\
&\qquad\qquad\times\lim_{N\to\infty}\mathbb{E}\left[\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\sum_{j=1}^{k}c_N(r_j)\prod_{i=1}^{k}c_N(r_i)\right].
\end{aligned}$$

The middle terms vanish due to (4.10)–(4.12) and the expression becomes

$$\begin{aligned}
&\lim_{N\to\infty}\mathbb{E}\left[\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\left(\prod_{i=1}^{k}p_{r_i}\right)\left(\prod_{\substack{r=1\\ \notin\{r_1,\ldots,r_k\}}}^{\tau_N(t)}(1-p_r)\right)\right]\le\alpha_n^k\lim_{N\to\infty}\mathbb{E}\left[\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\prod_{i=1}^{k}c_N(r_i)\right]\\
&\qquad+\alpha_n^k\sum_{l=1}^{\infty}(-\alpha_n)^l\frac{1}{l!}t^l\lim_{N\to\infty}\mathbb{E}\left[\mathbb{1}_{\{\tau_N(t)\ge k+l\}}\mathbb{1}_{E_N^1\cap E_N^2}\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\prod_{i=1}^{k}c_N(r_i)\right]\\
&\qquad+\exp[\alpha_n(1+B_n')(t+1)]\alpha_n^{k+1}(1+B_n')(1+B_n)^k\\
&\qquad\qquad\times\lim_{N\to\infty}\mathbb{E}\left[\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\sum_{j=1}^{k}c_N(r_j)\prod_{i=1}^{k}c_N(r_i)\right],
\end{aligned}\tag{4.28}$$

where passing the limit and expectation inside the infinite sum is justified by dominated

convergence and Fubini; see Lemma 4.16. To simplify the last line,

$$\begin{aligned}\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\sum_{j=1}^{k}c_N(r_j)\prod_{i=1}^{k}c_N(r_i) &\le \frac{1}{k!}\sum_{\substack{r_1,\ldots,r_k\\ \text{all distinct}}}^{\tau_N(t)}\sum_{j=1}^{k}c_N(r_j)\prod_{i=1}^{k}c_N(r_i)\\ &= \frac{1}{k!}\sum_{\substack{r_1,\ldots,r_k\\ \text{all distinct}}}^{\tau_N(t)}\sum_{j=1}^{k}c_N(r_j)^2\prod_{i\ne j}c_N(r_i)\\ &\le \frac{1}{k!}\sum_{j=1}^{k}\sum_{s=1}^{\tau_N(t)}c_N(s)^2\sum_{\substack{r_1,\ldots,r_{k-1}\\ \text{all distinct}}}^{\tau_N(t)}\prod_{i=1}^{k-1}c_N(r_i)\\ &\le \frac{1}{(k-1)!}\sum_{s=1}^{\tau_N(t)}c_N(s)^2(t+1)^{k-1},\end{aligned}$$

using Lemma 4.4(a) for the final inequality. Hence

$$\lim_{N\to\infty}\mathbb{E}\left[\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\sum_{s\in\{r_1,\ldots,r_k\}}c_N(s)\prod_{i=1}^{k}c_N(r_i)\right]\le\frac{1}{(k-1)!}(t+1)^{k-1}\lim_{N\to\infty}\mathbb{E}\left[\sum_{s=1}^{\tau_N(t)}c_N(s)^2\right]$$

which equals 0 by (4.11). By Lemmata 4.12, 4.13 and 4.14, $\lim_{N\to\infty}\mathbb{P}[\{\tau_N(t)\ge k+l\}\cap E_N^1\cap E_N^2]=1$, so we can apply Lemma 4.10 to the remaining expectations in (4.28), yielding

$$\begin{aligned}&\lim_{N\to\infty}\mathbb{E}\left[\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\left(\prod_{i=1}^{k}p_{r_i}\right)\left(\prod_{\substack{r=1\\ \notin\{r_1,\ldots,r_k\}}}^{\tau_N(t)}(1-p_r)\right)\right]\\ &\qquad\le\alpha_n^k\sum_{l=0}^{\infty}(-\alpha_n)^l\frac{1}{l!}t^l\sum_{\substack{i_1\le\cdots\le i_{k-1}\\ \in\{0,\ldots,k\}:\\ i_j\ge j\forall j}}\prod_{j=1}^{k}\frac{(t_j-t_{j-1})^{i_j-i_{j-1}}}{(i_j-i_{j-1})!}\\ &\qquad=\alpha_n^k e^{-\alpha_n t}\sum_{\substack{i_1\le\cdots\le i_{k-1}\\ \in\{0,\ldots,k\}:\\ i_j\ge j\forall j}}\prod_{j=1}^{k}\frac{(t_j-t_{j-1})^{i_j-i_{j-1}}}{(i_j-i_{j-1})!}\end{aligned}$$

as required. ■

Lemma 4.9 (Induction step lower bound)**.** *Assume* (4.1) *holds. Fix* $k \in \mathbb{N}$, $i_0 := 0$, $i_k := k$. *For any sequence of times* $0 = t_0 \le t_1 \le \cdots \le t_k \le t$,

$$\lim_{N\to\infty} \mathbb{E}\left[\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}} \left(\prod_{i=1}^{k} p_{r_i}\right)\left(\prod_{\substack{r=1\\ \notin\{r_1,\ldots,r_k\}}}^{\tau_N(t)} (1-p_r)\right)\right] \ge \alpha_n^k e^{-\alpha_n t} \sum_{\substack{i_1\le\cdots\le i_{k-1}\\ \in\{0,\ldots,k\}:\\ i_j\ge j\forall j}} \prod_{j=1}^{k} \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!}.$$

Proof. Firstly,

$$\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}} \left(\prod_{i=1}^{k} p_{r_i}\right)\left(\prod_{\substack{r=1\\ \notin\{r_1,\ldots,r_k\}}}^{\tau_N(t)} (1-p_r)\right) \ge \sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}} \left(\prod_{i=1}^{k} p_{r_i}\right)\left(\prod_{r=1}^{\tau_N(t)} (1-p_r)\right). \quad (4.29)$$

Now the second product does not depend on $r_1, \ldots, r_k$, and we can use the lower bound from (4.21):

$$\begin{aligned}\prod_{r=1}^{\tau_N(t)} (1-p_r) \ge \sum_{l=0}^{\tau_N(t)} (-\alpha_n)^l 1_N \frac{1}{l!} t^l \mathbb{1}_{E_N^3} &- \left(\sum_{s=1}^{\tau_N(t)} c_N(s)^2\right) \frac{1}{2}\alpha_n^2 \exp[\alpha_n 1_N (t+1)] \\ &- c_N(\tau_N(t)) \exp[\alpha_n 1_N (t+1)] \\ &- \left(\sum_{s=1}^{\tau_N(t)} D_N(s)\right) \exp[\alpha_n 1_N (t+1)(1+B_n)] \end{aligned} \quad (4.30)$$

where E_N^3 is defined as in (4.20). We will also need an upper bound on this product, which

is formed from (4.17) with a further deterministic bound:

$$\begin{aligned}\prod_{r=1}^{\tau_N(t)}(1-p_r) &\le \sum_{l=0}^{\tau_N(t)}(-\alpha_n)^l 1_N \frac{1}{l!} t^l \mathbb{1}_{\{\tau_N(t)\ge l\}} \mathbb{1}_{E_N^1\cap E_N^2} + c_N(\tau_N(t))\exp[\alpha_n 1_N(t+1)] \\ &\quad + \left(\sum_{s=1}^{\tau_N(t)} c_N(s)^2\right)\frac{1}{2}\alpha_n^2 \exp[\alpha_n 1_N(t+1)] \\ &\quad + \left(\sum_{s=1}^{\tau_N(t)} D_N(s)\right)\exp[\alpha_n 1_N(t+1)(1+B_n')] \\ &\le \exp[\alpha_n 1_N t] + \exp[\alpha_n 1_N(t+1)] \\ &\quad + \frac{1}{2}\alpha_n^2(t+1)\exp[\alpha_n 1_N(t+1)] + (t+1)\exp[\alpha_n 1_N(t+1)(1+B_n')] \\ &\le \left(2+\frac{\alpha_n^2(t+1)}{2}\right)\exp[\alpha_n 1_N(t+1)] + (t+1)\exp[\alpha_n 1_N(t+1)(1+B_n')].\end{aligned} \tag{4.31}$$

Now let us consider the remaining sum-product on the right-hand side of (4.29). We use the same bound on p_r as in (4.13):

$$p_r = 1 - p_{\Delta\Delta}(r) \ge \alpha_n 1_N \left[c_N(r) - B_n' D_N(r)\right] \tag{4.32}$$

where the 1_N term does not depend on r. When N is large enough for the factor of 1_N to be non-negative, the condition that the bound in (4.32) is non-negative holds on the event E_N^2 that was defined in (4.15). Then

$$\prod_{i=1}^{k} p_{r_i} \ge \alpha_n^k 1_N \prod_{i=1}^{k}\left[c_N(r_i) - B_n' D_N(r_i)\right]\mathbb{1}_{E_N^2}.$$

Applying a modification of Lemma 4.6 where the sum is over ordered indices rather than distinct indices,

$$\begin{aligned}\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\prod_{i=1}^{k} p_{r_i} &\ge \alpha_n^k 1_N \sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\prod_{i=1}^{k}\left[c_N(r_i) - B_n' D_N(r_i)\right]\mathbb{1}_{E_N^2} \\ &\ge \alpha_n^k 1_N \Bigg\{ \sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\prod_{i=1}^{k} c_N(r_i)\mathbb{1}_{E_N^2} \\ &\qquad - \frac{1}{k!}\left(\sum_{s=1}^{\tau_N(t)} D_N(s)\right)(t+1)^{k-1}(1+B_n')^k\Bigg\}.\end{aligned}$$

The above expression is already split into positive and negative terms; a lower bound on (4.29) can be formed by multiplying the positive terms by the lower bound (4.30) and the

negative terms by the upper bound (4.31). Thus

$$\begin{aligned}
&\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}} \left(\prod_{i=1}^{k} p_{r_i}\right)\left(\prod_{\substack{r=1\\ \notin\{r_1,\dots,r_k\}}}^{\tau_N(t)} (1-p_r)\right) \\
&\qquad \ge \alpha_n^k 1_N \sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}} \prod_{i=1}^{k} c_N(r_i)\mathbb{1}_{E_N^2}\Bigg\{ \sum_{l=0}^{\tau_N(t)} (-\alpha_n)^l 1_N \frac{1}{l!} t^l \mathbb{1}_{E_N^3} \\
&\qquad\qquad - \left(\sum_{s=1}^{\tau_N(t)} c_N(s)^2\right)\frac{1}{2}\alpha_n^2 \exp[\alpha_n 1_N(t+1)] \\
&\qquad\qquad - c_N(\tau_N(t))\exp[\alpha_n 1_N(t+1)] \\
&\qquad\qquad - \left(\sum_{s=1}^{\tau_N(t)} D_N(s)\right)\exp[\alpha_n 1_N(t+1)(1+B_n)]\Bigg\} \\
&\qquad - \left(\sum_{s=1}^{\tau_N(t)} D_N(s)\right)\alpha_n^k 1_N \frac{1}{k!}(t+1)^{k-1}(1+B_n')^k \Bigg\{ \\
&\qquad\qquad \left(2+\frac{\alpha_n^2(t+1)}{2}\right)\exp[\alpha_n 1_N(t+1)] \\
&\qquad\qquad + (t+1)\exp[\alpha_n 1_N(t+1)(1+B_n')]\Bigg\}.
\end{aligned}$$

Due to (4.10)–(4.12), all but the first line on the right-hand side of the above have vanishing expectation, leaving

$$\begin{aligned}
&\lim_{N\to\infty}\mathbb{E}\left[\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}} \left(\prod_{i=1}^{k} p_{r_i}\right)\left(\prod_{\substack{r=1\\ \notin\{r_1,\dots,r_k\}}}^{\tau_N(t)} (1-p_r)\right)\right] \\
&\qquad \ge \lim_{N\to\infty}\mathbb{E}\left[\alpha_n^k 1_N \sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}} \prod_{i=1}^{k} c_N(r_i)\mathbb{1}_{E_N^2} \sum_{l=0}^{\tau_N(t)} (-\alpha_n)^l 1_N \frac{1}{l!} t^l \mathbb{1}_{E_N^3}\right] \\
&\qquad = \alpha_n^k \sum_{l=0}^{\infty}(-\alpha_n)^l \frac{1}{l!} t^l \lim_{N\to\infty}\mathbb{E}\left[\mathbb{1}_{\{\tau_N(t)\ge l\}}\mathbb{1}_{E_N^2\cap E_N^3} \sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}} \prod_{i=1}^{k} c_N(r_i)\right].
\end{aligned} \tag{4.33}$$

Passing the limit and expectation inside the infinite sum is justified by dominated convergence and Fubini; see Lemma 4.16. Lemmata 4.12 and 4.14 establish that $\lim_{N\to\infty}\mathbb{P}[E_N^2\cap E_N^3] = 1$ and Lemma 4.13 deals with the other indicator. We can therefore apply

Lemma 4.10 to conclude that

$$\begin{aligned}\lim_{N\to\infty}\mathbb{E}\left[\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\left(\prod_{i=1}^{k}p_{r_i}\right)\left(\prod_{\substack{r=1\\ \notin\{r_1,\ldots,r_k\}}}^{\tau_N(t)}(1-p_r)\right)\right]&\\ \ge\alpha_n^k\sum_{l=0}^{\infty}(-\alpha_n)^l\frac{1}{l!}t^l\sum_{\substack{i_1\le\cdots\le i_{k-1}\\ \in\{0,\ldots,k\}:\\ i_j\ge j\forall j}}\prod_{j=1}^{k}\frac{(t_j-t_{j-1})^{i_j-i_{j-1}}}{(i_j-i_{j-1})!}&\\ =\alpha_n^k e^{-\alpha_n t}\sum_{\substack{i_1\le\cdots\le i_{k-1}\\ \in\{0,\ldots,k\}:\\ i_j\ge j\forall j}}\prod_{j=1}^{k}\frac{(t_j-t_{j-1})^{i_j-i_{j-1}}}{(i_j-i_{j-1})!}&\end{aligned}$$

as required. ■

Lemma 4.10. *Assume* (4.1) *holds. Fix* $k\in\mathbb{N}$, $i_0:=0$, $i_k:=k$. *Let* E_N *be a sequence of events such that* $\lim_{N\to\infty}\mathbb{P}[E_N]=1$. *Then for any sequence of times* $0=t_0\le t_1\le\cdots\le t_k\le t$,

$$\lim_{N\to\infty}\mathbb{E}\left[\mathbb{1}_{E_N}\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\prod_{i=1}^{k}c_N(r_i)\right]=\sum_{\substack{i_1\le\cdots\le i_{k-1}\\ \in\{0,\ldots,k\}:\\ i_j\ge j\forall j}}\prod_{j=1}^{k}\frac{(t_j-t_{j-1})^{i_j-i_{j-1}}}{(i_j-i_{j-1})!}.$$

Proof. As pointed out by Möhle (1999, p.460), the sum-product on the left hand side can be expanded as

$$\sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}}\prod_{i=1}^{k}c_N(r_i)=\sum_{\substack{i_1\le\cdots\le i_{k-1}\\ \in\{0,\ldots,k\}:\\ i_j\ge j\forall j}}\prod_{j=1}^{k}\frac{1}{(i_j-i_{j-1})!}\sum_{\substack{r_{i_{j-1}+1},\ldots,r_{i_j}\\ =\tau_N(t_{j-1})+1\\ \text{all distinct}}}^{\tau_N(t_j)}\prod_{i=i_{j-1}+1}^{i_j}c_N(r_i).$$

By a modification of the upper bound in Lemma 4.4(b) where the lower limit of the sum is a general time rather than 1,

$$\sum_{\substack{r_{i_{j-1}+1},\ldots,r_{i_j}\\ =\tau_N(t_{j-1})+1\\ \text{all distinct}}}^{\tau_N(t_j)}\prod_{i=i_{j-1}+1}^{i_j}c_N(r_i)\le(t_j-t_{j-1})^{i_j-i_{j-1}}+c_N(\tau_N(t_j))(t_j-t_{j-1}+1)^{i_j-i_{j-1}}$$

Now, taking the product on the outside,

$$\begin{aligned}
&\prod_{j=1}^{k} \frac{1}{(i_j - i_{j-1})!} \sum_{\substack{r_{i_{j-1}+1},\dots,r_{i_j} \\ =\tau_N(t_{j-1})+1 \\ \text{all distinct}}}^{\tau_N(t_j)} \prod_{i=i_{j-1}+1}^{i_j} c_N(r_i) \\
&\qquad \le \prod_{j=1}^{k} \left\{ \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!} + c_N(\tau_N(t_j)) \frac{(t_j - t_{j-1} + 1)^{i_j - i_{j-1}}}{(i_j - i_{j-1})!} \right\} \\
&\qquad \le \prod_{j=1}^{k} \left\{ \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!} + c_N(\tau_N(t_j))(t_j - t_{j-1} + 1)^{i_j - i_{j-1}} \right\} \\
&\qquad = \sum_{\mathcal{I} \subseteq [k]} \left(\prod_{j \in \mathcal{I}} \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!} \right) \left(\prod_{j \notin \mathcal{I}} c_N(\tau_N(t_j))(t_j - t_{j-1} + 1)^{i_j - i_{j-1}} \right).
\end{aligned}$$

Separating the term where $\mathcal{I} = [k]$, this becomes

$$\begin{aligned}
&\prod_{j=1}^{k} \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!} \\
&\qquad + \sum_{\mathcal{I} \subset [k]} \left(\prod_{j \in \mathcal{I}} \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!} \right) \left(\prod_{j \notin \mathcal{I}} c_N(\tau_N(t_j))(t_j - t_{j-1} + 1)^{i_j - i_{j-1}} \right) \\
&\le \prod_{j=1}^{k} \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!} + \sum_{\mathcal{I} \subset [k]} \left(\prod_{j \in \mathcal{I}} t^{i_j - i_{j-1}} \right) \left(\prod_{j \notin \mathcal{I}} c_N(\tau_N(t_j))(t+1)^{i_j - i_{j-1}} \right) \\
&\le \prod_{j=1}^{k} \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!} + \sum_{\mathcal{I} \subset [k]} c_N(\tau_N(t_{j^\star(\mathcal{I})})) \prod_{j=1}^{k} (t+1)^{i_j - i_{j-1}} \\
&= \prod_{j=1}^{k} \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!} + \sum_{\mathcal{I} \subset [k]} c_N(\tau_N(t_{j^\star(\mathcal{I})}))(t+1)^k
\end{aligned}$$

where, say, $j^\star(\mathcal{I}) := \min\{j \notin \mathcal{I}\}$. Now we are in a position to evaluate the desired limit:

$$\begin{aligned}
&\lim_{N \to \infty} \mathbb{E} \left[\mathbb{1}_{E_N} \sum_{\substack{r_1 < \dots < r_k: \\ r_i \le \tau_N(t_i) \forall i}} \prod_{i=1}^{k} c_N(r_i) \right] \le \lim_{N \to \infty} \mathbb{E} \left[\sum_{\substack{r_1 < \dots < r_k: \\ r_i \le \tau_N(t_i) \forall i}} \prod_{i=1}^{k} c_N(r_i) \right] \\
&\qquad \le \sum_{\substack{i_1 \le \dots \le i_{k-1} \\ \in \{0,\dots,k\}: \\ i_j \ge j \forall j}} \prod_{j=1}^{k} \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!} + \sum_{\substack{i_1 \le \dots \le i_{k-1} \\ \in \{0,\dots,k\}: \\ i_j \ge j \forall j}} \sum_{\mathcal{I} \subset [k]} \lim_{N \to \infty} \mathbb{E} \left[c_N(\tau_N(t_{j^\star(\mathcal{I})})) \right] (t+1)^k \\
&\qquad = \sum_{\substack{i_1 \le \dots \le i_{k-1} \\ \in \{0,\dots,k\}: \\ i_j \ge j \forall j}} \prod_{j=1}^{k} \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!}
\end{aligned}$$

using (4.10). For the corresponding lower bound, by a modification of the lower bound in Lemma 4.4(b) where the lower limit of the sum is a general time rather than 1,

$$\begin{aligned}
&\sum_{\substack{r_{i_{j-1}+1},\ldots,r_{i_j}=\tau_N(t_{j-1})+1\\ \text{all distinct}}}^{\tau_N(t_j)} \prod_{i=i_{j-1}+1}^{i_j} c_N(r_i) \\
&\quad \geq (t_j - t_{j-1})^{i_j - i_{j-1}} - \binom{i_j - i_{j-1}}{2} \left(\sum_{s=\tau_N(t_{j-1})+1}^{\tau_N(t_j)} c_N(s)^2 \right) (t_j - t_{j-1} + 1)^{i_j - i_{j-1} - 2} \\
&\quad \geq (t_j - t_{j-1})^{i_j - i_{j-1}} - (i_j - i_{j-1})! \left(\sum_{s=\tau_N(t_{j-1})+1}^{\tau_N(t_j)} c_N(s)^2 \right) (t_j - t_{j-1} + 1)^{i_j - i_{j-1} - 2}.
\end{aligned}$$

Define the events

$$E_N^4(j) = \left\{ \left(\sum_{s=\tau_N(t_{j-1})+1}^{\tau_N(t_j)} c_N(s)^2 \right) \leq \frac{1}{(i_j - i_{j-1})!} \left(\frac{t_j - t_{j-1}}{t_j - t_{j-1} + 1} \right)^{i_j - i_{j-1}} \right\},$$

which is sufficient to ensure the j^{th} term in the following product is non-negative, and define $E_N^4 := \bigcap_{j=1}^k E_N^4(j)$. If $t_j = t_{j-1}$ then $E_N^4(j)$ has probability one automatically; otherwise the constant on the right is strictly positive and so satisfies the conditions of Lemma 4.15. Now, taking a product over j,

$$\begin{aligned}
&\prod_{j=1}^k \frac{1}{(i_j - i_{j-1})!} \sum_{\substack{r_{i_{j-1}+1},\ldots,r_{i_j}=\tau_N(t_{j-1})+1\\ \text{all distinct}}}^{\tau_N(t_j)} \prod_{i=i_{j-1}+1}^{i_j} c_N(r_i) \\
&\quad \geq \prod_{j=1}^k \left\{ \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!} - \left(\sum_{s=\tau_N(t_{j-1})+1}^{\tau_N(t_j)} c_N(s)^2 \right) (t_j - t_{j-1} + 1)^{i_j - i_{j-1} - 2} \right\} \mathbb{1}_{E_N^4} \\
&\quad = \sum_{\mathcal{I} \subseteq [k]} (-1)^{k - |\mathcal{I}|} \left(\prod_{j \in \mathcal{I}} \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!} \right) \\
&\qquad \times \left(\prod_{j \notin \mathcal{I}} \left(\sum_{s=\tau_N(t_{j-1})+1}^{\tau_N(t_j)} c_N(s)^2 \right) (t_j - t_{j-1} + 1)^{i_j - i_{j-1} - 2} \right) \mathbb{1}_{E_N^4}.
\end{aligned}$$

Separating the term with $\mathcal{I} = [k]$, this becomes

$$\begin{aligned}
&\prod_{j=1}^{k} \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!} \mathbb{1}_{E_N^4} \\
&\quad + \sum_{\mathcal{I} \subset [k]} (-1)^{k-|\mathcal{I}|} \left(\prod_{j \in \mathcal{I}} \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!} \right) \\
&\qquad\qquad \times \left(\prod_{j \notin \mathcal{I}} \left(\sum_{s=\tau_N(t_{j-1})+1}^{\tau_N(t_j)} c_N(s)^2 \right) (t_j - t_{j-1} + 1)^{i_j - i_{j-1} - 2} \right) \mathbb{1}_{E_N^4} \\
&\geq \prod_{j=1}^{k} \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!} \mathbb{1}_{E_N^4} \\
&\quad - \sum_{\mathcal{I} \subset [k]} \left(\prod_{j \in \mathcal{I}} \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!} \right) \\
&\qquad\qquad \times \left(\prod_{j \notin \mathcal{I}} \left(\sum_{s=\tau_N(t_{j-1})+1}^{\tau_N(t_j)} c_N(s)^2 \right) (t_j - t_{j-1} + 1)^{i_j - i_{j-1} - 2} \right) \\
&\geq \prod_{j=1}^{k} \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!} \mathbb{1}_{E_N^4} \\
&\quad - \sum_{\mathcal{I} \subset [k]} \left(\prod_{j \in \mathcal{I}} t^{i_j - i_{j-1}} \right) \left(\prod_{j \notin \mathcal{I}} \left(\sum_{s=\tau_N(t_{j-1})+1}^{\tau_N(t_j)} c_N(s)^2 \right) (t+1)^{i_j - i_{j-1} - 2} \right).
\end{aligned}$$

Using parts (a) and (d) of Proposition 3.1 to upper bound all but one of the $\sum c_N(s)^2$ terms, and arbitrarily setting $j^\star(\mathcal{I}) := \min\{j \notin \mathcal{I}\}$, this is further bounded by

$$\begin{aligned}
&\prod_{j=1}^{k} \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!} \mathbb{1}_{E_N^4} \\
&\quad - \sum_{\mathcal{I} \subset [k]} \left(\sum_{s=\tau_N(t_{j^\star(\mathcal{I})-1})+1}^{\tau_N(t_{j^\star(\mathcal{I})})} c_N(s)^2 \right) \left(\prod_{j \in \mathcal{I}} t^{i_j - i_{j-1}} \right) \left(\prod_{j \notin \mathcal{I}} (t+1)^{i_j - i_{j-1} - 1} \right) \\
&\geq \prod_{j=1}^{k} \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!} \mathbb{1}_{E_N^4} - \sum_{\mathcal{I} \subset [k]} \left(\sum_{s=\tau_N(t_{j^\star(\mathcal{I})-1})+1}^{\tau_N(t_{j^\star(\mathcal{I})})} c_N(s)^2 \right) \prod_{j=1}^{k} (t+1)^{i_j - i_{j-1}} \\
&= \prod_{j=1}^{k} \frac{(t_j - t_{j-1})^{i_j - i_{j-1}}}{(i_j - i_{j-1})!} \mathbb{1}_{E_N^4} - \sum_{\mathcal{I} \subset [k]} \left(\sum_{s=\tau_N(t_{j^\star(\mathcal{I})-1})+1}^{\tau_N(t_{j^\star(\mathcal{I})})} c_N(s)^2 \right) (t+1)^k.
\end{aligned}$$

We can now evaluate the limit:

$$
\begin{aligned}
&\lim_{N\to\infty} \mathbb{E}\left[\mathbb{1}_{E_N} \sum_{\substack{r_1<\cdots<r_k:\\ r_i\le\tau_N(t_i)\forall i}} \prod_{i=1}^{k} c_N(r_i)\right] \\
&\quad\ge \lim_{N\to\infty} \mathbb{E}\left[\mathbb{1}_{E_N\cap E_N^4} \sum_{\substack{i_1\le\cdots\le i_{k-1}\\ \in\{0,\dots,k\}:\\ i_j\ge j\forall j}} \prod_{j=1}^{k} \frac{(t_j-t_{j-1})^{i_j-i_{j-1}}}{(i_j-i_{j-1})!}\right] \\
&\qquad - \lim_{N\to\infty} \mathbb{E}\left[\mathbb{1}_{E_N} \sum_{\substack{i_1\le\cdots\le i_{k-1}\\ \in\{0,\dots,k\}:\\ i_j\ge j\forall j}} \sum_{\mathcal{I}\subset[k]} \left(\sum_{s=\tau_N(t_{j^\star(\mathcal{I})-1})+1}^{\tau_N(t_{j^\star(\mathcal{I})})} c_N(s)^2\right)(t+1)^k\right] \\
&\quad\ge \sum_{\substack{i_1\le\cdots\le i_{k-1}\\ \in\{0,\dots,k\}:\\ i_j\ge j\forall j}} \prod_{j=1}^{k} \frac{(t_j-t_{j-1})^{i_j-i_{j-1}}}{(i_j-i_{j-1})!} \lim_{N\to\infty} \mathbb{E}\left[\mathbb{1}_{E_N\cap E_N^4}\right] \\
&\qquad - \lim_{N\to\infty} \mathbb{E}\left[\sum_{\substack{i_1\le\cdots\le i_{k-1}\\ \in\{0,\dots,k\}:\\ i_j\ge j\forall j}} \sum_{\mathcal{I}\subset[k]} \left(\sum_{s=\tau_N(t_{j^\star(\mathcal{I})-1})+1}^{\tau_N(t_{j^\star(\mathcal{I})})} c_N(s)^2\right)(t+1)^k\right] \\
&\quad= \sum_{\substack{i_1\le\cdots\le i_{k-1}\\ \in\{0,\dots,k\}:\\ i_j\ge j\forall j}} \prod_{j=1}^{k} \frac{(t_j-t_{j-1})^{i_j-i_{j-1}}}{(i_j-i_{j-1})!} \lim_{N\to\infty} \mathbb{P}\left[E_N\cap E_N^4\right] \\
&\qquad - \sum_{\substack{i_1\le\cdots\le i_{k-1}\\ \in\{0,\dots,k\}:\\ i_j\ge j\forall j}} \sum_{\mathcal{I}\subset[k]} \lim_{N\to\infty} \mathbb{E}\left[\sum_{s=\tau_N(t_{j^\star(\mathcal{I})-1})+1}^{\tau_N(t_{j^\star(\mathcal{I})})} c_N(s)^2\right](t+1)^k \\
&\quad= \sum_{\substack{i_1\le\cdots\le i_{k-1}\\ \in\{0,\dots,k\}:\\ i_j\ge j\forall j}} \prod_{j=1}^{k} \frac{(t_j-t_{j-1})^{i_j-i_{j-1}}}{(i_j-i_{j-1})!}
\end{aligned}
$$

where for the last equality we use (4.11) to show that the second sum vanishes and Lemma 4.15 to show that $\lim_{N\to\infty}\mathbb{P}[E_N\cap E_N^4]=1$. We have shown that the upper and lower bounds coincide, so the proof is complete. ■

4.3 Indicators

Many of the preceding results make use of indicator functions in order to control the sign of certain terms. It was claimed that the probabilities of the corresponding events converge to 1 as $N \to \infty$, so that the indicators do not have any effect in the limit. These claims are proved in this section. The first result (Lemma 4.11) shows that it is sufficient to prove the limits separately for each event, even if we are actually taking a product of indicators on two or more events. The remainder of this section is split into four more lemmata, each of which deals with events of a certain form, showing that their probabilities converge to 1 as $N \to \infty$.

Lemma 4.11. *Let $(A_N), (B_N)$ be sequences of events. If $\lim_{N\to\infty} \mathbb{P}[A_N] = 1$ and $\lim_{N\to\infty} \mathbb{P}[B_N] = 1$ then $\lim_{N\to\infty} \mathbb{P}[A_N \cap B_N] = 1$.*

Proof.

$$\begin{aligned}
&\lim_{N\to\infty} \mathbb{P}[A_N] = 1 \textbf{ and } \lim_{N\to\infty} \mathbb{P}[B_N] = 1 \\
\Leftrightarrow &\lim_{N\to\infty} \mathbb{P}[A_N^c] = 0 \textbf{ and } \lim_{N\to\infty} \mathbb{P}[B_N^c] = 0 \\
\Rightarrow &\lim_{N\to\infty} \{\mathbb{P}[A_N^c] + \mathbb{P}[B_N^c]\} = 0 \\
\Rightarrow &\lim_{N\to\infty} \mathbb{P}[A_N^c \cup B_N^c] = 0 \\
\Leftrightarrow &\lim_{N\to\infty} \mathbb{P}[A_N \cap B_N] = 1.
\end{aligned}$$

■

Lemma 4.12. *Assume (4.11) holds. Fix $t > 0$. Let $K > 0$ be a constant which may depend on n, N but not on r, such that $K^{-2} = O(1)$ as $N \to \infty$. Define the events $E_N(r) := \{c_N(r) < K\}$ and denote $E_N := \bigcap_{r=1}^{\tau_N(t)} E_N(r)$. Then $\lim_{N\to\infty} \mathbb{P}[E_N] = 1$.*

Proof.

$$\begin{aligned}
\mathbb{P}[E_N] = 1 - \mathbb{P}[E_N^c] = 1 - \mathbb{P}\left[\bigcup_{r=1}^{\tau_N(t)} E_N^c(r)\right] &= 1 - \mathbb{E}\left[\mathbb{1}_{\bigcup E_N^c(r)}\right] \geq 1 - \mathbb{E}\left[\sum_{r=1}^{\tau_N(t)} \mathbb{1}_{E_N^c(r)}\right] \\
= 1 - \mathbb{E}\left[\sum_{r=1}^{\tau_N(t)} \mathbb{E}\left[\mathbb{1}_{E_N^c(r)} \,\middle|\, \mathcal{F}_{r-1}\right]\right] &= 1 - \mathbb{E}\left[\sum_{r=1}^{\tau_N(t)} \mathbb{P}[E_N^c(r) \mid \mathcal{F}_{r-1}]\right] \qquad (4.34)
\end{aligned}$$

where for the second line we apply Lemma 3.2 with $f(r) = \mathbb{1}_{E_N^c(r)}$. By the generalised Markov inequality,

$$\mathbb{P}[E_N^c(r) \mid \mathcal{F}_{r-1}] = \mathbb{P}[c_N(r) \geq K \mid \mathcal{F}_{r-1}] \leq K^{-2}\mathbb{E}[c_N(r)^2 \mid \mathcal{F}_{r-1}].$$

Substituting this into (4.34) and applying Lemma 3.2 again, this time with $f(r) = c_N(r)^2$,

$$\mathbb{P}[E_N] \geq 1 - K^{-2}\mathbb{E}\left[\sum_{r=1}^{\tau_N(t)} \mathbb{E}[c_N(r)^2 \mid \mathcal{F}_{r-1}]\right] = 1 - K^{-2}\mathbb{E}\left[\sum_{r=1}^{\tau_N(t)} c_N(r)^2\right].$$

Applying (4.11), the limit is

$$\lim_{N\to\infty} \mathbb{P}[E_N] = 1 - O(1) \times 0 = 1$$

as required. ■

Lemma 4.13. *Fix $t > 0$. For any $l \in \mathbb{N}$, $\lim_{N\to\infty} \mathbb{P}[\tau_N(t) \geq l] = 1$.*

Proof. We can replace the event $\{\tau_N(t) \geq l\}$ with an event of the form of E_N in Lemma 4.12:

$$\{\tau_N(t) \geq l\} = \left\{\min\left\{s \geq 1 : \sum_{r=1}^{s} c_N(r) \geq t\right\} \geq l\right\} = \left\{\sum_{r=1}^{l-1} c_N(r) < t\right\}$$
$$\supseteq \bigcap_{r=1}^{l-1}\left\{c_N(r) < \frac{t}{l}\right\} \supseteq \bigcap_{r=1}^{\tau_N(l)}\left\{c_N(r) < \frac{t}{l}\right\}$$

since $\tau_N(l) \geq l$ (Proposition 3.1(f)). Hence

$$\lim_{N\to\infty} \mathbb{P}[\tau_N(t) \geq l] \geq \lim_{N\to\infty} \mathbb{P}\left[\bigcap_{r=1}^{\tau_N(l)}\left\{c_N(r) < \frac{t}{l}\right\}\right] = 1$$

by applying Lemma 4.12 with $K = t/l$. ■

Lemma 4.14. *Assume (4.12) holds. Fix $t > 0$. Let K be a constant not depending on N, r, but which may depend on n.*

$$\lim_{N\to\infty} \mathbb{P}\left[\bigcap_{r=1}^{\tau_N(t)} \{c_N(r) \geq KD_N(r)\}\right] = 1.$$

Proof.

$$\begin{aligned}
\mathbb{P}\left[\bigcap_{r=1}^{\tau_N(t)}\{c_N(r)\geq KD_N(r)\}\right] &\geq \mathbb{P}\left[\bigcap_{r=1}^{\tau_N(t)}\{c_N(r)>KD_N(r)\}\right] \\
&= 1-\mathbb{P}\left[\bigcup_{r=1}^{\tau_N(t)}\{c_N(r)\leq KD_N(r)\}\right] \\
&= 1-\mathbb{E}\left[\mathbb{1}_{\bigcup\{c_N(r)\leq KD_N(r)\}}\right] \\
&\geq 1-\mathbb{E}\left[\sum_{r=1}^{\tau_N(t)}\mathbb{1}_{\{c_N(r)\leq KD_N(r)\}}\right] \\
&= 1-\mathbb{E}\left[\sum_{r=1}^{\tau_N(t)}\mathbb{P}[c_N(r)\leq KD_N(r)\mid\mathcal{F}_{r-1}]\right] \qquad (4.35)
\end{aligned}$$

where the final inequality is an application of Lemma 3.2 with $f(r)=\mathbb{1}_{\{c_N(r)\leq KD_N(r)\}}$.

Fix $0<\epsilon<1/(2K)$ and assume $N>\max\{\epsilon^{-1},(K^{-1}-2\epsilon)^{-1}\}$. For each r,i define the event $A_i(r):=\{\nu_r^{(i)}\leq N\epsilon\}$. We have (almost surely)

$$\begin{aligned}
D_N(r) = &\frac{1}{N(N)_2}\sum_{i=1}^{N}(\nu_r^{(i)})_2\left[\nu_r^{(i)}+\frac{1}{N}\sum_{j\neq i}(\nu_r^{(j)})^2\right]\mathbb{1}_{A_i^c(r)} \\
&+\frac{1}{N(N)_2}\sum_{i=1}^{N}(\nu_r^{(i)})_2\left[\nu_r^{(i)}+\frac{1}{N}\sum_{j\neq i}(\nu_r^{(j)})^2\right]\mathbb{1}_{A_i(r)}.
\end{aligned}$$

For the first term,

$$\begin{aligned}
&\frac{1}{N(N)_2}\sum_{i=1}^{N}(\nu_r^{(i)})_2\left[\nu_r^{(i)}+\frac{1}{N}\sum_{j\neq i}(\nu_r^{(j)})^2\right]\mathbb{1}_{A_i^c(r)} \\
&\quad=\sum_{i=1}^{N}\mathbb{1}_{A_i^c(r)}\left\{\frac{1}{N(N)_2}(\nu_r^{(i)})_2\left[\nu_r^{(i)}+\frac{1}{N}\sum_{j\neq i}(\nu_r^{(j)})^2\right]\right\} \\
&\quad\leq\sum_{i=1}^{N}\mathbb{1}_{A_i^c(r)}D_N(r)\leq\sum_{i=1}^{N}\mathbb{1}_{A_i^c(r)}.
\end{aligned}$$

For the second term,

$$\begin{aligned}
&\frac{1}{N(N)_2}\sum_{i=1}^{N}(\nu_r^{(i)})_2\left[\nu_r^{(i)}+\frac{1}{N}\sum_{j\neq i}(\nu_r^{(j)})^2\right]\mathbb{1}_{A_i(r)}\\
&\qquad\leq \frac{1}{N(N)_2}\sum_{i=1}^{N}(\nu_r^{(i)})_2\nu_r^{(i)}\mathbb{1}_{A_i(r)}+\frac{1}{N^2(N)_2}\sum_{i=1}^{N}(\nu_r^{(i)})_2\sum_{j=1}^{N}(\nu_r^{(j)})^2\mathbb{1}_{A_i(r)}\\
&\qquad\leq \frac{1}{N}c_N(r)N\epsilon+\frac{1}{N^2(N)_2}\sum_{i=1}^{N}(\nu_r^{(i)})_2\sum_{j=1}^{N}(\nu_r^{(j)})_2\mathbb{1}_{A_i(r)}\\
&\qquad\quad+\frac{1}{N^2(N)_2}\sum_{i=1}^{N}(\nu_r^{(i)})_2\sum_{j=1}^{N}(\nu_r^{(j)})\mathbb{1}_{A_i(r)}\\
&\qquad\leq \epsilon c_N(r)+\frac{1}{N^2}\sum_{i=1}^{N}\nu_r^{(i)}N\epsilon c_N(r)+\frac{1}{N^2}c_N(r)N\\
&\qquad= c_N(r)\left(2\epsilon+\frac{1}{N}\right).
\end{aligned}$$

Altogether we have

$$D_N(r)\leq c_N(r)\left(2\epsilon+\frac{1}{N}\right)+\sum_{i=1}^{N}\mathbb{1}_{A_i^c(r)}.$$

Hence

$$\begin{aligned}
\{c_N(r)\leq KD_N(r)\}&\subseteq\left\{c_N(r)\leq Kc_N(r)(2\epsilon+N^{-1})+K\sum_{i=1}^{N}\mathbb{1}_{A_i^c(r)}\right\}\\
&=\left\{K^{-1}-2\epsilon-\frac{1}{N}\leq\sum_{i=1}^{N}\frac{\mathbb{1}_{A_i^c(r)}}{c_N(r)}\right\}
\end{aligned}$$

where the ratio $\mathbb{1}_{A_i^c(r)}/c_N(r)$ is well-defined because

$$A_i^c(r)\Rightarrow c_N(r):=\frac{1}{(N)_2}\sum_{j=1}^{N}(\nu_r^{(j)})_2\geq\frac{1}{(N)_2}(\nu_r^{(i)})_2\geq\frac{\epsilon(N\epsilon-1)}{N-1}\geq\epsilon\left(\epsilon-\frac{1}{N}\right)>0.$$

Hence by Markov's inequality (the conditions on ϵ, N ensuring the constant is always

strictly positive),

$$\begin{aligned}\mathbb{P}\left[c_N(r) \le KD_N(r) \mid \mathcal{F}_{r-1}\right] &\le \mathbb{P}\left[\sum_{i=1}^{N} \mathbb{1}_{A_i^c(r)} \ge \left(K^{-1} - 2\epsilon - \frac{1}{N}\right)\epsilon\left(\epsilon - \frac{1}{N}\right)\middle| \mathcal{F}_{r-1}\right] \\ &\le \frac{1}{\left(K^{-1} - 2\epsilon - \frac{1}{N}\right)\epsilon\left(\epsilon - \frac{1}{N}\right)}\mathbb{E}\left[\sum_{i=1}^{N} \mathbb{1}_{A_i^c(r)}\middle| \mathcal{F}_{r-1}\right] \\ &\le \frac{1}{\left(K^{-1} - 2\epsilon - \frac{1}{N}\right)\epsilon\left(\epsilon - \frac{1}{N}\right)}\mathbb{E}\left[\sum_{i=1}^{N} \frac{(\nu_r^{(i)})_3}{(N\epsilon)_3}\middle| \mathcal{F}_{r-1}\right] \\ &\le \frac{1}{\left(K^{-1} - 2\epsilon - \frac{1}{N}\right)\epsilon\left(\epsilon - \frac{1}{N}\right)}\mathbb{E}\left[\frac{N(N)_2}{(N\epsilon)_3} D_N(r)\middle| \mathcal{F}_{r-1}\right].\end{aligned}$$

Applying Lemma 3.2 once more, with $f(r) = D_N(r)$,

$$\begin{aligned}&\mathbb{E}\left[\sum_{r=1}^{\tau_N(t)} \mathbb{P}[c_N(r) \le KD_N(r) \mid \mathcal{F}_{r-1}]\right] \\ &\qquad \le \frac{1}{\left(K^{-1} - 2\epsilon - \frac{1}{N}\right)\epsilon\left(\epsilon - \frac{1}{N}\right)}\frac{N(N)_2}{(N\epsilon)_3}\mathbb{E}\left[\sum_{r=1}^{\tau_N(t)} \mathbb{E}[D_N(r) \mid \mathcal{F}_{r-1}]\right] \\ &\qquad = \frac{1}{\left(K^{-1} - 2\epsilon - \frac{1}{N}\right)\epsilon\left(\epsilon - \frac{1}{N}\right)}\frac{N(N)_2}{(N\epsilon)_3}\mathbb{E}\left[\sum_{r=1}^{\tau_N(t)} D_N(r)\right] \\ &\qquad \underset{N\to\infty}{\longrightarrow} \frac{1}{(K^{-1} - 2\epsilon)\epsilon^5} \times 0 = 0\end{aligned}$$

due to (4.12). Substituting this back into (4.35) concludes the proof. ∎

Lemma 4.15. *Assume* (4.11) *holds. Fix $k \in \mathbb{N}$, a sequence of times $0 = t_0 \le t_1 \le \cdots \le t_k \le t$, and let $K_1, \ldots, K_k$ be strictly positive constants. Define the event*

$$E_N := \bigcap_{j=1}^{k} \left\{ \sum_{s=\tau_N(t_{j-1})+1}^{\tau_N(t_j)} c_N(s)^2 \le K_j \right\}.$$

Then $\lim_{N\to\infty} \mathbb{P}[E_N] = 1$.

Proof.

$$\begin{aligned}\mathbb{P}[E_N] = 1 - \mathbb{P}[E_N^c] &= 1 - \mathbb{P}\left[\bigcup_{j=1}^{k} \left\{ \sum_{s=\tau_N(t_{j-1})+1}^{\tau_N(t_j)} c_N(s)^2 > K_j \right\}\right] \\ &\ge 1 - \sum_{j=1}^{k} \mathbb{P}\left[\sum_{s=\tau_N(t_{j-1})+1}^{\tau_N(t_j)} c_N(s)^2 \ge K_j\right].\end{aligned}$$

Applying Markov's inequality,

$$\mathbb{P}[E_N] \geq 1 - \sum_{j=1}^{k} K_j^{-1} \mathbb{E}\left[\sum_{s=\tau_N(t_{j-1})+1}^{\tau_N(t_j)} c_N(s)^2\right] \underset{N\to\infty}{\longrightarrow} 1 - \sum_{j=1}^{k} K_j^{-1} \times 0 = 1$$

by (4.11). ■

4.4 Fubini & dominated convergence conditions

There are a few instances where Fubini's Theorem and the Dominated Convergence Theorem are needed in order to pass a limit and expectation through an infinite sum. Now we verify the conditions of these theorems. This result, analogous to that in Koskela et al. (2018, p.24), is used once in Lemma 4.8 at (4.28) and once in Lemma 4.9 at (4.33).

Lemma 4.16. *For any fixed $t > 0$, for N sufficiently large,*

$$\mathbb{E}\left[\sum_{l=0}^{\infty}\left|(-\alpha_n)^l 1_N \frac{1}{l!} t^l \sum_{\substack{r_1<\cdots<r_k:\\ r_i\leq\tau_N(t_i)\forall i}} \prod_{i=1}^{k} c_N(r_i)\right|\right] < \infty.$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\sum_{l=0}^{\infty}\left|(-\alpha_n)^l 1_N \frac{1}{l!} t^l \sum_{\substack{r_1<\cdots<r_k:\\ r_i\leq\tau_N(t_i)\forall i}} \prod_{i=1}^{k} c_N(r_i)\right|\right] &= \mathbb{E}\left[\sum_{l=0}^{\infty} \alpha_n^l 1_N \frac{1}{l!} t^l \sum_{\substack{r_1<\cdots<r_k:\\ r_i\leq\tau_N(t_i)\forall i}} \prod_{i=1}^{k} c_N(r_i)\right] \\
&\leq \mathbb{E}\left[\sum_{l=0}^{\infty} \alpha_n^l 1_N \frac{1}{l!} t^l \sum_{r_1\neq\cdots\neq r_k}^{\tau_N(t)} \prod_{i=1}^{k} c_N(r_i)\right] \\
&\leq \mathbb{E}\left[\sum_{l=0}^{\infty} \alpha_n^l 1_N \frac{1}{l!} t^l (t+1)^k\right] \\
&= \mathbb{E}[\exp\{\alpha_n t 1_N\}(t+1)^k] \\
&= \exp\{\alpha_n t 1_N\}(t+1)^k < \infty
\end{aligned}$$

for N sufficiently large, where the bound on the sum-product comes from Lemma 4.4(a). ■

5 Applications

Earth's crammed with heaven,
And every common bush afire with God,
But only he who sees takes off his shoes;
The rest sit round and pluck blackberries.

Elizabeth Barrett Browning

Theorem 4.1 gives verifiable conditions under which interacting particle systems with dynamics in the form of Algorithm 2.1 have asymptotically Kingman genealogies. The work was motivated by SMC algorithms, which have the required form. However, certain choices of state space and dynamics within the context of Algorithm 2.1 yield systems that are not very SMC-like but may have applications in other fields such as population genetics. For instance, we have generally required that the resampling scheme is unbiased, but this is by no means necessary for Theorem 4.1 (or indeed Theorem 3.6); it is just that biased resampling schemes are of little use in SMC.

The applications presented in this chapter are all motivated by SMC, but an interesting area of future research would be to explore the implications of Theorem 4.1 in other contexts. From the population genetics point-of-view, Theorem 4.1 may be seen as a complement to the convergence criteria for neutral models discussed in Section 2.2.4, so it would be interesting to construct some corollaries for classical non-neutral population models.

For many of the following results it will be necessary to compute filtered expectations $\mathbb{E}_t[\cdot] \equiv \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$, which are generally difficult to compute directly. To simplify the computations we introduce a sequence of σ-algebras $(\mathcal{H}_t)$, defined below, such that filtered expectations can be written in terms of conditional expectations given $\mathcal{H}_t$.

Figure 5.1 shows a section of the conditional dependence graph implied by Algorithm 2.1, as in Figure 2.2, except that time is now labelled in reverse. The σ-algebra

$$\mathcal{H}_t := \sigma(X_{t-1}^{(1:N)}, X_t^{(1:N)}, w_{t-1}^{(1:N)}, w_t^{(1:N)}) \tag{5.1}$$

at each time t forms a separatrix, in the sense of d-separation (Verma and Pearl 1988), between the parental indices $a_t^{(1:N)}$ and the previous σ-algebra $\mathcal{F}_{t-1}$ in the filtration.

That is, $a_t^{(1:N)}$ is conditionally independent of $\mathcal{F}_{t-1}$ given $\mathcal{H}_t$. The practical upshot of this is that we can use the tower rule along with conditional independence to write filtered expectations as

$$\mathbb{E}_t\left[f(\nu_t^{(1:N)})\right] = \mathbb{E}_t\left[\mathbb{E}[f(\nu_t^{(1:N)}) \mid \mathcal{H}_t, \mathcal{F}_{t-1}]\right] = \mathbb{E}_t\left[\mathbb{E}[f(\nu_t^{(1:N)}) \mid \mathcal{H}_t]\right]. \tag{5.2}$$

As we will see, this enables us to compute bounds on the filtered expectations of interest relatively easily.

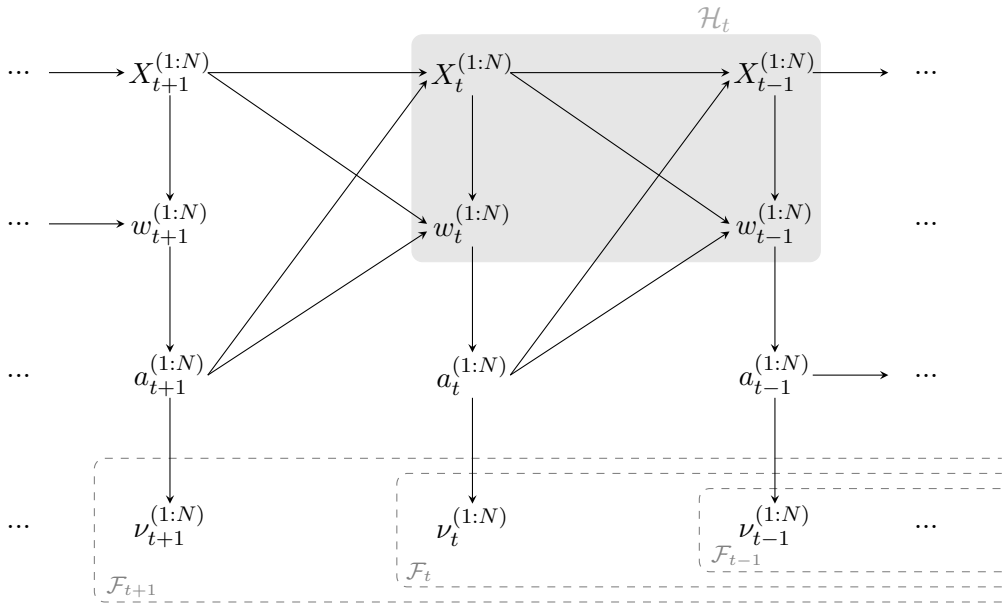


Figure 5.1: Part of the conditional dependence graph implied by Algorithm 2.1 illustrating the construction of $\mathcal{H}_t$. The direction of time is from left to right. The reverse-time filtration is indicated by the dashed areas. The nodes highlighted in grey generate the separatrix $\mathcal{H}_t$ between $a_t^{(1:N)}$ and $\mathcal{F}_{t-1}$.

5.1 Multinomial resampling

Multinomial resampling is often preferred in theoretical studies of SMC, because it renders the parental indices conditionally i.i.d. given the weights, making it relatively simple to analyse the resulting algorithm. The convergence of finite-dimensional distributions for multinomial resampling was proved in Koskela et al. (2018, Corollary 1), but we are now able to prove an analogous weak convergence result. The following proof also demonstrates the relative ease with which we can verify the conditions of Theorem 3.6 as opposed to those of Koskela et al. (2018, Theorem 1).

Corollary 5.1. *Consider an SMC algorithm using multinomial resampling, such that* (A1) *is satisfied. Assume there exist constants* $\varepsilon \in (0, 1], a \in [1, \infty)$ *and probability density* h *such that for all* x, x', t,

$$\frac{1}{a} \leq g_t(x, x') \leq a, \quad \varepsilon h(x') \leq q_t(x, x') \leq \frac{1}{\varepsilon} h(x'). \tag{5.3}$$

Let $(G_t^{(n,N)})_{t \geq 0}$ *denote the genealogy of a random sample of size* n *among the* N *terminal particles in the output of the algorithm. Then, for any fixed* n*, the time-scaled genealogy* $(G_{\tau_N(t)}^{(n,N)})_{t \geq 0}$ *converges weakly to Kingman's* n*-coalescent as* $N \to \infty$.

The bounds on g_t and q_t in (5.3) are rather strong; they can only reasonably be expected to hold if the state space is compact. However, they are widespread in the literature, where they are known as the *strong mixing conditions* (Del Moral 2004, Section 3.5.2), because they greatly facilitate the theoretical analysis of SMC algorithms. It is often possible to relax these conditions at the expense of considerable technical complication. The conditions on g_t in (5.3) ensure that the weights are all $O(N^{-1})$, none of them being too close to zero or one. Together with the bounds on q_t, this is enough to control the relative rate of multiple mergers, as seen in the following proof.

Proof. Define $\mathcal{H}_t$ as in (5.1). Conditional on $\mathcal{H}_t$ the parental indices are independent, with conditional law

$$\mathbb{P}\left[a_t^{(i)} = a_i \,\middle|\, \mathcal{H}_t\right] \propto g_t(X_{t+1}^{a_{t+1}^{(a_i)}}, X_t^{(a_i)}) q_{t-1}(X_t^{(a_i)}, X_{t-1}^{(i)}) \tag{5.4}$$

for each i, so the joint law is

$$\mathbb{P}\left[a_t^{(1:N)} = a_{1:N} \,\middle|\, \mathcal{H}_t\right] \propto \prod_{i=1}^{N} g_t(X_{t+1}^{a_{t+1}^{(a_i)}}, X_t^{(a_i)}) q_{t-1}(X_t^{(a_i)}, X_{t-1}^{(i)}).$$

Using the bounds (5.3) and the balls-in-bins coupling of Koskela et al. (2018, Proof of Lemma 3), we can obtain bounds on expectations of functions of $a_t^{(1:N)}$. For any $k \in \mathbb{N}$ the function $a_t^{(1:N)} \to (\nu_t^{(i)})_k$ is $\{i\}$-increasing in the sense of Koskela et al. (2018, p.19), so we may apply the bounds

$$\mathbb{E}\left[(V_L^{(i)})_k\right] \leq \mathbb{E}\left[(\nu_t^{(i)})_k \,\middle|\, \mathcal{H}_t\right] \leq \mathbb{E}\left[(V_U^{(i)})_k\right],$$

where

$$V_L^{(i)} \overset{d}{=} \text{Binomial}\left(N, \frac{\varepsilon/a}{(\varepsilon/a) + (N-1)(a/\varepsilon)}\right),$$
$$V_U^{(i)} \overset{d}{=} \text{Binomial}\left(N, \frac{a/\varepsilon}{(a/\varepsilon) + (N-1)(\varepsilon/a)}\right).$$

independently for each i and independently of $\mathcal{F}_\infty$. Furthermore, using the moments of

the Binomial distribution (see for example Mosimann 1962, p.67)

$$\mathbb{E}\left[(V_L^{(i)})_k\right] = (N)_k \left(\frac{\varepsilon/a}{(\varepsilon/a)+(N-1)(a/\varepsilon)}\right)^k \geq (N)_k \left(\frac{\varepsilon/a}{N(a/\varepsilon)}\right)^k = \frac{(N)_k}{N^k}\frac{\varepsilon^{2k}}{a^{2k}}.$$

Similarly,

$$\mathbb{E}\left[(V_U^{(i)})_k\right] \leq \frac{(N)_k}{N^k}\frac{a^{2k}}{\varepsilon^{2k}}.$$

We therefore have the bounds

$$\frac{(N)_k}{N^k}\frac{\varepsilon^{2k}}{a^{2k}} \leq \mathbb{E}\left[(\nu_t^{(i)})_k \,\middle|\, \mathcal{H}_t\right] \leq \frac{(N)_k}{N^k}\frac{a^{2k}}{\varepsilon^{2k}}.$$

for each $k \in \mathbb{N}$. Consequently,

$$\frac{1}{(N)_2}\sum_{i=1}^{N} \mathbb{E}\left[(\nu_t^{(i)})_2 \,\middle|\, \mathcal{H}_t\right] \geq \frac{\varepsilon^4}{Na^4} \tag{5.5}$$

and

$$\frac{1}{(N)_3}\sum_{i=1}^{N} \mathbb{E}\left[(\nu_t^{(i)})_3 \,\middle|\, \mathcal{H}_t\right] \leq \frac{a^6}{N^2\varepsilon^6}. \tag{5.6}$$

Applying (5.2) to (5.5) and (5.6) we find

$$\frac{\frac{1}{(N)_3}\sum_{i=1}^{N}\mathbb{E}_t[(\nu_t^{(i)})_3]}{\frac{1}{(N)_2}\sum_{i=1}^{N}\mathbb{E}_t[(\nu_t^{(i)})_2]} \leq \frac{a^6/(N^2\varepsilon^6)}{\varepsilon^4/(Na^4)} = \frac{a^{10}}{N\varepsilon^{10}} =: b_N \underset{N\to\infty}{\longrightarrow} 0.$$

Thus (4.1) is satisfied. It remains to show that, for N sufficiently large, $\mathbb{P}[\tau_N(t) = \infty] = 0$ for all finite t, a technicality which is proved in Lemma 5.2. Applying Theorem 4.1 then yields the result. ■

Lemma 5.2. *Consider an SMC algorithm using multinomial resampling, satisfying* (A1) *and* (5.3). *Then, for all $N > 2$, $\mathbb{P}[\tau_N(t) = \infty] = 0$ for all finite t.*

Proof. Since $c_N(t) \in [0,1]$ almost surely and has strictly positive expectation, for any fixed N the distribution of $c_N(t)$ with given expectation that maximises $\mathbb{P}[c_N(t) = 0 \mid \mathcal{F}_{t-1}]$ is two atoms, at 0 and 1 respectively. To ensure the correct expectation, the atom at 1 should have mass $\mathbb{P}[c_N(t) = 1 \mid \mathcal{F}_{t-1}] = \mathbb{E}_t[c_N(t)]$, which is bounded below by (5.5). If $c_N(t) > 0$ then $c_N(t) \geq 2/(N)_2 > 2/N^2$. Hence, in general $\mathbb{P}[c_N(t) > 2/N^2 \mid \mathcal{F}_{t-1}] \geq \mathbb{E}_t[c_N(t)]$. Applying (5.5) along with (5.2), we have for any finite N

$$\sum_{t=0}^{\infty}\mathbb{P}[c_N(t) > 2/N^2 \mid \mathcal{F}_{t-1}] \geq \sum_{t=0}^{\infty}\mathbb{E}_t[c_N(t)] \geq \sum_{t=0}^{\infty}\frac{\varepsilon^4}{Na^4} = \infty.$$

By a filtered version of the second Borel–Cantelli lemma (see for example Durrett 2019, Theorem 4.3.4), this implies that $c_N(t) > 2/N^2$ for infinitely many t, almost surely. This

ensures, for all $t < \infty$, that $\mathbb{P}\left[\exists s < \infty : \sum_{r=1}^{s} c_N(r) \geq t\right] = 1$, which by definition of $\tau_N(t)$ is equivalent to $\mathbb{P}[\tau_N(t) = \infty] = 0$. ■

5.2 Stratified resampling

Corollary 5.3. *Consider an SMC algorithm using stratified resampling, such that* (A1) *is satisfied. Assume that there exists a constant* $a \in [1, \infty)$ *such that for all* x, x', t,

$$\frac{1}{a} \leq g_t(x, x') \leq a. \tag{5.7}$$

Assume that $\mathbb{P}[\tau_N(t) = \infty] = 0$ *for all finite* t. *Let* $(G_t^{(n,N)})_{t \geq 0}$ *denote the genealogy of a random sample of size* n *among the* N *terminal particles in the output of the algorithm. Then, for any fixed* n, *the time-scaled genealogy* $(G_{\tau_N(t)}^{(n,N)})_{t \geq 0}$ *converges weakly to Kingman's* n*-coalescent as* $N \to \infty$.

Stratified resampling is, by design, much more restrictive than multinomial resampling. Once the weights are known there is little freedom in the offspring counts, so it is not surprising that control over the weights such as (5.7) provides is sufficient without any additional control over the transition densities q_t. Indeed the transition kernels need not even admit densities. This is in contrast to multinomial resampling (Corollary 5.1), where g_t and q_t are more or less on an equal footing, and we require both to be bounded.

It is not immediately clear that the finite time scale condition $\mathbb{P}[\tau_N(t) = \infty] = 0$ holds under conditions (5.7), so it is included in the statement of the corollary. Proposition 5.6 presents some sufficient conditions for the finite time scale, but these are by no means necessary.

Proof. Define the σ-algebras $\mathcal{H}_t$ as in (5.1). With stratified resampling, conditional on the weights each offspring count almost surely takes one of four values: $\nu_t^{(i)} \in \{\lfloor N w_t^{(i)} \rfloor - 1, \lfloor N w_t^{(i)} \rfloor, \lfloor N w_t^{(i)} \rfloor + 1, \lfloor N w_t^{(i)} \rfloor + 2\}$. Define for each $k \in \mathbb{Z}$

$$p_k^{(i)} := \mathbb{P}\left[\nu_t^{(i)} = \lfloor N w_t^{(i)} \rfloor + k \,\middle|\, \mathcal{H}_t\right]. \tag{5.8}$$

Then $p_k^{(i)} \equiv 0$ for $k \notin \{-1, 0, 1, 2\}$. Now

$$\begin{aligned}\mathbb{E}\left[(\nu_t^{(i)})_2 \,\middle|\, \mathcal{H}_t\right] = p_{-1}^{(i)}(\lfloor N w_t^{(i)} \rfloor - 1)_2 + p_0^{(i)}(\lfloor N w_t^{(i)} \rfloor)_2 + p_1^{(i)}(\lfloor N w_t^{(i)} \rfloor + 1)_2 \\ + p_2^{(i)}(\lfloor N w_t^{(i)} \rfloor + 2)_2\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left[(\nu_t^{(i)})_3 \,\middle|\, \mathcal{H}_t\right] &= p_{-1}^{(i)}(\lfloor Nw_t^{(i)}\rfloor - 1)_3 + p_0^{(i)}(\lfloor Nw_t^{(i)}\rfloor)_3 + p_1^{(i)}(\lfloor Nw_t^{(i)}\rfloor + 1)_3 \\
&\qquad + p_2^{(i)}(\lfloor Nw_t^{(i)}\rfloor + 2)_3 \\
&= p_{-1}^{(i)}(\lfloor Nw_t^{(i)}\rfloor - 3)(\lfloor Nw_t^{(i)}\rfloor - 1)_2 + p_0^{(i)}(\lfloor Nw_t^{(i)}\rfloor - 2)(\lfloor Nw_t^{(i)}\rfloor)_2 \\
&\qquad + p_1^{(i)}(\lfloor Nw_t^{(i)}\rfloor - 1)(\lfloor Nw_t^{(i)}\rfloor + 1)_2 + p_2^{(i)}\lfloor Nw_t^{(i)}\rfloor(\lfloor Nw_t^{(i)}\rfloor + 2)_2 \\
&\leq \lfloor Nw_t^{(i)}\rfloor\Big\{p_{-1}^{(i)}(\lfloor Nw_t^{(i)}\rfloor - 1)_2 + p_0^{(i)}(\lfloor Nw_t^{(i)}\rfloor)_2 + p_1^{(i)}(\lfloor Nw_t^{(i)}\rfloor + 1)_2 \\
&\qquad + p_2^{(i)}(\lfloor Nw_t^{(i)}\rfloor + 2)_2\Big\} \\
&= \lfloor Nw_t^{(i)}\rfloor\mathbb{E}\left[(\nu_t^{(i)})_2 \,\middle|\, \mathcal{H}_t\right] \\
&\leq a^2\mathbb{E}\left[(\nu_t^{(i)})_2 \,\middle|\, \mathcal{H}_t\right]. \qquad (5.9)
\end{aligned}$$

The last line uses the almost sure bound $w_t^{(i)} \leq a^2/N$ which follows from (5.7) along with the form of the weights in Algorithm 2.1. Some terms in the above expressions may be equal to zero when $w_t^{(i)}$ is small enough, but the bound still holds in these cases. Since (5.9) holds for all i, applying the tower rule as in (5.2) we have

$$\frac{1}{(N)_3}\sum_{i=1}^{N}\mathbb{E}_t\left[(\nu_t^{(i)})_3\right] \leq \frac{a^2}{N-2}\frac{1}{(N)_2}\sum_{i=1}^{N}\mathbb{E}_t\left[(\nu_t^{(i)})_2\right],$$

satisfying (4.1) with $b_N := a^2/(N-2) \to 0$. The result follows by applying Theorem 4.1. ■

Proposition 5.4. *Consider an SMC algorithm using stratified resampling. Suppose that there exists a constant $\varepsilon \in (0,1]$ and a probability density h such that*

$$\varepsilon h(x') \leq q_t(x,x') \leq \varepsilon^{-1}h(x')$$

uniformly in x, t, and that there exist $\zeta > 0$ and $\delta \in (0,1)$ such that

$$\mathbb{P}\left[\max_i w_t^{(i)} - \min_i w_t^{(i)} \geq 2\delta/N \,\middle|\, \mathcal{F}_{t-1}\right] \geq \zeta \qquad (5.10)$$

for infinitely many t. Then, for all $N > 1$, $\mathbb{P}[\tau_N(t) = \infty] = 0$ for all finite t.

We now assume q_t is bounded above and away from zero, as in (5.3). We saw that such a condition was not necessary for Corollary 5.3, and we do not believe it to be necessary here either; it is merely a convenient way to control the contributions from the transition density. Indeed, the terms in ε appearing in the bounds established in the following proof are rather crude. In fact, the stated condition is stronger than necessary: we only need the bounds on q_t to hold for infinitely many t, rather than for all t. We use this stronger statement to avoid complicating the proof.

The second condition (5.10) is required to ensure that, at least infinitely often, the weights are not equal to $(1, \ldots, 1)/N$, since stratified resampling is degenerate under equal weights, which could cause the time scale to explode. It is hardly conceivable that any real SMC algorithm would fail to satisfy this very mild condition, which effectively ensures that the weights cannot be "too well-behaved".

Proof. As argued in Lemma 5.2, it is sufficient to prove that under the stated conditions

$$\sum_{r=0}^{\infty} \mathbb{P}[c_N(r) > 2/N^2 \mid \mathcal{F}_{r-1}] = \infty.$$

Firstly,

$$\begin{aligned}\mathbb{P}[c_N(t) \leq 2/N^2 \mid \mathcal{H}_t] = \mathbb{P}[c_N(t) = 0 \mid \mathcal{H}_t] &= \mathbb{P}[\nu_t^{(i)} = 1 \, \forall i \in \{1, \ldots, N\} \mid \mathcal{H}_t] \\ &\leq \mathbb{P}[\nu_t^{(i^\star)} = 1 \mid \mathcal{H}_t], \end{aligned} \tag{5.11}$$

where $i^\star := \operatorname{argmax}_i\{w_t^{(i)}\}$ (but note that the inequality holds when $i^\star$ is taken to be any particular index). Define $p_k^{(i)}$ as in (5.8) and recall that, under stratified resampling, $p_k^{(i)} \equiv 0$ for $k \notin \{-1, 0, 1, 2\}$ and

$$\sum_{k=-1}^{2} p_k^{(i)} = \sum_{k=-1}^{2} \mathbb{P}\left[\nu_t^{(i)} = \lfloor N w_t^{(i)} \rfloor + k \,\middle|\, w_t^{(1:N)}\right] = 1.$$

Up to a proportionality constant C,

$$\begin{aligned} p_k^{(i)} = C\, \mathbb{P}\left[\nu_t^{(i)} = \lfloor N w_t^{(i)} \rfloor + k \,\middle|\, w_t^{(1:N)}\right] \\ \times \sum_{\substack{a_{1:N} \in \{1,\ldots,N\}^N: \\ |\{j : a_j = i\}| = \lfloor N w_t^{(i)} \rfloor + k}} \mathbb{P}\left[a_t^{(1:N)} = a_{1:N} \,\middle|\, \nu_t^{(i)}, w_t^{(1:N)}\right] \prod_{j=1}^{N} q_{t-1}(X_t^{(a_j)}, X_{t-1}^{(j)}) \end{aligned}$$

for each $k \in \{-1, 0, 1, 2\}$. We can bound each probability above and below using the almost sure bounds on q_{t-1} from the statement of the Proposition. Once the bounds on q_{t-1} are brought outside, the remaining sum of probabilities is equal to one:

$$p_k^{(i)} \geq C\, \mathbb{P}\left[\nu_t^{(i)} = \lfloor N w_t^{(i)} \rfloor + k \,\middle|\, w_t^{(1:N)}\right] \varepsilon^N \prod_{j=1}^{N} h(X_{t-1}^{(j)}),$$

$$p_k^{(i)} \leq C\, \mathbb{P}\left[\nu_t^{(i)} = \lfloor N w_t^{(i)} \rfloor + k \,\middle|\, w_t^{(1:N)}\right] \varepsilon^{-N} \prod_{j=1}^{N} h(X_{t-1}^{(j)}).$$

We then eliminate the proportionality constant C by normalising, to obtain lower bounds

$$\begin{aligned} p_k^{(i)} &\geq \frac{C\,\mathbb{P}[\nu_t^{(i)} = \lfloor N w_t^{(i)} \rfloor + k \mid w_t^{(1:N)}]\varepsilon^N \prod_{j=1}^N h(X_{t-1}^{(j)})}{\sum_{j=-1}^{2} C\,\mathbb{P}[\nu_t^{(i)} = \lfloor N w_t^{(i)} \rfloor + j \mid w_t^{(1:N)}]\varepsilon^{-N} \prod_{j=1}^N h(X_{t-1}^{(j)})} \\ &= \mathbb{P}[\nu_t^{(i)} = \lfloor N w_t^{(i)} \rfloor + k \mid w_t^{(1:N)}]\varepsilon^{2N} \end{aligned} \tag{5.12}$$

for each k, which also imply

$$1 - p_k^{(i)} \geq \left(1 - \mathbb{P}[\nu_t^{(i)} = \lfloor N w_t^{(i)} \rfloor + k \mid w_t^{(1:N)}]\right) \varepsilon^{2N}. \tag{5.13}$$

Suppose that $\max_i w_t^{(i)} - \min_i w_t^{(i)} \geq 2\delta/N$. Then that at least one of $\{\max_i w_t^{(i)} \geq (1+\delta)/N\}$ and $\{\min_i w_t^{(i)} \leq (1-\delta)/N\}$ occurs. We will now examine each of these possibilities.

We can always write the maximum weight as $w_t^{(i^\star)} = \frac{1+\gamma}{N}$ for some $\gamma \geq 0$. Then, using (5.11),

$$\mathbb{P}[c_N(t) > 2/N^2 \mid \mathcal{H}_t] \geq 1 - \mathbb{P}[\nu_t^{(i^\star)} = 1 \mid \mathcal{H}_t] = \begin{cases} 0 & \text{if } \gamma = 0 \\ 1 - p_0^{(i^\star)} & \text{if } \gamma \in (0,1) \\ 1 - p_{-1}^{(i^\star)} & \text{if } \gamma \in [1,2) \\ 1 & \text{if } \gamma \geq 2. \end{cases}$$

If $\gamma \in (0,1)$ then the overhang in the sense of Figure 2.7 is γ, and

$$1 - p_0^{(i^\star)} \geq \frac{3\gamma}{4}\varepsilon^{2N}$$

using Table 2.1 (upper bound on p_0) and (5.13). Similarly, if $\gamma \in [1,2)$ then the overhang is $\gamma - 1$ and by Table 2.1 (upper bound on p_{-1}),

$$1 - p_{-1}^{(i^\star)} \geq \left(1 - \frac{1}{4}\right)\varepsilon^{2N} \geq \frac{3}{4}\varepsilon^{2N}.$$

Overall, under the constraint $\max_i w_t^{(i)} \geq (1+\delta)/N$, we have

$$\mathbb{P}[c_N(t) > 2/N^2 \mid \mathcal{H}_t] \geq \min_{\gamma \geq \delta} \left\{ \frac{3\gamma}{4}\varepsilon^{2N}\mathbb{1}_{\{\gamma \in [0,1)\}} + \frac{3}{4}\varepsilon^{2N}\mathbb{1}_{\{\gamma \in [1,2)\}} + \mathbb{1}_{\{\gamma \geq 2\}} \right\} = \frac{3}{4}\delta\varepsilon^{2N},$$

since $\delta < 1$.

We now construct a similar argument for the minimum weight. Let $j^\star := \operatorname{argmin}_i\{w_t^{(i)}\}$ and write $w_t^{(j^\star)} = \frac{1-\gamma}{N}$, for some $\gamma \in [0,1]$. Then by (5.11) we have

$$\mathbb{P}[c_N(t) > 2/N^2 \mid \mathcal{H}_t] \geq 1 - \mathbb{P}[\nu_t^{(j^\star)} = 1 \mid \mathcal{H}_t] = \begin{cases} 1 - p_1^{(j^\star)} & \text{if } \gamma \in (0,1] \\ 0 & \text{if } \gamma = 0. \end{cases}$$

If $\gamma \in (0,1]$ then the overhang in the sense of Figure 2.7 is $1-\gamma$, and

$$1 - p_1^{(j^\star)} \geq \left(1 - \frac{1+(1-\gamma)}{2}\right)\varepsilon^{2N} = \frac{\gamma}{2}\varepsilon^{2N},$$

using Table 2.1 (upper bound on p_1). Therefore, under the constraint $\min_i w_t^{(i)} \leq (1-\delta)/N$, we have

$$\mathbb{P}[c_N(t) > 2/N^2 \mid \mathcal{H}_t] \geq \min_{\gamma \geq \delta}\left\{\frac{\gamma}{2}\varepsilon^{2N}\right\} = \frac{1}{2}\delta\varepsilon^{2N}.$$

Combining both cases, we find for arbitrary r

$$\mathbb{P}[c_N(r) > 2/N^2 \mid \mathcal{H}_r] \geq \frac{1}{2}\delta\varepsilon^{2N}\mathbb{1}_{\{\max_i w_r^{(i)} - \min_i w_r^{(i)} \geq 2\delta/N\}}$$

so, by the tower rule and conditional independence,

$$\begin{aligned}\mathbb{P}[c_N(r) > 2/N^2 \mid \mathcal{F}_{r-1}] &= \mathbb{E}_r\left[\mathbb{P}[c_N(r) > 2/N^2 \mid \mathcal{H}_r]\right] \\ &\geq \frac{1}{2}\delta\varepsilon^{2N}\mathbb{P}[\max_i w_r^{(i)} - \min_i w_r^{(i)} \geq 2\delta/N \mid \mathcal{F}_{r-1}] \\ &\geq \frac{1}{2}\delta\varepsilon^{2N}\zeta > 0\end{aligned}$$

for infinitely many r. Hence

$$\sum_{r=0}^{\infty}\mathbb{P}[c_N(r) > 2/N^2 \mid \mathcal{F}_{r-1}] = \infty$$

as required. ■

5.3 Stochastic rounding

Corollary 5.5. *Consider an SMC algorithm using any stochastic rounding as its resampling scheme, such that* (A1) *is satisfied. Assume that there exists a constant* $a \in [1,\infty)$ *such that for all* x, x', t,

$$\frac{1}{a} \leq g_t(x,x') \leq a.$$

Assume that $\mathbb{P}[\tau_N(t) = \infty] = 0$ *for all finite* t. *Let* $(G_t^{(n,N)})_{t\geq 0}$ *denote the genealogy of a random sample of size* n *among the* N *terminal particles in the output of the algorithm. Then, for any fixed* n, *the time-scaled genealogy* $(G_{\tau_N(t)}^{(n,N)})_{t\geq 0}$ *converges weakly to Kingman's* n*-coalescent as* $N \to \infty$.

Proof. We can apply exactly the proof of Corollary 5.3, except that stochastic rounding is more restrictive than stratified resampling, so that conditional on $w_t^{(1:N)}$ the only possible offspring counts (almost surely) are $\nu_t^{(i)} \in \{\lfloor Nw_t^{(i)}\rfloor, \lfloor Nw_t^{(i)}\rfloor + 1\}$. We simply set $p_{-1}^{(i)} =$

$p_2^{(i)} = 0$ in the proof of Corollary 5.3 to see that

$$\frac{1}{(N)_3}\sum_{i=1}^{N}\mathbb{E}_t\left[(\nu_t^{(i)})_3\right] \leq \frac{a^2}{N-2}\frac{1}{(N)_2}\sum_{i=1}^{N}\mathbb{E}_t\left[(\nu_t^{(i)})_2\right]$$

as required. The result then follows by applying Theorem 4.1. ■

We can also show, under additional conditions, that the assumption $\mathbb{P}[\tau_N(t) = \infty] = 0$ for all finite t holds.

Proposition 5.6. *Consider an SMC algorithm using any stochastic rounding as its resampling scheme. Suppose that there exists a constant $\varepsilon \in (0, 1]$ and a probability density h such that*

$$\varepsilon h(x') \leq q_t(x, x') \leq \varepsilon^{-1} h(x')$$

uniformly in x, t, and that there exist $\zeta > 0$ and $\delta \in (0, 1)$ such that

$$\mathbb{P}\left[\max_i w_t^{(i)} - \min_i w_t^{(i)} \geq 2\delta/N \,\middle|\, \mathcal{F}_{t-1}\right] \geq \zeta$$

for infinitely many t. Then, for all $N > 1$, $\mathbb{P}[\tau_N(t) = \infty] = 0$ for all finite t.

This result was published in Brown et al. (2021, Lemma B.1) with the slightly stronger conditions where the bounds on q_t are also uniform in x'. It has since been noted that the conditions given here are sufficient; the h terms can be cancelled as in (5.12). As was the case for Proposition 5.4, for convenience the conditions on q_t are made stronger than necessary.

Proof. Define $p_k^{(i)}$ for $k \in \mathbb{Z}$ as in (5.8). In the case of stochastic rounding, $p_k^{(i)} \equiv 0$ for all $k \notin \{0, 1\}$, and we also have

$$\begin{aligned}
\mathbb{P}\left[\nu_t^{(i)} = \lfloor N w_t^{(i)} \rfloor \,\middle|\, w_t^{(1:N)}\right] &= 1 - N w_t^{(i)} + \lfloor N w_t^{(i)} \rfloor, \\
\mathbb{P}\left[\nu_t^{(i)} = \lfloor N w_t^{(i)} \rfloor + 1 \,\middle|\, w_t^{(1:N)}\right] &= N w_t^{(i)} - \lfloor N w_t^{(i)} \rfloor.
\end{aligned}$$

Combining this with (5.12),

$$\begin{aligned}
p_0^{(i)} &\geq (1 - N w_t^{(i)} + \lfloor N w_t^{(i)} \rfloor)\varepsilon^{2N}, \\
p_1^{(i)} &\geq (N w_t^{(i)} - \lfloor N w_t^{(i)} \rfloor)\varepsilon^{2N}.
\end{aligned} \tag{5.14}$$

Define $i^\star := \operatorname{argmax}_i\{w_t^{(i)}\}$ and write $w_t^{(i^\star)} = \frac{1+\gamma}{N}$, for some $\gamma \geq 0$. Then, using (5.11),

$$\mathbb{P}\left[c_N(t) > 2/N^2 \,\middle|\, \mathcal{H}_t\right] \geq 1 - \mathbb{P}\left[\nu_t^{(i^\star)} = 1 \,\middle|\, \mathcal{H}_t\right] = \begin{cases} 1 - p_0^{(i^\star)} & \text{if } \gamma \in [0, 1) \\ 1 & \text{if } \gamma \geq 1. \end{cases}$$

In the case $\gamma \in [0,1)$ we have $Nw_t^{(i^\star)} - \lfloor Nw_t^{(i^\star)} \rfloor = \gamma$, so

$$\mathbb{P}\left[c_N(t) > 2/N^2 \,\middle|\, \mathcal{H}_t\right] \geq 1 - p_0^{(i^\star)} = p_1^{(i^\star)} \geq \gamma\varepsilon^{2N},$$

due to (5.14). Therefore, subject to $\max_i w_t^{(i)} \geq (1+\delta)/N$,

$$\mathbb{P}\left[c_N(t) > 2/N^2 \,\middle|\, \mathcal{H}_t\right] \geq \min_{\gamma \geq \delta}\left\{\gamma\varepsilon^{2N}\right\} = \delta\varepsilon^{2N}.$$

Similarly, write $j^\star := \operatorname{argmin}_i\{w_t^{(i)}\}$ and $w_t^{(j^\star)} = \frac{1-\gamma}{N}$, for some $\gamma \in [0,1]$. Then, again using (5.11),

$$\mathbb{P}\left[c_N(t) > 2/N^2 \,\middle|\, \mathcal{H}_t\right] \geq 1 - \mathbb{P}\left[\nu_t^{(j^\star)} = 1 \,\middle|\, \mathcal{H}_t\right] = \begin{cases} 0 & \text{if } \gamma = 0 \\ 1 - p_1^{(j^\star)} & \text{if } \gamma \in (0,1) \\ 1 & \text{if } \gamma = 1. \end{cases}$$

If $\gamma \in (0,1)$ then $Nw_t^{(i^\star)} - \lfloor Nw_t^{(i^\star)} \rfloor = 1 - \gamma$, so

$$1 - p_1^{(j^\star)} = p_0^{(j^\star)} \geq (1 - (1-\gamma))\varepsilon^{2N} = \gamma\varepsilon^{2N}.$$

Therefore, subject to $\min_i w_t^{(i)} \leq (1-\delta)/N$,

$$\mathbb{P}\left[c_N(t) > 2/N^2 \,\middle|\, \mathcal{H}_t\right] \geq \min_{\gamma \geq \delta}\left\{\gamma\varepsilon^{2N}\right\} = \delta\varepsilon^{2N}.$$

Combining the cases for the maximum and minimum weight we have that

$$\mathbb{P}\left[c_N(t) > 2/N^2 \,\middle|\, \mathcal{H}_t\right] \geq \delta\varepsilon^{2N}\mathbb{1}_{\{\max_i w_t^{(i)} - \min_i w_t^{(i)} \geq 2\delta/N\}}$$

and we conclude as in Proposition 5.4. ■

5.4 Residual resampling with stratified residuals

Corollary 5.7. *Consider an SMC algorithm using residual resampling with stratified residuals, such that* (A1) *is satisfied. Assume that there exists a constant $a \in [1,\infty)$ such that for all x, x', t,*

$$\frac{1}{a} \leq g_t(x,x') \leq a.$$

Assume that $\mathbb{P}[\tau_N(t) = \infty] = 0$ for all finite t. Let $(G_t^{(n,N)})_{t\geq 0}$ denote the genealogy of a random sample of size n among the N terminal particles in the output of the algorithm. Then, for any fixed n, the time-scaled genealogy $(G_{\tau_N(t)}^{(n,N)})_{t\geq 0}$ converges weakly to Kingman's n-coalescent as $N \to \infty$.

Proof. We can apply exactly the proof of Corollary 5.3, except that residual-stratified

resampling is more restrictive than stratified resampling, so that conditional on $w_t^{(1:N)}$ the only possible offspring counts (almost surely) are $\nu_t^{(i)} \in \{\lfloor Nw_t^{(i)} \rfloor, \lfloor Nw_t^{(i)} \rfloor + 1, \lfloor Nw_t^{(i)} \rfloor + 2\}$. We simply set $p_{-1}^{(i)} = 0$ in the proof of Corollary 5.3 to see that

$$\frac{1}{(N)_3} \sum_{i=1}^{N} \mathbb{E}_t \left[(\nu_t^{(i)})_3 \right] \leq \frac{a^2}{N-2} \frac{1}{(N)_2} \sum_{i=1}^{N} \mathbb{E}_t \left[(\nu_t^{(i)})_2 \right]$$

as required. The result then follows by applying Theorem 4.1. ■

We can also show, under additional conditions, that the assumption $\mathbb{P}[\tau_N(t) = \infty] = 0$ for all finite t holds.

Proposition 5.8. *Consider an SMC algorithm using residual resampling with stratified residuals. Suppose that there exists a constant $\varepsilon \in (0, 1]$ and a probability density h such that*

$$\varepsilon h(x') \leq q_t(x, x') \leq \varepsilon^{-1} h(x')$$

uniformly in x, and that there exist $\zeta > 0$ and $\delta \in (0, 1)$ such that

$$\mathbb{P}\left[\max_i w_t^{(i)} - \min_i w_t^{(i)} \geq 2\delta/N \,\middle|\, \mathcal{F}_{t-1} \right] \geq \zeta$$

for infinitely many t. Then, for all $N > 1$, $\mathbb{P}[\tau_N(t) = \infty] = 0$ for all finite t.

Proof. Define $p_k^{(i)}$ for $k \in \mathbb{Z}$ as in (5.8). In the case of residual resampling with stratified residuals, $p_k^{(i)} \equiv 0$ for all $k \notin \{0, 1, 2\}$. Define $i^\star := \text{argmax}_i\{w_t^{(i)}\}$ and write $w_t^{(i^\star)} = \frac{1+\gamma}{N}$, for some $\gamma \geq 0$. Then, using (5.11),

$$\mathbb{P}\left[c_N(t) > 2/N^2 \,\middle|\, \mathcal{H}_t \right] \geq 1 - \mathbb{P}\left[\nu_t^{(i^\star)} = 1 \,\middle|\, \mathcal{H}_t \right] = \begin{cases} 0 & \text{if } \gamma = 0 \\ 1 - p_0^{(i^\star)} & \text{if } \gamma \in (0, 1) \\ 1 & \text{if } \gamma \geq 1. \end{cases}$$

In the case $\gamma \in (0, 1)$ we have

$$1 - p_0^{(i^\star)} = p_1^{(i^\star)} + p_2^{(i^\star)} \geq p_1^{(i^\star)} \geq \mathbb{P}\left[\nu_t^{(i^\star)} = \lfloor Nw_t^{(i^\star)} \rfloor + 1 \,\middle|\, w_t^{(1:N)} \right] \varepsilon^{2N}$$

by (5.12). Also, the residual weight in this case is $r_{i^\star} = \gamma/R$, for some $R \in \{1, \ldots, N-1\}$ (since $\gamma > 0$, $R \neq 0$). Therefore $\mathbb{P}[\nu_t^{(i^\star)} = \lfloor Nw_t^{(i^\star)} \rfloor + 1 \mid w_t^{(1:N)}]$ is the probability that stratified resampling with R individuals assigns exactly 1 offspring to a parent with weight γ/R. According to Table 2.1 (lower bound on p_1), this probability is at least $\gamma/2$. Hence

$$\mathbb{P}\left[c_N(t) > 2/N^2 \,\middle|\, \mathcal{H}_t \right] \geq \frac{\gamma}{2} \varepsilon^{2N}.$$

This means that, subject to $\max_i w_t^{(i)} \geq (1+\delta)/N$,

$$\mathbb{P}\left[c_N(t) > 2/N^2 \,\middle|\, \mathcal{H}_t\right] \geq \min_{\gamma \geq \delta} \left\{\frac{\gamma}{2}\varepsilon^{2N}\right\} = \frac{1}{2}\delta\varepsilon^{2N}.$$

Now a similar calculation for the minimum weight: let $j^\star := \operatorname{argmin}_i\{w_t^{(i)}\}$ and write $w_t^{(j^\star)} = \frac{1-\gamma}{N}$, for some $\gamma \in [0,1]$. Using (5.11),

$$\mathbb{P}\left[c_N(t) > 2/N^2 \,\middle|\, \mathcal{H}_t\right] \geq 1 - \mathbb{P}\left[\nu_t^{(j^\star)} = 1 \,\middle|\, \mathcal{H}_t\right] = \begin{cases} 0 & \text{if } \gamma = 0 \\ 1 - p_1^{(j^\star)} & \text{if } \gamma \in (0,1) \\ 1 & \text{if } \gamma = 1. \end{cases}$$

If $\gamma \in (0,1)$ then $r_{j^\star} = (1-\gamma)/R$, for some $R \in \{1, \ldots, N-1\}$, and

$$1 - p_1^{(j^\star)} = p_0^{(j^\star)} + p_2^{(j^\star)} \geq p_0^{(j^\star)} \geq \mathbb{P}\left[\nu_t^{(j^\star)} = \lfloor N w_t^{(j^\star)} \rfloor \,\middle|\, w_t^{(1:N)}\right] \varepsilon^{2N}$$

by (5.12). Now $\mathbb{P}[\nu_t^{(j^\star)} = \lfloor N w_t^{(j^\star)} \rfloor \mid w_t^{(1:N)}]$ is the probability that stratified resampling with R individuals assigns exactly 0 offspring to a parent with weight $(1-\gamma)/R$. According to Table 2.1 (lower bound on p_0), this probability is at least $\gamma/2$. Hence

$$\mathbb{P}\left[c_N(t) > 2/N^2 \,\middle|\, \mathcal{H}_t\right] \geq \frac{\gamma}{2}\varepsilon^{2N}.$$

Therefore, using (5.12), we have that subject to $\min_i w_t^{(i)} \leq (1-\delta)/N$

$$\mathbb{P}\left[c_N(t) > 2/N^2 \,\middle|\, \mathcal{H}_t\right] \geq \min_{\gamma \geq \delta} \left\{\frac{\gamma}{2}\varepsilon^{2N}\right\} = \frac{1}{2}\delta\varepsilon^{2N}.$$

Combining the cases for the maximum and minimum weight we have

$$\mathbb{P}\left[c_N(t) > 2/N^2 \,\middle|\, \mathcal{H}_t\right] \geq \frac{1}{2}\delta\varepsilon^{2N} \mathbb{1}_{\{\max_i w_t^{(i)} - \min_i w_t^{(i)} \geq 2\delta/N\}}$$

and we conclude as in Proposition 5.4. ■

5.5 Residual resampling with multinomial residuals

We believe that an analogous result holds when the resampling scheme used is residual resampling with multinomial residuals. Considering the ordering by variance presented in Proposition 2.3, the residual-multinomial scheme sits between the multinomial scheme and the residual-stratified scheme, both of which admit the desired convergence result (Corollaries 5.1 and 5.7).

However, we have so far been unable to prove a similar corollary for the residual-multinomial scheme. The techniques used for other residual schemes (see Section 5.4) fail here because the number of offspring assigned to each individual is not upper bounded

by $\lfloor N w_t^{(i)} \rfloor$ plus a constant; as many as $R = O(N)$ residual offspring may be assigned to a single individual. The technique used for multinomial resampling (Section 5.1) also fails here: although we have a closed-form expression for the joint distribution of parental indices, it is not a straightforward product form because of the additional dependence between offspring counts induced by the deterministic assignments, so it is unclear how to recover the marginal distributions.

5.6 Star resampling

One might ask the question: is it possible to construct an SMC algorithm whose genealogies converge to some non-trivial limit other than the n-coalescent? The answer is yes, as we now illustrate.

Recall that star resampling assigns all of the offspring to a single parent which is sampled from the Categorical distribution parametrised by $w_t^{(1:N)}$. It is easy enough to show that such a resampling scheme does not satisfy (4.1). The vector of offspring counts is at every generation some permutation of $(N, 0, \ldots, 0)$, and hence we calculate

$$\frac{1}{(N)_2} \sum_{i=1}^{N} \mathbb{E}\left[(\nu_t^{(i)})_2 \,\middle|\, \mathcal{H}_t \right] = \frac{1}{(N)_2} (N)_2 = 1,$$
$$\frac{1}{(N)_3} \sum_{i=1}^{N} \mathbb{E}\left[(\nu_t^{(i)})_3 \,\middle|\, \mathcal{H}_t \right] = \frac{1}{(N)_3} (N)_3 = 1,$$

so no suitable sequence b_N can be found. Now we know that Theorem 3.6 does not apply, but this is not enough because condition (4.1) was not proved to be necessary. But in fact we know exactly what the genealogy of n particles from this SMC algorithm looks like (Figure 5.2). Whatever time scale is used, we cannot get away from the fact that this

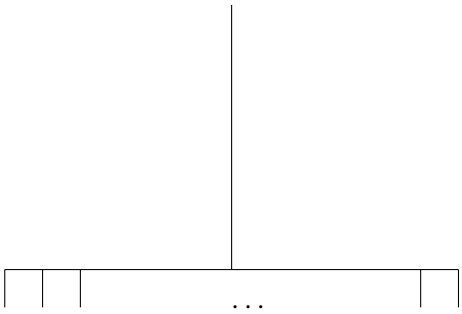


Figure 5.2: Sample genealogy induced by star resampling

genealogy involves multiple mergers; it cannot converge to the n-coalescent.

The limiting genealogy is more like a *star coalescent* (Griffiths and Mano 2016; Pitman 1999). This is the coalescent process comprising an Exp(1)-distributed event time at which all of the lineages merge into one.

In the case of star resampling we have $c_N(t) \equiv 1$, so the time-scaling function $\tau_N(t)$

defined in (3.2) is the identity function $\tau(t) \equiv t$ for all N, and this does not yield a continuous-time limit. Under any time scale that results in a continuous-time limiting process, the coalescent event time converges to 0, rather than the usual Exp(1)-distributed random variable. The resulting genealogy is a variant star coalescent where the distribution of the event time is a point mass at 0. An interesting consequence of this is that this coalescent comes down from infinity instantaneously, while the classical star coalescent does not.

5.7 Conditional SMC

In conditional SMC, one "immortal" particle is treated differently to the others when it comes to assigning offspring to parents. The immortal particle is guaranteed at least one offspring, and has on average one more offspring than each of the other parents in each generation. This results in genealogies that are qualitatively different to those of a corresponding standard SMC algorithm. For one thing, the population MRCA is *guaranteed* to be an immortal particle; there is a sense in which the immortal lineage *attracts* coalescence events.

Given this, we should not have been surprised if conditional SMC genealogies converged to a quite different coalescent process, perhaps a *structured coalescent* (Notohara 1990). As it turns out, we still recover Kingman's n-coalescent in the large population limit (Corollary 5.9). The explanation for this is that, as $N \to \infty$, the probability of a given sample of size n interacting with the immortal lineage (before its within-sample MRCA) vanishes, leaving a process that looks very much like the one induced by the corresponding standard SMC algorithm.

Corollary 5.9. *Consider a conditional SMC algorithm using multinomial resampling, such that* (A1) *is satisfied. Assume there exist constants* $\varepsilon \in (0, 1]$ *and* $a \in [1, \infty)$ *and probability density* h *such that for all* x, x', t,

$$\frac{1}{a} \leq g_t(x, x') \leq a, \quad \varepsilon h(x') \leq q_t(x, x') \leq \frac{1}{\varepsilon} h(x'). \tag{5.15}$$

Let $(G_t^{(n,N)})_{t \geq 0}$ *denote the genealogy of a random sample of size* n *among the* N *terminal particles in the output of the algorithm. Then, for any fixed* n*, the time-scaled genealogy* $(G_{\tau_N(t)}^{(n,N)})_{t \geq 0}$ *converges weakly to Kingman's* n*-coalescent as* $N \to \infty$.

We restrict here to the case of multinomial resampling, which seems to be the most commonly-used resampling scheme within conditional SMC. Implementing other resampling schemes while maintaining the immortal lineage is more involved, though by no means impossible (for details see Lee, Murray, and Johansen 2019, for example). We conjecture that similar results hold for conditional SMC with other resampling schemes, as in the preceding corollaries.

The conditions (5.15) are, as one might expect, identical to those assumed in the case of standard SMC with multinomial resampling (Corollary 5.1). These should be interpreted as holding uniformly in the choice of immortal trajectory.

Proof. Assume, without loss of generality, that the immortal particle takes index 1 in each generation. This assumption is valid due to (A1), and significantly lightens the notation, but the same argument holds if the immortal indices are taken to be $a^\star_{0:T}$ rather than $(1, \dots, 1)$.

Define $\mathcal{H}_t$ as in (5.1). The parental indices are conditionally independent given $\mathcal{H}_t$, as in standard SMC with multinomial resampling, but we have to treat $i = 1$ as a special case. The conditional law on the i^{th} parental index is

$$\mathbb{P}\left[a_t^{(i)} = a_i \,\middle|\, \mathcal{H}_t\right] \propto \begin{cases} \mathbb{1}_{a_i=1} & i = 1 \\ g_t(X_{t+1}^{a_t^{(a_i)}}, X_t^{(a_i)}) q_{t-1}(X_t^{(a_i)}, X_{t-1}^{(i)}) & i = 2, \dots, N, \end{cases}$$

resulting in the joint law

$$\mathbb{P}\left[a_t^{(1:N)} = a_{1:N} \,\middle|\, \mathcal{H}_t\right] \propto \mathbb{1}_{a_1=1} \prod_{i=2}^{N} g_t(X_{t+1}^{a_t^{(a_i)}}, X_t^{(a_i)}) q_{t-1}(X_t^{(a_i)}, X_{t-1}^{(i)}).$$

As in Corollary 5.1, under (5.15) we have bounds

$$\mathbb{E}\left[(V_L^{(i)})_k\right] \leq \mathbb{E}\left[(\nu_t^{(i)})_k \,\middle|\, \mathcal{H}_t\right] \leq \mathbb{E}\left[(V_U^{(i)})_k\right],$$

where now

$$V_L^{(i)} \stackrel{d}{=} \mathbb{1}_{i=1} + \text{Binomial}\left(N-1, \frac{\varepsilon/a}{(\varepsilon/a) + (N-1)(a/\varepsilon)}\right),$$
$$V_U^{(i)} \stackrel{d}{=} \mathbb{1}_{i=1} + \text{Binomial}\left(N-1, \frac{a/\varepsilon}{(a/\varepsilon) + (N-1)(\varepsilon/a)}\right).$$

independently for each i and independently of $\mathcal{F}_\infty$. Furthermore, using the Binomial moments and the identity $(X+1)_2 \equiv 2(X)_1 + (X)_2$, one can show that

$$\mathbb{E}\left[(V_L^{(i)})_2\right] \geq \begin{cases} \frac{(N-1)_2}{N^2}\frac{\varepsilon^4}{a^4} + \frac{2(N-1)}{N}\frac{\varepsilon^2}{a^2} & \text{if } i = 1 \\ \frac{(N-1)_2}{N^2}\frac{\varepsilon^4}{a^4} & \text{if } i \neq 1. \end{cases}$$

Using the identity $(X+1)_3 \equiv 3(X)_2 + (X)_3$, we also have

$$\mathbb{E}\left[(V_U^{(i)})_3\right] \leq \begin{cases} \frac{(N-1)_3}{N^3}\frac{a^6}{\varepsilon^6} + \frac{3(N-1)_2}{N^2}\frac{a^4}{\varepsilon^4} & \text{if } i = 1 \\ \frac{(N-1)_3}{N^3}\frac{a^6}{\varepsilon^6} & \text{if } i \neq 1. \end{cases}$$

We therefore have

$$\frac{1}{(N)_2}\sum_{i=1}^{N}\mathbb{E}\left[(\nu_t^{(i)})_2 \,\middle|\, \mathcal{H}_t\right] \geq \frac{1}{(N)_2}\sum_{i=1}^{N}\mathbb{E}\left[(V_L^{(i)})_2\right] \geq \frac{1}{(N)_2}\left[\frac{2(N-1)}{N}\frac{\varepsilon^2}{a^2} + \sum_{i=1}^{N}\frac{(N-1)_2}{N^2}\frac{\varepsilon^4}{a^4}\right]$$
$$= \frac{1}{N^2}\left[2\frac{\varepsilon^2}{a^2} + (N-2)\frac{\varepsilon^4}{a^4}\right] \geq \frac{\varepsilon^4}{Na^4} \tag{5.16}$$

and

$$\frac{1}{(N)_3}\sum_{i=1}^{N}\mathbb{E}\left[(\nu_t^{(i)})_3 \,\middle|\, \mathcal{H}_t\right] \leq \frac{1}{(N)_3}\sum_{i=1}^{N}\mathbb{E}\left[(V_U^{(i)})_3\right] \leq \frac{1}{(N)_3}\left[\frac{3(N-1)_2}{N^2}\frac{a^4}{\varepsilon^4} + \sum_{i=1}^{N}\frac{(N-1)_3}{N^3}\frac{a^6}{\varepsilon^6}\right]$$
$$= \frac{1}{N^3}\left[3\frac{a^4}{\varepsilon^4} + (N-3)\frac{a^6}{\varepsilon^6}\right] \leq \frac{a^6}{N^2\varepsilon^6}.$$

Hence, applying (5.2), we can upper bound the ratio

$$\frac{\frac{1}{(N)_3}\sum_{i=1}^{N}\mathbb{E}_t[(\nu_t^{(i)})_3]}{\frac{1}{(N)_2}\sum_{i=1}^{N}\mathbb{E}_t[(\nu_t^{(i)})_2]} \leq \frac{a^{10}}{N\varepsilon^{10}} =: b_N \underset{N\to\infty}{\longrightarrow} 0$$

so (4.1) is satisfied. Proof that the time scale is finite is relegated to Lemma 5.10, whence we conclude by applying Theorem 4.1.

■

Lemma 5.10. *Consider a conditional SMC algorithm using multinomial resampling, satisfying* (A1) *and* (5.15)*. Then, for all $N > 2$, $\mathbb{P}[\tau_N(t) = \infty] = 0$ for all finite t.*

Proof. The proof is identical to that of Lemma 5.2, since (5.16) gives us exactly the same lower bound on $\mathbb{E}_t[c_N(t)]$ that we had in standard SMC with multinomial resampling. ■

5.7.1 Effect of ancestor sampling

Ancestor sampling breaks up the immortal lineage into sections, so it is not really a lineage any more. We can still trace genealogies of terminal particles by tracing back their lineages as usual, except that some parts of these lineages may be ancestor-sampled links rather than links from the original forward lineages (see Figure 2.13b).

Using the parent sampling probabilities specified in (2.17), now with time reversed so future information must be incorporated, we obtain

$$\mathbb{P}\left[a_t^{(i)} = a_i \,\middle|\, \mathcal{H}_t\right] \propto \begin{cases} w_t^{(a_i)} q_{t-1}(X_t^{(a_i)}, X_{t-1}^{(i)}) & i \in \text{non-immortal particles} \\ w_t^{(a_i)} q_{t-1}(X_t^{(a_i)}, x_{t-1}^{\star}) & i = \text{immortal particle.} \end{cases}$$

But when i is the index of the immortal particle, $X_{t-1}^{(i)} = x_{t-1}^{\star}$, so the above simplifies to

$$\mathbb{P}\left[a_t^{(i)} = a_i \,\middle|\, \mathcal{H}_t\right] \propto w_t^{(a_i)} q_{t-1}(X_t^{(a_i)}, X_{t-1}^{(i)})$$

for each i, which is exactly (5.4), the law on parental indices under standard SMC with multinomial resampling!

In other words, when parental indices are chosen, the immortal particle is treated exactly like all of the other particles; it has completely lost its reproductive advantage. This means it is no more likely for lineages to coalesce onto the immortal lineage than onto any other lineage, so we do not see the behaviour of Figure 2.12 which caused the particle Gibbs chain to mix slowly over the sequential component. This supports the claim of Section 2.5.3: particle Gibbs with ancestor sampling still experiences ancestral degeneracy, but this no longer causes the sequential component to get stuck.

6 Discussion

> Oh, there's such a lot of things to do
> and such a lot to be
> That there's always lots of cherries
> on my little cherry tree!
>
> A. A. Milne

We have provided a simple sufficient condition for genealogies of SMC particle systems to converge weakly to Kingman's n-coalescent in the large population limit. This result complements existing work not only in the SMC literature but also in mathematical population genetics, where it shows that non-neutral population models can produce n-coalescents in the limit, under conditions analogous to those required for neutral models.

We have demonstrated that our convergence condition is verifiable in a range of settings, including SMC algorithms using many of the most popular resampling schemes. Convergence to a coalescent limit requires a random rescaling of time, governed by the function τ_N, which can be viewed as encoding the genealogical behaviour of each algorithm. Information about this time-scale function could therefore be used to directly compare the ancestral degeneracy of different algorithms, solve tuning problems, or quantify asymptotic behaviour of SMC estimators.

I believe that the main limitation of the work, therefore, is our lack of information about τ_N. An interesting topic of future research would be to characterise this function a priori, say for a particular tractable class of models. From there it would be possible to find the limiting distributions of many statistics of interest, such as the time to full coalescence or the probability of maintaining a certain number of distinct lineages over a given time window. It would also allow a direct comparison of the genealogies arising from different resampling schemes.

I will finish by describing three more open questions raised by the current work, which I believe to be interesting avenues for future research. These problems are, in my opinion, less critical than that of characterising the time-scale function, but probably easier to tackle. I hope that a future researcher may find these to be interesting diversions.

In neutral models, the neutral version of our main condition has been shown to be necessary and sufficient for convergence to the n-coalescent. This raises the question: are

our conditions necessary as well as sufficient, or else what alterations are needed to render them necessary and sufficient?

We have shown that our convergence theorems apply to a range of SMC algorithms, encompassing most of the resampling schemes that are routinely used by practitioners. A notable exception is residual resampling with multinomial residuals, which, although not generally recommended by theorists, is frequently used in practice. There is no reason to believe that the convergence results should not apply in this case: we have seen that by various metrics residual-multinomial resampling lies between multinomial and, say, residual-stratified resampling, both of which have been shown to satisfy the conditions of the theorems. However, we have not yet succeeded in proving a corollary for residual-multinomial resampling.

Adaptive resampling is routinely used to improve the performance of SMC, and mitigates the problem of ancestral degeneracy. It is not immediately clear, however, what exactly is the effect of adaptive resampling on the resulting genealogies. Del Moral, Doucet, and Jasra (2012) show that the random resampling times converge almost surely to some unspecified deterministic resampling times as the number of particles tends to infinity, suggesting that adaptive resampling should simply slow the coalescent time scale by a factor corresponding to the frequency of these resampling times. However, another effect of adaptive resampling is that whenever resampling occurs the weights necessarily have variability above a certain threshold, so the resulting coalescences tend to be larger than those resulting from non-adaptive resampling. It may be that conditions such as those in Corollary 5.1 impose sufficient regularity to nonetheless yield a Kingman coalescent limit under adaptive resampling, albeit on a slower time scale than its non-adaptive analogue.

Bibliography

Andrieu, Christophe, Arnaud Doucet, and Roman Holenstein (2010). "Particle Markov Chain Monte Carlo Methods". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.3, pp. 269–342.

Baum, Leonard E. et al. (1970). "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains". In: *The Annals of Mathematical Statistics* 41, pp. 164–171.

Brown, Suzie et al. (2021). "Simple Conditions for Convergence of Sequential Monte Carlo Genealogies with Applications". In: *Electronic Journal of Probability* 26.1, pp. 1–22. ISSN: 1083-6489. DOI: 10.1214/20-EJP561.

Cannings, Chris (1974). "The Latent Roots of Certain Markov Chains Arising in Genetics: A New Approach, I. Haploid Models". In: *Advances in Applied Probability* 6.2, pp. 260–290.

– (1975). "The Latent Roots of Certain Markov Chains Arising in Genetics: A New Approach, II. Further Haploid Models". In: *Advances in Applied Probability* 7.2, pp. 264–282.

Carpenter, James, Peter Clifford, and Paul Fearnhead (1999). "Improved Particle Filter for Nonlinear Problems". In: *IEE Proceedings — Radar, Sonar and Navigation* 146.1, pp. 2–7.

Chau, Thomas C. P. et al. (2012). "Adaptive Sequential Monte Carlo Approach for Real-Time Applications". In: *22nd International Conference on Field Programmable Logic and Applications (FPL)*. IEEE, pp. 527–530.

Chopin, Nicolas (2004). "Central Limit Theorem for Sequential Monte Carlo Methods and its Application to Bayesian Inference". In: *The Annals of Statistics* 32.6, pp. 2385–2411.

Chopin, Nicolas and Omiros Papaspiliopoulos (2020). *An Introduction to Sequential Monte Carlo*. Springer.

Corenflos, Adrien et al. (2021). *Differentiable Particle Filtering via Entropy-Regularized Optimal Transport*. arXiv: 2102.07850v3.

Crisan, Dan, Pierre Del Moral, and Terry Lyons (1999). "Discrete Filtering Using Branching and Interacting Particle Systems". In: *Markov Processes and Related Fields* 5.3, pp. 293–318.

Crisan, Dan and Arnaud Doucet (2002). "A Survey of Convergence Results on Particle Filtering Methods for Practitioners". In: *IEEE Transactions on Signal Processing* 50.3, pp. 736–746.

Crisan, Dan and Terry Lyons (1999). "A Particle Approximation of the Solution of the Kushner–Stratonovitch Equation". In: *Probability Theory and Related Fields* 115.4, pp. 549–578.

Del Moral, Pierre (2004). *Feynman–Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer.

– (2013). *Mean Field Simulation for Monte Carlo Integration*. Chapman and Hall/CRC.

Del Moral, Pierre, Arnaud Doucet, and Ajay Jasra (2012). "On Adaptive Resampling Strategies for Sequential Monte Carlo Methods". In: *Bernoulli* 18.1, pp. 252–278.

Del Moral, Pierre and Alice Guionnet (1999). "Central Limit Theorem for Nonlinear Filtering and Interacting Particle Systems". In: *Annals of Applied Probability*, pp. 275–297.

Del Moral, Pierre, Laurent Miclo, et al. (2009). "The Convergence to Equilibrium of Neutral Genetic Models". In: *Stochastic Analysis and Applications* 28.1, pp. 123–143.

Devroye, Luc (1986). *Non-Uniform Random Variate Generation*. New York: Springer Science+Business Media.

Douc, Randal, Olivier Cappé, and Eric Moulines (2005). "Comparison of Resampling Schemes for Particle Filtering". In: *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis*. IEEE, pp. 64–69.

Doucet, Arnaud, Simon J. Godsill, and Christophe Andrieu (2000). "On Sequential Monte Carlo Sampling Methods for Bayesian Filtering". In: *Statistics and Computing* 10.3, pp. 197–208.

Doucet, Arnaud and Adam M. Johansen (2011). "A Tutorial on Particle Filtering and Smoothing: Fifteen Years Later". In: *Handbook of Nonlinear Filtering*. OUP, pp. 656–704.

Durrett, Richard (2008). *Probability Models for DNA Sequence Evolution*. Springer Science & Business Media.

– (2019). *Probability: Theory and Examples*. 5th ed. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. DOI: 10.1017/9781108591034.

Efron, Bradley and Robert J. Tibshirani (1994). *An Introduction to the Bootstrap*. CRC press.

Ethier, Stewart N. and Thomas G. Kurtz (1986). *Markov Processes: Characterization and Convergence*. John Wiley & Sons.

Evensen, Geir (1994). "Sequential Data Assimilation with a Nonlinear Quasi-Geostrophic Model Using Monte Carlo Methods to Forecast Error Statistics". In: *Journal of Geophysical Research: Oceans* 99.C5, pp. 10143–10162.

Fearnhead, Paul and Peter Clifford (2003). "On-line Inference for Hidden Markov Models via Particle Filters". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.4, pp. 887–899.

Fisher, Ronald Aylmer (1923). "On the Dominance Ratio". In: *Proceedings of the Royal Society of Edinburgh* 42, pp. 321–341.

Fisher, Ronald Aylmer (1930). "The Distribution of Gene Ratios for Rare Mutations". In: *Proceedings of the Royal Society of Edinburgh* 50, pp. 205–220.

Forbes, Catherine et al. (2011). *Statistical Distributions*. John Wiley & Sons.

Fox, Dieter (2003). "Adapting the Sample Size in Particle Filters through KLD-Sampling". In: *The International Journal of Robotics Research* 22.12, pp. 985–1003.

Fu, Yun-Xin (2006). "Exact Coalescent for the Wright–Fisher Model". In: *Theoretical Population Biology* 69 (4), pp. 385–394.

Gandy, Axel and F. Din-Houn Lau (2016). "The Chopthin Algorithm for Resampling". In: *IEEE Transactions on Signal Processing* 64.16, pp. 4273–4281.

Gerber, Mathieu, Nicolas Chopin, and Nick Whiteley (2019). "Negative Association, Ordering and Convergence of Resampling Methods". In: *The Annals of Statistics* 47.4, pp. 2236–2260.

Godsill, Simon J., Arnaud Doucet, and Mike West (2004). "Monte Carlo Smoothing for Nonlinear Time Series". In: *Journal of the American Statistical Association* 99.465, pp. 156–168.

Gordon, Neil J., David J. Salmond, and Adrian F. M. Smith (1993). "Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation". In: *IEE Proceedings F (Radar and Signal Processing)*. Vol. 140. 2. IET, pp. 107–113.

Griffiths, Robert and Shuhei Mano (2016). "The Star-Shaped Λ-coalescent and Fleming–Viot Process". In: *Stochastic Models* 32.4, pp. 606–631.

Hardy, Godfrey Harold and Srinivasa Aaiyangar Ramanujan (1918). "Asymptotic Formulae in Combinatory Analysis". In: *Proceedings of the London Mathematical Society* s2-17.1, pp. 75–115.

Hol, Jeroen D., Thomas B. Schön, and Fredrik Gustafsson (2006). "On Resampling Algorithms for Particle Filters". In: *Nonlinear Statistical Signal Processing Workshop*. IEEE, pp. 79–82.

Huang, Chaofan, Vengazhiyil Roshan Joseph, and Simon Mak (2020). *Population Quasi-Monte Carlo*. arXiv: `2012.13769v1`.

Jacob, Pierre E., Lawrence M. Murray, and Sylvain Rubenthaler (2015). "Path Storage in the Particle Filter". In: *Statistics and Computing* 25.2, pp. 487–496.

Jazwinski, Andrew H. (2007). *Stochastic Processes and Filtering Theory*. Courier Corporation.

Joag-Dev, Kumar and Frank Proschan (1983). "Negative Association of Random Variables with Applications". In: *The Annals of Statistics*, pp. 286–295.

Kalman, Rudolph Emil (1960). "A New Approach to Linear Filtering and Prediction Problems". In: *Journal of Basic Engineering* 82.1, pp. 35–45.

Kingman, John F. C. (1982a). "Exchangeability and the Evolution of Large Populations". In: *Proceedings of the International Conference on Exchangeability in Probability and Statistics, Rome, 6th-9th April, 1981, in Honour of Professor Bruno de Finetti*. North-Holland, Amsterdam.

Kingman, John F. C. (1982b). "On the Genealogy of Large Populations". In: *Journal of Applied Probability* 19.A, pp. 27–43.

– (1982c). "The Coalescent". In: *Stochastic Processes and Their Applications* 13.3, pp. 235–248.

Kitagawa, Genshiro (1996). "Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models". In: *Journal of Computational and Graphical Statistics* 5.1, pp. 1–25.

Kon Kam King, Guillaume, Omiros Papaspiliopoulos, and Matteo Ruggiero (2021). "Exact inference for a class of hidden Markov models on general state spaces". In: *Electronic Journal of Statistics* 5.1, pp. 2832–2875. DOI: 10.1214/21-EJS1841.

Koskela, Jere et al. (2018). *Asymptotic Genealogies of Interacting Particle Systems with an Application to Sequential Monte Carlo*. arXiv: 1804.01811v7.

Kuipers, Lauwerens and Harald Niederreiter (1974). *Uniform Distribution of Sequences*. John Wiley & Sons.

Lee, Anthony, Lawrence M. Murray, and Adam M. Johansen (2019). "Resampling in Conditional SMC Algorithms". Unpublished.

Lee, Anthony and Nick Whiteley (2016). "Forest Resampling for Distributed Sequential Monte Carlo". In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 9.4, pp. 230–248.

– (2018). "Variance Estimation in the Particle Filter". In: *Biometrika* 105.3, pp. 609–625.

Li, Yichao et al. (2020). *Stratification and Optimal Resampling for Sequential Monte Carlo*. arXiv: 2004.01975v2.

Lin, Ming, Rong Chen, and Jun S. Liu (2013). "Lookahead Strategies for Sequential Monte Carlo". In: *Statistical Science* 28.1, pp. 69–94.

Lindsten, Fredrik and Thomas B. Schön (2013). "Backward Simulation Methods for Monte Carlo Statistical Inference". In: *Foundations and Trends in Machine Learning* 6.1, pp. 1–143.

Liu, Jun S. and Rong Chen (1995). "Blind Deconvolution via Sequential Imputations". In: *Journal of the American Statistical Association* 90.430, pp. 567–576.

– (1998). "Sequential Monte Carlo Methods for Dynamic Systems". In: *Journal of the American Statistical Association* 93.443, pp. 1032–1044.

Liu, Jun S., Rong Chen, and Tanya Logvinenko (2001). "A Theoretical Framework for Sequential Importance Sampling with Resampling". In: *Sequential Monte Carlo Methods in Practice*. Springer, pp. 225–246.

Möhle, Martin (1998). "Robustness Results for the Coalescent". In: *Journal of Applied Probability* 35.2, pp. 438–447.

– (1999). "Weak Convergence to the Coalescent in Neutral Population Models". In: *Journal of Applied Probability* 36.2, pp. 446–460.

– (2000). "Total Variation Distances and Rates of Convergence for Ancestral Coalescent Processes in Exchangeable Population Models". In: *Advances in Applied Probability*, pp. 983–993.

Möhle, Martin and Serik Sagitov (1998). "A Characterization of Ancestral Limit Processes Arising in Haploid Population Genetics Models". In:

– (2001). "A Classification of Coalescent Processes for Haploid Exchangeable Population Models". In: *The Annals of Probability* 29.4, pp. 1547–1562.

Möhle, Martin and Serik Sagitov (2003). "Coalescent Patterns in Exchangeable Diploid Population Models". In: *Journal of Mathematical Biology* 47, pp. 337–352.

Moran, Patrick Alfred Pierce (1958). "Random Processes in Genetics". In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 54. 1. Cambridge University Press, pp. 60–71.

Mosimann, James E. (1962). "On the Compound Multinomial Distribution, the Multivariate β-Distribution, and Correlations among Proportions". In: *Biometrika* 49.1/2, pp. 65–82.

Murray, Lawrence M., Anthony Lee, and Pierre E. Jacob (2016). "Parallel Resampling in the Particle Filter". In: *Journal of Computational and Graphical Statistics* 25.3, pp. 789–805.

Myers, Aaron et al. (2021). "Sequential Ensemble Transform for Bayesian Inverse Problems". In: *Journal of Computational Physics* 427, p. 110055.

Notohara, Morihiro (1990). "The Coalescent and the Genealogical Process in Geographically Structured Population". In: *Journal of Mathematical Biology* 29.1, pp. 59–75.

Pitman, Jim (1999). "Coalescents with Multiple Collisions". In: *Annals of Probability* 27.4, pp. 1870–1902.

Pitt, Michael K. and Neil Shephard (1999). "Filtering via Simulation: Auxiliary Particle Filters". In: *Journal of the American Statistical Association* 94.446, pp. 590–599.

Rauch, Herbert E., C. T. Striebel, and F. Tung (1965). "Maximum Likelihood Estimates of Linear Dynamic Systems". In: *AIAA Journal* 3.8, pp. 1445–1450.

Reich, Sebastian (2013). "A Nonparametric Ensemble Transform Method for Bayesian Inference". In: *SIAM Journal on Scientific Computing* 35.4, A2013–A2024.

Rubin, Donald B. (1987). "Discussion on 'The Calculation of Posterior Distributions by Data Augmentation'". In: *Journal of the American Statistical Association* 82.398, pp. 543–546. DOI: 10.2307/2289460.

Sagitov, Serik (1999). "The General Coalescent with Asynchronous Mergers of Ancestral Lines". In: *Journal of Applied Probability*, pp. 1116–1125.

Saunders, Ian W., Simon Tavaré, and G. A. Watterson (1984). "On the Genealogy of Nested Subsamples from a Haploid Population". In: *Advances in Applied Probability*, pp. 471–491.

Spouge, John L. (2014). "Within a Sample from a Population, the Distribution of the Number of Descendants of a Subsample's Most Recent Common Ancestor". In: *Theoretical Population Biology* 92, pp. 51–54.

Srinivasan, Aravind (2001). "Distributions on Level-Sets with Applications to Approximation Algorithms". In: *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science*. IEEE, pp. 588–597.

Verma, Thomas and Judea Pearl (1988). "Causal Networks: Semantics and Expressiveness". In: *Proceedings of the 4th Workshop on Uncertainty in Artificial Intelligence*. Minneapolis, MN, Mountain View, CA, pp. 352–359.

Vidoni, Paolo (1999). "Exponential Family State Space Models Based on a Conjugate Latent Process". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.1, pp. 213–221.

Wan, Eric A. and Rudolph van der Merwe (2000). "The Unscented Kalman Filter for Nonlinear Estimation". In: *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium*. IEEE, pp. 153–158.

Webber, Robert J. (2019). *Unifying Sequential Monte Carlo with Resampling Matrices*. arXiv: `1903.12583v1`.

Whiteley, Nick (2010). "Discussion on 'Particle Markov Chain Monte Carlo Methods'". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.3, pp. 306–307.

Whiteley, Nick, Anthony Lee, and Kari Heine (2016). "On the Role of Interaction in Sequential Monte Carlo Algorithms". In: *Bernoulli* 22.1, pp. 494–529.

Whitley, Darrell (1994). "A Genetic Algorithm Tutorial". In: *Statistics and Computing* 4.2, pp. 65–85.

Wright, Sewall (1931). "Evolution in Mendelian Populations". In: *Genetics* 16.2, pp. 97–159.