Branching and Fragmenting Particle Systems in Biology

Terence Tsui Ho Lung

St. John's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

November 17, 2023

Abstract

We introduce and study two new classes of branching and fragmenting particle systems in biology. The first class of branching particle systems with interactions model spatially-structured populations with density-dependent birth and death mechanisms. Our spatial model describes spatially heterogeneous populations using point measures, considering birth and death rates based on spatial position and local population density. Notably, our model includes a juvenile phase, where offspring mature with a density-dependent probability. This results in novel scaling limits: an interacting superprocess, a nonlocal partial differential equation (PDE), and a classical PDE (in which the diffusion term is nonlinear). Furthermore, using a lookdown representation, we retain information about genealogies and, in the case of deterministic limiting models, use this to deduce the backwards in time motion of the ancestral lineage of a sampled individual. We observe that knowing the history of the population density is not enough to determine the motion of ancestral lineages in our model.

In our second class of models, motivated by using cell-free DNA (cfDNA) as a biomarker, we capture three mechanisms governing cfDNA fragmentation in the bloodstream. Our Markovian model, FRIME (Fragmentation, Immigration, and Exit), allows the simulation of cfDNA fragment profiles by sampling from its stationary distribution. We validate the model by comparing simulated profiles with mitochondrial and genomic cfDNA fragments, finding consistency with experimental observations. We employ probabilistic tools and the theory of fragmentation processes to establish the first moments of FRIME processes mathematically. This simulation framework improves our computational understanding of DNA fragmentation dynamics in the bloodstream and has the potential to underpin tools for the analysis of liquid biopsy data.

Branching and Fragmenting Particle Systems in Biology



Terence Tsui Ho Lung

St. John's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

November 17, 2023

This thesis is dedicated to
my late father Andy Tsui Lap Fung
who throughout his life encouraged me to pursue my dreams and make a difference.

Abstract

We introduce and study two new classes of branching and fragmenting particle systems in biology. The first class of branching particle systems with interactions model spatially-structured populations with density-dependent birth and death mechanisms. Our spatial model describes spatially heterogeneous populations using point measures, considering birth and death rates based on spatial position and local population density. Notably, our model includes a juvenile phase, where offspring mature with a density-dependent probability. This results in novel scaling limits: an interacting superprocess, a nonlocal partial differential equation (PDE), and a classical PDE (in which the diffusion term is nonlinear). Furthermore, using a lookdown representation, we retain information about genealogies and, in the case of deterministic limiting models, use this to deduce the backwards in time motion of the ancestral lineage of a sampled individual. We observe that knowing the history of the population density is not enough to determine the motion of ancestral lineages in our model.

In our second class of models, motivated by using cell-free DNA (cfDNA) as a biomarker, we capture three mechanisms governing cfDNA fragmentation in the bloodstream. Our Markovian model, FRIME (Fragmentation, Immigration, and Exit), allows the simulation of cfDNA fragment profiles by sampling from its stationary distribution. We validate the model by comparing simulated profiles with mitochondrial and genomic cfDNA fragments, finding consistency with experimental observations. We employ probabilistic tools and the theory of fragmentation processes to establish the first moments of FRIME processes mathematically. This simulation framework improves our computational understanding of DNA fragmentation dynamics in the bloodstream and has the potential to underpin tools for the analysis of liquid biopsy data.

Declaration

I hereby declare that, except where specific reference is made to the work of others, the contents of this dissertation are original.

The work presented in Chapter 2-5 is a collaboration between the author and his collaborators Alison Etheridge, Thomas Kurtz, Peter Ralph and Ian Letter, and resulted in a pre-print submitted to the Electronic Journal of Probability, see `https://arxiv.org/abs/2305.14488`. The author of this thesis contributed substantially to the first draft of this work, especially on tightness arguments and the application of lookdown representations.

The work presented in Chapter 6-7 is a collaboration between the author and his collaborators Phil Xie, Salvador Chulian and Victor Perez-Garcia, and is based on a pre-print submitted to PLoS computational biology, see `https://www.biorxiv.org/content/10.1101/2023.07.06.547953v1`. The author of this thesis took a leading role in the development of the project and conducted all of the statistical and mathematical analysis involved.

Terence Tsui Ho Lung
November 17, 2023

Acknowledgements

First, I would like to thank my supervisor Alison Etheridge who has provided me with superb academic, professional and personal support at every step of my DPhil journey. Her rigour as an academic, vision as my career coach and humour as my mentor made my time at Oxford very colourful.

I would also like to thank my collaborators Tom Kurtz, Ian Letter, Peter Ralph, Phil Xie, Salvador Chulian and Victor Perez-Garcia. Their commitment to our collaboration enabled me to do interesting translational research, and they taught me how to become a successful academic.

I would like to thank my colleagues in Oxford and my peers during my stay at the Hausdorff Institute of Mathematics. I would like to thank Adrian, Bastian, Kim, and many more for their support and friendship at the statistics department, and Jere Koskela and Jaromir Sant for their feedback and suggestions on my computational work.

I would like to thank all the institutions who funded and supported my research. They include the EPSRC Centre for Doctoral Training in Mathematics of Random Systems: Analysis, Modelling and Simulation (EP/S023925/1), the Hausdorff Institute of Mathematics under the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2047/1 – 390685813, the Mathematical Oncology Laboratory under Ministerio de Ciencia e Innovación, Spain (grant numbers PID2019-110895RB-I00), St. John's College Oxford and the Rhodes Trust.

Finally, I would like to thank my dear friends and family members who supported me during the pandemic. I would like to thank Joshua Low and Amer Baroudi for hosting me at their home during the second lockdown, Peter Holderrieth and Edward Yee for their support and friendship, my partner Cherie for her constant encouragement, and most importantly, my brother Keith, my mom and my late father, whose love and warmth provide me with the strength to overcome life's many obstacles.

Contents

Notations for Chapters 2-5 **vii**

Notations for Chapters 6-7 **ix**

1 Introduction **1**
1.1 Two perspectives of mathematical genetics 1
1.2 Research Overview 3
1.3 Structure of the Thesis 9

2 Branching Particle Systems with Local Regulation: Model & Heuristics **11**
2.1 Introduction 11
2.2 Model and main results 16
2.3 Examples and applications 34
2.4 Heuristics 42

3 Branching Particle Systems with Local Regulation: 2-Step Convergence Proof **47**
3.1 Preliminaries 47
3.2 Proof of Theorem 2.10: convergence for the nonlocal process 49
3.3 Convergence of some nonlocal equations to classical PDEs 57

4 Branching Particle Systems with Local Regulation: Simultaneous Convergence Proof **68**
4.1 Overview 68
4.2 Moment bounds for $\rho_\epsilon * \eta$ 69
4.3 Continuity estimates for $\rho_\epsilon * \eta$ 77
4.4 Identification of the limit 82

5 Branching Particle Systems with Local Regulation: Lookdown Representation **89**
5.1 The lookdown process 89
5.2 Proofs of results for the lookdown process and ancestral lineages 103

6 FRIME - A Markov Model for Cell-free DNA Fragmentomics **114**
6.1 Introduction 114
6.2 Materials and methods 118
6.3 Results 129

6.4 Discussion 137

7 FRIME - A Mathematical Perspective **142**
7.1 Definition of the FRIME Processes 142
7.2 Stationary Profile of the FRIME Process 146
7.3 Sample Mean and Hydrodynamic Limit of FRIME Processes 148
7.4 Solutions to the FRIME equations 151

8 Conclusion **157**
8.1 Further Research Questions 158
8.2 Final Remarks 162

A Appendix **165**
A.1 Markov Mapping Theorem 165
A.2 Technical Lemmas on Probability Kernels 171

B FRIME - Appendix **176**
B.1 Data pre-processing techniques 176
B.2 Supplementary Figures 177

Notations for Chapters 2-5

$(X(t))_{t\geq 0}$	Population process with one unit of mass at the location of each individual.
$p_{\epsilon^2}(x)$	Gaussian interaction kernel with *interaction distance* $\epsilon > 0$
$\epsilon_\gamma, \epsilon_r, \epsilon_F$	Interaction distance for birth, establishment and net reproductive rate
$\rho_r * X/N$	Convolution of $p_{\epsilon_r^2}(x)$ against X/N, similarly defined for $\rho_\gamma * X/N$, $\rho_F * X/N$
$\gamma(x, m)$	Birth rate of adults at location x with local population density $m = \rho_\gamma * X/N$.
$r(y, m)$	Establishment probability of a juvenile offspring at birth location y with local population density m
$\mu(x, m)$	Death rate of adults at location x with local population density m.
$\theta, 1/N$	Time scale of events and mass of an individual under scaling limit
α	$\lim_{N\to\infty} \theta/N$
$F(x, m)$	$\theta(r(x, m)\gamma(x, m) - \mu(x, m))$
$q_\theta(x, dy)$	Gaussian kernel for location of juvenile offspring with parent at x
$\vec{b}(x)/\theta, \mathbf{C}(x)/\theta$.	Mean and covariance matrix of $q_\theta(x, dy)$
$(\eta^N(t))_{t\geq 0}$ . . .	Population process under re-scaling, i.e. $(\eta^N(t))_{t\geq 0} = (X(\theta t)/N)_{t\geq 0}$
$\gamma(x, \eta)$	Short hand for $\gamma(x, \rho_\gamma * \eta^N)$, similar notation for r, F, μ
$\mathcal{B}$	Operator for the diffusive limit of offspring dispersion, $$\mathcal{B}f(x) = \sum_{ij} \mathbf{C}(x)_{ij}\partial_{x_i}\partial_{x_j} f(x) + \sum_i \vec{b}(x)_i \partial_{x_i} f(x)$$
$\mathcal{B}^*$	Adjoint operator of $\mathcal{B}$
$(L_s)_{s=0}^T$	Ancestral lineage of an individual up to time T in the past
$(Q_s)_{s\geq 0}, \mathcal{L}_{\varphi_{T-s}}$.	Time inhomogeneous semi-group and generator of ancestral lineage motion
$\mathcal{P}^N, \mathcal{P}^\infty$	Generator of the spatial model $(\eta^N(t))_{t\geq 0}$ and its scaling limit $(\eta^\infty(t))_{t\geq 0}$, see Lemma 3.5
$\xi(t), \xi^N(t)$ . . .	Lookdown process corresponding to $\eta(t), \eta^N(t)$, also known as the spatial-level process

$b_\theta(x,\eta)$ . . . . . Local net difference between reproduction and death of η at x,

$$b_\theta(x,\eta) := \theta\left(\gamma(x,\eta)\int_{\mathbb{R}^d} r(y,\eta)q_\theta(x,dy) - \mu_\theta(x,\eta)\right)$$

$c_\theta(x,\eta)$ . . . . . Local rate of production of successful offspring,

$$c_\theta(x,\eta) := \frac{\theta}{N}\gamma(x,\eta)\int_{\mathbb{R}^d} r(y,\eta)q_\theta(x,dy)$$

A^N, A . . . . . Generator of the lookdown processes and its limit, see Definition 5.1, 5.2.

$X_a(t), U_a(t)$ . . Location and level of the lineage with label a in the lookdown process

Notations for Chapters 6-7

C_F, α_F	Fragmentation speed and fragmentation power of the FRIME model
r, a, b	Fragmentation ratio for FRIME model, with $r \sim Beta(a, b)$
L	Maximum length of cfDNA in the FRIME model
C_I	Immigration speed of the FRIME model
$\mathcal{E}(x), \ell$	Exit function and length of immediate clearance for the FRIME model
α_E, B_E	Exit power and exit boundary of the FRIME model
$(F_X(t))_{t\geq 0}$ . . .	Fragmentation with Exit process
$(F_{IX}(t))_{t\geq 0}$ . . .	FRIME process
$\mathbf{U}^{FI}_{stat}$	Stationary Distribution of the FRIME process
$(\mu^n(t))_{t\geq 0}$	FRIME process with mass rescaled by $1/n$
$(u_t(x))_{t\geq 0}$	Hydrodynamic limit of FRIME processes, also solution to the FRIME equation (7.23)
$S(x)$	Stationary distribution of the FRIME equation

1 Introduction

1.1 Two perspectives of mathematical genetics

The history of our understanding of genetics is a remarkable journey that begins with ancient observations of heredity and culminates with the burgeoning field of genomics today. Along this journey, mathematics has played a significant role in providing qualitative and quantitative descriptions of biological phenomena by modelling the dynamics of genes at both population and microscopic scale.

Early human societies recognised the phenomenon of heredity through the adage "like father, like son." However, it wasn't until the mid-19th century that an Augustinian friar named Gregor Mendel conducted groundbreaking experiments on pea plants, uncovering the fundamental laws of inheritance. He proposed the Law of Segregation, the Law of Independent Assortment, and the Law of Dominance, demonstrating how traits are inherited from parents by offspring. In the early 20th century, R.A. Fisher's work in statistics and mathematical genetics provided probabilistic models for the way in which the frequencies of different types of genes spread through a population. For example, the Wright-Fisher model developed independently by S. Wright and R.A. Fisher provides insights into genetic variation within populations by describing the change in allele (variants of genes resulting in different biological characteristics) frequencies in a population over generations due to the randomness associated with birth and death events. This model

considers the process of random sampling of alleles during birth and death events and can for example predict the probability of fixation or loss of alleles. It has been extensively used to study the effects of genetic drift, migration, and selection on genetic diversity in populations. On the other hand, the Fisher-KPP equation

$$\partial u/\partial t = \Delta u + ru(1-u),$$

with u being the frequency of a genetic type at time t and spatial location x, describes the spatial spread of invasive species, biological invasions, and the front propagation of beneficial or harmful genetic mutations in populations. These models inspired numerous directions of research in mathematical genetics for decades to come, where branching and interacting particle processes play an essential role in modelling the spread of genetic types across space and generations [Mor58; Bru+07; BEV10]. Furthermore, the Fisher-KPP equation can be seen as a scaling limit of individual based models in the spirit of Wright-Fisher models. To produce tractable descriptions of the evolution of traits at population scales, one often passes from individual-based models to a scaling limit to recover information about macroscopic phenomena through solutions to partial differential equations [McK67; KX99; EVY20].

With the discovery of DNA and the advent of genome-sequencing technologies, the study of genes can now be conducted within the human body, in contrast to the scale on which population geneticists typically work. The double helix structure of the DNA proposed by F. Crick and J. Watson in 1953, followed by the deciphering of the genetic code, revealed how human biology can be understood through the lens of polymer dynamics. In particular, the theory of fragmentation processes can be applied to understand the kinetics of DNA depolymerisation [ZM85; BK00]. Furthermore, Next-Generation-Sequencing techniques enable efficient sampling and sequencing of DNA, offering the possibility of personalised medicine via the statistical analysis of genomic data through, for example, Genome-Wide Association Studies or Single-Cell Genomics. As a result, the mathematics of genetics in this new era focuses on the fusion of statistical modelling, computational methodologies and biological interpretation.

In this thesis, we will present examples of how the mathematics of branching and fragmenting particle systems can help understand genetics from both perspectives. In the

first project, we study a new class of branching particle systems with local regulation at the population scale to uncover novel scaling limits in which, as an approximation, we investigate the dynamics of ancestral lineages. In the second project, we study a new class of fragmenting particle systems to simulate and analyse cfDNA fragment profiles in the bloodstream that are sampled through the nascent nanopore sequencing technology. These works have resulted in a 102 page paper submitted to the Electronic Journal of Probability (see `https://arxiv.org/abs/2305.14488`) and another paper submitted to PLoS computational biology (see `https://www.biorxiv.org/content/10.1101/2023.07.06.547953v1`), and this thesis will present both of them in a coherent manner in the framework of interacting particle systems.

1.2 Research Overview

The first part of this thesis uses the mathematical theory of branching particle systems with local interaction to address questions on genealogies in population genetics. This is joint work with Alison Etheridge and Ian Letter at the University of Oxford, Tom Kurtz at the University of Wisconsin-Madison, and Peter Ralph at the University of Oregon. First, we introduce a broad class of spatial models to describe how spatially heterogeneous populations live, die, and reproduce. A novelty of our model is that we explicitly incorporate juvenile phase with the reproductive cycle. We then apply mathematical tools such as the theory of weak convergence to discover three different scaling limits of our population. In particular, to retain information about genealogies of our population models at large scales, we employ a lookdown representation and, in the case of deterministic limiting models, deduce the backwards in time motion of the ancestral lineage of a sampled individual. We observe that knowing the history of the population density is not enough to determine the motion of ancestral lineages in our model, and furthermore, the ancestral lineages can behave wildly differently from those in the Fisher-KPP model. Our work is a first rigorous attempt to model mathematically how a juvenile phase impacts the long term evolution of a population. This approach not only introduces a new class of nonlinear diffusions as possible models for evolutionary

biologists to study the dynamics of population spread, but it also provides geneticists with a new framework in which they can trace genealogies for spatially-structured populations.

The second part of this thesis is an interdisciplinary collaboration with Phil Xie at the Ludwig Institute of Cancer Research at Oxford, Salvador Chulián and Victor M. Pérez-García at the Mathematical Oncology Lab (MOLAB) at the University of Castille-La Mancha to translate Monte Carlo simulation techniques and the theory of fragmentation processes into novel clinical analysis of cell-free DNA fragment profiles. Cell-free circulating DNA (cfDNA) are short sections of DNA that are released by cells undergoing stress or other damaging or pathological events. In the human bloodstream, cfDNAs undergo degradation by enzymes (DNase) through which they are broken into smaller pieces. Finally, they are cleared out constantly through the urinary system. Recently, oncologists have started to use a static snapshot of the cfDNA length distribution, also known as the cfDNA fragment profile, to infer the presence of cancer DNA. However, biologists have little understanding of how the dynamics of cfDNA fragmentation affect the overall fragment length profile. To address this issue, this thesis will present a Markovian model that captures the three leading mechanisms governing cfDNA fragmentation in the bloodstream: FRagmentation, IMmigration, and Exit (FRIME). The FRIME model enables the simulation of cfDNA fragment profiles by sampling from the stationary distribution of Markov processes. Simulation results are compared with mitochondrial and genomic cfDNA fragment profiles, demonstrating consistency with experimental and clinical observations. Finally, mathematical expressions that fit well with simulated FRIME profiles can be proved rigorously with techniques similar to those employed in the first research project. The FRIME simulation framework is a significant step towards understanding DNA fragmentation dynamics in the bloodstream, uncovering potential physico-chemical differences between mitochondrial and genomic cfDNA, and aiding the analysis of liquid biopsy data.

The following sections will introduce the model used in each project in further detail, and highlight the main results to be established in later chapters.

1.2.1 Scaling Limits and Ancestral Lineages of Branching Particle Systems with Local Regulations

Spatial Model. We are interested in the evolution of biological populations at large temporal and spatial scales when the birth and death rates of individuals are locally regulated through forces such as competition. Furthermore, we want to investigate how the introduction of a juvenile stage will impact the overall dynamics. To do so, we think of the population as evolving in continuous time and giving birth to one offspring at a time. We denote the population by X, and assign a mass of $\frac{1}{N}$ to each individual. First consider birth rates, defined by a nonnegative function $\gamma(x, m) : \mathbb{R}^d \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ of location x and local population density m. Local population density is defined as the convolution of X/N with a smooth (non-negative, integrable) kernel $\rho_\gamma(\cdot)$. We write this convolution as $\rho_\gamma * X/N$. Then, when the state of the population is X, an individual at location x gives birth to a single juvenile offspring at rate $\gamma(x, \rho_\gamma * X(x)/N)$. The location of the juvenile is determined by a multivariate Gaussian, with mean $\vec{b}(x)/\theta$ and covariance matrix $\mathbf{C}(x)/\theta$.

Similarly, the establishment probability of an offspring at location y is is $r(y, \rho_r * X(y)/N)$, where $r(y, m) : \mathbb{R}^d \times \mathbb{R}_{\geq 0} \to [0, 1]$ and again $\rho_r * X$ is the convolution of X/N with the smooth kernel ρ_r.

We shall write $\mu_\theta(x, X/N)$ for the per-capita death rate of mature individuals in the population, and define it to be

$$\mu_\theta(x) = \max\left\{0, r(x, \rho_r * X(x)/N)\gamma(x, \rho_\gamma * X(x)/N) - \frac{1}{\theta}F(x, \rho_F * X(x)/N)\right\},$$

where this equation defines $F(x, m) : \mathbb{R}^d \times \mathbb{R}_{\geq 0} \to \mathbb{R}$, and ρ_F is again a smooth kernel. Finally, we assume that the mollifiers $\rho_\gamma, \rho_r, \rho_F$ are parameterised by the interaction distances $\varepsilon_\gamma, \varepsilon_r, \varepsilon_F$, which are bounded above by some $\varepsilon > 0$.

With the notations above, we can define our pre-limit model to be the process with mass scaled by N and time scaled by θ,

$$\left(\eta_t^N\right)_{t\geq 0} := \left(X(\theta t)/N\right)_{t\geq 0},$$

and characterise it through a martingale problem, see Definition 2.4.

Scaling Limits. Under mild regularity assumptions given by Assumption 2.8, we show that when we send $\theta(N) \to \infty, N \to \infty$ in such a way that $\theta/N \to \alpha > 0$ and fix the mollifiers $\rho_\gamma, \rho_r, \rho_F$, our re-scaled models converge to a superprocess with a generator that depends on the local population density, see Theorem 2.10. In particular, the incorporation of a juvenile stage leads to a much wider class of superprocess models which we call *non-linear superprocesses* with all parameters dependent on local density.

On the other hand, if $\theta/N \to 0$, our model converges to a weak solution to a non-local partial differential equation that depends on $\varepsilon_\gamma, \varepsilon_r, \varepsilon_F$. Furthermore, we prove that in some important special cases, by sending $\varepsilon \to 0$, the weak solutions to the non-local partial differential equations converge to weak solutions to classical partial differential equations, see Proposition 2.15.

Finally, in the special case in which the limit is a classical reaction-diffusion equation, we show how to simultaneously scale ε with θ, N in such a way that we have 'one-step' convergence to the pde limit, see Theorem 2.21.

Lookdown Representation. The lookdown representation of our spatial model incorporates an extra level process, so the i-th particle has a spatial location x_i and an associated level $u_i \in [0, N]$. The particle system is now characterised by the measure $\xi^N = \sum_i \delta_{x_i,u_i}$, and we choose the dynamics of the particles carefully such that the corresponding scaled spatial projection of ξ_t, denoted as $\eta_t(\cdot) = \frac{1}{N}\xi_t(\cdot \times [0, N])$ has the same distribution as the spatial model $(\eta_t^N)_{t\geq 0}$ described above.

Specifically, a particle at location x and level u, at rate $2\theta(1 - u/N)\gamma(x, \eta)$, produces an offspring that disperses according to the kernel $q_\theta(x, dy)$ (same as the spatial model). A new level u_1 is sampled uniformly from $[u, N]$, and the parent and the offspring are assigned (in random order) to the levels $\{u, u_1\}$. Finally, each individual level evolves according to the ODE $\dot{u} = c_\theta(x, \eta)u^2 - b_\theta(x, \eta)u$, where

$$b_\theta(x, \eta) := \theta\left(\gamma(x, \eta)\int_{\mathbb{R}^d} r(y, \eta)q_\theta(x, dy) - \mu_\theta(x, \eta)\right),$$

and

$$c_\theta(x, \eta) := \frac{\theta}{N}\gamma(x, \eta)\int_{\mathbb{R}^d} r(y, \eta)q_\theta(x, dy).$$

A detailed explanation of this model can be found in Section 5.1.1, and in Definition 5.1, we characterise the lookdown representation of the model as a solution to a Martingale Problem.

Ancestral Lineages. The lookdown representation provides us with methods to trace individual lines of descent as diffusion processes even in the infinite density limit. In Section 5.1.2, with the use of the lookdown representation, we explicitly construct lines of descent in the pre-limit model, and in Proposition 5.6 provide an explicit expression for lines of descent in the limiting case. As an application, we obtain an explicit expression for the dynamics of ancestral lineages within travelling wave solutions to deterministic semilinear PDEs, see Theorem 2.24. As a corollary, we identify two populations with identical stationary profiles but distinct dynamics for the lines of descent, illustrating the importance of proving joint convergence of spatial dynamics and genealogies, see Section 2.3.4. Moreover, this phenomenon can only be observed if the two populations have different establishment rates, highlighting the significance of a juvenile stage in our model.

1.2.2 FRIME - A Markov Model for cfDNA Fragment Profiles

FRIME Model. Since cfDNA production and digestion occur very quickly, we model the length frequency profile of mt-cfDNA fragments as the long-term equilibrium measure of a Markov model that encompasses fragmentation, immigration, and exit mechanisms. We study the Fragmentation processes with Immigration and Exit (FRIME processes), with parameters $(C_F, \alpha_F, a, b, \ell, L, \mathcal{E}, C_I)$, denoted as $(F_{IX}(t))_{t\geq 0}$, which satisfies the following properties:

- The process starts off with a fragment with length $x_i(0) \sim U(0, L)$, the uniform distribution between $[0, L]$.
- Fragments of length x, at rate $C_F x^{\alpha_F}$, fragment into two pieces of length xr and $x(1-r)$, where $r \sim Beta(a, b)$. Both fragments remain in the system and evolve independently.
- Fragments of length $x < B_E$ leave the system at a size-dependent rate $\mathcal{E}(x)$. All fragments with length below ℓ leave the system instantaneously.

- New fragments immigrate into the system as a Poisson process with intensity C_I. The length z of a new fragment is independently distributed as $z \sim U(0, L)$.

A detailed discussion of the biological basis of the FRIME model can be found in Section 6.2.1, and the algorithm used to simulate FRIME profiles is presented in 6.2.4. Furthermore, the FRIME process can be constructed rigorously as a Markov Process with Immigration, as presented in Definition 7.4.

Numerical Simulation and Clinical Analysis The FRIME model suggests multiple hypotheses on cfDNA biology. Numerical simulations of FRIME processes identify the exit mechanism as the determining factor for the existence and position of a fragmentation profile peak, see Section 6.3.2. This suggests a previously unknown insight that the removal mechanism of mitochondrial cfDNA may be size-dependent, and such dependence could explain the existence of a fragment peak in the cfDNA profile.

Statistically, we show that clinical cfDNA fragment profiles exhibit linear decay in the log-log scale, see Section 6.3.1, and demonstrate that simulated profiles from the FRIME model can fit both mitochondrial and genomic cfDNA fragment profiles from clinical data, see Section 6.3.4. In particular, the best-fitting coefficients for mtDNA and gDNA fragment profiles are different (see Section 6.3.5), which strongly suggests potential physico-chemical differences in the two distinct types of DNA.

Mathematical Theory of FRIME Processes The numerical results from Monte Carlo simulations of FRIME processes are well supported by mathematical theory. In Section 7.1, we defined the measure-valued FRIME process $(F_{IX}(t))_{t\geq 0}$ rigorously as a Markov Process with Immigration of the type introduced in [Haa05]. By a colouring scheme presented in Definition 7.2, the FRIME process can be coupled with, and stochastically dominated by, the Fragmentation with Immigration process introduced in the same paper. This allows one to translate results on stationarity for Fragmentation with Immigration process into that for the FRIME process, see Section 7.2. Finally, by finding the hydrodynamic limit of a specific class of FRIME processes, we prove that the expected evolution of a

FRIME process $(\mu_t)_{t\geq 0}$, defined through the characterisation

$$\left(\int f(x)\mu_t(dx)\right)_{t\geq 0} = (\mathbb{E}[\langle f, F_{IX}(t)\rangle])_{t\geq 0}$$

for all $f \in \mathcal{C}([0,L])$, satisfies a *fragmentation with immigration and exit equation* of the form

$$\begin{aligned}\frac{\partial}{\partial t}\int f(x)\mu_t^\infty(dx) =& C_F \int x^\alpha \mathbb{E}_r\left[\{f(x(1-r)) + f(xr) - f(x)\}\right]\mu_t^\infty(dx) \\ &- \int \mathcal{E}(x)f(x)\mu_t^\infty(dx) + C_I/L\int_0^L f(z)dz,\end{aligned}$$

where $\mathbb{E}_r[g(r)]$ is short hand for the integral expression

$$\mathbb{E}_r[g(r)] = \int_0^1 g(r)\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}r^{a-1}(1-r)^{b-1}dr.$$

If we assume that the solution $\mu_t^\infty(dx)$ has a density, i.e. $\mu_t^\infty(dx) = u_t(x)dx$, then the density $u_t(x)$ solves the equation

$$\frac{\partial}{\partial t}u_t(x) = C_F x^\alpha\left\{\mathbb{E}_r\left[\frac{u_t\left(x/(1-r)\right)}{(1-r)^{\alpha+1}} + \frac{u_t\left(x/r\right)}{r^{\alpha+1}}\right] - u_t(x)\right\} - \mathcal{E}(x)u_t(x) + C_I/L.$$

By studying its stationary solutions, we can establish mathematical expressions that best fit the FRIME fragment profiles established from numerical simulations. See Section 7.3 for a detailed derivation.

1.3 Structure of the Thesis

Chapter 2 introduces the spatial model for locally-regulated populations in detail and presents some of our main results for the model (Theorems 2.10, 2.21, and 2.24) in Section 2.2. Next, in Section 2.3, we discuss a few striking results regarding the behaviour of genealogies in travelling waves, the appearance of periodic "clumps" in seemingly homogeneous population models, and identifiability of the underlying dynamics from a stationary population profile. In Section 2.4, we provide heuristic explanations of why the theorems ought to be true, and some key ideas behind them. The results are illustrated with individual-based simulations, made using SLiM [HM19], but these are provided for visualization and we do not embark on numerical study.

Chapter 3 proves two different scaling limits of our models, and shows how certain limits characterised as solutions to nonlocal partial differential equations can be further scaled to obtain solutions of classical partial differential equations.

Chapter 4 gives, in an important class of special cases, the more difficult proof of convergence when interaction distances simultaneously go to zero as we pass to the limit.

Chapter 5 covers the lookdown representation of our spatial model, and explains how it allows us to mathematically trace the dynamics of ancestral lineages.

Chapter 6 introduces the FRIME model in a computational biology context and discusses its impact on biology through simulation and statistical tests. In Section 6.2, we describe the biological factors that determine the dynamics of cfDNA and propose the FRagmentation with IMmigration and Exit (FRIME) framework, along with an algorithm to implement it. In Section 6.3, we use the FRIME framework to simulate the distribution of cfDNA fragment lengths and test various biological assumptions by adjusting the parameters of the model. We then compare our simulated results with real-world cfDNA data obtained using state-of-the-art sequencing techniques. In Section 6.4, we explore the biological implications of our results.

Chapter 7 puts the study of the FRIME process on a firm mathematical footing. We define the FRIME process rigorously in Section 7.1. We prove existence of stationary distributions of FRIME processes in Section 7.2, and, in Section 7.4, we relate the stationary distribution to solutions of a class of partial differential equations.

Finally, the thesis ends with the concluding Chapter 8 which discusses some potential future research questions and some final remarks.

Section A.1 of the Appendix A explains the Markov Mapping Theorem essential to the lookdown construction. Section A.2 proves some technical lemmas on probability kernels. Sections B.1 and B.2 include supplementary simulations and clinical data analysis results on the FRIME model.

2

Branching Particle Systems with Local Regulation: Model & Heuristics

2.1 Introduction

As one takes a journey, long or short, the landscape changes: forests thicken or thin or change their composition; even in flat plains springtime grasslands host intergrading mosaics of different types of flowers. The aim of this chapter is to introduce and study a broad class of mechanistic spatial models that might describe how spatially heterogeneous populations live, die, and reproduce. Questions that we (start to) address include: How does population density change across space and time? How might we learn about the underlying dynamics from genealogical or genetic data? And, how does genetic ancestry spread across geography when looking back through time in these populations?

Reproduction of individuals naturally leads to spatial branching process models, including branching random walk, branching Brownian motion, and the Dawson-Watanabe superprocesses. However, as a result of the branching assumption (once born, individuals behave independently of one another), a population evolving according to any of these models will either die out or grow without bound and, in so doing, can develop clumps of arbitrarily large density and extent. Our starting point here is an individual-based model of a single species in continuous space in which birth, death, and establishment

may all depend on local population density as well as on spatial location, allowing for stable populations through density-dependent feedback.

Although it is often mathematically convenient to assume that individuals follow Brownian motion during their lifetime, in our model, offspring are thrown off according to some spatial distribution centred on the location of the parent and do not subsequently move. This is particularly appropriate for modelling plant populations, in which this dispersal of offspring around the parent is the only source of spatial motion.

Often models do not distinguish between juveniles and adults, so, for example, the number of adults produced by a single parent is determined only by the degree of crowding at the location of the parent. Although we shall similarly only follow the adult population, in formulating the dynamics of the models we shall distinguish between production of juveniles, which will depend upon the location of the adult, and their successful establishment, which will depend on the location in which a juvenile lands. The result is that not only the absolute number, but also the spatial distribution around their parent, of those offspring that survive to adulthood will depend upon the local population density.

We shall consider three different classes of scaling limits for our model. The first yields a class of (generalised) superprocesses in which coefficients governing both the spatial motion and the branching components of the process can depend on local population density; the second is a corresponding class of deterministic non-local differential equations; and the third are classical PDEs. We measure local population density around a point by convolving with a smooth kernel $\rho(\cdot)$, which may differ for the two stages of reproduction. When the limiting population process is deterministic, it is a (weak) solution of an equation of the form

$$\partial_t \varphi_t(x) = r\left(x, \varphi_t\right) \mathcal{B}^* \left[\varphi_t(\cdot)\gamma\Big(\cdot, \varphi_t\Big)\right](x) + \varphi_t(x) F\left(x, \varphi_t\right), \tag{2.1}$$

where $\varphi_t(x)$ can be thought of as the population density at x (although the limit may be a measure without a density), and $\mathcal{B}^*$ is (the adjoint of) a strictly uniformly elliptic second order differential operator, typically the Laplacian. The dependence of each of the terms r, γ, and F on φ is only through the local density at x, e.g., $F(x, \varphi) = F(x, \rho * \varphi(x))$. We shall be more specific about the parameters below.

By replacing ρ by $\rho^\epsilon(\cdot) = \rho(\cdot/\epsilon)/\epsilon^d$, we can also scale the "width" of the region over which we measure local population density. When the population follows (2.1), we expect that if we take a second limit of $\epsilon \to 0$, thus scaling the kernels appearing in r, γ, and F and making interactions pointwise, we should recover a nonlinear PDE. We verify that this is indeed the case in two important examples: a special case of the porous medium equation with a logistic growth term, in which the limiting equation takes the form

$$\partial_t \varphi = \Delta(\varphi^2) + \varphi(1 - \varphi); \tag{2.2}$$

and a wide class of semi-linear PDEs of the form

$$\partial_t \varphi = \mathcal{B}^* \varphi + \varphi F(\varphi), \tag{2.3}$$

which includes the Fisher-KPP equation and the Allen-Cahn equation. Equations of this form have been studied extensively in the context of spatial ecology (see for instance [LL23] and [CC04]) and in many other fields; for instance, [GG22] derive a stochastic version of (2.3) to describe abundances of mutant bacteria strains along the human gut, while [LBS22] study the effects of nonlinear diffusion on long-term survival of a lattice-based interacting particle system. However, we do not study the effect of movement of adults, which can additionally affect the limiting equations: see for instance [Hol+94] or [PB23].

It is of interest to understand under what conditions we can replace the two-step limiting process described above by one in which we simultaneously scale the kernels and the other parameters in our population model to arrive at the PDE limit. This is mathematically much more challenging, but we establish such one-step convergence in cases for which the limit is a classical reaction-diffusion equation of the form (2.3) with $\mathcal{B} = \Delta$, and ρ is a Gaussian density. We allow a wide class of reaction terms, F, so that the Fisher-KPP equation (that is equation (2.3) with $\mathcal{B} = \Delta$ and $F(\varphi) = 1 - \varphi$) emerges as a special case.

Such results on (one-step) convergence to reaction-diffusion equation limits have been achieved for a variety of interacting particle systems. Following the now classical contributions of [DFL86; DP06; Oel85], much of this work has focused on lattice based models with one particle per site, or on systems with a fixed number, N, of interacting diffusions as $N \to \infty$. For systems of proliferating particles, as considered for example

by [Oel89; FLO19; FH21], an additional challenge (also apparent in our models), is the control of concentration of particles. We follow [Oel89; FLO19] in considering 'moderate interactions', meaning that the number of individuals in the neighbourhood over which we measure local population density tends to infinity, whereas [FH21] also consider the situation in which that number remains finite. We refer to [FH21] for a more thorough literature review, but note that both our model and scaling differ from those considered in the body of work discussed there: whereas in those settings, the only scalings are the number of particles in the system and the size of the neighbourhood over which individuals interact with one another, in keeping with the vast literature on continuous state branching models, we also scale time and so must ensure that births are adequately compensated by deaths to prevent the population from exploding.

The history of a natural population is often only accessible indirectly, through patterns of genetic diversity that have been laid down; from genetic data, one can try to infer the genealogical trees that relate individuals in a sample from the population, and these have been shaped by its history [see e.g., NA93; Kel+19]. It is therefore of interest to establish information about the distribution of genealogical trees under our population model, which we do with a lookdown construction. Lookdown constructions were first introduced in [DK96] to provide a mechanism for retaining information about genealogical relationships between individuals sampled from a population evolving according to the Moran model when passing to the infinite population limit. Since then, they have been extended to a wide range of models. Of particular relevance to our work here are the papers of [KR11] and [EK19], in which lookdown constructions are provided for a wide variety of population models, including spatially structured branching processes.

In general, even armed with a lookdown construction, calculation of relevant statistics of the genealogy remains a difficult question. However, in special circumstances, some progress can be made. As an illustration, we shall consider a scenario that has received a great deal of attention in recent years, in which a population is expanding into new territory as a travelling wave. In Section 2.3.2 we shall describe the motion of a single ancestral lineage relative to three different (deterministic) wavefronts across $\mathbb{R}^1$.

Most work on the topic of "waves" of expanding populations has focused on models that caricature the classical Fisher-KPP equation with a stochastic term, i.e.

$$dw = \Big(\Delta w + sw(1-w)\Big)dt + \sqrt{\frac{\alpha(w)}{N}}W(dt, dx),$$

where W is space-time white noise, and N is a measure of the local population density. The coefficient $\alpha(w)$ is generally taken to be either w, corresponding to a superprocess limit, or $w(1-w)$ giving a spatial analogue of a Wright-Fisher diffusion. Starting with the pioneering work of [Bru+06], a considerable body of evidence has been amassed to underpin the conjecture that for this, and a wide class of related models, genealogies converge on suitable timescales in the infinite density limit to a Bolthausen-Sznitman coalescent. This reflects the fact that, for this equation, ancestral lineages become trapped in the wavefront, where the growth rate of the population is highest. Once there, they will experience rapid periods of coalescence corresponding to significant proportions of individuals in the front being descended from particularly reproductively successful ancestors.

If one replaces the logistic growth term of the classical Fisher-KPP equation with a nonlinearity that reflects cooperative behaviour in the population, such as

$$wF(w) = w(1-w)(Cw-1), \tag{2.4}$$

then, for sufficiently large C (strong cooperation), the nature of the deterministic wave changes from "pulled" to "pushed", [BHK18; BHK21], and so the genealogies will be quite different from the Fisher-KPP case. For example, [EP22] show that for a discrete space model corresponding to this nonlinearity with $C > 2$, after suitable scaling, the genealogy of a sample converges not to a Bolthausen-Sznitman coalescent, but to a Kingman coalescent. The reason, roughly, is that ancestral lineages settle to a stationary distribution relative to the position of the wavefront which puts very little weight close to the 'tip' of the wave, so that when ancestral lineages meet it is typically at a location in which population density is high, where no single ancestor produces a disproportionately large number of descendants in a short space of time.

The shape of the wave is not determined solely by the reaction term. For example, as a result of the nonlinear diffusion, for suitable initial conditions, the solution to the one-dimensional porous medium equation with logistic growth (2.2) converges to a travelling

wave with a sharp cut-off; i.e., in contrast to the classical Fisher KPP equation, the solution at time t vanishes beyond $x = x_0 + ct$ for some constant wavespeed $c > 0$ [KR04]. As a first step towards understanding what we should expect in models with nonlinear diffusion, one can ask about the position of an ancestral lineage relative to the wavefront in the deterministic models. In Section 2.3.2 we shall see that in our framework, even with logistic growth, the nonlinear diffusion corresponding to the porous medium equation results in a stationary distribution for the ancestral lineage that is concentrated behind the wavefront, leading us to conjecture that in the stochastic equation the cooperative behaviour captured by the nonlinear diffusion will also result in a qualitatively different pattern of coalescence to that seen under the stochastic Fisher-KPP equation. Indeed, we believe that it should be feasible to show that in an appropriate limit one recovers a Kingman coalescent.

2.2 Model and main results

Our model is one of individuals distributed across a continuous space which we shall take to be $\mathbb{R}^d$. For applications, $d = 1$ or $d = 2$ (or even $d = 3$ for cells within the body), but our main results apply more generally. At time zero, the population is distributed over a bounded region, with $\mathcal{O}(N)$ individuals per unit area in that region, so the total number of individuals will also be $\mathcal{O}(N)$. The population changes in continuous time, and we encode the state of the population at time t by a counting measure $X(t)$, which assigns one unit of mass to the location of each individual.

Population dynamics are controlled by three quantities, birth (γ), establishment (r), and death (μ), each of which can depend on spatial location and local population density in a way specified below. Each individual gives birth at rate γ to a single (juvenile) offspring, which is dispersed according to a kernel $q(x, \cdot)$ away from the location x of the parent. We assume that q is the density of a multivariate Gaussian, allowing a nonzero mean and anisotropic variance. Both the mean and covariance of q can change across space, but do not depend on population density. The offspring does not necessarily survive to be counted in the population: it "establishes" with probability r, or else it dies immediately. Independently, each individual dies with rate μ.

We aim to capture universal behaviour by passing to a scaling limit. Specifically, we shall take the "density", N, to infinity, and also scale time by a factor of $\theta = \theta(N)$, in such a way that defining $\eta^N(t) = X(\theta t)/N$, the process $\{\eta^N(t)\}_{t\geq 0}$ will converge to a suitable measure-valued process as N and θ tend to infinity, with the nature of the limit depending on how they tend to infinity together. Evidently, we also need to scale the dispersal kernel if we are to obtain a nontrivial limit, for which we use $q_\theta(x,\cdot)$, the density of the multivariate Gaussian obtained by multiplying the mean and variance components of $q(x,\cdot)$ by $1/\theta$.

Birth, establishment, and death can depend on the location of the individual and the local population density. Since we would like the population density to scale with N, these are functions of X/N, i.e., the counting measure with mass $1/N$ placed at the location of each individual. First consider birth rates, defined by a nonnegative function $\gamma(x,m) : \mathbb{R}^d \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ of location x and local population density m. Local population density is defined as the convolution of X/N with a smooth (non-negative, integrable) kernel $\rho_\gamma(\cdot)$. We write this convolution as $\rho_\gamma * X/N$. Then, when the state of the population is X, an individual at location x gives birth to a single juvenile offspring at rate $\gamma(x, \rho_\gamma * X(x)/N)$. Similarly, the establishment probability of an offspring at location y is is $r(y, \rho_r * X(y)/N)$, where $r(y,m) : \mathbb{R}^d \times \mathbb{R}_{\geq 0} \to [0,1]$ and again $\rho_r * X$ is the convolution of X/N with the smooth kernel ρ_r.

We shall write $\mu_\theta(x, X/N)$ for the per-capita death rate of mature individuals in the population. In order for the population density to change over timescales of order θ, we should like the net per capita reproductive rate to scale as $1/\theta$. In classical models, in which r, γ, and μ are constant, this quantity is simply $r\gamma - \mu$. Here, because production of juveniles and their establishment are mediated by population density measured relative to different points, the net reproductive rate will take a more complicated form. In particular, the total rate of production of mature offspring by an individual at x will be

$$\gamma\Big(x, \rho_\gamma * X(x)/N\Big) \int r\Big(y, \rho_r * X(y)/N\Big) q_\theta(x, dy). \tag{2.5}$$

Nonetheless, it will be convenient to define the death rate μ_θ in terms of its deviation from $r\gamma$. To this end, we define the death rate of an individual at x to be

$$\mu_\theta(x) = \max\left\{0, r(x, \rho_r * X(x)/N)\gamma(x, \rho_\gamma * X(x)/N) - \frac{1}{\theta} F(x, \rho_F * X(x)/N)\right\}, \tag{2.6}$$

where this equation defines $F(x,m) : \mathbb{R}^d \times \mathbb{R}_{\geq 0} \to \mathbb{R}$, and ρ_F is again a smooth kernel. The function F is nearly the net per capita reproductive rate, scaled by θ, and would be equal to it in a nonspatial model; but, as can be seen from (2.5), differs because an offspring's establishment probability is measured at their new location rather than that of their parent. For the most part, we work with F instead of μ_θ.

So, each of the three demographic parameters r, γ, and F, depends on local density, measured by convolution with a smooth kernel, each of which can be different. As a result, death rate depends (in principle) on population densities measured in three different ways, so that we could write $\mu_\theta(x) = \mu_\theta(x, \rho_\gamma * X(x)/N, \rho_r * X(x)/N, \rho_F * X(x)/N)$. This may seem unnecessarily complex. However, not only is it natural from a biological perspective, it also turns out to be convenient for capturing nontrivial examples in the scaling limit.

Remark 2.1 *Although this model allows fairly general birth and death mechanisms, there are a number of limitations. Perhaps most obviously, to simplify the notation individuals give birth to only one offspring at a time, although this restriction could be easily lifted [as in Section 3.4 of EK19]. Furthermore, individuals do not move during their lifetime, and the age of an individual does not affect its fecundity or death rate. Finally, there is no notion of mating (although limitations on reproduction due to availability of mates can be incorporated into the birth rate, γ), so the lineages we follow will be uniparental. For these reasons, the model is most obviously applicable to bacterial populations or selfing plants, although we do not anticipate that incorporation of these complications will change the general picture.*

For each N and θ, we study primarily the process with mass scaled by N and time scaled by θ,

$$\left(\eta_t^N\right)_{t\geq 0} := \left(X(\theta t)/N\right)_{t\geq 0},$$

which takes values in the space of càdlàg paths in $\mathcal{M}_F(\mathbb{R}^d)$ (the space of finite measures on $\mathbb{R}^d$ endowed with the weak topology). In fact η_t^N will be a purely atomic measure comprised of atoms of mass $1/N$.

Notation 2.2 *Expressions like $\gamma(x, \rho_\gamma * \eta(x))$ will appear repeatedly in what follows. To make formulae more readable, we overload notation to define*

$$\gamma(x, \eta) := \gamma(x, \rho_\gamma * \eta(x)),$$

and similarly write $r(x, \eta)$ for $r(x, \rho_r * \eta(x))$, $F(x, \eta)$ for $F(x, \rho_F * \eta(x))$, and $\mu_\theta(x, \eta)$ for the expression of equation (2.6). When convenient, we may also suppress the arguments completely, writing simply γ, r, F, and μ_θ for these quantities.

Terminology 2.3 *In our prelimiting model, the population is represented by a point measure in which each individual is assigned a mass $1/N$. We use the term "population density" for this process, as it is supposed to measure population size relative to a nominal occupancy of N individuals per unit area. There is no implication that the measure representing the population is absolutely continuous with respect to Lebesgue measure; indeed in the prelimit it is certainly not.*

In summary, at each time t, η_t^N is purely atomic, consisting of atoms of mass $1/N$ (which are the individuals). At instantaneous rate $\theta\gamma(x, \eta_t^N)N\eta_t^N(dx)$ an offspring of mass $1/N$ is produced at location x; it disperses to a location y offset from x by an independent Gaussian random variable with mean $\vec{b}(x)/\theta$ and covariance matrix $\mathbf{C}(x)/\theta$, and once there establishes instantaneously with probability $r(y, \eta_t^N)$, or else dies. At instantaneous rate $\theta\mu_\theta(x, \eta_t^N)N\eta_t^N(dx)$ an individual at location x dies. Note that the process $\left(\eta_t^N\right)_{t\geq 0}$, which records numbers and locations of adult individuals, is just a scaled spatial birth and death process. If, for example, we insist that $\gamma(x, m)$ is bounded, then existence (and in particular non-explosion) is guaranteed by comparison with a pure birth process. We do not dwell on this, as we shall require more stringent conditions if we are to pass to the limit as θ and N tend to infinity.

It is convenient to characterise the process as a solution to a martingale problem. We write $C_b^\infty(\mathbb{R}^d)$ for the space of bounded smooth functions on $\mathbb{R}^d$, and, where convenient, we write $\langle f, \eta\rangle = \int_{\mathbb{R}^d} f(x)\eta(dx)$.

Definition 2.4 (Martingale Problem Characterisation) *For each value of N and θ, and each purely atomic $\eta_0^N \in \mathcal{M}_F(\mathbb{R}^d)$ with atoms of mass $1/N$, $(\eta_t^N)_{t\geq 0}$ is the (scaled) empirical measure of a birth-death process with càdlàg paths in $\mathcal{M}_F(\mathbb{R}^d)$ for which, for all $f \in C_b^\infty(\mathbb{R}^d)$, writing $q_\theta(x, dy)$ for the Gaussian kernel with mean $x + \vec{b}(x)/\theta$ and covariance $\mathbf{C}(x)/\theta$,*

$$\begin{aligned} M_t^N(f) := \langle f, \eta_t^N\rangle - \langle f, \eta_0^N\rangle \\ - \int_0^t \left\{ \left\langle \left(\int \theta \left(f(z)r(z,\eta_s^N) - f(x)r(x,\eta_s^N)\right) q_\theta(x,dz)\right) \gamma(x,\eta_s^N), \eta_s^N(dx)\right\rangle \right. \\ \left. + \left\langle f(x)F(x,\eta_s^N), \eta_s^N(dx)\right\rangle \right\} ds \end{aligned} \tag{2.7}$$

is a martingale (with respect to the natural filtration), with angle bracket process

$$\begin{aligned} \left\langle M^N(f)\right\rangle_t = & \frac{\theta}{N}\int_0^t \left\{ \left\langle \gamma(x,\eta_s^N) \int f^2(z) r(z,\eta_s^N) q_\theta(x,dz), \eta_s^N(dx)\right\rangle \right. \\ & \left. + \left\langle \mu_\theta(x,\eta_s^N) f^2(x), \eta_s^N(dx)\right\rangle \right\} ds. \end{aligned} \tag{2.8}$$

The angle bracket process (or, "conditional quadratic variation") is the unique previsible process making $(M^N(f)_t)^2 - \langle M^N(f)\rangle_t$ a martingale with respect to the natural filtration. It differs from the usual quadratic variation (usually denoted $[M^N(f)]_t$) because the process has jumps; for the (continuous) limit the two notions will coincide. The use of angle brackets for both integrals and this process is unfortunately standard but should not cause confusion, since the angle bracket process always carries a subscript for time.

The form of (2.7) and (2.8) is explained in Section 2.4. Note that since (juvenile) individuals are produced at rate $N\gamma\eta$, but each has mass $1/N$, these factors of N cancel in (2.7). Under our scaling, N and $\theta = \theta(N)$ will tend to infinity in such a way that $\alpha := \lim_{N\to\infty} \theta(N)/N$ exists and is finite. From the expression (2.8) it is easy to guess that whether the limiting processes will be deterministic or stochastic is determined by whether α is zero or nonzero.

It is convenient to record some notation for the generator of the diffusion limit of a random walk with jump distribution determined by $q_\theta(x, dy)$.

Definition 2.5 (Dispersal generator) *As above, we define the dispersal kernel, $q_\theta(x, dy)$, to be the density of a multivariate Gaussian with mean $\vec{b}(x)/\theta$ and covariance matrix*

$\mathbf{C}(x)/\theta$ *(although often we omit the dependence of* $\vec{b}$ *and* $\mathbf{C}$ *on* x*). Furthermore, we define for* $f \in C_b^\infty(\mathbb{R}^d)$*,*

$$\mathcal{B}f(x) = \frac{1}{2}\sum_{ij} \mathbf{C}(x)_{ij}\partial_{x_i}\partial_{x_j}f(x) + \sum_i \vec{b}(x)_i \partial_{x_i} f(x) \tag{2.9}$$

and denote the adjoint of $\mathcal{B}$ *by*

$$\begin{aligned}\mathcal{B}^* f(x) &= \frac{1}{2}\sum_{ij}\partial_{x_i}\partial_{x_j}(\mathbf{C}(x)_{ij}f(x)) - \sum_i \partial_{x_i}(f(x)\vec{b}(x)_i) \\ &= \frac{1}{2}\sum_{ij} C_{ij}(x)\partial_{x_i}\partial_{x_j}f(x) + \frac{1}{2}\sum_i \left(\sum_j \partial_{x_j}C_{ij}(x) - \vec{b}_i(x)\right)\partial_{x_i}f(x) \\ &\quad + \left(\frac{1}{2}\sum_{ij}\partial_{x_i}\partial_{x_j}C_{ij}(x) - \sum_i \partial_{x_i}\vec{b}_i(x)\right) f(x).\end{aligned}$$

Remark 2.6 $\mathcal{B}$ *is defined so that*

$$\theta \int (f(y) - f(x))\, q_\theta(x, dy) \to \mathcal{B}f(x) \qquad \text{as } \theta \to \infty.$$

Remark 2.7 *An equivalent way to describe the model would be to say that when the state of the population is* η*, an individual at* x *gives birth at rate*

$$\theta\gamma(x,\eta)\int r(y,\eta)q(x,dy),$$

and that offspring disperse according to the kernel

$$q_\theta^{\mathfrak{m}}(x,\eta,dy) := \frac{r(y,\eta)q_\theta(x,dy)}{\int r(z,\eta)q_\theta(x,dz)}.$$

Clearly, the random walk driven by this dispersal kernel is biased towards regions of higher establishment probability. For comparison with future results, it is interesting to write down the limiting generator:

$$\lim_{\theta\to\infty}\theta\int (f(y)-f(x))q_\theta^{\mathfrak{m}}(x,\eta,dy) = \frac{\mathcal{B}\left[f(\cdot)r(\cdot,\eta)\right](x) - f(x)\mathcal{B}\left[r(\cdot,\eta)\right](x)}{r(x,\eta^N)}. \tag{2.10}$$

In the simplest case of unbiased isotropic dispersal (i.e., $\vec{b} = 0$ *and* $\mathbf{C} = 2I$*),* $\mathcal{B} = \Delta$*, and so (2.10) is equal to*

$$\Delta f(x) + 2\nabla f(x)\cdot\nabla \log r(\cdot, \rho_r * \eta(\cdot))(x).$$

One might guess that the spatial motion described by following the ancestral lineage of an individual back through time would be described (in the limit) by the adjoint of this generator. However, we will see in Section 2.2.2 that this is not in fact the case.

In order to pass to a scaling limit, we will need to impose some conditions on the parameters of our model.

Assumptions 2.8 *We shall make the following assumptions on the parameters of our model.*

Dispersal generator: *We assume that*

1. $\vec{b}(x)$ *and* $\mathbf{C}(x)$ *are* α*-Hölder continuous for some* $\alpha \in (0,1]$ *and uniformly bounded in each component, and*

2. *the operator* $\mathcal{B}$ *is uniformly strictly elliptic, i.e.,* $\inf_x \inf_{y:\|y\|=1} \sum_{ij} y_i C(x)_{ij} y_j > 0$.

Reproduction parameters: *We assume that*

3. *The function* $F(x,m)$ *satisfies*

   (a) $F(x,m)$ *is locally Lipschitz in* m*;*

   (b) $F(x,m)$ *is uniformly bounded above (but not necessarily below);*

   (c) *for each fixed* m, $\sup_{x\in\mathbb{R}^d} \sup_{k\leq m} |F(x,k)| < \infty$*;*

4. *The functions* $r(x,m)$, $\gamma(x,m)$ *have bounded first and second derivatives in both arguments;*

5. $\gamma(x,m)$ *is uniformly bounded;*

6. *For each* $f \in C_b^2(\mathbb{R}^d)$, *there is a* C_f *such that*

$$|\gamma(x,\eta)\theta \int (r(y,\eta)f(y) - r(x,\eta)f(x))q_\theta(x,dy)| \leq C_f(1+|f(x)|)$$

   for all $x \in \mathbb{R}^d$ *and* $\eta \in \mathcal{M}_F(\mathbb{R}^d)$. *Furthermore,* C_f *only depends on the norm of the first two derivatives of* f, *i.e.,*

$$C_f = C(\sup_x \sup_{\|z\|=1} \max(\sum_i z_i \partial_{x_i} f(x), \sum_{ij} z_i z_j \partial_{x_i x_j} f(x))).$$

To keep expressions manageable, we shall also assume that

$$\mu_\theta(x) = r(x,\eta)\gamma(x,\eta) - \frac{1}{\theta}F(x,\eta),$$

that is, this expression is non-negative so that there is no need to take the maximum with zero in (2.6). (This is anyway true for sufficiently large θ.)

Since we take bounded f, for most situations the bound $C_f(1+|f(x)|)$ in Condition 6 above can be safely replaced simply by C_f; however, this will be useful in certain situations where we consider a sequence of f with increasing upper bounds. We now give two concrete situations in which Condition 6 is satisfied. The proof is in Section 3.1.

Lemma 2.9 *Assume that Conditions 2.8 are satisfied, except for Condition 6. If either*

1. *$|\partial_{x_i} r(x,\eta)|$ and $|\partial_{x_i x_j} r(x,\eta)|$ are uniformly bounded for $x \in \mathbb{R}^d$, $\eta \in \mathcal{M}_F(\mathbb{R}^d)$;*

2. *or, $m^2\gamma(x,m)$ is uniformly bounded and there exists $C < \infty$ such that for θ sufficiently large, and all $x \in \mathbb{R}^d$, $\eta \in \mathcal{M}_F(\mathbb{R}^d)$,*

$$\theta \int \Big(\rho_r * \eta(y) - \rho_r * \eta(x)\Big) q_\theta(x, dy) \le C \rho_\gamma * \eta(x),$$

and

$$\theta \int \Big(\rho_r * \eta(y) - \rho_r * \eta(x)\Big)^2 q_\theta(x, dy) \le C (\rho_\gamma * \eta(x))^2,$$

then Condition 6 is also satisfied.

The purpose of the conditions that we have placed on the reproduction parameters is to ensure that the net per capita reproduction rate (before time scaling) is order $1/\theta$. As remarked above, because of the non-local reproduction mechanism, it no longer suffices to assume that $r(x,\eta)\gamma(x,\eta) - \mu_\theta(x)$ is of order $1/\theta$. Perhaps the simplest example in which we can see this is the case where $\gamma \equiv 1$ and $F \equiv 0$, so that $\mu_\theta = r$, and $\eta = \delta_x$ (i.e., the population has all individuals at a single location), so that $\rho_r * \eta(y) = \rho_r(y)$. In this case, the mean rate of change of the total population size is $\int (r(y, \rho_r(y)) - r(x, \rho_r(x))) q_\theta(x, dy)$; the first condition of Lemma 2.9 would ensure this is of order $1/\theta$.

If $r(x,m)$ is independent of m, then the conditions are easy to satisfy; they just require some regularity of r as a function of x. Condition 1 of Lemma 2.9 is also satisfied if

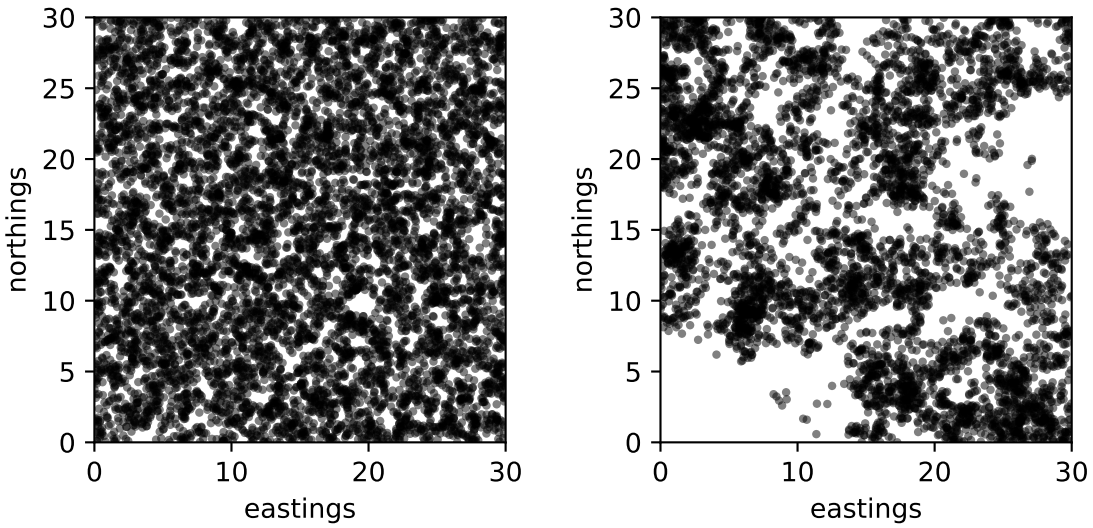


Figure 2.1: Snapshots of two simulations, with small $\alpha = \theta/N$ (left) and large $\alpha = \theta/N$ (right). Simulations are run with a Fisher-KPP-like parameterization: birth and establishment are constant, while death increases linearly with density, at slope $1/\theta$. Left: $\alpha = 0.1$. Right: $\alpha = 10$. Other parameters were the same: dispersal and interactions distance were set to 1, and the equilibrium density is 10 individuals per unit area.

for example $\|\nabla \rho_r\| \leq C\rho_r$ and $m\partial_m r(x,m)$, $m^2\partial_{mm} r(x,m)$ are bounded. This is the case, for instance, if ρ_r decays exponentially. On the other hand, it might seem more natural to take ρ_r to be a Gaussian density with parameter σ_r, say. Then, as we check in Lemma A.5, Condition 2 of Lemma 2.9 is satisfied if ρ_γ is also Gaussian with parameter σ_γ and $\sigma_\gamma > \sigma_r$. For large enough θ, this condition guarantees that $\sigma_r + 1/\theta < \sigma_\gamma$, so that the establishment probability of a juvenile is controlled by individuals that are already 'felt' by the fecundity-regulating kernel ρ_γ at the location of their parent.

2.2.1 Scaling limits of the population process

Our main results depend on two dichotomies: Is the limiting process deterministic or a (generalized) superprocess? And, are interactions pointwise in the limit or nonlocal? See Figure 2.1 for snapshots of the population from direct simulation of the process using SLiM [HM19] illustrating this first dichotomy. Below we have results for deterministic limits with pointwise and nonlocal interactions, and for superprocess limits with nonlocal interactions.

Scaling limits with nonlocal interactions: Recall that the process $(\eta_t^N)_{t\geq 0}$ takes its values in the space $\mathcal{D}_{[0,\infty)}(\mathcal{M}_F(\mathbb{R}^d))$ of càdlàg paths on $\mathcal{M}_F(\mathbb{R}^d)$. We endow $\mathcal{M}_F(\mathbb{R}^d)$ with the topology of weak convergence and $\mathcal{D}_{[0,\infty)}(\mathcal{M}_F(\mathbb{R}^d))$ with the Skorohod topology. A sequence of processes taking values in $\mathcal{D}_{[0,\infty)}(\mathcal{M}_F(\mathbb{R}^d))$ is said to be tight if the corresponding sequence of distributions is tight, i.e., that any infinite subsequence has a weakly convergent subsubsequence. Our first main result establishes tightness of our rescaled population processes in the case in which interactions remain nonlocal under the scaling, and characterises limit points as solutions to a martingale problem.

Theorem 2.10 *Let $(\eta_t^N)_{t\geq 0}$ be as defined in Definition 2.4 and assume that as $N \to \infty$, $\theta(N) \to \infty$ in such a way that $\theta(N)/N \to \alpha$. (However, the kernels ρ_r, ρ_γ, and ρ_F remain fixed.) Suppose that Assumptions 2.8 hold and, further, that $\{\eta_0^N\}_{N\geq 1}$ is a sequence of purely atomic measures, with η_0^N comprised of atoms of mass $1/N$, which is tight in $\mathcal{M}_F(\mathbb{R}^d)$. Also assume there exists a nonnegative $f_0 \in C(\mathbb{R}^d)$ with uniformly bounded first and second derivatives (i.e., with $\sup_x \sup_{\|z\|=1} \sum_i \partial_{x_i} f_0(x) z_i$ and $\sup_x \sup_{\|z\|=1} \sum_{ij} \partial_{x_i x_j} f_0(x) z_i z_j$ both finite) and $f_0(x) \to \infty$ as $|x| \to \infty$ for which $\langle f_0(x), \eta_0^N(dx)\rangle < C < \infty$ for some C independent of N. Then the sequence of processes $(\eta_t^N)_{t\geq 0}$ is tight, and for any limit point $(\eta_t)_{t\geq 0}$, for every $f \in C_b^\infty(\mathbb{R}^d)$,*

$$\begin{aligned} M_t(f) := \langle f(x), \eta_t(dx)\rangle - \langle f(x), \eta_0(dx)\rangle \\ - \int_0^t \Big\langle \gamma(x,\eta_s)\mathcal{B}\left(f(\cdot) r(\cdot,\eta_s)\right)(x) + f(x)F(x,\eta_s), \eta_s(dx)\Big\rangle ds \end{aligned} \tag{2.11}$$

is a martingale (with respect to the natural filtration), with angle bracket process

$$\langle M(f)\rangle_t = \alpha \int_0^t \Big\langle 2\gamma\left(x,\eta_s\right) r\left(x,\eta_s\right) f^2(x), \eta_s(dx)\Big\rangle ds. \tag{2.12}$$

If $\alpha = 0$ the limit is deterministic.

Recall when interpreting (2.11) that, for instance, $r(x,\eta_s) = r(x, \rho_r * \eta_s(x))$, and so $\mathcal{B}(fr)(x) = \mathcal{B}(f(\cdot) r(\cdot, \rho_r * \eta_s(\cdot)))(x)$. The proof of this theorem appears in Section 3.2.

Theorem 2.10 provides tightness of the rescaled processes. If the limit points are unique, then this is enough to guarantee convergence.

Corollary 2.11 *Under the assumptions of Theorem 2.10, if the martingale problem defined by equations* (2.11) *and* (2.12) *has a unique solution, then* $(\eta_t^N)_{t\geq 0}$ *converges weakly to that solution as* $N \to \infty$.

When $\alpha > 0$, the limit points can be thought of as interacting superprocesses. For example, when r and γ are constant, and $F(x, \eta_s) = 1 - \rho_F * \eta_s(x)$, we recover a superprocess with nonlinear death rates corresponding to logistic growth [Eth04] that is a continuous limit of the Bolker-Pacala model [BP97; BP99]. We are not aware of a general result to determine when we will have uniqueness of solutions to the martingale problem of Theorem 2.10 when $\alpha > 0$. However, the Dawson–Girsanov transform tells us that we have uniqueness in this special case of the superprocess with nonlinear death rates, and Perkins stochastic calculus (and its adaptation to a lookdown setting) provides uniqueness for cases with interactions in the dispersal mechanism of the superprocess. We refer to [Daw93; Per92], and [DK99] for approaches to showing that these sorts of martingale problems are well-posed.

For the deterministic case of $\alpha = 0$, the limiting process is a weak solution to a nonlocal PDE. We next describe some situations in which more is known about uniqueness and whether the solution is close to the corresponding local PDE. First, recall the following notion of solution to a PDE.

Definition 2.12 (Weak solutions) *We say that* $(\eta_t)_{t\geq 0}$, *with* $\eta_t \in \mathcal{M}_F(\mathbb{R}^d)$, *is a* weak solution *to the PDE*

$$\partial_t \varphi = r\mathcal{B}^*(\gamma\varphi) + \varphi F \tag{2.13}$$

(where r, γ *and* F *can all be functions of* φ*) if, for all* $f \in C_b^\infty(\mathbb{R}^d)$,

$$\frac{d}{dt}\langle f, \eta_t\rangle = \langle \gamma\mathcal{B}(rf) + fF, \eta_t\rangle.$$

The notation φ is meant to be suggestive of a density, and recall that equation (2.13) has made dependencies on x and φ implicit; written out more explicitly, (2.13) is

$$\partial_t \varphi_t(x) = r\left(x, \rho_r * \varphi_t(x)\right) \mathcal{B}^*\left[\varphi_t(\cdot)\gamma\Big(x, \rho_\gamma * \varphi_t(\cdot)\Big)\right](x) + \varphi_t(x) F\left(x, \rho_F * \varphi_t(x)\right).$$

Because Theorem 2.10 only tells us about weak convergence, in the case $\alpha = 0$ we can only deduce that any limit point η_t is a weak solution to this nonlocal PDE.

Specialising the results of [KX99] to the deterministic setting provides general conditions under which we have existence and uniqueness of solutions to (2.13) which have an L^2-density with respect to Lebesgue measure. Recall that the Wasserstein metric, defined by

$$\rho(\nu_1, \nu_2) = \sup\left\{\left|\int f d\nu_1 - \int f d\nu_2\right| : \sup_x |f(x)| \leq 1, |f(x) - f(y)| \leq \|x - y\|\right\},$$

determines the topology of weak convergence on $\mathcal{M}_F(\mathbb{R}^d)$. We write $r(x,\eta)\gamma(x,\eta)\mathbf{C}(x) = J(x,\eta)J(x,\eta)^T$, and $\beta(x,\eta) = r(x,\eta)\gamma(x,\eta)\Big(\vec{b}(x) + 2\mathbf{C}(x)\nabla \log r(x,\eta)\Big)$ (quantities that will appear in Proposition 5.6). If J, β, and F are bounded and Lipschitz in the sense that

$$|J(x_1,\nu_1)-J(x_2,\nu_2)|, |\beta(x_1,\nu_1)-\beta(x_2,\nu_2)|, |F(x_1,\nu_1)-F(x_2,\nu_2)| \leq C(\|x_1-x_2\|+\rho(\nu_1,\nu_2)) \tag{2.14}$$

for some $C > 0$, the methods of [KX99] show that if the initial condition η_0 for our population process has an L^2 density, then so does η_t for $t > 0$. Although the necessary estimates (for which we refer to the original paper) are highly nontrivial, the idea of the proof is simple. Take a solution to the equation and use it to calculate the coefficients r, γ and F that depend on local population density. Then η solves the *linear* equation obtained by regarding those values of r, γ and F as given. It remains to prove that the solution to the linear equation has a density which is achieved by obtaining L^2 bounds on its convolution with the heat semigroup at time δ and letting $\delta \to 0$. We also have the following uniqueness result.

Theorem 2.13 (Special case of [KX99], Theorem 3.5) *Suppose J, β, and F are bounded and Lipschitz in the sense of (2.14). If η_0 has an $L^2(\mathbb{R}^d)$-density, then there exists a unique $L^2(\mathbb{R}^d)$-valued solution of (2.13) in the sense of Definition 2.12.*

Remark 2.14 *[KX99] considers an infinite system of stochastic differential equations for the locations and weights of a collection of particles that interact through their weighted empirical measure, which is shown to be the unique solution to a stochastic PDE. As we shall see through our lookdown representation in Section 5.1, the solution to our deterministic equation can be seen as the empirical measure of a countable number of particles (all with the same weight) which, in the notation above, evolve according to*

$$X(t) = X(0) + \int_0^t \beta\Big(X(s), \eta_s\Big)ds + \int_0^t J\Big(X(s), \eta_s\Big)dW(s)$$

(with an independent Brownian motion W for each particle).

Two-step convergence to PDE: Although the coefficients at x in (2.13) are nonlocal, we can choose our kernels ρ_γ, ρ_r, and ρ_F in such a way that they depend only on the population in a region close to x, and so we expect that under rather general conditions solutions of the nonlocal PDE will be close to the corresponding classical PDE. The following propositions provide two concrete situations in which this is true. In the first, the PDE is a reaction-diffusion equation, and in the proof in Section 3.3.1 we borrow an idea from [Pen17] to express the solutions to both the nonlocal equation and the classical PDE through a Feynman-Kac formula.

Proposition 2.15 *Let $\rho_F^\epsilon(x) = \rho_F\Big(x/\epsilon)/\epsilon^d$. Assume $\varphi_0 \in L^2(\mathbb{R}^d)$ is a positive, uniformly Lipschitz, and uniformly bounded function. Suppose that $\varphi^\epsilon \in L^2(\mathbb{R}^d)$ is a weak solution to the equation*

$$\partial_t \varphi^\epsilon = \mathcal{B}^* \varphi^\epsilon + \varphi^\epsilon F(\rho_F^\epsilon * \varphi^\epsilon), \qquad x \in \mathbb{R}^d,\, t > 0, \tag{2.15}$$

with initial condition $\varphi_0(\cdot)$, and that φ is a weak solution to the equation

$$\partial_t \varphi = \mathcal{B}^* \varphi + \varphi F(\varphi), \qquad x \in \mathbb{R}^d,\, t > 0, \tag{2.16}$$

also with initial condition $\varphi_0(\cdot)$. Suppose further that F is a Lipschitz function which is bounded above, and that $\vec{b}(x)$ and $\mathbf{C}(x)$, the drift and covariance matrix of $\mathcal{B}$, satisfy the conditions of Assumptions 2.8. Finally, assume that $\mathcal{B}^*$ is a generator of a Markov process. Then, for all $T > 0$ there exists a constant $K = K(T, \|\varphi_0\|_\infty) < \infty$ and a function $\delta(\epsilon)$ (dependent on ρ_F) with $\delta(\epsilon) \to 0$ as $\epsilon \to 0$, such that, for all $0 \le t \le T$, and ϵ small enough,

$$\|\varphi_t(\cdot) - \varphi_t^\epsilon(\cdot)\|_\infty \le K\delta(\epsilon).$$

In particular, as $\epsilon \to 0$, we have that φ^ϵ converges uniformly in compact intervals of time to φ.

Remark 2.16 *When we assume that $\mathcal{B}^*$ is a generator of a Markov process, we are imposing that*

$$\frac{1}{2}\sum_{ij} \partial_{x_i}\partial_{x_j} C_{ij}(x) - \sum_i \partial_{x_i} \vec{b}_i(x) = 0. \tag{2.17}$$

In words, at each point x in space, the total change on coefficients of the mean drift $\vec{b}$ due to perturbation in spatial location must equal to that of the covariance matrices.

Remark 2.17 *Note that Theorem 2.13 guarantees uniqueness of solutions to equation (2.15).*

Our second example in which we know solutions to the nonlocal PDE converge to solutions of the local PDE as interaction distances go to zero is a nonlocal version of a porous medium equation with logistic growth. That is, we consider non-negative solutions to the equation

$$\partial_t \psi^\epsilon = \Delta\left(\psi^\epsilon \, \rho_\gamma^\epsilon * \psi^\epsilon\right) + \psi^\epsilon\left(1 - \rho_\gamma^\epsilon * \psi^\epsilon\right). \tag{2.18}$$

The case without the reaction term (and with $\mathbb{R}^d$ replaced by a torus) is considered by [LM01] who use it as a basis for a particle method for numerical solution of the porous medium equation. Of course this does not quite fit into our framework, since in the notation of our population models this would necessitate $\gamma(x, m) = \rho_\epsilon * m$ which is not bounded. However, this can be overcome by an additional layer of approximation (c.f. our numerical experiments of Section 2.3.1) and we do not allow this to detain us here. Existence and uniqueness of solutions to (2.18) can be obtained using the approach of [LM01], so we should like to prove that as $\epsilon \to 0$ we have convergence to the solution to the porous medium equation with logistic growth:

$$\partial_t \psi = \Delta\left(\psi^2\right) + \psi\left(1 - \psi\right). \tag{2.19}$$

Notation 2.18 *We use $\rightharpoonup$ to denote weak convergence in the sense of analysts; that is, $\psi^\epsilon \rightharpoonup \psi$ in L^1 means $\int \psi^\epsilon v dx \to \int \psi v dx$ for all $v \in L^\infty$.*

We write $L^2_t(H^1)$ for functions for which the H^1 norm in space is in L^2 with respect to time, i.e.

$$\int_0^T \int \left\{\psi_t(x)^2 + \|\nabla \psi_t(x)\|^2\right\} dx dt < \infty,$$

and $C_t(L^1)$ will denote functions for which the L^1 norm in space is continuous in time.

Proposition 2.19 *Suppose that we can write $\rho_\gamma = \zeta * \check{\zeta}$, where $\check{\zeta}(x) = \zeta(-x)$ and $\zeta \in \mathcal{S}(\mathbb{R}^d)$ (the Schwartz space of rapidly decreasing functions). Furthermore, suppose that $\psi_0^\epsilon \geq 0$ is such that there exist $\lambda \in (0, 1)$ and $C \in (0, \infty)$ (independent of ϵ) such that*

$$\int \exp(\lambda \|x\|) \psi_0^\epsilon(x) dx < C, \quad \text{and } \sup_\epsilon \int \psi_0^\epsilon |\log \psi_0^\epsilon| dx < \infty,$$

with $\psi_0^\epsilon \rightharpoonup \psi_0$ *as* $\epsilon \to 0$. *Then writing* ψ^ϵ *for the solution to (2.18) on* $[0,T] \times \mathbb{R}^d$ *with initial condition* ψ_0^ϵ, $\psi^\epsilon \rightharpoonup \psi$ *as* $\epsilon \to 0$ *where* $\psi \in L_t^2(H^1) \cap C_t(L^1)$, $\int \psi |\log \psi| dx < \infty$, *and* ψ *solves (2.19) on* $[0,T] \times \mathbb{R}^d$.

The example that we have in mind for the kernel ρ_γ is a Gaussian kernel. For the proof, see Section 3.3.2.

Remark 2.20 *Although it seems hard to formulate an all-encompassing result, Propositions 2.15 and 2.19 are by no means exhaustive. When the scaling limit is deterministic, one can expect analogous results under rather general conditions. However, when the limit points are stochastic, they resemble "nonlinear superprocesses" and so one cannot expect a density with respect to Lebesgue measure in* $d \geq 2$. *It is then not reasonable to expect to be able to make sense of the limit if we scale the kernels in this way. Moreover, in one dimension, where the classical superprocess does have a density with respect to Lebesgue measure, the form of (2.11) suggests that even if one can remove the local averaging from* γ, *it will be necessary to retain averaging of* r *in order to obtain a well-defined limit.*

One-step convergence to PDE: Theorem 2.10, combined with Proposition 2.15 or 2.19 implies that we can take the limit $N \to \infty$ followed by the limit $\epsilon \to 0$ to obtain solutions to the PDE (2.16). However, it is of substantial interest to know whether we can take those two limits simultaneously. The general case seems difficult, but we prove such "diagonal" convergence in the following situation. The proof is provided in Section 4.

Theorem 2.21 (Convergence to a PDE) *Let* $(\eta_t^N)_{t \geq 0}$ *be as defined in Definition 2.4 with* $r(x,m) \equiv 1 \equiv \gamma(r,m)$, $F(x,m) \equiv F(m)$, ρ_F^ϵ *a symmetric Gaussian density with variance parameter* ϵ^2, *and* $\mathcal{B} = \Delta/2$. *Further suppose that* $F(m)$ *is a polynomial with* $F(m)\mathbf{1}_{m \geq 0}$ *bounded above. Assume that* $\langle 1, \eta_0^N \rangle$ *is uniformly bounded, and that for all* $x \in \mathbb{R}^d$ *and* $k \in \mathbb{N}$,

$$\limsup_{\epsilon \to 0} \mathbb{E}\Big[\rho_F^\epsilon * \eta_0(x)^k\Big] < \infty,$$

and

$$\limsup_{\epsilon \to 0} \int \mathbb{E}\Big[\rho_F^\epsilon * \eta_0(x)^k\Big] dx < \infty.$$

Finally assume that $N \to \infty$, $\theta \to \infty$ and $\epsilon \to 0$ in such a way that

$$\frac{1}{\theta\epsilon^2} + \frac{\theta}{N\epsilon^d} \to 0. \tag{2.20}$$

Then the sequence of $\mathcal{D}_{[0,\infty)}(\mathcal{M}_F(\mathbb{R}^d))$-valued stochastic processes $\left(\rho_F^\epsilon * \eta_t^N(x)dx\right)_{t\geq 0}$ converges weakly to a measure-valued process with a density $\varphi(t,x)$ that solves

$$\partial_t \varphi(t,x) = \frac{1}{2}\Delta\varphi(t,x) + \varphi(t,x)F(\varphi(t,x)). \tag{2.21}$$

Remark 2.22 *In fact, our proof goes through without significant change under the conditions that $F(m)\mathbf{1}_{m\geq 0}$ is bounded above (but not necessarily below), and that for all $m, n \in [0,\infty)$*

$$|F(m)| \leq \sum_{j=1}^{k} a_j m^j, \qquad \text{and } |F(n) - F(m)| \leq |n-m| \sum_{j=1}^{k'} b_j \Big(n^j + m^j\Big),$$

for some non-negative constants $\{a_j\}_{j=0}^k$, $\{b_j\}_{j=0}^{k'}$. We take F to be polynomial to somewhat simplify notation in the proof.

2.2.2 Ancestral lineages in the scaling limit

Now that we have established what we can say about how population density changes with time, we turn to results on ancestral lineages, i.e., how genealogical ancestry can be traced back across the landscape. Informally, a *lineage* $(L_t^N)_{t\geq 0}$, begun at a spatial location $L_0^N = x$ where there is a focal individual in the present day can be obtained by, for each time t, setting L_t^N to be the spatial location of the individual alive at time t before the present from whom the focal individual is descended. Since in our model individuals have only one parent, this is unambiguous. Although we did not explicitly retain such information, it is clear that for finite N, since individuals are born one at a time, one could construct the lineage $(L_t^N)_{t=0}^T$ given the history of the population $(\eta_t^N)_{t=0}^T$, for each starting location to which η_T^N assigns positive mass. It is less clear, however, how to formally retain such information when we pass to the scaling limit.

The *lookdown construction* in Section 5.1 will enable us to recover information about ancestry in the infinite population limit. Roughly speaking, each particle is assigned a unique "level" from $[0,\infty)$ that functions as a label and thus allows reconstruction

of lineages. The key to the approach is that levels are assigned in such a way as to be exchangeable, so that sampling a finite number, k say, of individuals from a given region is equivalent to looking at the individuals in that region with the k lowest levels. Moreover, as we pass to the infinite population limit, the collection of (individual, level) pairs converges, as we show in Theorem 5.4. See [EK19] for an introduction to these ideas. In particular, even in the infinite population limit, we can sample an individual from a region (it will be the individual in that region with the lowest level) and trace its line of descent. This will allow us to calculate, for each x and $y \in \mathbb{R}^d$, the proportion of the population at location x in the present day population that are descended from a parent who was at location y at time t in the past. To make sense of this in our framework, in Section 5.2.2, we justify a weak reformulation of this idea.

We are interested in two questions. First, when is the motion of an ancestral lineage, given complete knowledge of the population process, a well-defined process? In other words, is knowledge of the process $(\eta_t)_{t=0}^T$ that records numbers of individuals but not their ancestry sufficient to define the distribution of $(L_t)_{t=0}^T$? Second, does the process have a tractable description?

We focus on the simplest situation, that in which the population process is deterministic. However, the results here apply when the population process solves either a nonlocal or a classical PDE. There will be no coalescence of ancestral lineages in the deterministic limit, but understanding motion of single lineages is useful in practice, and our results can be seen as a first step towards understanding genealogies for high population densities.

Proofs of results in this section are found in Section 5.2.

Definition 2.23 (Ancestral lineage) *Let $(\varphi_t(x))_{0\le t\le T}$ denote the density of the scaling limit of our population model, solving* (2.13)*, and let y be a point with $\varphi_T(y) > 0$. We define $(L_s)_{s=0}^T$, the ancestral lineage of an individual sampled from the population at y at time T, by setting $L_0 = y$ and L_s to be the position of the unique ancestor of that individual at time $T - s$. We define $(Q_s)_{s\ge 0}$ to be the time inhomogeneous semigroup satisfying*

$$Q_s f(y) := \mathbb{E}_y[f(L_s)].$$

Our next result identifies the ancestral lineage as a diffusion by characterizing its generator.

Theorem 2.24 *For $\varphi : \mathbb{R}^d \to \mathbb{R}$, define*

$$\mathcal{L}_\varphi f = \frac{r}{\varphi}\left[\mathcal{B}^*(\gamma\varphi f) - f\mathcal{B}^*(\gamma\varphi)\right] \tag{2.22}$$

$$= r\gamma\left[\frac{1}{2}\sum_{ij}\mathbf{C}_{ij}\partial_{x_i x_j} f + \sum_j \vec{m}_j \partial_{x_j} f\right], \tag{2.23}$$

where $\vec{m}$ is the vector

$$\vec{m}_j = \sum_i C_{ij}\partial_{x_i}\log(\gamma\varphi) + \sum_i \partial_{x_i} C_{ij} - \vec{b}_j.$$

Then the generator of the semigroup Q_s of Definition 2.23 is given by $\partial_s Q_s f(y) = \mathcal{L}_{\varphi_{T-s}} Q_s f(y)$.

Remark 2.25 *As usual, to make the generator readable, we've written it in concise notation, omitting the dependencies on location and population density, which itself changes with time. When interpreting this, remember that everything depends on location and density at that location and time – for instance, "r" is actually $r(x, \varphi(x))$ (in the classical case), or $r(x, \rho_r * \eta(x))$ (in the nonlocal case).*

Moreover, we haven't proved any regularity of the population density process φ, so, as written, the generator (2.22) may not make sense. Instead, it should be interpreted in a weak sense which is made precise in Section 5.2.2.

Corollary 2.26 *In addition to the assumptions of Theorem 2.24, if the covariance of the dispersal process is isotropic (i.e., $\mathbf{C} = 2\sigma^2 I$), then*

$$\mathcal{L}_\varphi f = r\gamma\left(\sigma^2\Delta f + \left(2\sigma^2\nabla\log(\gamma\varphi) - \vec{b}\right)\cdot\nabla f\right). \tag{2.24}$$

(However, $\vec{b}$ can still depend on location.)

In other words, the lineage behaves as a diffusion driven by Brownian motion run at speed σ^2 multiplied by the local per-capita production of mature offspring ($r\gamma$) in a potential tilted by migration bias ($(\varphi_s\gamma)^2\exp(-\vec{b}\cdot x/\sigma^2)$, whose gradient appears in the drift term of the generator). In particular, lineages are drawn to regions of high fecundity (production of juveniles), but their speed is determined by the rate of production of mature offspring. This can be compared to Remark 2.7.

Corollary 2.27 *In addition to the assumptions of Corollary 2.26, if the population process is stationary (so $\varphi_t \equiv \varphi$), and $\vec{b}(x) = \nabla h(x)$ for some function h, then Y is reversible with respect to*

$$\pi(x) = \frac{\gamma}{r}\varphi(x)^2 e^{-h(x)/\sigma^2}. \tag{2.25}$$

Long-term fitness of an individual is proportional to the fraction of lineages from the distant future that pass through the individual, and hence the total long-term fitness at a location is proportional to the stationary distribution of Y there, if it exists. Therefore, if π is integrable then the per-capita long-term fitness of an individual at x is proportional to $\pi(x)/\varphi(x)$.

Corollary 2.28 *In addition to the assumptions of Corollary 2.26, suppose that the population process is described by a travelling wave with velocity $\mathfrak{c}$, i.e., the population has density $\varphi(t,x) = w(x - t\mathfrak{c})$ where w solves*

$$r\mathcal{B}^*(\gamma w) + wF + \mathfrak{c}\cdot\nabla w = r\sigma^2\Delta(\gamma w) - \vec{b}\cdot\nabla(\gamma w) + \mathfrak{c}\cdot\nabla w = 0.$$

Then the semigroup Q_s of the motion of a lineage in the frame that is moving at speed $\mathfrak{c}$ is time-homogeneous with generator

$$\mathcal{L}f = \sigma^2 r\gamma\left(\Delta f + 2\nabla\log(\gamma w)\cdot\nabla f\right) + (\mathfrak{c} - r\gamma\vec{b})\cdot\nabla f. \tag{2.26}$$

2.3 Examples and applications

We now discuss some consequences of these results.

2.3.1 Beyond linear diffusion

Equation (2.13) is a nonlocal version of a reaction-diffusion equation; the diffusion is nonlinear if γ depends on population density: in other words, if the diffusivity of the population depends on the population density. Passing to the classical limit, we recover equations like (2.19). Such equations are widely used in a number of contexts in biology in which motility within a population varies with population density. For example, density dependent dispersal is a common feature in spatial models in ecology, eukaryotic cell

biology, and avascular tumour growth; see [She10] and references therein for further discussion. In particular, it has been suggested as a model for the expansion of a certain type of bacteria on a thin layer of agar in a Petri dish [Coh+99]. We shall pay particular attention to the case in which the equation can be thought of as modelling the density of an expanding population. We focus on the monostable reaction of (2.19).

Comparing with (2.13), we see that to set up a limit in which the population density φ follows the porous medium equation with logistic growth of (2.19), we need $r = 1$, $\gamma = \varphi$, and $F = 1-\varphi$. Consulting equation (2.6), this implies that $\mu_\theta = \max\left(0, (1+1/\theta)\varphi - 1/\theta\right)$. In other words, establishment is certain and birth rates increase linearly with population density, but to compensate, death rates increase slightly faster (also linearly). Alert readers will notice that the condition from Assumptions 2.8 that $\gamma(x, m)$ be uniformly bounded is violated. This can be corrected by use of a cut-off, and in fact the downwards drift provided by the logistic control of the population size prevents m from getting too big. In practice the simulations shown in Figure 2.2 take discrete time steps of length dt (with dt suitably small), and have each individual reproduce and die with probabilities, respectively,

$$p_{\text{birth}}(m) = \left(1 - e^{-m dt}\right) \qquad p_{\text{death}}(m) = \left(1 - e^{-(m(1+1/\theta)-1/\theta)dt}\right),$$

where m is the local density at their location. This makes $\gamma(x, m) = p_{\text{birth}}(m)/dt \approx m$ and

$$F(x, m) = \theta(\gamma(x, m) - \mu(x, m))/dt \approx 1 - m.$$

Birth and death rates are equal at density $m = 1$, corresponding to an unscaled density of N individuals per unit area.

In one dimension, equation (2.19) has an explicit travelling wave solution

$$w^P(t, x) := \left(1 - e^{\frac{1}{2}(x - x_0 - t)}\right)_+. \tag{2.27}$$

Notice that the wave profile has a sharp boundary at $x = x_0 + t$. There are also travelling wave solutions with $c > 1$ [GK05], which lack this property. However, for initial conditions that decay sufficiently rapidly at infinity, such as one might use in modelling a population invading new territory, the solution converges to (2.27) [KR04]. In Figure 2.2 we show simulations of the individual based model described above, which display travelling wave solutions qualitatively similar to solutions of (2.19), with better agreement for smaller θ/N (but in both cases, N is reasonably large).

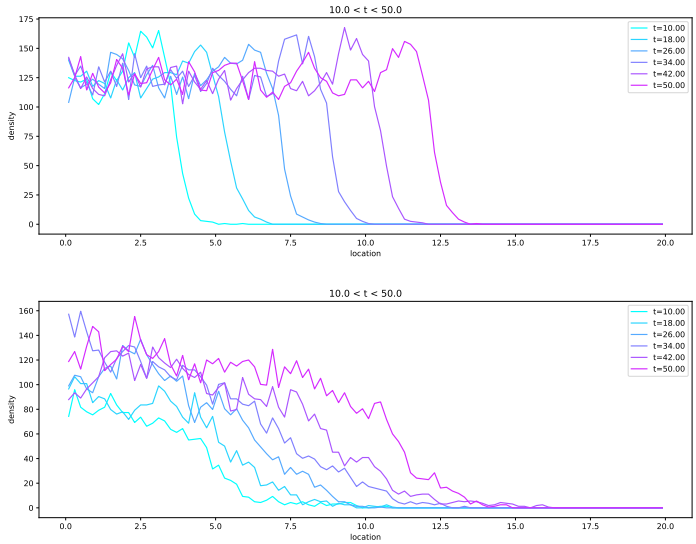


Figure 2.2: Simulated populations under a porous medium equation with logistic growth (2.19) in $d = 1$, θ/N small in the top, large in the bottom figure. Values of θ in top and bottom figures are 1 and 100, respectively, and both have N set so that the density is roughly 100 individuals per unit of habitat (as displayed on the vertical axis). See text for details of the simulations.

2.3.2 Ancestry in different types of travelling waves

Although it remains challenging to establish the distribution of genealogical trees relating individuals sampled from our population model, as described in Section 2.1, we can gain some insight by investigating the motion of a single ancestral lineage. Here we do that in the context of a one-dimensional population expanding into new territory as a travelling wave. We focus on three cases in which we have explicit information about the shape of the travelling wave profile: the Fisher-KPP equation, a special case of the Allen-Cahn equation with a bistable nonlinearity, and the porous media equation with logistic growth, equation (2.19). We work here in one dimension, and take $\sigma^2 = 1$ and $\vec{b} = 0$.

Fisher-KPP equation: Consider the classical Fisher-KPP equation,

$$\partial_t \varphi = \partial_{xx} \varphi + \varphi(1 - \varphi). \tag{2.28}$$

Even though we do not have an explicit formula for the wave shape in this case, our methods provide information about ancestral lineages. The equation has non-negative travelling wave solutions of speed c for all $c \geq 2$, but, started from any compact perturbation

of a Heaviside function, the solution will converge to the profile w^F with the minimal wavespeed, $c = 2$ [KPP37; FM77; Bra83]. No matter what initial condition, for any $t > 0$ the support of the solution will be the whole real line. In this case, we must have $r = \gamma = 1$, and $F(x, m) = 1 - m$ so $\mu_\theta(x, m) = 1 + (m - 1)/\theta$. By Corollary 2.28, the generator of the motion of an ancestral lineage is

$$\mathcal{L}_F f = \partial_{xx} f + 2\frac{\partial_x w^F}{w^F}\partial_x f + 2\partial_x f. \tag{2.29}$$

Near the tip of the wave (for x large), $w^F(x) \sim e^{-x}$, so (2.29) implies that the motion of a lineage is close to unbiased Brownian motion. On the other hand, in the "bulk", a lineage behaves approximately as Brownian motion with drift at rate two to the right. This implies that ancestral lineages are pushed into the tip of the wave, and there is no stationary distribution, so that long-term dynamics of genetic inheritance depend on the part of the wave not well-approximated by a smooth profile, in agreement with the previous results referred to in Section 2.1.

Allen-Cahn equation: Now take the Allen-Cahn equation:

$$\partial_t \varphi = \partial_{xx} \varphi + \varphi(1 - \varphi)(2\varphi - 1 + s), \tag{2.30}$$

for a given $s \in (0, 2)$. Once again we have taken $r = \gamma = 1$, but now the reaction term $F(x, m) = (1-m)(2m-1+s)$ is bistable. This equation can be used to model the motion of so-called hybrid zones in population genetics; see, for example, [Bar79], [Goo18], and [EGL22]. This equation has an explicit travelling wave solution with speed s and shape

$$w^A(x) = (1 + e^x)^{-1},$$

i.e., $\phi_t(x) = w^A(x - st)$ solves (2.30). Substituting w^A in place of w^F in (2.29), we find that the generator of an ancestral lineage relative to the wavefront is now,

$$\begin{aligned}\mathcal{L}_A f &= \partial_{xx} f + 2\frac{\partial_x w^A}{w^A}\partial_x f + s\partial_x f \\ &= \partial_{xx} f - 2\frac{e^x}{1 + e^x}\partial_x f + s\partial_x f,\end{aligned}$$

so lineages in the tip are pushed leftwards into the bulk of the wave at a rate $s-2e^x/(1+e^x)$. The density of the speed measure for this diffusion is

$$m_A(x) \propto e^{sx}(1+e^x)^{-2},$$

which is integrable, and so determines the unique stationary distribution. Thus the position of the ancestral lineage relative to the wavefront will converge to a stationary distribution which is maximised away from the extreme tip of the wave. This is consistent with [EP22], who consider an analogous stochastic population model, although the stronger result there (that the genealogy of a sample from behind the wavefront is approximately a Kingman coalecsent) requires the stronger condition $s < 1$.

Porous Medium equation with logistic growth: Finally, consider equation (2.19). Setting $x_0 = 0$ (for definiteness) and substituting the form of w^P from equation (2.27) into Corollary 2.28, with $c = 1$, $\gamma(x,m) = m$, $r(x,w) = 1$, and $F(x,m) = (1-m)$, the generator of the diffusion governing the position of the ancestral lineage relative to the wavefront is, for $x < 0$,

$$\begin{aligned}\mathcal{L}_P f &= w^P \left(\partial_{xx} f + 2\frac{\partial_x((w^P)^2)}{(w^P)^2}\partial_x f \right) + \partial_x f \\ &= \left(1 - e^{\frac{1}{2}x}\right)\partial_{xx} f - 2e^{\frac{1}{2}x}\partial_x f + \partial_x f.\end{aligned}$$

The speed measure corresponding to this diffusion has density

$$\begin{aligned}m_P(\xi) &\propto \frac{1}{2(1-e^{\xi/2})} \exp\left(\int_\eta^\xi \left\{ 1 - \frac{e^{x/2}}{1-e^{x/2}} \right\} dx \right) \\ &\propto e^\xi \left(1 - e^{\xi/2}\right), \quad \text{for } \xi < 0\end{aligned}$$

and $m_P(\xi) = 0$ for $\xi \geq 0$, which is integrable and so when suitably normalised gives the unique stationary distribution. Notice that even though we have the same reaction term as in the Fisher-KPP equation, with this form of nonlinear diffusion, at stationarity the lineage will typically be significantly behind the front, suggesting a different genealogy.

It is interesting to compare the stationary distribution we have obtained here to the expression that we'd get by setting $\vec{b} = -\mathfrak{c}$ and using Corollary 2.27, i.e., by giving each offspring a mean displacement that offsets the motion of the wave. In the Fisher-KPP and

Allen-Cahn cases above, we get the same expressions, but this is only the case because $r \equiv \gamma \equiv 1$ in both. For the porous medium equation with $\vec{b} = 1$ we have $\mathcal{B}^* = \Delta + \nabla$ and so the equation solved by the population density is

$$\partial_t \varphi = \Delta\varphi + 2\varphi\nabla\varphi + \varphi(1-\varphi),$$

which has a traveling wave solution of the same shape but moving at half the speed, $\varphi(x,t) = w^P(x - t/2)$, and a stationary distribution of the lineage relative to the wavefront of $\pi(\xi) \propto e^{3\xi/2}(1 - e^{\xi/2})^3$.

2.3.3 Clumping from nonlocal interactions

Simulating these processes and exploring parameter space, one sooner or later comes upon a strange observation: with certain parameter combinations, the population spontaneously forms a regular grid of stable, more or less discrete patches, separated by areas with nearly no individuals, as shown in Figure 2.3. The phenomenon is discussed in Section 16.10 of [HM22], and has been described in similar models, e.g., by [Bri90; Sas97; HL04; YRS01], and [Ber+09]. For example, if the density-dependent effects of individuals extend farther (but not too much farther) than the typical dispersal distance, then depending on the interaction kernel new offspring landing between two clumps can effectively find themselves in competition with *both* neighbouring clumps, while individuals within a clump compete with only one.

More mathematically, consider the case in which $\mathcal{B} = \sigma^2\Delta$ and all parameters are spatially homogeneous, so that $r(x,\eta) = r(\rho_r * \eta(x))$, and similarly for γ and F. If φ_0 is such that $F(\varphi_0) = 0$ and $F'(\varphi_0) < 0$, then the constant solution $\varphi \equiv \varphi_0$ is a nontrivial equilibrium of (2.1). However, this constant solution may not be unique, it may be unstable, and a stable solution may have oscillations on a scale determined by the interaction distance.

To understand the stability of the constant solution $\varphi \equiv \varphi_0$, we linearise (2.1) around φ_0: let $\varphi_t(x) = \varphi_0 + \psi_t(x)$, and (informally) $r(x) \approx r(\varphi_0) + r'(\varphi_0)\rho_r * \psi(x)$. Writing $r_0 = r(\varphi_0)$ and $r_0' = r'(\varphi_0)$, with analogous expressions for γ and F,

$$\partial_t \psi \approx \sigma^2\varphi_0 r_0 \gamma_0' \Delta\rho_\gamma * \psi + \sigma^2 r_0\gamma_0\Delta\psi + \varphi_0 F_0' \rho_F * \psi.$$

Letting $\widehat{f}(u) = \int e^{2\pi i u x} f(x) dx$ denote the Fourier transform,

$$\partial_t \widehat{\psi}(u) \approx \left\{ -4\pi^2 u^2 \sigma^2 \varphi_0 r_0 \gamma_0' \widehat{\rho}_\gamma(u) - 4\pi^2 u^2 \sigma^2 r_0 \gamma_0 + \varphi_0 F_0' \widehat{\rho}_F(u) \right\} \widehat{\psi}(u). \tag{2.31}$$

In the simplest case, in which γ is constant, so $\gamma_0' = 0$, this reduces to

$$\partial_t \widehat{\psi}(u) \approx \left(-4\pi^2 u^2 \sigma^2 r_0 \gamma_0 + \varphi_0 F_0' \widehat{\rho}_F(u) \right) \widehat{\psi}(u). \tag{2.32}$$

If we take $\rho_F = p_{\epsilon^2}$, then $\widehat{\rho}_F(u) = \exp(-2\pi\epsilon^2 u^2)$ and (recalling that $F_0' < 0$) the term in brackets is always negative, and we recover the well-known fact that in this case the constant solution is stable. If, on the other hand, $\widehat{\rho}_F$ changes sign, there may be values of u for which the corresponding quantity is positive. For example, if $d = 1$ and $\rho_F(x) = \mathbf{1}_{[-\epsilon,\epsilon]}(x)/2\epsilon$, then $\widehat{\rho}_F(u) = \sin(2\pi\epsilon u)/(2\pi\epsilon u)$, which is negative for $u \in (1/(2\epsilon), 1/\epsilon)$ (and periodically repeating intervals). Setting $v = \epsilon u$, the bracketed term on the right hand side of (2.32) becomes

$$\varphi_0 F_0' \frac{1}{2\pi v} \sin(2\pi v) - 4\pi^2 \frac{\sigma^2}{\epsilon^2} v^2 r_0 \gamma_0,$$

and we see that if σ^2/ϵ^2 is sufficiently small, there are values of v for which this is positive. In other words, in keeping with our heuristic above, if dispersal is sufficiently short range relative to the range over which individuals interact, there are unstable frequencies that scale with the interaction distance ϵ. In two dimensions, replacing the indicator of an interval by that of a ball of radius ϵ, a similar analysis applies, except that the sine function is replaced by a Bessel function.

Now suppose that γ is not constant. Then, from (2.31), if we take $\rho_\gamma = \rho_F = p_\epsilon^2$,

$$\partial_t \widehat{\psi}(u) = e^{-2\pi^2\epsilon^2 u^2} \left\{ -\sigma^2 \varphi_0 r_0 \gamma_0' 4\pi^2 u^2 - \sigma^2 r_0 \gamma_0 4\pi^2 u^2 e^{2\pi^2\epsilon^2 u^2} + \varphi_0 F_0' \right\} \widehat{\psi}(u).$$

If we make the (reasonable) assumption that $\gamma_0' < 0$, then we see that even when the Fourier transform of ρ does not change sign, there may be parameter values for which the constant solution is unstable. As before, we set $v = \epsilon u$. The term in brackets becomes

$$4\pi^2 \frac{\sigma^2}{\epsilon^2} v^2 r_0 \left(-\varphi_0 \gamma_0' - \gamma_0 e^{2\pi^2 v^2} \right) + \varphi_0 F_0',$$

and, provided $-\varphi_0 \gamma_0'/\gamma_0 > 1$, for sufficiently small v the term in round brackets is positive. We now see that if σ^2/ϵ^2 is sufficiently *large*, the equilibrium state $\varphi \equiv \varphi_0$ is

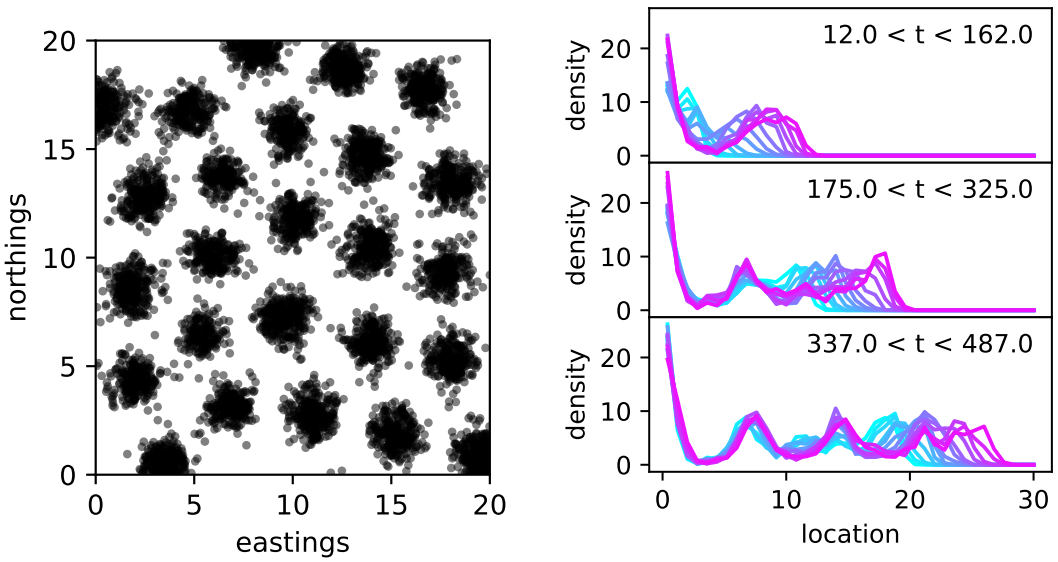


Figure 2.3: Left: A snapshot of individual locations in a two-dimensional simulation in which the constant density is unstable and a stable, periodic pattern forms. **Right:** Population density in an expanding wave in a one-dimensional simulation forming a periodic pattern; each panel shows the wavefront in three periods of time; within each period of time the wavefront at earlier times is shown in blue and later times in pink. In both cases, $\gamma(m) = 3/(1+m)$, $\mu \equiv 0.3$, and $r \equiv 1$; dispersal is Gaussian with $\sigma = 0.2$ and density is measured with $\rho_\gamma(x) = p_9(x)$, i.e., using a Gaussian kernel with standard deviation 3.

unstable. As before, the unstable frequencies will scale with ϵ and for given F, r and γ, whether or not such unstable frequencies exist will be determined by σ^2/ϵ^2, but in this case of Gaussian kernels, it is interaction distance being sufficiently small relative to dispersal that will lead to instability.

2.3.4 Lineage motion is not uniquely determined by population density

It is natural for applications to wonder about identifiability: when can the observed quantities like population density or certain summaries of lineage movement uniquely determine the underlying demographic parameters? Consider a deterministic, continuous population generated by parameters γ, r, and F, with $\vec{b} = 0$ and $\mathbf{C} = 2I$. Suppose it has a stationary profile $w(x)$, that must satisfy

$$r\Delta(\gamma w) + Fw = 0.$$

It is easy to see that w does not uniquely specify γ, F, and r: let $\lambda(x)$ be a smooth, nonnegative function on $\mathbb{R}^d$, and let $\widetilde{r}(x,m) = \lambda(x)r(x,m)$ and $\widetilde{F}(x,m) = \lambda(x)F(x,m)$ (and, let $\widetilde{\gamma} = \gamma$). Since $\mu = r\gamma - F/\theta$, this corresponds to multiplying both establishment probabilities and death rates by λ. Then the population with parameters $\widetilde{\gamma}$, $\widetilde{r}$, and $\widetilde{F}$ has the same stationary profile(s) as the original population.

Can these two situations be distinguished from summaries of lineage movement? The first has lineage generator

$$f \mapsto \mathcal{L}f = r\gamma\left(\Delta f + 2\nabla \log(\gamma w) \cdot \nabla f\right),$$

while the second has lineage generator $f \mapsto \lambda(x)\mathcal{L}f(x)$. In other words, although the stationary profile of the population is unchanged when we scale local establishment and death by λ, the motion of lineages is sped up locally by λ. This corresponds to making areas with $\lambda > 1$ more "sink-like" and $\lambda < 1$ more "source-like": if $\lambda(x) > 1$, then at x both the death rate and probability of establishment of new individuals are higher. As a result, lineages in the second model spend more time in areas with $\lambda < 1$, i.e., those areas have higher reproductive value, something that is, in principle, discernible from genetic data (because, for instance, making reproductive value less evenly distributed reduces long-term genetic diversity).

2.4 Heuristics

In this section we perform some preliminary calculations and use them to provide heuristic arguments for our main results, to build intuition before the formal proofs.

2.4.1 The population density

We reiterate that in our prelimiting model, the population is represented by a point measure η^N in which each individual is assigned a mass $1/N$. We use the term "population density" for this process, as it is supposed to measure population size relative to a nominal occupancy of N individuals per unit area, but it is not absolutely continuous with respect to Lebesgue measure.

We write $\mathcal{P}^N$ for the generator of the scaled population process η^N of Definition 2.4 acting on test functions of the form $G(\langle f, \eta\rangle)$, where $f \geq 0$ is smooth and bounded on $\mathbb{R}^d$ and $G \in C^\infty([0,\infty))$. Recall that $\theta = \theta(N) \to \infty$ as $N \to \infty$ in such a way that $\theta(N)/N \to \alpha$.

A Taylor expansion allows us to write

$$\begin{aligned}\mathcal{P}^N G(\langle f,\eta\rangle) = G'(\langle f,\eta\rangle) \lim_{\delta t \downarrow 0} \frac{1}{\delta t}\mathbb{E}\left[\langle f, \eta_{\delta t}\rangle - \langle f,\eta\rangle \,|\, \eta_0 = \eta\right] \\ + \frac{1}{2}G''(\langle f,\eta\rangle) \lim_{\delta t \downarrow 0} \frac{1}{\delta t}\mathbb{E}\left[\Big(\langle f, \eta_{\delta t}\rangle - \langle f,\eta\rangle\Big)^2 \Big| \, \eta_0 = \eta\right] + \epsilon_N(f,G,\eta), \quad (2.33)\end{aligned}$$

where the terms that make up $\epsilon_N(f,G,\eta)$ will be negligible in our scaling limit (at least if $G''' < \infty$).

Mean measure

Recall that in our parameterization only death rates μ_θ and the dispersal kernel q_θ depend on θ. For a suitable test function f, we find

$$\begin{aligned}\mathcal{P}^N\langle f,\eta\rangle &= \lim_{\delta t \downarrow 0} \frac{1}{\delta t}\mathbb{E}\left[\langle f, \eta_{\delta t}\rangle - \langle f,\eta\rangle \,|\, \eta_0 = \eta\right] \\ &= \theta \int\int f(z) r(z,\eta) q_\theta(x,dz)\gamma(x,\eta)\eta(dx) - \theta\int f(x)\mu_\theta(x,\eta)\eta(dx).\end{aligned} \quad (2.34)$$

The first term is the increment in $\langle f,\eta\rangle$ resulting from a birth event (recalling that we don't kill the parent) integrated against the rate of such events, and the second reflects death events. The factor of θ appears from the time rescaling. In both terms, the rate of events has a factor of N (because events happen at a rate proportional to the number of individuals, whereas η has mass $1/N$ for each individual) which is offset by the fact that the birth or loss of a single individual at the point y, say, changes $\langle f,\eta\rangle$ by $f(y)/N$.

We use the fact that $\int q_\theta(x,dz) = 1$ to rewrite (2.34) as

$$\begin{aligned}\int\left(\int \theta\left(f(z)r(z,\eta) - f(x)r(x,\eta)\right) q_\theta(x,dz)\right)\gamma(x,\eta)\eta(dx) \\ + \int f(x)\theta\Big(r(x,\eta)\gamma(x,\eta) - \mu_\theta(x,\eta)\Big)\eta(dx).\end{aligned} \quad (2.35)$$

We have defined μ_θ so that the second term is simple:

$$\theta\Big(r(x,\eta)\gamma(x,\eta) - \mu_\theta(x,\eta)\Big) = F(x,\eta).$$

Furthermore, recall from Remark 2.6 that

$$\int \theta\Big(r(z,\eta)f(z) - r(x,\eta)f(x)\Big)q_\theta(x,dz) \qquad \stackrel{\theta\to\infty}{\longrightarrow} \qquad \mathcal{B}\Big(r(\cdot,\eta)f(\cdot)\Big)(x). \tag{2.36}$$

In particular, if dispersal is determined by a standard multivariate Gaussian with mean zero and covariance $2\sigma^2 I/\theta$, then $\mathcal{B} = \sigma^2\Delta$, where Δ denotes the Laplacian.

In summary, equation (2.35) converges to

$$\int \gamma(x,\eta)\mathcal{B}\Big(f(\cdot)r(\cdot,\eta)\Big)(x)\eta(dx) + \int f(x)F(x,\eta)\eta(dx), \tag{2.37}$$

which explains the form of the martingale of Theorem 2.10.

Quadratic variation

We now look at the second order term in (2.33), which will converge to the quadratic variation of the limiting process. An individual at location x gives birth to a surviving offspring at y at rate

$$\gamma(x,\eta)r(y,\eta)q_\theta(x,dy),$$

and since this increments $\langle f,\eta\rangle$ by $f(y)/N$, the contribution to the quadratic variation from birth events, which occur at rate θ per individual (so, rate $N\theta|\eta|$ overall), is

$$\int N\theta\gamma(x,\eta)\int \frac{1}{N^2}f^2(y)r(y,\eta)q_\theta(x,dy)\eta(dx).$$

Similarly, the increment in $\langle f,\eta\rangle$ resulting from the death of an individual at x is $-f(x)/N$, and so combining with the above, the second order term in the generator takes the form

$$\begin{aligned} &G''(\langle f,\eta\rangle)\frac{1}{2}N\theta\left\{\int\gamma(x,\eta)\int\frac{1}{N^2}f^2(y)r(y,\eta)q_\theta(x,dy)\eta(dx) + \int\mu_\theta(x,\eta)\frac{1}{N^2}f^2(x)\eta(dx)\right\} \\ &\quad = \frac{1}{2}G''(\langle f,\eta\rangle)\frac{\theta}{N}\int\left\{\gamma(x,\eta)\int f^2(y)r(y,\eta)q_\theta(x,dy) + f^2(x)\mu_\theta(x,\eta)\right\}\eta(dx). \end{aligned}$$

Since $\int f^2(y)r(y,\eta)q_\theta(x,dy) \to f^2(x)r(x,\eta)$ and $r\gamma + \mu_\theta = 2r\gamma - F/\theta \to 2r\gamma$ as $\theta \to \infty$, this converges to

$$\frac{\alpha}{2}G''(\langle f,\eta\rangle)\Big\langle 2r(x,\eta)\gamma(x,\eta),\eta(dx)\Big\rangle.$$

An entirely analogous argument shows that if G''' is bounded, then the term $\epsilon_{\theta,N}(f,G,\eta)$ in (2.33) will be $\mathcal{O}(\theta/N^2)$.

If we hold ρ_γ, ρ_r, ρ_F fixed, then by taking $\theta/N \to 0$, the second order term in the generator will vanish and we expect a deterministic limit, for which $\partial_t\langle f, \eta_t\rangle$ is equal to (2.37). In other words, the limit is a weak solution to the deterministic equation

$$\partial_t\varphi_t(x) = r(x,\varphi_t)\mathcal{B}\Big(\gamma(\cdot,\varphi_t)\varphi_t(\cdot)\Big)(x) + F(x,\varphi_t)\varphi_t(x) \tag{2.38}$$

in the sense of Definition 2.12, where φ_t is the density of η_t, if it has a density. On the other hand, if $N = \alpha\theta$ for some $\alpha > 0$, the second order term remains, and we expect a "generalised superprocess" limit. The limiting quadratic variation is exactly as seen in Theorem 2.10.

One-step convergence: In order to pass directly to a classical PDE limit in Theorem 2.21 we impose the stronger condition that $\theta/N\epsilon^d \to 0$ and also require that $\theta\epsilon^2 \to \infty$. Recall that in this case, we take ρ_F^ϵ to be a symmetric Gaussian density with variance ϵ^2. The condition $\theta\epsilon^2 \to \infty$ ensures that ϵ^2 is large enough relative to $1/\theta$ that the regularity gained by smoothing our population density by convolution with ρ_ϵ is preserved under the dynamics dictated by q_θ. To understand the first condition, note that we are aiming to obtain a deterministic expression for the limiting population density. It is helpful to think about a classical Wright-Fisher model (with no spatial structure and just two types, say). We know then that if the timescale θ is on the same order as population size N, we see stochastic fluctuations in the frequencies of the two types in the limit as $N \to \infty$; to obtain a deterministic limit, we look over timescales that are short relative to population size. In our setting, the total population size is replaced by the local population size, as measured by convolution with ρ_ϵ, which we expect to be of order $N\epsilon^d$, and so in order to ensure a deterministic limit we take $\theta/(N\epsilon^d) \to 0$.

2.4.2 Motion of ancestral lineages

Although our proof of Theorem 2.24 uses an explicit representation in terms of the lookdown process, the result can be understood through informal calculations. Suppose that we have traced a lineage back to an individual at location y at time t. Looking further back through time, at the time of the birth of that individual, the lineage will jump

to the location of the parent of the individual. Now, the rate at which new individuals are born to parents at x and establish at y is

$$\theta N \eta_t^N(dx)\gamma(x,\eta_t^N)q_\theta(x,dy)r(y,\eta_t^N).$$

Suppose that η^N did have a density (in the prelimit it does not), say $\eta_t^N(dx) = \varphi_t^N(x)dx$. Informally, since the number of individuals near y is $N\varphi_t^N(y)dy$, the probability that a randomly chosen individual near y is a new offspring from a parent at x in $[t, t+dt)$ is

$$\frac{\theta\varphi_t^N(x)\gamma(x,\eta_t^N)r(y,\eta_t^N)}{\varphi_t^N(y)}\frac{q_\theta(x,dy)}{dy}dxdt. \tag{2.39}$$

Leaving aside questions of whether a lineage can be treated as a randomly chosen individual, we define a continuous-time jump process whose transition rates, conditional on $(\varphi_t^N)_{t=0}^T$, are given by (2.39). Because we are tracing the lineage backwards in time we make the substitution $s = T-t$ and write $(L_s^N)_{s=0}^T$ for the location of a lineage that moves according to these jump rates. Then, abusing notation to write $q_\theta(x,y)$ for the density of $q_\theta(x,dy)$,

$$\begin{aligned}&\mathbb{E}[f(L_{s+ds}^N) - f(y) \mid L_s^N = y]\\ &\quad = ds\,\theta\int (f(x)-f(y))\,\frac{\varphi_{T-s}^N(x)\gamma(x,\eta_{T-s}^N)r(y,\eta_{T-s}^N)}{\varphi_{T-s}^N(y)}q_\theta(x,y)dx.\end{aligned} \tag{2.40}$$

(Note that this integral is with respect to x.) Referring back to Remark 2.6, a quick calculation shows that as $N\to\infty$,

$$\begin{aligned}&\theta\int \Big(f(x)-f(y)\Big)g(x)q_\theta(x,y)dx\\ &\quad = \theta\int\Big\{(f(x)g(x)-f(y)g(y)) - f(y)(g(x)-g(y))\Big\}q_\theta(x,y)dx\\ &\quad \to \mathcal{B}^*(fg)(y) - f(y)\mathcal{B}^*g(y).\end{aligned}$$

Applying this to (2.40) with $g = \varphi_{T-s}\gamma$, this suggests that the generator of the limiting process is

$$\mathcal{L}_s f = \frac{r}{\varphi_{T-s}}\left\{\mathcal{B}^*(\gamma\varphi_{T-s}f) - f\mathcal{B}^*(\gamma\varphi_{T-s})\right\}. \tag{2.41}$$

This agrees with Theorem 2.24.

3

Branching Particle Systems with Local Regulation: 2-Step Convergence Proof

In this chapter we present formal proofs of the first two of our three scaling limits. In Subsection 3.2 we prove Theorem 2.10, to obtain (both stochastic and deterministic) limits in which interactions between individuals in the population are nonlocal. In Subsection 3.3 we show how, in two important examples in which the nonlocal limit is respectively a deterministic solution to a non-local equation of reaction-diffusion type and a deterministic solution to a nonlocal porous medium equation with an additional logistic growth term, one can pass to a further limit to obtain a classical PDE.

3.1 Preliminaries

Below we will have frequent use for the quantity

$$B_f^\theta(x,\eta) = \theta \int_{\mathbb{R}^d} (f(y)r(y,\eta) - f(x)r(x,\eta))q_\theta(x,dy). \tag{3.1}$$

First, we prove Lemma 2.9.

Proof [of Lemma 2.9] Here, we need to prove that $|\gamma(x,\eta)B_f^\theta(x,\eta)|$ is bounded, uniformly

over x and η. First suppose that Condition 1 of Lemma 2.9 is satisfied. We write

$$\begin{aligned}
r(y,\eta)f(y) - r(x,\eta)f(x) &= r(y,\eta)(f(y)-f(x)) + (r(y,\eta)-r(x,\eta))f(x) \\
&= r(y,\eta)\left(\sum_i (y-x)_i \partial_{x_i} f(x) + \sum_{ij}(y-x)_i(y-x)_j \partial_{x_i x_j} f(z_1)\right) \\
&\quad + f(x)\left(\sum_i (y-x)_i \partial_{x_i} r(x,\eta) + \sum_{ij}(y-x)_i(y-x)_j \partial_{x_i x_j} r(z_2,\eta)\right) \\
&= \left(r(x,\eta) + \sum_j (y-x)_j \partial_{x_j} r(z_3,\eta)\right)\left(\sum_i (y-x)_i \partial_{x_i} f(x)\right) \\
&\quad + r(y,\eta)\sum_{ij}(y-x)_i(y-x)_j \partial_{x_i x_j} f(z_1) \\
&\quad + f(x)\left(\sum_i (y-x)_i \partial_{x_i} r(x,\eta) + \sum_{ij}(y-x)_i(y-x)_j \partial_{x_i x_j} r(z_2,\eta)\right),
\end{aligned}$$

for some $z_i = \kappa_i x + (1-\kappa_i)y$. Integrating this against $q(x,dy)$, we get that

$$\begin{aligned}
&\left|\theta \int (r(y,\eta)f(y) - r(x,\eta)f(x))\, q_\theta(x,dy)\right| \\
&\quad \le \left|\sum_i \left(r(x,\eta)\partial_{x_i} f(x) + f(x)\partial_{x_i} r(x,\eta)\right)\theta \int (y-x)_i q_\theta(x,dy)\right| \\
&\qquad + \left|f(x)\theta \int \sum_{ij} \partial_{x_i x_j} r(z_2,\eta)(y-x)_i(y-x)_j q_\theta(x,dy)\right| \\
&\qquad + \left|\theta \int \sum_{ij}(y-x)_i(y-x)_j \left(\partial_{x_i} f(x)\partial_{x_j} r(z_3,\eta) + r(y,\eta)\partial_{x_i x_j} f(z_1)\right) q_\theta(x,dy)\right|.
\end{aligned}$$

Since $q_\theta(x,dy)$ is the density of a Gaussian with mean $\vec{b}(x)/\theta$ and covariance $\mathbf{C}(x)/\theta$, and both $\vec{b}(x)$ and $\mathbf{C}(x)$ are uniformly bounded, so that $\theta\int(y-x)_i q_\theta(x,dy)$ is bounded as well. Furthermore, a change of variables that diagonalizes $\mathbf{C}(x)$ shows for any $g:\mathbb{R}^d \to \mathbb{R}^{d+d}$, that if $C_g = \sup_y \sup_{\|z\|=1} \sum_{ij} g(y)_{ij} z_i z_j$ and $\lambda_* = \sup_y \sup_{\|z\|=1} \sum_{ij} \mathbf{C}(y)_{ij} z_i z_j$ then

$$\theta \int \sum_{ij} g(y)_{ij}(y-x)_i(y-x)_j q_\theta(x,dy) \le C_g \lambda_*.$$

Condition 1 gives uniform bounds on the derivatives of $r(x,\eta) = r(x, \rho_r * \eta(x))$ in this expression and so, provided f also has uniformly bounded first and second derivatives, we have a bound of the form

$$|B_f^\theta| \le K_1 + K_2|f(x)|,$$

for suitable constants K_1, K_2 that depend only on the derivatives of f.

Now suppose instead that Condition 2 of Lemma 2.9 is satisfied. First note that

$$\begin{aligned}|B_f^\theta| &= \left|\theta \int_{\mathbb{R}^n} \left\{f(y)r\Big(y, \rho_r{*}\eta(y)\Big) - f(x)r\Big(x, \rho_r{*}\eta(x)\Big)\right\} q_\theta(x, dy)\right| \\ &\leq \left|\theta \int_{\mathbb{R}^n} \left\{f(y)r\Big(y, \rho_r{*}\eta(y)\Big) - f(x)r\Big(x, \rho_r{*}\eta(y)\Big)\right\} q_\theta(x, dy)\right| \\ &\quad + \left|\theta \int_{\mathbb{R}^n} \left\{f(x)r\Big(x, \rho_r{*}\eta(y)\Big) - f(x)r\Big(x, \rho_r{*}\eta(x)\Big)\right\} q_\theta(x, dy)\right|. \end{aligned} \tag{3.2}$$

(Note the extra term introduced here, $r(x, \rho_r{*}\eta(y))$, has the two arguments to r "at different locations", contrary to the usual pattern.)

Writing $K_3 = \sup_{x,m} \max_i |\partial_{x_i} f(x) r(x, m)|$ and $K_4 = \sup_{x,m} \max_{i,j} |\partial_{x_i x_j} f(x) r(x, m)|$, the first term is bounded exactly as above. For the second,

$$\begin{aligned}&r(x, \rho_r{*}\eta(y)) - r(x, \rho_r{*}\eta(x)) \\ &\qquad = (\rho_r{*}\eta(y) - \rho_r{*}\eta(x)) r'(x, \rho_r{*}\eta(x)) + \frac{1}{2}(\rho_r{*}\eta(y) - \rho_r{*}\eta(x))^2 r''(x, \overline{m}),\end{aligned}$$

where $\overline{m} = \kappa' \rho_r * \eta(x) + (1 - \kappa')\rho_r * \eta(y)$ for some $0 \leq \kappa' \leq 1$, and we have used r' and r'' to denote the first and second derivatives of $r(x, m)$ with respect to the second argument. So, writing $K_5 = \|r'\|_\infty$ and $K_6 = \|r''\|_\infty$, the second term in (3.2) is bounded by $f(x)$ multiplied by

$$K_5 \left|\theta \int_{\mathbb{R}^d} (\rho_r{*}\eta(y) - \rho_r{*}\eta(x)) q_\theta(x, dy)\right| + K_6 \left|\theta \int_{\mathbb{R}^d} (\rho_r{*}\eta(y) - \rho_r{*}\eta(x))^2 q_\theta(x, dy)\right|.$$

Under Condition 2 of Assumptions 2.8, this is bounded by a constant times $\rho_\gamma{*}\eta(x) + (\rho_\gamma{*}\eta(x))^2$ and $\sup_x m^2\gamma(x, m)$ is bounded, so therefore $|\gamma(x, \eta) B_f^\theta(x, \eta)| \leq K_7 + K_8 f(x)$, where K_7 comes from K_3, K_4, and the supremum of γ, while K_8 comes from K_5, K_6, and the supremum of $m^2\gamma(x, m)$. □

3.2 Proof of Theorem 2.10: convergence for the non-local process

In this section we prove Theorem 2.10. This would be implied by convergence of the lookdown process (see [KR11] and [EK19]); however in our setting, because the parameters in the lookdown process depend on the empirical distribution, we actually

use tightness of the sequence of population processes in the proofs of tightness for the corresponding lookdown processes.

Proof [Proof of Theorem 2.10.] The proof follows a familiar pattern. First we extend $\mathbb{R}^d$ to its one-point compactification $\overline{\mathbb{R}}^d$ and establish, in Lemma 3.2, compact containment of the sequence of scaled population processes in $\mathcal{M}_F(\overline{\mathbb{R}}^d)$ (for which, since we have compactified $\mathbb{R}^d$, it suffices to consider the sequence of total masses); armed with this, tightness of the population processes in $\mathcal{D}_{[0,\infty)}(\mathcal{M}_F(\overline{\mathbb{R}}^d))$ follows from tightness of the real-valued processes $(H(\eta_t))_{t\geq 0}$ for a sufficiently large class of test functions H, which we establish through an application of the Aldous-Rebolledo criterion in Lemma 3.3. These ingredients are gathered together in Proposition 3.4 to deduce tightness of the scaled population processes in the larger space $\mathcal{D}_{[0,\infty)}(\mathcal{M}_F(\overline{\mathbb{R}}^d))$.

We then characterise limit points as solutions to a martingale problem in Lemma 3.6; finally in Lemma 3.7 we check that in the process of passing to the limit, no mass 'escaped to infinity', so that in fact the limit points take values in $\mathcal{D}_{[0,\infty)}(\mathcal{M}_F(\mathbb{R}^d))$. □

As advertised, we work with the one-point compactification of $\mathbb{R}^d$ and consider $(\eta_t^N)_{t\geq 0}$ as a sequence of $\mathcal{M}_F(\overline{\mathbb{R}}^d)$-valued processes. Since, for each $K > 0$, $\{\eta : \langle 1, \eta\rangle \leq K\}$ is a compact set in $\mathcal{M}_F(\overline{\mathbb{R}}^d)$, we shall focus on controlling $(\langle 1, \eta_t^N\rangle)_{t\geq 0}$. The key is that Assumptions 2.8 are precisely chosen to guarantee boundedness of the net per-capita reproduction rate.

Lemma 3.1 *Under Assumptions 2.8, for all $f \in C_b^2(\mathbb{R}^d)$ with uniformly bounded first and second derivatives, and all $T > 0$, there exists a $C = C(f, T) < \infty$, independent of N, such that*

$$\mathbb{E}[\langle f, \eta_t^N\rangle] \leq C\mathbb{E}[\langle 1, \eta_0^N\rangle] \tag{3.3}$$

for all $N \geq 1$.

Proof Consider the semimartingale decomposition from equation (2.7):

$$\langle f, \eta_t^N\rangle = \langle f, \eta_0^N\rangle + \int_0^t \int_{\mathbb{R}^d} \left\{\gamma(x, \eta_s^N)B_f^\theta(x, \eta_s^N) + f(x)F(x, \eta_s^N)\right\}\eta_s^N(dx)ds + M_t^N(f), \tag{3.4}$$

where $M_t^N(f)$ is a martingale and B_f^θ is defined in (3.1). First note that Condition 6 of Assumptions 2.8 stipulates that $|\gamma B_f^\theta|$ is uniformly bounded by a contant times $1 + f$,

and so recalling that F is bounded above, we conclude that under Assumptions 2.8 $\gamma(x,\eta)B_f^\theta(x,\eta) + f(x)F(x,\eta) \leq C_f(1+f(x))$ for some C_f.

Now, taking expectations in (3.4),

$$\mathbb{E}\left[\langle f, \eta_t^N\rangle\right] \leq \mathbb{E}\left[\langle f, \eta_0^N\rangle\right] + C_f \int_0^t \mathbb{E}\left[\langle 1+f, \eta_s^N\rangle\right] dt. \tag{3.5}$$

The bound (3.3) then follows by first applying Gronwall's inequality in the case $f = 1$, which yields

$$\mathbb{E}\Big[\langle 1, \eta_t^N\rangle\Big] \leq e^{Ct}\mathbb{E}\Big[\langle 1, \eta_0^N\rangle\Big],$$

with C independent of N, and then substituting the resulting bound on $\mathbb{E}\Big[\langle 1, \eta_s^N\rangle\Big]$ into the expression above. □

With a bound on per-capita net growth rate in hand, bounds on the expectation of the supremum of the total population size over a finite time interval also follow easily.

Lemma 3.2 (Compact containment for the population process) *Under the assumptions of Theorem 2.10, for each $T > 0$, there exists some constant C_T, independent of N, such that*

$$\mathbb{E}\left[\sup_{0\leq t\leq T}\langle 1, \eta_t^N\rangle\right] \leq C_T\mathbb{E}[\langle 1, \eta_0\rangle]. \tag{3.6}$$

In particular, for any $\delta > 0$, there exists $K_\delta > 0$ such that

$$\limsup_{N\to\infty}\mathbb{P}\left\{\sup_{s\in[0,T]}\langle 1, \eta_s^N\rangle > K_\delta\right\} \leq \frac{C_T}{K_\delta} < \delta. \tag{3.7}$$

Proof First note that by Lemma 3.1, $\mathbb{E}[\langle 1, \eta_t^N\rangle] \leq \mathbb{E}[\langle 1, \eta_0^N\rangle]e^{Ct}$ for some C (independent of N). Now, let $M_t^{N*}(f) = \sup_{0\leq s\leq t} M_t^N(f)$, and as before let $\langle M^N(f)\rangle_t$ be the angle bracket process of $M_t^N(f)$. The Burkholder-Davis-Gundy inequality says that there is a K for which $\mathbb{E}\left[M_t^{N*}(1)\right] \leq K\mathbb{E}[\sqrt{[M^N(1)]_t}]$, where $[M^N(1)]_t$ is the quadratic variation of $M^N(1)$. Furthermore, as discussed by [HJ22], the expectation of the quadratic variation of a local martingale is bounded by a (universal) constant multiple of the expectation of its angle bracket process [BJY86, Item (4.b'), Table 4.1, p. 162]. Now, since $\sqrt{x} \leq 1 + x$,

in the notation of Lemma 3.1, there is a C' such that

$$\begin{aligned}
\mathbb{E}\left[M_t^{N*}(1)\right] &\le C'\left(1+\mathbb{E}\left[\left\langle M^N(1)\right\rangle_t\right]\right)\\
&= C'\left(1+\frac{\theta}{N}\mathbb{E}\left[\int_0^t\left\langle\left\{\gamma(x,\eta_s^N)\int_{\mathbb{R}^d} r(y,\eta_s^N)q_\theta(x,dy)+\mu_\theta(x,\eta_s^N)\right\},\eta_s^N(dx)\right\rangle ds\right]\right)\\
&= C'\Big(1+\mathbb{E}\Big[\int_0^t\Big\langle\Big\{\frac{2\theta}{N}\gamma(x,\eta_s^N)r(x,\eta_s^N)\\
&\qquad\qquad +\frac{\gamma(x,\eta_s^N)}{N}B_1^\theta(x,\eta_s^N)-\frac{1}{N}F(x,\eta_s^N)\Big\},\eta_s^N(dx)\Big\rangle ds\Big]\Big).
\end{aligned}$$

We have not assumed that F is bounded below, but to see that the term involving $-F$ does not cause us problems, we rearrange equation (3.4) with $f=1$ to see that

$$\begin{aligned}
\mathbb{E}\left[\int_0^t\left\langle -F(x,\eta_s^N),\eta_s^N(dx)\right\rangle ds\right] &= \mathbb{E}[\langle 1,\eta_0^N\rangle]-\mathbb{E}[\langle 1,\eta_t^N\rangle]\\
&\quad+\mathbb{E}\left[\int_0^t\left\langle\gamma(x,\eta_s^N)B_1^\theta(x,\eta_s^N),\eta_s^N(dx)\right\rangle ds\right],
\end{aligned}\tag{3.8}$$

which is bounded since $\gamma(x,\eta)$ and $B_1^\theta(x,\eta)$ are both bounded and $\langle 1,\eta_t\rangle\ge 0$. Since $\theta/N\to\alpha<\infty$, combining constants, we obtain that for some C'',

$$\mathbb{E}\left[M_t^{N*}(1)\right]\le C'+C''\mathbb{E}[\langle 1,\eta_0^N\rangle]e^{tC}.$$

Taking suprema and expectations on both sides of equation (3.4), then again using the fact that $\gamma(x,\eta)B_1^\theta(x,\eta)+F(x,\eta)\le C$,

$$\begin{aligned}
\mathbb{E}\left[\sup_{0\le s\le T}\langle 1,\eta_s^N\rangle\right] &\le \mathbb{E}[\langle 1,\eta_0^N\rangle]+\mathbb{E}\left[\sup_{0\le t\le T}\int_0^t\left\langle\left\{\gamma(x,\eta_s^N)B_1^\theta(x,\eta_s^N)+F(x,\eta_s^N)\right\},\eta_s^N(dx)\right\rangle ds\right]\\
&\quad+\mathbb{E}[M_t^{N*}(1)]\\
&\le \mathbb{E}[\langle 1,\eta_0^N\rangle]+C\mathbb{E}\left[\int_0^T\sup_{0\le s\le t}\langle 1,\eta_s^N\rangle dt\right]+C'+C''\mathbb{E}[\langle 1,\eta_0^N\rangle]e^{tC}.
\end{aligned}$$

Once again applying Gronwall's inequality,

$$\mathbb{E}\left[\sup_{0\le s\le T}\langle 1,\eta_s^N\rangle\right]\le C'''\left(1+\mathbb{E}[\langle 1,\eta_0^N\rangle]\right)e^{2TC}.$$

For any T, the quantity on the right is bounded above by a constant $C(T)$ independent of N. As a result, for any $K>0$,

$$\limsup_{N\to\infty}\mathbb{P}\left[\sup_{0\le s\le T}\langle 1,\eta_s^N\rangle\ge K\right]\le\frac{C(T)}{K}.$$

□

Our next task is to show tightness of $(\langle f,\eta_t^N\rangle)_{t\ge 0}$ for $f\in C_b^\infty(\overline{\mathbb{R}}^d)$.

Lemma 3.3 (Tightness of $(\langle f, \eta_t^N\rangle)_{t>0}$) *For each $f \in C_b^\infty(\overline{\mathbb{R}}^d)$, the collection of processes $(\langle f, \eta_t^N\rangle)_{t\geq 0}$ for $N = 1, 2, \ldots$ is tight as a sequence of càdlàg, real-valued processes.*

Proof The Aldous-Rebolledo criterion (Theorem A.6) applied to the semimartingale representation of $\langle f, \eta_t^N\rangle$ of equation (3.4), tells us that it suffices to show that for each $T > 0$, (a) for each fixed $0 \leq t \leq T$, the sequence $\{\langle f, \eta_t^N\rangle\}_{N\geq 1}$ is tight, and (b) for any sequence of stopping times τ_N bounded by T, and for each $\nu > 0$, there exist $\delta > 0$ and $N_0 > 0$ such that

$$\sup_{N>N_0}\sup_{t\in[0,\delta]} \mathbb{P}\left\{\left|\int_\tau^{\tau+t}\int_{\mathbb{R}^d}\left\{\gamma(x,\eta_s^N)B_f^\theta(x,\eta_s^N) + f(x)F(x,\eta_s^N)\right\}\eta_s^N(dx)ds\right| > \nu\right\} < \nu, \quad (3.9)$$

$$\text{and} \qquad \sup_{N>N_0}\sup_{t\in[0,\delta]} \mathbb{P}\left\{\left|\langle M^N(f)\rangle_{\tau+t} - \langle M^N(f)\rangle_\tau\right| > \nu\right\} < \nu. \quad (3.10)$$

Tightness of $\langle f, \eta_t^N\rangle$ for fixed t follows from Lemma 3.1 and Markov's inequality, so we focus on the remaining conditions.

The proof of Lemma 3.1 provides a uniform bound on γB_f^θ, but we only know that F is bounded above. However, by assumption, for each fixed value of m, $\sup_{k\leq m}|F(x,k)|$ is uniformly bounded as a function of x. Noting that $\rho_F * \eta \leq \langle 1, \eta\rangle\|\rho_F\|_\infty$, we can use Lemma 3.2 to choose N_0 and K such that if $N > N_0$, then

$$\mathbb{P}\left\{\sup_{0\leq s\leq T}\langle 1, \eta_s^N\rangle \geq K\right\} < \nu/2,$$

we now choose δ_1 so that

$$\delta_1\|f\|_\infty \sup\left\{\sup_x |F(x,k)| : k \leq K\|\rho_F\|_\infty\right\} < \nu/4, \qquad \sup_{x,\eta}\gamma(x,\eta)B_f^\theta(\,\eta)\delta_1 < \nu/4,$$

so that (3.9) is satisfied with $\delta = \delta_1$.

Similarly,

$$\begin{aligned}
&\left|\langle M^N(f)\rangle_{\tau+t} - \langle M^N(f)\rangle_\tau\right| \\
&\quad = \left|\int_\tau^{\tau+t}\frac{\theta}{N}\int_{\mathbb{R}^d}\left\{\gamma(x,\eta_s^N)\int_{\mathbb{R}^d} f^2(y)r(y,\eta_s^N)q_\theta(x,dy) + \mu_\theta(x,\eta_s^N)f^2(x)\right\}\eta_s^N(dx)ds\right| \\
&\quad = \left|\int_\tau^{\tau+t}\frac{\theta}{N}\int_{\mathbb{R}^d}\left\{\gamma(x,\eta_s^N)\left(2f^2(x)r(x,\eta_s^N) + B_{f^2}^\theta(x,\eta_s^N)\right) - f^2(x)\frac{F(x,\eta_s^N)}{\theta}\right\}\eta_s^N(dx)ds\right|,
\end{aligned}$$

and so using the fact that $\theta/N \to \alpha < \infty$, an argument entirely analogous to that for (3.9) yields a δ_2 for which (3.10) is satsified. Taking $\delta = \min\{\delta_1, \delta_2\}$, the result follows.

□

We collect the implications of the last two lemmas into a proposition.

Proposition 3.4 (Tightness of $(\eta_t^N)_{t\geq 0}$) *The collection of measure-valued processes $\{(\eta_t^N)_{t\geq 0} : N \geq 1\}$ is tight in $\mathcal{D}_{[0,\infty)}(\mathcal{M}_F(\overline{\mathbb{R}}^d))$.*

Proof Theorem 3.9.1 in [EK86] says that if the collection of E-valued processes satisfies a compact containment condition (for any $\epsilon > 0$ and $T > 0$, there is a compact set such that the processes stay within that set up to time T with probability at least $1 - \epsilon$), then the collection is relatively compact (which is equivalent to tightness since we are working on a Polish space) if and only if $\{(f(\eta_t^N))_{t\geq 0} : N \geq 1\}$ is relatively compact for all f in a dense subset of $C_b(E)$ under the topology of uniform convergence in compact sets.

Since $\{\nu : \langle 1, \nu\rangle \leq K\}$ is compact in $\mathcal{M}_F(\overline{\mathbb{R}}^d)$, Lemma 3.2 gives compact containment. Lemma 3.3 shows that the real-valued processes $\langle f, \eta_t^N\rangle$ are relatively compact for all $f \in \mathcal{C}_b^\infty(\overline{\mathbb{R}}^d)$. Since by the Stone-Weierstrass theorem, the algebra of finite sums and products of terms of this form is dense in the space of bounded continuous functions on $\mathcal{M}_F(\overline{\mathbb{R}}^d)$, and tightness of $\langle f, \eta_t^N\rangle$ extends to sums and products of this form by Lemma A.7, we have relative compactness in $\mathcal{D}_{[0,\infty)}(\mathcal{M}_F(\overline{\mathbb{R}}^d))$. □

We wish to characterise the limit points of $\{(\eta_t^N)_{t>0}\}_{N\geq 1}$ as solutions to a martingale problem with generator $\mathcal{P}^\infty$ which we now identify. Most of the work was done in Section 2.4. First, we record an equivalent formulation of the martingale problems, which were essentially laid out in Subsection 2.4.1.

Lemma 3.5 *For $G \in \mathcal{C}^\infty(\mathbb{R})$ with $\|G'''\|_\infty < \infty$, and $f \in \mathcal{C}_b^\infty(\overline{\mathbb{R}}^d)$, define the function G_f by $G_f(\eta) := G(\langle f, \eta\rangle)$. Let $\mathcal{P}^N$ be the generator given by*

$$\begin{aligned}\mathcal{P}^N G_f(\eta) := \theta N \Big\langle \gamma(x,\eta) \int \big(G(\langle f,\eta\rangle + f(z)/N) - G(\langle f,\eta\rangle)\big)\, r(z,\eta) q_\theta(x,dz) \\ + \big(G(\langle f,\eta\rangle - f(x)/N) - G(\langle f,\eta\rangle)\big)\, \mu_\theta(x,\eta), \eta(dx)\Big\rangle.\end{aligned} \tag{3.11}$$

The process $(\eta_t^N)_{t\geq 0}$ of Definition 2.4 is the unique solution to the $(\mathcal{P}^N, \eta_0)$-martingale problem, i.e., for all such test functions

$$M_t := G_f(\eta_t^N) - G_f(\eta_0^N) - \int_0^t \mathcal{P}^N G_f(\eta_s^N) ds$$

is a martingale (with respect to the natural σ-field).

Furthermore, let $\mathcal{P}^\infty$ be the generator given by

$$\begin{aligned}\mathcal{P}^\infty G_f(\eta) := G'(\langle f,\eta\rangle)\Big\langle \gamma(x,\eta)\mathcal{B}\left(f(\cdot)r(\cdot,\eta)\right)(x) + f(x)F(x,\eta), \eta(dx)\Big\rangle \\ + \alpha G''(\langle f,\eta\rangle)\Big\langle \gamma\left(x,\eta\right) r\left(x,\eta\right) f^2(x), \eta(dx)\Big\rangle.\end{aligned} \tag{3.12}$$

A process $(\eta_t^\infty)_{t\geq 0}$ satisfies the martingale characterization of equations (2.11) *and* (2.12) *if it is a solution to the $(\mathcal{P}^\infty, \eta_0^\infty)$-martingale problem, i.e., for all such test functions*

$$M_t := G_f(\eta_t^\infty) - G_f(\eta_0^\infty) - \int_0^t \mathcal{P}^\infty G_f(\eta_s^\infty) ds$$

is a martingale (with respect to the natural σ-field).

Lemma 3.6 (Characterisation of limit points) *Suppose that $(\eta_0^N)_{N\geq 1}$ converges weakly to η_0 as $N\to\infty$. Then any limit point of $\{(\eta_t^N)_{t\geq 0}\}_{N\geq 1}$ in $\mathcal{D}_{[0,\infty)}(\mathcal{M}_F(\overline{\mathbb{R}}^d))$ is a solution to the martingale problem for $(\mathcal{P}^\infty, \eta_0)$.*

Proof We use Theorem 4.8.2 in [EK86]. First observe that the set of functions $\{G_f(\eta) := G(\langle f,\eta\rangle),\ G\in\mathcal{C}^\infty(\mathbb{R}), \|G'''\|_\infty<\infty,\ f\in\mathcal{C}_b^\infty(\overline{\mathbb{R}}^d)\}$ is separating on $\mathcal{M}_F(\overline{\mathbb{R}}^d)$. Therefore, it suffices to show that for any $t>0$ and $\tau>0$ that

$$\lim_{N\to\infty}\mathbb{E}\left[\left(G_f(\eta_{t+\tau}^N) - G_f(\eta_t^N) - \int_t^{t+\tau}\mathcal{P}^\infty G_f(\eta_s^N)ds\right)\prod_{i=1}^k h_i(\eta_{t_i}^N)\right] = 0 \tag{3.13}$$

for all $k\geq 0$, $0\leq t_1<t_2<\ldots,t_k\leq t<t+\tau$, and bounded continuous functions $h_1,\ldots,h_k$ on $\mathcal{M}_F(\overline{\mathbb{R}}^d)$. Since $(\eta_t^N)_{t\geq 0}$ is Markov, the tower property gives that, for each N,

$$\mathbb{E}\left[\left(G_f(\eta_{t+\tau}^N) - G_f(\eta_t^N) - \int_t^{t+\tau}\mathcal{P}^N G_f(\eta_s^N)ds\right)\prod_{i=1}^k h_i(\eta_{t_i}^N)\right] = 0. \tag{3.14}$$

Therefore, it suffices to show that

$$\lim_{N\to\infty}\mathbb{E}\left[\int_t^{t+\tau}\left|\mathcal{P}^N G_f(\eta_s^N) - \mathcal{P}^\infty G_f(\eta_s^N)\right| ds\prod_{i=1}^k h_i(\eta_{t_i}^N)\right] = 0, \tag{3.15}$$

and, again using the tower property, since the functions h_i are bounded, this will follow if

$$\lim_{N\to\infty}\mathbb{E}\left[\int_t^{t+\tau}\left|\mathcal{P}^N G_f(\eta_s^N) - \mathcal{P}^\infty G_f(\eta_s^N)\right| ds\Big|\mathcal{F}_t\right] = 0 \tag{3.16}$$

(where $\{\mathcal{F}_t\}_{t\geq 0}$ is the natural σ-field).

We rewrite $\mathcal{P}^N G_f(\eta_s^N)$ using a Taylor series expansion up to third order for $G\Big(\langle f,\eta\rangle \pm f(y)/N\Big)$ around $G(\langle f,\eta\rangle)$. As in Section 2.4 (except that now we are more explicit about the error term), we find

$$\begin{aligned}\mathcal{P}^N G_f(\eta) :=& G'(\langle f,\eta\rangle)\int_{\mathbb{R}^d}\theta\Big\langle \gamma(x,\eta)\int_{\mathbb{R}^d} f(y)r(y,\eta)q_\theta(x,dy) - f(x)\mu_\theta(x,\eta),\eta(dx)\Big\rangle \\ &+ \frac{1}{2}\frac{\theta}{N}G''(\langle f,\eta\rangle)\int_{\mathbb{R}^d}\Big\langle \gamma(x,\eta)\int_{\mathbb{R}^d} f^2(y)r(y,\eta)q_\theta(x,dy) + f^2(x)\mu_\theta(x,\eta),\eta(dx)\Big\rangle \\ &+ \frac{1}{6}\frac{\theta}{N^2}\Big\langle G'''(w)\gamma(x,\eta)\int_{\mathbb{R}^d} f^3(y)r(y,\eta)q_\theta(x,dy) - G'''(v)f^3(x)\mu_\theta(x,\eta),\eta(dx)\Big\rangle \end{aligned} \tag{3.17}$$

for some $w, v \in [\langle f,\eta\rangle - \|f\|_\infty/N, \langle f,\eta\rangle + \|f\|_\infty/N]$.

Combining with equation (2.37), and the fact that $\mu_\theta(x,\eta) \to r(x,\eta)\gamma(x,\eta)$ as $\theta \to \infty$, we have pointwise convergence:

$$\lim_{N\to\infty} |\mathcal{P}^N G(\langle f,\eta\rangle) - \mathcal{P}^\infty G(\langle f,\eta\rangle)| = 0. \tag{3.18}$$

To conclude convergence of the expectation, we would like to apply the Dominated Convergence Theorem in (3.15). Recall that f and G and their derivatives are bounded, and $\gamma(x,\eta)$ is bounded independent of θ. Since $\theta/N^2 \to 0$, rearranging as in (2.35) and using the convergence of (2.36), we deduce that we can dominate $\Big|\mathcal{P}^N G_f(\eta_s^N) - \mathcal{P}^\infty G_f(\eta_s^N)\Big|$ by a constant multiple of $\langle 1 + |F(x)|, \eta_s(dx)\rangle$. Since F is bounded above, there is a constant K such that $|F| \le K - F$ so that, exactly as in equation (3.8), we can check that

$$\mathbb{E}\Big[\int_t^{t+\tau}\Big\langle |F(x,\eta_s^N)|, \eta_s^N(dx)\Big\rangle ds\Big|\mathcal{F}_t\Big] < \infty,$$

which concludes our proof. □

The last step in the proof of Theorem 2.10 is to check that any limit point $(\eta_t)_{t\ge0}$ of $\{(\eta_t^N)_{t\ge0}\}_{N\ge1}$ actually takes its values in $\mathcal{M}_F(\mathbb{R}^d)$, that is, "no mass has escaped to infinity".

Lemma 3.7 *Under the assumptions of Theorem 2.10, let $(\eta_t)_{t\ge0}$ be a limit point of $\{(\eta_t^N)_{t\ge0}\}_{N\ge1}$. For any $\delta > 0$, and any $T > 0$,*

$$\mathbb{P}\Big[\sup_{t\in[0,T]} \eta_t\Big(\{\|x\| > R\}\Big) > \delta\Big] \to 0 \qquad \text{as } R \to \infty.$$

Proof [Sketch] Take $f_0(x)$ as in the statement of Theorem 2.10, i.e., f_0 is nonnegative, grows to infinity as $x \to \infty$, has uniformly bounded first and second derivatives, and has $\langle f_0, \eta_0^N \rangle$ uniformly bounded in N. We take a sequence of test functions f_n that increase to the function f_0 and having uniformly bounded first and second derivatives, so that there is a (single) C from Condition 6 of Assumptions 2.8 such that $\gamma(x,\eta)B_{f_n}^{\theta}(x,\eta) \leq C(1+f_n(x))$ for all x, η, and f_n. Then, just as we arrived at equation (3.5),

$$\mathbb{E}[\langle f_n(x), \eta_t^N(dx)\rangle] \leq \mathbb{E}[\langle f_n(x), \eta_0^N(dx)\rangle] + C\int_0^t \mathbb{E}\Big[\Big\langle f_n(x), \eta_s^N(dx)\Big\rangle\Big]ds,$$

with the same constant for all n and all N. Gronwall's inequality then implies that $\mathbb{E}[\langle f_n, \eta_t^N\rangle] \leq C'$ for some C' independent of n, N, and $t \in [0,T]$. By first taking $N \to \infty$ and then $n \to \infty$, we find that $\mathbb{E}[\langle f_0, \eta_t(dx)\rangle] \leq C'$ for $t \in [0,T]$, and since $f_0 \to \infty$ as $|x| \to \infty$, an application of Markov's inequality tells us that for any $\delta > 0$,

$$\mathbb{P}\Big\{\eta_t(\{x : \|x\| > R\}) > \delta\Big\} \leq \frac{\mathbb{E}[\langle f_0, \eta_t\rangle]}{\inf_{\{x:\|x\|\geq R\}} f_0(x)} \to 0 \qquad \text{as } R \to \infty.$$

We have now established control for arbitary fixed time $t > 0$. To control the whole process, we must show that $\mathbb{E}[\sup_{t\leq T}\langle f_n, \eta_t^N\rangle] < K'$ for some K' independent of n, N. To do so, we adopt the same BDG-inequality techniques used to prove Lemma 3.2, but applied to $M_t^{N*}(f_n)$, to give

$$\mathbb{E}\left[\sup_{0\leq s\leq T}\langle f_n, \eta_s^N\rangle\right] \leq C'''\left(1 + \mathbb{E}[\langle f_n, \eta_0^N\rangle]\right)e^{2TC}.$$

for constant C, C''' independent of n, N. Since $\mathbb{E}[\langle f_n, \eta_0^N\rangle]$ also has a uniform bound over n, we have established the uniform control required for this proof.

□

3.3 Convergence of some nonlocal equations to classical PDEs

It is natural to conjecture that when the limit of the rescaled population process that we obtained in the previous section solves a nonlocal PDE, if we further scale the kernels ρ_r, ρ_γ, and ρ_F by setting $\rho^\epsilon(\cdot) = \rho(\cdot/\epsilon)/\epsilon^d$, as $\epsilon \to 0$, the corresponding solutions should converge to a limiting population density that solves the corresponding "classical" PDE. We verify

this in two examples; in the first the nonlocal equation is a reaction-diffusion equation with the "nonlocality" only appearing in the reaction term; in the second the nonlocal PDE is a special case of a nonlinear porous medium equation. These, in particular, capture the examples that we explored in Section 2.3.2.

3.3.1 Reaction–diffusion equation limits

In this subsection we prove Proposition 2.15. *The conditions of the proposition are in force throughout this subsection. In particular, $\mathcal{B}^*$ is assumed to be a Markovian generator.* The proof rests on a Feynman-Kac representation. We write $(Z_t)_{t\geq 0}$ for a diffusion with generator $\mathcal{B}^*$ and denote its transition density by $f_t(x,y)$. The first step is a regularity result for this density.

Lemma 3.8 *Fix $T > 0$. There exists a constant $K = K(T) > 0$ such that, for any $x, y \in \mathbb{R}^d$ and $t \in [0,T]$,*

$$\int |f_t(x,z) - f_t(y,z)|dz \leq \frac{\|x-y\|}{\sqrt{t}} K. \tag{3.19}$$

Proof We first use the Intermediate Value Theorem to obtain the bound

$$\int |f_t(x,z) - f_t(y,z)|dz \leq \int \|x-y\| \|\nabla f_t(w,z)\| dz$$

where ∇ acts on the first coordinate only and w is in the line segment $[x,y]$ joining x to y. Under our assumptions on b and C, equation (1.3) of [She91], gives existence of constants $\lambda = \lambda(T) > 0$ and K such that,

$$\|\nabla f_t(w,z)\| \leq \frac{K}{\sqrt{t}} p_{\lambda t}(w,z),$$

where $p_s(x,y)$ is the Brownian transition density. Hence,

$$\int |f_t(x,z) - f_t(y,z)|dz \leq K\frac{\|x-y\|}{\sqrt{t}} \int p_{\lambda t}(w,z)dz = K\frac{\|x-y\|}{\sqrt{t}}.$$

□

Lemma 3.9 *Fix $T > 0$. Let $x, y \in \mathbb{R}^d$, $t \in [0,T]$, and denote by $(Z_t^y)_{t\geq 0}$ and $(Z_t^x)_{t\geq 0}$ independent copies of the diffusion $(Z_t)_{t\geq 0}$ starting from y and x respectively. There exists a constant $K = K(T) > 0$ such that,*

$$\mathbb{E}[\|Z_t^y - Z_t^x\|] \leq K(\sqrt{t} + \|y - x\|).$$

Proof First we write,

$$\mathbb{E}[\|Z_t^y - Z_t^x\|] = \int\int \|u-v\| f_t(y,u) f_t(x,v) du dv.$$

Under our regularity assumptions on C, b, using equation (1.2) of [She91], there exist constants K, $\lambda = \lambda(T) > 0$ for which,

$$f_t(y,u) \leq K p_{\lambda t}(y,u).$$

It then follows that,

$$\mathbb{E}[\|Z_t^y - Z_t^x\|] \leq \int\int \|u-v\| K^2 p_{\lambda t}(y,u) p_{\lambda t}(x,v) dv du = K^2 \mathbb{E}[\|B_{\lambda t}^y - B_{\lambda t}^x\|], \tag{3.20}$$

where $(B_t^y)_{t\geq 0}$ and $(B_t^x)_{t\geq 0}$ are independent Brownian motions starting at y and x respectively. Using the triangle inequality, and writing $(B_t^0)_{t\geq 0}$ for a Brownian motion started from the origin,

$$\mathbb{E}[\|B_{\lambda t}^y - B_{\lambda t}^x\|] \leq \|y-x\| + \mathbb{E}[\|B_{2\lambda t}^0\|] \leq \|y-x\| + C\sqrt{t}. \tag{3.21}$$

Substituting (3.21) in (3.20) gives the result. □

We use the representations of the solutions to equations (2.16) and (2.15) respectively:

$$\varphi_t(x) = \mathbb{E}_x\Big[\varphi_0(Z_t) + \int_0^t \varphi_s(Z_s) F(\varphi_s(Z_s)) ds\Big], \tag{3.22}$$

$$\varphi_t^\epsilon(x) = \mathbb{E}_x\Big[\varphi_0(Z_t) + \int_0^t \varphi_s^\epsilon(Z_s) F(\rho_F^\epsilon * \varphi_s^\epsilon(Z_s)) ds\Big], \tag{3.23}$$

from which

$$\varphi_t(x) - \varphi_t^\epsilon(x) = \mathbb{E}_x\left[\int_0^t \Big(\varphi_s(Z_s) F\Big(\varphi_s(Z_s)\Big) - \varphi_s^\epsilon(Z_s) F\Big(\rho_\epsilon * \varphi_s^\epsilon(Z_s)\Big)\Big) ds\right], \tag{3.24}$$

where $\mathbb{E}_x$ denotes expectation for Z with $Z_0 = x$. The key to our proof of Proposition 2.15 will be to replace $F(\varphi_s(Z_s))$ by $F(\rho_F^\epsilon * \varphi_s(Z_s))$ in this expression. We achieve this through three lemmas.

First we need a uniform bound on φ and φ^ϵ.

Lemma 3.10 *For any $T > 0$ there exists $M = M(T, \|\varphi_0\|) > 0$ such that, for all $0 \leq t \leq T$:*

$$\max\{\|\varphi_t(\cdot)\|_\infty, \|\varphi_t^\epsilon(\cdot)\|_\infty\} < M.$$

Proof Using that φ_0 and F are bounded above, from the representation (3.22), we have

$$\varphi_t(x) \leq \|\varphi_0\|_\infty + K\mathbb{E}\Big[\int_0^t \varphi_s(Z_s)ds\Big].$$

Taking supremum norm on both sides, and applying Gronwall's inequality,

$$\|\varphi_t(\cdot)\|_\infty \leq \|\varphi_0\|_\infty \exp(KT).$$

Similarly, $\|\varphi_t^\epsilon(\cdot)\|_\infty \leq \|\varphi_0\|_\infty \exp(KT)$. □

We also need a continuity estimate for φ.

Lemma 3.11 *Let $T > 0$. There exists a constant $K = K(T, \|\varphi_0\|_\infty) > 0$ and $\delta_0 = \delta_0(T, \|\varphi_0\|_\infty) > 0$ such that for all $0 < \delta < \delta_0$ and $0 \leq t \leq T$,*

$$\|x - y\| < \delta^3 \Rightarrow |\varphi_t(x) - \varphi_t(y)| < K\delta.$$

Proof First we need some notation. Fix $T > 0$ and write M for the corresponding constant from Lemma 3.10. Let $\|F\|_M = \sup_{m\in[0,M]} |F(m)|$. We reserve $\widehat{K}$ for the constant on the right hand side of equation (3.19) and $\widetilde{K}$ for the constant in Lemma 3.9, and write K_{φ_0} for the Lipschitz constant of φ_0. Set

$$\delta_0 = \min\Big(\frac{1}{\|F\|_M^2}, \frac{1}{Me\big(2\|F\|_M + \widehat{K}\big)}, \frac{1}{\widetilde{K}K_{\varphi_0} + 2\|F\|_M M}, 1\Big).$$

In what follows we take $0 < \delta < \delta_0$.

We first prove that the result holds if $t < \delta^2$. As before let Z_t^x and Z_t^y be independent copies of the diffusion Z_t starting at x and y respectively. From our representation (3.22) and Lemma 3.10, we can write:

$$\begin{aligned}
|\varphi_t(x) - \varphi_t(y)| &\leq \Big|\mathbb{E}_x[\varphi_0(Z_t)] - \mathbb{E}_y[\varphi_0(Z_t)]\Big| + 2\|F\|_M Mt \\
&\leq \mathbb{E}[|\varphi_0(Z_t^x) - \varphi_0(Z_t^y)|] + 2\|F\|_M Mt \\
&\leq K_{\varphi_0}\mathbb{E}[\|Z_t^x - Z_t^y\|] + 2\|F\|_M Mt \\
&\leq \widetilde{K}K_{\varphi_0}(\sqrt{t} + \|y - x\|) + 2\|F\|_M Mt \\
&\leq \widetilde{K}K_{\varphi_0}(\delta + \delta^3) + 2\|F\|_M M\delta^2 \leq (\widetilde{K}K_{\varphi_0} + 1)\delta,
\end{aligned}$$

where we have used Lemma 3.9 in the fourth inequality and the definition of δ_0 in the last inequality.

Suppose now that $\delta^2 < t$. We will follow the pattern in Lemma 2.2 of [Pen17]. First, note that by the Feynman-Kac formula we have an alternative representation for $\varphi_t(x)$: for any $t' < t$,

$$\varphi_t(x) = \mathbb{E}_x\Big[\varphi_{t-t'}(Z_{t'}) \exp\Big(\int_0^{t'} F(\varphi_{t-s}(Z_s))ds\Big)\Big].$$

Therefore, setting $t' = \delta^2$ and using Lemma 3.10, for all z,

$$e^{-\delta^2\|F\|_M}\mathbb{E}_z\Big[\varphi_{t-\delta^2}(Z_{\delta^2})\Big] \le \varphi_t(z) \le e^{\delta^2\|F\|_M}\mathbb{E}_z\Big[\varphi_{t-\delta^2}(Z_{\delta^2})\Big].$$

We can then deduce that

$$\begin{aligned}
\varphi_t(x) - \varphi_t(y) &\le e^{\delta^2\|F\|_M}\mathbb{E}_x\Big[\varphi_{t-\delta^2}(Z_{\delta^2})\Big] - e^{-\delta^2\|F\|_M}\mathbb{E}_y\Big[\varphi_{t-\delta^2}(Z_{\delta^2})\Big] \\
&= e^{\delta^2\|F\|_M}\Big(\mathbb{E}_x\Big[\varphi_{t-\delta^2}(Z_{\delta^2})\Big] - \mathbb{E}_y\Big[\varphi_{t-\delta^2}(Z_{\delta^2})\Big]\Big) \\
&\qquad + \Big(e^{\delta^2\|F\|_M} - e^{-\delta^2\|F\|_M}\Big)\mathbb{E}_y\Big[\varphi_{t-\delta^2}(Z_{\delta^2})\Big] \\
&\le e^{\delta^2\|F\|_M}\Big(\mathbb{E}_x\Big[\varphi_{t-\delta^2}(Z_{\delta^2})\Big] - \mathbb{E}_y\Big[\varphi_{t-\delta^2}(Z_{\delta^2})\Big]\Big) + M\Big(e^{\delta^2\|F\|_M} - e^{-\delta^2\|F\|_M}\Big).
\end{aligned} \tag{3.25}$$

To bound the differences of the expected values in the last equation note that, by using again Lemma 3.10,

$$\begin{aligned}
&\mathbb{E}_x\Big[\varphi_{t-\delta^2}(Z_{\delta^2})\Big] - \mathbb{E}_y[\varphi_{t-\delta^2}(Z_{\delta^2})\Big] \\
&\quad = \int \varphi_{t-\delta^2}(z)(f_{\delta^2}(x,z) - f_{\delta^2}(y,z))dz \\
&\quad \le M\int \Big|f_{\delta^2}(x,z) - f_{\delta^2}(y,z)\Big|dz \\
&\quad \le M\widehat{K}\frac{\|x-y\|}{\delta} \le M\widehat{K}\delta^2,
\end{aligned}$$

where we have used Lemma 3.8 and that $\|x-y\| < \delta^3$. Substituting in (3.25),

$$\begin{aligned}
\varphi_t(x) - \varphi_t(y) &\le e^{\delta^2\|F\|_M}\left(M\widehat{K}\delta^2 + M - Me^{-2\delta^2\|F\|_M}\right) \\
&\le e^{\delta^2\|F\|_M}\left(M\widehat{K}\delta^2 + 2M\delta^2\|F\|_M\right) \\
&\le e\left(M\widehat{K} + 2M\|F\|_M\right)\delta^2 \le \delta,
\end{aligned}$$

where the last two inequalities follow from the definition of δ. Interchanging x and y yields the same bound for $\varphi_t(y) - \varphi_t(x)$, and the result follows. $\square$

We proceed to control the difference between $F(\varphi)$ and $F(\rho_F^\epsilon * \varphi)$. Note first that since $\rho_F \in L^1$,

$$I(\epsilon) := \int_{\{\|y\| > \epsilon^{3/4}\}} \rho_F^\epsilon(y)dy = \int_{\{\|y\| > \epsilon^{-1/4}\}} \rho_F(y)dy \to 0 \qquad \text{as } \epsilon \to 0.$$

Lemma 3.12 *Let $T > 0$. There exists a constant $C = C(T, \|\varphi_0\|_\infty) > 0$ such that, for all $0 \le t \le T$, for all ϵ small enough,*

$$\|\varphi_t(\cdot) - \rho_F^\epsilon * \varphi_t(\cdot)\|_\infty \le C(I(\epsilon) + \epsilon^{1/4}). \tag{3.26}$$

Furthermore, there is a constant $\widetilde{C}(T, \|\varphi_0\|_\infty) = \widetilde{C}$ such that, for all $0 \le t \le T$,

$$\|F(\varphi_t(\cdot)) - F(\rho_F^\epsilon * \varphi_t(\cdot))\|_\infty \le \widetilde{C}\Big(I(\epsilon) + \epsilon^{1/4}\Big). \tag{3.27}$$

Proof Let $\epsilon < \delta_0^4$, with δ_0 from Lemma 3.11. Then,

$$\begin{aligned}
|\varphi_t(x) - \rho_F^\epsilon * \varphi_t(x)| &\le \int_{\|x-y\| > \epsilon^{3/4}} \rho_F^\epsilon(x-y)|\varphi_t(y) - \varphi_t(x)|dy \\
&\quad + \int_{\|x-y\| \le \epsilon^{3/4}} \rho_F^\epsilon(x-y)|\varphi_t(y) - \varphi_t(x)|dy \\
&\le 2M \int_{\|x-y\| > \epsilon^{3/4}} \rho_F^\epsilon(x-y)dy + \int_{\|x-y\| \le \epsilon^{3/4}} \rho_F^\epsilon(x-y)K\epsilon^{1/4}dy \\
&\le 2MI(\epsilon) + K\epsilon^{1/4},
\end{aligned}$$

where we used the estimates of Lemma 3.10 and Lemma 3.11. This proves (3.26). For (3.27), let L_M be the (uniform) Lipschitz constant of F on $[0, M]$, with M still taken from Lemma 3.10. Then,

$$\begin{aligned}
\|F(\varphi_t(\cdot)) - F(\rho_F^\epsilon * \varphi_t(\cdot))\|_\infty &\le L_M \|\varphi_t(\cdot)) - (\rho_F^\epsilon * \varphi_t(\cdot))\|_\infty \\
&\le L_M(2MI(\epsilon) + K\epsilon^{1/4}),
\end{aligned}$$

which proves (3.27). $\square$

Proof [of Proposition 2.15] Let ϵ be small enough that Lemma 3.12 holds. We use the notation $\widehat{\delta}(\epsilon)$ for the quantity on the right hand side of (3.27). Then from the

representation (3.24) and Lemma 3.12 we can write,

$$
\begin{aligned}
&|\varphi_t(x) - \varphi_t^\epsilon(x)| \\
&\leq \mathbb{E}_x \left[\int_0^t \left| \varphi_s(Z_s) F(\rho_F^\epsilon * \varphi_s(Z_s)) - \varphi_s^\epsilon(Z_s) F\Big(\rho_F^\epsilon * \varphi_s^\epsilon(Z_s)\Big) \right| ds \right] + Mt\widehat{\delta}(\epsilon) \\
&\leq \mathbb{E}_x \left[\int_0^t \left| F(\rho_F^\epsilon * \varphi_s^\epsilon(Z_s)) \right| \cdot \left| \varphi_s^\epsilon(Z_s) - \varphi_s(Z_s) \right| ds \right] \\
&\qquad + \mathbb{E}_x \left[\int_0^t |\varphi_s(Z_s)| \left| F(\rho_F^\epsilon * \varphi_s^\epsilon(Z_s)) - F(\rho_F^\epsilon * \varphi_s(Z_s)) \right| ds \right] + Mt\widehat{\delta}(\epsilon) \\
&\leq \|F\|_M \int_0^t \|\varphi_s^\epsilon(\cdot) - \varphi_s(\cdot)\|_\infty ds + ML_M \int_0^t \|\rho_F^\epsilon * \varphi_s^\epsilon(\cdot) - \rho_F^\epsilon * \varphi_s(\cdot)\|_\infty ds + Mt\widehat{\delta}(\epsilon) \\
&\leq (\|F\|_M + ML_M) \int_0^t \|\varphi_s^\epsilon(\cdot) - \varphi_s(\cdot)\|_\infty ds + Mt\widehat{\delta}(\epsilon),
\end{aligned}
$$

where the second inequality is the triangle inequality, and the third is Lemma 3.10. An application of Gronwall's inequality then yields,

$$
\begin{aligned}
\|\varphi_t^\epsilon(\cdot) - \varphi_t(\cdot)\|_\infty &\leq Mt\widehat{\delta}(\epsilon) \exp(t(\|F\|_M + ML_M)) \\
&\leq MT\widehat{\delta}(\epsilon) \exp(T(\|F\|_M + ML_M)),
\end{aligned}
$$

giving the result, since $\widehat{\delta} \to 0$ as $\epsilon \to 0$. □

3.3.2 Porous Medium Equation

In this subsection we prove Proposition 2.19. To ease notation, we present the proof in $d = 1$ (although we retain the notation ∇). However, to recall the dependence on ϵ we write ρ^ϵ for ρ_γ. It should be clear that it extends almost without change to higher dimensions.

Recall that we are concerned with non-negative solutions to the equation (2.18):

$$
\partial_t \psi_t^\epsilon(x) = \Delta \left(\psi_t^\epsilon \, \rho^\epsilon * \psi_t^\epsilon \right)(x) + \psi_t^\epsilon(x) \left(1 - \rho^\epsilon * \psi_t^\epsilon(x)\right),
$$

and we assume that $\rho = \zeta * \check{\zeta}$ with ζ a rapidly decreasing function and $\check{\zeta}(x) = \zeta(-x)$. The example we have in mind is ζ (and therefore ρ) being the density of a mean zero Gaussian random variable. We shall prove that under the assumptions of Proposition 2.19, as $\epsilon \to 0$, we have convergence to the solution to the porous medium equation with logistic growth, equation (2.2):

$$
\partial_t \psi_t(x) = \Delta \left(\psi_t^2 \right)(x) + \psi_t(x) \left(1 - \psi_t(x)\right).
$$

We work on the time interval $[0,T]$. We will require a lower bound on $\int \psi_t^\epsilon(x)\log\psi_t^\epsilon(x)dx$ which we record as a lemma.

Lemma 3.13 *Suppose that there exists $\lambda \in (0,1)$ and $C<\infty$, both independent of ϵ, such that $\int \exp(\lambda|x|)\psi_0^\epsilon(x)dx < C$. Then there exists a constant $K<\infty$, independent of ϵ, such that $\int \psi_t^\epsilon(x)\log\psi_t^\epsilon(x)dx > -K$ for all $t\in[0,T]$.*

Proof First observe that, since $x\log x$ is bounded below, $\int_{-1}^1 \psi_t^\epsilon(x)\log\psi_t^\epsilon(x)dx$ is bounded below, and recall that $\psi_t^\epsilon(x)\geq 0$.

Now consider

$$\frac{d}{dt}\int \exp(\lambda x)\psi_t^\epsilon(x)dx = \int \exp(\lambda x)\Delta\Big(\psi_t^\epsilon\,\rho^\epsilon{*}\psi_t^\epsilon\Big)(x)dx + \int \exp(\lambda x)\psi_t^\epsilon(x)\Big(1-\rho^\epsilon{*}\psi_t^\epsilon(x)\Big)dx$$
$$= \int(\lambda^2-1)\exp(\lambda x)\psi_t^\epsilon(x)\rho_\epsilon{*}\psi_t^\epsilon(x)dx + \int \exp(\lambda x)\psi_t^\epsilon(x)dx \leq \int \exp(\lambda x)\psi_t^\epsilon(x)dx, \quad (3.28)$$

and so, by Gronwall's inequality, $\int \exp(\lambda x)\psi_t^\epsilon(x)dx$ is uniformly bounded on $[0,T]$. In particular, combining with the Mean Value Theorem, we find

$$\int_x^{x+1} \psi_t^\epsilon(y)dy \leq C\exp(-\lambda x),$$

where the constant C is independent of $x\geq 1$. A fortiori,

$$\int_x^{x+1} \psi_t^\epsilon(y)\mathbf{1}_{\psi_t^\epsilon(y)\leq 1}dy \leq C\exp(-\lambda x). \quad (3.29)$$

Now the function $\psi \mapsto \mathbf{1}_{0\leq\psi\leq 1}\psi|\log\psi|$ is concave, and so using Jensen's inequality and (3.29),

$$\int_x^{x+1} \psi_t^\epsilon(y)|\log\psi_t^\epsilon(y)|\mathbf{1}_{\psi_t^\epsilon(y)\leq 1}dy \leq C'x\exp(-\lambda x).$$

Evidently a symmetric argument applies for $x\leq -1$. Summing over x, and using that $\psi\log\psi \geq -\psi|\log\psi|\mathbf{1}_{\psi\leq 1}$, we find

$$\int \psi_t^\epsilon(x)\log\psi_t^\epsilon(x)dx \geq -C''\sum_{x=1}^\infty x\exp(-\lambda x) > -K > -\infty,$$

as required. □

Proof [Proof of Proposition 2.19] Note that

$$\int \psi_t^\epsilon(x)\, \rho^\epsilon * \psi_t^\epsilon(x)dx = \int\int\int \psi_t^\epsilon(x)\psi_t^\epsilon(x-y)\zeta^\epsilon(y-z)\check{\zeta}^\epsilon(z)dzdydx$$
$$= \int\int\int \psi_t^\epsilon(\widetilde{x}-\widetilde{z})\psi_t^\epsilon(\widetilde{x}-\widetilde{y})\zeta^\epsilon(\widetilde{y})\zeta^\epsilon(\widetilde{z})d\widetilde{z}d\widetilde{y}d\widetilde{x} = \int \left(\zeta^\epsilon * \psi_t^\epsilon(x)\right)^2 dx,$$

where we have set $\widetilde{x} = x - z$, $\widetilde{y} = y - z$, $\widetilde{z} = -z$. Now note that

$$\frac{d}{dt}\int \psi_t^\epsilon(x)dx = \int \Delta\Big(\psi_t^\epsilon\, \rho^\epsilon * \psi_t^\epsilon\Big)(x)dx + \int \psi_t^\epsilon(x)\Big(1 - \rho^\epsilon * \psi_t^\epsilon(x)\Big)dx$$
$$= \int \psi_t^\epsilon(x)dx - \int \Big(\zeta^\epsilon * \psi_t^\epsilon(x)\Big)^2 dx.$$

Thus, Gronwall's inequality implies that $\int \psi_t^\epsilon(x)dx$ is uniformly bounded above in ϵ and $t \in [0,T]$. Note that this also then gives a uniform bound on the rate of change of $\int \psi_t^\epsilon(x)dx$, and since we are working on $[0,T]$ this will be enough to give continuity in time of the L^1 norm of the limit when we pass to a convergent subsequence. Now consider

$$\begin{aligned}
\frac{d}{dt}\int \psi_t^\epsilon \log \psi_t^\epsilon dx &= \int (1+\log\psi_t^\epsilon)\left[\Delta\Big(\psi_t^\epsilon\, \rho^\epsilon * \psi_t^\epsilon\Big) + \psi_t^\epsilon\Big(1 - \rho^\epsilon * \psi_t^\epsilon\Big)\right] dx \\
&= \int (1+\log\psi_t^\epsilon) \times \qquad (3.30)\\
&\qquad \left[\nabla\Big(\psi_t^\epsilon\, \nabla(\rho^\epsilon * \psi_t^\epsilon) + \nabla\psi_t^\epsilon\, \rho^\epsilon * \psi_t^\epsilon\Big) + \psi_t^\epsilon\Big(1 - \rho^\epsilon * \psi_t^\epsilon\Big)\right] dx \\
&= \int \Bigg[- \frac{\nabla\psi_t^\epsilon}{\psi_t^\epsilon}\Big(\psi_t^\epsilon\, \nabla(\rho^\epsilon * \psi_t^\epsilon) + \nabla\psi_t^\epsilon\, \rho^\epsilon * \psi_t^\epsilon\Big) \qquad (3.31)\\
&\qquad +(1+\log\psi_t^\epsilon)\psi_t^\epsilon(1 - \rho^\epsilon * \psi_t^\epsilon)\Bigg] dx \\
&= -\int \left(\nabla(\zeta^\epsilon * \psi_t^\epsilon)\right)^2 dx - \int (\nabla\psi_t^\epsilon)^2 \frac{\rho^\epsilon * \psi_t^\epsilon}{\psi_t^\epsilon} dx \\
&\quad + \int \left[\psi_t^\epsilon + \psi_t^\epsilon \log\psi_t^\epsilon\Big(1 - \rho^\epsilon * \psi_t^\epsilon\Big) - \psi_t^\epsilon\, \rho^\epsilon * \psi_t^\epsilon\right] dx \\
&= -\int \left(\nabla(\zeta^\epsilon * \psi_t^\epsilon)\right)^2 dx - \int (\nabla\psi_t^\epsilon)^2 \frac{\rho^\epsilon * \psi_t^\epsilon}{\psi_t^\epsilon} dx - \int (\zeta^\epsilon * \psi_t^\epsilon)^2 dx \\
&\quad + \int \left[\psi_t^\epsilon + \psi_t^\epsilon \log\psi_t^\epsilon\Big(1 - \rho^\epsilon * \psi_t^\epsilon\Big)\right] dx. \qquad (3.32)
\end{aligned}$$

The first three terms are negative; and we already saw that the L^1 norm of ψ_t^ϵ is uniformly bounded. Moreover, since $\psi_t^\epsilon \log\psi_t^\epsilon$ is uniformly bounded below and $\int \rho^\epsilon(x)dx = 1$,

$$-\int \psi_t^\epsilon \log\psi_t^\epsilon\, \rho^\epsilon * \psi_t^\epsilon dx \le C\int \rho^\epsilon * \psi_t^\epsilon dx = C\int \psi_t^\epsilon dx.$$

From this and (3.32), we see immediately that $\int \psi_t^\epsilon \log \psi_t^\epsilon dx$ is uniformly bounded above in ϵ and $t \in [0,T]$. Combining with Lemma 3.13, we deduce that we have a uniform bound on $\int \psi_s^\epsilon(x)|\log \psi_s^\epsilon(x)|dx$. From (3.32), this in turn means that both $\int_0^t \int (\zeta^\epsilon * \psi_s^\epsilon(x))^2 dxds$ and $\int_0^t \int \left(\nabla(\zeta^\epsilon * \psi_s^\epsilon(x))\right)^2 dxds$ are uniformly bounded in ϵ and $t \in [0,T]$.

We shall next show that $\zeta^\epsilon * \psi_t^\epsilon$ solves (2.2) up to a remainder of order ϵ. To simplify derivations, we now subsume the integral over time, and control our integrals over space as if we have point-wise estimates. In reality, readers should understand the following inequalities as bounds on the functions integrated over time. See Remark 3.14 for further discussions.

Note that

$$\int \Delta\left((\rho^\epsilon * \psi_t^\epsilon)\psi_t^\epsilon\right)\phi dx = -\int \nabla(\rho^\epsilon * \psi_t^\epsilon)\,\psi_t^\epsilon\,\nabla\phi dx - \int \rho^\epsilon * \psi_t^\epsilon\,\nabla\psi_t^\epsilon\,\nabla\phi dx. \tag{3.33}$$

We would like to show that this is close to $\int(\zeta^\epsilon * \psi_t^\epsilon)^2\Delta\phi dx$. For the first term

$$\begin{aligned}
\int \nabla(\rho^\epsilon * \psi_t^\epsilon)\,\psi_t^\epsilon\,\nabla\phi dx &= \int\int\int \nabla\psi_t^\epsilon(x-y)\zeta^\epsilon(y-z)\check{\zeta}^\epsilon(z)\psi_t^\epsilon(x)\nabla\phi(x)dzdydx \\
&= \int\int\int \nabla\psi_t^\epsilon(\widetilde{x}-\widetilde{y})\zeta^\epsilon(\widetilde{y})\zeta^\epsilon(\widetilde{z})\psi_t^\epsilon(\widetilde{x}-\widetilde{z})\nabla\phi(\widetilde{x}-\widetilde{z})d\widetilde{z}d\widetilde{y}d\widetilde{x} \\
&= \int (\nabla\zeta^\epsilon * \psi_t^\epsilon)\;(\zeta^\epsilon * (\psi_t^\epsilon\nabla\phi))\,dx \\
&= \frac{1}{2}\int \nabla((\zeta^\epsilon * \psi_t^\epsilon)^2)\,\nabla\phi dx \\
&\quad + \int \nabla(\zeta^\epsilon * \psi_t^\epsilon)\,[\zeta^\epsilon * (\psi_t^\epsilon\,\nabla\phi) - \nabla\phi\,(\zeta^\epsilon * \psi_t^\epsilon)]\,dx.
\end{aligned} \tag{3.34}$$

Here we applied the same substitution for $\widetilde{x}, \widetilde{y}$ and $\widetilde{z}$. By the Intermediate Value Theorem,

$$\left|\int [\psi_t^\epsilon(x-y)\nabla\phi(x-y)\zeta^\epsilon(y) - \nabla\phi(x)\psi_t^\epsilon(x-y)\zeta^\epsilon(y)]\,dy\right| \le C\|\Delta\phi\|_\infty \int \psi_t^\epsilon(x-y)|y|\zeta^\epsilon(y)dy.$$

Since $\zeta \in \mathcal{S}(\mathbb{R})$, the integral in (3.34) is $\mathcal{O}(\epsilon)$ independent of x.

Similarly, the integral $\int |\zeta^\epsilon * (\psi_t^\epsilon\,\nabla\phi)(x) - \nabla\phi\,(\zeta^\epsilon * \psi_t^\epsilon)(x)|\,dx$ is bounded above by

$$\begin{aligned}
&\left|\int\int [\psi_t^\epsilon(x-y)\nabla\phi(x-y)\zeta^\epsilon(y) - \nabla\phi(x)\psi_t^\epsilon(x-y)\zeta^\epsilon(y)]\,dydx\right| \\
&\qquad\qquad \le C\|\Delta\phi\|_\infty \int \left[\int \psi_t^\epsilon(x-y)dx\right]|y|\zeta^\epsilon(y)dy,
\end{aligned}$$

which is again $\mathcal{O}(\epsilon)$. In other words, the function $||\zeta^\epsilon * (\psi_t^\epsilon\,\nabla\phi) - \nabla\phi\,(\zeta^\epsilon * \psi_t^\epsilon)||_{L^2} = \mathcal{O}(\epsilon)$.

To bound the last term in Equation (3.34), we apply the Cauchy-Schwarz inequality:

$$\int |\nabla(\zeta^\epsilon * \psi_t^\epsilon)\,[\zeta^\epsilon * (\psi_t^\epsilon\,\nabla\phi) - \nabla\phi\,(\zeta^\epsilon * \psi_t^\epsilon)]|\,dx \\ \leq ||\nabla(\zeta^\epsilon * \psi_t^\epsilon(x))||_{L^2}||\zeta^\epsilon * (\psi_t^\epsilon\,\nabla\phi) - \nabla\phi\,(\zeta^\epsilon * \psi_t^\epsilon)||_{L^2} = \mathcal{O}(\epsilon). \quad (3.35)$$

Similarly, for the second term in (3.33),

$$\begin{aligned}\int (\rho^\epsilon * \psi_t^\epsilon)\,\nabla\psi_t^\epsilon\,\nabla\phi dx &= \int\int\int \psi_t^\epsilon(x-y)\zeta^\epsilon(y-z)\check{\zeta}^\epsilon(z)\nabla\psi_t^\epsilon(x)\nabla\phi(x)dzdydx \\ &= \int\int\int \psi_t^\epsilon(\widetilde{x}-\widetilde{y})\zeta^\epsilon(\widetilde{y})\zeta^\epsilon(\widetilde{z})\nabla\psi_t^\epsilon(\widetilde{x}-\widetilde{z})\nabla\phi(\widetilde{x}-\widetilde{z})d\widetilde{z}d\widetilde{y}d\widetilde{x} \\ &= \int (\zeta^\epsilon * \psi_t^\epsilon)\ (\zeta^\epsilon * (\nabla\psi_t^\epsilon\,\nabla\phi))\,dx \\ &= \frac{1}{2}\int \nabla((\zeta^\epsilon * \psi_t^\epsilon)^2)\,\nabla\phi dx \\ &\quad + \int (\zeta^\epsilon * \psi_t^\epsilon)\,[\zeta^\epsilon * (\nabla\psi_t^\epsilon\,\nabla\phi) - \nabla\phi\,(\nabla\zeta^\epsilon * \psi_t^\epsilon)]\,dx, \qquad (3.36)\end{aligned}$$

which is controlled in the same way as (3.34): using the Intermediate Value Theorem,

$$\left|\int [\nabla\psi_t^\epsilon(x-y)\nabla\phi(x-y)\zeta^\epsilon(y) - \nabla\phi(x)\nabla\psi_t^\epsilon(x-y)\zeta^\epsilon(y)]\,dy\right| \\ \leq C\|\Delta\phi\|_\infty \left|\int \nabla\psi_t^\epsilon(x-y)y\zeta^\epsilon(y)dy\right|, \quad (3.37)$$

which again is $\mathcal{O}(\epsilon)$. With the same Cauchy-Schwarz argument, we can give a bound on the last term of Equation (3.36). To conclude, the calculations above yield both a uniform (in ϵ) bound on $\zeta^\epsilon * \psi_t^\epsilon$ in $L^1 \cap L^2\Big([0,T]\times\mathbb{R}\Big)$, and that

$$\int \psi_t^\epsilon(x)\phi(x)dx - \int \psi_0^\epsilon(x)\phi(x)dx = \int_0^t \int (\zeta^\epsilon * \psi_s^\epsilon(x))^2\Delta\phi(x)dxds \\ + \int_0^t \int \zeta^\epsilon * \psi_s^\epsilon(x)\,(1 - \zeta^\epsilon * \psi_s^\epsilon(x))\,\phi(x)dxds + \mathcal{O}(\epsilon) \quad (3.38)$$

(for sufficiently regular ϕ). Since $\int \psi_t^\epsilon(x)\phi(x)dx - \int \zeta^\epsilon * \psi_t^\epsilon(x)\phi(x)dx$ is order ϵ, if we replace ψ^ϵ by $\zeta^\epsilon * \psi^\epsilon$ on the left hand side, then (3.38) says that $\zeta^\epsilon * \psi^\epsilon$ solves (2.19) weakly up to order ϵ. Therefore, $\zeta^\epsilon * \psi^\epsilon$ converges weakly to ψ in L^1, where ψ is the (unique) solution to equation (2.19) and, so, therefore, does ψ^ϵ. □

Remark 3.14 *We have only shown bounds and controls of our functions over integrals over time, not pointwise at every time spot. Therefore, we have really shown convergence among PDE solutions that are weak in both time and space.*

4

Branching Particle Systems with Local Regulation: Simultaneous Convergence Proof

4.1 Overview

In this chapter we prove Theorem 2.21, which proves convergence in the case that the width of the interaction kernel ρ_F simultaneously scales along with the parameters θ and N, in the special case in which $r \equiv 1 \equiv \gamma$, $q_\theta(x, dy)$ is isotropic with zero mean, the kernel ρ_F is Gaussian, and the scaling limit is a reaction-diffusion equation.

To simplify notation, in this section we shall write

$$\rho_\epsilon * \eta(x) = \rho_F^\epsilon * \eta(x) = \langle p_{\epsilon^2}(x, y), \eta(dy) \rangle,$$

where $p_t(x, y)$ denotes the heat semigroup. The assumptions of Theorem 2.21 will be in force throughout, in particular,

$$\epsilon^2 \theta \to \infty, \qquad \text{and } \frac{\theta}{N\epsilon^d} \to 0. \tag{4.1}$$

That $N, \theta \to \infty$ and $\epsilon \to 0$ simultaneously will be implicit, so for example if we write $\lim_{\epsilon \to 0}$, it should be understood that $\theta, N \to \infty$ in such a way that (4.1) is satisfied. Moreover, where there is no risk of confusion, except where it is helpful for emphasis, we suppress dependence of η on N.

The first part of the proof mirrors that of Theorem 2.10: in Subsection 4.2 we establish bounds on the moments of $\rho_\epsilon * \eta_t(x)$ that are sufficient to imply tightness and then apply standard results on convergence of Markov processes from [EK86]. The challenge comes in identifying the limit points. This is much more intricate than the case in which we do not scale the interaction kernel, as weak convergence will no longer be sufficient to guarantee the form of the nonlinear terms in the limiting equation. Identification of the limit will rest on regularity inherited from continuity estimates for a random walk with Gaussian jumps which we prove in Subsection 4.3, before identifying the limit points in Subsection 4.4.

4.2 Moment bounds for $\rho_\epsilon * \eta$

Let us write $\mathcal{L}^\theta f(x) := \theta \int (f(y) - f(x)) q_\theta(x, y) dy$ where q_θ is a Gaussian kernel of mean 0 and variance $1/\theta$. We note that $\mathcal{L}^\theta$ is the generator of a continuous (time and space) random walk, which makes jumps of mean 0 and variance $1/\theta$ at rate θ. In what follows we write $\psi_t^{\epsilon,x}(y)$ for the solution of

$$\partial_t \psi_t^{\epsilon,x} = \mathcal{L}^\theta \psi_t^{\epsilon,x}, \tag{4.2}$$

with initial condition $\psi_0^{\epsilon,x}(y) = \rho_\epsilon(y - x) = p_{\epsilon^2}(x, y)$.

To see why $\psi_t^{\epsilon,x}$ is useful, first note that for any time-dependent function $\phi_t(x)$ with time derivative $\dot{\phi}_t(x) = \partial_t \phi_t(x)$,

$$\langle \phi_t(x), \eta_t(dx) \rangle = \langle \phi_0(x), \eta_0(dx) \rangle + M_t(\phi) + \int_0^t \Big\langle \mathcal{L}^\theta \phi_s(x) + \dot{\phi}_s(x), \eta_s(dx) \Big\rangle ds + \int_0^t \Big\langle \phi_s(x) F(x, \eta_s), \eta_s(dx) \Big\rangle ds, \tag{4.3}$$

where $M_t(\phi)$ is a martingale (with respect to the natural filtration) with angle bracket process given by (2.8) with f replaced by $\phi_s(\cdot)$. So, taking $\phi_s(\cdot) = \psi_{t-s}^{\epsilon,x}(\cdot)$ for $0 \le s \le t$,

$$\begin{aligned} \rho_\epsilon * \eta_t(x) &= \langle \psi_0^{\epsilon,x}(y), \eta_t(dy) \rangle \\ &= \langle \psi_t^{\epsilon,x}(y), \eta_0(dy) \rangle + \int_0^t \Big\langle \psi_{t-s}^{\epsilon,x}(y) F\Big(\rho_\epsilon * \eta_s(y)\Big), \eta_s(dy) \Big\rangle ds + M_t(x), \end{aligned} \tag{4.4}$$

where $M_t(x)$ has mean zero and a second moment we can easily write down.

Lemma 4.1 *Fix $t > 0$, let $(\Pi(s))_{s\geq 0}$ be a rate one Poisson process, and let $T(t) = \Pi(\theta t)/\theta$. Then*

$$\psi_t^{\epsilon,x}(y) = \mathbb{E}\left[p_{\epsilon^2+T(t)}(x,y)\right],$$

and, moreover, since under our assumptions $\theta\epsilon^2$ is bounded below, there is a C independent of ϵ or t such that

$$\|\psi_t^{\epsilon,x}\|_\infty \leq \frac{C}{(\epsilon^2+t)^{d/2}}.$$

Proof The first claim is immediate from the definition of the random walk with generator $\mathcal{L}^\theta$.

For the second claim, first define $\tau(t) = T(t) - t$. Since if $\tau(t) \geq -(\epsilon^2+t)/2$, then $1/(\epsilon^2+T(t)) \leq 2/(\epsilon^2+t)$, while $\epsilon^2 + T(t) \geq \epsilon^2$ always, partitioning over $\{\tau(t) \geq -(\epsilon^2+t)/2\}$ and its complement,

$$\begin{aligned}\|\psi^{\epsilon,x}\|_\infty &= \mathbb{E}\left[\frac{1}{\Big(2\pi(\epsilon^2+T(t))\Big)^{d/2}}\right] \\ &\leq \frac{C}{(\epsilon^2+t)^{d/2}} + \frac{C}{\epsilon^d}\mathbb{P}\left\{\tau(t) < -(\epsilon^2+t)/2\right\}. \end{aligned} \tag{4.5}$$

Now, observe that since $\mathbb{E}[e^{-\Pi(\theta t)}] = \exp(-\theta t(1-e^{-1}))$, by Markov's inequality,

$$\begin{aligned}\mathbb{P}\left\{\tau(t) < -\frac{\epsilon^2+t}{2}\right\} &= \mathbb{P}\left\{e^{-\Pi(\theta t)} > e^{-\theta(t-\epsilon^2)/2}\right\} \\ &\leq \frac{\mathbb{E}[\exp\Big(-\Pi(\theta t)\Big)]}{\exp\Big(-\theta(t-\epsilon^2)/2\Big)} \\ &= \frac{\exp(-\theta t(1-e^{-1}))}{\exp(-\theta(t-\epsilon^2)/2)} \\ &= \exp\left\{-\chi\theta t - \frac{\theta\epsilon^2}{2}\right\}, \end{aligned} \tag{4.6}$$

where $\chi = 1/2 - e^{-1} > 0$. The second term in (4.5) is therefore bounded by

$$C\left(1+\frac{t}{\epsilon^2}\right)^{d/2} e^{-\chi\theta t}\frac{1}{(\epsilon^2+t)^{d/2}}e^{-\epsilon^2\theta/2}.$$

Now observe that the derivative (with respect to t) of $e^{-\chi\theta t}(1+t/\epsilon^2)^{d/2}$ is

$$\left(\frac{d}{2\epsilon^2} - \left(1+\frac{t}{\epsilon^2}\right)\chi\theta\right)\left(1+\frac{t}{\epsilon^2}\right)^{d/2-1} e^{-\chi\theta t},$$

which is negative if $\theta(\epsilon^2 + t) > d/2\chi$. At the maximum, $(1 + t/\epsilon^2) = d/(2\chi\theta\epsilon^2)$, and so this quantity is bounded uniformly over not only t but also ϵ (since we've assumed that $\theta\epsilon^2$ is bounded below). Therefore, we have the bound

$$\frac{1}{\epsilon^d}\mathbb{P}\left\{\tau(t) < -(\epsilon^2 + t)/2\right\} \leq \frac{C}{(\epsilon^2 + t)^{d/2}} e^{-\epsilon^2\theta/2}. \tag{4.7}$$

Substituting this into (4.5) yields the result. □

Lemma 4.2 *Let $\{\mathcal{F}_t\}_{t\geq 0}$ denote the natural filtration. Under the assumptions of Theorem 2.21, for each $T \in [0, \infty)$, and $k \in \mathbb{N}$, there exist constants $C = C(k, T)$ and $\widetilde{C} = \widetilde{C}(k, T)$, independent of ϵ, such that for all $x \in \mathbb{R}^d$ and all $u, t \in [0, T]$ with $u < t$,*

$$\mathbb{E}\Big[\Big(\rho_\epsilon * \eta_t(x)\Big)^k \Big| \mathcal{F}_u\Big] \leq C\langle \psi^{\epsilon,x}_{t-u}(z), \eta_u(dz)\rangle^k + C\frac{\theta}{N\epsilon^d}\langle \psi^{\epsilon,x}_{t-u}(z), \eta_u(dz)\rangle; \tag{4.8}$$

and

$$\mathbb{E}\Big[\int_u^t \langle \psi^{\epsilon,x}_{t-s}(z), \eta_s(dz)\rangle^{k-1}\Big\langle \psi^{\epsilon,x}_{t-s}(z)|F(\rho_\epsilon * \eta_s(z))|, \eta_s(dz)\Big\rangle ds \Big| \mathcal{F}_u\Big] \\ \leq \widetilde{C}\langle \psi^{\epsilon,x}_{t-u}(z), \eta_u(dz)\rangle^k + \widetilde{C}\frac{\theta}{N\epsilon^d}\langle \psi^{\epsilon,x}_{t-u}(z), \eta_u(dz)\rangle; \tag{4.9}$$

where the function $\psi^{\epsilon,x}_t(\cdot)$ was defined in (4.2). In particular, under the assumptions of Theorem 2.21, the expected values of the quantities on the right hand side of (4.8) and (4.9) are both integrable with respect to Lebesgue measure.

Proof To simplify our expressions, we shall consider the case $u = 0$, but the proof goes through unchanged for other values of u.

We proceed by induction. Taking expectations in (4.4), using that F is bounded above, and applying Gronwall's inequality to $\langle \psi^{\epsilon,x}_{t-s}, \eta_s\rangle$ we obtain $\mathbb{E}[\langle \psi^{\epsilon,x}_0, \eta_t\rangle] \leq C\mathbb{E}[\langle \psi^{\epsilon,x}_t, \eta_0\rangle]$, which implies (4.8) in the case $k = 1$. Moreover, rearranging (4.4) we find

$$-\int_0^t \Big\langle \psi^{\epsilon,x}_{t-s}(y)F(\rho_\epsilon * \eta_s(y), \eta_s(dy)\Big\rangle ds = \langle \psi^{\epsilon,x}_t(y), \eta_0(dy)\rangle - \langle \psi^{\epsilon,x}_0(y), \eta_t(dy)\rangle + M_t(x), \tag{4.10}$$

and taking expectations again, since $\langle \psi^{\epsilon,x}_0, \eta_t\rangle > 0$, and $M_0(x) = 0$, this yields

$$\mathbb{E}\Big[-\int_0^t \langle \psi^{\epsilon,x}_{t-s}(y)F(\rho_\epsilon * \eta_s(y)), \eta_s(dy)\rangle ds \Big| \mathcal{F}_0\Big] \leq \langle \psi^{\epsilon,x}_t(y), \eta_0(dy)\rangle.$$

Since F is bounded above, there exists a constant K such that $|F| \leq K - F$ and so combined with the bound on $\mathbb{E}[\langle \psi_0^{\epsilon,x}(y), \eta_t(dy)\rangle]$ just obtained, this in turn yields

$$\mathbb{E}\Big[\int_0^t \Big\langle \psi_{t-s}^{\epsilon,x}(y)|F(\rho_\epsilon * \eta_s(y))|, \eta_s(dy)\Big\rangle ds\Big|\mathcal{F}_0\Big] \leq \tilde{C}\langle \psi_t^{\epsilon,x}(y), \eta_0(dy)\rangle,$$

which is (4.9) in the case $k = 1$.

Now suppose that we have established (4.8) and (4.9) for all exponents $j < k$. First we apply the generator $\mathcal{P}^N$ of our scaled population process to functions of the form $\langle f, \eta\rangle^k$. Recalling that each jump of the process involves the birth or death of a single individual, and so increments $\langle f, \eta\rangle$ by $\pm f/N$ at the location of that individual and that $r \equiv \gamma \equiv 1$, we find

$$\begin{aligned}\mathcal{P}^N\Big(\langle f, \eta\rangle^k\Big) &= \Big\langle \int \theta N \sum_{j=1}^k \binom{k}{j} \frac{f(y)^j}{N^j} \langle f, \eta\rangle^{k-j} q_\theta(x, dy), \eta(dx)\Big\rangle \\ &\quad + \Big\langle \theta N\Big(1 - \frac{F(\rho_\epsilon * \eta(x))}{\theta}\Big) \sum_{j=1}^k \binom{k}{j}(-1)^j \frac{f(x)^j}{N^j}\langle f, \eta\rangle^{k-j}, \eta(dx)\Big\rangle. \quad (4.11)\end{aligned}$$

Mimicking what we did above, we set $f(\cdot) = \psi_t^{\epsilon,x}(\cdot)$ and write

$$\begin{aligned}\mathbb{E}\Big[\langle \psi_0^{\epsilon,x}, \eta_t\rangle^k\Big|\mathcal{F}_0\Big] &= \langle \psi_t^{\epsilon,x}(y), \eta_0(dy)\rangle^k + \mathbb{E}\Big[\int_0^t \mathcal{P}^N\Big(\langle \psi_{t-s}^{\epsilon,x}(y), \eta_s(dy)\rangle^k\Big) ds \\ &\quad - \int_0^t \langle k\dot{\psi}_{t-s}^{\epsilon,x}(y), \eta_s(dy)\rangle \Big\langle \psi_{t-s}^{\epsilon,x}(y), \eta_s(dy)\Big\rangle^{k-1} ds\Big|\mathcal{F}_0\Big]. \quad (4.12)\end{aligned}$$

Since $\dot{\psi}_s^{\epsilon,x} = \mathcal{L}_\theta \psi_s^{\epsilon,x}$, the $j = 1$ terms from $\mathcal{P}^N(\langle \psi_{t-s}^{\epsilon,x}(y), \eta_s(dy)\rangle^k)$ combines with the last term in (4.12) to yield

$$\int_0^t k\langle \psi_{t-s}^{\epsilon,x}, \eta\rangle^{k-1}\langle F(\rho_\epsilon * \eta_s(y))\psi_{t-s}^{\epsilon,x}(y), \eta_s(dy)\rangle ds.$$

As for the remaining terms, using (from Lemma 4.1) that $\sup_s \|\psi_s^{\epsilon,x}(\cdot)\|_\infty = C/\epsilon^d$, $N\epsilon^d > 1$, and our inductive hypothesis, we find

$$\begin{aligned}&\mathbb{E}\Big[\Big\langle \int_0^t \theta N \sum_{j=2}^k \binom{k}{j} \int \frac{\psi_{t-s}^{\epsilon,x}(z)^j}{N^j}\langle \psi_{t-s}^{\epsilon,x}, \eta_s\rangle^{k-j} q_\theta(y, dz), \eta_s(dy)\Big\rangle ds \\ &\quad + \Big\langle \int_0^t \theta N \sum_{j=2}^k \binom{k}{j} \frac{\psi_{t-s}^{\epsilon,x}(y)^j}{N^j}\langle \psi_{t-s}^{\epsilon,x}, \eta_s\rangle^{k-j}(-1)^j\Big(1 - \frac{F(\rho_\epsilon * \eta_s(y))}{\theta}\Big), \eta_s(dy)\Big\rangle ds\Big|\mathcal{F}_0\Big] \\ &\leq C\mathbb{E}\Big[\Big\langle \int_0^t \sum_{j=2}^k \frac{\theta}{N\epsilon^d}\Big(\frac{1}{(N\epsilon^d)^{j-2}}\Big)\Big\langle \psi_{t-s}^{\epsilon,x}(y)\Big(2 + \frac{|F(\rho_\epsilon * \eta_s(y))|}{\theta}\Big), \eta_s(dy)\Big\rangle \langle \psi_{t-s}^{\epsilon,x}, \eta_s\rangle^{k-j} ds\Big|\mathcal{F}_0\Big] \\ &\leq C'\frac{\theta}{N\epsilon^d}\sum_{j=1}^{k-1}\langle \psi_t^{\epsilon,x}(y), \eta_0(dy)\rangle^j \leq C''\frac{\theta}{N\epsilon^d}\Big(\langle \psi_t^{\epsilon,x}(y), \eta_0(dy)\rangle^k + \langle \psi_t^{\epsilon,x}(y), \eta_0(dy)\rangle\Big).\end{aligned}$$

Combining this with (4.11) and (4.12), using once again the fact that F is bounded above, we find

$$\mathbb{E}\Big[\langle\psi_0^{\epsilon,x},\eta_t\rangle^k\Big|\mathcal{F}_0\Big] \le \langle\psi_t^{\epsilon,x}(y),\eta_0(dy)\rangle^k + \widetilde{C}\mathbb{E}\Big[\int_0^t \langle\psi_{t-s}^{\epsilon,x}(y),\eta_s(dy)\rangle^k ds\Big|\mathcal{F}_0\Big] + C''\frac{\theta}{N\epsilon^d}\Big(\langle\psi_t^{\epsilon,x}(y),\eta_0(dy)\rangle^k + \langle\psi_t^{\epsilon,x}(y),\eta_0(dy)\rangle\Big),$$

and (4.8) follows from Gronwall's inequality. Rearranging exactly as in the case $k = 1$, we recover (4.9) and the inductive step is complete. □

We shall also need the following consequence of the bounds that we obtained in Lemma 4.2:

Corollary 4.3 *Under the assumptions of Theorem 2.21, for each $k \ge 1$, $T > 0$, there is a $C(k,T)$ such that*

$$\mathbb{E}\Big[\Big\langle(\rho_\epsilon * \eta_t)^k,\eta_t\Big\rangle\Big] < C(k,T) < \infty, \qquad \textit{for all } t \in [0,T]. \tag{4.13}$$

Proof [Sketch] First observe that if $A \in (0,1)$, then

$$p_{A\epsilon^2}(x,y) = \frac{1}{A^{d/2}}p_{\epsilon^2}(x,y)\exp\Big(-\frac{\|x-y\|^2}{2\epsilon^2}\Big(\frac{1}{A}-1\Big)\Big) \le \frac{1}{A^{d/2}}p_{\epsilon^2}(x,y). \tag{4.14}$$

Now consider

$$\begin{aligned}\mathbb{E}\Big[\langle\rho_\epsilon * \eta_t(x),\eta_t(dx)\rangle\Big] &= \mathbb{E}\Big[\int\int p_{\epsilon^2}(x,z)\eta_t(dz)\eta_t(dx)\Big]\\ &= \mathbb{E}\Big[\int\int\int p_{\epsilon^2/2}(x,y)p_{\epsilon^2/2}(y,z)dy\eta_t(dz)\eta_t(dx)\Big]\\ &= \mathbb{E}\Big[\int\Big(p_{\epsilon^2/2} * \eta_t(y)\Big)^2 dy\Big]\\ &\le C\int\mathbb{E}\Big[\Big(\rho_\epsilon * \eta_t(x)\Big)^2\Big]dx,\end{aligned}$$

where we used (4.14) in the last line. Using Lemma 4.2 and our assumptions on η_0, this quantity is finite.

To illustrate the inductive step, now consider

$$\begin{aligned}&\mathbb{E}\Big[\langle\rho_\epsilon * \eta_t(x)^2,\eta_t(dx)\rangle\Big] = \mathbb{E}\Big[\int\int\int p_{\epsilon^2}(x,z_1)p_{\epsilon^2}(x,z_2)\eta_t(dz_1)\eta_t(dz_2)\eta_t(dx)\Big]\\ &= \mathbb{E}\Big[\int\cdots\int p_{\epsilon^2/2}(x,y_1)p_{\epsilon^2/2}(x,y_2)p_{\epsilon^2/2}(y_1,z_1)p_{\epsilon^2/2}(y_2,z_2)\eta_t(dz_1)\eta_t(dz_2)dy_1dy_2\eta_t(dx)\Big].\end{aligned} \tag{4.15}$$

We use the identity

$$p_{\epsilon^2/2}(x,y_1)p_{\epsilon^2/2}(x,y_2) = p_{\epsilon^2}(y_1,y_2)p_{\epsilon^2/4}\Big(x,\frac{y_1+y_2}{2}\Big)$$

to rewrite (4.15) as

$$\mathbb{E}\Big[\int\int p_{\epsilon^2/2}*\eta_t(y_1)\,p_{\epsilon^2/2}*\eta_t(y_2)\,p_{\epsilon^2/4}*\eta_t\Big(\frac{y_1+y_2}{2}\Big)\,p_{\epsilon^2}(y_1,y_2)dy_1dy_2\Big]$$
$$\leq \mathbb{E}\Big[\int\int\Big\{\Big(p_{\epsilon^2/2}*\eta_t(y_1)\Big)^3+\Big(p_{\epsilon^2/2}*\eta_t(y_2)\Big)^3+\Big(p_{\epsilon^2/4}*\eta_t\Big(\frac{y_1+y_2}{2}\Big)\Big)^3\Big\}p_{\epsilon^2}(y_1,y_2)dy_1dy_2\Big],$$

where we have used that for any non-negative real numbers β_1, β_2, β_3, $\beta_1\beta_2\beta_3 \leq \beta_1^3+\beta_2^3+\beta_3^3$. For the first two terms in the sum we integrate with respect to y_2 and y_1 respectively to reduce to an expression of the form considered in Lemma 4.2. For the final term, the change of variables $z_1 = y_1+y_2$, $z_2 = y_1-y_2$ in the integral similarly allows us to integrate out the heat kernel, and we conclude that the result holds for $k=2$.

We can proceed in the same way for larger values of k, using repeatedly that

$$p_{t_1}(x,y_1)p_{t_2}(x,y_2) = p_{\frac{t_1t_2}{t_1+t_2}}\Big(x,\frac{t_2y_1+t_1y_2}{t_1+t_2}\Big)p_{t_1+t_2}(y_1,y_2)$$

to write

$$\prod_{j=1}^k p_\tau(y,y_j) = \prod_{j=2}^k p_{\frac{j\tau}{j-1}}\Big(y_j,Y_{j-1}\Big)p_{\frac{\tau}{k}}(y,Y_k)$$

where

$$Y_1 = y_1, \qquad Y_j = \frac{j-1}{j}Y_{j-1}+\frac{1}{j}y_j, \text{for } j\geq 2.$$

Writing $p_{\epsilon^2}(x,z_j) = \int p_{\epsilon^2/2}(x,y_j)p_{\epsilon^2/2}(y_j,z_j)dy_j$ and using the above with $\tau=\epsilon^2/2$, this yields

$$\Big\langle\Big(\rho_\epsilon*\eta_t(x)\Big)^k,\eta_t(dx)\Big\rangle = \int\cdots\int\prod_{j=2}^k p_{\epsilon^2 j/2(j-1)}(y_j,Y_{j-1})\prod_{i=1}^k p_{\epsilon^2/2}*\eta_t(y_i)p_{\epsilon^2/2k}*\eta_t(Y_k)dy_1\ldots dy_k$$
$$\leq\int\cdots\int\prod_{j=2}^k p_{\epsilon^2 j/2(j-1)}(y_j,Y_{j-1})\Big\{\sum_{i=1}^k\Big(p_{\epsilon^2/2}*\eta_t(y_i)\Big)^{k+1}+\Big(p_{\epsilon^2/2k}*\eta_t(Y_k)\Big)^{k+1}\Big\}dy_1\ldots dy_k,$$

and once again we can change variables in the integrals and use (4.14) to bound this by a constant multiple of $\int\mathbb{E}\Big[\Big(\rho_\epsilon*\eta_t(x)\Big)^{k+1}\Big]dx$, and the inductive step is complete. □

Corollary 4.4 (Tightness of $\{(\rho_\epsilon * \eta_t^N(x)dx)_{t\geq 0}\}$) *Under the assumptions of Theorem 2.21, the sequence of measure valued processes $\{\rho_\epsilon * \eta_t^N(x)dx\}_{t\geq 0}$ (taking values in $\mathcal{D}_{[0,T]}(\mathcal{M}_F(\mathbb{R}^d))$) is tight.*

Proof

First observe that the proof, from Lemma 3.2, that $\mathbb{E}[\sup_{0\leq t\leq T}\langle 1, \eta_t^N\rangle]$ is bounded goes through unchanged, and since $\langle 1, \rho_\epsilon * \eta_t^N(x)dx\rangle = \langle 1, \eta_t^N\rangle$, compact containment follows.

As in the nonlocal case, it suffices to prove that for $T > 0$, and any $f \in C_b^\infty(\mathbb{R}^d)$ with bounded second derivatives and $\int |f(x)|dx < \infty$, the sequence of real-valued processes $\Big\{\Big(\int f(x)\rho_\epsilon * \eta_t^N(x)dx\Big)_{t\geq 0}\Big\}_{N\geq 1}$ is tight. Let us temporarily write $X_f^N(t)$ for $\int f(x)\rho_\epsilon * \eta_t^N(x)dx$ and set

$$w'\Big(X_f^N, \delta, T\Big) = \inf_{\{t_i\}} \max_i \sup_{s,t\in[t_{i-1},t_i)} \Big|X_f^N(t) - X_f^N(s)\Big|,$$

where $\{t_i\}$ ranges over all partitions of the form $0 = t_0 < t_1 < \cdots < t_{n-1} < T \leq t_n$ with $\min_{1\leq i\leq n}(t_i - t_{i-1}) > \delta$ and $n \geq 1$. Using Corollary 3.7.4 of [EK86], to prove tightness of the sequence of real-valued processes X_f^N it suffices to check compact containment of the sequence $\{\int f(x)\rho_\epsilon * \eta_t^N(x)dx\}_{N\geq 1}$ at any rational time t and that for every $\nu > 0$ and $T > 0$, there exists $\delta > 0$ such that

$$\limsup_{N\to\infty} \mathbb{P}\Big[w'\Big(X_f^N, \delta, T\Big) > \nu\Big] < \nu.$$

Evidently this will follow if we can show that this condition is satisfied when we replace the minimum over all partitions with mesh at least δ in the definition of w', by the partition into intervals of length exactly δ.

We have

$$\begin{aligned}\Big|\langle \rho_\epsilon * f, \eta_t^N\rangle - \langle \rho_\epsilon * f, \eta_s^N\rangle\Big| \leq \Big|\int_s^t \Big\langle \theta \int \Big(\rho_\epsilon * f(y) - \rho_\epsilon * f(x)\Big) q_\theta(x, dy), \eta_u^N(dx)\Big\rangle du\Big| \\ + \int_s^t \Big\langle |F\Big(\rho_\epsilon * \eta_u^N(x)\Big)|\rho_\epsilon * |f|(x), \eta_u^N(dx)\Big\rangle du + 2\sup_{0\leq u\leq T} |\widehat{M}^N(f)_u|, \quad (4.16)\end{aligned}$$

where $\widehat{M}^N(f)$ is the martingale of (3.4) with the test function f replaced by $\rho_\epsilon * f$. We control each of the three terms on the right hand side separately.

By the Intermediate Value Theorem, using T_t to denote the heat semigroup, there exists $s \in (0, 1/\theta)$ such that

$$\Big|\theta \int \big(\rho_\epsilon * f(y) - \rho_\epsilon * f(x)\big) q_\theta(x, dy)\Big| = \Big|\theta\Big(T_{\epsilon^2+1/\theta} f(x) - T_{\epsilon^2} f(x)\Big)\Big| \\ = |\partial_s T_{\epsilon^2+s} f(x)| = |T_{\epsilon^2+s} \Delta f(x)| \le \|\Delta f\|_\infty.$$

The first term in (4.16) is therefore bounded by

$$\|\Delta f\|_\infty |t - s| \sup_{0 \le u \le T} \langle 1, \eta_u^N \rangle.$$

We follow the approach of Lemma 3.2. Consulting (2.8), the angle bracket process of $\widehat{M}^N(f)$ satisfies $\mathbb{E}[\langle \widehat{M}_f^N \rangle_T] \le C(\theta/N) \int_0^T \mathbb{E}[\langle 1, \eta_s \rangle] ds \le C'\theta/N$ for some constants C and C'. Now, using the Burkholder-Davis-Gundy inequality and [BJY86], $\mathbb{E}[\sup_{0 \le u \le T} |\widehat{M}^N(f)_u|^2] \le C'' \mathbb{E}[\langle \widehat{M}^N(f) \rangle_u]$, and so using Markov's inequality,

$$\limsup_{N \to \infty} \mathbb{P}\Big[2 \sup_{0 \le u \le T} |\widehat{M}^N(f)_u| > \frac{\nu}{3}\Big] \le \limsup_{N \to \infty} \frac{36}{\nu^2} C'' \mathbb{E}\Big[\langle \widehat{M}^N(f) \rangle_T\Big] \le \limsup_{N \to \infty} \frac{36}{\nu^2} \frac{C' C'' \theta}{N \epsilon^d} = 0. \tag{4.17}$$

Now consider

$$\mathbb{E}\Big[\Big(\int_s^t \Big\langle \rho_\epsilon * |f|(x) \Big| F\Big(\rho_\epsilon * \eta_u^N(x)\Big)\Big|, \eta_u^N(dx)\Big\rangle du\Big)^2\Big] \\ = 2\mathbb{E}\Big[\int_s^t \Big\langle \rho_\epsilon * |f|(x) \Big| F\Big(\rho_\epsilon * \eta_u^N(x)\Big)\Big|, \eta_u^N(dx)\Big\rangle \int_u^t \Big\langle \rho_\epsilon * |f|(x) \Big| F\Big(\rho_\epsilon * \eta_r^N(x)\Big)\Big|, \eta_r^N(dx)\Big\rangle dr du\Big]. \tag{4.18}$$

Since F is polynomial, we use the approach of Corollary 4.3, the tower property, and Lemma 4.2, to bound this in terms of sums of terms of the form

$$\mathbb{E}\Big[\int_s^t (t-u) \int \rho_\epsilon * |f|(x) \rho_\epsilon * \eta_u^N(x)^j dx \int \rho_\epsilon * |f|(y) \rho_\epsilon * \eta_u^N(y)^k dy du\Big].$$

Now observe that, again using Lemma 4.2, since for nonnegative a and b, $a^j b^k \le a^{j+k} + b^{j+k}$,

$$\mathbb{E}\Big[\int \int \rho_\epsilon * |f|(x) \rho_\epsilon * \eta_u^N(x)^j \rho_\epsilon * |f|(y) \rho_\epsilon * \eta_u^N(y)^k dx dy\Big] \\ \le \mathbb{E}\Big[\int \int \|f\|_\infty \rho_\epsilon * \eta_u^N(x)^{j+k} \rho_\epsilon * |f|(y) dx dy + \int \int \rho_\epsilon * |f|(x) \|f\|_\infty \rho_\epsilon * \eta_u^N(y)^{j+k} dx dy\Big] \\ \le C \int |f|(x) dx.$$

Thus the quantity (4.18) is bounded by $C(t-s)^2$ for a new constant C which we can take to be independent of s, t and ϵ. Markov's inequality then gives

$$\mathbb{P}\Big[\|f\|_\infty \int_s^t \Big\langle \Big|F\Big(\rho_\epsilon * \eta_u^N(x)\Big)\Big|, \eta_u^N(dx)\Big\rangle du \geq \frac{\nu}{3}\Big] \leq C\frac{(t-s)^2}{\nu^2}.$$

A union bound gives that

$$\mathbb{P}\Big[\max_i \|f\|_\infty \int_{t_{i-1}}^{t_i} \Big\langle \Big|F\Big(\rho_\epsilon * \eta_u^N(x)\Big)\Big|, \eta_u^N(dx)\Big\rangle du \geq \frac{\nu}{3}\Big] \leq C\frac{T\delta}{\nu^2}. \tag{4.19}$$

Now using Markov's inequality, we can choose K so that

$$\mathbb{P}\Big[\|\Delta f\|_\infty \sup_{0\leq t\leq T}\langle 1, \eta_t^N\rangle > K\Big] < \frac{\nu}{3},$$

and so choosing δ so that $K\delta < \nu/3$ in this expression and $C\delta < \nu^3/3T$ in (4.19), combining with (4.17), the result follows. □

4.3 Continuity estimates for $\rho_\epsilon * \eta$

To identify the limit point of any convergent subsequence of $\{\rho_\epsilon * \eta^N(x)\}$, we will require some control on the spatial continuity of the functions $\rho_\epsilon * \eta^N(x)$. This will be inherited from the regularity of the transition density of the Gaussian random walk with generator $\mathcal{L}^\theta$, which in turn follows from its representation as that of a Brownian motion evaluated at the random time $T(t)$ defined in Lemma 4.1. Our approach will be to approximate $\psi_t^{\epsilon,x}(\cdot)$ by $p_{\epsilon^2+t}(x,\cdot)$, and to control the error that this introduces we need to control $T(t)-t$.

Lemma 4.5 *In the notation of Lemma 4.1, for any $A > 1$,*

$$\mathbb{P}\left\{T(t) - t > A(\epsilon^2+t)\right\} \leq \exp\left(-\frac{\theta A}{4}\Big(\epsilon^2+t\Big)\right).$$

Proof This is just a Chernoff bound. With Π a rate one Poisson process as in Lemma 4.1, for any $A > 1$,

$$\begin{aligned}
\mathbb{P}\left\{T(t) - t > A(\epsilon^2+t)\right\} &= \mathbb{P}\Big\{\Pi(\theta t) > \theta\Big(t + A(\epsilon^2+t)\Big)\Big\} \\
&\leq \frac{\mathbb{E}\left[\exp\left(\alpha\Pi(\theta t)\right)\right]}{\exp\Big(\alpha\theta\Big(t+A(\epsilon^2+t)\Big)\Big)} \\
&= \exp\Big(\theta t\Big(e^\alpha - 1\Big) - \alpha\theta\Big(t + A(\epsilon^2+t)\Big)\Big) \\
&\leq \exp\left(\theta t\Big(e^\alpha - \alpha - 1 - \frac{A\alpha}{2}\Big) - \frac{A\alpha}{2}\theta(\epsilon^2+t)\right).
\end{aligned}$$

Now set $\alpha = 1/2$. Since $A > 1$, $e^\alpha - \alpha - 1 - A\alpha/2 < 0$ and the result follows. $\square$

As advertised, we wish to control the difference between $\psi_t^{\epsilon,x}(y)$ and $p_{\epsilon^2+t}(x,y)$.

Lemma 4.6 *In the notation of Lemma 4.1, there exists a $C < \infty$ such that*

$$|\psi_t^{\epsilon,x}(y) - p_{\epsilon^2+t}(x,y)| \le \frac{C}{(\epsilon^2\theta)^{1/2}} p_{6(\epsilon^2+t)}(x,y) + \frac{C}{(\epsilon^2+t)^{d/2}} \exp(-\epsilon^2\theta/2). \tag{4.20}$$

Proof Still using the notation of Lemma 4.1, we partition into three events according to the value of $\tau(t)$. Let $A_1 = \{\tau(t) < -(\epsilon^2+t)/2\}$, $A_2 = \{\tau(t) > 2(\epsilon^2+t)\}$, and A_3 the remaining event, $\{-(\epsilon^2+t)/2 \le \tau(t) \le 2(\epsilon^2+t)\}$. Then,

$$\begin{aligned}
|\psi_t^{\epsilon,x}(y) - p_{\epsilon^2+t}(x,y)| &= \left|\mathbb{E}\left[p_{\epsilon^2+t+\tau(t)}(x,y) - p_{\epsilon^2+t}(x,y)\right]\right| \\
&\le \mathbb{E}\left[(1_{A_1} + 1_{A_2} + 1_{A_3})\left|p_{\epsilon^2+t+\tau(t)}(x,y) - p_{\epsilon^2+t}(x,y)\right|\right].
\end{aligned}$$

For the first term, note that if $a < b$ then

$$\begin{aligned}
|p_a(x,y) - p_b(x,y)| &= \frac{1}{(2\pi)^{d/2}}\left|\frac{1}{a^{d/2}}e^{-\|x-y\|^2/2a} - \frac{1}{b^{d/2}}e^{-\|x-y\|^2/2b}\right| \\
&= \frac{1}{(2\pi a^2)^{d/2}}e^{-\|x-y\|^2/2b}\left|e^{-\|x-y\|^2\left(\frac{1}{2a}-\frac{1}{2b}\right)} - \left(\frac{a}{b}\right)^{d/2}\right| \\
&\le C\left(\frac{b}{a}\right)^{d/2} p_b(x,y),
\end{aligned}$$

where the inequality follows because both terms under the absolute value are less than 1. Since, on the event A_1, $\tau(t) < 0$, we can apply this with $a = \epsilon^2 + t + \tau(t)$ and $b = \epsilon^2 + t$, and, using the bound (4.6),

$$\begin{aligned}
\mathbb{E}\left[1_{A_1}|p_a(x,y) - p_b(x,y)|\right] &\le C\left(\frac{\epsilon^2+t}{\epsilon^2}\right)^{d/2} p_{\epsilon^2+t}(x,y)\mathbb{P}\left\{\tau(t) < -\frac{\epsilon^2+t}{2}\right\} \\
&\le C\frac{1}{\epsilon^d}\mathbb{P}\left\{\tau(t) < -\frac{\epsilon^2+t}{2}\right\} \\
&\le \frac{C}{(\epsilon^2+t)^{d/2}}\exp\left(-\frac{\theta\epsilon^2}{2}\right).
\end{aligned}$$

For the third term, we will first collect some facts. Observe that on the event A_3, $\epsilon^2 + t + \tau(t)$ is between $(\epsilon^2+t)/2$ and $3(\epsilon^2+t)$, and for any s in this interval,

$$\begin{aligned}
p_{2s}(y) &\le \left(\frac{6(\epsilon^2+t)}{\epsilon^2+t}\right)^{d/2} p_{6(\epsilon^2+t)}(x,y) \\
&= 6^{d/2} p_{6(\epsilon^2+t)}(x,y).
\end{aligned} \tag{4.21}$$

Moreover, since $ue^{-u} \leq e^{-1}$ for all $u \geq 0$,

$$\begin{aligned}\frac{\|x-y\|^2}{s} p_s(x,y) &= \frac{4}{(2\pi s)^{d/2}} e^{-\frac{\|x-y\|^2}{4s}} \frac{\|x-y\|^2}{4s} e^{-\frac{\|x-y\|^2}{4s}} \\ &\leq C p_{2s}(x,y). \end{aligned} \tag{4.22}$$

Now, by the Intermediate Value Theorem,

$$\left| p_{\epsilon^2+t+\tau(t)}(x,y) - p_{\epsilon^2+t}(x,y) \right| = |\tau(t)| \left| \frac{\partial p_s(x,y)}{\partial s} \right| \tag{4.23}$$

for some s between $\epsilon^2 + t + \tau(t)$ and $\epsilon^2 + t$. Since

$$\begin{aligned}\partial_s p_s(x,y) &= \partial_s \left(\frac{1}{(2\pi s)^{d/2}} \exp\left(-\frac{\|x-y\|^2}{2s} \right) \right) \\ &= -\frac{d}{2s} p_s(x,y) + \frac{\|x-y\|^2}{2s^2} p_s(x,y),\end{aligned}$$

applying the inequality (4.22), using the fact that $p_s(x,y) \leq 2^{d/2} p_{2s}(x,y)$, and then (4.21), we have that for any $s \in ((\epsilon^2+t)/2, 3(\epsilon^2+t))$,

$$\left| \frac{\partial}{\partial s} p_s(x,y) \right| \leq \frac{C}{s} p_{2s}(x,y) \leq \frac{C}{\epsilon^2+t} p_{6(\epsilon^2+t)}(x,y).$$

Therefore, recalling that $\mathbb{E}[\tau(t)^2] = t/\theta$, substituting into (4.23),

$$\begin{aligned}\mathbb{E}\left[1_{A_3} \left| p_{\epsilon^2+t+\tau(t)}(x,y) - p_{\epsilon^2+t}(x,y) \right| \right] &\leq \frac{C}{\epsilon^2+t} p_{6(\epsilon^2+t)}(x,y) \mathbb{E}\left[|\tau(t)| \right] \\ &\leq \frac{C}{\epsilon^2+t} p_{6(\epsilon^2+t)}(x,y) \mathbb{E}\left[\tau(t)^2 \right]^{1/2} \\ &= \left(\frac{Ct}{\theta(\epsilon^2+t)^2} \right)^{1/2} p_{6(\epsilon^2+t)}(x,y) \\ &\leq \frac{C}{\sqrt{\theta\epsilon^2}} p_{6(\epsilon^2+t)}(x,y),\end{aligned}$$

where the last inequality follows from $2\epsilon^2 t \leq (\epsilon^2+t)^2$.

Finally, on the event $A_2 = \{\tau(t) > 2(\epsilon^2+t)\}$, we simply use

$$\left| p_{\epsilon^2+t+\tau(t)}(x,y) - p_{\epsilon^2+t}(x,y) \right| \leq \frac{C}{(\epsilon^2+t)^{d/2}},$$

so that

$$\mathbb{E}\left[1_{A_2} \left| p_{\epsilon^2+t+\tau(t)}(x,y) - p_{\epsilon^2+t}(x,y) \right| \right] \leq \frac{C}{(\epsilon^2+t)^{d/2}} \mathbb{P}\left\{ \tau(t) > 2(\epsilon^2+t) \right\},$$

and apply Lemma 4.5 with $A = 2$. □

The last result will be useful when combined with the next bound for the heat kernel.

Lemma 4.7 *Let $s > 0$, and $x, y, z \in \mathbb{R}^d$. The following estimate holds:*

$$|p_s(x,z) - p_s(y,z)| \leq \frac{C\|x-y\|}{\sqrt{s}} \left(p_{2s}(x,z) + p_{2s}(y,z)\right),$$

where the constant C does not depend on x, y, z or s.

Proof

Expanding the difference of two squares,

$$e^{-\frac{\|y-z\|^2}{2s}} - e^{-\frac{\|x-z\|^2}{2s}} = \left(e^{-\frac{\|y-z\|^2}{4s}} - e^{-\frac{\|x-z\|^2}{4s}}\right)\left(e^{-\frac{\|y-z\|^2}{4s}} + e^{-\frac{\|x-z\|^2}{4s}}\right).$$

Now, thinking of the first term in brackets as a function of a single variable x on the line segment $[y,z]$ connecting y to z, we can apply the Intermediate Value Theorem and take the modulus to bound this expression by

$$\|y-x\| \left(\frac{2\|w-z\|}{4s} \exp\left(-\frac{\|w-z\|^2}{4s}\right)\right) (4\pi s)^{d/2} \left(p_{2s}(y,z) + p_{2s}(x,z)\right)$$

for some $w \in [y,z]$. Using the fact that xe^{-x^2} is uniformly bounded, we can bound the first bracket in the last equation by $C/\sqrt{s}$, and the result follows. □

We now have the ingredients that we need to write down a continuity estimate for $\rho_\epsilon * \eta$. We fix $\delta > 0$ and suppose that $s > \delta$. Let us write $\widehat{\epsilon}(\delta, \epsilon, \theta) := \frac{1}{(\epsilon^2+\delta)^{d/2}} e^{-\epsilon^2\theta/2}$, and note that under the assumption that $\epsilon^2\theta \to \infty$, for each fixed $\delta > 0$, $\lim_{\epsilon\to 0, \theta\to\infty} \widehat{\epsilon}(\delta, \epsilon, \theta) = 0$.

Using the semimartingale decomposition (4.4), and Lemma 4.6, we have

$$
\begin{aligned}
|\rho_\epsilon * \eta_s(y) - \rho_\epsilon * \eta_s(w)| &= |\langle p_{\epsilon^2}(y,z) - p_{\epsilon^2}(w,z), \eta_s(dz)\rangle| \\
&\leq \Big\langle |p_{\epsilon^2+s}(y,z) - p_{\epsilon^2+s}(w,z)|, \eta_0(dz)\Big\rangle \\
&\quad + \int_0^{s-\delta} \Big\langle |p_{s-r+\epsilon^2}(y,z) - p_{s-r+\epsilon^2}(w,z)||F(\rho_\epsilon * \eta_r(z)|, \eta_r(dz)\Big\rangle dr \\
&\quad + \Big\langle \frac{C}{(\theta\epsilon^2)^{1/2}}\Big(p_{6(\epsilon^2+s)}(y,z) + p_{6(\epsilon^2+s)}(w,z)\Big) + C\widehat{\epsilon}(\delta,\epsilon,\theta), \eta_0(dz)\Big\rangle \\
&\quad + \int_0^{s-\delta} \Big\langle \Big\{ \frac{C}{(\epsilon^2\theta)^{1/2}} |p_{6(s-r+\epsilon^2)}(y,z) + p_{6(s-r+\epsilon^2)}(w,z)| + \widehat{\epsilon}(\delta,\epsilon,\theta)\Big\} |F(\rho_\epsilon * \eta_r(z)|, \eta_r(dz)\Big\rangle dr \\
&\quad + \int_{s-\delta}^{s} \Big\langle |\psi^{\epsilon,y}_{s-r}(z) + \psi^{\epsilon,w}_{s-r}(z)||F(\rho_\epsilon * \eta_r(z)|, \eta_r(dz)\Big\rangle dr + |M_s(y)| + |M_s(w)| \\
&\leq \Big\langle \frac{\|y-w\|}{\sqrt{s+\epsilon^2}}\Big(p_{2(s+\epsilon^2)}(y,z) + p_{2(s+\epsilon^2)}(w,z)\Big), \eta_0(dz)\Big\rangle \\
&\quad + \int_0^{s-\delta} \Big\langle \frac{\|y-w\|}{\sqrt{s-r+\epsilon^2}}\Big(p_{2(s-t+\epsilon^2)}(y,z) + p_{2(s-r+\epsilon^2)}(w,z)\Big) |F(\rho_\epsilon * \eta_r(z)|, \eta_r(dz)\Big\rangle dr \\
&\quad + \Big\langle \frac{C}{(\theta\epsilon^2)^{1/2}}\Big(p_{6(\epsilon^2+s)}(y,z) + p_{6(\epsilon^2+s)}(w,z)\Big) + C\widehat{\epsilon}(\delta,\epsilon,\theta), \eta_0(dz)\Big\rangle \\
&\quad + \int_0^{s-\delta} \Big\langle \Big\{ \frac{C}{(\epsilon^2\theta)^{1/2}} (p_{6(s-r+\epsilon^2)}(y,z) + p_{6(s-r+\epsilon^2)}(w,z)) + \widehat{\epsilon}(\delta,\epsilon,\theta)\Big\} |F(\rho_\epsilon * \eta_r(z)|, \eta_r(dz)\Big\rangle dr \\
&\quad + \int_{s-\delta}^{s} \Big\langle |\psi^{\epsilon,y}_{s-r}(z) + \psi^{\epsilon,w}_{s-r}(z)||F(\rho_\epsilon * \eta_r(z)|, \eta_r(dz)\Big\rangle dr + |M_s(y)| + |M_s(w)|.
\end{aligned}
\tag{4.24}
$$

Although this expression is lengthy, we have successfully isolated the terms involving $\|y-w\|$, which will control the regularity as we pass to the limit. Asymptotically, we don't expect the martingale terms to contribute, since their quadratic variation scales with $\theta/(N\epsilon^d)$; under the assumption that $\epsilon^2\theta \to \infty$, for any fixed $\delta > 0$, the terms arising from approximating the transition density $\psi^{\epsilon,\cdot}_{s-r}(\cdot)$ of the Gaussian walk by $p_{s-r+\epsilon^2}(\cdot,\cdot)$ at times with $s - r > \delta$ will tend to zero; and the moment bounds of Lemma 4.2 will allow us to control the integral over $[s-\delta, s]$. There is some technical work to be done to rigorously identify the limit points of $\rho_\epsilon * \eta^N$, but it really amounts to applying the tower property and our moment bounds from Lemma 4.2 and Corollary 4.3.

4.4 Identification of the limit

We now turn to the identification of the limit points of the sequence $\Big\{\Big(\rho_\epsilon * \eta_t^N(x)dx\Big)_{t\geq 0}\Big\}_{N\geq 1}$. We would like to show that any limit point solves (2.21) in the limit, i.e.,

$$\langle f(x), \varphi(t,x)dx\rangle = \int_0^t \Big\langle \frac{1}{2}\Delta f(x) + f(x)F(\varphi(s,x)), \varphi(s,x)dx\Big\rangle ds. \tag{4.25}$$

Since $\langle f, \rho_\epsilon * \eta_t^N(x)dx\rangle = \langle \rho_\epsilon * f(x), \eta_t^N(dx)\rangle$, and the limit is deterministic, this will follow if we can show that each of the terms in the semimartingale decomposition (2.7), with the test function f replaced by $\rho_\epsilon * f(\cdot)$, converges to the corresponding term in (4.25).

The linear term is straightforward. Write $T_\cdot$ for the heat semigroup, so that $\rho_\epsilon * f(x) = T_{\epsilon^2} f(x)$. By a Taylor expansion,

$$\begin{aligned}\int_0^t \Big\langle \mathcal{L}^\theta T_{\epsilon^2} f, \eta_s^N(dx)\Big\rangle ds &= \int_0^t \Big\langle \frac{1}{2}\Delta T_{\epsilon^2} f(x), \eta_s^N(dx)\Big\rangle + \mathcal{O}\Big(\frac{1}{\theta}\Big) \\ &= \int_0^t \Big\langle \frac{1}{2}\Delta f(x), \rho_\epsilon * \eta_s^N(x)dx\Big\rangle ds + \mathcal{O}\Big(\frac{1}{\theta}\Big).\end{aligned}$$

Thus, from weak convergence we can deduce that under our scaling, for any (weakly) convergent subsequence $\{\rho_\epsilon * \eta^N(x)dx\}_{N\geq 1}$,

$$\int_0^t \Big\langle \mathcal{L}^\theta T_{\epsilon^2} f, \eta_s^N(dx)\Big\rangle ds \to \int_0^t \Big\langle \frac{1}{2}\Delta f(x), \varphi(s,x)dx\Big\rangle ds.$$

The nonlinear term in the semimartingale decomposition is more intricate. It takes the form

$$\mathbb{E}\Big[\int_0^t \Big\langle T_{\epsilon^2} f(y) F\Big(\rho_\epsilon * \eta_s(y)\Big), \eta_s(dy)\Big\rangle ds\Big]$$

and we should like to show that this converges to

$$\int_0^t \int f(y)F(\varphi(s,y))\varphi(s,y)dyds.$$

We proceed in stages. First we should like to transfer the heat semigroup from $T_{\epsilon^2} f$ onto η_s. Since f is smooth, this will follow easily if we can show that

$$\mathbb{E}\Big[\int_0^t \Big\langle T_{\epsilon^2} f(y) F\Big(\rho_\epsilon * \eta_s(y)\Big), \eta_s(dy)\Big\rangle ds\Big] \sim \mathbb{E}\Big[\int_0^t \Big\langle T_{\epsilon^2} f(y) F\Big(\rho_\epsilon * \eta_s(y)\Big), \rho_\epsilon * \eta_s(y)dy\Big\rangle ds\Big].$$

This is the content of Proposition 4.8.

Proposition 4.8 *Under the conditions of Theorem 2.21,*

$$\lim_{\epsilon\to 0}\mathbb{E}\Big[\Big|\int_0^t \Big\langle T_{\epsilon^2}f(y)F(\rho_\epsilon * \eta_s(y)),\eta_s(y)\Big\rangle - \Big\langle T_{\epsilon^2}f(y)F(\rho_\epsilon * \eta_s(y)),\rho_\epsilon * \eta_s(y)dy\Big\rangle ds\Big|\Big] = 0. \tag{4.26}$$

Proof In fact we are going to fix $\delta > 0$, with $t > \delta$, and show that the expression on the left hand side of (4.26) is less than a constant times δ, with a constant independent of δ, N, and ϵ. Since δ is arbitrary, the result will follow.

We first note that,

$$\begin{aligned}
&\langle T_{\epsilon^2}f(y)F(\rho_\epsilon * \eta_s(y)),\rho_\epsilon * \eta_s(y)dy\rangle - \langle T_{\epsilon^2}f(y)F(\rho_\epsilon * \eta_s(dy)),\eta_s(dy)\rangle\\
&= \langle \int T_{\epsilon^2}f(y)F(\rho_\epsilon * \eta_s(y))\rho_\epsilon(y-w)dy,\eta_s(dw)\rangle - \langle T_{\epsilon^2}f(w)F(\rho_\epsilon * \eta_s(w)),\eta_s(dw)\rangle\\
&= \langle \int \Big\{T_{\epsilon^2}f(y)F(\rho_\epsilon * \eta_s(y)) - T_{\epsilon^2}f(w)F(\rho_\epsilon * \eta_s(w))\Big\}\rho_\epsilon(w-y)dy,\eta_s(dw)\rangle.
\end{aligned}$$

Let us denote the integral against dy in the last expression by I, that is

$$I := \int \{T_{\epsilon^2}f(y)F(\rho_\epsilon * \eta_s(y)) - T_{\epsilon^2}f(w)F(\rho_\epsilon * \eta_s(w))\}\rho_\epsilon(w-y)dy,$$

and note that $|I|$ is bounded by

$$\begin{aligned}
&\int \Big\{|F(\rho_\epsilon * \eta_s(y)) - F(\rho_\epsilon * \eta_s(w))|T_{\epsilon^2}f(y) + F(\rho_\epsilon * \eta_s(w))|T_{\epsilon^2}f(y) - T_{\epsilon^2}f(w))|\Big\}\rho_\epsilon(w-y)dy\\
&\le \int \|f\|_\infty\Big|F(\rho_\epsilon * \eta_s(y)) - F(\rho_\epsilon * \eta_s(w))\Big|\rho_\epsilon(w-y)dy + C\epsilon\|f'\|_\infty|F(\rho_\epsilon * \eta_s(w))|, \quad (4.27)
\end{aligned}$$

where we have used that

$$\int |T_{\epsilon^2}f(y) - T_{\epsilon^2}f(w)|p_{\epsilon^2}(w,y)dy \le \|f'\|_\infty \int |y-w|p_{\epsilon^2}(w,y)dy.$$

Now recall that F is a polynomial of degree n, and so there exist real numbers b_k such that $F(a) - F(b) = (a-b)\sum_{k=1}^{n-1} b_k a^k b^{n-1-k}$ and so

$$|F(\rho_\epsilon * \eta_s(y)) - F(\rho_\epsilon * \eta_s(w))| \le |\rho_\epsilon * \eta_s(y) - \rho_\epsilon * \eta_s(w)|\sum_{k=1}^{n-1}|b_k|\left(\rho_\epsilon * \eta_s(y)^{n-1} + \rho_\epsilon * \eta_s(w)^{n-1}\right).$$

Combining the above, we have reduced the problem to showing that for any $k \ge 0$,

$$\lim_{\epsilon\to 0}\mathbb{E}\Big[\int_0^t \Big\langle \int |\rho_\epsilon * \eta_s(y) - \rho_\epsilon * \eta_s(w)|\left(\rho_\epsilon * \eta_s(y))^k + \rho_\epsilon * \eta_s(w)^k\right)p_{\epsilon^2}(w,y)dy,\eta_s(dw)\Big\rangle ds\Big] = 0. \tag{4.28}$$

We are going to use the estimate (4.24). First note that by Lemma 4.2 the contribution to (4.28) from the integral over the time interval $[0,\delta]$ is $\mathcal{O}(\delta)$. We focus instead on the interval $(\delta,t]$.

The first term in (4.24) gives

$$\int_\delta^t \mathbb{E}\Big[\Big\langle \int \Big\langle \frac{\|y-w\|}{\sqrt{s+\epsilon^2}}\Big(p_{2(s+\epsilon^2)}(y,z)+p_{2(s+\epsilon^2)}(w,z)\Big),\eta_0(dz)\Big\rangle \\ \Big(\rho_\epsilon * \eta_s(y)^k + \rho_\epsilon * \eta_s(w)^k\Big)p_{\epsilon^2}(w,y)dy,\eta_s(dw)\Big\rangle\Big]ds.$$

We "borrow" from the exponential term to see that $\|y-w\|p_{\epsilon^2}(w,y) \le C\epsilon p_{2\epsilon^2}(w,y)$ and so bound this by

$$C\int_\delta^t \frac{\epsilon}{\sqrt{s+\epsilon^2}}\mathbb{E}\Big[\Big\langle \int \Big\langle \Big(p_{2(s+\epsilon^2)}(y,z)+p_{2(s+\epsilon^2)}(w,z)\Big),\eta_0(dz)\Big\rangle \\ \Big(\rho_\epsilon * \eta_s(y)^k + \rho_\epsilon * \eta_s(w)^k\Big)p_{2\epsilon^2}(w,y)dy,\eta_s(dw)\Big\rangle ds\Big]. \quad (4.29)$$

The four terms in the product are taken separately, according to the combinations of w and y appearing. First,

$$\int_\delta^t \frac{\epsilon}{\sqrt{s+\epsilon^2}}\mathbb{E}\Big[\Big\langle \int \Big\langle p_{2(s+\epsilon^2)}(y,z),\eta_0(dz)\Big\rangle \rho_\epsilon * \eta_s(y)^k p_{2\epsilon^2}(w,y)dy,\eta_s(dw)\Big\rangle\Big]ds$$

can be rewritten as

$$\int_\delta^t \frac{\epsilon}{\sqrt{s+\epsilon^2}}\mathbb{E}\Big[\int\int \Big\langle p_{2(s+\epsilon^2)}(y,z),\eta_0(dz)\Big\rangle \rho_\epsilon * \eta_s(y)^k p_{\epsilon^2}(x,y)\rho_\epsilon * \eta_s(x)dydxds\Big] \\ \le \int_\delta^t \frac{\epsilon}{\sqrt{s+\epsilon^2}}\mathbb{E}\Big[\int\int \Big\langle p_{2(s+\epsilon^2)}(y,z),\eta_0(dz)\Big\rangle \Big(\rho_\epsilon * \eta_s(y)^{k+1} + \rho_\epsilon * \eta_s(x)^{k+1}\Big)p_{\epsilon^2}(x,y)dydx\Big]ds,$$

and using Lemma 4.2 and the tower property, and integrating with respect to s, under our assumptions on η_0, this is bounded by

$$C\epsilon\int_\delta^t \frac{1}{\sqrt{s+\epsilon^2}}\mathbb{E}\Big[\int\int \Big\langle p_{2(s+\epsilon^2)}(y,z),\eta_0(dz)\Big\rangle \\ \Big(\rho_\epsilon * \eta_0(y) + \rho_\epsilon * \eta_0(y)^{k+1} + \rho_\epsilon * \eta_0(x) + \rho_\epsilon * \eta_0(x)^{k+1}\Big)p_{\epsilon^2}(x,y)dydx\Big]ds \\ \le C'\epsilon\int_\delta^t \frac{1}{(s+\epsilon^2)^{d/2}}ds.$$

For fixed δ, this bound tends to zero as $\epsilon \to 0$. The term involving $\Big\langle p_{2(s+\epsilon^2)}(w,z),\eta_0(dz)\Big\rangle \rho_\epsilon * \eta_s(w)^k$ is handled similarly.

On the other hand

$$\Big\langle \int \langle p_{2(s+\epsilon^2)}(y,z), \eta_0(dz)\rangle \rho_\epsilon * \eta_s(w)^k p_{2\epsilon^2}(w,y)dy, \eta_s(dw)\Big\rangle$$
$$\leq \frac{C}{(s+\epsilon^2)^{d/2}} \langle 1, \eta_0\rangle \Big\langle \rho_\epsilon * \eta_s(w)^k, \eta_s(dw)\Big\rangle,$$

and since $\langle 1, \eta_0\rangle$ is uniformly bounded we apply Corollary 4.3 to obtain a bound on the contribution to (4.29) from this term of the same form as the others.

Now consider the contribution to the left hand side of (4.28) from the second term in (4.24). Since F is a polynomial, it is bounded by a sum of terms of the form

$$\int_\delta^t \int_0^{s-\delta} \Big\langle \int \frac{\|y-w\|}{\sqrt{s-r+\epsilon^2}} \Big\langle \Big(p_{2(s-r+\epsilon^2)}(y,z) + p_{2(s-r+\epsilon^2)}(w,z)\Big)\rho_\epsilon * \eta_r(z)^j, \eta_r(dz)\Big\rangle$$
$$\rho_\epsilon * \eta_s(y)^k p_{\epsilon^2}(y,w)dy, \eta_s(dw)\Big\rangle drds$$
$$\leq C\epsilon \int_\delta^t \int_0^{s-\delta} \frac{1}{\sqrt{s-r+\epsilon^2}} \Big\langle \int \Big\langle \Big(p_{2(s-r+\epsilon^2)}(y,z) + p_{2(s-r+\epsilon^2)}(w,z)\Big)\rho_\epsilon * \eta_r(z)^j, \eta_r(dz)\Big\rangle$$
$$\rho_\epsilon * \eta_s(y)^k p_{2\epsilon^2}(y,w)dy, \eta_s(dw)\Big\rangle drds,$$

where as usual we have "borrowed" from the exponential term in $p_\epsilon^2(y,w)$ to replace $\|y-w\|$ by a constant times ϵ.

Once again, our approach is to rearrange terms so that we can apply Lemma 4.2 or Corollary 4.3 to obtain a bound on the contribution to (4.28) from these terms of the form $C\epsilon$ (where C may depend on δ but not ϵ).

For example, using the Chapman-Kolmogorov equation to rewrite

$$\int \Big\langle \Big\langle p_{2(s-r+\epsilon^2)}(y,z)\rho_\epsilon * \eta_r(z)^j, \eta_r(dz)\Big\rangle \rho_\epsilon * \eta_s(y)^k p_{2\epsilon^2}(y,w), \eta_s(dw)\Big\rangle dy$$

as

$$\int\int \Big\langle p_{2(s-r+\epsilon^2)}(y,z)\rho_\epsilon * \eta_r(z)^j, \eta_r(dz)\Big\rangle \rho_\epsilon * \eta_s(y)^k p_{\epsilon^2}(y,x)\rho_\epsilon * \eta_s(x)dxdy,$$

and using Lemma 4.2 and the tower property, we are led to control terms of the form

$$\mathbb{E}\Big[\int \Big\langle p_{2(s-r+\epsilon^2)}(y,z)\rho_\epsilon * \eta_r(z)^j, \eta_r(dz)\Big\rangle \rho_\epsilon * \eta_r(y)^{k+1} dy\Big].$$

This, in turn, is at most

$$\mathbb{E}\Big[\Big\langle \rho_\epsilon * \eta_r(z)^{j+k+1}, \eta_r(dz)\rangle\Big] + \mathbb{E}\Big[\int\int p_{2(s-r)}(y,x)\rho_\epsilon * \eta_r(x)\rho_\epsilon * \eta_r(y)^{j+k+1}dydx\Big]$$
$$\leq \mathbb{E}\Big[\Big\langle \rho_\epsilon * \eta_r(z)^{j+k+1}, \eta_r(dz)\rangle\Big] + 2\int \mathbb{E}\Big[\rho_\epsilon * \eta_r(x)^{j+k+2}\Big]dx,$$

which is bounded by Lemma 4.2.

We now turn to the contribution arising from the martingale terms in (4.24):

$$\mathbb{E}\Big[\int_\delta^t \Big\langle \int \Big(|M_s(y)| + |M_s(w)|\Big)\Big(\rho_\epsilon * \eta_s(y)^k + \rho_\epsilon * \eta_s(w)^k\Big) p_{\epsilon^2}(w,y)dy, \eta_s(dw)\Big\rangle ds\Big].$$

Since $\psi^{\epsilon,x}_{t-s}(y) = \mathbb{E}[p_{T(t-s)+\epsilon^2}(x,y)]$, rearranging (4.4) we see that we can pull a convolution with $p_{\epsilon^2/2}$ out of our expressions for $M_s(y)$ and $M_s(w)$ and so all the manipulations that we used to control terms above with still be valid. To deal with the two terms in the product involving $|M_s(y)|$, we write the first as $\int |M_s(y)|\rho_\epsilon * \eta_s(y)^{k+1}dy$ and then use Hölder's inequality, Lemma 4.2, and the fact that $\mathbb{E}\Big[|M_s(y)|^2\Big]$ is $\mathcal{O}\Big(\theta/(N\epsilon^d)\Big)$ to see that the contribution from this term tends to zero in the limit. For the second, we use the idea of the proof of Corollary 4.3 to reduce to a form to which we can apply Hölder's inequality.

Control of the terms arising from approximating $\psi^{\epsilon,x}$ by the heat kernel follows in an entirely analogous way.

Combining the above, we see that given $\delta > 0$,

$$\lim_{\epsilon\to 0}\mathbb{E}\Big[\Big|\int_0^t \Big\langle T_{\epsilon^2} f(y)F(\rho_\epsilon * \eta_s(y)), \eta_s(y)\Big\rangle ds - \int_0^t \Big\langle T_{\epsilon^2} f(y)F(\rho_\epsilon * \eta_s(y)), \rho_\epsilon * \eta_s(y)dy\Big\rangle ds\Big|\Big] < C\delta,$$

where the constant C is independent of δ. Since δ was arbitrary, the proof is complete. □

Since f is smooth, $T^2_\epsilon f - f$ is $\mathcal{O}(\epsilon)$, with an application of the triangle inquality,

$$\lim_{\epsilon\to 0}\mathbb{E}\Big[\Big|\int_0^t \Big\langle T_{\epsilon^2} f(y)F(\rho_\epsilon * \eta_s(y)), \eta_s(y)\Big\rangle ds - \int_0^t \Big\langle f(y)F(\rho_\epsilon * \eta_s(y)), \rho_\epsilon * \eta_s(y)dy\Big\rangle ds\Big|\Big] < C\delta,$$

now follows immediately. Thus to complete the characterisation of the limit, it remains to show that if we take a convergent subsequence $\Big\{\Big(\rho_\epsilon * \eta^N_t(dx)\Big)_{t\geq 0}\Big\}$ converging to a limit point $\Big(\varphi(t,x)dx\Big)_{t\geq 0}$, then

$$\int_0^t \int f(x)\rho_\epsilon * \eta_s(x)F(\rho_\epsilon * \eta^N_s(x))dxds \to \int_0^t \int f(x)\varphi(s,x)F(\varphi(s,x))dxds.$$

Since F is a polynomial, we consider powers of $\rho_\epsilon * \eta$. To illustrate the approach, we first prove that

$$\int_0^t \int f(x)\rho_\epsilon * \eta^N_s(x)^2 dxds \to \int_0^t \int f(x)\varphi(s,x)^2 dxds. \tag{4.30}$$

The convergence of higher powers will follow in an entirely analogous manner, but with more complex expressions.

The approach is standard. We fix $\tau > 0$ and, in keeping with our notation ρ_ϵ, in this subsection, use ρ_τ to denote the symmetric Gaussian kernel with variance parameter τ^2. Our strategy is to show that, up to an error that tends to zero as $\tau \to 0$,

$$\int_0^t \int f(z)\rho_\epsilon * \eta_s(z)^2 dzds \sim \int_0^t \int \int f(z)(\rho_\epsilon * \eta_s)(z)\rho_\tau(z-y)(\rho_\epsilon * \eta_s)(y)dzdyds. \quad (4.31)$$

Analogously, also up to an error that vanishes as $\tau \to 0$,

$$\int_0^t \int f(z)\varphi(s,z)^2 dzds \sim \int_0^t \int \int f(z)\varphi(s,z)\rho_\tau(z-y)\varphi(s,y)dzdyds. \quad (4.32)$$

On the other hand, weak convergence of $\rho_\epsilon * \eta$ (plus continuity of the mapping $(z,y) \to f(z)\rho_\tau(z-y)$) gives that

$$\int_0^t \int \int f(z)(\rho_\epsilon * \eta_s)(z)\rho_\tau(z-y)(\rho_\epsilon * \eta_s)(y)dzdyds \to \int_0^t \int \int f(z)\varphi(s,z)\rho_\tau(z-y)\varphi(s,y)dzdyds. \quad (4.33)$$

Since τ is arbitrary, the convergence (4.30) will follow.

Proposition 4.9 *Under the conditions of Theorem 2.21, we have that along any convergent subsequence,*

$$\limsup_{\epsilon \to 0} \mathbb{E}\Big[\Big| \int_0^t \int f(y)\rho_\epsilon * \eta_s(y)^2 dyds - \int_0^t \int \int f(z)(\rho_\epsilon * \eta_s)(z)\rho_\tau(z-y)(\rho_\epsilon * \eta_s)(y)dzdyds\Big|\Big] \leq C\tau, \quad (4.34)$$

where C is independent of τ.

Proof First note,

$$\int_0^t \mathbb{E}\left[\left|\langle f(y), (\rho_\epsilon * \eta_s(y))^2)dy\rangle - \int \int f(y)(\rho_\epsilon * \eta_s)(z)\rho_\tau(z-y)(\rho_\epsilon * \eta_s)(y)dzdy\right|\right] ds$$
$$\leq \|f\|_\infty \int \int_0^t \mathbb{E}\left[\int \left\{\left|(\rho_\epsilon * \eta_s)(y) - (\rho_\epsilon * \eta_s)(z)\right| \rho_\tau(z-y)dz\right\} (\rho_\epsilon * \eta_s)(y)\right] dsdy. \quad (4.35)$$

Now proceed exactly as in the proof of Proposition 4.8. The only distinction is that $|p_{\epsilon^2}(y,z) - p_{\epsilon^2}(w,z)|$ is replaced by $|p_\tau(y,z) - p_\tau(w,z)|$ and the estimate $\|y-w\|p_{\tau^2}(y,w) \leq C\tau p_{2\tau^2}(y,w)$ replaces the corresponding statement with ϵ^2 replacing τ^2 in our previous argument. $\square$

The extension of Proposition 4.9 to higher moments is straightforward, if notationally messy. For fixed (but arbitrary) τ, one shows that

$$\limsup_{\epsilon\to 0} \mathbb{E}\Big[\Big| \int_0^t \int f(y)\rho_\epsilon * \eta_s^N(y)^k dy ds \\ - \int_0^t \int \cdots \int f(y_1)\rho_\epsilon * \eta_s^N(y_1) \prod_{i=2}^{k} \rho_\tau(y_i - y_{i-1})\rho_\epsilon * \eta_s^N(y_i) dy_k \ldots dy_1 ds\Big|\Big] \leq C\tau,$$

as well as a corresponding statement with $\rho_\epsilon * \eta_s^N(x)$ replaced by $\varphi(s,x)$ and then uses weak convergence to see that, up to an error of order τ, any limit point of the sequence $\{\rho_\epsilon * \eta^N(x)dx\}$ solves (the weak form of) equation (2.21). Since τ was arbitrary, the proof of Theorem 2.21 is complete.

5

Branching Particle Systems with Local Regulation: Lookdown Representation

5.1 The lookdown process

Our characterisation of the motion of lines of descent (from which we establish that of ancestral lineages) when we pass to the scaling limit in our model will be justified via a lookdown construction. In this section we present such a construction for the general population model of Definition 2.4. It will be in the spirit of [KR11]. The general set-up is as follows. Each individual will be labelled with a "level", a number in $[0, N]$. We will still encode the process embellished by these levels as a point measure: if the i^{th} individual's spatial location is x_i and level is u_i, then we will write

$$\xi^N = \sum_i \delta_{x_i, u_i},$$

which is a measure on $\mathbb{R}^d \times [0, N]$. Note that each individual contributes mass 1 to the measure, not $1/N$ as above. If we assign mass $1/N$ to each individual and ignore the levels we will recover our population model. Moreover, at any time, the levels of individuals in a given spatial region will be exchangeable and conditionally uniform on $[0, N]$: in particular, choosing the k individuals with the lowest levels in that region is equivalent to taking a uniform random sample of size k from the population in the region. However, this exchangeability is only as regards the *past*: an individual's level encodes information about

their future reproductive output, since individuals with lower levels tend to live longer, and have more offspring. For more explanation of the set-up and how this is possible, see [KR11] and [EK19] (and note that our N corresponds to the λ of those papers). The power of this approach is that we can pass to a limit under the same scalings as described in Theorem 2.10, and the limiting "spatial-level" process will still be a point measure, and so we explicitly retain the notion of individuals and lineages in the infinite-population limit.

5.1.1 Lookdown representation of the model of Definition 2.4

For the remainder of this section, when there is no risk of ambiguity we shall suppress the superscript N on the processes η and ξ.

In this subsection, we'll define the process $(\xi_t)_{t\geq 0}$ in terms of the dynamics of labelled particles, and write down its generator. The dynamics depend on the spatial locations of particles, and in this section η_t is the corresponding spatial measure, i.e.,

$$\eta_t(\cdot) = \frac{1}{N}\xi_t(\cdot \times [0, N]).$$

A nontrivial consequence of the way we define ξ_t will be that this process has the same distribution as the process $(\eta_t)_{t\geq 0}$ of Definition 2.4. (This provides our justification for using the same notation for both.)

Following [EK19], we build the generator step by step from its component parts. Suppose that the initial population is composed of $O(N)$ particles with levels uniformly distributed on $[0, N]$, and that the current state of the population is ξ, with spatial projection η.

An individual at spatial location x with level u produces one juvenile offspring at rate

$$2\theta\left(1 - \frac{u}{N}\right)\gamma(x, \eta),$$

which disperses to a location relative to x drawn from the kernel $q_\theta(x, \cdot)$. Averaging over the uniform distribution of the level u, we recover the birth rate $\theta\gamma(x, \eta)$. This juvenile – suppose its location is y – either survives, with probability $r(y, \eta)$, or immediately dies. (As before, "maturity" is instantaneous.) If it survives, a new level u_1 is sampled independently and uniformly from $[u, N]$, and the parent and the offspring are assigned

in random order to the levels $\{u, u_1\}$. This random assignment of levels to parent and offspring will ensure that assignment of individuals to levels remains exchangeable.

Evidently this mechanism increases the proportion of individuals with higher levels. To restore the property that the distribution of levels is conditionally uniform given η, we impose that the level v of an individual at location x evolves according to the differential equation

$$\dot{v} = -\theta \frac{v}{N} (N - v) \gamma(x, \eta) \int_{\mathbb{R}^d} r(y, \eta) q_\theta(x, dy).$$

Since $v \in [0, N]$, this moves levels down; see [EK19], Section 3.4 for a detailed explanation.

Levels never cross below 0, while particles whose levels move above N are regarded as dead (and are removed from the population). Therefore, in order to incorporate death, the level of the individual at location x with level u moves upwards at an additional rate $\theta \mu_\theta(x, \eta) u$. Since levels are uniform, it is easy to check that if μ_θ were constant, this would imply an exponential lifetime for each individual; see [EK19], Section 3.1 for more general justification.

Putting these together, the level u of an individual at x evolves according to:

$$\dot{u} = -\theta \frac{u}{N} (N - u) \gamma(x, \eta) \int_{\mathbb{R}^d} r(y, \eta) q_\theta(x, dy) + \theta \mu_\theta(x, \eta) u. \tag{5.1}$$

We shall write

$$b_\theta(x, \eta) := \theta \left(\gamma(x, \eta) \int_{\mathbb{R}^d} r(y, \eta) q_\theta(x, dy) - \mu_\theta(x, \eta) \right),$$

which captures the local net difference between reproduction and death, and

$$c_\theta(x, \eta) := \frac{\theta}{N} \gamma(x, \eta) \int_{\mathbb{R}^d} r(y, \eta) q_\theta(x, dy), \tag{5.2}$$

which captures the local rate of production of successful offspring. Recall from equation (2.6) that $F(x, \eta) = \theta(r(x, \eta)\gamma(x, \eta) - \mu_\theta(x, \eta))$, and so

$$b_\theta(x, \eta) = \theta \gamma(x, \eta) \int_{\mathbb{R}^d} (r(y, \eta) - r(x, \eta)) \, q_\theta(x, dy) + F(x, \eta). \tag{5.3}$$

Under Assumptions 2.8, as $\theta \to \infty$,

$$b_\theta(x, \eta) \to \gamma(x, \eta) \mathcal{B} r(x, \eta) + F(x, \eta), \tag{5.4}$$

and $c_\theta(x,\eta)$ will tend to $\alpha\gamma(x,\eta)r(x,\eta)$.

We can then rewrite the differential equation governing the dynamics of the level of each individual as

$$\begin{aligned}\dot{u} &= \theta\gamma(x,\eta)\int_{\mathbb{R}^d} r(y,\eta)q_\theta(x,dy)\left\{-\frac{u}{N}(N-u)+u\right\} - b_\theta(x,\eta)u \\ &= c_\theta(x,\eta)u^2 - b_\theta(x,\eta)u. \end{aligned} \tag{5.5}$$

Now, we can write down the generator for $(\xi_t)_{t\geq 0}$, the lookdown process. In what follows, we will write sums (and, products) over "$(x,u)\in\xi$" to mean a sum over the (location, level) pairs of each individual in the population. Test functions for ξ will take the form

$$f(\xi) = \prod_{(x,u)\in\xi} g(x,u) = \exp\left(\int \log g(x,u)\xi(dx,du)\right), \tag{5.6}$$

where $g(x,u)$ is differentiable in u and smooth in x. We will also assume that $0\leq g(x,u)\leq 1$ for all $u\in[0,N]$, and $g(x,u)\equiv 1$ for $u\geq N$. In the expressions that follow, we shall often see one or more factor of $1/g(x,u)$; it should be understood that if $g(x,u)=0$, then it simply cancels the corresponding factor in $f(\xi)$.

First consider the terms in the generator that come from birth events. When a birth successfully establishes, a new level is generated above the parent's level, and this new level is assigned to either the offspring or the parent. Since the probability of each is $1/2$, the contribution of birth to the generator maps $f(\xi)$ to

$$f(\xi)\sum_{(x,u)\in\xi} 2\frac{\theta}{N}\gamma(x,\eta)\int_u^N\int_{\mathbb{R}^d}\left(\frac{1}{2}\left\{g(y,u_1)+\frac{g(y,u)g(x,u_1)}{g(x,u)}\right\}-1\right)r(y,\eta)q_\theta(x,dy)du_1 \tag{5.7}$$

$$\begin{aligned} = f(\xi)\sum_{(x,u)\in\xi} 2\gamma(x,\eta)\Bigg\{&\frac{1}{2N}\int_u^N g(x,u_1)du_1\frac{\theta\int_{\mathbb{R}^d}(g(y,u)-g(x,u))r(y,\eta)q_\theta(x,dy)}{g(x,u)} \\ &+\frac{\theta}{N}\int_u^N\int_{\mathbb{R}^d}\left(\frac{g(y,u_1)+g(x,u_1)}{2}-1\right)r(y,\eta)q_\theta(x,dy)du_1\Bigg\}. \end{aligned} \tag{5.8}$$

In (5.7), u_1 is the new level and y is the offspring's location, and so the two terms in the integral correspond to the two situations: in the first, we have added an individual at (y,u_1), while in the second, we replace an individual at (x,u) by one at (x,u_1) and another at (y,u). We've rewritten it in the form (5.8) because each of the two pieces naturally converges to a separate term in the limit.

The remaining term in the generator is due to the motion of particles' levels. Reading off from (5.5), it takes the form

$$f(\xi) \sum_{(x,u)\in\xi} \Big(c_\theta(x,\eta)u^2 - b_\theta(x,\eta)u\Big) \frac{\partial_u g(x,u)}{g(x,u)}. \tag{5.9}$$

We can now define the spatial-level process explicitly as a solution to a martingale problem, whose generator is just the sum of (5.8) and (5.9). We need some notation. Write $\mathcal{C} = \mathcal{C}(\mathbb{R}^d \times [0,\infty))$ for the counting measures on $\mathbb{R}^d \times [0,\infty)$ and $\mathcal{C}_N$ for the subset consisting of counting measures on $\mathbb{R}^d \times [0,N]$.

Definition 5.1 (Martingale Problem Characterisation) *For given positive values of N and θ, define the generator A^N by*

$$\begin{aligned} &A^N f(\xi) \\ &= f(\xi) \sum_{(x,u)\in\xi} 2\gamma(x,\eta)\Bigg\{ \frac{1}{2N}\int_u^N g(x,u_1)du_1 \frac{\theta \int_{\mathbb{R}^d}(g(y,u)-g(x,u))r(y,\eta)q_\theta(x,dy)}{g(x,u)} \\ &\qquad\qquad + \frac{\theta}{N}\int_u^N \int_{\mathbb{R}^d} \left(\frac{g(y,u_1)+g(x,u_1)}{2} - 1\right) r(y,\eta)q_\theta(x,dy)du_1 \Bigg\} \\ &\quad + f(\xi) \sum_{(x,u)\in\xi} \Big(c_\theta(x,\eta)u^2 - b_\theta(x,\eta)u\Big) \frac{\partial_u g(x,u)}{g(x,u)}, \end{aligned} \tag{5.10}$$

where $f(\xi) = \prod_{(x,u)\in\xi} g(x,u)$ is as defined in (5.6), and $\eta(\cdot) = \xi(\cdot \times [0,N])/N$ as before. Given $\xi_0 \in \mathcal{C}_N$, we say that a $\mathcal{D}_{[0,\infty)}(\mathcal{C}_N)$-valued process $(\xi_t)_{t\geq 0}$ is a solution to the (A^N, ξ_0) martingale problem if $f(\xi_t) - f(\xi_0) - \int_0^t A^N f(\xi_s)ds$ is a martingale (with respect to the natural filtration) for all test functions f as defined above.

The martingale problem for finite N has a unique solution. Next we state the limiting martingale problem, for which we do not necessarily have uniqueness. As before, the parameter α will correspond to $\lim_{N\to\infty} \theta(N)/N$. Whereas for finite N, conditional on the population process η_t^N, the levels of particles are independent and uniformly distributed on $[0,N]$, in the infinite population limit, conditional on η_t, the process ξ_t is Poisson distributed on $\mathbb{R}^d \times [0,\infty)$ with mean measure $\eta_t \times \lambda$, where λ is Lebesgue measure.

Definition 5.2 (Martingale Problem Characterisation, scaling limit) *Fix $\alpha \in [0,\infty)$, and define test functions f by $f(\xi) = \prod_{(x,u)\in\xi} g(x,u)$ with g differentiable in u, smooth in x, satisfying $0 \le g(x,u) \le 1$ and such that there exists a u_0 with $g(x,u) = 1$ for all $u > u_0$. Then, define the operator A on such test functions by*

$$\begin{aligned} Af(\xi) = f(\xi) &\sum_{(x,u)\in\xi} \gamma(x,\eta)\frac{\mathcal{B}(g(\cdot,u)r(\cdot,\eta))(x) - g(x,u)\mathcal{B}r(x,\eta)}{g(x,u)} \\ &+ f(\xi) \sum_{(x,u)\in\xi} 2\alpha\gamma(x,\eta)r(x,\eta)\int_u^\infty (g(x,u_1)-1)du_1 \\ &+ f(\xi) \sum_{(x,u)\in\xi} \left(\alpha\gamma(x,\eta)r(x,\eta)u^2 - \{\gamma(x,\eta)\mathcal{B}r(x,\eta) + F(x,\eta)\}\, u\right) \frac{\partial_u g(x,u)}{g(x,u)}. \end{aligned} \tag{5.11}$$

We say that a $\mathcal{D}_{[0,\infty)}(\mathcal{C})$-valued process $(\xi_t)_{t\ge 0}$ is a solution to the (A,ξ_0) martingale problem if it has initial distribution ξ_0 and $f(\xi_t) - f(\xi_0) - \int_0^t Af(\xi_s)ds$ is a martingale (with respect to the natural filtration) for all test functions f as defined above.

The lookdown processes have been carefully constructed so that observations about the past spatial positions of individuals in the population do not give us any information about the assignment of individuals to levels. In other words, the dynamics of the lookdown process preserve the conditionally uniform (or in the limit, conditionally Poisson) structure – if started with uniform levels, levels are uniform at all future times. Moreover, if we average over levels in the expression for the generator (equation (5.10) or (5.11)) we recover the generator for the population process. Once this is verified (along with some boundedness conditions) the Markov Mapping Theorem (Theorem A.1; also see [EK19]) tells us that by "removing labels" from the lookdown process ξ we recover the population process η.

To make this precise, define the spatial projection maps $\kappa^N : \mathcal{M}(\mathbb{R}^d \times [0,N]) \to \mathcal{M}(\mathbb{R}^d)$ by $\kappa^N(\xi^N)(\cdot) = \xi^N(\cdot \times [0,N])/N$, and $\kappa : \mathcal{M}(\mathbb{R}^d \times [0,\infty)) \to \mathcal{M}(\mathbb{R}^d)$ by $\kappa(\xi)(\cdot) = \lim_{u_0\to\infty} \xi(\cdot \times [0,u_0])/u_0$. We will also need an inverse notion: for a measure ξ^N on $\mathbb{R}^d \times [0,N]$ and a σ-field $\mathcal{F}$, we say that ξ^N *is conditionally uniform given* $\mathcal{F}$ if $\kappa^N(\xi)$ is $\mathcal{F}$-measurable and for all compactly supported f,

$$\mathbb{E}[e^{-\langle f,\xi\rangle} \mid \mathcal{F}] = e^{-\langle H_f^N, \kappa^N(\xi)\rangle}, \tag{5.12}$$

where

$$H_f^N(x) = -N \log \frac{1}{N} \int_0^N e^{-f(x,u)} du.$$

In other words, the $[0, N]$ components of ξ are independent, uniformly distributed on $[0, N]$, and independent of $\kappa^N(\xi)$. Similarly, for a measure ξ on $\mathbb{R}^d \times [0, \infty)$ we say that ξ *is a conditionally Poisson random measure given* $\mathcal{F}$ if $\kappa(\xi)$ is $\mathcal{F}$-measurable and for all compactly supported f,

$$\mathbb{E}[e^{-\langle f,\xi\rangle} \mid \mathcal{F}] = e^{-\langle \int_0^\infty (1-e^{-f(x,u)})du, \kappa(\xi)(dx)\rangle}. \tag{5.13}$$

In other words, ξ is conditionally Poisson with Cox measure $\kappa(\xi) \times \lambda$, where λ is Lesbegue measure.

Proposition 5.3 *If $\widetilde{\eta}^N$ is a solution of the martingale problem of Definition 2.4 with initial distribution η_0^N then there exists a solution ξ^N of the (A^N, ξ_0^N)-martingale problem of Definition 5.1 such that $\eta^N = \kappa^N \circ \xi^N$ has the same distribution on $D_{\mathcal{M}_F(\mathbb{R}^d)}[0, \infty)$ as $\widetilde{\eta}^N$. Furthermore, for each t, ξ_t^N is conditionally uniform given $\mathcal{F}_t^{\eta^N}$ in the sense of* (5.12).

Similarly, if $\widetilde{\eta}$ is a solution of the limiting martingale problem of Theorem 2.10 with initial distribution η_0 then there exists a solution ξ of the martingale problem of of Definition 5.2 such that $\eta = \kappa \circ \xi$ has the same distribution on $D_{\mathcal{M}_F(\mathbb{R}^d)}[0, \infty)$ as $\widetilde{\eta}$. Furthermore, ξ_t is conditionally Poisson given $\mathcal{F}_t^\eta$ in the sense of (5.13).

Now we can present the main convergence theorem that is analogous to Theorem 2.10 for the population process.

Theorem 5.4 *Let (ξ_t^N) satisfy Definition 5.1 and assume that as $N \to \infty$, $\theta \to \infty$ in such a way that $\theta/N \to \alpha$. Let $\eta_0^N = \kappa(\xi_0^N)$ and suppose also that $\eta_0^N \to \eta_0$ in $\mathcal{M}_F(\mathbb{R}^d)$, and that for each N, ξ_0^N is conditionally uniform given η_0^N in the sense of* (5.12). *Then, $(\xi_t^N)_{t\geq 0}$ has a subsequence which converges in distribution as $N \to \infty$ to a measure-valued process $(\xi_t)_{t\geq 0}$ with ξ_t conditionally Poisson given $\eta_t = \kappa(\xi_t)$ for each t in the sense of* (5.13), *that is a solution to the martingale problem of Definition 5.2.*

Both results are proved in Section 5.2.

5.1.2 Explicit construction of lines of descent

The main interest in using a lookdown construction for our population processes is that it allows us to retain information about the relatedness of individuals as we pass to the infinite population limit. In order to exploit this, in this section we write down stochastic equations for the locations and levels of individuals in the prelimiting lookdown model. We will then be able to pass to the scaling limit. This provides an explicit description of the solution to the limiting martingale problem of Definition 5.2 which will enable us to identify all individuals in the current population that are descendants of a given ancestor at time zero. In theory at least, this allows us to recover all the information about genealogies relating individuals sampled from the present day population. This idea draws on the notion of "tracers", popular in statistical physics and used in population genetics by a number of authors including [HN08], [DF16], and [BEK21].

We will construct the process using a Ulam-Harris indexing scheme. First, we assign each individual alive at time 0 a unique label from $\mathbb{N}$. Suppose an individual with label a and level u reproduces, and as a result there are two individuals, one with level u and one with a new level $u_1 > u$. The parent individual, previously labeled a, might be assigned either level. We will track chains of descendant individuals forwards through time by following levels, rather than individuals, and will call this a *line of descent*. So, after reproduction, we give a new label to *only* the individual that is given the new level u_1, retaining the label a for the individual with the old level u. In this way, at each birth event, a unique label is assigned to the resulting individual with the higher level, and the label of an individual may change throughout its lifetime.

Concretely, then: for each label a in $\mathcal{I} = \bigcup_{k\geq 1} \mathbb{N}^k$, let Π_a be an independent Poisson process on $[0,\infty)^2 \times \mathbb{R}^d \times \{0,1\}$. The mean measure of each Π_a is a product of Lebesgue measure on $[0,\infty)^2$, the density of the standard Gaussian on $\mathbb{R}^d$, and $(\delta_0 + \delta_1)/2$ on $\{0,1\}$. It will also be convenient to suppose that for each label a we have an enumeration of the points in Π_a, so we may refer to "the j^{th} point in Π_a", although the precise order of this enumeration is irrelevant. If (τ, v, z, κ) is the j^{th} point in Π_a, then τ will determine a possible birth time, v will determine the level of the offspring, z will determine the spatial displacement of the offspring relative to the parent, κ will be used to determine whether

parent or offspring is assigned the new level, and the new label produced will be $a \oplus j$, i.e., the label a with j appended (so, if $a = (a_1, \ldots, a_k)$ then $a \oplus j = (a_1, \ldots, a_k, j)$). Each label a has a birth time τ_a, when it is first assigned, and a (possibly infinite) death time σ_a, when its level first hits N. For any $\tau_a \le t \le \sigma_a$ we denote by $X_a(t)$ and $U_a(t)$ the spatial location and level of the individual carrying label a at time t, respectively. Furthermore, define

$$\eta_t^N = \frac{1}{N} \sum_{a:\tau_a \le t < \sigma_a} \delta_{X_a(t)} \qquad \text{and} \qquad \xi_t^N = \sum_{a:\tau_a \le t < \sigma_a} \delta_{(X_a(t), U_a(t))}.$$

Now, since we have defined labels so that the level does not jump, U_a satisfies (5.5) for $\tau_a \le t \le \sigma_a$, i.e.,

$$\begin{aligned} U_a(t) &= U_a(\tau_a) \\ &\quad + \int_{\tau_a}^t \Big(c_\theta(X_a(s), \eta_s) U_a(s)^2 - b_\theta(X_a(s), \eta_s) U_a(s) \Big)\, ds, \end{aligned} \tag{5.14}$$

and, of course, $\sigma_a = \inf\{t \ge \tau_a : U_a(t) > N\}$.

Potential reproduction events occur at times τ for each point $(\tau, v, z, \kappa) \in \Pi_a$ with $\tau_a \le \tau < \sigma_a$. (We say "potential" since if the level of the resulting offspring is greater than N, the event does not happen.) If this is the j^{th} point in Π_a, the potential new label is $a \oplus j$, the birth time is $\tau_{a \oplus j} = \tau$, and the spatial displacement of the potential offspring is $y(X(\tau-), z)$, where

$$y(x, z) := \frac{1}{\theta} \vec{b}(x) + \frac{1}{\sqrt{\theta}} K(x) z,$$

and $K(x) K^T(x) = \mathbf{C}(x)$.

Next we must choose the new level created at the birth event. We would like an individual with level u and at spatial position x to produce offspring at y at instantaneous rate

$$2\Big(1 - \frac{u}{N}\Big) \theta \gamma(x, \eta) r(x + y, \eta). \tag{5.15}$$

To do this we will associate the point $(\tau, v, z, \kappa) \in \Pi_a$ with level $u + v\ell$, where ℓ is chosen so that the rate of appearance of points in Π_a with level below N, that is points with $v\ell < N - u$, is given by (5.15). Since the mean measure of Π_a is Lebesgue measure in the t and v directions, we must take

$$\ell(x, y, \eta) = \frac{N - u}{2(1 - u/N)\theta \gamma(x, \eta) r(x + y, \eta)} = \frac{1}{2N^{-1} \theta \gamma(x, \eta) r(x + y, \eta)}, \tag{5.16}$$

and, using this, the (potential) new level is

$$U_{a\oplus j}(\tau) = U_a(\tau) + v\ell\Big(X_a(\tau-), y(X_a(\tau-), z), \eta_{\tau-}\Big). \tag{5.17}$$

If $U_{a\oplus j}(\tau) < N$, the new individual labeled $a \oplus j$ is produced, and κ determines which label, a or $a \oplus j$, is associated with the new location, so

$$X_{a\oplus j}(\tau) = X_a(\tau-) + (1-\kappa)y\Big(X_a(\tau-), z\Big).$$

On the other hand if $U_{a\oplus j}(\tau) \geq N$, then X_a is unchanged and $X_{a\oplus j}$ is undefined, so

$$X_a(\tau) = X_a(\tau-) + \kappa y(X_a(\tau-), z)\mathbf{1}_{U_{a\oplus j}(\tau)<N}. \tag{5.18}$$

Recall that the parental *individual* always retains their spatial location, so that $\kappa = 1$ corresponds to the parent being assigned a new level, and our line of descent switching to the offspring. Combining these observations, X_a, for $\tau_a \leq t < \sigma_a$, solves the equation

$$\begin{aligned} X_a(t) =& X_a(\tau_a) \\ &+ \int_{[\tau_a,t)\times[0,\infty)\times\mathbb{R}\times[0,1]} y(X_a(\tau-), z)\kappa\mathbf{1}_{U_a(\tau)+v\ell(X_a(\tau-),y(X_a(\tau-),z),\eta_{\tau-})<N} d\Pi_a(\tau, v, z, \kappa). \end{aligned}$$

Although we have described the evolution of a line of descent only for a given label (i.e., for $\tau_a \leq t < \sigma_a$), we can extend the definition to times $0 \leq t < \sigma_a$ by setting $X_a(t)$ equal to $X_{[a]_t}(t)$, where $[a]_t$ is the label of the ancestor of label a alive at time t, and similarly for $U_a(t)$. It is then straightforward, albeit tedious, to write down the time evolution of $(X_a(t), U_a(t))$ for all time back to $t = 0$ in terms of the driving Poisson processes.

Remark 5.5 *Although we have a single construction that couples the processes across all N, unlike in [KR11] the actual trajectories, $X_a(\cdot)$, do not necessarily coincide for different values of N, since they are affected by the whole population process. However, this does suggest approximating the genealogies in the infinite density limit by simulating up until a sufficiently high level that we have a good approximation to the population process.*

5.1.3 Limiting processes for lines of descent

The previous section constructed the lookdown process using the same underlying Poisson processes $\{\Pi_a\}_{a\in\mathcal{I}}$ for different values of N. As a result, if the spatial projections η converge, then individual lines of descent converge pointwise as $N \to \infty$. To see this, first note that if the Poisson processes are fixed then the set of events with which a given label $a \in \mathcal{I}$ is associated is also fixed – this is the sequence $(\tau_k, v_k, z_k, \kappa_k)$ associated with the label a. To conclude that the lines of descent converge, first, we clearly need that the spatial projections η converge. Supposing that they do, consider how a line of descent $(X_a(t), U_a(t))$ evolves. It throws off a new line of descent at a higher level when there is a point (τ, v, z, κ) in Π_a with $\tau > \tau_a$ and

$$v < 2\frac{\Big(N - U_a(\tau)\Big)}{N}\theta\gamma(X_a(\tau-), \eta_{\tau-})r\Big(X_a(\tau-) + y\Big(X_a(\tau-), z\Big), \eta_{\tau-}\Big). \tag{5.19}$$

Since the mean measure of the v coordinate is Lebesgue measure, $\theta/N \to \alpha$, and $q_\theta(x, dy) \to \delta_x(dy)$, this corresponds in the limit to new lines of descent being thrown off according to a Poisson process with intensity

$$2\alpha\gamma(X_a(t), \eta_t)r(X_a(t), \eta_t)dt \times du.$$

Now consider the location of the line of descent: at each birth event, with probability one half the line of descent jumps to $X_a(t) + y$. Taking g to be a suitable test function on $\mathbb{R}^d$, and rewriting (5.19), when the level is u and the state of the population is η, the generator of the spatial motion of the line of descent applied to $g(x)$ is

$$\begin{aligned}
&\left(1 - \frac{u}{N}\right)\gamma(x, \eta)\theta\int_{\mathbb{R}^d} r(x + y, \eta)(g(x + y) - g(x))q_\theta(x, dy) \\
&\quad = \left(1 - \frac{u}{N}\right)\gamma(x, \eta)\Bigg\{\theta\int_{\mathbb{R}^d}(r(x + y, \eta)g(x + y) - r(x, \eta)g(x))q_\theta(x, dy) \\
&\qquad\quad - \theta\int_{\mathbb{R}^d}(r(x + y, \eta) - r(x, \eta))g(x)q_\theta(x, dy)\Bigg\} \\
&\quad \to \gamma(x, \eta)\left(\mathcal{B}(rg)(x) - g(x)\mathcal{B}(r)(x)\right), \qquad \text{as } N, \theta \to \infty.
\end{aligned}$$

Notice that the factors of 2 have cancelled, and that the result is independent of u. Also recall that $r(x, \eta)$ depends on η only through $\rho_r * \eta(x)$, which is guaranteed to be smooth, so that $\mathcal{B}(r)$ and $\mathcal{B}(gr)$ are well-defined.

We write out the differential operator above in more detail. Recall that $\mathcal{B}g(x) = \sum_i \vec{b}_i \partial_i g(x) + \frac{1}{2}\sum_{ij} \mathbf{C}_{ij}\partial_{ij} g(x)$, and for the moment write $r(x)$ for $r(x,\eta)$, $\vec{b}(x) = \vec{b}$, and $\mathbf{C}(x) = \mathbf{C}$ so that

$$\begin{aligned}\mathcal{B}(rg)(x) - g(x)\mathcal{B}(r)(x) &= r(x)\sum_i \vec{b}_i \partial_i g(x) + \sum_{ij} \partial_i r(x)\mathbf{C}_{ij}\partial_j g(x) + r(x)\frac{1}{2}\sum_{ij}\mathbf{C}_{ij}\partial_{ij}g(x) \\ &= r(x)\left\{\left(\vec{b} + \mathbf{C}\nabla \log r(x)\right)\cdot \nabla g(x) + \frac{1}{2}\sum_{ij}\mathbf{C}_{ij}\partial_{ij}g(x)\right\}. \quad (5.20)\end{aligned}$$

The only thing that remains is to describe how the levels change, but this is immediate from applying limit (5.4) to equation (5.5).

We summarize the results in a proposition.

Proposition 5.6 (Line of descent construction) *Define $J(x,\eta)$ and $\beta(x,\eta)$ by*

$$\begin{aligned} r(x,\eta)\gamma(x,\eta)\mathbf{C}(x) &= 2J(x,\eta)J(x,\eta)^T \\ \beta(x,\eta) &= r(x,\eta)\gamma(x,\eta)\Big(\vec{b}(x) + \mathbf{C}(x)\nabla \log r(x,\eta)\Big).\end{aligned}$$

Associate with each label $a \in \mathcal{I} = \cup_{k\geq 1}\mathbb{N}^k$ an independent d-dimensional Brownian motion W_a and an independent Poisson process R_a on $[0,\infty)^2$ with Lebesgue mean measure, and with points ordered in some way. Given $\eta_0 \in \mathcal{M}_F(\mathbb{R}^d)$, let (x_i, u_i) be the points of a Poisson process on $\mathbb{R}^d \times [0,\infty)$ with mean measure $\eta_0 \times \lambda$ (the product of η_0 and Lebesgue measure). For each i, begin a line of descent with label i, location $X_i(0) = x_i$, level $U_i(0) = u_i$, and birth time $\tau_i = 0$.

Write τ_a for the birth time of the label a and $\sigma_a = \lim_{u_0\to\infty}\inf\{t \geq 0 : U_a(t) > u_0\}$ the time the level hits ∞. Finally, write $\eta_t = \lim_{u_0\to\infty}\frac{1}{u_0}\sum_{a:\tau_a\leq t<\sigma_a\ ;\ U_a(t)<u_0}\delta_{X_a(t)}$.

Suppose the spatial locations and level of each line of descent a solve, for $\tau_a \leq t < \sigma_a$,

$$\begin{aligned} X_a(t) &= X_a(\tau_a) + \int_{\tau_a}^t \beta(X_a(s),\eta_s)ds + \int_{\tau_a}^t J(X_a(s),\eta_s)dW_a(s) \\ U_a(t) &= U_a(\tau_a) + \int_{\tau_a}^t \Big(\alpha\gamma(X_a(s),\eta_s)r(X_a(s),\eta_s)U_a(s)^2 \\ &\qquad - \Big\{\gamma(X_a(s),\eta_s)\mathcal{B}r(X_a(s),\eta_s) + F(X_a(s),\eta_s)\Big\}U_a(s)\Big)ds. \end{aligned} \quad (5.21)$$

In other words, the evolution of the line of descents at time t depend on local density at its spatial location, measured by convolutions against the collection of lineages η_t.

Each point in each R_a denotes a potential birth time for a: if the j^{th} point in R_a is (τ, v), with $\tau_a \leq \tau < \sigma_a$, then a new line of descent with label $a \oplus j$ is produced, with birth time $\tau_{a\oplus j} = \tau$, location $X_{a\oplus j}(\tau) = X_a(\tau)$, and level

$$U_{a\oplus j}(\tau) = U_a(\tau) + \frac{v}{2\alpha\gamma(X_a(\tau), \eta_\tau)r(X_a(\tau), \eta_\tau)}, \tag{5.22}$$

if this is finite.

We write $(\widetilde{\eta}_s)_{s\geq 0}$, $(\widetilde{\xi}_s)_{s\geq 0}$ to be solutions to the martingale problems of Theorems 2.10 and 5.4, respectively. For any solution to the equations above, and for any fixed time $t \geq 0$,

$$\eta_t = \lim_{u_0\to\infty} \frac{1}{u_0} \sum_{a:\tau_a\leq t<\sigma_a \ ; \ U_a(t)<u_0} \delta_{X_a(t)} \qquad \text{and} \qquad \xi_t = \sum_{a:\tau_a\leq t<\sigma_a} \delta_{(X_a(t),U_a(t))}$$

has the same distribution as $\widetilde{\eta}_t, \widetilde{\xi}_t$.

In particular, note that if $\alpha = 0$, no new lines of descent are produced. More precisely, comparing with (5.16), they are produced, but "at infinity", and their trace is seen in the spatial motion of the line of descent which results from the production of these lineages.

Proof [of Proposition 5.6] The fact that a solution to the system of equations (5.21) is a solution to the martingale problem of Theorem 5.4 is an application of Itô's theorem. Furthermore, in Proposition 5.3 we showed that the conditional Poisson property of ξ_0 is preserved (i.e., holds for ξ_t for all t), and so for arbitrary $t > 0$, η_t almost surely exists and has the same distribution as $\widetilde{\eta}_t$, $\widetilde{\xi}_t$. □

Remark 5.7 *Although we know that η_t exists almost surely for fixed time $t > 0$, additional work is needed to ensure this holds simultaneously for all $t \geq 0$, such that*

$$(\eta(t))_{t\geq 0}, \ (X_a(t))_{t\geq 0}, \ (U_a(t))_{t\geq 0}$$

are well-defined as stochastic processes. To do so, we first note that in the pre-limit, i.e. for finite $n > 0$, the processes

$$(\eta^n(t))_{t\geq 0}, \ (X^n_a(t))_{t\geq 0}, \ (U^n_a(t))_{t\geq 0}$$

are all well-defined, and in particular, by the first half of Proposition 5.3 and Theorem 2.10, $(\eta^n(t))_{t\geq 0}$ converges in distribution along subsequences to the process $(\eta^n(t))_{t\geq 0}$. Therefore,

by the Skorokhod's representation theorem, there exists a common probability space $(\Omega_1, \mathbb{P}_1, \mathcal{F}_1)$ *and a subsequence* $(n_k)_{k\geq 1}$ *such that* $(\eta^{n_k}(t))_{t\geq 0} \to (\eta(t))_{t\geq 0}$ *almost surely as measure-valued stochastic processes.*

Next, the processes $(X_a^{n_k}(t), U_a^{n_k}(t))_{t\geq 0}$ *defined through Equations* (5.14), (5.18) *(with implicit dependence on* N *through* σ_a*) are tight given the control over the processes* $\{(\eta^{n_k}(t))_{t\geq 0} : n_k \geq 0\}$*, and once again there exists a sub-subsequence* $\{n_{k_j} : j \geq 1\}$ *and a further Skorokhod embedding* $(\Omega_a, \mathbb{P}_a, \mathcal{F}_a)$ *such that* $(X_a^{n_{k_j}}(t), U_a^{n_{k_j}}(t))_{t\geq 0} \to (X_a(t), U_a(t))_{t\geq 0}$ *almost surely as stochastic processes. In particular,* $(\eta(t))_{t\geq 0}$ *is well-defined in* $(\Omega_a, \mathbb{P}_a, \mathcal{F}_a) \supset (\Omega_1, \mathbb{P}_1, \mathcal{F}_1)$ *so Equation* (5.21) *is well-defined.*

Finally, $(\xi(t))_{t\geq 0}$ *is well-defined in the product space* $\prod_a \Omega_a$ *with the product measure.*

Proofs of further results on ancestral lineages are in Section 5.2.

Remark 5.8 *The process we consider is similar to the state-dependent branching processes of [KR11], so one might expect that the proofs there would carry over with little change. However, there is an important difference: Recall that the level* $U_a(t)$ *of a line of descent evolves as*

$$\dot{u} = c_\theta(x, \eta)u^2 - b_\theta(x, \eta)u, \tag{5.23}$$

where $b_\theta(x, \eta)$ *and* $c_\theta(x, \eta)$ *are defined in* (5.3) *and* (5.2) *respectively. Note that* $c_\theta(x, \eta) \geq 0$*, while* $b_\theta(x, \eta)$ *may take either sign. Assumptions 2.8 imply that* $c_\theta(x, \eta)$ *is bounded, while* $b_\theta(x, \eta)$*, because of* $F(x, \eta)$*, is bounded above but not necessarily below. In [KR11],* b_θ *was bounded above and* c_θ *was bounded away from zero, so they noted that if* $U_a(t) \geq b_\theta/c_\theta$ *for some label* a*, that line of descent would only move upwards from that time onwards. Furthermore, coefficients did not depend on the state of the process (i.e., on* η*), thus allowing the processes to be jointly and simultaneously constructed for all values of* N*, with a pointwise embedding of* $(\xi_t^N)_{t\geq 0}$ *within* $(\xi^M)_{t\geq 0}$ *for* $b_\theta/c_\theta < N < M$*. In other words, individuals with levels above* $N > b_\theta/c_\theta$ *at time* t_0 *do not affect* $(\xi_t^N)_{t\geq t_0}$*, thus allowing a comparison of the number of lines of descent below level* u_0 *to a branching process. Although we have provided a joint construction of* ξ^N *for all* N *in Section 5.1.2, it does not have this monotonicity: for one thing,* b_θ *and* c_θ *depend on the population process* η *and so all individuals can affect all other ones (even those with lower levels). Furthermore,*

in the deterministic case θ/N, and hence c, converges to zero, and so lines of descent with arbitrarily high level may drift back downwards. Indeed, this must be the case if the population persists, since in the deterministic case there is no branching.

5.2 Proofs of results for the lookdown process and ancestral lineages

Now we turn to results about the lookdown process, first establishing the basic connection between the population process η^N and the lookdown process ξ^N, Proposition 5.3, and then in the next section, convergence of the lookdown process itself.

Proof [of Proposition 5.3] This proposition is the content of the Markov Mapping Theorem, reproduced from [EK19] as Theorem A.1, applied to our situation. The function γ of that theorem is what we have called κ above, and the kernel α of that theorem is the transition function that assigns levels uniformly on $[0, N]$ (in the first case) or as a Poisson process with Lesbegue intensity (in the limiting case). We need a continuous $\psi^N(\xi) \geq 1$ such that $|A^N f(\xi)| \leq c_f \psi^N(\xi)$ for all f in the domain of A^N (and similarly a function ψ for A). We also need that applying the lookdown generator to a function and averaging over levels is equivalent to applying the population process generator to the function whose dependence on levels has been averaged out, a condition which we precisely state, and verify, in Lemmas A.2 and A.3 of the Appendix.

For finite N, taking $f(\xi)$ of the form (5.6), we can use

$$\psi^N(\xi) = \langle C(1 + u|F(x, \eta)|), \xi(dx, du)\rangle$$

for an appropriate constant C. For the scaling limit, recall that the test functions f are of the form $f(\xi) = \prod_{(x,u)\in\xi} g(x, u)$ with $g(x, u) = 1$ for $u \geq u_0$, and consulting (5.11), we see that most terms in $Af(\xi)$ can be bounded as above by constant multiples of $\langle 1, \eta\rangle$. However, the term involving F is, as usual, more troublesome. Since $0 \leq f(\xi)/g(x, u) \leq 1$ for any $(x, u) \in \xi$,

$$\begin{aligned}\Big|f(\xi) \sum_{(x,u)\in\xi} F(x, \eta)u\frac{\partial_u g(x, u)}{g(x, u)}\Big| &\leq \|\partial_u g\|_\infty \sum_{(x,u)\in\xi} |F(x, \eta)u\mathbf{1}_{u\leq u_0}| \\ &\leq \|\partial_u g\|_\infty e^{u_0} \sum_{(x,u)\in\xi} |F(x, \eta)ue^{-u}|.\end{aligned}$$

The first line would be just what we want, except that $\psi(\xi)$ cannot depend on f, and hence neither on u_0. So, the second line provides us with the required bound: we absorb $\|\partial_u g\|_\infty e^{u_0}$ into c_f and take $\psi(\xi) = 1 + \langle 1 + F(x,\eta)ue^{-u}, \xi(dx,du)\rangle$. □

5.2.1 Tightness of the Lookdown Process

Now we turn to the main theorem on convergence of the lookdown process, Theorem 5.4, whose proof follows a similar pattern to that of convergence for the population processes in Section 3.2.

We first give a description of the lookdown process ξ^N in terms of the lines of descent introduced in Section 5.1.2. Each line of descent gives birth to lines at higher levels at rate $2(N-u)c_\theta(x,\eta)$, and each such new line chooses a level uniformly from $[u,N]$, a spatial location y from the kernel

$$q^m(x,dy,\eta) = \frac{r(y,\eta)q(x,dy)}{\int_{\mathbb{R}^d} r(y,\eta)q(x,dy)}, \tag{5.24}$$

and the two lines swap spatial locations with probability 1/2; the level of each line of descent evolves according to equation (5.23).

It is evident from the description of the process (or, by differentiating in Definition 5.1) that

$$\begin{aligned}\langle f, \xi_t^N\rangle =& \langle f, \xi_0^N\rangle + M_t^f \\ &+ \int_0^t \Bigg\langle c_\theta(x,\eta_s^N) \int_u^N \int_{\mathbb{R}^d} \left(f(y,u_1) + f(x,u_1) + f(y,u) - f(x,u)\right) q^m(x,dy,\eta_s^N)du_1 \\ &\qquad + \left(c_\theta(x,\eta_s^N)u^2 - b_\theta(x,\eta_s^N)u\right)\frac{d}{du}f(x,u), \xi_s^N(dx,du)\Bigg\rangle ds,\end{aligned} \tag{5.25}$$

where M^f is a martingale with angle bracket process

$$\begin{aligned}\left\langle M^f\right\rangle_t = \int_0^t \Bigg\langle c_\theta(x,\eta_s^N) \int_u^N \int_{\mathbb{R}^d} \Big[&f(y,u_1)^2 \\ &+ \left(f(x,u_1) + f(y,u) - f(x,u)\right)^2\Big] du_1 q^m(x,dy,\eta_s^N), \xi_s^N(dx,du)\Bigg\rangle ds.\end{aligned} \tag{5.26}$$

Remark 5.9 *In addition to tightness of the measure-valued processes ξ^N, the bounds used in the proofs below also imply tightness of the number of lines of descent and the number*

of births below a fixed level, and of the motion of individual lines of descent. In other words, the limiting "line of descent" construction of Section 5.1.2 holds.

Proof [Proof of Theorem 5.4] As in Section 3.2, the theorem will follow from tightness and characterization of the limit points. This time, the processes ξ^N take values in $\mathcal{M}(\overline{\mathbb{R}^d} \times [0, \infty))$, the space of locally finite measures on space $\times$ levels. (They will in fact be point measures, including the limit, but that is a consequence of this theorem.) Again, tightness follows from a compact containment condition, tightness of one-dimensional distributions, and an application of [EK86] Theorem 3.9.1.

Lines of descent can escape to infinite level in finite time, and so we endow $\mathcal{M}(\overline{\mathbb{R}^d} \times [0, \infty))$ with the vague topology "in the level coordinate", induced by test functions on $\overline{\mathbb{R}^d} \times [0, \infty)$ of the form $g(x)h(u)$, where $g \in C_b(\overline{\mathbb{R}^d})$ is bounded and continuous and $h \in C_c([0, \infty))$ is compactly supported (following, e.g., [EK19], Condition 2.1). In several places below we require a dense subset of $C_b(\mathcal{M}(\overline{\mathbb{R}^d} \times [0, \infty)))$, the bounded, continuous functions on $\mathcal{M}(\overline{\mathbb{R}^d} \times [0, \infty))$. The functions $\xi \mapsto \exp(-\langle f, \xi \rangle)$ with $f \geq 0$ do not form not a dense subset of $C_b(\mathcal{M}(\overline{\mathbb{R}^d} \times [0, \infty)))$, but they do separate points and vanish nowhere, i.e., for any ξ_1 and ξ_2 there is an f with $\exp(-\langle f, \xi_1 \rangle) \neq \exp(-\langle f, \xi_2 \rangle)$, and a g such that $\exp(-\langle g, \xi_1 \rangle) \neq 0$. Therefore, by the Stone-Weierstrass theorem, the algebra they generate is dense in $C_b(\mathcal{M}(\overline{\mathbb{R}^d} \times [0, \infty)))$ (with respect to uniform convergence on compact subsets). Topologized in this way, the space $\mathcal{M}(\overline{\mathbb{R}^d} \times [0, \infty))$ is completely metrizable, and we may choose a countable set of nonnegative f_k, each supported on $\mathbb{R}^d \times [0, u_k]$ for some $u_k < \infty$, such that a subset $K \subset \mathcal{M}(\overline{\mathbb{R}^d} \times [0, \infty))$ is relatively compact if and only if $\sup_{\xi \in K} \langle f_k, \xi \rangle < \infty$ for each k. (To see this, use Theorem A.2.3 of [Kal97] and the first argument made in the proof of Lemma A.7.) Below, Lemma 5.12 proves exactly this, and therefore compact containment. Here we have compactified $\mathbb{R}^d$ for convenience (since it turned out to be straightforward to show that mass does not escape to infinity in space); however, we need to use the vague topology "in the level direction" because *levels* may escape to infinity in finite time in the limit.

In order to apply [EK86] Theorem 3.9.1 we require that $\{(F(\xi_t^N))_{t \geq 0}\}_N$ is tight as a sequence of real-valued càdlàg processes, for all F in a subset of $C_b(\mathcal{M}(\overline{\mathbb{R}^d} \times [0, \infty)))$ that is dense with respect to uniform convergence on compact subsets. Lemma 5.13 shows that

$\{\langle f, \xi_t^N\rangle\}_N$ is a tight sequence for any $f : \overline{\mathbb{R}^d} \times [0,\infty) \to \mathbb{R}$ with compact support in the level direction. Since as above the algebra generated by the functions $\xi \mapsto e^{-\langle f,\xi\rangle}$ is dense in $C_b(\mathcal{M}(\overline{\mathbb{R}^d} \times [0,\infty)))$, it suffices to show that tightness for the processes $\langle f, \xi_t^N\rangle$ extends to finite sums and products of these processes, which is shown in Lemma A.7. The fact that martingale properties are preserved under passage to the limit is straightforward, and can be proved in a way analogous to Lemma 3.6; we omit the proof. Finally, we must show that the limiting lookdown process ξ projects to the limiting process η, i.e., a solution of the martingale problem in Theorem 2.10. Let $N_k \to \infty$ be a sequence along which ξ^{N_k} converges. By Theorem 2.10, there is a subsequence $N_{k(j)}$ along which the projected population processes $\eta^{N_{k(j)}}$ converge, and the limit solves the martingale problem. Thus any limit point of ξ^N projects to a population process η solving the martingale problem of Theorem 2.10. □

What we need for compact containment will come from the following Lemma. The generality is unimportant – for concreteness one may take $h(u) = e^{-u}$.

Lemma 5.10 *Let h be a positive, continuous, nonincreasing, differentiable function on $[0,\infty)$ such that $\int_0^\infty \int_u^\infty h(v)dvdu$, $\int_0^\infty u^2|h'(u)|du$, and $\int_0^\infty h(u)^2du$ are all finite. Suppose that Assumptions 2.8 hold, and that $\theta/N \to \alpha$ and $\xi_0^N \to \xi_0$ weakly as $N \to \infty$, where each ξ_0^N is conditionally uniform given η_0^N in the sense of* (5.12) *and ξ_0 is conditionally Poisson given η_0 in the sense of* (5.13)*. Then for any T there exists a constant $K(T)$ such that for all $M > 0$,*

$$\limsup_{N\to\infty} \mathbb{P}\left\{\sup_{0\le t\le T} \langle h, \xi_t^N\rangle > M\right\} < \frac{K(T)}{M}.$$

We postpone the proof of this Lemma until we have shown how it yields compact containment. First, we show that this implies compact containment of the processes $(\langle f, \xi_t^N\rangle)_{0\le t\le T}$ for arbitrary compactly supported f.

Lemma 5.11 *Suppose $f \in C(\overline{\mathbb{R}}^d \times [0,\infty))$ and there is a u_f such that if $u \ge u_f$ then $\sup_x f(x,u) = 0$. Under the assumptions of Lemma 5.10, for any T there exists a constant $K(f,T)$ such that for all $M > 0$,*

$$\limsup_{N\to\infty} \mathbb{P}\left\{\sup_{0\le t\le T} \langle f, \xi_t^N\rangle > M\right\} < \frac{K(f,T)}{M}.$$

Proof [of Lemma 5.11] Let h be as in Lemma 5.10, so there is a $c_f < \infty$ such that $f(x,u) \le c_f h(u)$ for all x and u. Therefore, $\langle f, \xi \rangle \le \langle h, \xi \rangle$, and so by Lemma 5.10,

$$\limsup_{N\to\infty} \mathbb{P}\left\{ \sup_{0\le t\le T} \langle f, \xi_t^N \rangle > M \right\} \le \limsup_{N\to\infty} \mathbb{P}\left\{ \sup_{0\le t\le T} \langle h, \xi_t^N \rangle > M/c_f \right\} < \frac{K(T)c_f}{M}.$$

□

Lemma 5.12 (Compact containment for $\boldsymbol{\xi}$) *Let $f_1, f_2, \ldots$ be a sequence of functions each satisfying the conditions of Lemma 5.11. Under the assumptions of Lemma 5.10, for any T and $\delta > 0$ there exists a sequence $(C_1, C_2, \ldots)$ of finite constants such that*

$$\limsup_{N\to\infty} \mathbb{P}\left\{ \sup_{0\le t\le T} \langle f_k, \xi_t^N \rangle > C_k \text{ for some } k \ge 1 \right\} < \delta. \tag{5.27}$$

In other words, the processes ξ^N stay in the set

$$\left\{ \xi \in \mathcal{M}(\overline{\mathbb{R}^d} \times [0,\infty)) \ : \ \langle f_k, \xi \rangle \le C_k \text{ for all } k \ge 1 \right\},$$

for all $0 \le t \le T$ with uniformly high probability, a set which (as discussed in the proof of Theorem 5.4) is relatively compact for an appropriate choice of $\{f_k\}_{k\ge1}$.

Proof [of Lemma 5.12] By a union bound,

$$\mathbb{P}\left\{ \sup_{0\le t\le T} \langle f_k, \xi_t^N \rangle > C_k \text{ for some } k \ge 1 \right\} \le \sum_{k\ge1} \mathbb{P}\left\{ \sup_{0\le t\le T} \langle f_k, \xi_t^N \rangle > C_k \right\},$$

so (5.27) follows by taking $C_k = 2^{k-1} K(f_k, T)/\delta$ and using Lemma 5.11. □

Finally, we prove the key lemma.

Proof [of Lemma 5.10] Applied to $f(x,u) = h(u)$, the martingale representation (5.25) is

$$\begin{aligned}\langle h, \xi_t^N \rangle = \langle h, \xi_0^N \rangle + M_t^h \\ + \int_0^t \Big\langle 2c_\theta(x, \eta_s^N) \int_u^N h(v)dv, \xi_s^N(dx,du) \Big\rangle ds \\ + \int_0^t \Big\langle \left(c_\theta(x,\eta_s^N)u^2 - b_\theta(x,\eta_s^N)u \right) h'(u), \xi_s^N(dx,du) \Big\rangle ds,\end{aligned}$$

where M_t^h is a martingale with angle bracket process

$$\left\langle M^h \right\rangle_t = \int_0^t \langle 2c_\theta(x,\eta_s^N) \int_u^N h(v)^2 dv, \xi_s^N(dx,du) \rangle ds.$$

Now, note that $0 \leq c_\theta(x, \eta_x^N) \leq C_a < \infty$ and $b_\theta(x, \eta_s^N) \leq C_b < \infty$, and we have assumed that $h'(u) \leq 0$ (since h is nonincreasing), so we may bound

$$\begin{aligned}\langle h, \xi_t^N \rangle \leq \langle h, \xi_0^N \rangle + M_t^h \\ + \int_0^t \left\langle 2C_a \int_u^\infty h(v)dv + \left(C_a u^2 + C_b u\right) |h'(u)|, \xi_s^N(dx, du) \right\rangle ds.\end{aligned} \tag{5.28}$$

Now, since ξ_t^N is conditionally uniform given η_t^N in the sense of (5.12), we know that for compactly supported f, $\mathbb{E}[\langle f, \xi_t^N \rangle] = \mathbb{E}[\langle \widetilde{f}_N, \eta_t^N \rangle]$, where $\widetilde{f}_N(x) = \int_0^N f(x,u)du$. By our assumptions on h, we know that

$$\int_0^\infty \left(2C_a \int_u^\infty h(v)dv + \left(C_a u^2 + C_b u\right) |h'(u)|\right) du < C$$

for some $C < \infty$, and so (by dominated convergence)

$$\mathbb{E}\left[\langle h, \xi_t^N \rangle\right] \leq \mathbb{E}\left[\langle h, \xi_0^N \rangle\right] + C \int_0^t \mathbb{E}\left[\left\langle 1, \eta_s^N \right\rangle\right] ds,$$

which we know by Lemma 3.1 is bounded by $C_0 e^{C_1 t}$ for some other constants C_0 and C_1.

Now consider the maximum. By (5.28), using that the integrand is nonnegative,

$$\begin{aligned}\sup_{0\leq t\leq T} \langle h, \xi_t^N \rangle \leq \langle h, \xi_0^N \rangle + \sup_{0\leq t\leq T} M_t^h \\ + \int_0^T \left\langle 2C_a \int_u^\infty h(v)dv + \left(C_a u^2 + C_b u\right) |h'(u)|, \xi_s^N(dx, du) \right\rangle ds.\end{aligned}$$

As in the proof of Lemma 3.2, the Burkholder-Davis-Gundy inequality, [BJY86], and the fact that $\sqrt{x} \leq 1 + x$ for $x \geq 0$ tells us that there is a C' such that

$$\begin{aligned}\mathbb{E}\left[\sup_{0\leq t\leq T} M_t^h\right] &\leq C'\left(1 + \mathbb{E}\left[\langle M^h \rangle_T\right]\right) \\ &\leq C'\left(1 + \int_0^T \mathbb{E}\left[\langle 2c_\theta(x, \eta_s^N) \int_u^\infty h(v)^2 dv, \xi_s^N(dx, du)\rangle\right] ds\right) \\ &\leq C'\left(1 + 2C_a \int_0^\infty h(v)^2 dv \int_0^T \mathbb{E}\left[\langle 1, \xi_s^N(dx, du)\rangle\right] ds\right) \\ &\leq C_2 e^{C_1 T},\end{aligned}$$

for a constant C_2 which is finite by our assumption that $\int_0^\infty h(v)^2 dv < \infty$.

Therefore,

$$\mathbb{E}\left[\sup_{0\leq t\leq T} \langle h, \xi_t^N \rangle\right] \leq \mathbb{E}\left[\langle h, \xi_0^N \rangle\right] + (C_2 + C_0/C_1) e^{C_1 T},$$

and so

$$\mathbb{P}\left\{\sup_{0\le t\le T}\langle h,\xi_t^N\rangle > K\right\} \le \frac{\mathbb{E}\left[\langle h,\xi_0^N\rangle\right] + (C_2 + C_0/C_1)e^{C_1 T}}{K}.$$

□

Lemma 5.13 *Let f be a bounded, continuous real-valued function on $\mathbb{R}^d \times [0,\infty)$ with uniformly bounded first and second derivatives for which there exists a u_0 such that if $u > u_0$ then $f(x,u) = 0$. Then, the sequence of real-valued processes $(\langle f,\xi_t^N\rangle)_{t\ge 0}$ for $N \ge 1$ is tight in $\mathcal{D}_{[0,\infty)}(\mathbb{R})$.*

Proof [of Lemma 5.13] Again, we use the Aldous-Rebolledo criterion. Tightness of $\langle f,\xi_t\rangle$ for a fixed t follows from Lemma 5.11, so we need only prove conditions analogous to (3.9) and (3.10) applied to the martingale representation of equations (5.25) and (5.26). Rewriting (5.25) with $c_\theta = c_\theta(x,\eta_s)$,

$$\begin{aligned}\langle f,\xi_t\rangle = \langle f,\xi_0\rangle + M_t^f + \int_0^t \Big\langle c_\theta \int_u^N \int (f(y,u_1)+f(x,u_1))q^m(x,dy,\eta)du_1 \\ + c_\theta(N-u)\int_0^t (f(y,u)-f(x,u))q^m(x,dy,\eta) + (c_\theta u^2 - b_\theta u)\frac{d}{du}f,\xi_s\Big\rangle ds.\end{aligned}$$

The bounds analogous to (3.9) and (3.10) follow as in the proof of Lemma 3.3: for instance, observe that using that $c_\theta \le C_a$ for some C_a, the predictable part of this semimartingale decomposition is bounded by

$$\left\langle 2C_a\|f\|_\infty u_f + (1-u/N)\gamma B_f^\theta + (C_a u^2 - b_\theta u)\frac{d}{du}f,\xi_s\right\rangle,$$

the last term of which is bounded by

$$\langle C_a u_f^2 + \sup_x |b_\theta(x,\eta_s)|u_f\|\frac{d}{du}f\|_\infty\rangle,$$

which can be bounded as we did for (3.9). □

5.2.2 Motion of ancestral lineages

In this section we prove Theorem 2.24. The argument follows directly from the discussion in Section 5.1.3.

Proof [of Theorem 2.24] For brevity, in the proof we write $\gamma(x)$ or γ for $\gamma(x, \eta)$.

Here we have taken the high-density, deterministic limit (so, $\theta, N \to \infty$ and $\theta/N \to 0$). We first proceed informally, as if the limiting process has a density $\varphi_t(x)$ at location x and time t (which it may not), and follow this with an integration against test functions to make the argument rigorous. Let Y denote the spatial motion followed by a single line of descent. Above equation (5.20), we showed that Y is a diffusion with generator at time s

$$\mathcal{L}_s^Y g(x) = \gamma(x, \eta_s)(\mathcal{B}(r(\cdot, \eta_s)g(\cdot))(x) - g(x)\mathcal{B}r(x, \eta_s)).$$

The diffusion is time-inhomogeneous if the density is not constant in time. Let $\varphi_t(x)$ be the limiting density, which is a weak solution to (2.1), $\partial_t \varphi_t = r\mathcal{B}^*[\varphi_t \gamma] + \varphi_t F$. Formally, the intensity of individuals at y at time t that are descended from individuals that were at x at time s (with $s < t$) is

$$\varphi_s(x)\mathbb{E}_{s,x}\left[\exp\left(\int_s^t (F + \gamma\mathcal{B}r)(Y_u)du\right)\mathbf{1}_{Y_t=y}\right]dy, \tag{5.29}$$

where the subscript s, x in the expectation indicates that $Y_s = x$. To see why this should be true, suppose that an ancestor at time s has level v. Conditional on its spatial motion $\{Y_u\}_{s\le u\le t}$, its level at time t will be $v\exp(-\int_s^t (F + \gamma\mathcal{B}r)(Y_u)du)$. This will be less than a given level λ if $v < \lambda\exp(\int_s^t (F + \gamma\mathcal{B}r)(Y_u)du)$. The intensity of levels at y that are descended from individuals at x can therefore be obtained as the limit as $\lambda \to \infty$ of $1/\lambda$ times the number of levels at x at time s with $u < \lambda\exp(\int_s^t (F+\gamma\mathcal{B}r)(Y_u)du)$ and for which the corresponding individual is at y at time t, which is precisely the quantity in (5.29).

By our construction in Section 5.1.3, when we integrate (5.29) with respect to x we recover $\varphi_t(y)dy$. Consider an individual sampled at location y at time t, and write $p(t, s, y, x)$ for the probability density that their ancestor at time s was at x. As a consequence of (5.29), still formally,

$$p(t, s, y, x) = \frac{\varphi_s(x)}{\varphi_t(y)}\mathbb{E}_{s,x}\left[\exp\left(\int_s^t (F + \gamma\mathcal{B}r)(Y_u)du\right)\mathbf{1}_{Y_t=y}\right] \qquad \text{for } s < t. \tag{5.30}$$

To make (5.30) meaningful, we multiply by suitable test functions f and g and integrate.

$$\int\int f(y)\varphi_t(y)p(t,s,y,x)g(x)dydx$$
$$= \int g(x)\varphi_s(x)\mathbb{E}_{x,s}\left[\exp\left(\int_s^t (F+\gamma\mathcal{B}r)(Y_u)du\right)f(Y_t)\right]dx.$$

Writing $\widehat{T}_{t,s}$ for the time-inhomogeneous semigroup corresponding to the motion of ancestral lineages backwards in time (that is, $\widehat{T}_{t,s}f(y) = \int p(t,s,x,y)f(x)dy$), we can write this as

$$\int f(y)\varphi_t(y)\widehat{T}_{t,s}g(y)dy = \int g(x)\varphi_s(x)\mathbb{E}_{s,x}\left[\exp\left(\int_s^t (F+\gamma\mathcal{B}r)(Y_u)du\right)f(Y_t)\right]dx. \quad (5.31)$$

Next, we will differentiate this equation with respect to t. There are two terms in the product on the left-hand side that depend on t, so if we use that $\partial_t\varphi_t = r\mathcal{B}^*[\varphi_t\gamma] + \varphi F$ (in a weak sense), and write $\mathcal{L}_u$ for the generator of $\widehat{T}_{t,s}$ at time $t = u$ so that $\partial_t\widehat{T}_{t,s}g(y)\Big|_{t=s} = \mathcal{L}_s g(y)$, then

$$\frac{d}{dt}\int f(y)\varphi_t(y)\widehat{T}_{t,s}g(y)dy\Big|_{t=s}$$
$$= \int f(y)\left\{\varphi_s(y)\mathcal{L}_s g(y) + [r(y)\mathcal{B}^*(\gamma\varphi_s)(y) + \varphi_s(y)F(y)]\,g(y)\right\}dy.$$

As for the right-hand side, since $Y_s = x$ under $\mathbb{E}_{x,s}$,

$$\frac{d}{dt}\mathbb{E}_{x,s}\left[\exp\left(\int_s^t (F+\gamma\mathcal{B}r)(Y_u)du\right)f(Y_t)\right]\Bigg|_{t=s} = [F(x)+\gamma(x)\mathcal{B}r(x)]\,f(x) + \mathcal{L}_s^Y f(x).$$

Therefore, the derivative of (5.31) (with respect to t, evaluated at $t = s$) is

$$\int f(y)\left\{\varphi_s(y)\mathcal{L}_s g(y) + (r(y)\mathcal{B}^*(\gamma\varphi_s)(y) + \varphi_s(y)F(y))\,g(y)\right\}dy$$
$$= \int g(x)\varphi_s(x)\left(\mathcal{L}_s^Y f(x) + [F(x)+\gamma(x)\mathcal{B}r(x)]\,f(x)\right)dx$$
$$= \int f(x)\left((\mathcal{L}_s^Y)^*(\varphi_s g)(x) + [F(x)+\gamma(x)\mathcal{B}r(x)]\,\varphi_s(x)g(x)\right)dx,$$

where $(\mathcal{L}_s^Y)^*$ is the adjoint of $\mathcal{L}_s^Y$. Since f was arbitrary,

$$\mathcal{L}_s g = \frac{1}{\varphi_s}\left[(\mathcal{L}_s^Y)^*(\varphi_s g) + \gamma\varphi_s g\mathcal{B}(r) - rg\mathcal{B}^*(\gamma\varphi_s)\right].$$

(Note that the $\varphi_s Fg$ terms have cancelled.) Since the adjoint of $\mathcal{L}_s^Y$ is

$$(\mathcal{L}_s^Y)^* f = r\mathcal{B}^*(\gamma f) - \gamma f\mathcal{B}r,$$

we can rewrite the generator of a lineage as

$$\mathcal{L}_s g \;=\; \frac{r}{\varphi_s}\left[\mathcal{B}^*(\gamma\varphi_s g) - g\mathcal{B}^*(\gamma\varphi_s)\right].$$

This is equation (2.22).

To simplify to equation (2.23), first define $\mathcal{D}f(x) = \frac{1}{2}\sum_{ij}\mathbf{C}_{ij}\partial_{ij}f(x)$, and so the adjoint of $\mathcal{D}$ is

$$\mathcal{D}^* f(x) = \frac{1}{2}\sum_{ij}\partial_{ij}(\mathbf{C}_{ij}f(x)).$$

Note that $\mathcal{D}^*$ satisfies the following identity:

$$\begin{aligned}\mathcal{D}^*(fg) &= \frac{1}{2}\sum_{ij}\left\{g\partial_{ij}(\mathbf{C}_{ij}f) + 2f\partial_i(\mathbf{C}_{ij})\partial_j(g) + 2\mathbf{C}_{ij}\partial_i(f)\partial_j(g) + \mathbf{C}_{ij}f\partial_{ij}g\right\}\\ &= g\mathcal{D}^* f + f\vec{c}\cdot\nabla g + (\mathbf{C}\nabla f)\cdot\nabla g + f\mathcal{D}g,\end{aligned}$$

where $\vec{c}_j = \sum_i \partial_j \mathbf{C}_{ij}$. So, with $f = \gamma\varphi_s$,

$$\begin{aligned}\mathcal{L}_s g &= \frac{r}{\varphi_s}\left[\mathcal{D}^*(\gamma\varphi_s g) - \nabla\cdot(\gamma\varphi_s g\vec{b}) - g\mathcal{D}^*(\gamma\varphi_s) + g\nabla\cdot(\gamma\varphi_s\vec{b})\right]\\ &= \frac{r}{\varphi_s}\left[\gamma\varphi_s\mathcal{D}g + \gamma\varphi_s\vec{c}\cdot\nabla g + (\mathbf{C}\nabla(\gamma\varphi_s))\cdot\nabla g - \gamma\varphi_s\vec{b}\cdot\nabla g\right]\\ &= r\gamma\left[\mathcal{D}g + \vec{c}\cdot\nabla g + (\mathbf{C}\nabla\log(\gamma\varphi_s))\cdot\nabla g - \vec{b}\cdot\nabla g\right],\end{aligned}$$

which is equation (2.23). □

Proof [of Corollary 2.27] For the moment, we will write $r(x)$ for $r(x,\eta)$ and $\gamma(x)$ for $\gamma(x,\eta)$. First note that since in this case the semigroup does not depend on time, we can write $\mathcal{L} = \mathcal{L}_s$, and

$$\mathcal{L}f = \sigma^2 r\gamma\left(\Delta f + \nabla(2\log(\gamma\varphi) - h/\sigma^2)\cdot\nabla f\right).$$

Now, observe that

$$\int_{\mathbb{R}^d} e^{H(x)} f(x)(\Delta + \nabla H(x)\cdot\nabla)g(x)dx = -\int_{\mathbb{R}^d} e^{H(x)}\left\{\nabla f(x)\cdot\nabla g(x)\right\}dx,$$

so that by choosing $H(x) = 2\log(\gamma(x)\varphi(x)) - h(x)/\sigma^2$ and

$$\pi(x) = e^{H(x)}/(\sigma^2 r(x)\gamma(x)) = \frac{\gamma(x)\varphi(x)^2 e^{-h(x)/\sigma^2}}{\sigma^2 r(x)},$$

we have that

$$\int_{\mathbb{R}^d} \pi(x) f(x) \mathcal{L} g(x) dx = -\int_{\mathbb{R}^d} e^{H(x)} \nabla f(x) \cdot \nabla g(x) dx.$$

Since this is symmetric in f and g, the process Y is reversible with respect to π; the constant factor of σ^2 does not affect the result. $\square$

6

FRIME - A Markov Model for Cell-free DNA Fragmentomics

6.1 Introduction

Cell-free DNA (cfDNA) is extracellular DNA present in biofluids, particularly blood, that originates mainly from dying cells as part of normal cellular turnover or pathological processes, including cancer [HAS20]. Assays, such as PCR and sequencing, can detect cfDNA in the bloodstream, providing clinicians with molecular information on the cell-of-origin. This created a new field called “liquid biopsy", which is more accessible and less risky than traditional biopsies, which require invasive procedures that may cause complications such as infection, bleeding, or pain. The decreasing cost of sequencing technologies has made cfDNA a promising material for prenatal genetic diagnosis, early cancer detection, and cancer treatment decision-making [Lo+21]. Currently, over 1000 clinical trials are exploring the potential of cfDNA, and there are a few FDA-approved tests [Cis+22].

Apart from being low-risk and easier to obtain, cfDNA can provide real-time monitoring of the disease status, allowing for prompt therapy adjustments. However, several challenges must be addressed before cfDNA can become a standard clinical tool. These include the need for further validation of the methods used for cfDNA detection, identification of the most appropriate clinical applications, optimization of the cost-effectiveness of cfDNA

analysis and an improved understanding of the processes behind cfDNA generation and dynamics [Rol+20].

There are many molecular features in cfDNA data, such as DNA mutation, copy number variation [Ada+17], DNA methylation [Liu+20], and histone modifications [Sad+21]. When focusing on the oncological applications of cfDNA, identifying pathological signals from physiological activities in cfDNA is non-trivial. This is because tumor-derived cfDNA usually makes up less than 10% of plasma cfDNA even at late tumor stages [MNT17], resulting in low signal-to-noise ratio. Sometimes, tumor cfDNA cannot be detected despite using highly sensitive methods such as digital PCR [Bet+14]. Strategies to overcome this problem include enriching for tumor-derived DNA [Mou+18], as well as combining multi-modal information to increase the number of tumor-specific features that can be captured or analyzed [Sie+21; Pen+21].

CfDNA fragment length profile is one of the keys in both strategies. It has been demonstrated in mice xenograft models and cancer patients that tumor-derived cfDNA are in general shorter and more fragmented [Und+16; Mou+18], therefore selecting for short DNA fragments between 90-150 base pairs (bp) can enrich tumor signals. The short-to-long fragment ratio has also been proposed to be a diagnostic marker [Cri+19]. Yet, studies so far have been focusing on fragment sizes under approximately 200bp, and fragment length profiles beyond this have not been thoroughly described. Moreover, how to best integrate the fragmentation profile with other multi-modal information remains an open question.

Currently, cfDNA data is commonly perceived by biologists as a fixed snapshot, with testing for specific markers of interest, such as mutation frequency. Nevertheless, cfDNA is part of a dynamic equilibrium similar to other biological processes, constantly being generated, broken down, and eliminated from the body. Although the kinetics (rate of change in concentration) of cfDNA can potentially be analyzed during tumor resection [Die+08] or caesarean section [Yu+13], these approaches are not practical for diagnostic or molecular characterization purposes.

Fragmentation patterns have been proposed to have a high potential of containing information of clinical relevance in what it is currently being referred to as cfDNA fragmentomics. Further insights into the fragmentomics of plasma cfDNA may shed light

on the origin and fragmentation mechanisms of cfDNA during pathological processes in diseases and enhance our ability to take the advantage of plasma cfDNA as a molecular diagnostic tool. This includes extracting information from fragment length, end motifs, jagged ends, preferred end coordinates, as well as nucleosome footprints, open chromatin region, and gene expression inferred by the cfDNA fragmentation pattern across the genome [Qi+23; Liu22; Mou+21].

However to unveil the cfDNA properties, including that of fragmentation patterns, as a potential biomarker for early diagnosis, diagnosis and prognosis, it is crucial to characterize the intrinsic kinetics of cfDNA within the organism. This is so especially when using cfDNA analysis as longitudinal diagnostic tool for monitoring the course of treatment [KL18]. Consequently, computational models that capture the dynamics of cfDNA may serve as a valuable tool to extract biological insights from the static cfDNA data.

Mathematicians and physical chemists have used fragmentation equations to model the degradation of polymers [ZM85; EBE81]. Fragmentation equations are differential equations that describe how the concentration of fragments at different length scales changes over time. However, these equations did not account for the biological noise that occurs naturally or the advent of DNA sequencing technology. While fragmentation with immigration processes have been studied [BK00; Haa05], no mathematical or computational model has been built including an exit mechanism to study the effect of cfDNA clearance mechanisms on the fragment profile. Simple approaches have estimated basic averaged kinetic parameters for different processes [KL18] but without incorporating the complexities of fragmentation processes. Recent studies based on non-dynamical approaches have correlated tumor detection size with circulating tumor DNA shedding [Ava+20] and developed a mathematical equation relating the distribution profile of a stochastically fragmented DNA sample to the probability that a DNA fragment within that sample can be amplified by any PCR assay of arbitrary length [Joh+22]. Finally, data analytic mathematical and machine learning methods have been used extensively to analyze cfDNA fragmentation datasets [Ren+22; Zho+23; Mos+23]. Nevertheless, those approaches have not incorporated mechanistically the dynamical interplay of the different processes playing a role in the fragmentation dynamics.

CfDNA samples collected before and after treatment can provide an expanded view of the genetic response of a patient's tumor, including the dynamic changes in the mutational landscape as well as the heterogeneity that develops due to the selective pressure of therapy. In addition, it is clear that the potential of cfDNA analyses is enhanced when combined with results obtained from tissue biopsies and integrated with mathematical modelling of tumor evolution [Kha+18].

Therefore, simulation-based computational frameworks are required to gain a deeper understanding of the biological processes that occur during DNA fragmentation and to provide a more robust clinical analysis, and thus develop biomarkers, of sequenced cfDNA fragment profiles.

In this study, we develop a mathematical/computational stochastic Markovian model and propose that the dynamic behaviour of cfDNA can be retrieved from the fragment length distribution of cfDNA, obtained from sequencing. When a long DNA molecule is cleaved by DNase, the resulting shorter molecules remain in the system, which would be further split into even smaller molecules. When the system reaches dynamic equilibrium, the abundance of molecules of each size would be dependent on the fragmentation behaviour.

The structure of this chapter is as follows: In the Methods section, we describe the biological factors that determine the dynamics of cfDNA and propose the FRagmentation with IMmigration and Exit (FRIME) framework, along with an algorithm to implement it. In the Results section, we use the FRIME framework to simulate the distribution of cfDNA fragment lengths and test various biological assumptions by adjusting the parameters of the model. We then compare our simulated results with real-world cfDNA data obtained using state-of-the-art sequencing techniques. In the Discussion section, we explore the biological implications of our results. Our simulations elucidate the shape of the entire cfDNA fragmentation profile, including the fragment counts of high molecular weight fragments. Moreover, our study identifies a difference in fragmentation kinetics between genomic and mitochondrial cfDNA, revealing new insights into how cfDNA is metabolized. We also provide a novel framework to study the dynamic parameters of cfDNA from existing data, which may improve analytic strategies. By shedding light on

the underlying biological mechanisms of cfDNA fragmentation, our study contributes to the development of cfDNA as a noninvasive biomarker in clinical settings.

6.2 Materials and methods

6.2.1 Biological determinants of cfDNA dynamics

Various pathways of cell death, degradation, and regulated extrusion, partial or complete genomes of various origins (e.g., host cells, fetal cells, and infiltrating viruses and microbes) contribute to the shedding into human body fluids, most notably blood, of segmented cell-free DNA (cfDNA) molecules [Gra+20]. Active and passive release and elimination mechanisms contribute to the composition of cfDNA. In this chapter we assumed these macromolecules to be continuously produced, metabolized, and removed from blood through the interplay of the following processes:

Production. CfDNA originates mainly from dying cells, which can be part of normal cellular turnover or pathological processes such as cancer [HAS20]. When cells die, mitochondrial and genomic DNAs are released into the bloodstream. There are variations in blood cfDNA levels among patients with different tumor types and stages. In general patients with benign lesions or with early-stage cancer have lower amounts of cfDNA compared to patients with advanced or metastatic tumors of comparable size, which could be a result of the increased cell turnover and metabolic properties of progressing cancer [Kus+19]. It has been found that cfDNA levels are correlated with metabolic disease volume, estimated with ^{18}F-labelled fluorodeoxyglucose positron emission tomography, in some cancers. In our modelling framework, cfDNA input mechanisms, whatever their origin, are incorporated as an input term feeding large fragments sizes into the system.

Fragmentation rate. CfDNA can be digested both intracellularly during cell death and extracellularly through different processes. The fragmentation rate of cfDNA depends on the DNase activity, which can differ across different body fluids [NYK93], health conditions [Tam+06], and individuals. In the context of plasma cfDNA, it is digested extracellularly by circulating DNases that break partially digested cfDNA into even smaller fragments.

Fragmentation pattern. Healthy and diseased cells exhibit a structured genomic DNA arrangement, with nucleosomes tightly binding most of the DNA except for short linker stretches [Iva+15]. DNases selectively digest the linker DNA, creating a periodic pattern of fragment sizes, where each peak corresponds to DNA associated with nucleosomes. This pattern is referred to as a nucleosomal pattern. In contrast, mitochondrial DNA is not bound by nucleosomes and lacks a nucleosomal fragmentation pattern.

Elimination. The elimination mechanisms of circulating cfDNA are not fully understood. Nucleosome-bound cfDNA is thought to mainly be cleared by the liver, whereas naked single-stranded DNA may be cleared through the kidney in a size-dependent manner [Kus+19]. Since the concentration of cfDNA is more or less in a dynamic equilibrium, it can also be inferred that elimination of cfDNA is concentration-dependent. The reported half-life of cfDNA is between 0.25 to 2 hours [Yu+13; Die+08].

6.2.2 Technical determinants of cfDNA fragmentation data

There are multiple factors that may cause a systematic bias in cfDNA fragment lengths readings, including the choice of cfDNA extraction kit [Dev+14], single-stranded versus double-stranded library preparation [Sny+16], and most notably, short-read versus long-read sequencing platform. Short-read sequencing platforms are designed for short DNA molecules, and are not ideal for our study for a few reasons. Firstly, the sequencing length is not guaranteed to cover the entire DNA molecule, and fragment length is inferred from the reference genome after sequence alignment. Thus, fragment length inference is affected by alignment errors as well as host-specific genetic variations. Secondly, there is a fragment length bias in short-read sequencing, where large DNA fragments are captured very inefficiently. This may be an effect of PCR during library preparation or the sequencer itself [Goh+19]. On the other hand, long-read sequencing is capable of directly measuring the fragment length by reading through the whole DNA molecule. It can also capture large DNA fragments efficiently with a modest size bias [Yu+23].

6.2.3 Computational model

To model the dynamics of cell-free DNA fragments in the bloodstream, we propose the FRIME framework which integrates the mechanisms mentioned above into a continuous time Markov process.

See Fig 6.1 for a schematic diagram of the process.

FRagmentation mechanisms. CfDNA fragmentation is known to occur through DNase digestion, however it remains unclear how exactly these biomolecules are broken down and how the reaction rate of cfDNA digestion depends on their size. Reaction rates of other polymers are known to be proportional to polymer sizes raised to some power [BEE78]. Similar results can also be deduced mathematically under the Zimm chain model which considers random movement of polymers [DEE88]. In this chapter we consider different types of fragmentation rate and test their outcomes against human data to see which one provides a faithful description of the observed phenomenology. Such comparison with clinical data can help us find possible biological mechanisms for cfDNA digestion.

We also assume that a fragment of length x reacts with an enzyme at a rate $C_F x^{\alpha_F}$. Upon reaction, the fragment breaks into two pieces of length $xr, x(1-r)$, where $r \sim Beta(a, 1)$. In this chapter, we will call the random variable r *the fragmentation ratio*, and the parameters C_F, α_F will be referred as the *fragmentation speed* and *fragmentation power* respectively.

The case of $\alpha_F < 0$ corresponds to an increase of fragmentation rate with shorter fragment lengths. This corresponds to the hypothesis that the reaction rate of cfDNA polymers are limited by diffusion, and longer cfDNA polymers diffuse slower than shorter polymers due to their size and possible secondary structures.

The case $\alpha_F > 0$, conversely, assumes that fragmentation rate increases with longer fragment length. This scenario corresponds to cfDNA digestion by surface exposure, where the rate would be proportional to the number of available reaction sites.

The parameter a determines the distribution of the fragmentation ratio r. When $a = 1$, the fragmentation ratio r is given by a Beta(1,1) distribution, and is thus uniformly distributed in $[0, 1]$. Biologically, this would correspond to the situation where cfDNA is

digested by endonucleases, i.e. fragments cut points are equally likely to be anywhere. When $a \gg 1$, the fragmentation ratio r is biased to values close to 1. This last situation would be where the dominant fragmentation mechanism is digestion by exonucleases, i.e. fragments are cut near the edges. It is known that the major intracellular and extracellular DNases responsible for cutting cfDNA are not exonucleases[HAS20; Han+20], thus in this chapter we chose $a = 1$. As described in the paragraph *Fragmentation pattern*, genomic DNA is nucleosome-protected, and fragmentation is not uniform. However, we still assumed uniform cutting on a nucleosomal scale.

Finally, in Fig B.3 in the appendix, we will show that the fragmentation profiles simulated with $r \sim Beta(a, a)$ do not resemble clinical data. Therefore, we will not study in detail the case where $r \sim Beta(a, a)$ for $a \neq 1$.

IMmigration mechanisms. We assume that the maximum possible length of an immigrated fragment is L, that would be equal to or smaller than the size of a full DNA molecule. Fragments of sizes between $[0, L]$ immigrate into the system with rate C_I. As we have no prior knowledge of the length distribution of freshly released cell-free DNA, the lengths of new fragments are assumed to be uniformly distributed in the interval $[0, L]$. The parameter C_I is chosen to denote the *Immigration speed*.

An alternative is to assume uniform cutting by intracellular DNases of DNA before entering the bloodstream. This scenario would result in more short fragments than long fragments, and could be modelled by imposing that the typical immigrant has length $x = Lr$, where $r \sim Beta(1, n)$. The stationary fragment distribution profiles under such an assumption are also studied in this chapter and described in detail in Figure B.3.

Exit mechanisms. We considered three different classes of exit mechanism: a constant exit rate (CON), Power decay with No Boundary (PNB), and a Power decay exit model with a Fixed Boundary (PFB).

For the CON model, we assumed that fragments of all sizes exit at the same rate C_E. This implies that the cfDNA clearance mechanism is agnostic to cfDNA size. The PNB model corresponds to the assumption that fragments of length x exit the system at a rate proportional to x^{α_E} with $\alpha_E < 0$. That functional dependence describes situations

in which all fragments leave the system but smaller fragments are easier to release from circulation. Finally, the PFB model is based on the assumption that only fragments smaller than B_E exit the system and do so at a rate proportional to $x^{\alpha_E} - B_E^{\alpha_E}$. This modelling scenario is in-line with the the renal clearance hypothesis, in which only short cell-free DNA fragments can pass through the glomerular capsules and be cleared.

In the following sections of this chapter, the parameter α_E is referred as the *exit power* and the parameter B_E as the *exit boundary*. A summary of all three exit mechanisms to be studied in this chapter can be found in Table 6.1. One of our goals in this chapter is to study which mechanism is more likely to be the one responsible for cfDNA clearance.

Exit Mechanism	$\mathcal{E}(x)$
CON	C_E
PNB	$L^{-\alpha_E} x^{\alpha_E},\ \alpha_E < 0$
PFB	$L^{-\alpha_E} \max\{x^{\alpha_E} - B_E^{\alpha_E}, 0\}$

Table 6.1: Functional dependencies of the exit rates $\mathcal{E}(x)$ on the fragment sizes x for the different exit mechanisms considered in this chapter.

FRIME Process and equilibrium measure. To model the cfDNA fragmentation profile, we study a Markov model with the above fragmentation, immigration, and exit mechanisms. Fragment profiles are sampled from the equilibrium measure of the model. More specifically, Monte Carlo simulations were run for a continuous-time Markov Chain (FRIME process) satisfying the following properties:

- The process starts off with a fragment with length $x_i(0) \sim U(0, L)$.
- Fragments of length x at rate $C_F x^{\alpha_F}$ fragment into two pieces of length xr and $x(1-r)$, where $r \sim Beta(a, 1)$. Both fragments remain in the system and evolve independently.
- Fragments of length $x < B_E$ leave the system at rate $\mathcal{E}(x)$.
- New fragments immigrate into the system as a Poisson process with intensity $C_I L$. The length z of a new fragment is given by the uniform distribution on $[0, L]$.

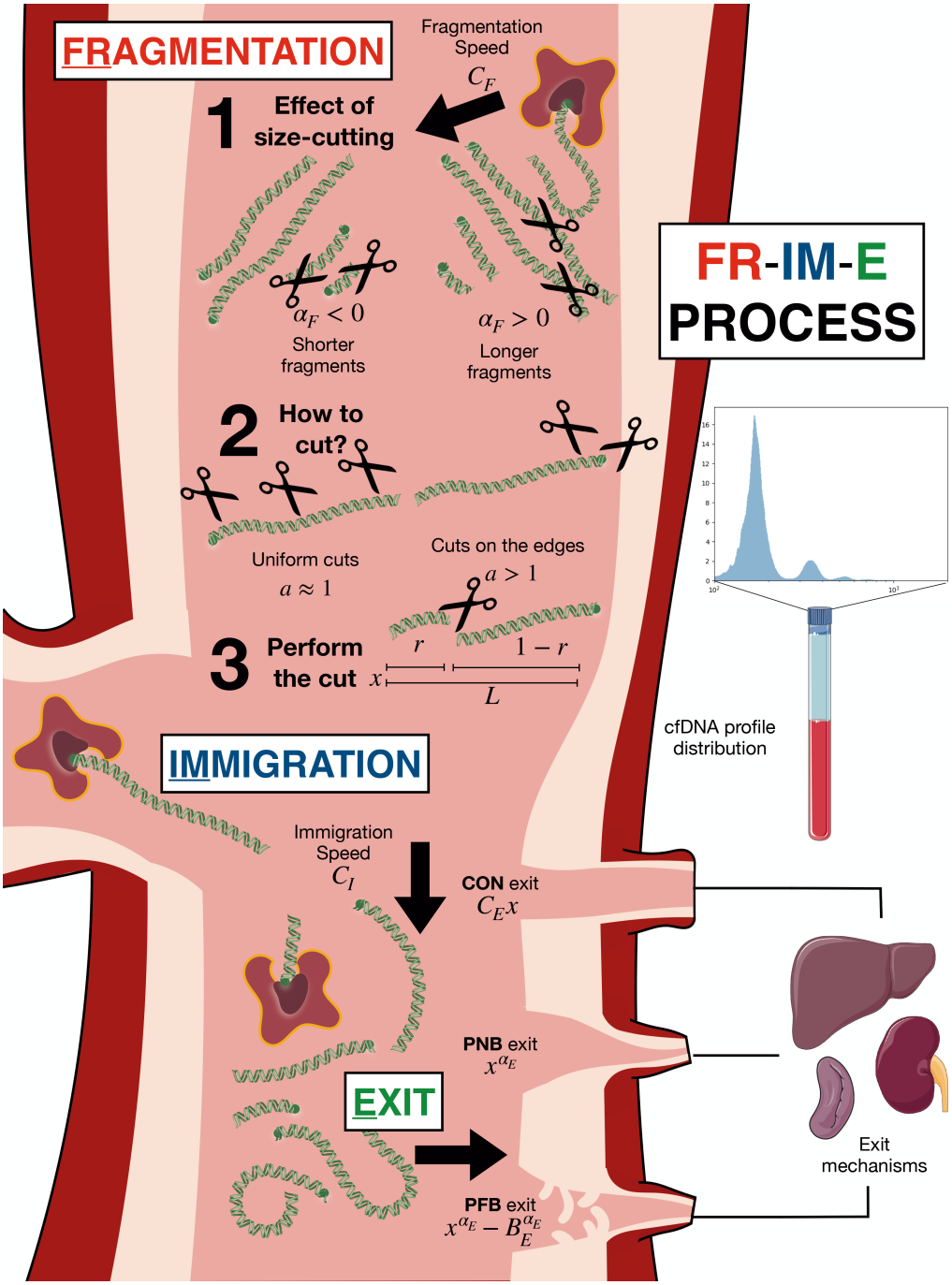


Figure 6.1: Schematic diagram of the FRIME process of the distribution of cfDNA fragments with length x. The FRagmentation mechanisms included the effect of: 1) fragmentation speed C_F and fragmentation power α_F (negativity and positivity related to shorter and longer fragment lengths, respectively); 2) the cut location, with $a = 1$ as uniform cuts, and $a \gg 1$ as cuts near the edges; 3) the fragmentation ratio r, which followed a distribution $Beta(a, 1)$ for any possible fragment of length L. The IMmigration processes were modelled by an immigration speed C_I. For the Exit mechanisms we considered three different possibilities. First a constant exit rate (CON, with C_E as the constant), next a Power decay with No Boundary (PNB, with α_E as the exit rate), and finally a Power decay exit with Fixed Boundary (PFB, with α_E as the exit rate and B_E as the boundary). Icons used are originally from `https://bioicons.com`, with license `https://creativecommons.org/licenses/by/3.0/`.

6.2.4 Algorithm

The following algorithm was used to generate one simulation of a stationary profile of the FRIME process for the numerical results.

1. Generate $x_0 \sim U(0, L)$. Assign $\mathcal{L} = [x_0]$. Assign $\mathcal{L}^{hist} = \mathcal{L}$. Assign $d = 1, T = 0, N_{step} = 0$ and pick a suitable value for T_{scale}.

2. While $d > \kappa$:

   (a) $N_{step} += 1$

   (b) While $T < N_{step} * T_{scale}$:

      i. Generate $t_{imm} \sim Exp(C_I L), t_{frag} \sim Exp(\sum_{x \in \mathcal{L}} C_F x^{\alpha_F})$ and $t_E \sim Exp(\sum_{x \in \mathcal{L}} \mathcal{E}(x))$

      ii. Assign $t_{min} = \min(t_{imm}, t_{frag}, t_{exit})$. Assign $T = T + t_{min}$.

      - If $t_{min} = t_{frag}$, sample $x_{frag} \in \mathcal{L}$, such that the probability of sampling x_i is proportional to $x_i^{\alpha_F}$.
        Generate $r \sim Beta(a, 1)$.
        Assign $\mathcal{L} = \mathcal{L} - [x_{frag}] + [rx_{frag}, (1 - r)x_{frag}]$
      - If $t_{min} = t_{imm}$, generate $x_{new} \sim U(0, L)$. Assign $\mathcal{L} = \mathcal{L} + [x_{new}]$.
      - If $t_{min} = t_E$, sample $x_E \in \mathcal{L}$, such that the probability of sampling x_i is proportional to $\mathcal{E}(x_i)$.
        Assign $\mathcal{L} = \mathcal{L} - [x_E]$.

   (c) Assign $\mathcal{L}^{hist} = \mathcal{L}^{hist} + \mathcal{L}$. Assign $d = KS_{Dist}(\mathcal{L}, \mathcal{L}^{hist}/N_{step})$.

In step 1, the initial conditions of the FRIME process are set up. The list $\mathcal{L}$ is the current fragment length distribution and the list $\mathcal{L}^{hist}$ is the list of fragment lengths throughout the history of the simulation. To ensure reproducibility of numerical experiment results, an initial seed of 42 is set to generate random numbers with numpy packages.

In the while loop, T corresponds to the current time step of the algorithm, and is updated whenever a new event is simulated. N_{step} is the number of time steps taken, which correspond to the number of times the fragment list $\mathcal{L}$ is updated. The fragment list is updated at intervals of T_{scale}. The value of T_{scale} is picked to be sufficiently large

to ensure at least $10 - 100$ immigration and fragmentation events between updates. κ is the threshold of tolerance for the metric d, which is the Kolmogorov-Smirnov distance between the current fragment list $\mathcal{L}$ and the historical average $\mathcal{L}^{hist}/N_{step}$.

In step 2(a), the next immigration, fragmentation and exit times t_{imm}, t_{frag}, t_E are generated and in step 2(b) the present time T is updated to $T + \min\{t_{imm}, t_{frag}, t_E\}$ by the theory of competing exponentials. The fragment list $\mathcal{L}$ is also updated by different mechanisms according to which simulated event comes first.

In step 2(c), the present fragment profile is loaded into the historical fragment list, and the Kolmogorov-Smirnov distance between the current fragment list $\mathcal{L}$ and the historical average $\mathcal{L}^{hist}/T$ is computed and stored. The process terminates when d is smaller than some preset tolerance threshold $\kappa > 0$.

Note that T is an auxiliary variable and does not reflect biological time; it is independent of the stationary profile of a FRIME process. Furthermore, because we only count fragment lengths when around 100+ fragments are immigrated, we ensure that both $\mathcal{L}, \mathcal{L}^{hist}/N_{step}$ contain a mixture of sub-fragments. In particular, for a typical simulation, $\mathcal{L}$ would contain the mixture of the evolution of more than 1000 immigrated fragments, and contain at least 10,000 fragments. Therefore, by law of large number, one can suppose that upon re-scaling and for large enough T, both $\mathcal{L}, \mathcal{L}^{hist}/N_{step}$ are close to data sampled from the stationary distribution of the underlying FRIME process.

On the other hand, to produce simulated fragment profiles for comparison with clinical data, we do not use κ as a threshold, but simply terminate the update of $\mathcal{L}$ after 10^6 fragmentation/immigration/exit mechanisms have occurred.

6.2.5 Computing machines and software

All simulations and clinical data analysis were run on Python 3.9.12 on a MacBook Air M2 2022 with 8GB Memory and the Apple M2 Chip with operating system macOS Ventura 13.3.1. Python packages required for the simulations are Numpy version 1.23.4, Scipy version 1.7.3, Matplotlib version 3.5.1, Sklearn version 1.1.3, Seaborn version 0.11.2 and Pandas version 1.4.2. Simulations of FRIME processes were computationally light; all

simulations were run within 3 hours. Clinical data analysis took less than 2 hours. All codes for this chapter are accessible on `https://github.com/thltsui/cfDNA-FRIME`.

6.2.6 Data availability and preprocessing

Published cfNano data from Katsman et al. [Kat+22b] were used in this study, comprising cfDNA Nanopore long-read sequencing data from 7 healthy control subjects and 6 lung adenocarcinoma patients. Shallow sequencing was done at about 0.2× genome coverage and varies across samples. Anonymised sequence alignment files in BAM format were downloaded from [Kat+22a].

Alignment and data pre-processing have been previously described in the original publication. Fragment lengths in base pairs were extracted from properly mapped reads using samtools version 1.8 by excluding unmapped reads, non-primary alignment reads, and supplementary alignment reads. Fragment length information is then aggregated to obtain the number of fragments for each unique length in genomic DNA (gDNA) and mitochondrial DNA (mtDNA). We considered gDNA and mtDNA separately because they exhibit very different fragmentation profiles, as the former is nucleosome-bound and has a multi-peak nucleosomal pattern, while the latter is not nucleosome-bound and only has one peak.

6.2.7 Model fitting and comparison

Numerical experiment. To study the effect of each parameter of the FRIME model on its stationary profile, numerical experiments were run where one parameter of interest was assigned multiple values with all other parameters fixed.

For each parameter regime, a stationary profile was generated by running 50 simulations with different seeds. For all experiments, we set $L = 1, \kappa = 0.05$. The sample average, lower quartile and upper quartile of fragment count across samples at each length scale were plotted. The simulated data were also compared to model-specific expected fragment counts derived analytically in Chapter 7. From now on, we will call such model-specific mathematical expressions as the *best fit curves* for our models.

For a detailed derivaation of these expected fragment count, refer to Section 7.4.

Model	Extra Assumptions	Expected Fragment Counts at Stationarity
CON	$a = 1$	$C_1(C_F x^{\alpha_F} + C_E)^{-\frac{\alpha_F+2}{\alpha_F}}$
PNB	$a = 1, \alpha_F = 1$	$C_3 z^{-\alpha_E}(C_F z^{1-\alpha_E} + 1)^{\frac{3-\alpha_E}{\alpha_E-1}}$
PFB	$a = 1,\ x > B_E$	$C_2 x^{-(\alpha_F+2)}$

Table 6.2: Expected fragment counts for fragment profiles under different exit mechanisms. For each model above, the expected number of fragments as a function of fragment length for the stationary distribution of the FRIME process is given under extra assumptions. For the CON model, we assumed that the fragmentation ratio was uniformly distributed, i.e. $a = 1$. For the PFB model, the best fit curve only applies on the region $x > B_E$ and only if the fragmentation ratio is uniform. For the PNB model, the best fit curve is applicable if the fragmentation power is $\alpha_F = 1$.

FRIME simulation fitting. To test how well the FRIME model fits clinical data, stationary profiles from simulated FRIME processes were compared to clinical mitochondrial DNA (mtDNA) fragment profiles from [Kat+22a].

As fragments longer than 10^4bp are scarce, and sequencing accuracy for very long fragments is low, all fragments longer than 10^4 in clinical data were discarded. FRIME processes were simulated with maximum length $L = 10^4$. Furthermore, since cfDNAs are digested through endonuclease, we take $a = 1$ for our simulations.

Four clinical mitochondrial cell-free DNA fragment profiles sequenced by nanopore technology with more than 500 fragment counts were analyzed in detail. These fragment profiles were tested against simulated fragment profiles from the FRIME model using the Kolmogorov-Smirnov test. The simulated fragment profiles were sampled after running a FRIME process under specific parameter regimes for 10^6 fragmentation / exit/ immigration events.

The genomic cell-free DNA fragment profile for sample ISPRO.bc05 was analyzed, which has the highest sequencing depth among healthy control samples. To adjust for the cyclical fragmentation pattern of genomic DNA (gDNA) data, fragment counts were aggregated according to nucleosomal bins (see Figure B.2 for details). Because there are fewer than 10 relevant nucleosomal bins for our data analysis, the similarity between simulated FRIME profiles and the genomic cfDNA fragment profile was evaluated through a χ^2 goodness of fit test and not a Kolomogorov-Smirnov test (used for continuous data). See Figure B.1 for further details.

PFB model fitting. Note that in Table 6.2, the best-fit curve for the PFB model is the simplest among the three options. Furthermore, the PFB model is useful for a huge range of α_F, while the PNB model is only applicable for $\alpha_F = 1$. Therefore, a total of 13 mitochondrial and genomic cell-free DNA fragment profiles in [Kat+22a] were fitted with the PFB model. Specifically, the tail distributions of the these fragment profiles are fitted with the best fit curve of the form Cx^{β}. For further details on how the best fit coefficients are estimated with interpolation methods, refer to paragraph Interpolation method to account for high 0 counts in Appendix B.1.

6.3 Results

6.3.1 CfDNA fragment profile exhibits a peak and a linear decay trend under log-log scale

During initial data exploration, we observed that the cell-free mtDNA fragment profile generated from Nanopore long-read sequencing exhibits a linear decay trend under log-log scale. While the gDNA fragment profile has a cyclical, nucleosomal pattern, the linear trend remains true for gDNA if fragment counts are averaged over nucleosomal bins. The findings were illustrated using a selected sample in Fig 6.2. Furthermore, a clear peak in fragment count was observed for all mitchondrial and genomic cfDNA fragment profiles around 100-200bp. Plots for the full set of data can be found in Fig B.7, Fig B.8 and Fig B.9.

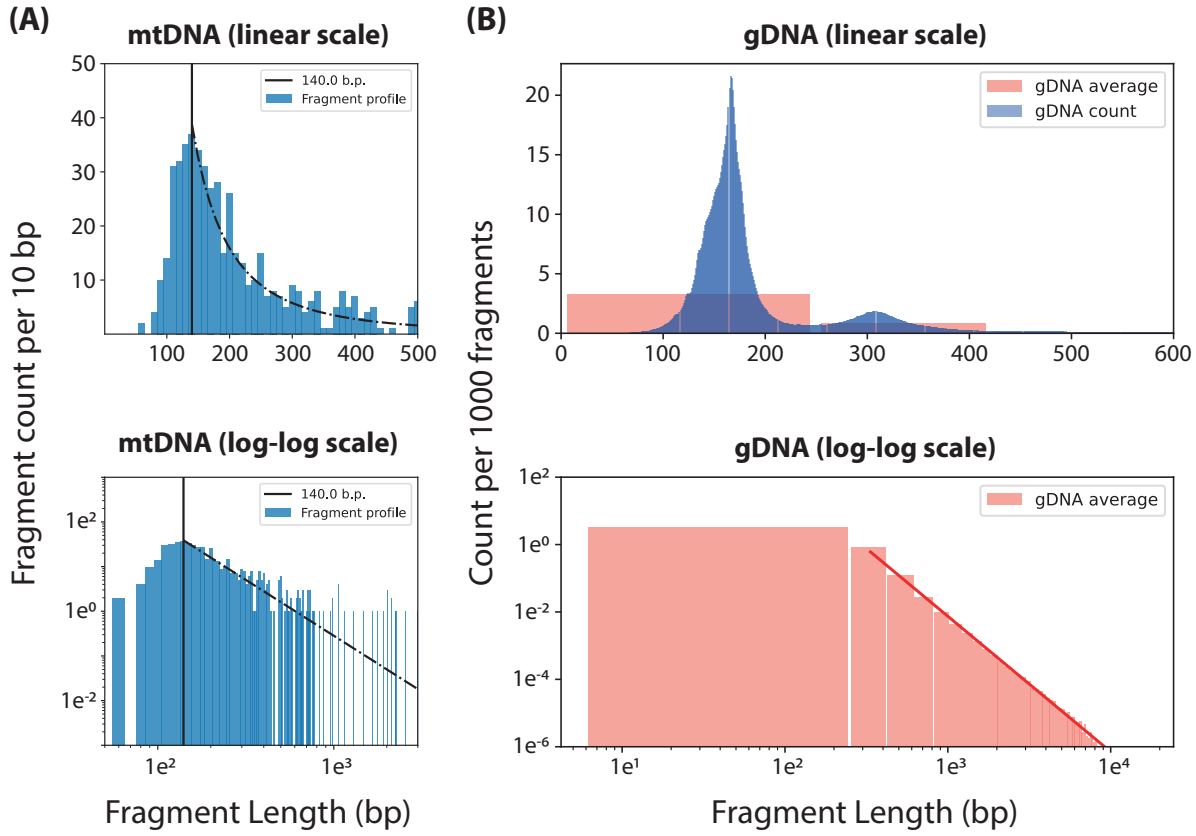


Figure 6.2: Fragment size profiles of a selected sample ISPRO.bc01 in linear and log-log scale. (A) Count per 10bp plotted against mtDNA fragment size shown in linear (top) and log-log scale (bottom). A straight line is fitted against interpolated counts in log-log scale, and is then plotted as a black dashed line in both linear and log-log scale. **(B)** Count per 1000 fragment plotted against gDNA fragment size, shown in linear (top) and log-log scale (bottom). The actual count is shown in blue, whereas the average count for each nucleosomal bin is shown in pink. A best-fit straight line for the average count is plotted in red for log-log scale.

6.3.2 Exit mechanism determines existence and position of fragmentation profile peak

To study the stationary fragment distribution as a function of fragment size, numerical experiments were run for different parameter values for each of the three classes of exit mechanisms described in the Section 6.2.3. Some results for specific parameter values are shown in Fig 6.3.

The results in Fig 6.3 showed relevant trends and features of the different models/exit mechanisms considered. Firstly the CON model led to a monotonic decrease of the fragment count with the fragment size and did not exhibit a peak in fragment count ubiquitous in clinical cfDNA profiles.

The PFB model showed a peak in the dependence of the fragment count with size. When the boundary played a substantial role, in this case $\alpha_E \leq -1$, the peak was obtained on the boundary B_E, and within the interval $[0, B_E]$ for larger α_E values. For a theoretical justification of the result, refer to Section 7.4.2.

Finally, the PNB model displayed a peak number of counts within the interval. The peak moved to smaller fragment counts and was more pronounced as the parameter α_E grew. We note that the log-log plots showed that the tail-distributions for PFB profiles were linearly decreasing in a broad range of small to medium fragment sizes. However, profiles obtained using the CON and PNB models did not exhibit linearity in the log-log plots.

The presence of peaks and the shape of the tail distribution in the log-log plots for the PFB model simulations are consistent with the previous data exploration. Thus, in the following, we perform the subsequent analyses focusing only on the PFB model.

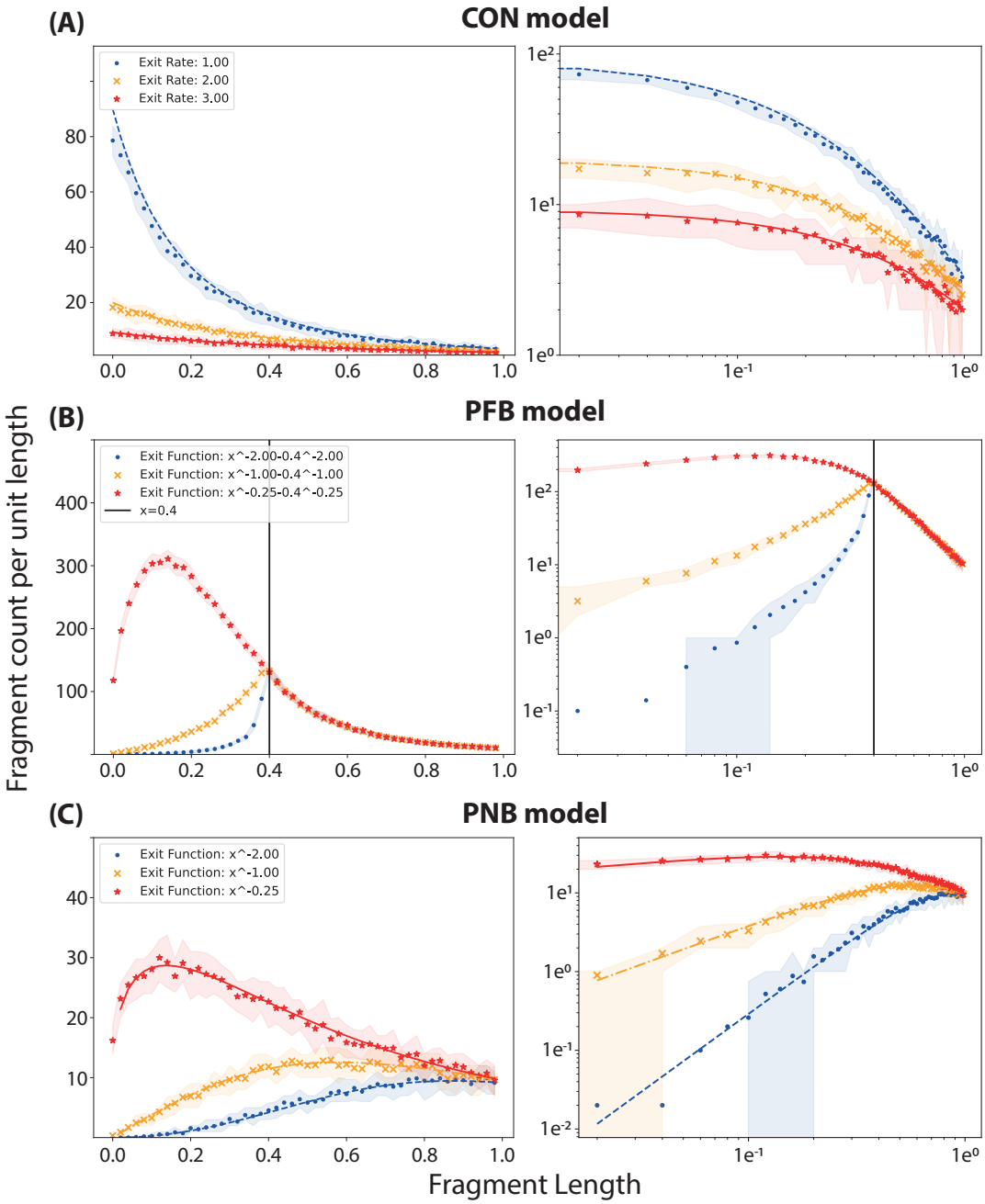


Figure 6.3: Simulated distribution of fragment sizes for different exit mechanisms. Stationary distributions of fragment sizes obtained from simulations of the FRIME process for parameter values $C_I = 500$, $C_F = 2$, $\alpha_F = 1$, $a = 1$, $\kappa = 0.05$. 50 points were plotted for each configuration. Each point corresponds to the fragment count per unit length over an interval of length 0.02. Different choices of the exit function were taken and 50 simulations run for each choice of exit function. The plots in the left column are in normal scale, while the plots in the right column are in log-log scale. **(A)** CON model with exit functions $\mathcal{E}(x) = 1$ (blue circles), 2 (orange crosses), 3 (red stars). **(B)** PFB model, $\mathcal{E}(x) = x^{\alpha_E} - B_E^{\alpha_E}$, with α_E = -2 (blue circles), -1 (orange crosses), -0.25 (red stars). Parameter B_E was taken to be equal to 0.4. The vertical line was plotted at the exit boundary $B_E = 0.4$. **(C)** PNB model, $\mathcal{E}(x) = x^{-2}$, (blue circles), x^{-1} (orange crosses), $x^{-0.25}$ (red stars). Shaded regions were plotted using upper and lower quartile fragment count across all simulations. Best fit lines were also plotted for comparison for the CON and PNB model.

6.3.3 Fragmentation mechanism determines tail distribution of fragment profile for PFB model

Fragmentation ratio and fragmentation power. To study the effect of the fragmentation mechanism on the stationary fragment profile, numerical experiments were performed for three different fragmentation ratios and three different fragmentation power rates, as seen in Fig 6.4.

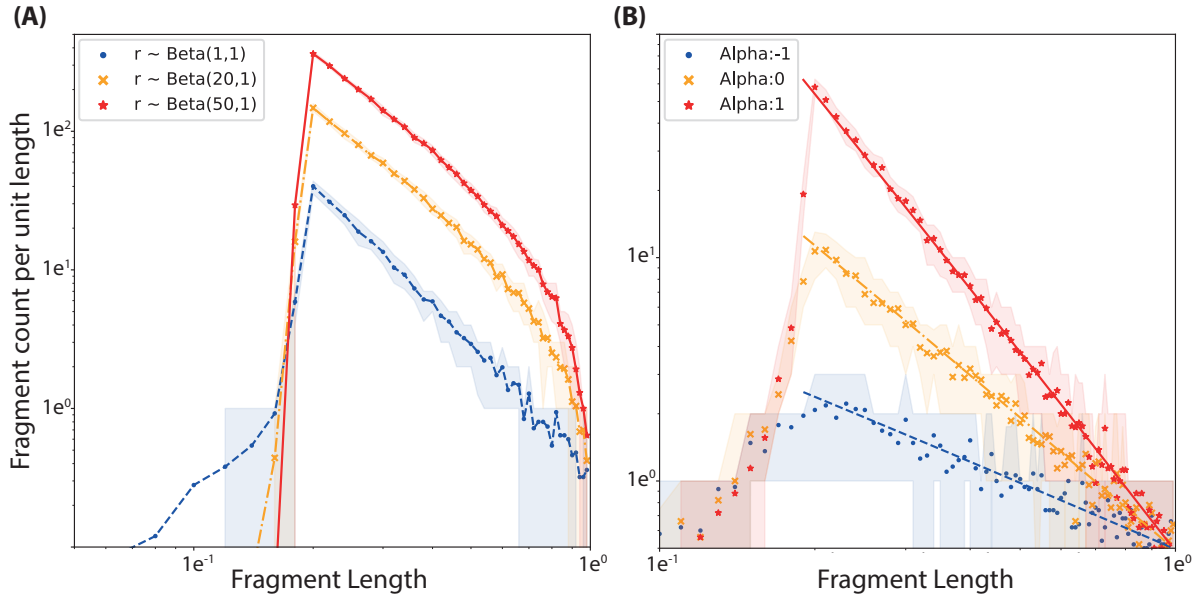


Figure 6.4: Simulated fragment profiles with different fragmentation ratios and fragmentation power rates. 50 simulations were run for each fragmentation mechanism. 50 points were plotted for each configuration corresponding to average fragment density at intervals sampled in range $[0, 1]$ of length 0.02. Shaded regions were plotted using upper and lower quartile fragment count across all simulations. PFB best fit lines were also plotted for comparison in panel **(B)** with blue-dashed, orange point-dashed and red-solid lines respectively for $\alpha_F \in \{-1, 0, 1\}$. All plots were in log-log scale. **(A)** The fragmentation ratios were $r \sim Beta(1, 1)$, $Beta(20, 1)$, $Beta(50, 1)$ respectively for blue-circled, orange-crossed and red-starred dots. Fixed parameters were ($C_I = 20$, $C_F = 1$, $\alpha_F = 1$, $\mathcal{E}(x) = x^{-2} - 0.2^{-2}$, $\kappa = 0.05$). **(B)** Fragmentation powers were $\alpha_F \in \{-1, 0, 1\}$ respectively for blue-circled, orange-crossed and red-starred dots. Fixed parameters were ($C_I = 20$, $C_F = 1$, $r \sim Beta(1, 1)$, $\mathcal{E}(x) = x^{-2} - 0.2^{-2}$, $\kappa = 0.05$).

In Fig 6.4(A), several fragmentation profiles with different fragmentation ratios were explored. A larger value in the first coefficient a of the $Beta(a, 1)$ distribution (corresponding to exonuclease hypothesis, see Fragmentation mechanisms in Computational model) leads to a steeper decay in the fragment profile tail distribution. In particular when $a = 1$, the fragment profile under the log-log scale is linear.

For the fragmentation profile corresponding to $r \sim Beta(1,1)$, the PFB best fit expression of the form $Cx^{-\alpha_F-2}$ lies within the shaded region, (see Fig 6.4(B)). This corresponds to a linear decrease under the log-log scale. Mathematically, if we denote by β_{PFB} the slope of the best-fit line in log-log scale, then

$$\beta_{PFB} = -\alpha_F - 2. \tag{6.1}$$

For fragmentation profiles where $r \sim Beta(n,n)$ for large n, refer to Fig B.4.

Fragmentation speed and immigration speed. To study the effect of fragmentation speed on the stationary fragment profile, numerical experiments were performed for three different fragmentation and three different immigration speeds, namely C_F and C_I. The fragmentation speeds were chosen as $C_F \in \{1,2,4\}$ with fixed $C_I = 400$. Similarly, the immigration rates were chosen as $C_I = 100, 200, 400$ with fixed $C_F = 1$.

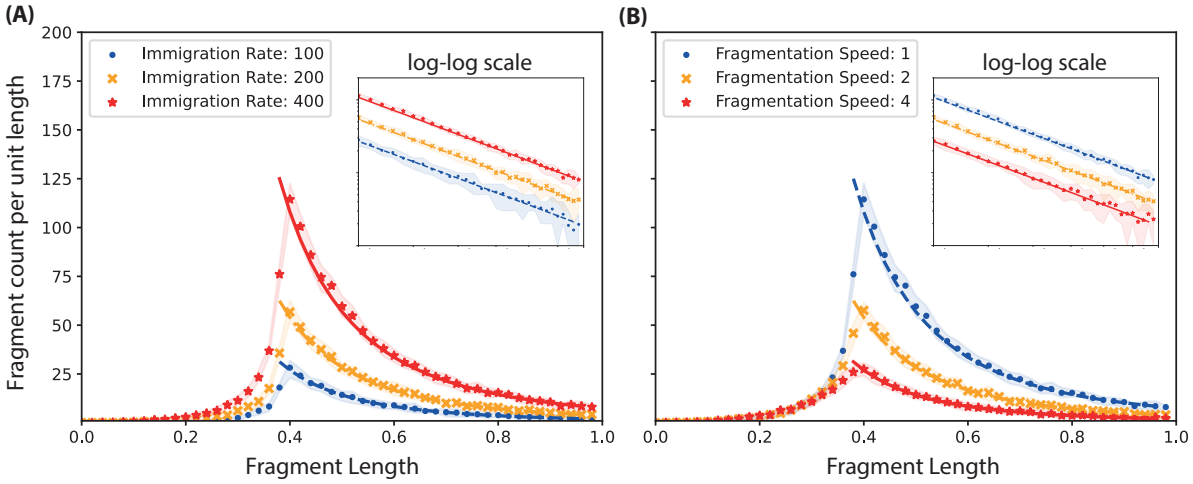


Figure 6.5: Simulated fragment profiles with different fragmentation and immigration speeds. 50 simulations were run for each fragmentation mechanism. 50 points were plotted for each configuration corresponding to average fragment density at [0,0.02], ...[0.98,1]. Shaded regions were plotted using upper and lower quartile fragment count across all simulations. The small plots were the tail distribution but in log-log scale. **(A)** The immigration rates were $C_I =$ 100, 200, 400 for blue circles, orange crosses, and red stars respectively. Fixed parameters were $C_F = 1$, $\alpha_F = 1$, $r \sim Beta(1,1)$, $\mathcal{E}(x) = x^{-2} - 0.4^{-2}$, $\kappa = 0.05$. **(B)** The fragmentation speeds were $C_f = 1$, 2, 4 for blue circles, orange crosses, and red stars respectively. Fixed parameters were $C_I = 400$, $\alpha_F = 1$, $r \sim Beta(1,1)$, $\mathcal{E}(x) = x^{-2} - 0.4^{-2}$, $\kappa = 0.05$.

The fragment profiles under the log-log scale are parallel to each other. The fragment profiles are scaled at the ratio C_I/C_F, where C_I is the immigration rate and C_F is the fragmentation speed.

For fragment profiles with non-uniform immigration distribution (e.g. Exponential, Normal distributions), refer to Fig B.3.

6.3.4 Simulated FRIME profiles fit clinical data

To check how well the FRIME model fits mtDNA clinical data, FRIME processes with different parameters were simulated and compared to the mtDNA fragment profile of the 4 samples ISPRO.S1, ISPRO.bc01, ISPRO.bc03 and ISPRO.bc05 from the nanopore study found in [Kat+22a]. The 4 samples were chosen according to the criteria described in Section 6.2.7.

For each of the 4 mtDNA fragment profiles, there are parameter regimes which produce simulated fragment profiles with p-values larger than 0.1 under the Kolmogorov-Smirnov test, see Table 6.3. As the p-values are larger than 0.1, the simulated fragment profiles cannot be considered statistically different from clinical data. See Fig 6.6(A-B) for an example of a simulated FRIME process that fits well with clinical data.

id	C_F	α_F	C_I	α_E	Model	B_E	d_{KS}	p
ISPRO.S1	10^6	-1.0	11.0	2.05	PNB	N/A	0.022	0.287
ISPRO.bc01	10^3	0.36	10.0	2.5	PFB	115	0.049	0.345
ISPRO.bc03	10^5	-0.5	10.0	2.10	PNB	N/A	0.046	0.261
ISPRO.bc05	7.5×10^5	-0.78	8.0	2.10	PNB	N/A	0.048	0.216

Table 6.3: Parameters of mtDNA best fit of the clinical datasets with FRIME processes. Simulated FRIME processes were run for 10^6 events with a fixed initial seed for generating FRIME events. The fragmentation ratio was fixed to be uniform, i.e. $r \sim Beta(1,1)$, and new fragment lengths were distributed as $U(0,L)$, with $L = 10^4$. The column id corresponds to the id of the patient whose mtDNA fragment profile was studied. The constant C_F is the fragmentation speed of the FRIME process and α_F is the fragmentation power. The constant C_I measures the immigration speed. The constant B_E is the exit boundary and α_E the exit power. For the PNB model, the exit function is of the form $L^{-\alpha_E}x^{\alpha_E}$; for the PFB model, the exit function is of the form $L^{-\alpha_E}(x^{\alpha_E} - B_E^{\alpha_E})$. The value d_{KS} is the Kolmogorov-Smirnov distance between the simulated fragment distribution selected (best fit) and the clinical mtDNA data. The value p refers to the p-value in the Kolmogorov-Smirnov test.

Similarly, simulated FRIME profiles are statistically close to genomic cfDNA fragment profiles when adjusted for nucleosomal cycles. Under the binning outlined in Fig B.2, simulated FRIME profiles and gDNA profiles are statistically similar (see Fig 6.6(C-E) for an example).

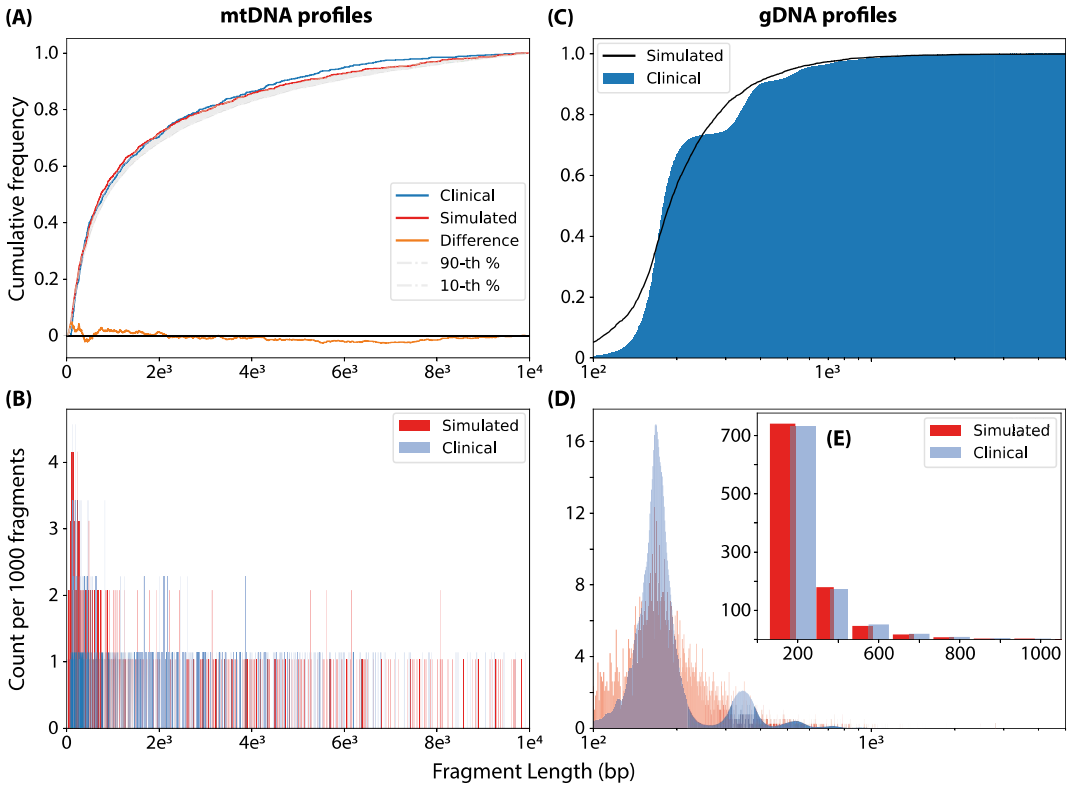


Figure 6.6: Fitting clinical data for sample ISPRO.bc05 with the FRIME model. **mtDNA -** The mtDNA fragment profile for sample ISPRO.bc05 in the nanopore study was compared to a simulated FRIME profile. The simulated profile was generated by running a FRIME model with parameters $C_F = 10^6$, $C_I = 11$, $\mathcal{E}(x) = L^{2.05}x^{-2.05}$, $\alpha_F = -1$, $L = 10^4$, $r \sim Beta(1, 1)$, with seed = 42 until 10^6 events had occurred. Fragment lengths were rounded to the nearest integer and plotted against clinical data. **(A)** The cumulative fragment count for the mtDNA profile of sample ISPRO.bc05 was plotted as the blue line. The mean cumulative fragment count for the simulated FRIME profile was plotted as the red line, and the shaded region was plotted using 10th and 90th percentile fragment count across all simulations. The difference between the two cumulative counts was plotted as the orange line. **(B)** The fragment count (per 1000 fragments) for the mtDNA profile of sample ISPRO.bc05 was plotted as a histogram in blue colour. The fragment count (per 1000 fragments) for the simulated FRIME profile was plotted as a histogram in red colour. Fragment length was measured in base pairs. **gDNA -** The gDNA fragment profile for sample ISPRO.bc05 was compared to a simulated FRIME profile. The simulated profile was generated by running a FRIME model with parameters $C_F = 2.5$, $C_I = 10$, $\mathcal{E}(x) = L^2(x^{-2} - 167^{-2})$, $\alpha_F = 1.2$, $L = 10^4$, $r \sim Beta(1, 1)$ until 10^6 events occurred. Fragment lengths were rounded to the nearest integer and plotted against clinical data. **(C)** The cumulative fragment count for the gDNA profile of sample ISPRO.bc05 was plotted with the blue area. The cumulative fragment count for the simulated FRIME profile was plotted as a black line. The x-axis is under the log scale. **(D)** The fragment count for the gDNA profile of sample ISPRO.bc05 was plotted as a histogram in blue colour. The fragment count for the simulated FRIME profile was plotted as a histogram in red colour. Fragment length is measured in base pairs and the x-axis is under the log scale. **(E)** The number of fragments in the region $[0, 250]$, $[250, 420]$, $[420, 620]$, $[620, 820]$, ..., $[1220, 1420]$ for both profiles were plotted as histograms. The simulated profile was plotted in red and the fragment profile for the gDNA profile of sample ISPRO.bc05 was plotted in blue.

6.3.5 Differences in PFB best-fit lines for gDNA and mtDNA profiles

The tail distribution of gDNA fragmentation profiles for all 13 samples in [Kat+22a] were fitted with the PFB best-fitting expression of the form Cx^{β}. Similarly, the mtDNA fragment profiles for 9 samples in [Kat+22a] were fitted. The PFB best fit lines align with the overall pattern for all of the 9 mtDNA and 13 gDNA fragment profiles.

Best-fit statistics for mtDNA data. For mtDNA data, the parameter β_{mtDNA} (slope of the best-fitting line under log-log scale) has an average value of -1.56 with range between $[-2.5, -0.69]$. According to Equation (6.1), this corresponds to a negative fragmentation power α_F of around -0.5. See Fig 6.7 for the best fit expressions of 6 mtDNA profiles. See Fig B.7, Fig B.8 for the best-fit curves for all profiles.

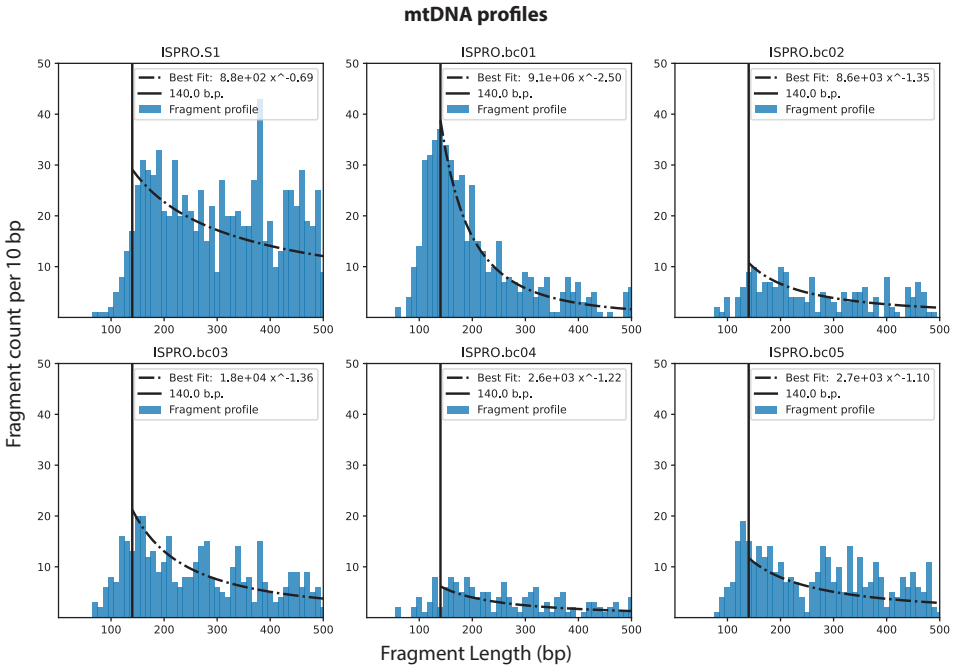


Figure 6.7: Cell-free mtDNA fragment profile with PFB best fit line. Mitochondrial DNA fragment profiles for samples ISPRO.S1, ISPRO.bc01, ISPRO.bc02, ISPRO.bc03, ISPRO.bc04, ISPRO.bc05 from nanopore studies are plotted as bar charts in blue on the top panels after linear interpolation at intervals of 10bp. Best fit curves of the expression Cx^{β} are plotted as black dot-dashed lines. The coefficients of the best fit curve are given by a linear fit under the log-log scale on the counts of fragments longer than 140bp. A black vertical line at 140bp indicating the fixed boundary is also plotted on each profile. Both x and y axes are in linear scale.

Best-fit statistics for gDNA data. For gDNA data, the parameter β_{gDNA} has an average value of -3.32 with range between $[-4, -2.5]$. According to Equation (6.1), this corresponds to a positive fragmentation power α_F of around 1. See Fig 6.8 for the best fit expressions of 6 gDNA profiles. See Fig B.9 for the best-fit curves for all profiles.

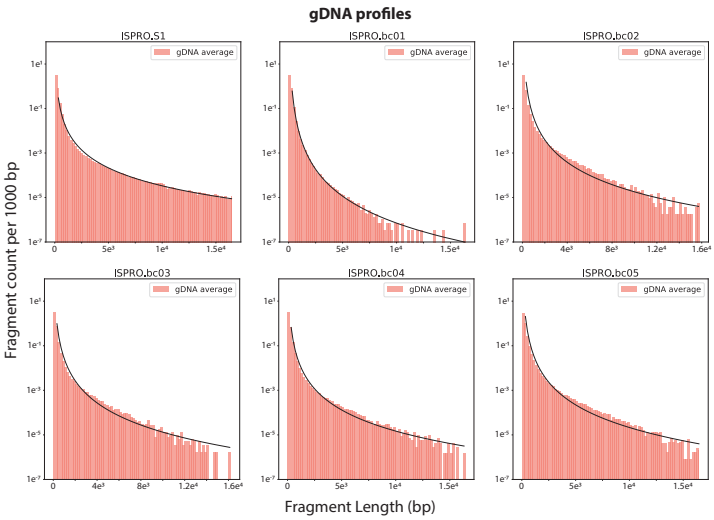


Figure 6.8: Cell-free gDNA fragment profile with PFB best fit line. Genomic DNA fragment profiles for samples ISPRO.S1, ISPRO.bc01, ISPRO.bc02, ISPRO.bc03, ISPRO.bc04, ISPRO.bc05 from nanopore studies are plotted as bar charts in pink. The coefficients C, β are estimated to plot the best fit curves of the form Cx^{β} in black. The coefficients of the best fit curve are given by a linear fit under the log-log scale after binning the data. The y-axis is in log-scale.

6.4 Discussion

In this chapter, we have proposed a Markov process (FRIME process) as a model for cfDNA production, digestion and clearance in the bloodstream. Under the FRIME model, we can examine how different assumptions on fragmentation, immigration, and exit mechanisms affect stationary fragment profiles and compare them to clinical data. The results suggest possible cfDNA fragment profiles as the equilibrium measure of a FRIME process. Using the FRIME model, we can examine how different assumptions on fragmentation, immigration and exit mechanisms produce stationary fragment profiles and compare them against real clinical data. As a result, we can deduce possible cfDNA digestion and clearance mechanisms that are impossible to study under an experimental setting. We now draw the following biological insights from our results.

6.4.1 Implications

Removal mechanism of mitochondrial cell-free DNA is size-dependent and explains the existence of fragment peak. We considered the cases of size-independent exit (CON), size-dependent exit with no boundary (PNB), and with fixed boundary (PFB). We show that size-independent exit results in no peak in the absence of other factors such as protein protection. On the other hand, a size-dependent removal mechanism of cfDNA is sufficient to explain the existence of a peak in fragment sizes. Circulating mtDNA, unlike gDNA, is not nucleosome-bound, but exhibits a single peak at around 100-150bp in the fragment profile, which is variable across individuals. The peak can be due to failure to capture ultra-short fragments during sequencing, or protection by protein. However, both cases should result in a peak at a same fragment length scale across individuals, which is not the case. We therefore propose that the existence of a modal fragment size in mitochondrial cfDNA can be explained by its removal mechanism, where shorter fragments are removed faster than longer fragments.

The tail distribution of cfDNA fragment profile provides rich information when examined on a log-log scale. In the initial data exploration, we observe that cfDNA fragment profiles exhibit a linear decay trend under log-log scale. In the subsequent modelling of the exit mechanism, we show that only the PFB model has a log-log linear tail distribution, i.e. to the right of the exit boundary. Furthermore, we examine the effect of endonuclease and exonuclease activity on the fragment profile. Log-log linearity is preserved only when all positions have the same chance of being cut (i.e. endonuclease model). In addition, we show that the slope of the tail distribution is determined by the fragmentation power, which reflects the effect of DNA length on reaction kinetics. Therefore, log-log linearity is a very unique property that can only be met by a few mechanisms in the FRIME model, which suggests cfDNAs degrade through endonuclease and are cleared only when they become small.

Fully nucleosome-bound cell-free DNA is rarely removed from circulation. While there are periodic peaks in genomic cfDNA data, we observe that the fragment distribution binned in nucleosomal units is linear on a log-log scale, until the first nucleosome. From our simulations, this suggests that genomic cfDNA has an exit boundary at around the size of 1 or few nucleosome(s), above which the clearance rate is 0 (corresponding to the PFB model), or very small compared to the fragmentation rate (corresponding to the PNB model with small α_E). We hypothesize that short gDNA, especially those shorter than 167bp, may no longer maintain a stable complex with histone proteins, which leads to a change in physical properties and clearance rate. Our result seems to contradict previous mice experiments, which have shown that the liver is capable of clearing, or at least trapping, circulating mononucleosomes and chromatin with radiolabelled histones [GTM96; Du +99]. However, we argue that previous studies only radiolabelled the histone component, and it is unknown whether the nucleosomes are cleared from the circulation as an intact unit with the DNA, so the experiment could not account for the rate of clearance of cfDNA shorter than 167bp.

mtDNA and gDNA undergo different size-biased fragmentation mechanisms. Parameter inference of the tail distribution from clinical data using PFB model shows that mtDNA and gDNA have a systematically different slopes, as shown in Fig B.6. The fragmentation power α_F can be derived from the slope, and is negative for mtDNA and positive for gDNA. Hence for each molecule, smaller mtDNA fragments react faster than longer mtDNA fragments. Meanwhile, gDNA exhibits the opposite reaction kinetics, and longer fragments react faster than smaller fragments. This suggests that the degradation rate of mtDNA may be diffusion-limited. Smaller fragments react faster because they have a higher diffusion rate than longer ones, which increases the chance of colliding with a DNase. On the other hand, although mtDNA and gDNA are chemically identical, gDNA forms a stable complex with histone proteins. This provides an explanation for the positive α_F, which suggests that gDNA fragmentation is limited by surface area instead of diffusion. The longer the DNA molecule, the higher the availability of fragmentation sites, hence higher chance of reacting with a DNase. Taken together with the discussion

on the clearance mechanism of nucleosome-bound cfDNA, we propose a unifying model where nucleosome-bound cfDNA are not free-floating in circulation, but may have low solubility and form aggregates or phase-separated droplets. This mesoscale organization may sequester the nucleosomes, making them difficult for the body to remove, and cfDNA fragmentation only happens at the contact surface with plasma. This is supported by the existing literature on cell-surface bound DNA, where high molecular weight DNA is found on surfaces of blood cells [Lak+04].

The FRIME model suggests the interpretation of cfDNA concentration as biomarker of tumor. Plasma cfDNA concentration alone is able to predict patient prognosis in multiple cancer types [Dzi+22; Fer+19; Che+21] and correlates well with overall tumor burden [Dzi+22]. Therefore, one hypothesis is that the increased plasma cfDNA comes from increased production. However, the literature also claimed that the DNase activity in cancer patients may be reduced, contributing to higher cfDNA concentration [Tam+06]. The factors contributing to cfDNA concentration have not been quantitatively studied holistically. From Fig 6.5, we show that under the FRIME model, when the cfDNA fragmentation kinetics remain unchanged, the rate of production of cfDNA is directly proportional to the total amount of fragments in a patient's bloodstream. This explains the linear correlation between cfDNA concentration and tumor burden. On the other hand, when the cfDNA production rate is unchanged, the fragmentation speed is inversely proportional to cfDNA fragment counts. The FRIME model, therefore, provides a numerical framework for interpreting the concentration of plasma cfDNA.

Best fit coefficients in log-log scale can serve as potential biomarkers. We show in Fig B.7, Fig B.8 that all fragment profiles in our study have an excellent fit under the PFB model visually. However, there are variations among individuals in the slope of linearity for such trends. We believe this to be a potential biomarker for diseases. Although we do not have enough data to fully examine this hypothesis, preliminary analysis suggests that best-fit gDNA coefficients of patients with lung adenocarcinoma have higher variance than healthy patients. For further detail, refer to Fig B.6.

6.4.2 Limitation and Future Work

We have generated several testable hypotheses from our FRIME model, including the exit mechanisms, physical and chemical kinetic properties for nucleosome-bound and free-floating cfDNA. We hope these ideas can be tested with *in vitro* and *in vivo* experiments in future. Apart from drawing biological insights, we are also able to infer different model parameters for individual patients, which are related to their cfDNA fragmentation and production rates, which may be more meaningful parameters than cfDNA concentration alone.

However, one of the biggest limitations here is the lack of high depth long read cfDNA sequencing data. Most of the published cfDNA sequencing data were obtained using short-read sequencing platforms, which are very inefficient in capturing DNA molecules beyond 1 kb, thus not suitable for comparing with our model. At the time of our study, only 1 long read cfDNA sequencing dataset was available without restrictions [Kat+22a]. Unfortunately, shallow sequencing was done for this dataset. We set a cutoff of at least 5000 mtDNA molecules being sequenced in order to statistically compare with our simulation, and only 4 samples pass this cutoff. With more clinical long read sequencing data we will be able to evaluate whether the fitted model parameters from the fragmentation profile are good biomarkers. Another limitation is that we are not able to reach a definite conclusion about the mode of the cfDNA exit mechanism. Although we obtain a good fit using the PNB model statistically under the FRIME framework, it cannot be guaranteed that the solution is the best fit.

Finally, our model inferred parameters are based on a time variable t set out in our algorithm with no biological meaning. For example, we may infer that a clinical profile can be simulated with x fragmentation events happen per unit time, but further work needs to be done to translate this into the commonly used units for DNase activity. Commonly used assays of DNase activity involve comparing samples to a reference with known activity [Sin+94; NYK93], which is often specified by the manufacturers as the amount of enzyme required to completely degrade 1 µg of plasmid DNA in the 10 minutes at optimal working conditions. We propose that a simulation can be performed to convert our auxiliary time unit to the degradation assay units, see Section 8.1.2.

7

FRIME - A Mathematical Perspective

7.1 Definition of the FRIME Processes

In this section, we introduce the mathematical formulation of *Fragmentation with Immigration and Exit (FRIME)* processes and compare them with *fragmentation with immigration* processes first introduced in [Haa05].

We first recall the definition of a self-similar fragmentation process and introduce the notion of an (α_F, a, b)-fragmentation process with finite dislocation measure. We then construct a fragmentation with exit process from an (α_F, a, b)-fragmentation process through a colouring scheme. Finally, define the FRIME process explicitly as a fragmentation process with Immigration.

7.1.1 Self-Similar Fragmentation Processes with Exit

We use $\mathcal{D}$ to denote the state space

$$\mathcal{D} = \{\mathbf{s} = (s_j)_{j\geq 1} : s_1 \geq s_2 \geq ... \geq 0, \lim_{j\to\infty} s_j = 0\}, \tag{7.1}$$

which, under the right topology, can be identified with the set of Radon counting measures on $(0, \infty)$ with bounded support endowed with the topology of vague convergence through the homeomorphism

$$\mathbf{s} \in \mathcal{D} \mapsto \sum_{j\geq 1} \delta_{s_j} \mathbb{1}_{s_j>0}.$$

In particular, we define $\mathcal{D}_1$ to be the subset of $\mathcal{D}$ such that $\sum_{j\geq 1} s_j \leq 1$. We call $\alpha \in \mathbb{R}$ the index of self-similarity, $c \geq 0$ an erosion coefficient and ν a dislocation measure on the set $\mathcal{D}_1$ that integrates $1 - s_1$. Furthermore, we restrict ourselves to finite dislocation measures. We now recap the notion of a standard (α, c, ν)-fragmentation first defined in [Ber02].

Self-Similar Fragmentation Processes: Under the (α, c, ν)-fragmentation process, a fragment with mass m splits into fragments with masses $m\mathbf{s}, \mathbf{s} \in \mathcal{D}_1$, at rate $m^\alpha \nu(d\mathbf{s})$, and undergoes a continuous erosion in size by a factor of e^{-ct} after a time $t \geq 0$. The subfragments then evolve independently from all other fragments with the same dynamics described above.

We now define the (C_F, α_F, a, b)-fragmentation process.

Definition 7.1 *The (C_F, α_F, a, b)-fragmentation process, denoted by $(F^{C_F,\alpha_F,a,b}(t))_{t\geq 0}$, is an $(\alpha_F, 0, \nu)$-fragmentation process, where a fragment of length x, at rate $C_F x^{\alpha_F}$, splits into fragments of lengths xr and $x(1-r)$, with $r \sim Beta(a, b)$. Specifically, ν has support on the set $\{(r, 1-r, 0, 0, ...) : r \in (0, 1)\} \subset \mathcal{D}_1$ with density*

$$d\nu((r, 1-r, 0, 0, ...)) = C_F \frac{1}{B(a, b)} r^{a-1}(1-r)^{b-1} dr. \tag{7.2}$$

We now introduce a colouring scheme to construct a fragmentation with exit process. This colouring scheme will also serve as a coupling argument for upcoming proofs.

Definition 7.2 *We consider a (C_F, α_F, a, b)-fragmentation process with an initial fragment of length $x_0 \sim Uni[0, L]$. We colour the fragment red if $x > \ell$, and blue otherwise. Whenever a fragmentation event occurs, fragments with size larger than ℓ inherit the same colour as their parent fragment, while fragments smaller than ℓ are immediately coloured blue. As the process evolves, red fragments of length x turn into blue fragments independently at an exponential rate with parameter $\mathcal{E}(x)$.*

We call the collection of red fragments the $(C_F, \alpha_F, a, b, \ell, L, \mathcal{E})$-fragmentation with exit process, and denote it $(F_X^{C_F,\alpha_F,a,b,\ell,L,\mathcal{E}}(t))_{t\geq 0}$.

Remark 7.3 *In particular, we only consider three types of exit function $\mathcal{E}(x)$, the constant (CON) model, the power exit with fixed boundary (PFB) model and the power exit with no boundary (PNB) model. For the CON model, $\mathcal{E}(x) = C_E$. For the PNB model, $\mathcal{E}(x) = L^{-\alpha_E} x^{\alpha_E}$, $\alpha_E < 0$. For the PFB model, $\mathcal{E}(x) = L^{-\alpha_E} \max\{x^{\alpha_E} - B_E^{\alpha_E}, 0\}$.*

Self-similar fragmentation with immigration processes have been studied extensively in [Haa05]. However, because the exist function $\mathcal{E}(x)$ is very general, the fragmentation with exit process cannot be considered as a self-similar fragmentation process. For example, when $\alpha_E = -1, \alpha_F = 1$, the rate of exit is inversely proportional to length and the rate of fragmentation is proportional to fragment size. It would then be impossible to see exit as dusting from fragmentation events.

When the parameters are clearly understood, we will suppress the parameters and denote the process as $(F_X(t))_{t\geq 0}$. Note that from the construction of the processes, we can infer immediately that the set of red fragments $F_X(t)$ is a subset of the whole process $F(t)$ at all time $t > 0$, and so $F_X(t)$ is stochastically dominated by $F(t)$.

Finally, we denote the set of continuous functions on the interval $[\ell, L]$ as $C([\ell, L])$ for any function $f \in C([\ell, L])$, we define

$$\langle f, F_X(t)\rangle := \sum_{i=1}^{N(t)} f(x_i) \tag{7.3}$$

to be the sum of the function f over all fragment lengths at time t.

7.1.2 FRIME Process as a Fragmentation Process with Immigration

We now characterise the FRIME process as a Fragmentation Process with Immigration.

Fragmentation Processes with Immigration: We introduce the notion of an *X-process with immigration*. A comprehensive construction of these processes is formulated in [Haa05].

We let $\mathcal{R}$ denote the set of Radon point measures on $(0, \infty)$ and equip it with the topology of vague convergence. We say that an $\mathcal{R}$-valued Markov process satisfies the

superposition property if for all $\mathbf{u}, \mathbf{v} \in \mathcal{R}$, the sum of two independent processes $X^{(\mathbf{u})}, X^{(\mathbf{v})}$ starting from $\mathbf{u}, \mathbf{v}$ respectively, is distributed as $X^{(\mathbf{u}+\mathbf{v})}$.

We also consider I, a nonnegative σ-finite measure on $\mathcal{R}$, and let $((\mathbf{s}(v_i), v_i), i \geq 1)$ be the atoms of a Poisson measure with intensity $I(d\mathbf{s})dt, t \geq 0$. Here $\mathbf{s}(v_i) = (s_1(v_i), s_2(v_i), ...)$ corresponds to the list of fragments with length $s_1(v_i), s_2(v_i), ...$ that immigrate into the system at time $v_i > 0$.

Then, the process

$$XI^{(\mathbf{u})}(t) := X^{(\mathbf{u})}(t) + \sum_{v_i \leq t} X^{(\mathbf{s}(v_i))}(t - v_i), \tag{7.4}$$

where the X-processes on the RHS evolve independently, is called the *X-process with immigration* starting from $\mathbf{u}$ with immigration measure $I(d\mathbf{s})$.

Characterisation of FRIME Processes: With the above terminology, we can define a FRIME process as an F_X-process with immigration. We consider an immigration measure $I(d\mathbf{s})$ which only has support on the set of Dirac measures, i.e. only one fragment immigrates into the system per immigration event. In this case, we can simplify $I(d\mathbf{s})$ to

$$I(d\mathbf{s}) = \int_0^\infty C_I p_I(s) \delta_{(s,0,0,...)}(d\mathbf{s}) ds, \tag{7.5}$$

where C_I is the intensity of the immigration process, and $p_I(s)$ is the probability density for the length of an immigrated fragment. In this chapter, we will often consider $I(ds) = C_I/L \mathbb{1}_{s \in [0,L]} ds$, i.e. uniformly distributed immigration.

Consequently, the atoms $((s_{new}(v_i), v_i), i \geq 1)$ of the Poisson measure with intensity $I(ds)dt$ corresponds to the immigration process with one fragment of length $s_{new}(v_i)$ immigrating into the system at time v_i.

Under these assumptions, we can define a FRIME process rigorously.

Definition 7.4 *The $(C_F, \alpha_F, a, b, \ell, L, \mathcal{E}, C_I, p_I)$-Fragmentation with Immigration and Exit (FRIME) Process with initial state μ_0, denoted as $(F_{IX}^{C_F,\alpha_F,a,b,\ell,L,\mathcal{E},C_I,p_I}(t))_{t \geq 0}$ in this chapter, is an $F_X^{C_F,\alpha_F,a,b,\ell,L,\mathcal{E}}$-process with immigration starting from a sequence of fragments μ_0, with immigration measure $I(ds) = C_I p_I(s) ds$.*

From now on, for simplicity we will suppress our parameters and use the notation $F_{IX}(t)$ for the FRIME process, and the notation $F_I(t)$ for the associated fragmentation with immigration process without exit mechanism.

7.2 Stationary Profile of the FRIME Process

In [Haa05], Haas identified conditions for an X-process with immigration to possess a stationary distribution.

Lemma 7.5 (Stationary Distribution of X-processes with Immigration) *For an X-process with immigration defined in* (7.4)*, we define the random variable*

$$\boldsymbol{U}_{stat}^{XI} := \sum_{t_i \geq 0} X^{\boldsymbol{s}(t_i)}(t_i), \tag{7.6}$$

where $(\boldsymbol{s}(t_i), t_i)$ *are the atoms of a Poisson measure with intensity* $I(d\boldsymbol{s})dt$*.*

If $\boldsymbol{U}_{stat}^{XI} \in \mathcal{R}$ *almost surely, then its distribution* $\mathcal{L}(\boldsymbol{U}_{stat}^{XI})$ *is a stationary distribution for* $(XI(t))_{t\geq 0}$*, and for any random* $\boldsymbol{u}$ *such that* $X^{\boldsymbol{u}}(t) \to 0$ *in probability as* $t \to \infty$*,*

$$XI^{(\boldsymbol{u})}(t) \to \boldsymbol{U}_{stat}^{XI}$$

in distribution.

Proof See Lemma 5 in [Haa05]. An illustration of the stationary distribution can be found in Fig 7.1. □

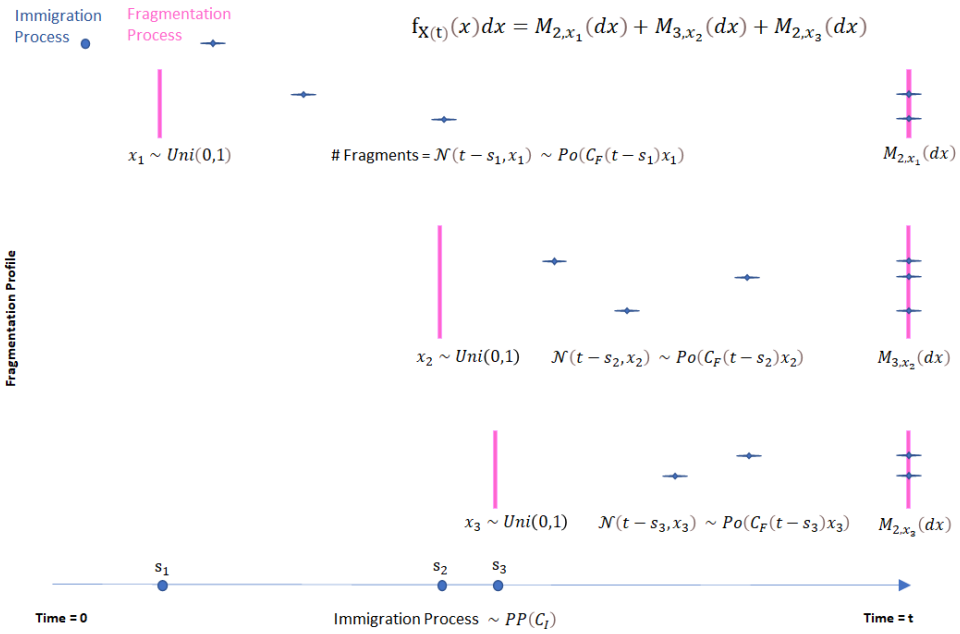


Figure 7.1: Illustration of the stationary profile of a FRIME process Fragments immigrate into the system at s_1, s_2, s_3 according to a Poisson point process with intensity $C_I L$. Fragmentation events occurs independently on each immigrated fragment. Since the immigration process is Poisson and thus reversible, the long term equilibrium for a FRIME process must be the infinite sum of the contributions of subfragment lengths by all past immigrants.

Furthermore, Haas established the existence of the stationary random variable $\mathbf{U}^{FI}_{stat}$ for an (α, c, ν)-fragmentation processes with immigration measure $I(d\mathbf{s})$ (see [Haa05] Theorem 7,8,9).

If we consider the specific case of (C_F, α_F, a, b)-fragmentation processes where the immigration measure satisfies $I(ds) = C_I p_I(s)ds$, we have the following statements on the existence of $\mathbf{U}^{FI}_{stat}$:

- For $\alpha < 0$, the stationary state $\mathbf{U}^{FI}_{stat} \in \ell^p$ a.s. for all $p > 1 + \alpha$ if

$$\int_{s\geq 1} s^{-\alpha} p_I(s)ds < \infty. \tag{7.7}$$

- For $\alpha = 0$, the stationary state $\mathbf{U}^{FI}_{stat} \in \ell^p$ for all $p > 1$ with probability one if

$$\int_{s\geq 1} \log(s) p_I(s)ds < \infty. \tag{7.8}$$

- For $\alpha > 0$, the stationary state $\mathbf{U}^{FI}_{stat} \in \ell^p$ for all $p > 1 + \alpha/(\gamma \wedge 1)$ almost surely if

$$\int_{s\geq 1} s^{\gamma} p_I(s)ds < \infty. \tag{7.9}$$

Existence of Stationary Distribution for FRIME Processes: By the coupling argument implicit in Definition 7.2, we know that the $(C_F, \alpha_F, a, b, \ell, L, \mathcal{E})$-fragmentation with exit process is stochastically dominated by the associated (C_F, α_F, a, b)-fragmentation process. Therefore, the stationary states for their corresponding processes with immigration satisfy

$$\mathbf{U}^{FI}_{stat} := \sum_{t_i \geq 0} F^{\mathbf{s}(t_i)}(t_i) \gg \sum_{t_i \geq 0} F_X^{\mathbf{s}(t_i)}(t_i) = \mathbf{U}^{F_{IX}}_{stat} \tag{7.10}$$

As a result, under Condition (7.7),(7.8) and (7.9), the stationary state for the FRIME process satisfies $\mathbf{U}^{F_{IX}}_{stat} \in \ell^p$ a.s.. By Lemma 7.5, the FRIME process also has a stationary distribution.

7.3 Sample Mean and Hydrodynamic Limit of FRIME Processes

In this section, we will apply the law of large numbers to equate the infinite sample mean of a FRIME process with its hydrodynamic limit. We will also show that the hydrodynamic limit is given by a *fragmentation with immigration and exit* equation.

7.3.1 Law of large numbers for FRIME processes

We now consider n independent and identically distributed FRIME processes $(F^{(i)}_{IX}(t))_{t\geq 0}$, $i = 1, .., n$ with initial state μ_0. For all continuous functions $f \in C([\ell, L])$ on the interval $[\ell, L]$, we define the random variable

$$\langle f, F^{(i)}_{IX}(t)\rangle := \sum_{x \in F^{(i)}_{IX}(t)} f(x) \tag{7.11}$$

to be the sum of f over lengths of all fragments at time t for the i-th FRIME process.

Note that the random variables $\langle f, F^{(i)}_{IX}(t)\rangle$, $i = 1, ..., n$ are real-valued, independent and identically distributed. Therefore, by the Strong Law of Large Numbers, the empirical average of n i.i.d samples of FRIME processes will converge to the expected value of another related FRIME process, i.e. almost surely,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \langle f, F^{(i)}_{IX}(t)\rangle = \mathbb{E}\langle f, F^{(1)}_{IX}(t)\rangle. \tag{7.12}$$

Therefore, to understand the mean of the FRIME process, it suffices for us to understand the limit of $\frac{1}{n}\sum_{i=1}^{n} F_{IX}^{(i)}(t)$ under the weak topology.

Note that for two independently-evolving fragmentation with exit processes (without immigration) satisfying $F_X^{\mu}(t) = \mu, F_X^{\nu}(t) = \nu$ at time t, for all $s > 0$

$$F_X^{\mu}(t+s) + F_X^{\nu}(t+s) =^{d} F_X^{\mu+\nu}(t+s), \tag{7.13}$$

where $(F_X^{\mu+\nu}(t+s))_{s\geq 0}$ is an independent fragmentation with exit process that satisfies $F_X^{\mu+\nu}(t) = \mu+\nu$. Furthermore, note that the sum of n independent Poisson point processes of intensity C_I has the same distribution as a Poisson point process of intensity nC_I.

Therefore, by the superposition property, the sum of n i.i.d $(C_F, \alpha_F, a, b, \ell, L, \mathcal{E}, C_I, p_I)$-FRIME process is again a FRIME process, albeit with a new immigration rate of nC_I. Explicitly, the process is a $(C_F, \alpha_F, a, b, \ell, L, \mathcal{E}, nC_I, p_I)$-FRIME process, with all but one parameters being identical to those of the original FRIME process, and a new immigration rate of nC_I. Similarly, the initial state of this new process is $n\mu_0$.

As a result, the mean of a FRIME process is the hydrodynamic limit when we scale intensity of immigration by a factor of n, initial state by a factor of n, and the mass of a fragment by a factor of $1/n$.

7.3.2 Hydrodynamic Limit of FRIME Processes

Prelimit Model: For a $(C_F, \alpha_F, a, b, \ell, L, \mathcal{E}, C_I, p_I)$-FRIME process $F_{IX}(t)$, we write $F_{IX}^n(t)$ for the $(C_F, \alpha_F, a, b, \ell, L, \mathcal{E}, nC_I, p_I)$-FRIME process with n times the immigration intensity and n times the initial state. Similarly, we write μ_t^n to be the re-scaled empirical distribution of fragment lengths at time t such that each fragment has a mass of $1/n$, i.e.

$$\mu_t^n := \frac{1}{n}\sum_{x\in F_{IX}^n(t)} \delta_x.$$

For each $f \in C([\ell, L])$, we define $X_t^{n,f}$ to be the real-valued projection of μ_t^n through f, i.e.

$$X_t^{n,f} := \frac{1}{n}\sum_{x_i\in F_{IX}^n(t)} f(x_i) = \int f(x)\mu_t^n(dx). \tag{7.14}$$

Identifying the Limit: We evaluate the evolution of $X_t^{n,f} := \int f(x)\mu_t^n(dx)$ for all $f \in \mathcal{C}([\ell, L])$.

To simplify our calculation, we use the notation

$$\mathbb{E}_r[g(r)] = \int_0^1 g(r)\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}r^{a-1}(1-r)^{b-1}dr \tag{7.15}$$

to denote the mean of a function $g(r)$ under the $Beta(a, b)$ distribution.

$$\begin{aligned}
&\lim_{\delta t\to 0}\mathbb{E}\left[\frac{X_{t+\delta t}^{n,f} - X_t^{n,f}}{\delta t}\middle|\mu_t^n = \nu\right]\\
&= \int_\ell^L C_F x^{\alpha_F}\mathbb{E}_r\left[\frac{f(xr)\mathbb{1}_{xr\geq\ell} + f(x(1-r))\mathbb{1}_{x(1-r)\geq\ell} - f(x)}{n}\right]\times n\nu(dx)\\
&\quad + nC_I\int_\ell^L \frac{f(x)}{n}p_I(x)dx - \int_\ell^L \mathcal{E}(x)\frac{f(x)}{n}\times n\nu(dx)\\
&= \int_\ell^L C_F x^{\alpha_F}\mathbb{E}_r\left[f(xr)\mathbb{1}_{xr\geq\ell} + f(x(1-r))\mathbb{1}_{x(1-r)\geq\ell} - f(x)\right]\nu(dx)\\
&\quad + C_I\int_\ell^L f(x)p_I(x)dx - \int_\ell^L \mathcal{E}(x)f(x)\nu(dx).
\end{aligned} \tag{7.16}$$

Similarly,

$$\begin{aligned}
&\lim_{\delta t\to 0}\mathbb{E}\left[\frac{(X_{t+\delta t}^{n,f} - X_t^{n,f})^2}{\delta t}\middle|\mu_t^n = \nu\right]\\
&= \int_\ell^L C_F x^{\alpha_F}\mathbb{E}_r\left[\frac{(f(xr)\mathbb{1}_{xr\geq\ell} + f(x(1-r))\mathbb{1}_{x(1-r)\geq\ell} - f(x))^2}{n^2}\right]\times n\nu(dx)\\
&\quad + nC_I\int_\ell^L \frac{f^2(x)}{n^2}p_I(x)dx + \int_\ell^L \mathcal{E}(x)\frac{f^2(x)}{n^2}\times n\nu(dx)\\
&= \frac{1}{n}\int_\ell^L C_F x^{\alpha_F}\mathbb{E}_r\left[(f(xr)\mathbb{1}_{xr\geq\ell} + f(x(1-r))\mathbb{1}_{x(1-r)\geq\ell} - f(x))^2\right]\nu(dx)\\
&\quad + \frac{1}{n}C_I\int_\ell^L f^2(x)p_I(x)dx + \frac{1}{n}\int_\ell^L \mathcal{E}(x)f^2(x)\nu(dx)
\end{aligned} \tag{7.17}$$

In other words, $(\mu_t^n)_{t\geq 0}$ is a solution to the following martingale problem:

For all $f \in C(\ell, L)$,

$$\begin{aligned}
M_t^{n,f} :=& \int f(x)\mu_t^n(dx) - \int f(x)\mu_0^n(dx)\\
&- \int_0^t\left\{\int_\ell^L C_F x^{\alpha_F}\mathbb{E}_r\left[f(xr)\mathbb{1}_{xr\geq\ell} + f(x(1-r))\mathbb{1}_{x(1-r)\geq\ell} - f(x)\right]\mu_s^n(dx)\right.\\
&\qquad \left. + C_I\int_\ell^L f(x)p_I(x)dx - \int_\ell^L \mathcal{E}(x)f(x)\mu_s^n(dx)\right\}ds
\end{aligned} \tag{7.18}$$

is a martingale, with quadratic variation

$$\langle M_t^{n,f}\rangle = \frac{1}{n}\int_0^t \left\{ \int_\ell^L C_F x^{\alpha_F} \mathbb{E}_r \left[(f(xr)\mathbb{1}_{xr\geq \ell} + f(x(1-r))\mathbb{1}_{x(1-r)\geq \ell} - f(x))^2 \right] \mu_s^n(dx) \right.$$
$$\left. + C_I \int_\ell^L f^2(x) p_I(x) dx + \int_\ell^L \mathcal{E}(x) f^2(x) \mu_s^n(dx) \right\} ds. \tag{7.19}$$

As $n \to \infty$, the quadratic variation term $\langle M_t^{n,f}\rangle \to 0$. Assuming tightness, we have $M_t^{n,f} \to 0$. As a result, the real-valued process $(X_t^{n,f})_{t\geq 0}$ converges in the Skorokhod topology to a deterministic process $(X_t^{\infty,f})_{t\geq 0} := (\int f(x)\mu_t^\infty(dx))_{t\geq 0}$, which solves the FRIME equation

$$\frac{\partial}{\partial t}\int f(x)\mu_t^\infty(dx) = C_F \int_\ell^L x^{\alpha_F} \mathbb{E}_r \left[f(xr)\mathbb{1}_{xr\geq \ell} + f(x(1-r))\mathbb{1}_{x(1-r)\geq \ell} - f(x) \right] \mu_t^\infty(dx)$$
$$- \int_\ell^L \mathcal{E}(x) f(x) \mu_t^\infty(dx) + C_I \int f(z) p_I(z) dz, \tag{7.20}$$

with initial condition $\mu_0^\infty(dx) = \mathbb{1}_{[\ell,L]}(x)\frac{1}{L-\ell}dx$.

Remark 7.6 *The equation without the exit term $-\int_\ell^L \mathcal{E}(x)f(x)\mu_t^\infty(dx)$ has been studied as the fragmentation with immigration equation in [Haa05] Equation (E).*

7.4 Solutions to the FRIME equations

We now consider more explicit expressions for solutions to the FRIME equation (7.20).

7.4.1 Density solution to FRIME equations

We assume that the solution $\mu_t^\infty(dx)$ is differentiable in space and time, i.e. there exists a function $u_t(x) : [0,\infty) \times [\ell, L] \mapsto \mathbb{R}$ differentiable in t and continuous in x such that $\mu_t^\infty(dx) = u_t^\infty(x)dx$.

Furthermore, recall that for $r \sim Beta(a,b)$, the ratio $\widetilde{r} := 1 - r$ is distributed as $Beta(b,a)$. For simplicity, we once again use the notation presented in Equation (7.15) and similarly, we write

$$\mathbb{E}_{\widetilde{r}}[g(\widetilde{r})] = \int_0^1 g(\widetilde{r})\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\widetilde{r}^{a-1}(1-\widetilde{r})^{b-1}d\widetilde{r} = \int_0^1 g(r)\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}r^{b-1}(1-r)^{a-1}dr. \tag{7.21}$$

Applying Fubini's theorem and substituting $s = r^{-1}$, we have

$$\begin{aligned}
&\int_\ell^L f(x)\left\{\frac{\partial}{\partial t}u_t(x) + C_F x^{\alpha_F} + \mathcal{E}(x) - C_I p_I(x)\right\}dx \\
=&C_F\mathbb{E}_r\left[\int_\ell^L x^{\alpha_F} f(xr)\mathbb{1}_{xr\geq\ell}u_t(x)dx\right] + C_F\mathbb{E}_r\left[\int_\ell^L x^{\alpha_F} f(x(1-r))\mathbb{1}_{x(1-r)\geq\ell}u_t(x)dx\right] \\
=&C_F\mathbb{E}_r\left[\int_\ell^L x^{\alpha_F} f(xr)\mathbb{1}_{r\geq x^{-1}\ell}u_t(x)dx\right] + C_F\mathbb{E}_{\widetilde{r}}\left[\int_\ell^L x^{\alpha_F} f(x\widetilde{r})\mathbb{1}_{\widetilde{r}\geq x^{-1}\ell}u_t(x)dx\right] \\
=&C_F\mathbb{E}_r\left[r^{-\alpha-1}\int_\ell^{rL} y^{\alpha_F} f(y)u_t(y/r)dy\right] + C_F\mathbb{E}_{\widetilde{r}}\left[\widetilde{r}^{-\alpha-1}\int_\ell^{\widetilde{r}L} y^{\alpha_F} f(y)u_t(y/\widetilde{r})dy\right] \\
=&C_F\int_0^1\left\{r^{-\alpha-1}\int_\ell^{rL} y^{\alpha_F} f(y)u_t(y/r)dy\right\}\times\left\{\frac{r^{a-1}(1-r)^{b-1}+r^{b-1}(1-r)^{a-1}}{B(a,b)}\right\}dr \\
=&\int_\ell^L C_F y^{\alpha_F}\left\{\int_{y/L}^1 r^{-(\alpha-1)}\frac{r^{a-1}(1-r)^{b-1}+r^{b-1}(1-r)^{a-1}}{B(a,b)}u_t(y/r)\times r^{-2}dr\right\}f(y)dy \\
=&\int_\ell^L C_F y^{\alpha_F}\left\{\int_1^{L/y}\frac{1}{B(a,b)}s^{\alpha-a-b+1}\left((s-1)^{b-1}+(s-1)^{a-1}\right)u_t(ys)ds\right\}f(y)dy.
\end{aligned} \tag{7.22}$$

As a result, we have established $\mu_t^\infty(dx)$ as a weak solution to the Equation

$$\begin{aligned}
\frac{\partial}{\partial t}u_t(x) =&C_F x^{\alpha_F}\left\{\int_1^{L/x}\frac{s^{\alpha-a-b+1}\left((s-1)^{b-1}+(s-1)^{a-1}\right)}{B(a,b)}u_t(xs)ds - u_t(x)\right\} \\
&-\mathcal{E}(x)u_t(x) + C_I p_I(x),
\end{aligned} \tag{7.23}$$

with initial condition $u_0(x) = \frac{1}{L-\ell}$ on $[\ell, L]$.

In particular, if we consider uniform fragmentation processes with $a = b = 1$, the fragmentation ratio is uniform, i.e. $r \sim Uni[0,1]$, and $u_t(x)$ solves

$$\frac{\partial}{\partial t}u_t(x) = C_F x^{\alpha_F}\left\{2\int_1^{L/y} s^{\alpha-1}u_t(xs)\,ds - u_t(x)\right\} - \mathcal{E}(x)u_t(x) + C_I p_I(x), \tag{7.24}$$

with initial condition $u_0(x) = \frac{1}{L-\ell}$ on $[\ell, L]$.

Remark 7.7 *One can also construct Equation* (7.24) *by considering $u_t(x)$ as the mass of fragments with length x at time t. The integral term on the R.H.S. of* (7.24) *corresponds to the increase in mass at length x due to fragmentation of larger fragments. The $-u_t(x)$ term corresponds to loss of mass at length x due to fragmentation. The last two terms correspond to fragments exiting the system at size-dependent rate $\mathcal{E}(x)$, and new fragments immigrating into the system with a size-dependent rate $C_I p_I(x)$.*

7.4.2 Stationary solution to FRIME equations

Since the mean measure of the FRIME process can be expressed as a solution to the FRIME Equation in (7.20), and its density (if it exists) satisfies the density equation (7.23), one can study the stationary distribution of the FRIME process with the stationary solution to (7.23).

Let $S(x)$ be a stationary solution to (7.23), then for $x \in [\ell, L]$,

$$\frac{C_F x^{\alpha_F}}{B(a,b)} \int_1^{L/x} s^{\alpha-a-b+1} \left((s-1)^{b-1} + (s-1)^{a-1}\right) S(xs)ds + C_I p_I(x) = (C_F x^{\alpha_F} + \mathcal{E}(x))S(x). \tag{7.25}$$

Furthermore, at stationarity, the rate of mass increase for fragments with size L due to immigration ($C_I p_I(L)$) must offset the rate of mass loss for fragments with size L due to exit ($\mathcal{E}(L)S(L)$) and fragmentation ($C_F L^{\alpha_F}$). Therefore, $S(x)$ must satisfy the boundary condition $S(L) = \frac{C_I p_I(L)}{C_F L^{\alpha_F} + \mathcal{E}(L)}$.

We first consider the case where $a = b = 1$. If we substitute $z = xs$, Equation (7.25) becomes

$$2C_F \int_x^L z^{\alpha-1} S(z)dz + C_I p_I(x) = (C_F x^{\alpha_F} + \mathcal{E}(x))S(x),$$

and taking the derivative with respect to x on both sides, we have

$$-2C_F x^{\alpha-1} S(x) + C_I p_I'(x) = (\alpha C_F x^{\alpha-1} + \mathcal{E}'(x))S(x) + (C_F x^{\alpha_F} + \mathcal{E}(x))S'(x),$$

and therefore

$$S'(x) + \frac{\mathcal{E}'(x) + (\alpha_F + 2)C_F x^{\alpha-1}}{C_F x^{\alpha_F} + \mathcal{E}(x)} S(x) = \frac{C_I p_I'(x)}{C_F x^{\alpha_F} + \mathcal{E}(x)}. \tag{7.26}$$

By the variation of parameters method for first order linear differential equations,

$$S(x) = A_1 \exp(-P(x)) + A_2 \exp(-P(x)) \int_\ell^x \frac{C_I p_I'(z)}{C_F z^{\alpha_F} + \mathcal{E}(z)} \exp(P(z))dz, \tag{7.27}$$

where $A_1, A_2 > 0$ are constants determined by the boundary conditions and the integrating factor $P(x)$ is given by

$$P(x) = \int_\ell^x \frac{\mathcal{E}'(z) + (\alpha_F + 2)C_F z^{\alpha-1}}{C_F z^{\alpha_F} + \mathcal{E}(z)} dz. \tag{7.28}$$

Finally, at stationarity, the net mass of fragments with length between $[L - dx, L]$ immigrating into the system must be equal to the net mass of fragments with length between $[L - dx, L]$ leaving the system through fragmentation, so $S(x)$ must satisfy the boundary condition

$$S(L)C_F L^{\alpha_F} + \mathcal{E}(L) = C_I. \tag{7.29}$$

We now consider the shape of the stationary profile for various scenarios of exit mechanisms.

No Exit Mechanism: We consider the case where there is only a killing boundary, i.e. $\mathcal{E}(x) = 0$ for $x \geq \ell$ and all fragments with length smaller than ℓ dies instantaneously. In this case,

$$P(x) = \int_\ell^x (\alpha_F + 2) z^{-1} dz = (\alpha_F + 2) \log(x) + C_1, \tag{7.30}$$

so $\exp(P(x)) = C_2 x^{(\alpha_F+2)}$, and

$$S(x) = A_1 x^{-(\alpha_F+2)} + A_2 x^{-(\alpha_F+2)} \int_\ell^x \frac{C_I}{C_F} p_I'(z) z^2 dz, \tag{7.31}$$

where A_1, A_2 are some constants determined by the boundary condition.

If we further assume that immigrating fragments are uniformly distributed in $[\ell, L]$, the probability density function for the immigration length satisfies $p_I(z) = \frac{1}{L-\ell} \mathbb{1}_{[\ell,L]}(z)$. As a result, the stationary profile satisfies $S(x) = Ax^{-\alpha-2}$. Finally, using the boundary condition we have $C_I dx = AL^{-\alpha_F-2} \times C_F L^{\alpha_F} dx$, and so $A = \frac{C_I}{C_F} L^2$.

To conclude, the stationary profile is of the form

$$S(x) = \frac{C_I}{C_F} L^2 x^{-\alpha-2}. \tag{7.32}$$

The same result was established in [Haa05] in full and in [BK00] for the special case where $\alpha = 1$.

Constant Exit Rate: (CON) We may also consider the case where $\mathcal{E}(x) \equiv C_E$. Then

$$P(x) = \int_\ell^x \frac{(\alpha_F + 2)C_F z^{\alpha_F - 1}}{C_F z^{\alpha_F} + C_E} dz = \frac{\alpha_F + 2}{\alpha_F} \log(C_F x^{\alpha_F} + C_E) + C, \tag{7.33}$$

and so

$$S(x) = A_1(C_F x^{\alpha_F} + C_E)^{-\frac{\alpha_F+2}{\alpha_F}} + A_2(C_F x^{\alpha_F} + C_E)^{-\frac{\alpha_F+2}{\alpha_F}} \int_\ell^x C_I p_I'(z)(C_F z^{\alpha_F} + C_E)^{\frac{2}{\alpha_F}} dz, \tag{7.34}$$

where $C_1, A_1, A_2 > 0$ are some constants.

If we consider uniform immigration, then $p_I'(z) \equiv 0$, so the stationary profile is of the form

$$S(x) = A(C_F x^{\alpha_F} + C_E)^{-\frac{\alpha_F+2}{\alpha_F}}, \tag{7.35}$$

with derivative

$$S'(x) = -\frac{\alpha_F + 2}{\alpha_F} A(C_F x^{\alpha_F} + C_E)^{-\frac{2\alpha_F+2}{\alpha_F}} C_F x^{\alpha_F - 1} < 0. \tag{7.36}$$

The stationary profile is therefore strictly decreasing and exhibitis no peak.

Power Law with Fixed Boundary (PFB): If we consider $\alpha_F = 1, L = 1$ and take the exit function as $\mathcal{E}(x) = \max\{x^{\alpha_E} - B_E^{\alpha_E}, 0\}$, then for $x < B_E$,

$$\log S(x) = C_1 - \int_\ell^x \frac{3C_F + \alpha_E z^{\alpha_E - 1}}{z^{\alpha_E} + C_F z - B_E^{\alpha_E}} dz, \tag{7.37}$$

where C_1 is some constant. A peak is formed at position $p < B_E$ only if $\frac{d}{dx} \log S(x) < 0$ for $x \in [p, B_E]$.

In Figure 6.3, $C_F = 1, B_E = 0.4$. Therefore,

$$\frac{d}{dx} \log S(x) = \frac{3 + \alpha_E x^{\alpha_E - 1}}{x^{\alpha_E} + x - 0.4^{\alpha_E}}.$$

The expression on the right has a root at $x = \frac{1}{3|\alpha_E|}^{\frac{1}{|\alpha_E|}}$. In particular, the root is smaller than 0.4 if and only if $|\alpha_E| < 1.097$. This aligns with the observation that in the PFB panel in Figure 6.3 in Chapter 6, the peak of the fragmentation profile is clearly to the left of the exit boundary when $\alpha_E = -0.25$.

Power Law with No Boundary (PNB): If we consider $\alpha = 1$, and consider the exit function $\mathcal{E}(x) = L^{-\alpha_E} x^{\alpha_E}$. In that case,

$$\log S(x) = C_1 - \int_\ell^x \frac{3C_F + L^{-\alpha_E}\alpha_E z^{\alpha_E - 1}}{z^{\alpha_E} L^{-\alpha_E} + C_F z} dz, \tag{7.38}$$

so $S(x) = C_2 x^{-\alpha_E}(C_F L^{\alpha_E} x^{1-\alpha_E} + 1)^{\frac{3-\alpha_E}{\alpha_E - 1}}$, where C_1, C_2 are constants determined by the boundary condition.

This gives the best fit line that is very close to simulated data in Figure 6.3 in Chapter 6.

Linear Model: Here we assume that $\mathcal{E}(z) = C_E - m_E z$ and $m_E \neq C_F$ for all $z \leq B_E := \frac{C_E}{m_E}$. If we take $\alpha = 1$, then

$$P(x) = \int_\ell^x \frac{3C_F - m_E}{C_E + (C_F - m_E)z} dz = -\frac{3C_F - m_E}{C_F - m_E} \log(C_E + x(C_F - m_E)) + C_1, \tag{7.39}$$

for some constant C_1,

$$\begin{aligned} S(x) =& A_1(C_E + x(C_F - m_E))^{\frac{3C_F - m_E}{C_F - m_E}} \\ &+ A_2(C_E + x(C_F - m_E))^{\frac{3C_F - m_E}{C_F - m_E}} \int_\ell^x \frac{C_I p_I'(z)}{C_F z^{\alpha_F} + \mathcal{E}(z)} (C_E + z(C_F - m_E))^{-\frac{3C_F - m_E}{C_F - m_E}} dz, \end{aligned} \tag{7.40}$$

where $A_1, A_2 > 0$ are positive constants.

If we assume uniform immigration, i.e. $p_I(x) = 1$, then, for some constant positive factor $A > 0$,

$$S(x) = A \times \left(B_E + x \left(\frac{C_F}{m_E} - 1 \right) \right)^{(m_E - 3C_F)/(C_F - m_E)} \tag{7.41}$$

which increases with x in $[0, B_E]$ only if $m_E > 3C_F$. In particular, when $m_E = 2C_F$, the stationary profile $S(x)$ decreases linearly from 0 to B_E, and when $m_E = 3C_F$, the stationary profile $S(x)$ is roughly constant in $[0, B_E]$. This aligns with the observation that B_E and the mode of the length frequencies coincide whenever $m_E > 3C_F$, see Figure B.5.

8
Conclusion

In this thesis, we have shown how the mathematical theory of branching and fragmenting particle systems can provide the biology community with novel qualitative insights. For population geneticists, we have created a new framework to study genealogy and population dynamics for a wide class of spatially-structured populations. For oncologists, we have shown in our clinical data analysis that the FRIME model highly resembles real life data with a few simple hypotheses, and can be used to draw inferences from mtDNA fragment profiles. Although gDNA fragment profiles have an inherent cyclical nature, once we adjust our fragment profile by binning data, i.e. by effectively measuring fragment lengths in units of nucleosomes, the gDNA fragment profile also fits into the FRIME model and the log-log linearity model. We believe that our model reveals insights about cfDNA kinetics that would otherwise be unethical to measure in patients. We look forward to the general application of the FRIME model in other means of liquid biopsy as well, such as circulating RNA and proteins.

8.1 Further Research Questions

8.1.1 Lookdown Representation and an analogue of the Wright-Malécot formula

With the description of ancestral lineages in Theorem 2.24, we can attempt a similar argument to find the probability that the ancestral lineages of individuals sampled at y and z at time t coalesce in an infinitesimal neighbourhood of the point x and time 0. Once again we work formally (with the understanding that integration against suitable test functions will be required to make a more rigorous statement) and we proceed by finding the intensity of individuals at y and z that are descended from parent-child pairs originating from reproduction events in the time interval $[0, \delta t)$ in which the parent was at x at time zero.

This time we need $1/N^2$ times the number of pairs $(U_{\tau'}, U_\tau)$ corresponding to a parent that is at x at time 0, and an offspring born in the infinitesimal time interval $[0, \delta t)$, for which $U_{\tau'} \exp(\int_0^t b(Y_s^{(1)})ds) < N$, $Y_t^{(1)} = y$; and $U_\tau \exp(\int_0^t b(Y_s^{(2)})ds) < N$, $Y_t^{(2)} = z$, or the symmetric event in which $Y_t^{(1)} = z$ and $Y_t^{(2)} = y$, where $Y^{(1)}$ is the spatial position of the parent and $Y^{(2)}$ is the spatial motion of the offspring. Recalling from (5.22) that $U_\tau = U_{\tau'} + v/(2\alpha\gamma(x, \Xi_0)r(x, \Xi_0))$, where $(s, v) \in \xi_\tau$ for some $s \in [0, \delta t)$, formally, knowing $\{\Xi_s\}_{0 \le s \le t}$, the intensity that we seek takes the form

$$\Xi_0(x) 2\alpha\gamma(x, \Xi_0) r(x, \Xi_0) \mathbb{E}_x \left[\exp\left(\int_0^t b(Y_s^{(1)}, \Xi_s) ds + \int_0^t b(Y_s^{(2)}, \Xi_s) ds \right) \mathbf{1}_{Y_t^{(1)} = y} \mathbf{1}_{Y_t^{(2)} = z} \right] dy dz.$$

To recover the (density for the) probability of coalescence for sampled individuals, we divide by $\Xi_t(y)\Xi_t(z)$.

With this we can present a framework to obtain the genealogies of a sample in the population. Recall our notation $\alpha = \lim_N \theta(N)/N$ from §2.2. If we consider the stochastic equation with α close to 0, then it is easy to believe that the evolution of the population will be approximated by the deterministic limit and, correspondingly, the motion of ancestral lineages will be close to that found for the deterministic limit in our calculations above. In that notation, the probability of coalescence of the ancestral lineages at x at time 0 is approximately

$$\frac{2\alpha\gamma(w(x))r(w(x))}{w(x)} p(t, y, x) p(t, z, x) dt dx.$$

On the contrary, if α is so large that on timescales of order α^{-1} the ancestral lineages have reached stationarity, then tracing back in time, we expect that the rate of pairwise coalescence will be approximately

$$\int \alpha \frac{2c(w(x))}{w(x)} m(x)^2 dx, \tag{8.1}$$

where m is the stationary distribution. We observe that this approximation does not work for the classical Fisher-KPP equation, for which the integral in (8.1) diverges. In that setting, lineages are forced into a front of width approximately $\log \alpha^{-1}$, within which they behave approximately as Brownian motion. The population size there is too small to ignore fluctuations, and indeed over timescales of the order of $(\log \alpha^{-1})^3$ we expect to see fluctuations in which a significant proportion of the population in the front is descended from a single individual. By contrast, in the case of the porous medium equation with logistic growth, the expression in (8.1) converges. We conjecture that in that setting, on timescales of order α^{-1}, as $\alpha \to 0$, the genealogy of a sample from a population expanding according to the porous medium equation with logistic growth (and indeed many other reaction diffusion equations with nonlinear diffusion) will converge to the Kingman coalescent.

To give a proper proof of the last results we would need to control the approximations made; namely, how fast we have convergence to Y and Ξ in the particle system and how close the stochastic front is to the deterministic for α close to 0. The rate of convergence of Y can be disentangled from the estimates in §3.2 and the rate of convergence to Ξ can be similarly deduced, at least in the case of semi-linear heat equations. This says that, in theory, by controlling how fast the stochastic front approximate the deterministic we should be able to fill in the details in the last computations, in a similar vein to what is done in [EP22]. The novelty would not only include the ability to be able to work with a large family of equations, but also that we only need to control rates of convergence and not the derivative of the population profiles. This idea will be pursued in the future.

8.1.2 FRIME - Fast Implementation and Fragmentation Rate Inference

The algorithm presented in §6.2.4 has two main limitations. First, it relies on calculations of Kolmogorov-Smirnov distances on the fly to determine convergence. Furthermore, it relies on the use of an auxiliary time variable t which has no biological interpretation. As a result, we cannot a priori determine the level of uncertainty before running our simulations. We also cannot relate the parameters C_I, C_F with existing notion of DNA fragmentation rates used in the biology community.

We will now propose a new approach to simulate stationary profiles, and demonstrate how that can help us relate our parameters to biological measurements.

Recall that the stationary profile of a FRIME process can be constructed according to Equation (7.6), i.e.

$$\sum_{t_i \geq 0} F_X^{\mathbf{s}(t_i)}(t_i) = \mathbf{U}_{stat}^{F_{IX}},$$

where F_X is the fragmentation with exit process, $t_i : i \geq 1$ are sampled from a Poisson point process with intensity C_I, and $\mathbf{U}_{stat}^{F_{IX}}$ is the stationary distribution of the FRIME process.

By construction, all fragments with length smaller than ℓ will exit the system immediately, so the fragmentation with exit process $(F_X^L(t))_{t\geq 0}$ starting with a fragment of length L satisifies $\lim_{t\to\infty} F_X^L(t) = 0$ a.s.. Furthermore, by the Borel-Cantelli Lemma, if we can establish that

$$\sum_{i=1}^{\infty} \mathbb{P}(F_X^{\mathbf{s}(t_i)}(t_i) \neq 0) < \infty,$$

then we can conclude that there exists almost surely some threshold $T < \infty$ such that $F_X^{\mathbf{s}(t_i)}(t_i) = 0$ for all $t_i > T$. In other words, we only need to simulate a finite number of immigrated fragments and their descendants to determine the stationary profile of a FRIME process.

Note however that this threshold T is not fixed for different realisations of FRIME. In practice, we may want to fix some $\widetilde{T}$ a priori, and only simulate fragments immigrated into the system at $\widetilde{T}$ unit of time ago, i.e. we compute the approximate stationary profile

$$\widetilde{\mathbf{U}}_{stat}^{F_{IX}} := \sum_{t_i \leq \widetilde{T}} F_X^{\mathbf{s}(t_i)}(t_i).$$

For our algorithm to be efficient and effective, we must pick some reasonably small $\widetilde{T}$ such that $\widetilde{U}_{stat}$ is close to the true U_{stat}. One way to pick this T is to refer to results in Ex.3. [Haa21] which controls the decay rate of the largest fragment for a Beta fragmentation processes. For example, the extinction time ζ for a binary fragmentation process with $\alpha = -1, a = 1, b = 1$ satisfies

$$\mathbb{P}(\zeta > t) \approx \exp(-t)t^2.$$

Therefore, to ensure our simulation produce a good estimated stationary profile, we should pick T_ϵ such that

$$\mathbb{E}\left[\sum_{t_i \geq T} 1_{F_X^{\mathbf{s}(t_i)}(t_i)} \neq 0\right] = C_I \int_{T_\epsilon}^{\infty} \exp(-t)t^2 dt < \epsilon. \tag{8.2}$$

This choice can be further improved if we also consider the effect of exit mechanism.

Finally, the same estimate on the size of a largest fragment can help us infer the real biological fragmentation rate. Experimentally, oncologists mix cell-free DNA fragments into DNAse and record the time until the length of all fragments are below a detection threshold (approximately 1-10 b.p.). By tuning parameters of our simulation with observed data, we can utilise the above estimates to infer parameters of C_F and C_I. This will provide the biology community with new biomarkers that have potentials in identifying cancers.

8.1.3 FRIME - A Branching Particle System Perspective

Finally, note that the calculations for uniform FRIME processes make use of the Poissonian properties of fragmentations. However, calculating the variance of fragment sizes will require a new method.

We can associate a FRIME process with a dyadic branching particle system with state-dependent deaths. The construction laid out here is similar to that in [CR21].

First, we consider fragments with length $x \in [0, \infty]$. We then let $X := \{x_1, .., x_N\}$ be the lengths for a collection of fragments. We will consider the empirical measure of the form

$$\mathcal{L}(X) := \sum_{x \in X} \delta_{\ell_x},$$

where $\ell_x := -\log(x)$.

Recall that a fragment of length x at rate $C_F x^\alpha$ fragments into two sub-fragments x_1, x_2 with lengths xr and $x(1-r)$ respectively, where $r \sim Beta(a, b)$.

Correspondingly, a particle at location ℓ_x in $\mathcal{L}(X)$ at rate $C_F \exp(-\alpha \ell_x)$ split into two particles ℓ_{x_1} and ℓ_{x_2}, whose locations are now $\ell_{x_1} - \log r$ and $\ell_{x_1} - \log(1-r)$ respectively. Note that when $r \sim Uni[0, 1]$, we have correlated exponential jumps.

Similarly, at some rate $R(\ell_x)$, ℓ_x dies and goes to some cemetery state δ.

Finally, with rate C_I, a particle immigrates into the system and its position is at $\ell_{x_i} \sim J$ for some non-negative random distribution J. Here $\exp(-J)$ corresponds to the distribution of an immigrating fragment.

We believe further studies and potential applications of many-to-few lemmas on this system will help study higher moments of fragment sizes.

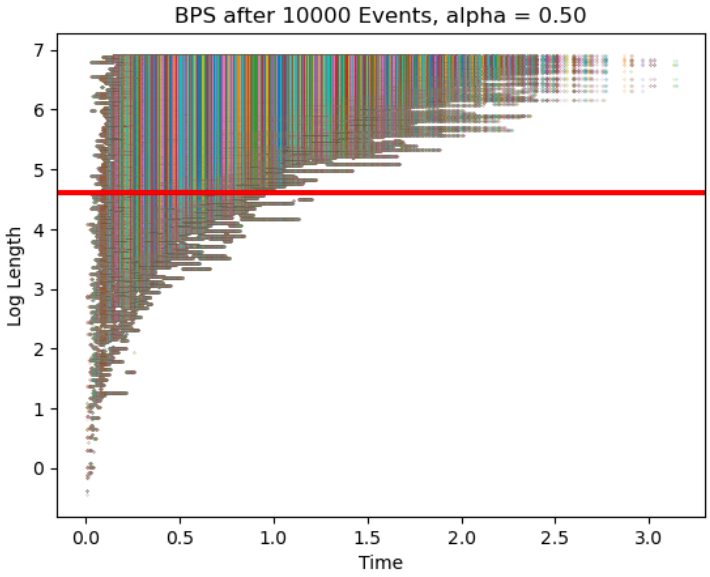


Figure 8.1: Illustration of FRIME process in log scale, $\alpha = 0.5$ This figure shows the evolution in time of fragment sizes under the length transformation $x \to -\log(x)$. As the particle systems evolves, the largest fragment decays as $t^{-\beta}$.

8.2 Final Remarks

There is still a long way to go before the mathematics established in this thesis can deliver truly transformative impact. We are still at the nascent age of genomics. There are only few spatially-structured genetic databases at a population level for biologists to verify and

apply the mathematics of lookdown representations. Further studies in cell-free DNA are needed to verify the biological assumptions underlying the FRIME framework.

Having said that, there are avenues where the mathematics presented here can make huge transformative impact. One such exciting field is spatial transcriptomics. Spatial transcriptomics is a cutting-edge technique in the field of genomics that enables researchers to study gene expression patterns within intact tissue samples while preserving their spatial information. Traditional transcriptomics methods typically involve isolating RNA from bulk tissue and analyzing gene expression levels averaged across all the cells within the sample. However, this approach overlooks the fact that cells within a tissue exhibit remarkable diversity and specialization, each performing unique functions based on their specific location within the tissue. Spatial transcriptomics seeks to overcome this limitation by providing a spatially resolved view of gene expression within the tissue. It allows researchers to observe which genes are expressed and at what levels in different regions or specific cell types of interest, providing valuable insights into the underlying biological processes and cellular interactions within the tissue.

The mathematics of population genetics can provide valuable insights and analytical tools for studying spatial transcriptomics. For example, the tools we used to analyse ancestral lineage dynamics can help researchers identify whether observed spatial patterns are due to genetic drift-like processes or other factors such as cell migration or developmental gradients. Similarly, spatial transcriptomics can reveal spatially defined subpopulations of cells with unique gene expression profiles. By applying our analysis on clustering phenomena and density-dependent regulation to spatial transcriptomics data, researchers can identify and characterize these subpopulations, unraveling the cellular heterogeneity within tissues.

Another exciting field where our mathematics can be helpful is nuclear engineering. The field of nuclear engineering is witnessing remarkable advances in next-generation reactor design, offering promising solutions for sustainable energy through technologies like small modular reactors. These advances rely on the mathematics of radiation transport, specifically the Boltzmann transport equation (BTE). To accurately predict radiation-matter interactions in complex environments such as nuclear reactors with

the BTE, Monte Carlo methods and branching particle systems are often employed. Despite their historical association with the civil nuclear era, these mathematical theories have evolved independently and found applications in population genetics (superprocess theory) and Bayesian statistics (Markov Chain Monte Carlo methods). Meanwhile, engineers have developed their own methods (e.g. Russian roulette and weight windows) to address challenges in reactor simulations. However, knowledge exchange between nuclear engineering and mathematical research communities has been limited, hindering interdisciplinary progress. Neutron clustering is a phenomenon observed in Monte Carlo simulations where, to maintain a stable particle count, some particles are randomly killed during fission events. Such models with killing, known as Fleming-Viot models (or Moran model), have been well understood by probabilists in the context of genetic drift. In particular, it is shown that critical branching particle systems exhibit clustering (the pain in the torus) contrary to the intuition of population geneticists. To circumvent this issue, new models including the spatial Lambda-Fleming-Viot (SLFV) process [BEV10] or branching particle systems with local interactions are proposed. The same clustering phenomenon troubles the nuclear physics community when it comes to fission source convergence. In short, algorithms to evaluate the criticality of nuclear reactor designs tend to underestimate the criticality due to clustering phenomenon. Furthermore, in decoupled heterogeneous media, such as reactors with fissile materials separated by water, events in the core take a long time to propagate, further complicating convergence assessment. Quantifying and predicting spatial correlations and clustering in heterogeneous environments is therefore crucial for improving the reliability of nuclear reaction simulations and enhancing the safety of existing and future reactors. In the future, I hope to facilitate cross-fertilization of ideas and methodologies from stochastic modelling of biological systems and nuclear physics to solve the above problems.

To conclude, the mathematics of branching and fragmenting particle systems can and will produce transformative research relevant to multiple scientific disciplines. I look forward to the day where the mathematics presented in this thesis serves as a basis or inspiration for outstanding scientific research to come.

A Appendix

A.1 Markov Mapping Theorem

The following appears as Theorem A.2 in [EK19], specialized slightly here to the case that the processes are càdlàg and have no fixed points of discontinuity. For an S_0-valued, measurable process Y, $\widehat{\mathcal{F}}_t^Y$ denotes the completion of the σ-algebra generated by $Y(0)$ and $\{\int_0^r h(Y(s))ds, r \leq t, h \in B(S_0)\}$. Also, let $D_S[0,\infty)$ denote the space of càdlàg, S-valued functions with the Skorohod topology, and $M_S[0,\infty)$ the space of Borel measurable functions from $[0,\infty)$ to S, topologized by convergence in Lesbegue measure. For other definitions see [EK19].

Theorem A.1 (Markov Mapping Theorem) *Let (S,d) and (S_0,d_0) be complete, separable metric spaces. Let $A \subset C_b(S) \times C(S)$ and $\psi \in C(S)$, $\psi \geq 1$. Suppose that for each $f \in \mathcal{D}(A)$ there exists c_f such that*

$$|Af(x)| \leq c_f \psi(x), \qquad x \in A,$$

and define $A_0 f(x) = Af(x)/\psi(x)$.

Suppose that A_0 is a countably determined pre-generator, and suppose that $\mathcal{D}(A) = \mathcal{D}(A_0)$ is closed under multiplication and is separating. Let $\gamma : S \to S_0$ be Borel measurable, and let α be a transition function from S_0 into S ($y \in S_0 \to \alpha(y,\cdot) \in \mathcal{P}(S)$ is Borel

measurable) satisfying $\int h \circ \gamma(x)\alpha(y,dx) = h(y)$ *for* $y \in S_0$ *and* $h \in B(S_0)$*, that is,* $\alpha(y,\gamma^{-1}(y)) = 1$*. Assume that* $\widetilde{\psi}(y) \equiv \int_S \psi(z)\alpha(y,dz) < \infty$ *for each* $y \in S_0$ *and define*

$$C = \{\int_S f(z)\alpha(\cdot,dz), \int_S Af(z)\alpha(\cdot,dz) \ : \ f \in \mathcal{D}(A)\}. \tag{1.1}$$

Let $\mu_0 \in \mathcal{P}(S_0)$ *and define* $\nu_0 = \int \alpha(y,\cdot)\mu_0(dy)$*.*

(a) If $\widetilde{Y}$ *satisfies* $\int_0^t \mathbb{E}[\widetilde{\psi}(\widetilde{Y}(s))]ds < \infty$ *for all* $t \geq 0$ *and* $\widetilde{Y}$ *is a solution of the martingale problem for* (C,μ_0)*, then there exists a solution* X *of the martingale problem for* (A,ν_0) *such that* $\widetilde{Y}$ *has the same distribution on* $M_{S_0}[0,\infty)$ *as* $Y = \gamma \circ X$*. If* Y *and* $\widetilde{Y}$ *are cádlág, then* Y *and* $\widetilde{Y}$ *have the same distribution on* $D_{S_0}[0,\infty)$*.*

(b) For $t \geq 0$*,*

$$\mathbb{P}\{X(t) \in \Gamma \mid \widehat{\mathcal{F}}_t^Y\} = \alpha(Y(t),\Gamma), \qquad \text{for } \Gamma \in \mathcal{B}(S).$$

(c) If, in addition, uniqueness holds for the martingale problem for (A,ν_0)*, then uniqueness holds for the* $M_{S_0}[0,\infty)$*-martingale problem for* (C,μ_0)*. If* $\widetilde{Y}$ *has sample paths in* $D_{S_0}[0,\infty)$ *then uniqueness holds for the* $D_{S_0}[0,\infty)$*-martingale problem for* (C,μ_0)*.*

(d) If uniqueness holds for the martingale problem for (A,ν_0) *then* Y *is a Markov process.*

In our application, we have taken S to be the space of locally finite counting measures on $\mathbb{R}^d \times [0,N)$ or on $\mathbb{R}^d \times [0,\infty)$, and S_0 the space of finite measures on $\overline{\mathbb{R}}^d$. Then, A corresponds to the generator for the lookdown process (i.e., either A^N or A), and C corresponds to the generator for the spatial population process (i.e., either $\mathcal{P}^N$ or $\mathcal{P}$). The "γ" of the theorem is our spatial projection operator that we have called κ^N or κ, and the "α" of the theorem will be named Γ^N or Γ below. Finally, "X" of the theorem is our lookdown process, ξ, and "Y" is our spatial process, η.

A.1.1 Lookdown Generators

In this section we verify one of the conditions of the Markov Mapping Theorem, namely, that "integrating out levels" in the generator of the lookdown process we obtain the generator of the projected process. In the notation of the theorem, we are verifying that C defined in (1.1) is in fact $\mathcal{P}^N$ (if defined with A^N) or $\mathcal{P}^\infty$ (if defined with A). We will work with test functions of the form

$$f(\xi) = \prod_{(x,u)\in\xi} g(x,u) = \exp\left(\langle \log g, \xi\rangle\right), \tag{1.2}$$

where $0 \le g \le 1$ and $g(x,u) = 1$ for all $u \ge u_g$ for some $u_g < \infty$. Furthermore, recall that $\kappa^N(\xi)(\cdot) = \xi(\cdot \times [0,N))/N$ is the "spatial projection operator", and define the transition function $\Gamma^N : \mathcal{M}_F(\mathbb{R}^d) \to \mathcal{M}(\mathbb{R}^d \times [0,N))$ so that for $\eta \in \mathcal{M}_F(\mathbb{R}^d)$, if $\widehat{g}_N(x) = \int_0^N g(x,u)du/N$, then

$$\begin{aligned} F_g^N(\eta) &:= \int f(\xi)\Gamma^N(\eta, d\xi) \\ &= \exp\left(N\left\langle \log\frac{1}{N}\int_0^N g(x,u)du, \eta(dx)\right\rangle\right) \\ &= \exp\left(N\left\langle \log \widehat{g}_N(x), \eta(dx)\right\rangle\right), \end{aligned}$$

i.e., Γ^N assigns independent labels on $[0,N]$ to each of the points in η. It follows from Lemma 3.5 that for test functions of this form the generator of η_t^N is

$$\begin{aligned} \mathcal{P}^N F_g^N(\eta) = F_g^N(\eta) N\theta \Big\langle \gamma(x,\eta)\int r(z,\eta)\left(\widehat{g}_N(z) - 1\right) q_\theta(x,dz) \\ + \mu_\theta(x,\eta)\left(\frac{1}{\widehat{g}_N(x)} - 1\right), \eta(dx)\Big\rangle. \end{aligned} \tag{1.3}$$

(Note that f here differs from the f used in Lemma 3.5 so as to agree with standard usage in the literature on lookdown processes.) The generator of ξ_t^N is A^N, defined in equation (5.10).

Lemma A.2 *For all finite counting measures η on $\mathbb{R}^d$, if f is of the form* (1.2)*, then*

$$\int A^N f(\xi)\Gamma^N(\eta, d\xi) = \mathcal{P}^N F_g^N(\eta). \tag{1.4}$$

For the limiting process, recall that $\kappa(\xi)(\cdot) = \lim_{u\to\infty} \xi(\cdot \times [0,u))/u$ is the "spatial projection operator", and define the probability kernel $\Gamma : \mathcal{M}_F(\mathbb{R}^d) \to \mathcal{M}(\mathbb{R}^d \times [0,\infty))$ so that for $\eta \in \mathcal{M}_F(\mathbb{R}^d)$, defining $\widetilde{g}(x) = \int_0^\infty (g(x,u)-1)du$,

$$\begin{aligned} F_g(\eta) &:= \int f(\xi)\Gamma(\eta, d\xi) \\ &= \exp\left(\left\langle \int_0^\infty (g(x,u)-1)du, \eta(dx)\right\rangle\right) \\ &= e^{\langle \widetilde{g}(x), \eta(dx)\rangle}. \end{aligned}$$

i.e., $\Gamma(\eta,\cdot)$ is the distribution of a conditionally Poisson process with intensity a product of η and Lebesgue measure. It again follows from Lemma 3.5 that for test functions of this form the generator of η_t is

$$\mathcal{P}^\infty F_g(\eta) = F_g(\eta)\left\langle \gamma(x,\eta)\mathcal{B}(\widetilde{g}(\cdot)r(\cdot))(x) + F(x,\eta)\widetilde{g}(x) + \alpha\gamma(x,\eta)r(x,\eta)\widetilde{g}^2(x), \eta(dx)\right\rangle. \tag{1.5}$$

The generator of ξ_t is A, defined in equation (5.11).

Lemma A.3 *For all $\eta \in \mathcal{M}_F(\mathbb{R}^d)$, if f is of the form* (1.2)*, then*

$$\int Af(\xi)\Gamma(\eta, d\xi) = \mathcal{P}^\infty F_g(\eta). \tag{1.6}$$

Proof [of Lemma A.2] First, break the generator A^N into three parts,

$$\begin{aligned} A_1^N f(\xi) &= f(\xi)\sum_{(x,u)\in\xi} 2c_\theta(x,\eta)\int_u^N \left(\frac{1}{2}\frac{g(x,v_1)}{g(x,u)}\int_{\mathbb{R}^d}(g(y,u)-g(x,u))q_\theta^m(x,dy,\eta)\right)dv_1, \\ A_2^N f(\xi) &= f(\xi)\sum_{(x,u)\in\xi} 2c_\theta(x,\eta)\int_u^N \left(\frac{1}{2}\int_{\mathbb{R}^d}\left(\frac{g(y,v_1)+g(x,v_1)}{2}-1\right)q_\theta^m(x,dy,\eta)\right)dv_1, \\ A_3^N f(\xi) &= f(\xi)\sum_{(x,u)\in\xi}\left(c_\theta(x,\eta)u^2 - b_\theta(x,\eta)u\right)\frac{\partial_u g(x,u)}{g(x,u)}, \end{aligned}$$

where q^m was defined in equation (5.24), so that

$$A^N f(\xi) = A_1^N f(\xi) + A_2^N f(\xi) + A_3^N f(\xi).$$

We now integrate each piece against Γ^N. First note that by the product form of f,

$$\int f(\xi)\sum_{(x,u)\in\xi}\frac{\ell(x,u)}{g(x,u)}\Gamma^N(\eta,d\xi) = F_g^N(\eta)\left\langle \frac{1}{N\widehat{g}_N(x)}\int_0^N \ell(x,u)du, N\eta(dx)\right\rangle.$$

Therefore,

$$\begin{aligned}
&\int A_1^N f(\xi)\Gamma^N(\eta, d\xi) \\
&\quad = F_g^N(\eta)\Big\langle \frac{c_\theta(x,\eta)}{\widehat{g}_N(x)} \int_0^N \left\{ \int_u^N \left(\frac{1}{2} g(x,v_1) \int_{\mathbb{R}^d} (g(y,u) - g(x,u)) q_\theta^m(x, dy, \eta) \right) dv_1 \right\} du, \eta(dx) \Big\rangle \\
&\quad = F_g^N(\eta)\Big\langle \frac{c_\theta(x,\eta)}{\widehat{g}_N(x)} \left\{ \int_{\mathbb{R}^d} \left\{ \int_0^N \int_u^N g(x,v_1)(g(y,u) - g(x,u)) dv_1 du \right\} q_\theta^m(x, dy, \eta), \eta(dx) \Big\rangle.
\end{aligned}$$

For the second generator, we have

$$\begin{aligned}
&\int A_2^N f(\xi)\Gamma^N(\eta, d\xi) \\
&\quad = F_g^N(\eta)\Big\langle \frac{c_\theta(x,\eta)}{\widehat{g}_N(x)} \int_0^N g(x,u) \left\{ \int_u^N \left(\int_{\mathbb{R}^d} \left(\frac{g(y,v_1) + g(x,v_1)}{2} - 1 \right) q_\theta^m(x, dy, \eta) \right) dv_1 \right\}, \eta(dx) \Big\rangle \\
&\quad = F_g^N(\eta)\Big\langle \frac{c_\theta(x,\eta)}{\widehat{g}_N(x)} \int_{\mathbb{R}^d} \left\{ \int_0^N \int_u^N g(x,u) \left(g(y,v_1) + g(x,v_1) - 2 \right) dv_1 du \right\} q_\theta^m(x, dy, \eta), \eta(dx) \Big\rangle.
\end{aligned}$$

For the third generator we have that

$$\begin{aligned}
&\int A_3^N f(\xi)\Gamma^N(\eta, d\xi) \\
&\quad = F_g^N(\eta)\Big\langle \frac{1}{\widehat{g}_N(x)} \int_0^N \left(c_\theta(x,\eta)u^2 - b_\theta(x,\eta)u \right) \partial_u g(x,u) du, \eta(dx) \Big\rangle \\
&\quad = F_g^N(\eta)\Big\langle \frac{1}{\widehat{g}_N(x)} \int_0^N \left(b_\theta(x,\eta) - 2c_\theta(x,\eta)u \right) (g(x,u) - 1) du, \eta(dx) \Big\rangle.
\end{aligned}$$

Note that $2\int_0^N \int_u^N g(x,v_1)g(y,u)dv_1du = N^2\widehat{g}_N(x)\widehat{g}_N(y)$, and so

$$\begin{aligned}
&\int_0^N \int_u^N g(x,v_1)(g(y,u) - g(x,u))dv_1du + \int_0^N \int_u^N g(x,u) \left(g(y,v_1) + g(x,v_1) - 2 \right) dv_1du \\
&\quad = N^2 \widehat{g_N}(x)(\widehat{g_N}(y) - 2) + 2\int_0^N ug(x,u)du.
\end{aligned}$$

Combining the last equations, and using the fact that $Nc_\theta(x,\eta) - b_\theta(x,\eta) = \theta\mu_\theta(x,\eta)$, we have

$$\begin{aligned}
&\int \left(A_1^N f(\xi) + A_2^N f(\xi) + A_3^N f(\xi) \right) \Gamma^N(\eta, d\xi) \\
&\quad = F_g^N(\eta)\Big\langle c_\theta(x,\eta)N^2 \int_{\mathbb{R}^d} (\widehat{g}_N(y) - 2) q^m(x, dy, \eta) + \frac{1}{\widehat{g}_N(x)} c_\theta(x,\eta) \int_0^N 2u du \\
&\qquad\quad + \frac{1}{\widehat{g}_N(x)} b_\theta(x,\eta) \int_0^N (g(x,u) - 1) du, \eta(dx) \Big\rangle \\
&\quad = F_g^N(\eta)\Big\langle c_\theta(x,\eta)N^2 \int_{\mathbb{R}^d} (\widehat{g}_N(y) - 1) q^m(x, dy, \eta) + N^2 c_\theta(x,\eta) \left(\frac{1}{\widehat{g}_N(x)} - 1 \right) \\
&\qquad\quad + Nb_\theta(x,\eta) \left(1 - \frac{1}{\widehat{g}_N(x)} \right), \eta(dx) \Big\rangle \\
&\quad = F_g^N(\eta) N \Big\langle N c_\theta(x,\eta) \int_{\mathbb{R}^d} (\widehat{g}_N(y) - 1) q^m(x, dy, \eta) + \theta\mu_\theta(x,\eta) \left(\frac{1}{\widehat{g}_N(x)} - 1 \right), \eta(dx) \Big\rangle.
\end{aligned}$$

This matches equation (1.3), as desired, because $Nc_\theta(x,\eta)q_m(x,dy,\eta) = \theta\gamma(x,\eta)q_\theta(x,dy)$. □

Before proving Lemma A.3, we recall an important equality for conditionally Poisson point processes ([KR11] Lemma A.3).

Lemma A.4 *If $\xi = \sum_i \delta_{Z_i}$ is a Poisson random measure with mean measure ν, then for $\ell \in L^1(\nu)$ and $g \geq 0$ with $\log g \in L^1(\nu)$,*

$$\mathbb{E}\left[\sum_j \ell(Z_j)\prod_i g(Z_i)\right] = \int \ell g d\nu e^{\int (g-1)d\nu}. \tag{1.7}$$

Proof [of Lemma A.3] By Lemma A.4,

$$\int f(\xi)\sum_{(x,u)\in\xi}\frac{\ell(x,u)}{g(x,u)}\Gamma(\eta,\xi) = F_g(\eta)\left\langle \int_0^\infty \ell(x,u)du, \eta(dx)\right\rangle.$$

Comparing this to the definition of A (equation (5.11)), we see that

$$\int Af(\xi)\Gamma(\eta,d\xi) = F_g(\eta)\left\langle \int_0^\infty (\ell_1(x,u)+\ell_2(x,u)+\ell_3(x,u))du, \eta(dx)\right\rangle,$$

where

$$\begin{aligned}\ell_1(x,u) &= \gamma(x,\eta)\left(\mathcal{B}(g(\cdot,u)r(\cdot,\eta))(x) - g(x,u)\mathcal{B}r(x,\eta)\right)\\ &= \gamma(x,\eta)\left(\mathcal{B}((g(\cdot,u)-1)r(\cdot,\eta))(x) - (g(x,u)-1)\mathcal{B}r(x,\eta)\right)\end{aligned}$$

and

$$\ell_2(x,u) = 2g(x,u)\alpha\gamma(x,\eta)r(x,\eta)\int_u^\infty (g(x,v)-1)dv$$

and

$$\ell_3(x,u) = \left(\alpha\gamma(x,\eta)r(x,\eta)u^2 - \{\gamma(x,\eta)\mathcal{B}r(x,\eta)+F(x,\eta)\}u\right)\partial_u g(x,u).$$

First note that since $\mathcal{B}$ acts on space, it commutes with the integral over levels, and so

$$\int_0^\infty \ell_1(x,u)du = \gamma(x,\eta)\left(\mathcal{B}(\widetilde{g}(\cdot)r(\cdot,\eta))(x) - \widetilde{g}(x)\mathcal{B}r(x,\eta)\right),$$

since $\widetilde{g}(x) = \int_0^\infty (g(x,u)-1)du$. Next,

$$\int_0^\infty \ell_2(x,u)du = \alpha\gamma(x,\eta)r(x,\eta)2\int_0^\infty g(x,u)\int_u^\infty (g(x,v)-1)dvdu.$$

Finally, integrating by parts,

$$\int_0^\infty \ell_3(x,u)du = -\alpha\gamma(x,\eta)r(x,\eta)\int_0^\infty 2u(g(x,u)-1)du + \{\gamma(x,\eta)\mathcal{B}r(x,\eta) + F(x,\eta)\}\widetilde{g}(x)$$

Now, note that

$$\begin{aligned}
&\int_0^\infty g(x,u)\int_u^\infty (g(x,v)-1)dvdu - \int_0^\infty u(g(x,u)-1)du \\
&\quad = \int_0^\infty g(x,u)\int_u^\infty (g(x,v)-1)dvdu - \int_0^\infty\int_v^\infty (g(x,u)-1)dudv \\
&\quad = \int_0^\infty (g(x,u)-1)\int_u^\infty (g(x,v)-1)dvdu \\
&\quad = \widetilde{g}(x)^2/2.
\end{aligned}$$

Adding these together, we get that

$$\begin{aligned}
&\int_0^\infty (\ell_1(x,u)+\ell_2(x,u)+\ell_3(x,u))du \\
&\quad = \gamma(x,\eta)\mathcal{B}(\widetilde{g}(\cdot)r(\cdot,\eta))(x) + F(x,\eta)\widetilde{g}(x) + \alpha\gamma(x,\eta)r(x,\eta)\widetilde{g}(x)^2,
\end{aligned}$$

which agrees with (1.5), as desired. □

A.2 Technical Lemmas on Probability Kernels

A.2.1 Constraints on kernel widths

Lemma A.5 *Suppose the first three conditions of Assumptions 2.8 hold, and furthermore the kernels $\rho_r = p_{\epsilon_r^2}$ and $\rho_\gamma = p_{\epsilon_\gamma^2}$ are each Gaussian with standard deviations ϵ_r and ϵ_γ respectively. Let $\lambda = \sup_x \sup_{y:\|y\|=1} y^T\mathbf{C}(x)y$ be the largest eigenvalue of $\mathbf{C}(x)$ across all x. If $\epsilon_r^2 + \frac{2\lambda}{\theta} < \epsilon_\gamma^2$, then there is a $C < \infty$ such that for all $x \in \mathbb{R}^d$, $\eta \in \mathcal{M}_F(\mathbb{R}^d)$,*

$$\left|\theta\int_{\mathbb{R}^d}(\rho_r * \eta(y) - \rho_r * \eta(x))q_\theta(x,dy)\right| \le C\rho_\gamma * \eta(x) \tag{1.8}$$

and

$$\theta\int_{\mathbb{R}^d}(\rho_r * \eta(y) - \rho_r * \eta(x))^2 q_\theta(x,dy) \le C(\rho_\gamma * \eta(x))^2. \tag{1.9}$$

Note that the right hand side of each is the average density over a wider region (since $\epsilon_\gamma > \epsilon_r$). The key assumption here is that the spatial scale over which local density affects birth rate is larger than the scale over which it affects establishment. In the simple case of $\vec{b} = 0$ and $\mathbf{C} = \sigma^2 I$, the condition is simply that $\epsilon_r^2 + 2\sigma^2/\theta < \epsilon_\gamma^2$. This gives a yet more concrete situation in which Condition 2 of Lemma 2.9 holds.

Proof [of Lemma A.5]

First we prove (1.8). Recall that $\rho_r * \eta(x) = \int p_{\epsilon_r^2}(x-w)\eta(dw)$, where p_t is the density of a Gaussian with mean 0 and variance t, so that applying Fubini, (1.8) is

$$\left| \int_{\mathbb{R}^d} \theta \int_{\mathbb{R}^d} (p_{\epsilon_r^2}(y-w) - p_{\epsilon_r^2}(x-w)) q_\theta(x, dy)\eta(dw) \right|.$$

Write $p_{s,x}(\cdot)$ for the density of a Gaussian with mean $s\vec{b}(x)$ and covariance $\epsilon_r^2 I + s\mathbf{C}(x)$, so that $\int p_{\epsilon_r^2}(y-w)q_\theta(x,dy) = p_{1/\theta,x}(w-x)$. It therefore suffices to show that for all x and $w \in \mathbb{R}^d$, there exists K such that

$$\left| \theta \int_{\mathbb{R}^d} (p_{\epsilon_r^2}(y-w) - p_{\epsilon_r^2}(x-w)) q_\theta(x, dy) \right| = \theta \left| p_{1/\theta,x}(w-x) - p_{0,x}(w-x) \right| \le K p_{\epsilon_\gamma^2}(w-x).$$

However, $\theta(p_{1/\theta,x}(z) - p_{0,x}(z)) = \partial_s p_{s,x}(z)$ for some $0 \le s \le 1/\theta$. Write $\Gamma(s,x) = \epsilon_r^2 I + s\mathbf{C}(x)$, so that

$$p_{s,x}(z) = \frac{1}{(2\pi|\Gamma(s,x)|)^{d/2}} \exp\left(-\frac{1}{2}(z - s\vec{b}(x))^T \Gamma(s,x)^{-1} (z - s\vec{b}(x)) \right),$$

and note that if λ_i are the eigenvalues of $\mathbf{C}(x)$ then $|\Gamma(s,x)| = \prod_i (\epsilon_r^2 + s\lambda_i)$, and $\partial_s |\Gamma(s,x)| = \sum_i \lambda_i |\Gamma(s,x)|/(\epsilon_r^2 + s\lambda_i)$. Therefore,

$$\partial_s p_{s,x}(z) = \Bigg(\vec{b}(x)^T \Gamma(s,x)^{-1}(z - s\vec{b}(x)) + (z - s\vec{b}(x))^T \Gamma(s,x)^{-1} \mathbf{C}(x) \Gamma(s,x)^{-1} (z - s\vec{b}(x)) - \sum_i \frac{\lambda_i}{\epsilon_r^2 + s\lambda_i} \Bigg) p_{s,x}(z),$$

where z^T is the transpose of z. This implies that

$$\frac{\theta \int_{\mathbb{R}^d} (p_{\epsilon_r^2}(y-w) - p_{\epsilon_r^2}(x-w)) q_\theta(x, dy)}{p_{\epsilon_\gamma^2}(x-w).} = h(x-w) e^{k(x-w)},$$

where $h(z)$ and $k(z)$ are quadratic polynomials in z whose coefficients depend on s and x but are uniformly bounded, and

$$k(z) = \frac{1}{2\epsilon_\gamma^2}\|z\|^2 - \frac{1}{2}(z - s\vec{b}(x))^T \Gamma(s,x)^{-1} (z - s\vec{b}(x)),$$

Since $\inf_z z^T\Gamma(s,x)^{-1}z/\|z\|^2 = 1/(s\lambda(x)+\epsilon_r^2)$, where $\lambda(x) = \sup z^T\mathbf{C}(x)z/\|z\|^2$ is the largest eigenvalue of $\mathbf{C}(x)$, this is negative for all z outside a bounded region, and so equation (1.8) follows from the assumption that $\epsilon_r^2 + 2\sup_x \lambda(x)/\theta < \epsilon_\gamma^2$. (Note that we do not yet need the factor of 2.)

Next we prove equation (1.9), in a similar way. Again applying Fubini,

$$\begin{aligned}
&\theta\int_{\mathbb{R}^d} \left(\rho_r * \eta(y) - \rho_r * \eta(x)\right)^2 q_\theta(x,dy) \\
&\quad = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \theta \int_{\mathbb{R}^d} (p_{\epsilon_r^2}(y-w) - p_{\epsilon_r^2}(x-w))(p_{\epsilon_r^2}(y-v) - p_{\epsilon_r^2}(x-v)) q_\theta(x,dy)\eta(dv)\eta(dw),
\end{aligned}$$

and so as before, equation (1.9) will follow if the integrand is bounded by $Kp_\gamma(x-w)p_\gamma(x-v)$. Now, let Y_1, Y_2, and Z be independent d-dimensional Gaussians with mean zero, where Y_1 and Y_2 have covariance $\epsilon_r^2 I$, and Z has covariance $\mathbf{C}(x)$. Write $p_{s,t,x}(\cdot,\cdot)$ for the joint density of $Y_1 + \sqrt{s}Z + s\vec{b}(x)$ and $Y_2 + \sqrt{t}Z + t\vec{b}(x)$. Then, observe that

$$\begin{aligned}
&\theta\int_{\mathbb{R}^d} (p_{\epsilon_r^2}(y-w) - p_{\epsilon_r^2}(x-w))(p_{\epsilon_r^2}(y-v) - p_{\epsilon_r^2}(x-v)) q_\theta(x,dy) \\
&\quad = \theta\Big(p_{1/\theta,1/\theta,x}(x-w,x-v) - p_{0,1/\theta,x}(x-w,x-v) \\
&\qquad\quad - p_{1/\theta,0,x}(x-w,x-v) + p_{0,0,x}(x-w,x-v)\Big) \\
&\quad = \partial_s p_{s,1/\theta,x}(x-w,x-v) - \partial_t p_{0,t,x}(x-w,x-v),
\end{aligned}$$

for some $0 \le s,t \le 1/\theta$. As before,

$$\frac{\theta\int_{\mathbb{R}^d}(p_{\epsilon_r^2}(y-w) - p_{\epsilon_r^2}(x-w))(p_{\epsilon_r^2}(y-v) - p_{\epsilon_r^2}(x-v))q_\theta(x,dy)}{p_{\epsilon_\gamma^2}(x-w)p_{\epsilon_\gamma^2}(x-v)} = h(x-w,x-v)e^{k(x-w,x-v)},$$

where $h(z_1,z_2)$ is a polynomial with uniformly bounded coefficients and

$$k(z_1,z_2) = (\|z_1\|^2 + \|z_2\|^2)/(2\epsilon_\gamma^2) - \frac{1}{2}[z_1,z_2]^T\Gamma(s,t,x)^{-1}[z_1,z_2],$$

where $[z_1,z_2]$ is the $\mathbb{R}^{2d}$ vector formed by concatenating z_1 and z_2, and $\Gamma(s,t,x)$ is the block matrix

$$\Gamma(s,t,x) = \begin{bmatrix} \epsilon_r^2 I + s\mathbf{C}(x) & \sqrt{st}\mathbf{C}(x) \\ \sqrt{st}\mathbf{C}(x) & \epsilon_r^2 I + t\mathbf{C}(x) \end{bmatrix}.$$

If $\mathbf{C}(x)u = au$ for some $a \in \mathbb{R}$, then $[u\sqrt{s}, u\sqrt{t}]$ is an eigenvector of $\Gamma(s,t,x)$ with eigenvalue $\epsilon_r^2 + (s+t)a$, and $[u\sqrt{t}, -u\sqrt{s}]$ is an eigenvector of $\Gamma(s,t,x)$ with eigenvalue

0. This implies the largest eigenvalue of $\Gamma(s,t,x)$ is equal to $\epsilon_r^2 + (s+t)\lambda(x)$, where $\lambda(x)$ is again the largest eigenvalue of $\mathbf{C}(x)$. Therefore, if $s + t \leq 2/\theta$,

$$\begin{aligned}(\|z_1\|^2 + \|z_2\|^2)/\epsilon_\gamma^2 - [z_1, z_2]^T \Gamma(s,t,x)^{-1}[z_1, z_2] \\ \leq (\|z_1\|^2 + \|z_2\|^2)\left(\frac{1}{\epsilon_\gamma^2} - \frac{1}{\epsilon_r^2 + 2\lambda(x)/\theta}\right),\end{aligned}$$

which is negative by assumption. Therefore, there is a K such that

$$\frac{\left|\partial_s p_{s,1/\theta,x}(x-w, x-v) - \partial_t p_{0,t,x}(x-w, x-v)\right|}{p_{\epsilon_\gamma^2}(x-w)p_{\epsilon_\gamma^2}(x-v)} \leq K$$

for all $\theta > 1$ and all x, v, and $w \in \mathbb{R}^d$, proving equation (1.9) and hence the lemma. □

A.2.2 Tightness of processes

Here we record, for completeness, the fact used above that tightness for a family of processes, if determined by the Aldous-Rebolledo criterion, extends to sums and products of those processes. We first record for reference one version of the Aldous-Rebolledo criteria for tightness of a sequence real-valued processes (as it appears in Theorem 1.17 of [Eth00]; see also Theorem 4.13 of [JS13]):

Theorem A.6 ([Reb80]) *Let $\{Y^{(n)}\}_{n\geq 1}$ be a sequence of real-valued processes with càdlàg paths. Suppose that the following conditions are satisfied.*

1. *For each fixed $t \in [0,T]$, $\{Y_t^{(n)}\}_{n\geq 1}$ is tight.*

2. *Given a sequence of stopping times τ_n, bounded by T, for each $\epsilon > 0$ there exists $\delta > 0$ and n_0 such that*

$$\sup_{n\geq n_0} \sup_{\theta\in[0,\min(\delta, T-\tau_n)]} \mathbb{P}\left\{\left|Y_{\tau_n+\theta}^{(n)} - Y_{\tau_n}^{(n)}\right| > \epsilon\right\} \leq \epsilon.$$

Then the sequence $\{(Y_t^{(n)})_{t=0}^T\}_{n\geq 1}$ is tight.

Lemma A.7 *Let $\{X^{(n)}\}_{n\geq 1}$ and $\{Y^{(n)}\}_{n\geq 1}$ be sequences of jointly defined real-valued processes with càdlàg paths satisfying the conditions of Theorem A.6. Then $\{X^{(n)}Y^{(n)}\}_{n\geq 1}$ and $\{X^{(n)} + Y^{(n)}\}_{n\geq 1}$ also satisfy the conditions of Theorem A.6.*

By "jointly defined" we mean that $X^{(n)}$ and $Y^{(n)}$ are defined on the same probability space, so that the products and sums make sense.

Proof [of Lemma A.7] The proof for $X^{(n)} + Y^{(n)}$ is similar to but more straightforward than for $X^{(n)}Y^{(n)}$, so on only prove the Lemma for the latter. First, note that for any $\epsilon > 0$, by tightness of $X_t^{(n)}$ and $Y_t^{(n)}$ there is a K such that $\mathbb{P}\{X_t^{(n)} > \sqrt{K}\}$ and $\mathbb{P}\{Y_t^{(n)} > \sqrt{K}\}$ are both less than $\epsilon/2$, and hence

$$\mathbb{P}\{X_t^{(n)}Y_t^{(n)} > K\} \leq \mathbb{P}\{X_t^{(n)} > \sqrt{K}\} + \mathbb{P}\{Y_t^{(n)} > \sqrt{K}\} \leq \epsilon.$$

Therefore, $X_t^{(n)}Y_t^{(n)}$ is tight.

Next, note that for $0 \leq \tau_n \leq T$,

$$\begin{aligned}
&\sup_{0\leq\theta\leq\min(\delta,T-\tau_n)} \left|X^{(n)}_{\tau_n+\theta}Y^{(n)}_{\tau_n+\theta} - X^{(n)}_{\tau_n}Y^{(n)}_{\tau_n}\right| \\
&\quad\leq \sup_{0\leq\theta\leq\min(\delta,T-\tau_n)} \left|X^{(n)}_{\tau_n+\theta}\right| \left|Y^{(n)}_{\tau_n+\theta} - Y^{(n)}_{\tau_n}\right| + \left|X^{(n)}_{\tau_n+\theta} - X^{(n)}_{\tau_n}\right| \left|Y^{(n)}_{\tau_n}\right| \\
&\quad\leq \sup_{0\leq t\leq T} \left|X_t^{(n)}\right| \sup_{0\leq\theta\leq\min(\delta,T-\tau_n)} \left|Y^{(n)}_{\tau_n+\theta} - Y^{(n)}_{\tau_n}\right| + \sup_{0\leq\theta\leq\min(\delta,T-\tau_n)} \left|X^{(n)}_{\tau_n+\theta} - X^{(n)}_{\tau_n}\right| \sup_{0\leq t\leq T} \left|Y_t^{(n)}\right|,
\end{aligned}$$

so that for any C,

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{0\leq\theta\leq\min(\delta,T-\tau_n)} \left|X^{(n)}_{\tau_n+\theta}Y^{(n)}_{\tau_n+\theta} - X^{(n)}_{\tau_n}Y^{(n)}_{\tau_n}\right| > \epsilon\right\} \\
&\qquad\leq \mathbb{P}\left\{\sup_{0\leq t\leq T} \left|X_t^{(n)}\right| > C\right\} + \mathbb{P}\left\{\sup_{0\leq\theta\leq\min(\delta,T-\tau_n)} \left|Y^{(n)}_{\tau_n+\theta} - Y^{(n)}_{\tau_n}\right| > \epsilon/C\right\} \qquad (1.10) \\
&\qquad + \mathbb{P}\left\{\sup_{0\leq\theta\leq\min(\delta,T-\tau_n)} \left|X^{(n)}_{\tau_n+\theta} - X^{(n)}_{\tau_n}\right| > \epsilon/C\right\} + \mathbb{P}\left\{\sup_{0\leq t\leq T} \left|Y_t^{(n)}\right| > C\right\}
\end{aligned}$$

Now, since $\max_{0\leq t\leq T} X_t^{(n)}$ is tight (and likewise for Y) (see, e.g., Remark 3.7.3 in [EK86]), we may choose a $C \geq 4$ for which

$$\mathbb{P}\left\{\sup_{0\leq t\leq T} \left|X_t^{(n)}\right| > C\right\} \leq \frac{\epsilon}{4}.$$

Similarly, by assumption we can choose a δ for which

$$\mathbb{P}\left\{\sup_{0\leq\theta\leq\min(\delta,T-\tau_n)} \left|X^{(n)}_{\tau_n+\theta} - X^{(n)}_{\tau_n}\right| > \epsilon/C\right\} \leq \frac{\epsilon}{C}.$$

If we choose C and δ that do this for both $X^{(n)}$ and $Y^{(n)}$, then each of the terms in equation (1.10) are bounded by $\epsilon/4$, and condition (2) is satisfied for the product process. □

B
FRIME - Appendix

B.1 Data pre-processing techniques

Interpolation method to account for high 0 counts To fit the PFB best-fitting curve to both mitochondrial and genomic fragment profile, we fitted log fragment count against log fragment length according to a linear model. However, long fragments are sparse and rare in our dataset, so the count of fragments with exactly x base pairs can be 0 if x is large. As a result, under the log scale, the log count becomes negative infinity. To adjust for this sparsity, we plotted the cumulative fragment count against fragment length, then interpolated the count at intervals of 10 bp to give the average fragment count per unit length. We then apply the linear model on the adjusted log fragment counts. Details of the method can be found in the github repository.

Evolution to stationarity To evaluate how well the stationary distribution of FRIME processes fit clinical data, we stop a simulated FRIME process after every 10,000 events. We then conduct a Kolomogorov-Smirnov two-sample test on the simulated fragment profile and the clinical data for mtDNA fragment profile. For gDNA fragment profile, we conduct a χ^2-two-sample test on the total fragment counts in the bins (0, 250], (250, 420], (420, 620], (620, 820]. We observe that the p-values are low initially when the FRIME process has not reached stationarity, and the p-value is consistently above 0.05 after many events. See Figure B.1.

B.2 Supplementary Figures

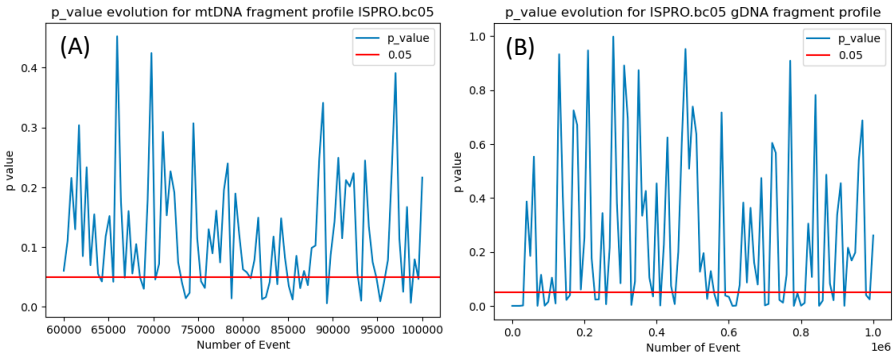


Figure B.1: Evolution of p-value under KS test on FRIME simulations We ran two FRIME simulations and compared them against the mtDNA fragment profile and gDNA fragment profile for sample ISPRO.bc05 under the two-sample Kolmogorov-Smirnov test. **(A)** A FRIME process with parameters $C_F = 7.5 * 10^5, \alpha_F = -0.78, L = 10^4, C_I = 8, \mathcal{E}(x) = L^{2.10}x^{-2.10}, r \sim Beta(1,1)$ is run for 10^5 events. After every 10^4 events, the fragment profile of the process is compared against the mtDNA fragment profile under the Kolmogorov-Smirnov test. The p-value of the result is plotted in the graph. **(B)** A FRIME process with parameters $C_F = 2.5, \alpha_F = 1.2, L = 10^4, C_I = 10, \mathcal{E}(x) = L^2x^{-2}, r \sim Beta(1,1)$ is run for 10^5 events. After every 10^4 events, the fragment profile of the process is compared against the gDNA fragment profile under the χ^2 two-sample test. The p-value of the result is plotted in the graph.

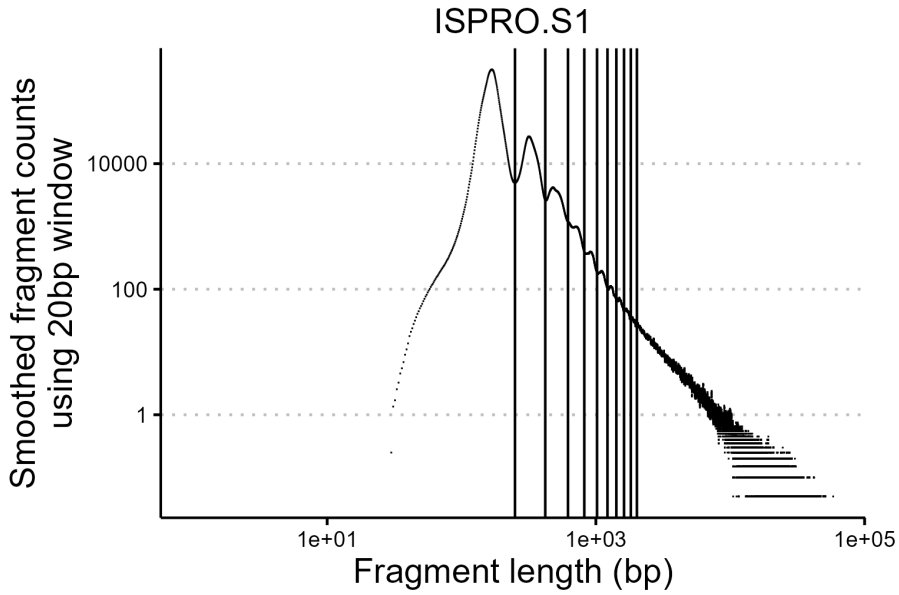


Figure B.2: Genomic DNA data on log-log scale, showing the first 9 partitions To study gDNA data, we binned the fragment counts according to specified nucleosomal bins. This binning is used to account for the cyclical nature of genomic cfDNA fragment profiles. The nucleosomal bins are defined manually as (0, 250], (250, 420], (420, 620], (620, 820], and every 200bp onward. The binning strategy has been visually confirmed to be reasonable.

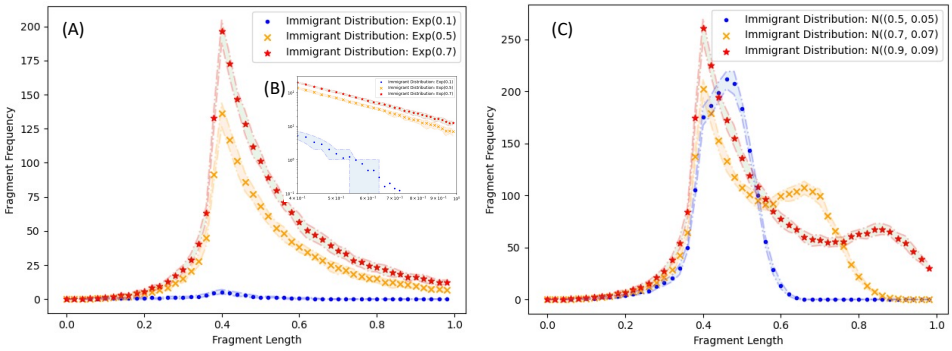


Figure B.3: FRIME profile with exponential and normal distributed immigration mechanism Stationary distributions of fragment sizes obtained from simulations of the FRIME process with immigration and fragmentation parameter values $C_I = 500, C_F = 1, \alpha_F = 1, a = 1$, PFB exit mechanism with parameters $\alpha_E = -2, B_E = 0.4$, and simulation threshold $\kappa = 0.05$. 50 points were plotted for each configuration. Each point corresponds to the fragment count per unit length over an interval of length 0.02. Different choices of the immigration function were taken and 50 simulations run for each choice of immigration function. Shaded regions were plotted using upper and lower quartile fragment count across all simulations. **(A)** EXP model with immigrating length $z \sim Exp(\lambda)$ where $\lambda = 0.1$ (dots), 0.5 ($\times$), 0.7 (stars). **(B)** The simulated fragment profiles look sufficiently linear under the log-log scale. **(C)** NORM model with immigrating length $z \sim N(\mu, 0.1\mu)$ with $\mu = 0.5$ (dots), 0.7 ($\times$) ,0.9 (stars). The simulated fragment profiles clearly exhibit a peak at the mean fragment length size.

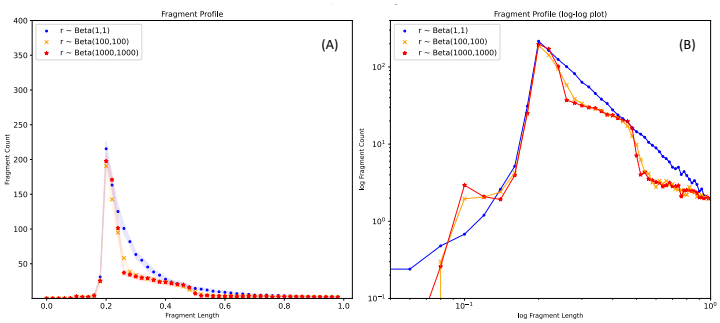


Figure B.4: FRIME profile with strong fragmentation around mid-point Stationary distributions of fragment sizes obtained from simulations of the FRIME process with uniform immigration and fragmentation parameter values $C_I = 100, C_F = 1, \alpha_F = 1$, PFB exit mechanism with parameters $\alpha_E = -2, B_E = 0.2$, and simulation threshold $\kappa = 0.05$. 50 points were plotted for each configuration. Each point corresponds to the fragment count per unit length over an interval of length 0.02. Three different choices of the fragmentation ratios were taken and 50 simulations run for each choice of fragmentation ratio. Shaded regions were plotted using upper and lower quartile fragment count across all simulations. Right panel is normal scale, left panel is log-log scale. **(A)** Fragmentation ratio $r \sim Beta(a, a)$ with $a = 1$ (dots), 100 ($\times$), 1000 (stars). For large a, fragments tend to be split into two roughly equal proportions. The simulated fragment profiles exhibit big drops and flat decays. **(B)** In log-log scale, the fragment profiles for large values of a look like steps.

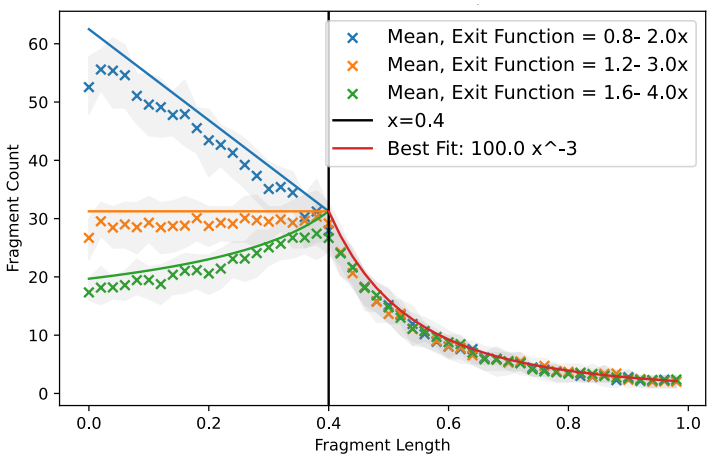


Figure B.5: FRIME profile with different linear exit mechanism Stationary distributions of fragment sizes obtained from simulations of the FRIME process with uniform immigration and fragmentation parameter values $C_I = 500, C_F = 1, \alpha_F = 1, a = 1, b = 1$ and simulation threshold $\kappa = 0.05$. 50 points were plotted for each configuration. Each point corresponds to the fragment count per unit length over an interval of length 0.02. Different choices of the immigration function were taken and 50 simulations were run for each choice of immigration function. Shaded regions were plotted using upper and lower quartile fragment count across all simulations. The simulations were run with linear exit mechanism $\mathcal{E}(x) = B_E m_E - m_E x$, where $B_E = 0.4$ and $m_E = 2$ (blue), 3 (orange), 4 (green). A vertical line at $B_E = 0.4$ was plotted in black. A best-fit line of $100x^3$ was plotted in red in the region $x > 0.4$. Note that when $m_E = 4 > 3C_F$, the peak coincides with 0.4 and when $m_E = 2 < 3C_F$, the peak is to the left of 0.4.

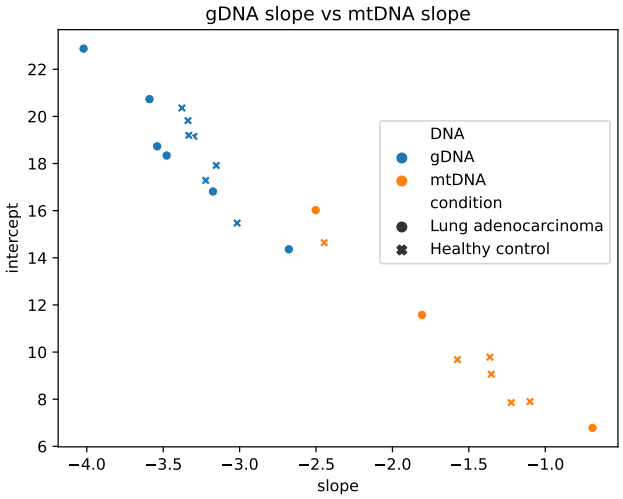


Figure B.6: gDNA best fit coefficients distribution plot. The best-fit coefficients for linearity in log-log scale were plotted as a scatter plot. The x-axis is the gDNA slope, and the y-axis is the gDNA intercept. Healthy patients are in blue and patients with lung adenocarcinoma is in pink. The scatter plot suggests that the data for patients with lung adenocarcinoma has higher variability.

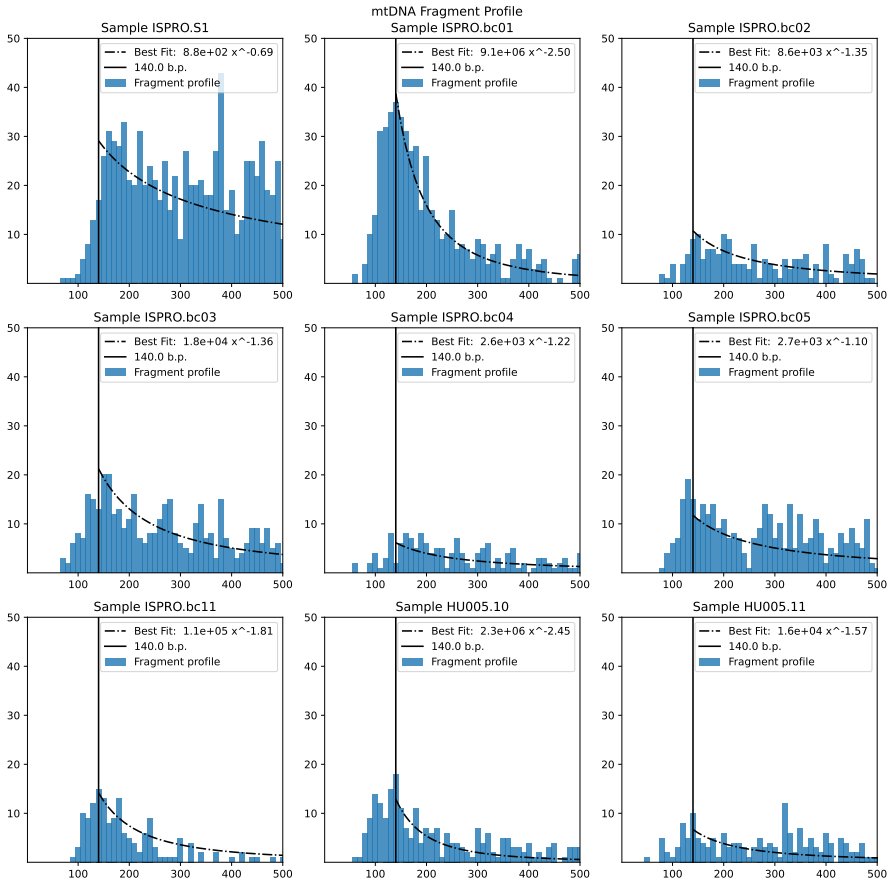


Figure B.7: Mitochondrial cfDNA fragment profiles with best fit curve. Mitochondrial DNA fragment profiles for all samples from nanopore studies are plotted out as bar charts in blue on the top panels. Best fit curves of the expression Cx^{β} are plotted as black dot-dashed lines (-.). The coefficients of the best fit curve are given by a linear fit under the log-log scale on the counts of fragments longer than 140bp. A black vertical line at 140bp is also plotted on each profile. The cumulative fragment count for all fragment profiles were plotted out as bar charts in blue. Black lines plotted on top of the bars were linear interpolation of the cumulative fragment count at intervals of 10bp.

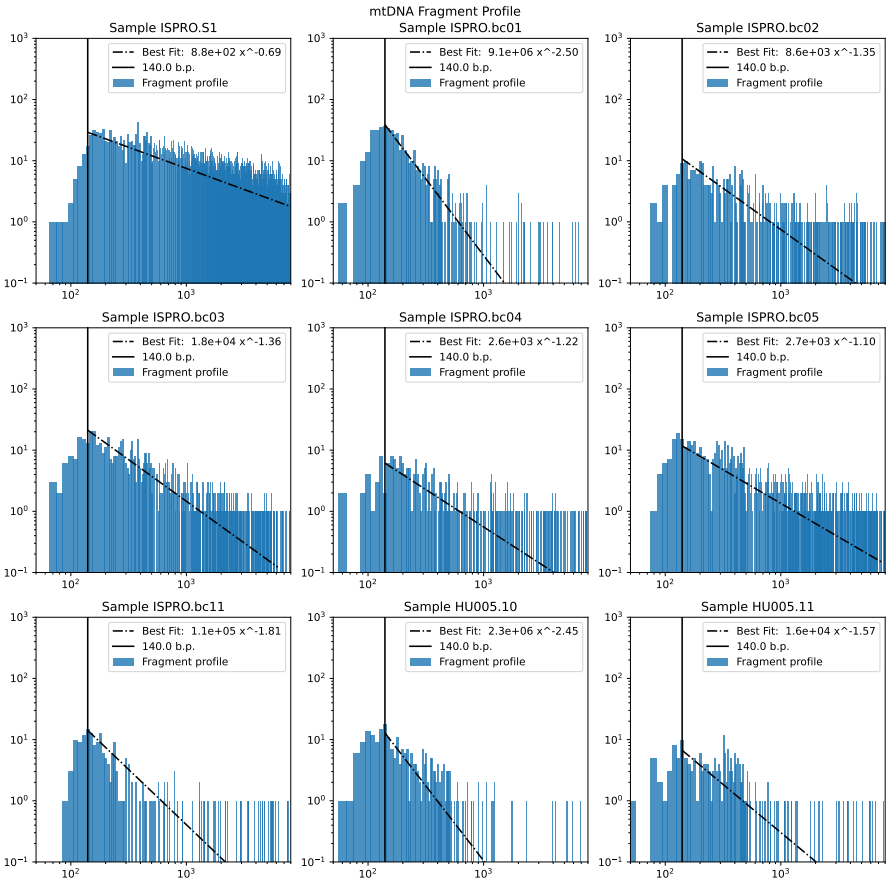


Figure B.8: Mitochondrial cfDNA fragment profiles with best fit curve, log-log scale. Mitochondrial DNA fragment profiles for all samples from nanopore studies are plotted out. Best fit curve of the expression Cx^{β} is also plotted out. The coefficients of the best fit curve are given by a linear fit under the log-log scale on the counts of fragments longer than 140bp. A black vertical line at 140bp is also plotted on each profile. Both axes are in log scale.

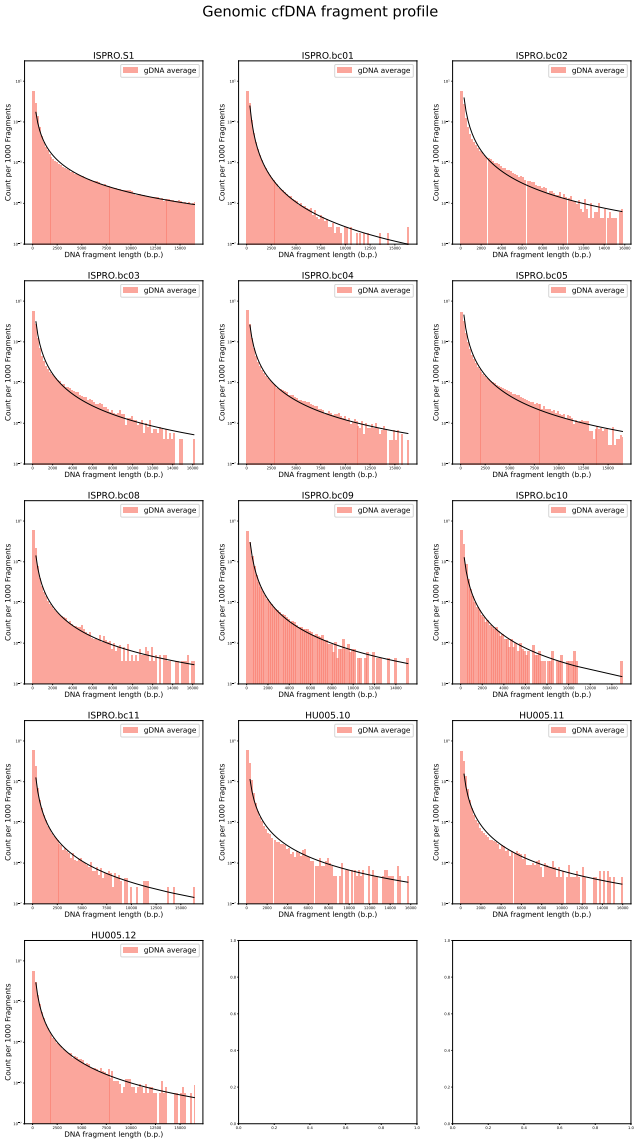


Figure B.9: Genomic cfDNA fragment profiles with best fit curve. Genomic DNA fragment profiles for all samples from nanopore studies are plotted out as bar charts in pink. Best fit curves of the expression Cx^{β} are also plotted out. The coefficients of the best fit curve are given by a linear fit under the log-log scale after binning the data. The x-axis is in log-scale.

Bibliography

[Ada+17] Viktor A. Adalsteinsson et al. "Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors". en. In: *Nature Communications* 8.1 (Nov. 2017). Number: 1 Publisher: Nature Publishing Group, p. 1324. URL: https://www.nature.com/articles/s41467-017-00965-y (visited on 12/22/2022).

[Ava+20] Stefano Avanzini et al. "A mathematical model of ctDNA shedding predicts tumor detection size". In: *Science advances* 6.50 (2020), eabc4308.

[BJY86] M. T. Barlow, S. D. Jacka, and M. Yor. "Inequalities for a Pair of Processes Stopped at a Random Time". In: *Proceedings of the London Mathematical Society* s3-52.1 (Jan. 1986), pp. 142–172. eprint: https://academic.oup.com/plms/article-pdf/s3-52/1/142/4387013/s3-52-1-142.pdf. URL: https://doi.org/10.1112/plms/s3-52.1.142.

[Bar79] N H Barton. "The dynamics of hybrid zones". In: *Heredity* 43.3 (1979), pp. 341–359.

[BEV10] N H Barton, A M Etheridge, and A Véber. "A new model for evolution in a spatial continuum". In: *Electron. J. Probab.* 15 (2010), pp. 162–216.

[BEE78] Arno Max Basedow, Klaus Heinrich Ebert, and Hanns Josef Ederer. "Kinetic Studies on the Acid Hydrolysis of Dextran". In: *Macromolecules* 11.4 (1978). _eprint: https://doi.org/10.1021/ma60064a031, pp. 774–781. URL: https://doi.org/10.1021/ma60064a031.

[BK00] E. Ben-Naim and P. L. Krapivsky. "Fragmentation with a steady source". en. In: *Physics Letters A* 275.1 (Oct. 2000), pp. 48–53. URL: https://www.sciencedirect.com/science/article/pii/S0375960100005703 (visited on 04/25/2023).

[Ber+09] Henri Berestycki et al. "The non-local Fisher-KPP equation: travelling waves and steady states". In: *Nonlinearity* 22.12 (2009), pp. 2813–2844. URL: http://stacks.iop.org/0951-7715/22/2813.

[Ber02] Jean Bertoin. "Self-similar fragmentations". en. In: *Annales de l'Institut Henri Poincare (B) Probability and Statistics* 38.3 (Jan. 2002), pp. 319–340. URL: https://www.sciencedirect.com/science/article/pii/S0246020300010736 (visited on 04/24/2023).

[Bet+14] Chetan Bettegowda et al. "Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies". In: *Science Translational Medicine* 6.224 (Feb. 2014). Publisher: American Association for the Advancement of Science, 224ra24–224ra24. URL: https://www.science.org/doi/10.1126/scitranslmed.3007094 (visited on 04/25/2023).

[BHK21] Gabriel Birzu, Oskar Hallatschek, and Kirill S Korolev. “Genealogical structure changes as range expansions transition from pushed to pulled”. In: *Proc. Natl. Acad. Sci. (USA)* 118.34 (2021), e2026746118.

[BHK18] Gabriel Birzu, Oskar Hallatschek, and Kirill S. Korolev. “Fluctuations uncover a distinct class of traveling waves”. In: *Proc. Natl. Acad. Sci. (USA)* 115.16 (2018), pp. 3645–3654.

[BEK21] Niloy Biswas, Alison Etheridge, and Aleksander Klimek. “The spatial Lambda-Fleming-Viot process with fluctuating selection”. In: *Electron. J. Probab.* 26 (2021), pp. 1–51. URL: https://doi.org/10.1214/21-EJP593.

[BP97] B M Bolker and S W Pacala. “Using moment equations to understand stochastically driven spatial pattern formation in ecological systems”. In: *Theor. Pop. Biol.* 52.3 (1997), pp. 179–197.

[BP99] B M Bolker and S W Pacala. “Spatial moment equations for plant competition: Understanding spatial strategies and the advantages of short dispersal”. In: *American Naturalist* 153.6 (1999), pp. 575–602.

[Bra83] M Bramson. “Convergence of solutions of the Kolmogorov equation to travelling waves”. In: *Mem. Amer. Math. Soc.* 44.285 (1983).

[Bri90] N. F. Britton. “Spatial Structures and Periodic Travelling Waves in an Integro-Differential Reaction-Diffusion Population Model”. In: *SIAM Journal on Applied Mathematics* 50.6 (1990), pp. 1663–1688.

[Bru+06] E Brunet et al. “Noisy travelling waves: effect of selection on genealogies”. In: *Europhys. Lett.* 76 (2006), pp. 1–7.

[Bru+07] E Brunet et al. “Effect of selection on ancestry: An exactly soluble case and its phenomenological generalization”. In: *Phys. Rev. E.* 76 (2007), p. 041104.

[CR21] Alice Callegaro and Matthew I. Roberts. *A spatially-dependent fragmentation process*. arXiv:2103.09761 [math]. June 2021. URL: http://arxiv.org/abs/2103.09761 (visited on 04/25/2023).

[CC04] Robert Stephen Cantrell and Chris Cosner. *Spatial ecology via reaction-diffusion equations*. John Wiley & Sons, 2004.

[Che+21] Emmalyn Chen et al. “Cell-free DNA concentration and fragment size as a biomarker for prostate cancer”. en. In: *Scientific Reports* 11.1 (Mar. 2021). Number: 1 Publisher: Nature Publishing Group, p. 5040. URL: https://www.nature.com/articles/s41598-021-84507-z (visited on 04/25/2023).

[Cis+22] M. Cisneros-Villanueva et al. “Cell-free DNA analysis in current cancer clinical trials: a review”. en. In: *British Journal of Cancer* 126.3 (Feb. 2022). Number: 3 Publisher: Nature Publishing Group, pp. 391–400. URL: https://www.nature.com/articles/s41416-021-01696-0 (visited on 12/22/2022).

[Coh+99] I Cohen et al. “Continuous and discrete models of cooperation in complex bacterial colonies”. In: *Fractals* 7 (1999), pp. 235–247.

[Cri+19] Stephen Cristiano et al. “Genome-wide cell-free DNA fragmentation in patients with cancer”. en. In: *Nature* 570.7761 (June 2019). Number: 7761 Publisher: Nature Publishing Group, pp. 385–389. URL: https://www.nature.com/articles/s41586-019-1272-6 (visited on 12/22/2022).

[Daw93] D A Dawson. “Measure-valued Markov processes”. In: *École d’été de probabilités de Saint Flour*. Vol. 1541. Springer-Verlag, 1993.

[DFL86] A. De Masi, P. A. Ferrari, and J. L. Lebowitz. “Reaction-diffusion equations for interacting particle system”. In: *J. Stat. Phys.* 44 (1986), pp. 589–644.

[DP06] Anna DeMasi and Errico Presutti. *Mathematical methods for hydrodynamic limits*. Springer, 2006.

[Dev+14] Alison S. Devonshire et al. “Towards standardisation of cell-free DNA measurement in plasma: controls for extraction efficiency, fragment size bias and quantification”. en. In: *Analytical and Bioanalytical Chemistry* 406.26 (Oct. 2014), pp. 6499–6512. URL: https://doi.org/10.1007/s00216-014-7835-3 (visited on 02/20/2023).

[Die+08] Frank Diehl et al. “Circulating mutant DNA to assess tumor dynamics”. en. In: *Nature Medicine* 14.9 (Sept. 2008). Number: 9 Publisher: Nature Publishing Group, pp. 985–990. URL: https://www.nature.com/articles/nm.1789 (visited on 12/22/2022).

[DEE88] Masao Doi, Samuel Frederick Edwards, and Samuel Frederick Edwards. *The theory of polymer dynamics*. Vol. 73. oxford university press, 1988.

[DK96] P J Donnelly and T G Kurtz. “A countable representation of the Fleming-Viot measure-valued diffusion”. In: *Ann. Probab.* 24 (1996), pp. 698–742.

[DK99] P J Donnelly and T G Kurtz. “Particle representations for measure-valued population models”. In: *Ann. Probab.* 27 (1999), pp. 166–205.

[Du +99] T W Du Clos et al. “Chromatin clearance in C57Bl/10 mice: interaction with heparan sulphate proteoglycans and receptors on Kupffer cells”. In: *Clinical and Experimental Immunology* 117.2 (Aug. 1999), pp. 403–411. URL: https://doi.org/10.1046/j.1365-2249.1999.00976.x (visited on 04/25/2023).

[DF16] R Durrett and W-T Fan. “Genealogies in expanding populations”. In: *Ann. Appl. Probab.* 26 (2016), pp. 3456–3490.

[Dzi+22] Rafal Dziadziuszko et al. “Circulating Cell-free DNA as a Prognostic Biomarker in Patients with Advanced ALK+ Non–small Cell Lung Cancer in the Global Phase III ALEX Trial”. In: *Clinical Cancer Research* 28.9 (May 2022), pp. 1800–1808. URL: https://doi.org/10.1158/1078-0432.CCR-21-2840 (visited on 04/25/2023).

[EBE81] H. J. Ederer, A. M. Basedow, and K. H. Ebert. “Modelling of Polymer Degradation Reactions”. In: *Modelling of Chemical Reaction Systems*. Ed. by Klaus H. Ebert, Peter Deuflhard, and Willi Jäger. Berlin, Heidelberg: Springer Berlin Heidelberg, 1981, pp. 189–215.

[EGL22] A Etheridge, M Gooding, and I Letter. “On the effects of a wide opening in the domain of the (stochastic) Allen-Cahn equation and the motion of hyrbid zones”. In: *Electron. J. Probab.* 27 (2022), pp. 1–52.

[Eth04] A M Etheridge. “Survival and extinction in a locally regulated population”. In: *Ann. Appl. Probab.* 14.1 (2004), pp. 188–214.

[EP22] A. Etheridge and S. Penington. “Genealogies in bistable waves”. In: *Electron. J. Probab.* 27.121 (2022), pp. 1–99.

[Eth00] Alison Etheridge. *An introduction to superprocesses*. 20. American Mathematical Society, 2000.

[EK19] Alison M. Etheridge and Thomas G. Kurtz. “Genealogical constructions of population models”. In: *Ann. Probab.* 47.4 (2019), pp. 1827–1910.

[EVY20] Alison M. Etheridge, Amandine Véber, and Feng Yu. “Rescaling limits of the spatial Lambda-Fleming-Viot process with selection”. In: *Electronic Journal of Probability* 25.none (2020), pp. 1–89. URL: https://doi.org/10.1214/20-EJP523.

[EK86] Stewart N. Ethier and Thomas G. Kurtz. *Markov processes – characterization and convergence*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. New York: John Wiley & Sons Inc., 1986, pp. x+534.

[Fer+19] Daniel Fernandez-Garcia et al. “Plasma cell-free DNA (cfDNA) as a predictive and prognostic marker in patients with metastatic breast cancer”. In: *Breast Cancer Research* 21.1 (Dec. 2019), p. 149. URL: https://doi.org/10.1186/s13058-019-1235-8 (visited on 04/25/2023).

[FM77] Paul C. Fife and J. B. McLeod. “The approach of solutions of nonlinear diffusion equations to travelling front solutions”. In: *Archive for Rational Mechanics and Analysis* 65.4 (Dec. 1977), pp. 335–361. URL: https://doi.org/10.1007/BF00250432.

[FH21] Franco Flandoli and Ruojun Huang. “The KPP equation as a scaling limit of locally interacting Brownian particles”. In: *J. Differential Equations* 303 (2021), pp. 608–644.

[FLO19] Franco Flandoli, Matti Leimbach, and Christian Olivera. “Uniform convergence of proliferating particles to the FKPP equation”. In: *Journal of Mathematical Analysis and Applications* 473.1 (2019), pp. 27–52.

[GTM96] V J Gauthier, L N Tyler, and M Mannik. “Blood clearance kinetics and liver uptake of mononucleosomes in mice.” In: *The Journal of Immunology* 156.3 (Feb. 1996), pp. 1151–1156. URL: https://doi.org/10.4049/jimmunol.156.3.1151 (visited on 04/25/2023).

[GG22] Olivia M. Ghosh and Benjamin H. Good. “Emergent evolutionary forces in spatial models of microbial growth in the human gut microbiota”. In: *Proc. Natl. Acad. Sci. (USA)* 119.28 (July 2022), e2114931119.

[GK05] B H Gilding and R Kersner. “A Fisher/KPP-type equation with density dependent diffusion and convection: travelling wave solutions”. In: *J. Phys. A: Math. Gen.* 38 (2005), pp. 337–3379.

[Goh+19] Daryl M. Gohl et al. “Measuring sequencer size bias using REcount: a novel method for highly accurate Illumina sequencing-based quantification”. In: *Genome Biology* 20.1 (Apr. 2019), p. 85. URL: https://doi.org/10.1186/s13059-019-1691-6 (visited on 05/24/2023).

[Goo18] M. Gooding. *Long term behaviour of spatial population models with heterozygous or asymmetric homozygous selection*. DPhil thesis, Oxford University, 2018. URL: https://ora.ox.ac.uk/objects/uuid:ef35b918-35ad-4f90-98ea-325fc692f1eb.

[Gra+20] Stefan Grabuschnig et al. "Putative origins of cell-free DNA in humans: a review of active and passive nucleic acid release mechanisms". In: *International journal of molecular sciences* 21.21 (2020), p. 8062.

[Haa05] Bénédicte Haas. "Equilibrium for Fragmentation with Immigration". In: *The Annals of Applied Probability* 15.3 (2005). Publisher: Institute of Mathematical Statistics, pp. 1958–1996. URL: http://www.jstor.org/stable/30038380 (visited on 04/24/2023).

[Haa21] Bénédicte Haas. "Tail asymptotics for extinction times of self-similar fragmentations". In: *arXiv preprint arXiv:2111.07757* (2021).

[HN08] O Hallatschek and D Nelson. "Gene surfing in expanding populations". In: *Theor. Pop. Biol.* 73 (2008), pp. 158–170.

[HM19] Benjamin C. Haller and Philipp W. Messer. "SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model". In: *Molecular Biology and Evolution* 36.3 (2019), pp. 632–637.

[HM22] Benjamin C. Haller and Philipp W. Messer. *SLiM: An Evolutionary Simulation Framework*. 2022. URL: https://messerlab.org/SLiM.

[Han+20] Diana S. C. Han et al. "The Biology of Cell-free DNA Fragmentation and the Roles of DNASE1, DNASE1L3, and DFFB". en. In: *The American Journal of Human Genetics* 106.2 (Feb. 2020), pp. 202–214. URL: https://www.sciencedirect.com/science/article/pii/S0002929720300082 (visited on 12/22/2022).

[HAS20] Ellen Heitzer, Lisa Auinger, and Michael R. Speicher. "Cell-Free DNA and Apoptosis: How Dead Cells Inform About the Living". English. In: *Trends in Molecular Medicine* 26.5 (May 2020). Publisher: Elsevier, pp. 519–528. URL: https://www.cell.com/trends/molecular-medicine/abstract/S1471-4914(20)30040-X (visited on 12/21/2022).

[HL04] Emilio Hernández-Garcıa and Cristóbal López. "Clustering, advection, and patterns in a model of population dynamics with neighborhood-dependent rates". In: *Phys. Rev. E* 70.1 (July 2004), p. 016216. URL: https://link.aps.org/doi/10.1103/PhysRevE.70.016216.

[HJ22] Ma. Elena Hernández-Hernández and Saul D. Jacka. "A generalisation of the Burkholder-Davis-Gundy inequalities". In: *Electronic Communications in Probability* 27.none (Jan. 2022). URL: https://doi.org/10.1214%2F22-ecp493.

[Hol+94] E. E. Holmes et al. "Partial Differential Equations in Ecology: Spatial Interactions and Population Dynamics". In: *Ecology* 75.1 (1994), pp. 17–29. URL: http://www.jstor.org/stable/1939378 (visited on 05/22/2023).

[Iva+15] Maxim Ivanov et al. "Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation". In: *BMC genomics* 16.13 (2015), pp. 1–12.

[JS13] Jean Jacod and Albert Shiryaev. *Limit theorems for stochastic processes*. Vol. 288. Springer Science & Business Media, 2013.

[Joh+22] Andrew D Johnston et al. “Modelling clinical DNA fragmentation in the development of universal PCR-based assays for bisulfite-converted, formalin-fixed and cell-free DNA sample analysis”. In: *Scientific reports* 12.1 (2022), p. 16051.

[Kal97] Olav Kallenberg. *Foundations of modern probability*. Vol. 2. Springer, 1997.

[KR04] S Kamin and P Rosenau. “Emergence of waves in a nonlinear convection-reaction-diffusion equation”. In: *Adv. Nonlinear Stud.* 4 (2004), pp. 251–272.

[Kat+22a] Efrat Katsman et al. *Data for for Detecting cell-of-origin and cancer-specific methylation features of cell-free DNA from Nanopore sequencing*. Apr. 2022.

[Kat+22b] Efrat Katsman et al. “Detecting cell-of-origin and cancer-specific methylation features of cell-free DNA from Nanopore sequencing”. In: *Genome Biology* 23.1 (July 2022), p. 158. URL: https://doi.org/10.1186/s13059-022-02710-1 (visited on 12/22/2022).

[Kel+19] Jerome Kelleher et al. “Inferring whole-genome histories in large population datasets”. In: *Nature Genetics* 51.9 (2019), pp. 1330–1338. URL: https://doi.org/10.1038/s41588-019-0483-y.

[Kha+18] Khurum H Khan et al. “Longitudinal liquid biopsy and mathematical modeling of clonal evolution forecast time to treatment failure in the PROSPECT-C phase II colorectal cancer clinical trial”. In: *Cancer discovery* 8.10 (2018), pp. 1270–1285.

[KL18] Sonia Khier and Laura Lohan. “Kinetics of circulating cell-free DNA for biomedical applications: critical appraisal of the literature”. In: *Future science OA* 4.4 (2018), FSO295.

[KPP37] A Kolomogorov, I Petrovsky, and N Piscounov. “Étude de l’equation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique”. In: *Moscow Univ. Math. Bull.* 1 (1937), pp. 1–25.

[KR11] Thomas G. Kurtz and Eliane R. Rodrigues. “Poisson representations of branching Markov and measure-valued branching processes”. In: *Ann. Probab.* 39.3 (May 2011), pp. 939–984. URL: https://doi.org/10.1214/10-AOP574.

[KX99] Thomas G. Kurtz and Jie Xiong. “Particle representations for a class of nonlinear SPDEs”. In: *Stoch. Proc. Appl.* 83.1 (1999), pp. 103–126. URL: https://www.sciencedirect.com/science/article/pii/S0304414999000241.

[Kus+19] Anatoli Kustanovich et al. “Life and death of circulating cell-free DNA”. In: *Cancer Biology & Therapy* 20.8 (Apr. 2019), pp. 1057–1067. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6606043/ (visited on 02/20/2023).

[Lak+04] Pavel P. Laktionov et al. “Cell-Surface-Bound Nucleic Acids: Free and Cell-Surface-Bound Nucleic Acids in Blood of Healthy Donors and Breast Cancer Patients”. en. In: *Annals of the New York Academy of Sciences* 1022.1 (2004). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1196/annals.1318.034, pp. 221–227. URL: https://onlinelibrary.wiley.com/doi/abs/10.1196/annals.1318.034 (visited on 04/27/2023).

[LL23] King-Yeung Lam and Yuan Lou. *Introduction to Reaction-Diffusion Equations: Theory and Applications to Spatial Ecology and Evolutionary Biology*. Springer International Publishing AG, 2023.

[LBS22] Y Li, P Buenzli, and M Simpson. “Interpreting how nonlinear diffusion affects the fate of bistable populations using a discrete modelling framework”. In: *Proc. Roy. Soc. A* 478 (2022).

[LM01] P-L Lions and S Mas-Gallic. “Une méthode particulaire déterministe pour des équations diffusives non linéaires”. In: *C. R. Acad. Sci. Paris* 332, Série I (2001), pp. 369–376.

[Liu+20] M. C. Liu et al. “Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA”. English. In: *Annals of Oncology* 31.6 (June 2020). Publisher: Elsevier, pp. 745–759. URL: `https://www.annalsofoncology.org/article/S0923-7534(20)36058-0/fulltext` (visited on 12/22/2022).

[Liu22] Yaping Liu. “At the dawn: cell-free DNA fragmentomics and gene regulation”. In: *British Journal of Cancer* 126.3 (2022), pp. 379–390.

[Lo+21] Y. M. Dennis Lo et al. “Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies”. In: *Science* 372.6538 (Apr. 2021). Publisher: American Association for the Advancement of Science, eaaw3616. URL: `https://www.science.org/doi/10.1126/science.aaw3616` (visited on 12/21/2022).

[McK67] H P McKean. “Propagation of chaos for a class of non-linear parabolic equations”. In: *Lecture Series in Differential Equations* 2 (1967), pp. 177–194.

[MNT17] Bhuvan Molparia, Eshaan Nichani, and Ali Torkamani. “Assessment of circulating copy number variant detection for cancer screening”. en. In: *PLOS ONE* 12.7 (July 2017). Publisher: Public Library of Science, e0180647. URL: `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0180647` (visited on 04/24/2023).

[Mor58] P A P Moran. “Random processes in genetics”. In: *Proc. Camb. Phil. Soc.* 54 (1958), pp. 60–71.

[Mos+23] Tina Moser et al. “Bridging biological cfDNA features and machine learning approaches”. In: *Trends in Genetics* (2023).

[Mou+18] Florent Mouliere et al. “Enhanced detection of circulating tumor DNA by fragment size analysis”. In: *Science Translational Medicine* 10.466 (Nov. 2018). Publisher: American Association for the Advancement of Science, eaat4921. URL: `https://www.science.org/doi/full/10.1126/scitranslmed.aat4921` (visited on 04/24/2023).

[Mou+21] Florent Mouliere et al. “Fragmentation patterns and personalized sequencing of cell-free DNA in urine and plasma of glioma patients”. In: *EMBO Molecular Medicine* 13.8 (2021), e12881.

[NYK93] D Nadano, T Yasuda, and K Kishi. “Measurement of deoxyribonuclease I activity in human tissues and body fluids by a single radial enzyme-diffusion method”. In: *Clinical Chemistry* 39.3 (Mar. 1993), pp. 448–452. URL: `https://doi.org/10.1093/clinchem/39.3.448` (visited on 10/19/2021).

[NA93] J E Neigel and J C Avise. “Application of a random walk model to geographic distributions of animal mitochondrial DNA variation”. In: *Genetics* 135.4 (Dec. 1993), pp. 1209–1220.

[Oel85] Karl Oelschläger. “A law of large numbers for moderately interacting diffusion processes”. In: *Z. Wahrsch. verw. Geb.* 69.2 (1985), pp. 279–322.

[Oel89] Karl Oelschläger. “On the derivation of reaction-diffusion equations as limit dynamics of systems of moderately interacting stochastic processes”. In: *Prob. Theor. Rel. Fields* 82.4 (1989), pp. 565–586.

[Pen+21] Peter Peneder et al. “Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden”. en. In: *Nature Communications* 12.1 (May 2021). Number: 1 Publisher: Nature Publishing Group, p. 3230. URL: https://www.nature.com/articles/s41467-021-23445-w (visited on 04/24/2023).

[Pen17] S Penington. “The spreading speed of solutions of the non-local Fisher-KPP equation”. In: *J. Functional Anal.* 275.12 (2017), pp. 3259–3302.

[Per92] E A Perkins. “Measure-valued branching diffusions with spatial interactions”. In: *Prob. Th. Rel. Fields* 94 (1992), pp. 189–245.

[PB23] Jonathan R. Potts and Luca Börger. “How to scale up from animal movement decisions to spatiotemporal patterns: An approach via step selection”. In: *Journal of Animal Ecology* 92.1 (2023), pp. 16–29. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/1365-2656.13832. URL: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2656.13832.

[Qi+23] Ting Qi et al. “Cell-Free DNA Fragmentomics: The Novel Promising Biomarker”. In: *International Journal of Molecular Sciences* 24.2 (2023), p. 1503.

[Reb80] R Rebolledo. “Sur l’existence de solutions á certains problèmes de semimartingales”. In: *C. R. Acad. Sci. Paris* 290 (1980).

[Ren+22] Gabriel Renaud et al. “Unsupervised detection of fragment length signatures of circulating tumor DNA using non-negative matrix factorization”. In: *Elife* 11 (2022), e71569.

[Rol+20] Christian Rolfo et al. “Challenges and opportunities of cfDNA analysis implementation in clinical practice: Perspective of the International Society of Liquid Biopsy (ISLB)”. In: *Critical Reviews in Oncology/Hematology* 151 (2020), p. 102978. URL: https://www.sciencedirect.com/science/article/pii/S1040842820301165.

[Sad+21] Ronen Sadeh et al. “ChIP-seq of plasma cell-free nucleosomes identifies gene expression programs of the cells of origin”. en. In: *Nature Biotechnology* 39.5 (May 2021). Number: 5 Publisher: Nature Publishing Group, pp. 586–598. URL: https://www.nature.com/articles/s41587-020-00775-6 (visited on 12/22/2022).

[Sas97] Akira Sasaki. “Clumped Distribution by Neighbourhood Competition”. In: *J. Theor. Biol.* 186.4 (June 1997), pp. 415–430. URL: https://doi.org/10.1006%2Fjtbi.1996.0370.

[She10] J A Sherratt. “On the form of smooth-front travelling waves in a reaction-diffusion equation with degenerate nonlinear diffusion”. In: *Math. Model. Nat. Phenom.* 5.5 (2010), pp. 64–79.

[She91] S J Sheu. “Some estimates of the transition density of a nondegenerate diffusion Markov process”. In: *Ann. Probab.* 19.2 (1991), pp. 538–561.

[Sie+21] Paulina Siejka-Zielińska et al. “Cell-free DNA TAPS provides multimodal information for early cancer detection”. In: *Science Advances* 7.36 (Sept. 2021). Publisher: American Association for the Advancement of Science, eabh0534. URL: https://www.science.org/doi/full/10.1126/sciadv.abh0534 (visited on 04/25/2023).

[Sin+94] D. Sinicropi et al. “Colorimetric Determination of DNase I Activity with a DNA-Methyl Green Substrate”. en. In: *Analytical Biochemistry* 222.2 (Nov. 1994), pp. 351–358. URL: https://www.sciencedirect.com/science/article/pii/S0003269784715028 (visited on 04/25/2023).

[Sny+16] Matthew W. Snyder et al. “Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin”. English. In: *Cell* 164.1 (Jan. 2016). Publisher: Elsevier, pp. 57–68. URL: https://www.cell.com/cell/abstract/S0092-8674(15)01569-X (visited on 12/22/2022).

[Tam+06] Svetlana N. Tamkovich et al. “Circulating DNA and DNase Activity in Human Blood”. en. In: *Annals of the New York Academy of Sciences* 1075.1 (2006). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1196/annals.1368.026, pp. 191–196. URL: https://onlinelibrary.wiley.com/doi/abs/10.1196/annals.1368.026 (visited on 04/25/2023).

[Und+16] Hunter R. Underhill et al. “Fragment Length of Circulating Tumor DNA”. en. In: *PLOS Genetics* 12.7 (July 2016). Publisher: Public Library of Science, e1006162. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006162 (visited on 04/24/2023).

[YRS01] W. R. Young, A. J. Roberts, and G. Stuhne. “Reproductive pair correlations and the clustering of organisms”. In: *Nature* 412.6844 (2001), pp. 328–331. URL: https://doi.org/10.1038/35085561.

[Yu+23] Stephanie C Y Yu et al. “Comparison of Single Molecule, Real-Time Sequencing and Nanopore Sequencing for Analysis of the Size, End-Motif, and Tissue-of-Origin of Long Cell-Free DNA in Plasma”. In: *Clinical Chemistry* 69.2 (Feb. 2023), pp. 168–179. URL: https://doi.org/10.1093/clinchem/hvac180 (visited on 05/24/2023).

[Yu+13] Stephanie CY Yu et al. “High-Resolution Profiling of Fetal DNA Clearance from Maternal Plasma by Massively Parallel Sequencing”. In: *Clinical Chemistry* 59.8 (Aug. 2013), pp. 1228–1237. URL: https://doi.org/10.1373/clinchem.2013.203679 (visited on 12/22/2022).

[Zho+23] Ze Zhou et al. “Fragmentation landscape of cell-free DNA revealed by deconvolutional analysis of end motifs”. In: *Proceedings of the National Academy of Sciences* 120.17 (2023), e2220982120.

[ZM85] R. M. Ziff and E. D. McGrady. “The kinetics of cluster fragmentation and depolymerisation”. In: *Journal of Physics A: Mathematical and General* 18.15 (Oct. 1985), p. 3027. URL: https://dx.doi.org/10.1088/0305-4470/18/15/026.