The spatial structure of genetic diversity under natural selection and in heterogeneous environments

Raphael Forien

NNT : 2017SACLX082

THESE DE DOCTORAT
DE L'UNIVERSITE PARIS-SACLAY

préparée à

L'ÉCOLE POLYTECHNIQUE

ÉCOLE DOCTORALE N°574

École doctorale de mathématiques Hadamard (EDMH)

Spécialité de doctorat : Mathématiques

par

Raphaël Forien

Structure spatiale de la diversité génétique : influence de la sélection naturelle et d'un environnement hétérogène

Thèse présentée et soutenue à Palaiseau, le 24 novembre 2017.

Après avis des rapporteurs :

Étienne Pardoux (IMM - Aix-Marseille Université)
Steven N. Evans (University of California Berkeley)

Composition du jury :

Jean-François Delmas	(CERMICS - École des Ponts ParisTech)	Président du jury
Étienne Pardoux	(IMM - Aix-Marseille Université)	Rapporteur
Sylvie Méléard	(CMAP - École Polytechnique)	Examinatrice
Emmanuel Schertzer	(LPMA - Université Paris 6 Jussieu)	Examinateur
Amandine Véber	(CMAP - École Polytechnique)	Directrice de thèse
Alison Etheridge	(Dept. of Statistics - University of Oxford)	Directrice de thèse

Thèse de doctorat

Structure spatiale de la diversité génétique: influence de la sélection naturelle et d'un environnement hétérogène

préparée par

Raphaël Forien

sous la direction de

Amandine Véber et Alison Etheridge

année universitaire 2017 - 2018

Contents

Remerciements

Après trois ans, neuf cahiers et douze stylos, je prends quelques lignes pour remercier toutes celles et ceux qui m'ont permis d'en arriver là.

Tout d'abord, cette thèse n'aurait pas été possible sans les conseils et les encouragements de mes deux directrices. Je ne peux résumer en quelques lignes tout ce qu'elles m'ont apporté au long de ces trois années, à la fois scientifiquement et humainement. Elles n'ont eu de cesse de me guider tout en me laissant une grande liberté dans mes recherches (en insistant tout de même pour que je travaille sur leurs projets respectifs et non ceux de l'autre !). J'ai énormément profité de leur expérience et j'espère pouvoir continuer à travailler avec elles dans les années à venir.

Je dois également beaucoup à mes collaborateurs, en particulier à Sarah Penington, sans qui je serais encore en train de changer les notations dans mon premier papier, à Harald Ringbauer, qui m'a fait découvrir les "massifs de la corrélation dans la mer d'improbabilité" et à Graham Coop, qui m'a si bien accueilli à Davis et avec qui j'ai eu beaucoup de plaisir à interagir.

Je remercie Steven Evans et Étienne Pardoux qui ont accepté de relire ma thèse. Leurs remarques ont permis de corriger certains arguments et ont grandement amélioré la présentation de certains chapitres. Je suis reconnaissant envers Jean-François Delmas, Sylvie Méléard, Étienne Pardoux et Emmanuel Schertzer qui me font l'honneur de participer à mon jury de thèse.

Je souhaite également remercier Nina Gantert, Tom Kurtz et Nick Barton, avec qui j'ai eu l'occasion de discuter de mes travaux et dont les remarques m'ont été très profitables.

Rien ne pourrait se faire au CMAP, pas même une thèse, sans le travail constant de l'équipe administrative. Je remercie donc en particulier Nasséra, Alex, Manoella et Vincent qui m'ont notamment permis de me sortir des méandres de la réinscription à Paris Saclay au début de chaque année. Je voudrais également mentionner Sylvain et Pierre, grâce à qui mon ordinateur revit après chaque reboot prendant une mise à jour.

Grâce à la chaire Modélisation Mathématiques et Biodiversité, financée par Véolia et dont le Museum d'Histoire Naturelle et la Fondation de l'École Polytechnique sont partenaires, j'ai pu profiter de nombreux échanges avec d'autres chercheurs et étudiants qui travaillent à l'interface entre mathématiques et biologie. Je remercie Sylvie pour tout le temps qu'elle consacre à cette chaire et tous ses membres qui en font un espace d'échanges riche et dynamique.

Je remercie également les doctorants du CMAP grâce à qui j'ai le courage d'affronter le RER (presque) tous les matins. En premier lieu le bureau 2016 : Jean-Bernard toujours à la recherche de

bureaux, Perle et ses mandalas, Hadrien jamais sans sa serviette, Mathieu, Florian, Cédric, Émile et Paul. Je remercie également Ludovic, Benoît, Matthieu, Aymeric et Céline avec qui j'ai organisé le séminaire des doctorant.e.s. Merci aussi à Aline avec qui j'ai partagé les galères de la rédaction et de la soutenance. Merci également à Tristan, José, Kévish, Dorian et tous les autres. Je n'oublie pas non plus tous ceux qui ont déjà quitté le CMAP mais qui y ont laissé leur marque : Aymeric, Étienne, Clément (qui a fourni le template de la page de garde de ce manuscrit), Hélène, Manon, Simona, Antoine, Romain et Massil. Enfin, merci à Lucas, doctorant par adoption et toujours fidèle à la pause café !

Au cours de mes différentes visites à Oxford, j'ai eu le plaisir d'être accueilli par une équipe de doctorants à laquelle j'ai pu facilement m'intégrer. Merci à Sarah, Mitch et Dominic pour les pub quizz et les parties de croquet, merci à Franz et Eleonora pour les sorties au Spin Jazz, merci également à Dane, Helmut, Luke, Daniel et Mikolaj. Enfin, je ne peux manquer de remercier Jane, Emma et surtout Beverley dont les cookies auront rythmé chacune de mes visites. J'ai également eu la chance de découvrir plusieurs coins de nature surprenamment vallonnés du Royaume-Uni grâce à l'Oxford University Walking Club et à Jaya, Michiel, Helen, Christian, Thomas, Nicola et à Laurent.

J'ai par ailleurs eu le plaisir de faire partie du Coop Lab pendant un mois l'an dernier, et je veux remercier Nancy et Doc pour m'avoir offert ma première séance d'ornithologie sur le continent américain. Merci à Emily de m'avoir dépanné d'une tente et d'un sac de couchage, et merci à Sivan, Vincent, Anita, Brandon, Kristin et Erin pour leur enthousiasme au tea tasting (même si je suis sûr que la compétition était truquée depuis le début en faveur du Yorkshire Tea !).

Il me reste à remercier mon frère avec qui j'ai toujours pu partager ma passion des maths, et mes parents, à qui je dois tout, et qui ont toujours fait en sorte que nous recevions la meilleure éducation possible. Enfin, Marthe, ton soutien est la chose la plus précieuse qui m'ait été offerte.

At present one may say that the mathematical theory of evolution is in a somewhat unfortunate position, too mathematical to interest most biologists, and not sufficiently mathematical to interest most mathematicians. Nevertheless, it is reasonable to suppose that in the next half century it will be developed into a respectable branch of Applied Mathematics.

J.B.S. Haldane, 1938

Summary of the Introduction

Most spieces occupy a spatially extended habitat and have done so for many generations. Since individuals travel some distance between each generation, the geographical structure of populations shapes their genetic diversity and individuals living close to each other tend to be more related and to share more genetic material than more distant individuals.

In this thesis, we study the fluctuations in the genetic composition of spatially extended populations under natural selection and in heterogeneous environments. The aim of this work is to be able to infer the main demographic and geographic features of a population from a sample of genetic sequences.

Spatial models in population genetics Mathematical models including spatial structure have a long history in population genetics [Wri43; Fis37; Mal48]. In discrete space, the stepping stone framework developed by Wright [Wri43] and Kimura [Kim64] has imposed itself as the standard way to model the evolution of allele frequencies in a structured population. Individuals are assumed to live in colonies or demes located on an integer lattice of dimension $d \in \{1, 2, 3\}$ and at each generation, they reproduce within each colony and exchange a number of migrants with neighbouring colonies. Let N_i be the number of individuals in deme $i \in \mathbb{Z}^d$ and $\tilde{m}_{ij}$ the probability that an individual in deme i has an ancestor in deme j in the previous generation. Then if $p(t, i)$ denotes the proportion of individuals carrying a given allele in deme i at time t, $(p(t, \cdot), t \geq 0)$ satisfies the following system of stochastic differential equations:

$$dp(t,i) = \sum_j \tilde{m}_{ij}(p(t,j) - p(t,i))dt + \sqrt{\frac{1}{N_i} p(t,i)(1-p(t,i))}dB_t^i, \qquad i \in \mathbb{Z}^d,$$

where $(B^i, i \in \mathbb{Z}^d)$ is a family of independent standard Brownian motions.

Stationary models for the evolution of populations living in a continuous geographical space have eluded population geneticists for some time [Fel75; BDE02]. Recently, however, a new framework was proposed in [Eth08; BEV10a] which overcomes the main difficulties raised by Felsenstein (see also [Kin77]). This model, the spatial Λ-Fleming-Viot process (SLFV), represents the genetic composition of a population by a measure on $\mathbb{R}^d \times K$, where K is the space of genetic types in the population. When only two types are present (say 0 and 1), this measure reduces to a random function $w_t : \mathbb{R}^d \to [0, 1]$ indicating the proportion of type 1 individuals at each point.

The SLFV evolves through successive reproduction events during which some individuals in a

given region die and are replaced by the offspring of a randomly choosen parent. More precisely, let μ be a measure on $(0,\infty)$ and for $r>0$ let ν_r be a probability measure on $(0,1]$ such that

$$\int_0^\infty \int_0^1 u r^d \nu_r(du)\mu(dr) < \infty.$$

The process $(w_t)_{t\geq 0}$ is then defined as follows.

Definition (spatial Λ-Fleming-Viot process)**.** *Let Π be a Poisson point process on $\mathbb{R}_+ \times \mathbb{R}^d \times (0,\infty) \times (0,1]$ with intensity measure $dt \otimes dx \otimes \mu(dr)\nu_r(du)$. For every $(t,x,r,u) \in \Pi$, a reproduction event with impact parameter u hits the ball of centre x and radius r at time t:*

i) pick a location y uniformly in $B(x,r)$ and pick a type $k \in \{0,1\}$ such that $k=1$ with probability $w_{t^-}(y)$ and $k=0$ otherwise,

ii) update w for $z \in B(x,r)$ with

$$w_t(z) = (1-u)w_{t^-}(z) + uk,$$

and leave w unchanged outside of $B(x,r)$.

Both the stepping stone model and the SLFV admit an equivalent description in terms of their dual which describes the genealogy of a random sample of individuals as a system of coalescing random walks (or coalescing ancestral lineages). See [BEV13a] for a more detailed discussion of duality and other results concerning the SLFV.

This dual provides a convenient way to study the large scale behaviour of these models. In particular, if we rescale time by n and space by $\sqrt{n}$ and if we let n tend to infinity, each of these lineages converges to Brownian motion. If u and r are assumed to be fixed (*i.e.* μ and ν_r are Dirac measures δ_r and δ_u), then Berestycki et al. [BEV13b] show that in the limit these lineages coalesce upon meeting in one spatial dimension, but never meet in higher dimensions. As a result, the SLFV (under the same rescaling) converges to the solution to the heat equation when $d \geq 2$ and to a system of Bernoulli random variables when $d=1$.

On the other hand, if u decreases as we rescale space and time, other regimes can be observed. Etheridge, Véber and Yu have investigated one of them in [EVY14], where they consider the SLFV with natural selection (see Subsection 2.1.2) and show that if u decreases as $\frac{1}{\sqrt{n}}$ and if we rescale space and time appropriately, the SLFVS converges in distribution to a process $(w_t^\infty)_{t\geq 0}$ as $n \to \infty$ (keeping the assumption that u and r are fixed for each n). In one spatial dimension, w^∞ is the solution to the following stochastic partial differential equation:

$$\partial_t w_t^\infty = \frac{\sigma^2}{2}\Delta w_t^\infty + s w_t^\infty(1-w_t^\infty) + \sqrt{\frac{1}{N} w_t^\infty(1-w_t^\infty)}\dot{W}_t,$$

where W is space-time white noise (see [Wal86]) and σ^2 and N are given by (denoting the volume of a ball of radius r by V_r)

$$\sigma^2 = u V_r \frac{2r^2}{d+2} \qquad \text{and} \qquad N = \frac{1}{V_r^2 u^2}.$$

If $d \geq 2$, then w^∞ is deterministic and solves the celebrated Fisher-KPP equation

$$\partial_t f_t = \frac{\sigma^2}{2}\Delta f_t + s f_t(1 - f_t).$$

Since, at least in high dimensions, the SLFV is close to being deterministic on large spatial and temporal scales, we wish to describe how it fluctuates around this limit.

A central limit theorem for the SLFV with selection With Sarah Penington (Oxford Universiy), we studied these fluctuations in a regime where the SLFV is close to the solution of the Fisher-KPP equation, even in one spatial dimension. To do this, we adapted the works of Norman [Nor74a; Nor75a], who studied the Wright-Fisher model in a strong selection regime. We show in Chapter 2 that the fluctuations of the SLFV around a sequence of deterministic functions converging to $(f_t)_{t\geq 0}$ are given by the solution to the following stochastic PDE:

$$\begin{cases} \partial_t z_t = \dfrac{\sigma^2}{2}\Delta z_t + s(1-2f_t)z_t + \sqrt{\dfrac{1}{N} f_t(1-f_t)}\dot{W}_t \\ z_0 = 0, \end{cases}$$

where W is space-time white noise.

We also studied another regime in which the radius of each reproduction event is drawn from a distribution in the domain of attraction of a stable law. In this case the motion of rescaled lineages converges to a stable Lévy process, and the results above hold after replacing the standard Laplacian with a fractionnal Laplacian and space-time white noise by a coloured noise (which is only correlated in space).

Note that the equation for the fluctuations can be obtained by linearizing the stochastic partial differential equation solved by w^∞ in one dimension when it is close to f_t. Also for any smooth and compactly supported function $\phi : \mathbb{R}^d \to \mathbb{R}$, $\langle z_t, \phi\rangle$ is a Gaussian random variable, which is why this result can be interpreted as a central limit theorem for the SLFV with selection.

We apply this result to the case of overdominance in order to estimate the drift load in spatially structured populations. Overdominance occurs when individuals are diploid and heterozygous individuals at a given locus produce more offspring than homozygous individuals. If two alleles A_1 and A_2 are present in the population, the relative fitnesses of the three possible genotypes are

A_1A_1	A_1A_2	A_2A_2
$1 - s_1$	1	$1 - s_2$.

In this configuration, a stable intermediate allele frequency is maintained in the population, for which the mean fitness of the population is maximised. If the population size N is small, the actual allele frequency deviates from this equilibrium as a result of genetic drift, thereby reducing the mean fitness of the population. This reduction in the mean fitness of the population is what we call drift load. Robertson [Rob70] showed that in a panmictic population, this reduction is close to $\frac{1}{4N}$, irrespective of the strength of natural selection.

Adapting our central limit theorem to this setting, we show in Chapter 2 that spatially structured populations follow a different regime. We study the SLFVS with overdominance with fixed radius r

and impact parameter u and we show that when both selection parameters s_1 and s_2 are small, the drift load is proportional to

$$\begin{cases} u\sqrt{s_1+s_2} & \text{if } d=1 \\ u(s_1+s_2)\,|\ln(s_1+s_2)| & \text{if } d=2 \\ u(s_1+s_2) & \text{if } d\geq 3. \end{cases}$$

We note that the drift load is smaller in spatially structured populations compared to the panmictic case, and that it decreases with the dimension of the habitat (the parameter u should be compared to $\frac{1}{2N}$ as it corresponds to the proportion of individuals replaced at each reproduction). This is due to the smoothing effect of migrations: allele frequencies which locally deviate from the equilibrium are brought back to it by an arrival of migrants which are on average close to the equilibrium.

Dispersal heterogeneity in the spatial Λ-Fleming-Viot process The above results suggest that the two main demographic parameters governing the genetic evolution of a population are σ^2 and N, *i.e.* the mean square displacement of individuals per generation and the population density. But these characteristics are seldom uniform across the whole range of a spieces. In particular, sharp transitions between regions with different demographic parameters can lead to qualitatively different behaviours compared to the homogeneous case.

In [Nag76], Nagylaki proposed a one dimensional stepping stone model with variable migration, where demes left of the origin exchange fewer migrants per generation than neighbouring demes right of the origin (see Figure 1.4). He showed that in the absence of genetic drift, the allele frequencies converge under a diffusive rescaling to the solution to the following equation

$$\begin{cases} \partial_t \rho(t,x) = \dfrac{\sigma(x)^2}{2}\partial_{xx}\rho(t,x) \\ (1+\beta)\partial_x\rho(t,0^+) = (1-\beta)\partial_x\rho(t,0^-), \end{cases}$$

where $\sigma(x) = \sigma_\pm$ if $\pm x > 0$ and $\beta \in (-1,1)$ depends on the details of the model near the origin.

In Chapter 3 of this thesis, we study the SLFV with heterogeneous dispersal, *i.e.* when the radius of a reproduction event depends on the halfspace in which its centre falls (see Figure 1.6). We note that its dual consists of a system of coalescing random walks which behave as symmetric random walks outside of a bounded region around the origin. We then adapt the results of Iksanov and Pilipenko [IP16] to show that, after rescaling time by n and space by $\sqrt{n}$, these random walks converge in distribution to skew Brownian motions.

Skew Brownian motion was introduced by Walsh [Wal78] and Itô and McKean [IM63] and can be roughly described as a process performing Brownian excursions to the left of the origin with probability $\frac{1-\beta}{2}$ and to the right with probability $\frac{1+\beta}{2}$, for some $\beta \in (-1,1)$. Harrison and Shepp [HS81] showed that skew Brownian motion can be characterised as the unique real-valued Markov process $(X_t)_{t\geq 0}$ solving

$$X_t = B_t + \beta L_t^0(X)$$

where $L^0(X)$ is the local time at 0 of X and B is standard Brownian motion.

In our setting, we show that ancestral lineages in the dual of the SLFV with heterogeneous

dispersal converge to solutions of

$$X_t^1 = X_0^1 + \int_0^t \sigma(X_s^1) dB_s^1 + \beta L_t^0(X^1)$$
$$X_t^i = X_0^i + \int_0^t \sigma(X_s^1) dB_s^i \qquad i \geq 2,$$

where $B = (B^1, \dots, B^d)$ is d-dimensional standard Brownian motion. Furthermore, as in the result of [BEV13b], ancestral lineages coalesce upon meeting in the limit when $d = 1$ but never meet in higher dimensions. As a consequence, we obtain the large scale behaviour of the SLFV with heterogeneous dispersal: in one spatial dimension, it converges to a system of Bernoulli random variables with parameters given by $(\rho(t, x), x \in \mathbb{R}^d)$ and in higher dimensions, it is deterministic and equals ρ.

Barriers to gene flow The genetic composition of populations can also be affected by physical obstacles which reduce migration between different parts of its habitat. Most current methods to quantify and measure the strength of these barriers to gene flow do not rely on spatially explicit models and are potentially subject to confounding by isolation by distance patterns.

In [Nag76], Nagylaki studied a one dimensional stepping stone model in which the two neighbouring demes closest to the origin exchange significantly fewer migrants at each generation than other demes (see Figure 1.8). As for the model with dispersal heterogeneity, he showed that the allele frequencies could be approximated by the solution of

$$\begin{cases} \partial_t p(t, x) = \dfrac{\sigma^2}{2} \partial_{xx} p(t, x) \\ \partial_x p(t, 0^+) = \partial_x p(t, 0^-) = \gamma(p(t, 0^+) - p(t, 0^-)), \end{cases}$$

for some $\gamma \in (0, \infty)$, see Figure 4.2. The parameter γ measures the permeability of the barrier, if $\gamma \to \infty$ then the barrier has no effect on allele frequencies, while if $\gamma \to 0$, the two populations on each side of the origin evolve independently of each other.

In Chapter 4, we are interested in the motion of ancestral lineages in this setting. We show that they can be approximated by a process that we call partially reflected Brownian motion, which behaves like reflected Brownian motion until its local time at the origin reaches an exponential random variable with some parameter $\gamma \in (0, \infty)$, after which it behaves like reflected Brownian motion on the other side of the origin until its local time reaches another independent exponential random variable, and so on.

We also provide another construction of partially reflected Brownian motion from the excursions of standard Brownian motion outside a region of length $\frac{1}{\gamma}$, we obtain an explicit formula for its transition density and we give a martingale problem characterisation of this process. This process was previously obtained in [MP16] as a limit of one dimensional diffusions and its transition density appears in [GVNL14] as the fundamental solution to Nagylaki's equation. In passing we recover Nagylaki's approximation of allele frequencies by duality. Finally, this model was used in [Rin+17] to detect barriers to gene flow from distorted isolation by distance patterns, hence allowing to fully take into account the spatial structure of a population when estimating the strength of a barrier.

Demographic inference in heterogeneous environments Two individuals who share at least one recent common ancestor typically inherit one or more continuous blocks of genome from this ancestor. These blocks, delimited by recombination events, are called blocks of identity by descent, or IBD blocks. Blocks inherited from very recent common ancestors tend to be longer than more ancient blocks, because they have gone through fewer recombinations. It follows that IBD blocks shared within a sample of individuals carry some information about their genealogy and the demographic history of the whole population.

Recently, Ralph and Coop [RC13] have been able to identify IBD blocks in the European subset of the POPRES dataset [Nel+08]. IBD sharing in Europe shows clear signs of isolation by distance, with different rates of decay of relatedness in different parts of the continent. Ringbauer et al. [RCB16] have used this pattern to infer both the mean squared displacement of individuals per generation and the effective population density in Eastern Europe. Their approach, based on a composite likelihood maximisation, found that the rate of dispersal in this region is close to 62 km/$\sqrt{\text{gen}}$.

In Chapter 5, we present work in progress with Harald Ringbauer (IST Austria) and Graham Coop (UC Davis California) to extend this method to allow both of these parameters to vary in different parts of space. We show how to compute numerically the expected number of IBD blocks of a given length shared between two individuals depending on their geographical separation, and we present some results on the performance of our method on simulated datasets.

Chapitre 1

Introduction

1.1 La géographie racontée par les gènes

Chaque être humain sur Terre possède 23 paires de chromosomes dans le noyau de ses cellules. Ceux-ci sont constitués d'une très longue molécule recroquevillée sur elle même : l'ADN, qui se présente sous la forme de deux brins enroulés l'un autour de l'autre en une double hélice. Chaque brin est formé d'une suite de quatre bases azotées : adénine (A), thymine (T), cytosine (C) et guanine (G). En mettant bout à bout les 46 chromosomes d'un humain, on obtient une séquence de plus de 3 milliards de paires de bases.

Cette séquence est le fruit de l'union des chromosomes maternels et paternels présents dans l'ovule et le spermatozoïde dont il est issu, elle est donc unique pour chaque individu (à l'exception des vrais jumeaux). En comparant les séquences de deux individus, on peut donc savoir s'ils sont frères et sœurs, mère et fille, cousins, etc. À une plus large échelle, en comparant les génomes d'individus issus de deux populations, on peut dire si des échanges ont lieu régulièrement entre ces deux populations et estimer leur durée et leur intensité.

L'état actuel de la diversité génétique au sein d'une population nous renseigne donc sur son histoire démographique. Pour déchiffrer le langage dans lequel cette histoire est écrite, on doit faire appel aux outils mathématiques de la génétique des populations. Les premiers modèles dans ce domaine on été introduits dans les années 1920-1930 par Ronald A. Fisher, John B. S. Haldane et Sewall Wright. Ils furent les premiers à mettre en équation l'évolution de la variabilité génétique au sein d'une population, et leurs résultats sont encore aujourd'hui largement utilisés par la communauté des généticiens.

Si différents individus d'une même population présentent des séquences génétiques différentes à une position donnée du génome, on s'intéresse aux variations du nombre d'individus portant la même version de cette séquence. La position de cette portion de séquence dans le génome est appelée *locus*, et les différentes versions de la séquence génétique à ce locus sont appelées allèles. Lorsque plusieurs allèles sont présents dans la population, on dit que celle-ci est polymorphique pour ce locus. La génétique des populations s'intéresse aux différentes forces qui font varier la fréquence de ces allèles au sein d'une population. Ces forces, dites évolutives, sont regroupées en quatre groupes : les

mutations, la sélection naturelle, les migrations et la dérive génétique.

Les mutations désignent les erreurs commises lors de la réplication de l'ADN qui sont transmises à la descendance d'un individu. On estime qu'à chaque génération, une base se substitue à une autre à une position donnée environ une fois sur 10^8 [NC00 ; Roa+10]. Ces mutations modifient les fréquences alléliques soit parce qu'elles créent de nouveaux allèles soit parce qu'elles produisent des individus portant un allèle déjà présent dans la population.

La sélection naturelle agit lorsqu'un allèle donné permet aux individus qui le portent de produire un plus grand nombre de descendants que les autres. En effet l'ADN joue un rôle prépondérant dans le fonctionnement d'un organisme et sert notamment à produire les différentes protéines qui régulent son activité. Certaines mutations peuvent ainsi avoir des effets drastiques sur la survie ou la reproduction des individus qui les portent. Les allèles correspondants disparaissent alors rapidement de la population. À l'inverse certaines mutations confèrent un avantage reproductif dans un environnement donné et leur fréquence dans la population augmente alors mécaniquement.

L'échange régulier d'individus entre plusieurs populations a pour effet d'harmoniser les fréquences alléliques entre ces populations. Une mutation avantagée par la sélection naturelle peut alors se répandre de population en population. Au contraire, l'isolement géographique permet à certaines populations de diverger en acquérant des mutations endémiques. De manière plus générale, on peut regrouper avec les migrations tous les effets de la structure interne d'une population, spatiale ou non.

La dérive génétique désigne les fluctuations aléatoires des fréquences alléliques dues à l'échantillonnage entre chaque génération. En effet si une population contient très peu d'individus, la mort ou la reproduction de l'un d'entre eux a un effet significatif sur les fréquences alléliques. Si en revanche de nombreux individus meurent et se reproduisent en même temps, les fréquences alléliques varieront moins. La dérive génétique est notamment responsable de la disparition de certains allèles dans les populations de faible effectif, même en l'absence de sélection naturelle.

Si aucune de ces quatre forces n'agit à un locus donné, alors les fréquences des différents allèles à ce locus resteront stables au cours du temps.

Parmi toutes les mutations possibles, certaines sont létales ou très délétères et sont rapidement purgées par la sélection naturelle. Un petit nombre est avantagé dans leur environnement et se fixe rapidement dans la population. Les autres mutations sont soit neutres soit si peu sélectionnées que leur fréquence dans la population évolue principalement sous l'effet de la dérive génétique. Toutes finiront soit par disparaître soit par se fixer, mais après un temps potentiellement très long. La majorité des polymorphismes génétiques observés à un instant donné est donc neutre ou quasi neutre.

Cette observation a été faite pour la première fois en 1968 par Kimura [Kim68] qui l'a ensuite étayée dans son ouvrage *The neutral theory of molecular evolution* [Kim84]. Cette théorie ne contredit en rien la vision Darwinienne de l'évolution selon laquelle les traits phénotypiques évoluent par accumulation progressive de mutations avantagées par la sélection naturelle. En effet un grand nombre de mutations sont sans conséquence sur la séquence d'acides aminés produite par la transcription de la séquence d'ADN (mutations dites synonymes) ou bien ne changent pas la fonction de la protéine correspondante (mutations dites silencieuses). De plus, une grande partie du génome ne code pas directement pour une séquence d'acides aminés (ce qui ne veut pas dire qu'elle ne joue aucun rôle dans l'expression des gènes). On observe d'ailleurs un plus grand nombre de mutations synonymes et silencieuses que de mutations affectant la fonction des protéines et les zones dites non codantes du

génome sont plus souvent polymorphiques que les autres. Ceci est attendu dans la théorie neutraliste de Kimura car les premières ont moins de chances d'être délétères et d'être purgées par la sélection naturelle.

En conséquence, on peut considérer que la plupart des mutations se fixent uniquement par hasard, sans affecter la démographie de la population. Elles laissent donc des traces qui nous renseignent sur la généalogie des individus. Cette diversité neutre peut alors être utilisée pour reconstruire l'histoire des populations. Cette tâche est loin d'être aisée, mais les données génétiques affluent avec le progrès des techniques de séquençage. Afin d'exploiter ces données, de nouveaux outils mathématiques sont nécessaires en complément de ceux développés depuis les années 1920 [Dur08 ; Kim64].

Cette thèse porte sur la structure géographique de la diversité génétique. Entre sa naissance et le moment où il se reproduit, un individu parcourt en moyenne un certain nombre de kilomètres. Comme il en va de même pour ses ancêtres et les ancêtres de ses ancêtres, deux individus sont d'autant plus éloignés génétiquement qu'ils le sont géographiquement. Cela dépend également de la densité de la population : plus celle-ci est élevée, moins deux individus, même proches, ont de chances de partager des ancêtres communs récents.

Dans un premier temps, nous étudions l'effet de la selection naturelle sur des populations structurées en espace. Nous considérons un régime pour lequel les fluctuations des fréquences alléliques sont faibles et nous montrons que celles-ci sont asymptotiquement égales à la solution d'une équation aux dérivées partielles stochastique.

Nous nous intéressons ensuite aux effets des hétérogénéités spatiales sur la distribution des allèles au sein d'une population. Deux situations nous intéressent particulièrement : le cas d'une dispersion hétérogène et celui d'une barrière au flux de gènes. Dans le cas d'une dispersion hétérogène, les individus se déplacent sur de plus grandes distances dans une région de l'espace que dans l'autre. Une barrière au flux de gènes est un obstacle physique qui réduit les échanges génétiques entre deux parties de l'espace. Dans les deux cas, nous étudions les généalogies des individus avec pour objectif la détection de ces hétérogénéités à partir de données génétiques. Ce problème d'inférence est d'ailleurs abordé à la fin de ce manuscrit dans le cas de la dispersion hétérogène.

Les travaux présentés ici apportent plusieurs résultats nouveaux qui éclairent la structure spatiale de la diversité génétique. Le Chapitre 2 traite des fluctuations dans les fréquences alléliques d'une population dans un espace homogène en présence de sélection naturelle. Une application au problème du fardeau de dérive (voir plus bas) y est également donnée. Ce chapitre est le fruit d'une collaboration avec Sarah Penington, de l'Université d'Oxford, et a fait l'objet d'une publication dans *Electronic Journal of Probability* [FP17]. Les Chapitres 3 et 4 traitent respectivement des généalogies en présence de dispersion hétérogène et d'une barrière au flux de gènes. Enfin, dans le Chapitre 5, nous présentons un travail en cours avec Harald Ringbauer (Institute of Science and Technology, Austria) et Graham Coop (UC Davis, California) sur une méthode d'inférence en environnement hétérogène.

Dans la section suivante, nous présentons quelques modèles élémentaires en génétique des populations, qui serviront d'exemples par la suite. Les modèles prenant en compte la structure spatiale des populations sont introduits dans la Section 1.3. C'est sur ces derniers que portent les travaux présentés dans cette thèse. Nous présentons ensuite dans la Section 1.4 les résultats sur les fluctuations des fréquences alléliques dans des populations structurées en espace (qui sont démontrés dans le Chapitre 2). La Section 1.5 introduit le problème de la dispersion hétérogène ainsi que les traitements

existants dans la littérature. Nous nous tournons ensuite vers les travaux concernant l'impact des barrières au flux de gènes et nous exposons les résultats correspondants dans la Section 1.6. Enfin, dans la Section 1.7, nous présentons une méthode d'inférence démographique développée par Ringbauer et al. [RCB16] qui sera adaptée à un environnement hétérogène dans le Chapitre 5.

1.2 Modèles mathématiques en génétique des populations

Les premiers modèles mathématiques décrivant l'évolution de la composition génétique d'une population ont été introduits dans les travaux de Sewall Wright, Ronald A. Fisher et John B. S. Haldane dans les années 1920-1930. Ces modèles ont ouvert la voie à tout un domaine des mathématiques et de la génétique à travers de nombreux raffinements et généralisations. Nous présentons ici quelques modèles classiques en préparation des modèles structurés en espace présentés dans la Section 1.3.

De manière générale, deux points de vue complémentaires coexistent en génétique des populations. Le premier consiste à décrire l'évolution de la composition génétique d'une population dans le sens normal du temps, génération après génération. Nous présentons deux modèles dans ce cadre plus bas : le modèle de Wright-Fisher et le modèle de Moran. La seconde classe de modèles décrit la généalogie dans le passé d'un échantillon aléatoire (uniforme) au sein de la population. Nous introduisons l'un d'eux dans la sous-section 1.2.4 : le coalescent de Kingman, et nous mettons en évidence ce qui le lie aux modèles précédents.

Tous ces modèles décrivent des situations très idéalisées, ce qui les rend attirants pour le mathématicien, mais qui peut poser problème au scientifique qui souhaite les utiliser pour comprendre un système biologique. Le but du généticien devient alors d'identifier les quelques paramètres qui résument le comportement d'une population et de chercher à estimer ces paramètres à partir d'un échantillon de la population en question. Cette approche sera présentée plus en détail dans la Section 1.7 et dans le Chapitre 5.

1.2.1 Le modèle de Wright-Fisher

Le modèle de Wright-Fisher décrit l'évolution d'une population haploïde (chaque individu ne porte qu'une seule copie de chaque gène), panmictique (la population n'est pas structurée) et qui évolue par générations discrètes (tous les individus meurent et se reproduisent en même temps).

Définition 1.2.1 (Modèle de Wright-Fisher neutre)**.** *On considère une population de N individus haploïdes. À chaque génération, une nouvelle population de N individus remplace la précédente et chaque individu de la nouvelle génération "choisit" son parent uniforémément au hasard au sein de la génération précédente, indépendement des autres individus.*

Si k individus portent un allèle donné à la génération n, le nombre d'individus portant ce même allèle à la génération suivante suit une loi binomiale de paramètres N et k/N. On note $X_n = k/N$ la proportion d'individus de ce type dans la population à la génération n. La suite $(X_n, n \geq 0)$ est alors une chaîne de Markov à espace d'état fini dont les états 0 et 1 sont absorbants.

Le modèle ci-dessus est dit neutre car le nombre moyen d'enfants d'un individu ne dépend pas de l'allèle qu'il porte. Si la sélection naturelle confère un avantage reproductif à l'un des types par

rapport aux autres, le modèle neutre ne suffit plus pour décrire l'évolution de la population. On suppose dans la suite que la population est composée d'individus de deux types, 0 et 1, et que la sélection naturelle avantage l'un de ces deux types par rapport à l'autre (un individu du type avantagé aura en moyenne plus d'enfants qu'un individu de l'autre type).

Définition 1.2.2 (Modèle de Wright-Fisher avec sélection)**.** *On considère une population de N individus haploïdes qui peuvent être de type 0 ou 1. À chaque génération, N nouveaux individus remplacent les précédents. S'il y a k individus de type 1 au sein de la génération n, alors chaque individu de la génération $n+1$ choisit un parent parmi ceux de la génération n de telle manière qu'un individu de type 1 est choisi avec probabilité $\frac{1+s}{N+sk}$ et un individu de type 0 est choisi avec probabilité $\frac{1}{N+sk}$.*

On note X_n la proportion d'individus de type 1 au sein de la population à la génération n. Sachant que $X_n = x$, NX_{n+1} suit une loi binomiale de paramètres N et

$$\frac{(1+s)x}{1+sx}.$$

On remarque que le rapport entre le nombre moyen d'enfants d'un individu de type 1 et celui d'un individu de type 0 est $1+s$, on dit alors que les types 0 et 1 ont une fitness relative de 1 et $1+s$.

Le modèle de Wright-Fisher a été largement étudié, et nous ne nous attardons pas plus dessus. Une présentation plus détaillée de ses propriétés est faite dans [Eth11].

1.2.2 Le modèle de Moran

Le modèle de Moran est un analogue du modèle de Wright-Fisher dans le cas où les individus ne se reproduisent pas de manière synchrone, mais meurent et naissent à des temps aléatoires.

Définition 1.2.3 (Modèle de Moran neutre)**.** *On considère une population de N individus haploides. Soit $(P_t)_{t\geq 0}$ un processus de Poisson de paramètre $\binom{N}{2}$. À chaque saut de P, on choisit deux individus uniformément au hasard (sans remise) dans la population. L'un d'eux meurt et l'autre se reproduit, chacun avec égale probabilité.*

Si l'on note X_t la proportion d'individus portant un allèle donné dans la population au temps t, alors $(X_t, t \geq 0)$ est un processus de Markov à temps continu, qui saute de x à $x + \frac{1}{N}$ et à $x - \frac{1}{N}$ à taux

$$\binom{N}{2} x(1-x).$$

La Figure 1.1 présente deux trajectoires de X_t pour deux tailles de populations différentes.

Remarque. *Dans le modèle de Moran, il faut en moyenne N reproductions pour renouveler toute la population. Puisque ces événements surviennent à taux $\binom{N}{2}$, une unité de temps dans le modèle de Moran correspond à environ N générations dans le modèle de Wright-Fisher.*

Pour inclure la sélection naturelle dans ce modèle, on introduit une autre suite d'événements de reproduction qui favorisent les individus du type avantagé.

Définition 1.2.4 (Modèle de Moran avec sélection)**.** *On considère une population de N individus qui peuvent être de type 0 ou 1. On suppose que l'allèle 1 confère un avantage sélectif à ceux qui le*

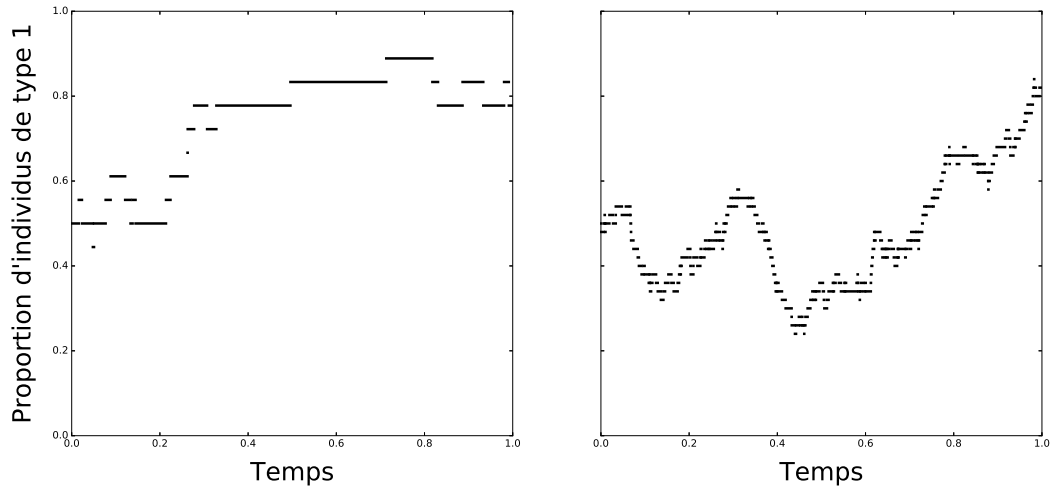

Figure 1.1 : Fréquences alléliques dans le modèle de Moran
Trajectoire de $(X_t)_{t\geq 0}$ dans le modèle de Moran neutre pour $N = 18$ (à gauche) et $N = 50$ (à droite).

portent. Des événements de reproduction neutres se produisent à taux $\binom{N}{2}$ *comme dans le modèle défini ci-dessus. De plus, des reproductions sélectives se produisent à taux* $s\binom{N}{2}$*, pour* $s > 0$*. Lors d'une reproduction sélective, on choisit deux individus au hasard dans la population. Si les deux sont du même type, rien ne se passe, sinon, celui de type 0 meurt et celui de type 1 se reproduit.*

En notant toujours X_t la proportion d'individus de type 1 dans la population à l'instant t, $(X_t, t \geq 0)$ est un processus de Markov de générateur infinitésimal

$$\mathcal{L}f(x) = \binom{N}{2} x(1-x)\left(f\left(x+\tfrac{1}{N}\right) + f\left(x-\tfrac{1}{N}\right) - 2f(x)\right) + s\binom{N}{2} 2x(1-x)\left(f\left(x+\tfrac{1}{N}\right) - f(x)\right). \quad (1.1)$$

1.2.3 La diffusion de Wright-Fisher

Lorsque la taille de la population que l'on considère est assez grande, les proportions des différents types au sein de celle-ci (autrement dit les fréquences alléliques) changent peu d'une génération à l'autre, en raison de la plus faible variance dans l'échantillonnage des parents. Sur une échelle de temps appropriée, il est alors possible de voir les fluctuations des fréquences alléliques comme un processus de diffusion : la diffusion de Wright-Fisher.

Dans le modèle de Wright-Fisher avec sélection, sachant que

$$X_{n+1} \sim \frac{1}{N} Bin\left(N, \frac{(1+s)X_n}{1+sX_n}\right),$$

on montre facilement que

$$\mathbb{E}\left[X_{n+1} - X_n \mid X_n = x\right] = sx(1-x) + o(s),$$
$$\mathbb{E}\left[(X_{n+1} - X_n)^2 \mid X_n = x\right] = \frac{1}{N}x(1-x) + o(s).$$

Après avoir vérifié plusieurs hypothèses supplémentaires, (voir [Eth11, Théoreme 3.6]), on obtient le résultat suivant, mentionné notamment par Feller [Fel51].

Proposition 1.2.5 (Approximation du modèle de Wright-Fisher par une diffusion)**.** *Soit* $(s_N, N \geq 1)$ *une suite de réels telle que*

$$Ns_N \underset{N\to\infty}{\longrightarrow} \bar{s}.$$

Pour $N \geq 1$*, soit* $(X_n^N, n \geq 0)$ *la suite des fréquences du type 1 dans le modèle de Wright-Fisher de taille* N *avec sélection, de coefficient* s_N *(Définition 1.2.2). On suppose que*

$$X_0^N \underset{N\to\infty}{\longrightarrow} \bar{x}_0$$

presque sûrement. Pour $t \geq 0$*, on pose*

$$X^N(t) = X^N_{\lfloor Nt \rfloor}. \tag{1.2}$$

Alors, lorsque $N \to \infty$*, la suite de processus* $(X^N, N \geq 1)$ *converge en loi dans* $\mathrm{D}([0,T], \mathbb{R})$ *pour tout* $T > 0$ *fixé vers la solution de l'équation différentielle stochastique*

$$\begin{cases} dx_t = \bar{s}x_t(1-x_t)dt + \sqrt{x_t(1-x_t)}dB_t, \\ x_0 = \bar{x}_0 \end{cases} \tag{1.3}$$

où B *est un mouvement brownien standard.*

En revenant à l'échelle temporelle originale du modèle de Wright-Fisher, cela revient à approcher $(X_n, n \geq 0)$ par la solution de

$$\begin{cases} dx_t = sx_t(1-x_t)dt + \sqrt{\frac{1}{N}x_t(1-x_t)}dB_t \\ x_0 = X_0. \end{cases} \tag{1.4}$$

Le processus $(x_t, t \geq 0)$ est appelé diffusion de Wright-Fisher, et permet d'étudier de nombreux aspects du comportement à grande échelle à la fois du modèle de Wright-Fisher, mais également du modèle de Moran. Rappelons qu'une unité de temps dans le modèle de Moran correspond à N générations du modèle de Wright-Fisher, il n'y a donc pas besoin de changer d'échelle temporelle pour observer la convergence de ce dernier vers la diffusion de Wright-Fisher.

Proposition 1.2.6 (Approximation du modèle de Moran par une diffusion)**.** *Soit* $(s_N, N \geq 1)$ *une suite de réels positifs telle que*

$$Ns_N \underset{N\to\infty}{\longrightarrow} \bar{s}.$$

Pour $N \geq 1$*, soit* $(X_t^N, t \geq 0)$ *le processus décrivant la fréquence du type 1 dans le modèle de Moran*

de taille de population N avec sélection de coefficient s_N (Définition 1.2.4). On suppose que

$$X_0^N \underset{N\to\infty}{\longrightarrow} \bar{x}_0$$

presque sûrement. Alors, lorsque $N \to \infty$, la suite $(X^N, N \geq 1)$ converge en loi dans $\mathrm{D}([0,T],\mathbb{R})$ pour tout $T > 0$ fixé vers la solution de (1.3).

Pour montrer ce résultat, il suffit de montrer que la suite $(X^N, N \geq 1)$ est tendue dans $\mathrm{D}([0,T],\mathbb{R})$ et que la suite des générateurs infinitésimaux converge lorsque $N \to \infty$. Or si f est deux fois continuement dérivable sur $[0,1]$, (1.1) devient

$$\mathcal{L}^N f(x) = \frac{1}{2}x(1-x)f''(x) + \bar{s}x(1-x)f'(x) + \mathcal{O}\left(\frac{1}{N}\right).$$

On reconnaît à droite le générateur de (1.3). La convergence du modèle de Moran vers la diffusion de Wright-Fisher est illustrée par la Figure 1.2.

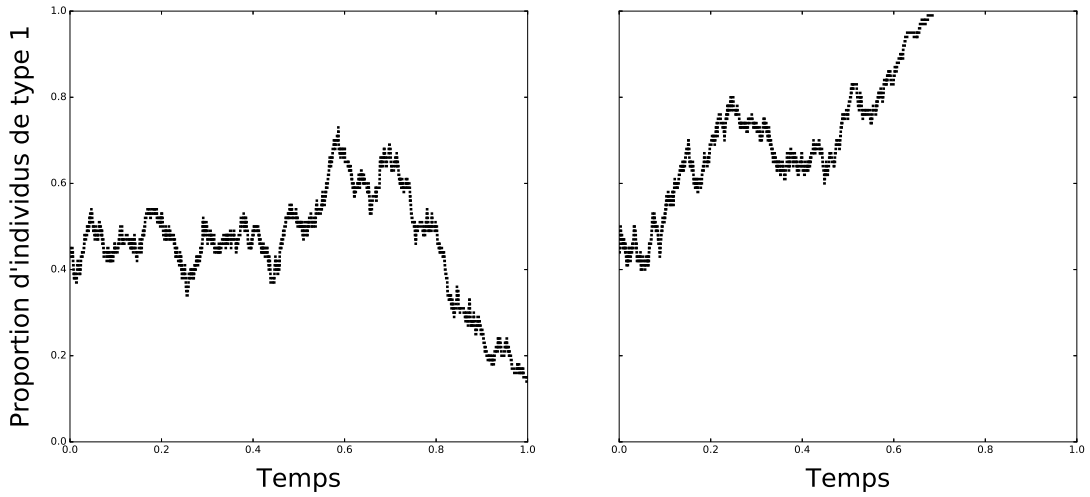

Figure 1.2 : Modèle de Moran en grande population
Exemple de trajectoire suivie par $(X_t, t \geq 0)$ dans le modèle de Moran avec N=100. À gauche le modèle est neutre ($s = 0$) et à droite on a pris $s = 0.05$.

1.2.4 Le coalescent de Kingman

Tournons nous à présent vers un troisième modèle, introduit par Kingman [Kin82], qui permet de décrire la composition génétique d'une population à travers les liens de parenté entre ses individus. Le coalescent de Kingman est un processus de Markov à valeurs dans l'espace des partitions de $\{0,\ldots,N\}$. Il décrit la généalogie d'un échantillon de N individus au sein de la population : deux individus sont dans le même bloc de la partition à l'instant t s'ils ont un ancêtre commun t unités de temps dans le passé.

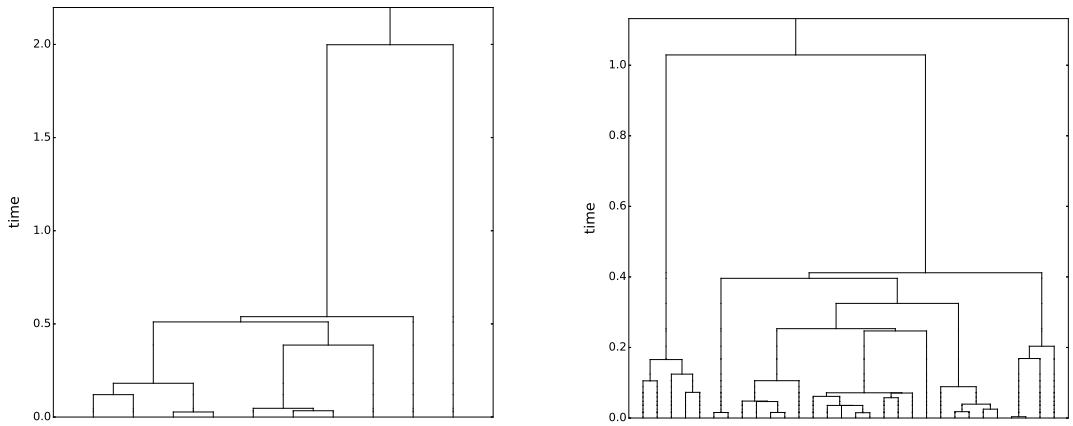

FIGURE 1.3 : Représentation du Coalescent de Kingman
Représentation du coalescent de Kingman de taille initiale 10 (à gauche) et 30 (à droite) sous la forme d'un arbre plan. On remarque que les premières colaescence sont très rapides et que la dernière coalescence prend au moins autant de temps que toutes les autres réunies.

Définition 1.2.7 (Coalescent de Kingman)**.** *Soit $N \geq 1$ et soit $(\Pi_t)_{t\geq 0}$ un processus de Markov à valeurs dans l'espace des partitions de $\{0, \dots, N\}$. On dit que Π est un coalescent de Kingman de taille initiale N si*

i) $\Pi_0 = \{\{0\}, \dots, \{N\}\}$,

ii) lorsque Π_t contient k blocs, après un temps aléatoire exponentiel de paramètre $\binom{k}{2}$, deux blocs choisis uniformément au hasard dans Π_t fusionnent.

De manière équivalente, on peut dire que chaque paire de blocs dans Π_t fusionne (ou coalesce) à taux 1. On peut représenter la trajectoire de $(\Pi_t, t \geq 0)$ sous la forme d'un arbre, comme illustré dans la Figure 1.3.

Il existe un lien profond entre le coalescent de Kingman et la diffusion de Wright-Fisher présentée plus haut. Nous présentons ici une facette de cette connexion sous la forme de la dualité de moments.

Définition 1.2.8 (Dualité de moments)**.** *Soit $(X_t)_{t\geq 0}$ un processus de Markov défini sur un espace $(\Omega, \mathcal{F}, \overrightarrow{\mathbb{P}})$ à valeurs dans un espace E_1 et soit $(Y_t)_{t\geq 0}$ un autre processus de Markov défini sur un espace $(\Omega', \mathcal{F}', \overleftarrow{\mathbb{P}})$ à valeurs dans E_2. Soit $h : E_1 \times E_2 \to \mathbb{R}$ une fonction mesurable. Si pour tout $t \geq 0$, $(x, y) \in E_1 \times E_2$,*

$$\overrightarrow{\mathbb{E}}_x \left[h(X_t, y)\right] = \overleftarrow{\mathbb{E}}_y \left[h(x, Y_t)\right], \tag{1.5}$$

alors on dit que X et Y sont duaux par rapport à h.

Il convient d'interpréter X comme un processus allant dans le sens normal du temps et Y comme un processus décrivant une évolution depuis un état présent vers un état passé, comme nous le verrons plus bas dans le cas du coalescent de Kingman. Une définition plus complète de la dualité pour les processus de Markov peut se trouver au Chapitre 4.4 de [EK86].

Il existe une caractérisation pratique de la dualité de moments à l'aide des générateurs infinitésimaux.

Proposition 1.2.9. *Soient $\overrightarrow{A}$ et $\overleftarrow{A}$ les générateurs infinitésimaux respectifs de X et de Y. On suppose que $h(\cdot, y)$ est dans le domaine de $\overrightarrow{A}$ pour tout $y \in E_2$ et inversement $h(x, \cdot)$ est dans le domaine de $\overleftarrow{A}$ pour tout $x \in E_1$. Alors X et Y sont duaux par rapport à h si et seulement si pour tout $(x, y) \in E_1 \times E_2$,*

$$\overrightarrow{A} h(x, y) = \overleftarrow{A} h(x, y). \tag{1.6}$$

Éléments de preuve. En supposant (1.5), on a

$$\begin{aligned}
\overrightarrow{A} h(x, y) &= \lim_{\delta t \downarrow 0} \frac{1}{\delta t} \overrightarrow{\mathbb{E}}_x \left[h(X_{\delta t}, y) - h(x, y) \right] \\
&= \lim_{\delta t \downarrow 0} \frac{1}{\delta t} \overleftarrow{\mathbb{E}}_y \left[h(x, Y_{\delta t}) - h(x, y) \right] \\
&= \overleftarrow{A} h(x, y),
\end{aligned}$$

Pour la réciproque, on suppose (1.6) et on construit un espace de probabilités

$$(\Omega^\times, \mathcal{F}^\times, \mathbb{P}) = (\Omega \times \Omega', \mathcal{F} \otimes \mathcal{F}', \overrightarrow{\mathbb{P}} \otimes \overleftarrow{\mathbb{P}})$$

de telle manière que sous $\mathbb{P}$, X et Y soient indépendants. On montre alors

$$\begin{aligned}
\frac{d}{ds} \mathbb{E} \left[h(X_s, Y_{t-s}) \right] &= \mathbb{E} \left[\overrightarrow{A} h(X_s, Y_{t-s}) - \overleftarrow{A} h(X_s, Y_{t-s}) \right] \\
&= 0.
\end{aligned}$$

On obtient (1.5) en intégrant entre 0 et t. □

Le coalescent de Kingman est lié par une relation de dualité à la diffusion de Wright-Fisher dans le cas neutre ($s = 0$). Cette dualité s'exprime de la manière suivante.

Proposition 1.2.10 (Dualité avec la diffusion de Wright-Fisher)**.** *Soit $(\Pi_t)_{t \geq 0}$ un coalescent de Kingman (sa taille initale sera précisée plus bas), et soit $(x_t, t \geq 0)$ la solution de l'équation différentielle stochastique*

$$dx_t = \sqrt{x_t(1 - x_t)} dB_t$$

où B est un mouvement brownien standard. Pour $t \geq 0$, on note N_t le nombre de blocs dans Π_t. Pour $n \geq 1$ et $x \in [0, 1]$, on a

$$\overrightarrow{\mathbb{E}}_x \left[x_t^n \right] = \overleftarrow{\mathbb{E}}_n \left[x^{N_t} \right], \tag{1.7}$$

où $\overrightarrow{\mathbb{E}}$ désigne l'espérance par rapport à la loi de $(x_t)_{t \geq 0}$ et $\overleftarrow{\mathbb{E}}$ celle par rapport à la loi de Π.

La relation (1.7) s'interprète aisément : si l'on échantillonne n individus dans une population au temps t, la probabilité qu'ils soient tous de type 1 est la probabilité que tous leurs ancêtres au temps 0 soient de type 1.

Démonstration. On pose $h(x,n) = x^n$. Le générateur de $(x_t)_{t\geq 0}$ agit sur h de la manière suivante

$$\begin{aligned}\overrightarrow{A}h(x,n) &= \frac{1}{2}x(1-x)\frac{\partial^2 h}{\partial x^2}(x,n)\\ &= \frac{1}{2}(1-x)n(n-1)x^{n-1}\\ &= \binom{n}{2}\left(h(x,n-1)-h(x,n)\right)\\ &= \overleftarrow{A}h(x,n).\end{aligned}$$

La proposition 1.2.10 découle alors de la proposition 1.2.9. □

Rappelons que dans la diffusion de Wright-Fisher, N générations s'écoulent à chaque unité de temps, où N est la taille de la population (voir (1.2)). En effet, plus une population est nombreuse, moins il y a de chances que deux individus pris au hasard aient un ancêtre commun récent. À l'inverse, une population de faible effectif contient de nombreux individus d'une même famille, et les coalescences sont donc plus rapides. Si l'on est capable d'estimer l'âge des ancêtres communs les plus récents pour différentes paires d'individus au sein d'une population, on peut alors estimer le rythme des coalescences et donc la taille de la population concernée.

Cette taille de population estimée est appelée la taille de la population efficace et correspond au nombre moyen d'individus qui se reproduisent à chaque génération. Celle-ci peut différer grandement de l'effectif total de la population pour diverses raisons (accès limité aux partenaires par exemple).

Cette taille peut également varier dans le temps, comme c'est le cas pour les populations humaines. La distribution des coalescences dans le passé permet alors d'estimer la taille d'une population à différentes périodes de son histoire (voir par exemple [LD11]).

Tous les modèles présentés dans cette section ont en commun le fait d'ignorer la structure spatiale des populations étudiées. Cette hypothèse n'est pas toujours raisonnable lorsque l'on cherche à étudier des données de terrain. Il est donc nécessaire d'introduire une autre famille de modèles qui prend en compte l'éloignement géographique comme source d'éloignement génétique.

1.3 Isolation par la distance, introduction de la structure spatiale

Lorsqu'une population occupe un espace étendu, ou bien de nombreux sites distants les uns des autres, les individus n'explorent qu'une petite portion de l'aire totale de répartition de la population entre leur naissance et le moment où ils se reproduisent. Les ancêtres d'un individu donné se répartissent donc dans une zone géographique d'autant plus restreinte qu'ils sont récents. De même, deux individus sont d'autant plus suceptibles de partager un ancêtre récent qu'ils sont proches géographiquement [Agu+17].

Afin de quantifier cet effet et de l'utiliser pour étudier l'histoire des populations biologiques, il importe de définir un modèle mathématique pour des populations structurées spatialement. Nous présentons ici deux classes de modèles que l'on retrouvera dans la suite de ce manuscrit : des modèles en espace discret, dits *stepping stone* ("pas japonais" en français) et un modèle en espace continu, le processus Λ-Fleming Viot spatial.

Tous ces modèles ont en commun de décrire l'isolation génétique résultant de la séparation géographique entre des sous-populations. Nous verrons dans la sous-section 1.3.2 comment les corrélations entre les fréquences alléliques à différents sites décroissent avec la distance géographique dans le modèle stepping stone. Nous retrouverons une approche similaire dans la Section 1.7 avec le processus Λ-Fleming Viot spatial. Nous présenterons également dans la sous-section 1.3.5 le comportement à grande échelle de ce dernier dans différents régimes.

De même que pour les modèles sans structure spatiale présentés dans la Section 1.2, les modèles spatiaux peuvent être décrits de deux façons : soit en précisant comment les fréquences alléliques évoluent au cours du temps, soit en étudiant la généalogie d'un échantillon aléatoire d'individus au sein de la population. Nous présentons ces deux approches pour chaque modèle spatial ci-dessous.

1.3.1 Populations en espace discret, le modèle stepping stone

Introduit pour la première fois dans les travaux de Wright [Wri43] puis repris et détaillé par Kimura [Kim53], le modèle stepping stone décrit l'évolution d'une population divisée en colonies (ou dèmes) situées aux sommets d'un graphe (typiquement $\mathbb{Z}^d$ avec $d = 1$ ou 2). Chaque colonie suit un modèle de Wright-Fisher à la différence près qu'une partie des individus de chaque dème descend d'un parent venant d'un autre dème. Pour reproduire l'effet d'isolation par la distance, ces migrants doivent provenir d'une région limitée autour de leur colonie d'immigration. C'est le cas par exemple si la migration ne se fait qu'entre plus proches voisins dans $\mathbb{Z}^d$.

Soit $G = \mathbb{Z}^d$, $d \in \{1, 2, 3\}$ et soit $(N_i, i \in G)$ une famille d'entiers non nuls. Soit $M = (m_{ij}, i, j \in G)$ une famille de réels positifs telle que, pour tout $i \in G$,

$$\sum_{j \in G} N_j m_{ji} = N_i \tag{1.8}$$

Le modèle stepping stone sur $\mathbb{Z}^d$ de matrice de migration M est défini de la manière suivante.

Définition 1.3.1 (Modèle stepping stone à générations discrètes)**.** *On considère une population répartie sur $\mathbb{Z}^d$ pour $d \in \{1, 2, 3\}$. Le dème $i \in \mathbb{Z}^d$ contient N_i individus. À chaque génération, tous les individus meurent et sont remplacés par de nouveaux individus. Un nouvel individu dans le dème i choisit un parent uniformément au hasard dans le dème j avec probabilité*

$$\tilde{m}_{ij} := \frac{N_j}{N_i} m_{ji}.$$

On note que dans ce modèle, chaque individu du dème j produit en moyenne m_{ji} enfants dans le dème i. L'équation (1.8) s'interprète alors comme une équation de conservation du nombre d'individus de chaque dème.

Pour $t \geq 0$ et $i \in G$, soit $p(t, i)$ la proportion d'individus portant un allèle donné dans le dème i au temps t. Si les N_i sont assez grand, il est naturel de chercher à approcher chaque $(p(t, i), t \geq 0)$ par une diffusion analogue à (1.4). Nous formalisons cette approche, due à Kimura [Kim53 ; Kim64] ci-dessous. Notons que celle-ci est valable lorsque la migration à chaque génération est faible, c'est-à-dire pour $\tilde{m}_{ii}$ suffisamment proche de 1 pour tout i.

Définition 1.3.2 (Modèle stepping stone de Kimura)**.** *Soient $d \in \{1,2,3\}$, $G = \mathbb{Z}^d$ et $(N_i, i \in G)$, $(m_{ij}, i,j \in G)$ deux familles de réels positifs vérifiant* (1.8)*. Fixons $p_0 : G \to [0,1]$. Soit alors $(p(t,i), t \geq 0, i \in G)$ la solution du système d'équations différentielles stochastiques suivant*

$$\begin{cases} dp(t,i) = \sum_{j \in G} \tilde{m}_{ij}(p(t,j) - p(t,i))dt + \sqrt{\frac{1}{N_i} p(t,i)(1-p(t,i))} dB_t^i, \\ p(0,i) = p_0(i) \end{cases} \tag{1.9}$$

où $\{B^i, i \in G\}$ est une famille de mouvements browniens standards indépendants.

Le modèle ci-dessus peut être vu comme une généralisation de la diffusion de Wright-Fisher à une population vivant dans un espace discret. De même que pour la diffusion de Wright-Fisher, on peut établir une relation de dualité entre le modèle de Kimura et un coalescent décrivant la généalogie d'un échantillon d'individus. Cette généalogie prend la forme d'un système de marches aléatoires coalescentes décrivant les trajectoires des lignées ancestrales de l'échantillon.

Définition 1.3.3 (Coalescent structuré)**.** *Soient $n \geq 1$ et $x_1, \ldots, x_n \in \mathbb{Z}^d$. Soit $(\Pi_t, t \geq 0)$ un système de particules,*

$$\Pi_t = \{\xi_t^1, \ldots, \xi_t^{N_t}\}, \quad \xi_t^i \in \mathbb{Z}^d, \quad N_t \geq 1,$$

tel que chaque particule en $i \in \mathbb{Z}^d$ migre vers le dème j à taux $\tilde{m}_{ij}$, indépendemment des autres particules. De plus, deux particules situées dans le même dème i coalescent après un temps exponentiel de paramètre $\frac{1}{N_i}$ (sauf si l'une d'elles migre avant).

On établit alors une relation de dualité entre ces deux modèles de la même manière que précédemment. Étant donnés une condition initiale $p_0(\cdot)$ et un ensemble de positions d'échantillonnage $\{x_1, \ldots, x_n\}$, on a [Eth11]

$$\mathbb{E}_{p_0}\left[\prod_{i=1}^{n} p(t, x_i)\right] = \mathbb{E}_{\{x_1,\ldots,x_n\}}\left[\prod_{i=1}^{N_t} p_0(\xi_t^i)\right]. \tag{1.10}$$

On peut donc suivre la trajectoire décrite par l'ensemble des positions des ancêtres d'un individu dans la population, on parlera alors de lignée ancestrale et on dira que deux lignées coalescent lorsque les individus correspondant ont un ancêtre commun dans le passé.

1.3.2 La décroissance des corrélations avec la distance

D'après la construction du coalescent structuré, il est facile de voir que plus deux individus sont éloignés géographiquement, moins ils ont de chances d'avoir un ancêtre commun récent. Deux populations sont donc d'autant plus éloignées génétiquement qu'elles le sont physiquement. Ce phénomène dit d'"isolation par la distance" a été illustré d'abord par Wright [Wri43], puis par Malécot [Mal48] et par Kimura et Weiss [KW64].

Ces derniers considèrent un modèle stepping stone analogue à celui présenté dans la définition 1.3.1, dans lequel tous les dèmes sont occupés par le même nombre d'individus et où la migration ne se fait qu'entre plus proches voisins. Ils supposent également qu'à chaque génération, une partie des individus migrent depuis une population source à l'équilibre. Si $p(t,i)$ désigne la proportion d'individus portant

un allèle donné au temps t dans la population $i \in \mathbb{Z}^d$ et si $\bar{p}$ est la proportion d'individus de ce type dans la population source, alors p obéit à l'équation suivante

$$dp(t,i) = \frac{m}{2}\sum_{j\sim i}(p(t,j)-p(t,i))dt + \mu(\bar{p}-p(t,i))dt + \sqrt{\frac{1}{N}p(t,i)(1-p(t,i))}dB_t^i, \tag{1.11}$$

où $N, m, \mu > 0$ et $\sum_{i\sim j}$ désigne la somme sur les plus proches voisins de i. (Kimura et Weiss considèrent en réalité un modèle à temps discret, mais les mêmes méthodes s'appliquent à l'équation ci-dessus.) On note $r(k)$ la corrélation entre les fréquences alléliques en deux points séparés par $k \in \mathbb{Z}^d$, donnée par

$$r(k) = \frac{\mathbb{E}\left[(p(t,i)-\bar{p})(p(t,i+k)-\bar{p})\right]}{\mathbb{E}\left[(p(t,i)-\bar{p})^2\right]}.$$

Kimura et Weiss montrent que, à l'équilibre,

$$\begin{aligned} &\text{si } d=1, && r(k) \simeq \exp\left(-\sqrt{\frac{2\mu}{m}}\,|k|\right), \\ &\text{si } d=2, && r(k) \simeq \frac{C_0}{\pi m}K_0\left(\sqrt{\frac{4\mu}{m}}\,|k|\right) \end{aligned} \tag{1.12}$$

où $|k| = \sqrt{k_1^2+k_2^2}$ est la norme euclidienne de k, C_0 une constante positive et K_0 désigne la fonction de Bessel modifiée de degré 0 [AS64]. Il est à noter que lorsque $|k|$ est grand devant $\sqrt{m}$, en dimension 2, $r(k)$ est proportionnel à

$$\frac{1}{\sqrt{|k|}}\exp\left(-\sqrt{\frac{4\mu}{m}}\,|k|\right).$$

Les corrélations décroissent donc plus rapidement avec la distance en dimension 2 qu'en dimension 1. Cela est lié au fait qu'en grande dimension, il est plus facile pour deux particules (ou lignées ancestrales) dans le coalescent structuré de s'éviter. Sawyer [Saw76 ; Saw77] a présenté une formulation plus précise de ce résultat, en donnant un sens rigoureux à cette approximation et en précisant la valeur de la constante C_0.

Les corrélations entre les fréquences alléliques peuvent également s'exprimer à l'aide du dual de l'équation (1.11). Ce dernier est analogue au coalescent structuré de la Définition 1.3.3 à ceci près que chaque particule migre vers la population source à taux μ, indépendemment des autres particules. Si l'on note T_c le temps de coalescence de deux lignées et T_∞ l'instant où l'une d'entre elles migre vers la population source, on a alors

$$\mathbb{E}\left[p(t,i)p(t,j)\right] = \mathbb{P}_{i,j}\left(T_c < T_\infty\right)\bar{p} + \mathbb{P}_{i,j}\left(Tc \geq T_\infty\right)\bar{p}^2.$$

Si l'on note

$$F(i,j) = \mathbb{P}_{i,j}\left(T_c < T_\infty\right) = \mathbb{E}_{i,j}\left[e^{-2\mu T_c}\right],$$

alors $r(k)$ s'écrit

$$r(k) = \frac{F(0,k)}{F(0,0)}. \tag{1.13}$$

Or d'après la formule dite de Wright-Malécot [BDE02], si $d = 2$,

$$\mathbb{E}_{0,k}\left[e^{-2\mu T_c}\right] \simeq \begin{cases} \dfrac{K_0\left(\sqrt{\frac{2\mu}{m}}\,|k|\right)}{\mathcal{N} + \log\left(\sqrt{\frac{m}{2\mu}}/\kappa\right)} & \text{pour } |k| > \kappa, \\ \dfrac{\log\left(\sqrt{\frac{m}{2\mu}}/\kappa\right)}{\mathcal{N} + \log\left(\sqrt{\frac{m}{2\mu}}/\kappa\right)} & \text{pour } |k| \leq \kappa, \end{cases} \tag{1.14}$$

où $\mathcal{N} = 2N\pi m$ est la taille de voisinage au sens de Wright et pour le modèle stepping stone avec migration entre plus proches voisins, $\kappa = 1/\sqrt{32}$. On retrouve alors (1.12) à partir de (1.13).

La quantité $F(i,j)$ est la probabilité que deux individus pris en i et en j aient un ancêtre commun plus récent que leur arrivée depuis la population source, on l'appelle la probabilité d'identité par descendance. En appliquant la propriété de Markov à un interval de temps infinitésimal, on montre que

$$2\mu F(i,j) = \frac{m}{2}\sum_{k\sim i}(F(k,j) - F(i,j)) + \frac{m}{2}\sum_{l\sim j}(F(i,l) - F(i,j)) + \frac{1}{N}(1 - F(i,j))\delta_{ij}. \tag{1.15}$$

Barton, Depaulis et Etheridge [BDE02] montrent comment obtenir la formule de Wright-Malécot (1.14) à partir de cette équation en remplacant formellement les deux laplaciens discrets par un laplacien usuel dans (1.15).

Cette décroissance des corrélations entre les fréquences alléliques avec la distance peut être utilisée pour l'inférence de paramètres démographiques. En effectuant une régression des corrélations empiriques dans les fréquences alléliques contre la distance géographique, on peut estimer la taille du voisinage $\mathcal{N}$ [Rou97]. Il est également possible de retrouver ce paramètre par une approche du maximum de vraisemblance [RL07 ; Bar+13].

1.3.3 Population en espace continu, le processus Λ-Fleming-Viot spatial

Toutes les populations ne sont pas structurées en un réseau régulier de sous-populations bien définies. Il est donc nécessaire de vérifier que les résultats obtenus pour le modèle stepping stone sont robustes à un changement de la géométrie de l'espace. Pour cela, on souhaite définir un modèle mathématique pour des populations occupant un epace continu ($\mathbb{R}^d$ avec $d \in \{1,2,3\}$ par exemple).

Un premier modèle dans ce sens étudié par Malécot [Mal48] suppose que la n-ième génération peut être représentée par un processus ponctuel G_n sur $\mathbb{R}^d$. À la génération suivante, chaque individu produit un nombre aléatoire d'enfants qui se déplacent ensuite indépendemment les uns des autres. Malécot suppose que les nombres d'enfants des différents individus sont des variables de Poisson indépendantes de paramètre 1.

Felsenstein [Fel75] a cependant montré que ce modèle ne pouvait présenter de comportement stationnaire. En effet, chaque famille constitue un processus de branchement critique. La plupart va donc s'éteindre rapidement tandis que celles qui survivent vont devenir très nombreuses, formant des îlots très peuplés (appelés "clumps" par Felsenstein) et de plus en plus espacés. Kingman [Kin77]

a de plus montré que pour qu'un modèle de ce type admette une loi stationnaire sous la forme d'un processus de Poisson, il fallait renoncer à l'indépendance entre les déplacements des différents individus.

Récemment, un nouveau modèle a été introduit par Barton, Etheridge et Véber [Eth08 ; BEV10a] qui permet de décrire l'évolution de la composition génétique d'une population vivant dans un espace continu. Ce modèle inclus un mécanisme strict de régulation locale de la taille de la population, et permet de considérer des situations à l'équilibre analogues à celle étudiée par Malécot.

Ce modèle, appelé le process Λ-Fleming-Viot spatial, décrit l'état de la population à un instant donné par une mesure sur $\mathbb{R}^d \times K$ dont la première marginale est la mesure de Lebesgue sur $\mathbb{R}^d$. (K est l'espace des types génétiques présents dans la population.) À une série d'instants aléatoires, une partie de la population occupant une région donnée meurt et est remplacée par la descendance d'un individu choisi au hasard dans cette région. Dans la suite, nous nous limiterons toujours à une population ne comportant que deux types d'individus ($K = \{0, 1\}$). La population est alors décrite par une fonction $w : \mathbb{R}_+ \times \mathbb{R}^d \to [0, 1]$, où $w_t(x)$ est la proportion d'individus portant l'allèle 1 en x à l'instant t.

Soit μ une mesure sur $(0, \infty)$ et pour $r \in (0, \infty)$, soit ν_r une mesure de probabilité sur $(0, 1]$. On suppose que

$$\int_0^\infty \int_0^1 u r^d \nu_r(du)\mu(dr) < \infty. \tag{1.16}$$

Définition 1.3.4 (processus Λ-Fleming-Viot spatial)**.** *Soit $w_0 : \mathbb{R}^d \to [0, 1]$ une fonction mesurable. Soit Π un processus ponctuel de Poisson sur $\mathbb{R}_+ \times \mathbb{R}^d \times (0, \infty) \times (0, 1]$ de mesure d'intensité $dt \otimes dx \otimes \mu(dr)\nu_r(du)$ [Kin93]. Pour chaque point (t, x, r, u) de Π, un événement de reproduction se produit dans la boule de centre x et de rayon r (notée $B(x, r)$) à l'instant t. Lors d'un tel événement, on choisit y uniformément au hasard dans $B(x, r)$ puis on choisit $k \in \{0, 1\}$ tel que $k = 1$ avec probabilité $w_{t^-}(y)$ et $k = 0$ sinon. Alors pour tout $z \in B(x, r)$,*

$$w_t(z) = (1 - u)w_{t^-}(z) + uk. \tag{1.17}$$

Hors de $B(x, r)$, w_t reste inchangé.

En d'autres termes, à chaque événement de reproduction, une fraction u des individus présents dans la région affectée meurt et est remplacée par la progéniture d'un individu pris au hasard au sein de cette région. Le réel u est appelé paramètre d'impact.

Remarquons que nous n'avons pas précisé l'espace dans lequel la variable aléatoire w_t prend ses valeurs, ni la tribu dont ce dernier est muni. En réalité, le processus $(w_t, t \geq 0)$ doit être vu comme un processus de Markov à valeurs dans l'espace quotient noté Ξ des fonctions mesurables de $\mathbb{R}^d$ dans $[0, 1]$ où l'on identifie les fonctions qui ne diffèrent qu'en un ensemble de mesure de Lebesgue nulle. Cet espace s'identifie à un sous-espace de l'ensemble des mesures sur $\mathbb{R}^d$ qui sont absoluement continue par rapport à la mesure de Lebesgue. On munit Ξ de la métrique d_Ξ suivante qui induit la convergence vague des mesures sur $\mathbb{R}^d$. Soit $(f_n)_{n\geq 1}$ une famille séparante et uniformément bornée de fonctions à valeurs réelles à support compact dans $\mathbb{R}^d$, alors pour $w, w' \in \Xi$,

$$d_\Xi(w, w') = \sum_{n=1}^{\infty} \frac{1}{2^n} \left| \langle w, f_n \rangle - \langle w', f_n \rangle \right|. \tag{1.18}$$

On note que le fait d'identifier les fonctions qui ne diffèrent que sur un ensemble de mesure nulle ne rend pas la Définition 1.3.4 caduque car celle-ci ne fait intervenir que la moyenne de w_{t^-} sur une boule (qui ne dépend pas du représentant que l'on choisit pour w_{t^-}).

Soit $C_c(\mathbb{R}^d)$ l'espace des fonctions continues à support compact de $\mathbb{R}^d$ dans $\mathbb{R}$. Pour $j \in \mathbb{N}^*$, $\psi \in C_c((\mathbb{R}^d)^j)$ et $w \in \Xi$, on pose

$$I(w,\psi) = \int_{(\mathbb{R}^d)^j} \prod_{i=1}^{j} w(x_i)\psi(x_1,\dots,x_j)dx_1\dots dx_j.$$

On note V_r le volume de la boule de rayon r dans $\mathbb{R}^d$. Alors le générateur infinitésimal de $(w_t, t \geq 0)$ agit sur les fonctions $I(\cdot,\psi)$ de la façon suivante [BEV10a],

$$\mathcal{L}I(w,\psi) = \int_{\mathbb{R}^d} dx \int_0^\infty \mu(dr) \int_0^1 \nu_r(du) \frac{1}{V_r} \int_{B(x,r)} dy \int_{(\mathbb{R}^d)^j} \psi(x_1,\dots,x_j)dx_1\dots dx_j$$
$$\prod_{i\in\mathcal{I}^c} w(x_i) \left[w(y) \prod_{i\in\mathcal{I}} ((1-u)w(x_i)+u) + (1-w(y)) \prod_{i\in\mathcal{I}} ((1-u)w(x_i)) - \prod_{i\in\mathcal{I}} w(x_i) \right], \quad (1.19)$$

où $\mathcal{I}$ désigne l'ensemble des indices i tels que $x_i \in B(x,r)$. Puisque l'ensemble des combinaisons linéaires des fonctions de la forme $I(\cdot,\psi)$ et de fonctions constantes est dense dans l'espace des fonctions continues sur Ξ [BEV10a, Lemme 4.1], le processus Λ-Fleming-Viot spatial (ou SLFV) est alors défini comme l'unique processus de Markov à valeurs dans Ξ qui est solution du problème de martingales pour le générateur $\mathcal{L}$ défini par (1.19). L'existence d'un tel processus est prouvée dans [BEV10a], en utilisant la relation de dualité que nous exposons ci-après.

1.3.4 Le dual du processus Λ-Fleming-Viot spatial

De même que pour le modèle stepping stone, on peut décrire la généalogie d'un échantillon aléatoire d'individus dans le SLFV sous la forme d'un système de particules coalescentes. Les particules à l'instant t correspondent aux ancêtres de l'échantillon t unités de temps dans le passé. Comme le processus Π de la définition 1.3.4 est invariant par retournement du temps, c'est le même processus qui régit les événements de reproduction passés.

Définition 1.3.5 (Λ-coalescent spatial)**.** *Soit Π un processus ponctuel de Poisson sur $\mathbb{R}_+ \times \mathbb{R}^d \times (0,\infty) \times (0,1]$ de mesure d'intensité $dt \otimes dx \otimes \mu(dr)\nu_r(du)$. Soit un système de particules qui évolue selon la dynamique suivante. Pour chaque point (t,x,r,u) de Π, un événement de reproduction se produit à l'instant t dans $B(x,r)$. Lors d'un tel événement, on choisit un point y uniformément au hasard dans $B(x,r)$. Chaque particule située dans $B(x,r)$ à l'instant t est ensuite marquée avec probabilité u, indépendemment des autres particules. Puis toutes les particules marquées coalescent en une particule située en y.*

On note N_t le nombre de particules au temps t dans le Λ-coalescent spatial, et on note $\xi_t^1,\dots,\xi_t^{N_t}$ leurs positions respectives. Le processus $(\mathcal{A}_t, t \geq 0)$ avec

$$\mathcal{A}_t = \{\xi_t^1,\dots,\xi_t^{N_t}\}$$

est alors un processus de Markov dont le générateur infinitésimal agit sur les fonctions de la forme

$$F_w(\{\xi^1,\dots,\xi^N\}) = \prod_{i=1}^{N} w(\xi^i), \quad w : \mathbb{R}^d \to \mathbb{R},$$

de la manière suivante

$$\mathcal{L}'F_w(\{\xi^1,\dots,\xi^N\}) = \int_{\mathbb{R}^d} dx \int_0^\infty \mu(dr) \int_0^1 \nu_r(du) \frac{1}{V_r} \int_{B(x,r)} dy \prod_{i\in\mathcal{I}^c} w(\xi^i) \left\{ \sum_{\substack{D\subset\mathcal{I}\\|D|\geq 1}} u^{|D|}(1-u)^{|\mathcal{I}\setminus D|} \left(w(y) \prod_{i\in\mathcal{I}\setminus D} w(\xi^i) - \prod_{i\in\mathcal{I}} w(\xi^i) \right) \right\}.$$

En d'autres termes, pour tout $\{x_1,\dots,x_j\} \in (\mathbb{R}^d)^j$,

$$\lim_{t\downarrow 0} \frac{1}{t} \mathbb{E}_{x_1,\dots,x_j} \left[F_w(\mathcal{A}_t) - F_w(\{x_1,\dots,x_j\}) \right] = \mathcal{L}'F_w(\{x_1,\dots,x_j\}),$$

où $\mathbb{E}_{x_1,\dots,x_j}$ désigne la loi de $(\mathcal{A}_t)_{t\geq 0}$ lorsque $\mathcal{A}_0 = \{x_1,\dots,x_j\}$. On peut alors montrer [BEV10a] (voir aussi [EVY14]) que pour $\psi \in C_c((\mathbb{R}^d)^j)$,

$$\mathcal{L}I(w,\psi) = \int_{(\mathbb{R}^d)^j} \mathcal{L}'F_w(\{x_1,\dots,x_j\})\psi(x_1,\dots,x_j)dx_1\dots dx_j.$$

D'après la Proposition 1.2.9, cela prouve la relation de dualité (de moments) entre le SLFV et le Λ-coalescent spatial.

Proposition 1.3.6 (Dualité pour le SLFV, [BEV10a])**.** *Pour tout $\psi \in C_c((\mathbb{R}^d)^j)$ et pour toute fonction mesurable $w_0 : \mathbb{R}^d \to [0,1]$, on a*

$$\mathbb{E}_{w_0}\left[I(w_t,\psi)\right] = \int_{(\mathbb{R}^d)^j} \mathbb{E}_{x_1,\dots,x_j} \left[\prod_{i=1}^{N_t} w_0(\xi_t^i) \right] \psi(x_1,\dots,x_j)dx_1\dots dx_j. \tag{1.20}$$

Dans [VW15], Véber et Wakolbinger montrent une relation de dualité plus forte entre ces deux processus, mais nous omettons de la détailler ici. Cette relation de dualité permet notamment de caractériser le comportement à grande échelle du SLFV, comme nous allons le voir dans la sous-section suivante.

Si l'on suppose de plus que des mutations se produisent le long de chaque lignée à un taux μ, on peut définir comme dans le modèle stepping stone la probabilité d'identité par descendance (c'est-à-dire la probabilité qu'aucun de leurs ancêtres n'ait subit de mutation depuis leur ancêtre commun le plus récent) de deux individus échantillonnés à une distance $r > 0$ l'un de l'autre, notée $F(r)$. Si les deux lignées correspondantes coalescent après un temps T_c, alors cette probabilité est $e^{-2\mu T_c}$, on a donc

$$F(r) = \mathbb{E}_{0,r}\left[e^{-2\mu T_c}\right].$$

On peut alors montrer que le Λ-coalescent spatial vérifie la formule de Wright-Malécot (1.14) où m et $\mathcal{N}$ peuvent être exprimés en fonction de μ et ν_r [BDE02 ; Bar+13].

1.3.5 Comportement à grande échelle du SLFV

En s'inspirant de l'approximation des modèles de Wright-Fisher et de Moran par la diffusion de Wright-Fisher, on est tenté d'approcher l'évolution des fréquences alléliques dans une population structurée en espace par la solution de l'équation aux dérivées partielles stochastique suivante

$$\partial_t w_t = \frac{\sigma^2}{2}\Delta w_t + \sqrt{\frac{1}{N} w_t(1-w_t)}\dot{W}_t, \tag{1.21}$$

où $\sigma^2 > 0$, $N > 0$ et W est un bruit blanc gaussien. L'équation (1.21) équivaut à dire que pour toute fonction $\phi : \mathbb{R}^d \to \mathbb{R}$ décroissant suffisamment rapidement à l'infini, en notant $\langle f, g\rangle = \int_{\mathbb{R}^d} f(x)g(x)dx$,

$$\langle w_t, \phi\rangle - \frac{\sigma^2}{2}\int_0^t \langle w_s, \Delta\phi\rangle\, ds$$

est une martingale locale continue (pour la filtration engendrée par $(w_t, t \geq 0)$) de variation quadratique

$$\frac{1}{N}\int_0^t \int_{\mathbb{R}^d} \phi(x)^2 w_s(x)(1-w_s(x))dxds.$$

On peut trouver une introduction détaillée aux équations aux dérivées partielles (EDP) stochastiques dans [Wal78 ; PZ14].

Malheureusement, cette approche n'est possible que lorsque la dimension de l'espace est 1. En dimensions supérieures, (1.21) ne possède pas de solution. Nous détaillons ici les conditions sous lesquelles cette équation constitue une bonne approximation du SLFV ainsi que les autres régimes possibles.

Paramètre d'impact fixé Considérons dans un premier temps le SLFV avec un rayon et un paramètre d'impact fixés, c'est-à-dire où μ et ν_r sont des mesures de Dirac, respectivement et $r > 0$ et $u \in (0, 1]$. Notons le processus correspondant $(w_t, t \geq 0)$ et soit w_0 une fonction mesurable de $\mathbb{R}^d$ dans $[0, 1]$. Pour $n \geq 1$, on pose

$$\mathbf{w}^n(t, x) = w_{nt}(\sqrt{n}x) \tag{1.22}$$

et on suppose que $\mathbf{w}^n(0, x) = w_0(x)$ pour tout $n \geq 1$. Soit $\rho(t, x), t \geq 0, x \in \mathbb{R}^d$ la solution de l'équation aux dérivées partielles

$$\begin{cases} \partial_t \rho(t, x) = \frac{\sigma^2}{2}\Delta\rho(t, x) \\ \rho(0, x) = w_0(x) \end{cases} \tag{1.23}$$

avec $\sigma^2 = uV_r\frac{2r^2}{d+2}$. Le résultat suivant est dû à Berestycki, Etheridge et Véber [BEV13b].

Théorème 1.1 ([BEV13b])**.** *Lorsque n tend vers l'infini, la suite de processus $(\mathbf{w}^n, n \geq 1)$ converge au sens des lois fini-dimensionnelles dans la topologie vague vers un processus $(p(t, \cdot), t \geq 0)$. En dimension 1, $p(t, x)$ est une variable de Bernoulli de paramètre $\rho(t, x)$ et les corrélations entre différentes valeurs de $p(t, \cdot)$ sont non triviales et sont données par* (1.24)*. En dimensions 2 et plus, $p(t, x)$ est déterministe et égal à $\rho(t, x)$.*

On prouve le Théorème 1.1 en étudiant le dual du SLFV. Notons que si l'on suit une lignée

ancestrale, celle-ci se comporte comme une marche aléatoire symétrique de variance

$$\sigma^2 = u\frac{1}{dV_r}\int_{B(0,r)} dx \int_{B(x,r)} |y|^2\, dy = uV_r\frac{2r^2}{d+2}.$$

Le changement d'échelle (1.22) revient alors à poser

$$\mathcal{A}_t^n = \left\{ \frac{1}{\sqrt{n}}\xi_{nt}^1, \ldots, \frac{1}{\sqrt{n}}\xi_{nt}^{N_{nt}} \right\}.$$

En laissant tendre n vers $+\infty$, $\mathcal{A}^n$ converge vers un système de mouvements browniens de variance σ^2 qui coalescent dès qu'ils se rencontrent [BEV13b]. En dimensions 2 et plus, deux mouvements browniens ne se rencontrent jamais et on n'observe donc pas de coalescence à la limite. Pour voir qu'en dimension 1 deux lignées coalescent instantanément lorsqu'elles se rencontrent, il suffit de noter qu'après le changement d'échelle (1.22), deux lignées situées à une distance inférieure à $\frac{2r}{\sqrt{n}}$ coalescent après un temps d'ordre $\frac{1}{\sqrt{n}}$ (pour qu'elles passent un temps d'ordre 1 à une distance inférieure à $\frac{2r}{\sqrt{n}}$ l'une de l'autre, voir [BEV10a, Proposition 6.4]). En passant à la limite, la coalescence se confond donc avec la rencontre des lignées.

Pour voir que ce résultat implique le Thèorème 1.1, on utilise la dualité de la Proposition 1.3.6 pour établir les corrélations entre les différentes valeurs de $p(t,\cdot)$. En dimensions 2 et plus, deux lignées (X^1, X^2) ne coalescent jamais et évoluent donc indépendemment l'une de l'autre. Alors

$$\begin{aligned}\mathbb{E}\left[p(t,x)^2\right] &= \mathbb{E}_{x,x}\left[w_0(X_t^1)w_0(X_t^2)\right] \\ &= \mathbb{E}_x\left[w_0(X_t^1)\right]^2 \\ &= \mathbb{E}\left[p(t,x)\right]^2,\end{aligned}$$

ce qui implique que $p(t,x)$ soit déterministe. En dimension 1, comme deux lignées coalescent dès qu'elles se rencontrent, à tout instant strictement positif chaque site n'est occupé que par un type d'individus. La population devient ainsi instantanément divisée en plusieurs zones homogènes dont les frontières forment un système de mouvements browniens qui s'annihilent dès qu'ils se touchent. En passant à la limite dans (1.20), on obtient

$$\mathbb{E}_{w_0}\left[I(p(t,\cdot),\psi)\right] = \int_{(\mathbb{R}^d)^j} \mathbb{E}_{x_1,\ldots,x_j}\left[\langle w_0, \mathcal{A}_t^\infty\rangle\right]\psi(x_1,\ldots,x_j)dx_1\ldots dx_j. \tag{1.24}$$

On en déduit les corrélations entre les différentes valeurs de $p(t,\cdot)$ en dimension 1.

Paramètre d'impact tendant vers 0 Si u est très proche de zéro, on peut observer des comportements limites différents. Pour cela, posons pour $n \geq 1$

$$u_n = \frac{u}{\sqrt{n}}$$

et supposons que $(w_t^n, t \geq 0)$ est le SLFV avec $\mu = \delta_r$ et $\nu_r = \delta_{u_n}$. On pose alors

$$\bar{w}^n(t,x) = \frac{1}{V_r}\int_{B(x,r)} w_t^n(y)dy$$

et

$$\mathbf{w}^n(t,x) = \bar{w}^n(n^{3/2}t, \sqrt{n}x). \tag{1.25}$$

On suppose de plus que $\mathbf{w}^n(0,x) = w_0(x)$ pour tout $n \geq 1$. Dans ce cas, Etheridge, Véber et Yu [EVY14] montrent le résultat suivant (ils considèrent en fait le SLFV avec sélection, voir Section 1.4). Pour un espace métrique Y, on note $\mathrm{D}(\mathbb{R}_+, Y)$ l'espace des fonctions càdlàg de $\mathbb{R}_+$ dans Y et on le munit de la topologie de Skorokhod usuelle [Bil99]. De même pour $T > 0$ on notera $\mathrm{D}([0,T], Y)$ l'espace des fonctions càdlàg de $[0,T]$ dans Y, muni de la topologie de Skorokhod.

Théorème 1.2 ([EVY14])**.** *Lorsque $n \to \infty$, la suite de processus $(\mathbf{w}^n, n \geq 1)$ converge en loi dans $\mathrm{D}(\mathbb{R}_+, \Xi)$ vers un processus $(w_t^\infty, t \geq 0)$ tel que $w_0^\infty = w_0$. Si $d = 1$ alors w^∞ est solution de* (1.21) *avec*

$$\sigma^2 = uV_r\frac{2r^2}{d+2} \qquad \text{et} \qquad N = \frac{1}{V_r^2u^2}. \tag{1.26}$$

Lorsque $d \geq 2$, w^∞ est déterministe et égal à ρ (1.23)*.*

En dimension 1, ce théorème est un analogue en espace continu du résultat de Müeller et Tribe [MT95] sur les limites d'échelles du modèle du votant.

Comme u_N tend vers zéro, l'échelle de temps considérée est différente. Sous le changement d'échelle (1.25), on pose pour le dual

$$\mathcal{A}_t^n = \left\{\frac{1}{\sqrt{n}}\xi^1_{n^{3/2}t}, \ldots, \frac{1}{\sqrt{n}}\xi^{N_{n^{3/2}t}}_{n^{3/2}t}\right\}.$$

Dans ce cas, en dimensions 2 et plus, $\mathcal{A}^n$ converge également vers un système de mouvements browniens indépendants de variance σ^2. De plus, si l'on considère deux lignées dans $\mathcal{A}$, alors elles coalescent à un temps T_c tel que le temps passé par ces deux lignées à une distance inférieure à $2r$ l'une de l'autre jusqu'à T_c est approximativement une variable exponentielle de paramètre $u^2V_r^2$ (voir par exemple [DR08]). En dimension 1, $\mathcal{A}^n$ converge donc vers un système de mouvements browniens qui coalescent lorsque le temps local à l'origine de leur différence atteint une variable exponentielle de paramètre $u^2V_r^2$. Il est montré dans [Lia09] qu'un tel système vérifie une relation de dualité du même type que (1.24) avec la solution de (1.21) dont les paramètres sont donnés par (1.26).

Enfin, on verra au Chapitre 2 que si u_N est négligeable devant $n^{-1/2}$, sous certaines conditions, le SLFV s'approche de ρ sous un bon changement d'échelle, même en dimension 1.

Il ressort des résultats ci-dessus que le paramètre d'impact détermine la variance des fréquences alléliques à grande échelle. On peut en conclure que la coalescence des lignées ancestrales dans le coalescent spatial est intimement liée aux corrélations dans les fluctuations des fréquences alléliques dans le SLFV. L'étude de celui-ci permet donc d'en déduire le comportement de celui-là, et inversement. Dans les chapitres suivants, nous ferons de nombreux allers-retours entre les deux points de vue, en mettant l'accent parfois plus sur l'un ou sur l'autre selon ce qui sera le plus commode mathématiquement.

1.4 Fluctuations dans la composition génétique d'une population structurée en espace en présence de sélection naturelle

Dans la Proposition 1.2.5, nous avons vu que lorsque Ns est suffisamment loin de 0 et de $+\infty$, le modèle de Wright-Fisher de taille N et de coefficient de sélection s peut être approché par une diffusion. Si en revanche Ns est très grand, alors la sélection naturelle est plus forte que les fluctuations dues à la dérive génétique. On s'attend donc à ce que les fréquences alléliques soient quasiment déterministes.

Ce problème a été étudié par Norman [Nor75a]. Il a montré que si $(X_t^N, t \geq 0)$ suit le modèle de Wright-Fisher avec sélection de taille N (Définition 1.2.2) de coefficient de sélection $s_N = \varepsilon_N s$ tel que

$$\varepsilon_N \underset{N\to\infty}{\longrightarrow} 0, \qquad\qquad \varepsilon_N N \underset{N\to\infty}{\longrightarrow} +\infty,$$

alors $(X^N_{t/\varepsilon_N}, t \geq 0)$ converge en loi vers la solution de l'équation differentielle ordinaire

$$\partial_t g_t = s g_t (1 - g_t).$$

De plus, si l'on pose

$$Z_t^N = (N\varepsilon_N)^{-1/2} \left(X^N_{t/\varepsilon_N} - g_t \right),$$

alors, lorsque $N \to \infty$, Z^N converge en loi vers un processus de diffusion $(z_t, t \geq 0)$ solution de

$$\begin{cases} dz_t = s(1-2g_t)z_t dt + \sqrt{g_t(1-g_t)} dB_t, \\ z_0 = 0, \end{cases} \tag{1.27}$$

où B est un mouvement brownien standard.

Le processus Z^N donne une mesure des fluctuations dans le modèle de Wright-Fisher autour du processus déterministe $(g_{\varepsilon_N t}, t \geq 0)$ lorsque Ns est grand. De plus on remarque que z_t suit une loi gaussienne centrée de variance σ_t^2 qui vérifie d'après la formule d'Itô,

$$\partial_t \sigma_t^2 = 2s(1-2g_t)\sigma_t^2 + g_t(1-g_t).$$

Le résultat de Norman peut donc être vu comme un théorème de la limite centrale pour le modèle de Wright-Fisher. Ce régime avait déjà reçu l'attention de Feller [Fel51] dans un cas particulier où X^N fluctue autour d'une constante dans $(0, 1)$.

Pour voir comment on obtient l'équation (1.27), on peut remplacer $(X_t^N, t \geq 0)$ par la diffusion de Wright-Fisher (1.4),

$$dx_t = s_N x_t(1-x_t)dt + \sqrt{\frac{1}{N} x_t(1-x_t)} dB_t.$$

En notant $x_t^N = x_{t/\varepsilon_N}$, on a

$$dx_t^N = s x_t^N (1-x_t^N) dt + \sqrt{\frac{1}{N\varepsilon_N} x_t^N (1-x_t^N)} dB_t.$$

Comme $\varepsilon_N N \to \infty$, le second terme est négligeable devant le premier et on retrouve la convergence de x^N vers g. De plus, en posant

$$Z_t^N = (\varepsilon_N N)^{1/2} \left(x_t^N - g_t\right),$$

on obtient

$$dZ_t^N = (\varepsilon_N N)^{1/2} s \left[x_t^N(1-x_t^N) - g_t(1-g_t)\right] dt + \sqrt{x_t^N(1-x_t^N)} dB_t.$$

En appliquant la formule de Taylor à $x \mapsto x(1-x)$, il vient

$$dZ_t^N = s(1-2g_t)Z_t^N dt - \frac{s}{\sqrt{\varepsilon_N N}}(Z_t^N)^2 dt + \sqrt{x_t^N(1-x_t^N)} dB_t. \tag{1.28}$$

Le deuxième terme est négligeable car $\varepsilon_N N \to \infty$ et on peut utiliser la convergence de x^N vers g dans le troisième terme pour obtenir (1.27) lorsque $N \to \infty$. Ces techniques sont également utilisée par Kurtz pour divers modèles en génétiques des populations [Kur71 ; Kur81].

1.4.1 Un théorème de la limite centrale pour le SLFV avec sélection

En collaboration avec Sarah Penington, nous avons étendu les travaux de Norman aux populations structurées spatiallement [FP17]. Nous avons vu dans la section précédente que, au moins en dimension 2 et plus, mais en réalité également en dimension 1 sous certaines conditions, le processus Λ-Fleming-Viot spatial (Définition 1.3.5) se comporte comme la solution de l'équation de la chaleur sur $\mathbb{R}^d$ (1.23). Nous avons montré que les fluctuations autour de cette limite déterministe sont asymptotiquement proches d'un processus solution d'une équation aux dérivées partielles stochastique.

Avant d'introduire plus en détail ces résultats, nous exposons comment prendre en compte la sélection naturelle dans le SLFV, d'après un modèle introduit par Etheridge, Véber et Yu [EVY14]. Cette construction s'inspire du mécanisme de sélection dans le modèle de Moran (Définition 1.2.4). On se limite de plus au cas où le paramètre d'impact u est fixé.

Définition 1.4.1 (SLFV avec sélection)**.** *Soient $u \in (0,1]$ et $s \in (0,1)$ fixés et soit μ une mesure sur $(0,\infty)$. Soient Π et Π^S deux processus ponctuels de Poisson indépendants sur $\mathbb{R}_+ \times \mathbb{R}^d \times (0,\infty)$ de mesure d'intensité respectivement $(1-s)dt \otimes dx \otimes \mu(dr)$ et $sdt \otimes dx \otimes \mu(dr)$. Le processus Λ-Fleming-Viot spatial avec sélection (SLFVS) est défini comme suit.*

Pour chaque point $(t,x,r) \in \Pi$, un événement de reproduction neutre se produit à l'instant t dans la boule $B(x,r)$, suivant le même principe que dans la définition (1.3.5)*.*

Pour $(t,x,r) \in \Pi^S$, un événement sélectif se produit à l'instant t dans la boule $B(x,r)$. On choisit alors deux points y_1 et y_2 indépendemment uniformément au hasard dans $B(x,r)$. On choisit ensuite un type k_i en chaque point y_i tel que $k_i = 1$ avec probabilité $w_{t^-}(y_i)$ et $k_i = 0$ sinon. On pose alors $k = k_1 \vee k_2$ et on met à jour w suivant l'équation (2.7)

On considère deux cas :

i) *rayon fixe*, où $\mu(dr) = \delta_R(dr)$ pour un $R > 0$ fixé,

ii) *cas stable*, où

$$\mu(dr) = \frac{\mathbb{1}_{\{r\geq 1\}}}{r^{d+\alpha+1}}dr \tag{1.29}$$

pour $\alpha \in (0, 2 \wedge d)$ fixé.

Dans les deux cas, (1.16) est vérifié et le SLFVS est bien défini (voir [EVY14]). Les deux Théorèmes cités ci-dessous disent la même chose, respectivement dans les cas (i) et (ii) : sous un bon changement d'échelle, le SLFVS se comporte comme la solution d'une équation aux dérivées partielles et les fluctuations autour de cette EDP sont données par la solution d'une EDP stochastique qui s'obtient en linéarisant le problème de martingales associé au SLFVS. Ces résultats sont démontrés dans le Chapitre 2 et ont fait l'objet d'une publication [FP17].

Rayon fixe Soit $w_0 : \mathbb{R}^d \to [0,1]$ une fonction quatre fois dérivable de dérivées uniformément bornées. Prenons deux suites $(\varepsilon_N)_{N\geq 1}$ et $(\delta_N)_{N\geq 1}$ de réels positifs dans $(0,1]$ décroissantes vers zéro et posons

$$s_N = \delta_N^2 s, \qquad u_N = \varepsilon_N u, \qquad r_N = \delta_N R, \qquad q_0^N(x) = w_0(\delta_N x).$$

Soit alors $(q_t^N, t \geq 0)$ le SLFVS de paramètre d'impact u_N, de coefficient de sélection s_N et de rayon de reproduction R, avec pour condition initiale q_0^N. Pour $x \in \mathbb{R}^d$, $t \geq 0$, on définit

$$\mathbf{q}_t^N(x) = q^N_{t/(\varepsilon_N \delta_N^2)}(x/\delta_N). \tag{1.30}$$

Soit $f : \mathbb{R}_+ \times \mathbb{R}^d \to [0,1]$ la solution de l'EDP suivante

$$\begin{cases} \partial_t f_t = uV_R \left[\frac{R^2}{d+2} \Delta f_t + s f_t(1-f_t) \right] \\ f_0 = w_0. \end{cases} \tag{1.31}$$

On note de plus $\mathcal{S}(\mathbb{R}^d)$ l'espace de Schwartz des fonctions lisses et décroissant rapidement à l'infini et $\mathcal{S}'(\mathbb{R}^d)$ l'espace des distributions tempérées.

Théorème 1.3 (Théorème de la limite centrale pour le SLFVS - rayon fixe [FP17])**.** *Si $\varepsilon_N = o\left(\delta_N^{d+2}\right)$, alors, pour tout $T > 0$ fixé,*

$$\mathbb{E}\left[\sup_{t\in[0,T]} d_\Xi\left(\mathbf{q}_t^N, f_t\right) \right] \underset{N\to\infty}{\longrightarrow} 0. \tag{1.32}$$

De plus, il existe une suite de fonctions $f^N : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ déterministes, convergeant uniformément vers f sur $[0,T] \times \mathbb{R}^d$, telle que si l'on pose

$$Z_t^N = (\varepsilon_N \delta_N^{d-2})^{-1/2} \left(\mathbf{q}_t^N - f_t^N \right),$$

alors la suite de processus $(Z^N, N \geq 1)$ converge dans $\mathrm{D}\left([0,T], \mathcal{S}'(\mathbb{R}^d)\right)$ (au sens défini par Walsh dans [Wal86]) vers la solution de l'EDP stochastique suivante

$$\begin{cases} \partial_t z_t = uV_R \left[\frac{R^2}{d+2} \Delta z_t + s(1-2f_t) z_t \right] + uV_R \sqrt{f_t(1-f_t)} \dot{W}_t \\ z_0 = 0, \end{cases} \tag{1.33}$$

où W est un bruit blanc gaussien sur $\mathbb{R}_+ \times \mathbb{R}^d$.

On reconnaît dans (1.31) l'équation de Fisher-KPP proposée par Fisher [Fis37] pour modéliser le front d'invasion d'un allèle avantagé par la sélection naturelle. Le SLFVS permet donc d'étudier l'effet de la dérive génétique sur cette équation.

Cas stable Dans le cas où $\mu(dr)$ est donnée par (1.29), on suppose que $w_0 : \mathbb{R}^d \to [0,1]$ est deux fois dérivable de dérivées uniformément bornées. Pour deux suites $(\varepsilon_N)_{N\geq 1}$ et $(\delta_N)_{N\geq 1}$ décroissantes vers zéro, on pose cette fois

$$s_N = \delta_N^\alpha s, \qquad u_N = \varepsilon_N u, \qquad q_0^N(x) = w_0(\delta_N x).$$

Soit alors $(q_t^N, t \geq 0)$ le SLFVS de paramètre d'impact u_N, de coefficient de sélection s_N et de mesure $\mu(dr)$ donnée par (1.29) avec pour condition initiale q_0^N. Pour $t \geq 0$, $x \in \mathbb{R}^d$, on définit

$$\mathbf{q}_t^N(x) = q^N_{t/(\varepsilon_N \delta_N^\alpha)}(x/\delta_N).$$

Le résultat dans le cas stable est analogue à celui présenté plus haut, à ceci près que le Laplacien dans (1.31) et (1.33) doit être remplacé par un Laplacien fractionnaire et que le bruit qui régit les fluctuations asymptotiques est maintenant corrélé en espace.

Soit $f : \mathbb{R}_+ \times \mathbb{R}^d \to [0,1]$ la solution de

$$\begin{cases} \partial_t f_t = u\left[\mathcal{D}^\alpha f_t + \frac{sV_1}{\alpha} f_t(1-f_t)\right] \\ f_0 = w_0, \end{cases}$$

où $\mathcal{D}^\alpha$ désigne, à une constante multiplicative près, le Laplacien fractionnaire d'indice α [SKM93].

Théorème 1.4 (Théorème de la limite centrale pour le SLFVS - cas stable [FP17])**.** *Si $\varepsilon_N = o\left(\delta_N^{2\alpha}\right)$, alors pour tout $T > 0$ fixé, on a* (1.32)*. De plus, il existe une suite de fonctions $f^N : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ déterministes convergeant uniformément vers f sur $[0,T] \times \mathbb{R}^d$ telle que, si l'on pose*

$$Z_t^N = \varepsilon_N^{-1/2}\left(\mathbf{q}_t^N - f_t^N\right),$$

alors la suite de processus $(Z^N, N \geq 1)$ converge en loi dans $\mathrm{D}\left([0,T], \mathcal{S}'(\mathbb{R}^d)\right)$ vers la solution de l'EDP stochastique suivante

$$\begin{cases} \partial_t z_t = u\left[\mathcal{D}^\alpha z_t + \frac{sV_1}{\alpha}(1-2f_t)z_t\right] + u\dot{W}_t^\alpha \\ z_0 = 0, \end{cases}$$

où W^α est un bruit gaussien non corrélé en temps et dont les corrélations spatiales sont précisées dans le Chapitre 2.

L'approche générale permettant de montrer ces résultats est analogue à celle mise en place par Norman pour le modèle de Wright-Fisher. La dimension spatiale du problème rend cependant le traitement bien plus technique.

En dimension 1, on peut exposer le fil conducteur de la preuve dans le cas du rayon fixe en remplaçant le SLFVS par la solution de l'équation suivante (voir [EVY14])

$$\partial_t q_t^N = u_N V_R \left[\frac{R^2}{d+2} \Delta q_t^N + s_N q_t^N (1 - q_t^N) \right] + u_N V_R \sqrt{q_t^N (1 - q_t^N)} \dot{W}_t. \tag{1.34}$$

Après le changement d'échelle (1.30), cette équation devient

$$\partial_t \mathbf{q}_t^N = u V_R \left[\frac{R^2}{d+2} \Delta \mathbf{q}_t^N + s \mathbf{q}_t^N (1 - \mathbf{q}_t^N) \right] + (\varepsilon_N \delta_N^{d-2})^{-1/2} u V_R \sqrt{\mathbf{q}_t^N (1 - \mathbf{q}_t^N)} \dot{W}_t.$$

Comme $\varepsilon_N = o\left(\delta_N^{d+2}\right)$, le dernier terme est négligeable lorsque N est grand, d'où la convergence de $\mathbf{q}^N$ vers f. De plus, si l'on retranche (1.31) à cette équation et qu'on la multiplie par $(\varepsilon_N \delta_N^{d-2})^{1/2}$, on obtient

$$\partial_t Z_t^N = u V_R \left[\frac{R^2}{d+2} \Delta Z_t^N + s (\varepsilon_N \delta_N^{d-2})^{1/2} \left(\mathbf{q}_t^N (1 - \mathbf{q}_t^N) - f_t (1 - f_t) \right) \right] + u V_R \sqrt{\mathbf{q}_t^N (1 - \mathbf{q}_t^N)} \dot{W}_t.$$

En appliquant la formule de Taylor comme en (1.28), on écrit

$$\partial_t Z_t^N = u V_R \left[\frac{R^2}{d+2} \Delta Z_t^N + s(1 - 2 f_t) Z_t^N - \frac{s}{(\varepsilon_N \delta_N^{d-2})^{1/2}} (Z_t^N)^2 \right] + u V_R \sqrt{\mathbf{q}_t^N (1 - \mathbf{q}_t^N)} \dot{W}_t.$$

En négligeant le terme quadratique en Z^N et en remplaçant $\mathbf{q}^N$ par f dans le dernier terme, on obtient (1.33). Le Théorème 2.1 rend ce raisonnement rigoureux quelque soit la dimension de l'espace (même lorsque (1.34) est mal posée).

Le raisonnement pour le SLFVS en dimensions supérieures est analogue, à certains points techniques près. Le contrôle du terme quadratique est plus délicat que dans le cas non spatial car $(Z^N, N \geq 1)$ converge en tant que processus à valeurs dans un espace de distributions, son carré n'a donc plus de sens à la limite. La condition $\varepsilon_N = o\left(\delta_N^{d+2}\right)$ permet d'assurer que ce terme est bien négligeable asymptotiquement (voir les détails au Chapitre 2).

Le raisonnement est analogue dans le cas stable, et nous ne le détaillons pas ici. On notera seulement que la forme des corrélations dans $\dot{W}^\alpha$ est similaire à celle obtenue [BEK06] comme limite de superpositions aléatoires de disques dont les rayons suivent une loi dans le domaine d'attraction d'une loi stable (comme c'est le cas pour $\mu(dr)$).

1.4.2 Application au fardeau de dérive dans une population structurée spatialement

Dans le mécanisme de sélection défini plus haut (Définition 1.4.1), un allèle est systématiquement avantagé par rapport à l'autre et finit par se fixer dans la population. Si chaque individu porte deux copies de chaque chromosome (on dit alors qu'ils sont diploïdes), d'autres comportements sont possibles. C'est notamment le cas si les individus qui portent deux allèles différents (appelés hétérozygotes) produisent plus de descendants que ceux qui portent deux copies du même allèle (ces derniers sont dits homozygotes), on parle alors de surdominance.

On suppose dans la suite qu'à un locus donné, deux allèles coexistent, que l'on note A_1 et A_2, et

que la fitness relative des trois génotypes est

$$\begin{array}{ccc} A_1A_1 & A_1A_2 & A_2A_2 \\ 1-s_1 & 1 & 1-s_2. \end{array} \tag{1.35}$$

Si le type A_1 est majoritaire dans la population, un individu portant l'allèle A_2 aura de bonnes chances de s'associer avec un allèle A_1 et produira donc plus de descendants que les autres individus dans la population. Réciproquement, si l'allèle A_2 est majoritaire, l'allèle A_1 augmentera en fréquence. La sélection naturelle agit donc de façon à maintenir une proportion stable de chaque allèle dans la population, pour laquelle ces deux effets se compensent exactement.

Si la fréquence de l'allèle A_1 dans la population est q (et celle de A_2 $1-q$), alors la fitness moyenne de la population est

$$\begin{aligned} K &= (1-s_1)q^2 + 2q(1-q) + (1-s_2)(1-q)^2 \\ &= 1 - \frac{s_1 s_2}{s_1+s_2} - (s_1+s_2)\left(q - \frac{s_2}{s_1+s_2}\right)^2. \end{aligned} \tag{1.36}$$

Si la population est de taille finie, alors la fréquence de A_1 fluctue autour de l'équilibre (pour lequel K est maximal) sous l'effet de la dérive génétique. Ces fluctuations font diminuer K d'autant plus que l'on s'éloigne de l'équilibre et donc d'autant plus que la taille de la population est petite. Cet écart entre la valeur maximale de K et sa vraie valeur se nomme le fardeau de dérive.

Robertson [Rob70] a étudié ce problème dans le cas d'une population panmictique. Il a montré que dans le modèle de Wright-Fisher avec un mécanisme de sélection suivant les fitness (1.35), si s_1 et s_2 sont proches de zéro, le fardeau de dérive est asymptotiquement égal à $\frac{1}{4N}$, où N est la taille de la population, quelle que soit la force de la sélection naturelle. On peut facilement s'en convaincre à partir de la diffusion de Wright-Fisher, qui s'écrit dans ce cas pour la fréquence de l'allèle A_1 :

$$dq_t = q_t(1-q_t)(s_2 - (s_1+s_2)q_t)dt + \sqrt{\frac{1}{2N}q_t(1-q_t)}dB_t.$$

(On a remplacé N par $2N$ car la population est diploïde, il y a donc $2N$ gènes dans la population.) On note $\lambda = \frac{s_2}{s_1+s_2}$ la fréquence de l'allèle A_1 à l'équilibre et on pose

$$Z_t = q_t - \lambda.$$

Pour q_t suffisamment proche de λ, en négligeant les termes d'ordre supérieur, on écrit

$$dZ_t = -(s_1+s_2)\lambda(1-\lambda)Z_t dt + \sqrt{\frac{\lambda(1-\lambda)}{2N}}dB_t.$$

(Un traitement plus rigoureux peut être fait en reprenant les travaux de Norman [Nor75a ; Nor74a ; Nor75b].) Le processus $(Z_t, t \geq 0)$ est donc un processus d'Ornstein-Uhlenbeck, dont la loi stationnaire est une gaussienne centrée de variance

$$\frac{1}{4N(s_1+s_2)}.$$

D'après (1.36), on obtient

$$\mathbb{E}[K] = 1 - \frac{s_1 s_2}{s_1 + s_2} - \frac{1}{4N}$$

et on retrouve ainsi le résultat de Robertson [Rob70].

Avec Sarah Penington [FP17], nous avons adapté ce résultat à une population structurée spatialement. Nous présentons dans le Chapitre 2 comment adapter la définition du SLFVS (Définition 1.4.1) pour permettre la sélection naturelle sur des individus diploïdes. Nous définissons alors la fitness moyenne locale $K(t,x)$ comme la fitness moyenne d'un individu dont les deux parents se situent dans la boule de centre x et de rayon R, où R est le rayon (supposé fixé) des événements de reproduction. Cette quantité s'écrit alors comme précédemment

$$K(t,x) = 1 - \frac{s_1 s_2}{s_1 + s_2} - (s_1 + s_2)\left(\overline{q}_t(x,R) - \lambda\right)^2,$$

où $\overline{q}_t(x,R)$ désigne la fréquence de l'allèle A_1 dans $B(x,R)$ à l'instant t. Le fardeau de dérive $\Delta(t,x)$ est alors défini par

$$\Delta(t,x) = (s_1 + s_2)\mathbb{E}\left[\left(\overline{q}_t(x,R) - \lambda\right)^2\right]. \tag{1.37}$$

Nous montrons sous des hypothèses précisées au Chapitre 2 (Théorème 2.3) que pour des constantes positives $(C_i, i \geq 1)$ qui ne dépendent que de λ, lorsque u, s_1 et s_2 sont proches de zéro,

$$\text{si } d = 1, \qquad \Delta(t,x) \sim C_1 u\sqrt{s_1 + s_2} \tag{1.38}$$

$$\text{si } d = 2, \qquad \Delta(t,x) \sim C_2 u(s_1 + s_2)\left|\ln(s_1 + s_2)\right| \tag{1.39}$$

$$\text{si } d \geq 3, \qquad \Delta(t,x) \sim C_d u(s_1 + s_2). \tag{1.40}$$

Le paramètre u joue ici le même rôle que $\frac{1}{2N}$ dans la diffusion de Wright-Fisher car il définit la part de la population totale qui est remplacée à chaque reproduction (on peut également comparer (1.21) et (1.41)). On observe alors que le fardeau de dérive est plus faible dans une population structurée que dans une population panmictique, et ce d'autant plus que la dimension de l'espace est grande. Cela est dû au fait si la fréquence de l'allèle A_1 s'éloigne de la fréquence d'équilibre en un point, la fréquence de cet allèle en des points voisins sera en moyenne proche de cet équilibre. La migration va donc empêcher les fréquences alléliques de trop s'éloigner de celui-ci.

Dans le Chapitre 2, nous supposons de plus que les individus peuvent muter d'un allèle vers l'autre avec une certaine probabilité à chaque génération. Cela empêche les fréquences alléliques de trop s'approcher de 1 ou de 0 dans une région donnée, auquel cas elles peuvent rester un long moment loin de la fréquence d'équilibre et l'approximation ci-dessus n'est plus valable.

On peut retrouver (1.38), (1.39), (1.40) à l'aide de l'équation suivante

$$\partial_t q_t = uV_R\left[\frac{R^2}{d+2}\Delta q_t + (s_1 + s_2)q_t(1 - q_t)(\lambda - q_t)\right] + uV_R\sqrt{q_t(1 - q_t)}\dot{W}_t. \tag{1.41}$$

Cette équation n'est bien posée que si $d = 1$, mais elle donne le bon résultat quelle que soit la dimension (le traitement rigoureux à l'aide du SLFVS est donné dans le Chapitre 2). Pour q_t suffisamment proche de λ, on pose

$$Z_t = q_t - \lambda$$

et on écrit, en négligeant les termes quadratiques,

$$\partial_t Z_t = uV_R \left[\frac{R^2}{d+2} \Delta Z_t - (s_1+s_2)\lambda(1-\lambda) Z_t \right] + uV_R \sqrt{\lambda(1-\lambda)} \dot{W}_t.$$

Contrairement à (1.41), cette équation a un sens quelle que soit la dimension de l'espace. Elle possède de plus une solution stationnaire pour laquelle on montre que

$$\left\langle Z_t, \frac{1}{V_R} \mathbb{1}_{\{B(x,R)\}} \right\rangle$$

suit une loi gaussienne centrée dont la variance Σ^2 satisfait

$$\begin{aligned} &\text{si } d = 1, && \Sigma^2 \sim C_1 u (s_1+s_2)^{-1/2}, \\ &\text{si } d = 2, && \Sigma^2 \sim C_2 u \left| \ln(s_1+s_2) \right|, \\ &\text{si } d \geq 3, && \Sigma^2 \sim C_d u, \end{aligned}$$

où C_1, C_2, C_d sont des constantes qui ne dépendent que de d et de λ. D'après (1.37), on retrouve bien (1.38), (1.39), (1.40).

1.5 Modélisation d'une population avec dispersion hétérogène

Tous les modèles évoqués jusqu'ici supposent que les individus se déplacent de la même façon en tous les points de l'espace et dans toutes les directions. Cela ne sera en général pas le cas pour des populations dont l'aire de répartition couvre plusieurs types d'habitat. Si les paramètres démographiques varient continuement dans l'espace, il est possible d'adapter les résultats précédents sans grande difficulté (voir par exemple le début de [Nag88]). En cas de changement abrupt de ces caractéristiques, on observe en revanche des comportements qualitativement différents du cas homogène. Nous nous intéressons ici au cas où les individus se déplacent plus facilement dans une région que dans l'autre.

1.5.1 Le modèle stepping stone avec dispersion hétérogène

Dans une série de travaux, Nagylaki et ses co-auteurs ont proposé et étudié un modèle stepping stone avec cette caractéristique [Nag76 ; Nag78 ; Nag88 ; NB88 ; ADN99]. Dans ce modèle, les individus sont répartis en colonies le long d'un habitat linéaire et les colonies échangent plus d'individus d'un côté de l'origine que de l'autre. Étant donnés m_+ et m_- deux réels positifs, on suppose que les colonies voisines à droite (resp. à gauche) de l'origine échangent $\frac{1}{2}m_+$ (resp. $\frac{1}{2}m_-$) individus à chaque génération, comme représenté sur la Figure 1.4. Nagylaki suppose de plus que toutes les colonies contiennent N individus, avec N suffisamment grand pour pouvoir négliger les effets de la dérive génétique.

Comme les effectifs des différentes colonies sont identiques, on a $m_{ij} = \tilde{m}_{ji}$. Si l'on note $p(t,i)$ la

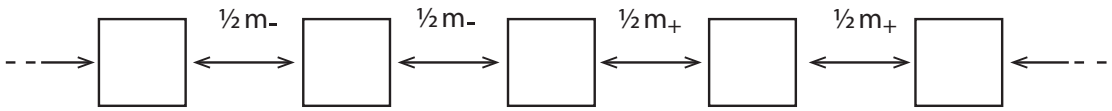

Figure 1.4 : Modèle stepping stone avec dispersion hétérogène

fréquence d'un allèle donné dans le dème i à l'instant t, celle-ci vérifie

$$p(t+1,i) - p(t,i) = \sum_{j \neq i} m_{ji}(p(t,j) - p(t,i)). \tag{1.42}$$

Nagylaki [Nag76] montre qu'en posant

$$p_n(t,x) = p(nt, \sqrt{n}x) \tag{1.43}$$

et en laissant tendre n vers $+\infty$, p_n converge vers p_∞ où

$$\begin{cases} \partial_t p_\infty(t,x) = \dfrac{m_\pm}{2} \partial_{xx} p_\infty(t,x) & \text{si } \pm x \geq 0, \\ p_\infty(t,0^+) = p_\infty(t,0^-) \\ m_+ \partial_x p_\infty(t,0^+) = m_- \partial_x p_\infty(t,0^-). \end{cases} \tag{1.44}$$

En effet, pour $\pm x > 0$, (1.42) implique

$$\partial_t p_n(t,x) = \frac{m_\pm}{2} \partial_{xx} p_n(t,x) + o(1).$$

Pour $x = 0$, on a

$$\partial_t p_n(t,0) = \sqrt{n}\left(m_+ \partial_x p_n(t,0^+) - m_- \partial_x p_n(t,0^-)\right) + o\left(\sqrt{n}\right).$$

Si p_n converge vers p_∞, il faut donc que

$$m_+ \partial_x p_n(t,0^+) - m_- \partial_x p_n(t,0^-) = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right),$$

ce qui donne bien (1.44). Cette équation peut s'interpréter comme une équation de conservation du flux de gène, analogue à la loi de Fick en conduction thermique.

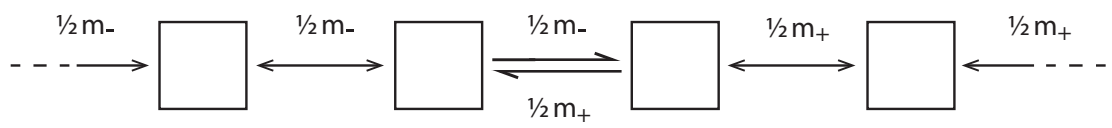

Figure 1.5 : Modèle stepping stone avec migration isotropes

Nagylaki remarque cependant dans un autre article [Nag88] que cette équation est sensible aux détails du modèle microscopique. En effet, si l'on considère un schéma de migration (que Nagylaki nomme isotrope) comme sur la Figure 1.5, on a à la place

$$m_+^2 \partial_x p_\infty(t, 0^+) = m_-^2 p_\infty(t, 0^-). \tag{1.45}$$

Cela peut se retrouver en écrivant (1.42) pour $i = \pm 1$. On a alors, après avoir posé (1.43),

$$\begin{aligned}
\partial_t p_n(t, \tfrac{1}{\sqrt{n}}) &= n\left(m_-(p_n(t, -\tfrac{1}{\sqrt{n}}) - p_n(t, \tfrac{1}{\sqrt{n}})) + m_+ \tfrac{1}{\sqrt{n}} \partial_x p_n(t, \tfrac{1}{\sqrt{n}})\right) + o\left(\sqrt{n}\right) \\
\partial_t p_n(t, -\tfrac{1}{\sqrt{n}}) &= n\left(m_+(p_n(t, \tfrac{1}{\sqrt{n}}) - p_n(t, -\tfrac{1}{\sqrt{n}})) - m_- \tfrac{1}{\sqrt{n}} \partial_x p_n(t, -\tfrac{1}{\sqrt{n}})\right) + o\left(\sqrt{n}\right).
\end{aligned}$$

En multipliant la première ligne par m_+, la seconde par m_- et en additionnant le tout, on obtient

$$m_+ \partial_t p_n(t, 0^+) + m_- \partial_t p_n(t, 0^-) = \sqrt{n}\left(m_+^2 \partial_x p_n(t, 0^+) - m_-^2 \partial_x p_n(t, 0^-)\right) + o\left(\sqrt{n}\right).$$

D'où (1.45) en passant à la limite lorsque $n \to \infty$. Nagylaki a également étudié une version de ce modèle dans laquelle chaque individu a une probabilité μ de muter à chaque génération. On suppose que chaque mutation produit un allèle nouveau, qui n'est porté par aucun autre individu dans la population. On définit alors la probabilité que deux individus pris au hasard dans les dèmes i et j portent le même allèle, c'est-à-dire qu'aucune mutation ne soit intervenue depuis leur dernier ancêtre commun. On appelle cette quantité la probabilité d'identité par descendance et on la note $F(i, j)$.

Si $\mu = \frac{\bar{\mu}}{n}$ et si l'on pose

$$F_n(x, y) = F(\sqrt{n}x, \sqrt{n}y),$$

alors Nagylaki [Nag88] montre que F_n converge vers $F_\infty : \mathbb{R}^2 \to [0, 1]$ qui vérifie

$$2\bar{\mu} F_\infty(x, y) = \frac{m(x)}{2} \partial_{xx} F_\infty(x, y) + \frac{m(y)}{2} \partial_{yy} F_\infty(x, y) + \frac{1}{N} \delta_0(x - y)(1 - F_\infty(x, y)), \tag{1.46}$$

où $m(x) = m_\pm$ si $\pm x > 0$. Il montre de plus que F_∞ est continue sur $\mathbb{R}^2$ et que, dans le cas du modèle de la Figure 1.4 (cas dit symétrique), on a, pour tout $x, y \in \mathbb{R}$,

$$\begin{cases}
m_+ \partial_x F_\infty(0^+, y) = m_- \partial_x F_\infty(0^-, y) \\
m_+ \partial_x F_\infty(x, 0^+) = m_- \partial_y F_\infty(x, 0^-).
\end{cases}$$

Dans le cas isotropique (celui de la Figure 1.5), on a en revanche

$$\begin{cases}
m_+^2 \partial_x F_\infty(0^+, y) = m_-^2 \partial_x F_\infty(0^-, y) \\
m_+^2 \partial_x F_\infty(x, 0^+) = m_-^2 \partial_y F_\infty(x, 0^-).
\end{cases}$$

Dans la suite de ses travaux [NB88 ; ADN99], Nagylaki étudie les conséquences d'une telle configuration sur l'existence de polymorphismes en présence ou non de sélection naturelle. On remarque que (1.46) généralise l'équation (1.15) au cas de la dispersion hétérogène.

1.5.2 Dispersion hétérogène en espace continu

Le raisonnement suivi par Nagylaki utilise les détails du modèle et du mode de migration pour obtenir le comportement à grande échelle des fréquences alléliques. On peut donc se demander si le même comportement à grande échelle peut être observé sous des hypothèses plus générales. Qu'advient-il en particulier si la transition entre les deux régions de l'espace se fait sur un nombre arbitraire de dèmes au lieu de un ou deux comme dans les Figures 1.4 et 1.5 ? Quel comportement observe-t-on en dimensions supérieures ? Enfin le résultat reste-t-il vrai si la population vit dans un espace continu ?

Pour répondre à ces questions, nous définissons dans le Chapitre 3 le SLFV avec dispersion hétérogène, ainsi que son dual décrivant la généalogie d'un échantillon d'individus dans la population. Le SLFV avec dispersion hétérogène est défini en modifiant la Définition 1.3.4 pour que le rayon des événements de reproduction dépende de la position de leur centre, comme représenté Figure 1.6.

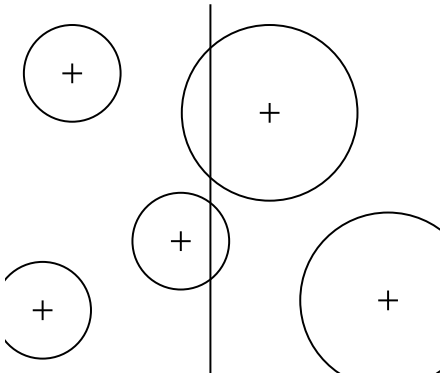

Figure 1.6 : Taille des événements de reproduction

Le rayon d'un événement de reproduction dépend du demi espace dans lequel son centre se situe ($x_1 > 0$ ou $x_1 < 0$).

L'espace $\mathbb{R}^d$ est ainsi divisé en deux demi espaces $\mathbb{H}^+$ et $\mathbb{H}^-$ tels que

$$\mathbb{H}^{\pm} = \{x \in \mathbb{R}^d : \pm x_1 > 0\}.$$

On suppose alors qu'un événement de reproduction dont le centre se trouve dans $\mathbb{H}^+$ (resp. dans $\mathbb{H}^-$) affecte une boule de rayon r_+ (resp. r_-) avec $0 < r_- < r_+$ pour fixer les idées. Ainsi, les individus situés dans le demi espace $\mathbb{H}^+$ peuvent produire des descendants dans une région plus grande que ceux vivants dans $\mathbb{H}^-$. On note que la probabilité d'être touché par un événement de reproduction (et donc de mourrir) dépend de la taille de ces derniers, or il semble raisonnable de supposer que tous les individus d'une même population ont la même espérance de vie. Il faut donc ajuster l'intensité des événements de reproduction de chaque côté pour compenser cette différence.

Soient $u \in (0, 1]$ et $0 < r_- < r_+ < +\infty$ fixés, on note comme dans la Section 1.3 $w_t(x)$ la proportion d'individus de type 1 en x à l'instant t.

Définition 1.5.1 (SLFV avec dispersion hétérogène)**.** *Soit Π^+ (resp. Π^-) un processus ponctuel de Poisson sur $\mathbb{R}_+ \times \mathbb{H}^+$ (resp. sur $\mathbb{R}_+ \times \mathbb{H}^-$) de mesure d'intensité $\frac{1}{V_{r_+}} dt \otimes dx$ (resp. $\frac{1}{V_{r_-}} dt \otimes dx$). Pour chaque point $(t, x) \in \Pi^{\pm}$, un événement de reproduction se produit dans la boule $B(x, r_{\pm})$ à l'instant t. On choisit alors un point y uniformément au hasard dans $B(x, r_{\pm})$ et on choisit un type $k \in \{0, 1\}$ tel que $k = 1$ avec probabilité $w_{t^-}(y)$ et $k = 0$ sinon. On pose alors pour $z \in B(x, r_{\pm})$,*

$$w_t(z) = (1 - u)w_{t^-}(z) + uk,$$

et on ne change pas w en dehors de $B(x, r_{\pm})$.

En adaptant le Théorème 4.2 de [BEV10b], on montre qu'il existe un unique processus de Markov $(w_t)_{t \geq 0}$ càdlàg à valeurs dans Ξ qui satisfait cette définition.

On définit le dual du SLFV avec dispersion hétérogène de manière analogue, en reprenant la Définition 1.3.5 et en remplaçant Π par le couple (Π^+, Π^-) (voir la Définition 3.3.1). Celui-ci vérifie la même relation de dualité que le SLFV en espace homogène, donnée par l'équation (1.20).

Le déplacement d'une lignée ancestrale dans le dual du SLFV avec dispersion hétérogène suit une marche aléatoire sur $\mathbb{R}^d$, qui se comporte comme une marche aléatoire simple de variance

$$\sigma_\pm^2 = u\frac{2r_\pm^2}{d+2} \tag{1.47}$$

dès qu'elle se trouve dans $\{x \in \mathbb{R}^d : \pm x_1 > r_+\}$ (car dans ce cas cette lignée ne peut être affectée par des événements de $\Pi^\mp$). Si l'on note ξ_t la position d'une lignée ancestrale t unités de temps dans le passé et que l'on pose, pour $n \geq 1$,

$$\xi_t^n = \frac{1}{\sqrt{n}}\xi_{nt}, \tag{1.48}$$

on s'attend donc à ce que, pour n assez grand, $(\xi_t^n)_{t\geq 0}$ se comporte comme un mouvement brownien de variance $\sigma_\pm^2$ dans le demi espace $\mathbb{H}^\pm$. Le comportement de ξ^n à l'interface est quant à lui déterminé par la loi de ξ lorsqu'il se trouve dans $\{x \in \mathbb{R}^d : x_1 \in [-r_+, r_+]\}$.

Comme $r_+ > r_-$, les individus situés dans cette région sont plus souvent affectés par des événements de reproduction dont le centre se situe dans $\mathbb{H}^+$. Une plus grande partie de leurs ancêtres vient donc de ce côté, et on s'attend à ce que ξ^n fasse plus d'excursions dans $\mathbb{H}^+$ que dans $\mathbb{H}^-$. Pour montrer et quantifier cet effet, nous devons faire appel à la théorie des temps locaux et définir le mouvement brownien de Walsh (*skew Brownian motion* en anglais).

1.5.3 Temps local d'un processus de Markov de trajectoires continues

Il est bien connu que pour un mouvement brownien $(B_t)_{t\geq 0}$, la mesure de Lebesgue de $\{t \in \mathbb{R}_+ : B_t = 0\}$ est nulle [RY13]. Il existe cependant une manière de mesurer le temps que B passe en un point donné grâce à la théorie des temps locaux.

La définition usuelle du temps local en zéro de B est

$$L_t^0(B) = \lim_{\varepsilon \downarrow 0} \frac{1}{2\varepsilon} \left|\{s \in [0,t] : |B_s| \leq \varepsilon\}\right|, \tag{1.49}$$

où $|A|$ désigne la mesure de Lebesgue de $A \subset \mathbb{R}$. Nous choisissons ici une approche différente qui, quoique moins classique, sera plus commode pour la suite. Une description plus complète de la théorie des temps locaux est disponible dans [RY13].

Commençons par énoncer le Lemme suivant, dû à Skorokhod [Sko61]. Pour $x \in \mathbb{R}$ on pose

$$\mathrm{sign}(x) = \mathbb{1}_{\{x>0\}} - \mathbb{1}_{\{x<0\}}, \qquad \text{et} \qquad x^\pm = \max(\pm x, 0).$$

Lemme 1.5.2 ([Sko61])**.** *Soit $f : \mathbb{R}_+ \to \mathbb{R}$ une fonction continue telle que $f(0) \leq 0$. Il existe une unique fonction $l : \mathbb{R}_+ \to \mathbb{R}$ continue telle que*

i) $X(t) := l(t) - f(t)$ est positif ou nul pour tout $t \geq 0$,

ii) $l(0) = 0$ et $t \mapsto l(t)$ est croissante,

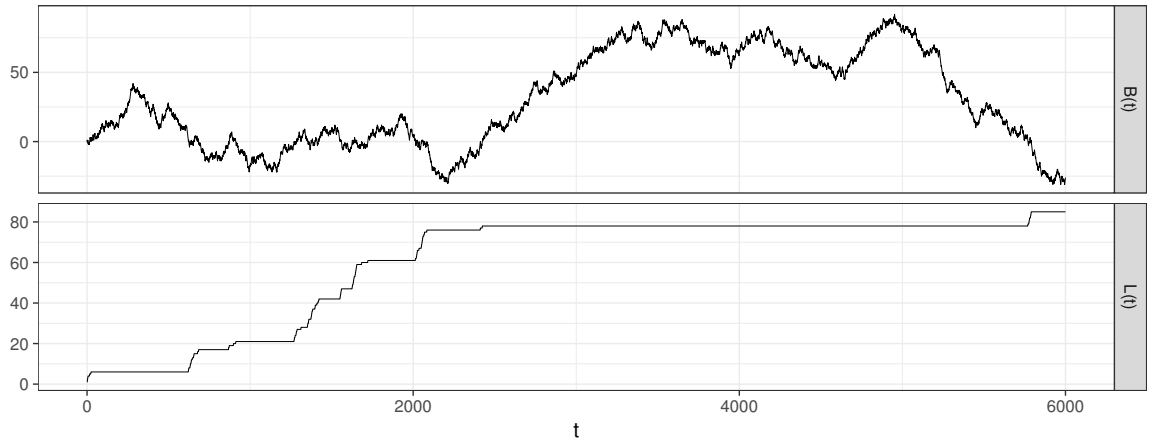

Figure 1.7 : Temps local du mouvement brownien
Mouvement brownien standard représenté avec son temps local en 0.

iii) $\int_0^\infty \mathbb{1}_{\{X(s)>0\}} dl(s) = 0.$

On dit alors que l est solution du problème de Skorokhod pour f. Cette fonction est donnée par

$$l(t) = \max_{0\leq s\leq t} (f(s))^+ .$$

Soit $(X(t), t \geq 0)$ un processus de Markov à valeurs réelles tel que $t \mapsto X(t)$ soit presque sûrement continue. On définit alors le temps local symétrique de X en un point x comme suit (on omettra parfois l'adjectif symétrique).

Définition 1.5.3 (Temps local symétrique)**.** *On appelle temps local symétrique de X en 0, que l'on note $(L_t^0(X), t \geq 0)$, la solution du problème de Skorokhod pour*

$$f(t) = -\int_0^t \text{sign}(X(s))dX(s).$$

Pour $x \in \mathbb{R}$, on pose $L_t^x(X) = L_t^0(X - x)$.

Une réalisation du mouvement brownien standard avec son temps local est donné Figure 1.7. On définit également le temps local à gauche et à droite d'un point x de la manière suivante.

Définition 1.5.4 (Temps local à gauche / à droite)**.** *On appelle temps local à gauche (resp. à droite) de X en 0, que l'on note $L_t^{0-}(X)$ (resp. $L_t^{0+}(X)$), la solution du problème de Skorokhod pour*

$$f(t) = 2\int_0^t \mathbb{1}_{\{X(s)<0\}} dX(s)$$

et respectivement

$$f(t) = -2\int_0^t \mathbb{1}_{\{X(s)>0\}} dX(s).$$

Pour $x \in \mathbb{R}$, on pose $L_t^{x\pm}(X) = L_t^{0\pm}(X - x)$.

On a alors les formules suivantes, dues à Tanaka [RY13],

$$|X(t)| = L_t^0(X) + \int_0^t \text{sign}(X(s))dX(s) \tag{1.50}$$

$$(X(t))^{\pm} = \frac{1}{2}L_t^{0\pm}(X) \pm \int_0^t \mathbb{1}_{\{\pm X(s)>0\}}dX(s). \tag{1.51}$$

On en déduit

$$L_t^0(X) = \frac{1}{2}\left(L_t^{0+}(X) + L_t^{0-}(X)\right),$$

ce qui justifie l'adjectif symétrique pour $L_t^0(X)$. On peut interpréter (1.51) en disant que $L_t^{0+}(X)$ (resp. $L_t^{0-}(X)$) "compte" le nombre de fois que X croise l'origine vers la droite (resp. vers la gauche). On retrouve la définition usuelle (1.49) grâce à la formule de densité d'occupation [RY13], pour $f : \mathbb{R} \to \mathbb{R}$ mesurable,

$$\int_0^t f(X(s))d\langle X\rangle_s = \int_{\mathbb{R}} f(x)L_t^x(X)dx \text{ p.s.} \tag{1.52}$$

On note que, pour le mouvement brownien, le temps local à gauche est égal au temps local à droite en tout point de l'espace et pour tout $t \geq 0$.

1.5.4 Le mouvement brownien de Walsh

Le mouvement brownien de Walsh est un processus de Markov à valeurs réelles qui ressemble à un mouvement brownien standard avec un comportement singulier à l'origine : il quitte l'origine d'un côté plus souvent que de l'autre. Comme en réalité le mouvement brownien atteint l'origine en une quantité indénombrable de points, cette description n'a pas de sens mathématique.

Le mouvement brownien de Walsh (*skew Brownian motion* en anglais) a été introduit par Walsh [Wal78] et Itô et McKean [IM63] puis repris par Harrison et Shepp [HS81]. De nombreux travaux ont suivi dont on trouvera un aperçu dans [Lej06].

Soit $(B_t)_{t\geq 0}$ un mouvement brownien standard. Dans [HS81], Harrison et Shepp montrent qu'il existe un unique processus de Markov $(X(t), t \geq 0)$ à valeurs réelles tel que

$$X(t) = B_t + \beta L_t^0(X)$$

si et seulement si $\beta \in [-1, 1]$. Un tel processus est un mouvement brownien de Walsh de paramètre β.

Walsh montre que $(X(t), t \geq 0)$ est solution du problème de martingales associé à l'opérateur $\mathcal{L}^\beta$ défini sur l'ensemble des fonctions $f : \mathbb{R} \to \mathbb{R}$ continues sur $\mathbb{R}$ et deux fois continuement dérivables sur $\mathbb{R} \setminus \{0\}$ telles que

$$(1 + \beta)f'(0^+) = (1 - \beta)f'(0^-),$$

où $\mathcal{L}^\beta f = \frac{1}{2}f''$.

Harrison et Shepp obtiennent $(X(t), t \geq 0)$ comme limite de marches aléatoires $(X_k^\beta, k \geq 0)$ sur $\mathbb{Z}$

telles que

$$\mathbb{P}\left(X^\beta_{k+1} = x \pm 1 \;\middle|\; X^\beta_k = x\right) = \frac{1}{2} \qquad \text{si } x \neq 0,$$
$$\mathbb{P}\left(X^\beta_{k+1} = \pm 1 \;\middle|\; X^\beta_k = 0\right) = \frac{1 \pm \beta}{2}.$$

Si l'on pose

$$X^n(t) = \frac{1}{\sqrt{n}} X^\beta_{\lfloor nt \rfloor},$$

alors X^n converge en loi vers X lorsque $n \to \infty$. Cette convergence donne un sens à la description donnée plus haut selon laquelle X quitte l'origine d'un côté plus souvent que de l'autre.

Ce résultat a été étendu par Iksanov et Pilipenko [IP16] à une famille plus large de marches aléatoires qui sont symétriques en dehors d'une région de taille finie autour de l'origine. Dans le Chapitre 3, nous étendons leurs arguments aux trajectoires suivies par les lignées ancestales dans le dual du SLFV avec dispersion hétérogène. Nous montrons que la suite de processus $(\xi^n, n \geq 1)$ définie par (1.48) converge lorsque $n \to \infty$ vers la solution du système d'équations différentielles stochastiques suivant

$$\begin{cases} X^1_t = \displaystyle\int_0^t \sigma(X^1_s) dB^1_s + \beta L^0_t(X^1) \\ X^i_t = \displaystyle\int_0^t \sigma(X^1_s) dB^i_s \qquad\qquad 2 \leq i \leq d. \end{cases} \tag{1.53}$$

où $\sigma(x) = \sigma_\pm \mathbb{1}_{\{\pm x > 0\}}$ et $B = (B^1, \ldots, B^d)$ est un mouvement brownien standard d-dimensionnel. Ce processus suit donc un mouvement brownien de variance $\sigma^2_\pm$ (donnée par (1.47)) dans chaque demi espace $\mathbb{H}^\pm$ et X^1 suit un mouvement brownien de Walsh sur $\mathbb{R}$.

L'existence et l'unicité d'un tel X^1 a été montrée par Le Gall [LG84a] et l'existence et l'unicité de X^i pour $i \geq 2$ en découle (voir également [Por79a]).

1.5.5 Comportement à grande échelle du SLFV avec dispersion hétérogène

Une fois connu le comportement à grande échelle d'une lignée ancestrale, on peut en déduire celui du coalescent spatial correspondant, puis par dualité du SLFV avec dispersion hétérogène. Si

$$\mathcal{A}_t = \left\{\xi^1_t, \ldots, \xi^{N_t}_t\right\}$$

désigne le dual du SLFV avec dispersion hétérogène (voir Définition 3.3.1), on pose pour $n \geq 1$,

$$\mathcal{A}^n_t = \left\{\frac{1}{\sqrt{n}}\xi^1_{nt}, \ldots, \frac{1}{\sqrt{n}}\xi^{N_{nt}}_{nt}\right\}.$$

Entre les événements de coalescence, la trajectoire suivie par chaque $\frac{1}{\sqrt{n}}\xi^i_{nt}$ s'approche d'un mouvement brownien de Walsh solution de (1.53). En reprenant les arguments de [BEV13b] pour la preuve du Théorème 1.1 cité ci-dessus, on montre que $\mathcal{A}^n$ converge au sens des lois fini-dimensionnelles vers un processus $(\mathcal{A}^\infty_t, t \geq 0)$. En dimension 1, $\mathcal{A}^\infty$ est un système de solutions de (1.53) qui coalescent dès qu'elles se rencontrent et en dimensions supérieures, $\mathcal{A}^\infty$ est un système de solutions de (1.53)

indépendantes (voir le Théorème 3.2 dans le Chapitre 3).

Grâce à la relation de dualité (1.20), on peut en déduire le comportement à grande échelle du SLFV avec dispersion hétérogène. Soit $(w_t)_{t\geq 0}$ le processus de la Définition 1.5.1, on pose pour $n \geq 1$, $t \geq 0$ et $x \in \mathbb{R}^d$,

$$w^n(t,x) = w_{nt}(\sqrt{n}x)$$

et on suppose que $w^n(0,x) = w_0(x)$ pour tout $n \geq 1$. Soit $\rho : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ la solution du système

$$\begin{cases} \partial_t \rho(t,x) = \dfrac{\sigma(x)^2}{2} \Delta\rho(t,x), \qquad \rho(0,x) = w_0(x), \\ (1-\beta) \left.\dfrac{\partial \rho(t,x)}{\partial x_1}\right|_{x_1=0^-} = (1+\beta) \left.\dfrac{\partial \rho(t,x)}{\partial x_1}\right|_{x_1=0^+} \end{cases} \tag{1.54}$$

Théorème 1.5 (Comportement à grande échelle du SLFV avec dispersion hétérogène)**.** *Lorsque $n \to \infty$, w^n converge au sens des lois fini dimensionnelles vers un processus $(p(t,\cdot), t \geq 0)$ à valeurs dans Ξ. Ce processus est le dual de $\mathcal{A}^\infty$ via la relation* (1.24)*. En dimension 1, $p(t,x)$ est une variable de Bernoulli de paramètre $\rho(t,x)$ et les corrélations entre les valeurs de $p(t,\cdot)$ en différents points sont données par* (1.24)*. En dimensions supérieures, $p(t,x)$ est déterministe et égal à $\rho(t,x)$.*

On retrouve dans l'équation (1.54) le résultat de Nagylaki suivant lequel, en l'absence de dérive génétique, les fréquences alléliques suivent l'équation (1.44). Ce résultat découle de la convergence de $\mathcal{A}^n$ en notant d'après (1.24) que

$$\rho(t,x) = \mathbb{E}_{w_0}\left[p(t,x)\right] = \mathbb{E}_x\left[w_0(X_t)\right],$$

où $(X_t)_{t\geq 0}$ est solution de (1.53). La fonction $\rho(t,\cdot)$ est donc le résultat de l'action du semi-groupe du mouvement brownien de Walsh sur w_0. Elle doit donc vérifier

$$\partial_t \rho = \mathcal{L}^\beta \rho, \qquad\qquad \rho(0,\cdot) = w_0,$$

où $\mathcal{L}^\beta$ est le générateur infinitésimal de $(X_t)_{t\geq 0}$. On en déduit alors (1.54).

Le fait que p soit déterministe dès que $d \geq 2$ et qu'il soit une famille de Bernoullis en dimension 1 découle de la forme de $\mathcal{A}^\infty$ de la même manière que pour le Théorème 1.1 dans le cas homogène.

D'autres régimes peuvent être considérés dans le cadre de la dispersion hétérogène, en particulier ceux des Théorèmes 1.2 et 2.1. Ayant déjà montré que les trajectoires suivies par les lignées ancestrales sont (à la limite) des mouvements browniens de Walsh (équation 1.53), il est facile de conjecturer comment les résultats des Théorèmes 1.2 et 2.1 s'adaptent en présence de dispersion hétérogène. Le rythme des coalescences est différent dans les deux demi-espaces, car la région dans laquelle se trouvent les ancêtres potentiels d'une particule est plus petite d'un côté que de l'autre.

Dans la section suivante, nous présentons un autre cas d'hétérogénéité spatiale qui affecte les trajectoires suivies par les lignées ancestrales, et donc la composition génétique des populations.

1.6 Influence d'une barrière géographique sur la composition génétique d'une population

De nombreuses espèces occupent un habitat fractionné, que cela soit dû à des activités humaines (déforestation, construction d'autoroutes ou de clôtures...) ou bien naturelles (chaînes de montagnes, canyons ou rivières). Un tel obstacle constitue une barrière au flux de gènes s'il diminue significativement les échanges génétiques entre les populations de chaque côté de celui-ci. De nombreuses études cherchent à déterminer si tel ou tel obstacle agit ou non comme une barrière au flux de gènes [Cas+00 ; Ril+06 ; Zal+09]. Ces études s'appuient le plus souvent sur des mesures de différentiation génétique ou bien sur des techniques de clustering qui ne prennent pas en compte la structure spatiale des populations de chaque côté de la barrière.

1.6.1 Le modèle stepping stone avec une barrière au flux de gènes

Nagylaki [Nag76] a proposé un modèle stepping stone présentant une barrière au flux de gènes et a étudié la structure spatiale de la diversité génétique en présence d'un obstacle à la migration. Il considère une population répartie en dèmes sur $\mathbb{Z} \setminus \{0\}$, où deux dèmes voisins échangent en moyenne $\frac{1}{2}m$ individus par génération, à l'exception des dèmes situés en -1 et 1 qui en échangent $\frac{1}{2}cm$, avec $m > 0$ et $c \in (0, 1)$, comme représenté sur la Figure 1.8.

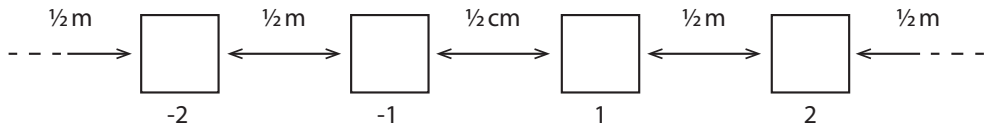

Figure 1.8 : Modèle stepping stone avec une barrière au flux de gènes

De plus, chaque dème contient un nombre N d'individus, supposé grand, de sorte qu'on peut négliger les effets de la dérive génétique. Si $p(t, i)$ désigne la fréquence d'un allèle donné au temps t dans le dème i, on a alors

$$p(t+1, i) - p(t, i) = \frac{m}{2} \left(p(t, i+1) + p(t, i-1) - 2p(t, i) \right) \tag{1.55}$$

pour tout $|i| > 1$ et

$$\begin{aligned} p(t+1, 1) &= \frac{m}{2} \left(p(t, 2) - p(t, 1) + c(p(t, -1) - p(t, 1)) \right) \\ p(t+1, -1) &= \frac{m}{2} \left(p(t, -2) - p(t, -1) + c(p(t, 1) - p(t, -1)) \right). \end{aligned}$$

Pour $n \geq 1$, on pose

$$p_n(t, x) = p(nt, \sqrt{n}x)$$

et on suppose que $\sqrt{n}c_n \to \gamma \in (0, \infty)$ lorsque $n \to \infty$. Nagylaki montre alors que p_n converge lorsque

$n \to \infty$ vers la solution de l'EDP suivante

$$\begin{cases} \partial_t p(t,x) = \dfrac{m}{2}\partial_{xx} p(t,x), & x \in \mathbb{R} \setminus \{0\} \\ \partial_x p(t,0^+) = \partial_x p(t,0^-) = \gamma(p(t,0^+) - p(t,0^-)). \end{cases} \tag{1.56}$$

La première partie de cette équation s'obtient par un passage à la limite dans (1.55) grâce à la formule de Taylor, comme dans la section précédente. Pour la condition au bord en 0, on écrit pour p_n

$$\partial_t p_n(t,0^\pm) = \pm\sqrt{n}\frac{m}{2}\left(\partial_x p_n(t,0^\pm) - \sqrt{n}c_n(p_n(t,0^+) - p_n(t,0^-))\right) + o\left(\sqrt{n}\right).$$

Si p_n converge vers p_∞, et $\sqrt{n}c_n \to \gamma$, on doit donc avoir

$$\partial_x p_\infty(t,0^\pm) = \gamma(p_\infty(t,0^+) - p_\infty(t,0^-)).$$

Le paramètre γ s'interprète comme une mesure de la perméabilité de la barrière. Si $\gamma \to \infty$, les fréquences alléliques sont continues en 0 et la barrière n'a plus d'effet sur la population alors que si $\gamma \to 0$, les fréquences alléliques évoluent indépendemment sur les deux demi droites avec une condition au bord de Neumann.

Nagylaki étudie également la probabilité d'identité par descendance dans ce modèle (voir la sous-section 1.3.2) [Nag88 ; NKD93]. En supposant que chaque individu mute vers un allèle inédit dans la population à chaque génération avec probabilité $\mu > 0$, on note $F(i,j)$ la probabilité qu'un individu en i soit du même type qu'un individu en j. Si on suppose que $\mu = \frac{\bar{\mu}}{n}$ pour $n \geq 1$, on pose

$$F_n(x,y) = F(\sqrt{n}x, \sqrt{n}y).$$

Alors Nagylaki [Nag88] montre que pour N assez grand, F_n s'approche de F_∞ lorsque $n \to \infty$, avec

$$\begin{cases} 2\mu F_\infty = \dfrac{m}{2}\partial_{xx}F_\infty + \dfrac{m}{2}\partial_{yy}F_\infty + \dfrac{1}{N}(1 - F_\infty)\delta_{xy}, \\ \partial_x F_\infty(0^+,y) = \partial_x F_\infty(0^-,y) = \gamma(F_\infty(0^+,y) - F_\infty(0^-,y)) \\ \partial_y F_\infty(x,0^+) = \partial_x F_\infty(x,0^-) = \gamma(F_\infty(x,0^+) - F_\infty(x,0^-)). \end{cases}$$

La solution de cette équation est également étudiée dans [NKD93] et Barton [Bar08] en donne une solution approchée en dimensions 1 et 2. Ce modèle stepping stone avec une barrière au flux de gènes est repris dans des travaux plus récents [NZ16 ; Nag16] où Nagylaki étudie les conditions pour le maintien d'un polymorphisme en présence de panmixie partielle.

1.6.2 Marches aléatoires avec obstacles

Comme pour le modèle de Kimura de la Définition 1.3.2, on peut définir un coalescent structuré en présence d'une barrière génétique. Étant donnés $m > 0$ et $c \in (0,1)$, on note m_{ij} le taux de migration entre les dèmes i et j dans la Figure 1.8. On fixe de plus $N > 0$. Le coalescent structuré avec barrière génétique est alors défini en suivant la Définition 1.3.3, chaque lignée ancestrale y suit une marche aléatoire sur $\mathbb{Z} \setminus \{0\}$ dont les taux de sauts sont donnés par la Figure 1.8.

Soit $(c_n)_{n\geq 1}$ une suite de réels dans $(0,1)$ telle que $\sqrt{n}c_n \to \gamma \in (0,\infty)$ lorsque $n \to \infty$ et soit

$\left(x_n^0\right)_{n\geq 1}$ une suite de $\mathbb{Z}\setminus\{0\}$. Soit alors $(\xi_t^n)_{t\geq 0}$ une marche aléatoire sur $\mathbb{Z}^*$ dont les taux de sauts sont donnés par les m_{ij} avec $c=c_n$ et $\xi_0^n = x_n^0$. Pour $n\geq 1$ on pose

$$X_n(t) = \frac{1}{\sqrt{n}}\xi_{nt}^n \tag{1.57}$$

et on cherche les limites possibles de la suite des processus $(X_n, n\geq 1)$. Dans le Chapitre 4, on montre le résultat suivant.

Théorème 1.6 (Limite d'échelle de marches aléatoires avec obstacles)**.** *On suppose que $\sqrt{n}x_n^0 \to x_0 \in (0,\infty)$ lorsque $n\to\infty$. Alors pour tout $T>0$ fixé, la suite $(X_n, n\geq 1)$ converge en loi dans $\mathrm{D}([0,T],\mathbb{R})$ vers un processus à valeurs réelles $(X_t)_{t\geq 0}$. Ce dernier se comporte comme un mouvement brownien réfléchi en 0 jusqu'à ce que son temps local à l'origine atteigne une variable exponentielle de paramètre γ, après quoi $(X_t)_{t\geq 0}$ se comporte comme un mouvement brownien réfléchi de l'autre côté de l'origine, jusqu'à ce que son temps local en 0 atteigne une autre variable exponentielle, etc.*

Nous appelons le processus $(X_t)_{t\geq 0}$ le mouvement brownien partiellement réfléchi. Il apparaît dans [MP16] comme limite d'une suite de diffusions dans $\mathbb{R}$ dont le terme de dérive s'approche de $+\infty$ en 0^+ et de $-\infty$ en 0^-.

Tant qu'elle ne traverse pas $\{-1,1\}$, la marche aléatoire $(\xi_t)_{t\geq 0}$ se comporte comme une marche aléatoire simple refléchie en ± 1, il est donc naturel qu'après le changement d'échelle (1.57), X_n converge vers un processus se comportant comme un mouvement brownien réfléchi. On note de plus que par la propriété de Markov, le nombre de visites de ξ en $\{-1,1\}$ entre deux traversées est une variable géométrique de paramètre

$$\frac{c_n}{1+2c_n} \sim \frac{\gamma}{\sqrt{n}}.$$

Divisé par $\sqrt{n}$, ce nombre de visites converge en loi vers le temps local accumulé en 0 entre deux traversées de l'origine par $(X_t)_{t\geq 0}$, qui est donc une variable exponentielle de paramètre γ.

1.6.3 Construction du mouvement brownien partiellement réfléchi

Nous montrons de plus dans le Chapitre 4 que le mouvement brownien partiellement réfléchi s'obtient à partir des excursions d'un mouvement brownien standard en dehors d'une région de largeur $\frac{1}{\gamma}$ (voir la Figure 4.3). Pour $\gamma\in(0,\infty)$, on pose

$$r(x) = \left(x-\tfrac{1}{2\gamma}\right)^+ - \left(x+\tfrac{1}{2\gamma}\right)^-.$$

Soit $(B_t)_{t\geq 0}$ un mouvement brownien et soit $\tau(t)$ le temps d'arrêt défini par

$$\tau(t) = \inf\left\{\tau>0 : \int_0^\tau \mathbb{1}_{\left\{|B_s|>\frac{1}{2\gamma}\right\}}ds > t\right\}.$$

Alors $X_t = r(B_{\tau(t)})$ définit un processus $(X_t)_{t\geq 0}$ qui n'est autre que le mouvement brownien partiellement réfléchi. Ce résultat est une conséquence d'un Théorème de Ray-Knight [SK91, Théorème 6.4.7] selon lequel le temps local à droite en $\frac{1}{2\gamma}$ accumulé avant que B n'atteigne $-\frac{1}{2\gamma}$ est une variable exponentielle de paramètre γ.

Cette construction permet d'obtenir une caractérisation du mouvement brownien partiellement réfléchi comme la solution d'un problème de martingales. Le processus $(X_t)_{t\geq 0}$ défini plus haut n'est pas markovien sur $\mathbb{R}$, mais il le devient si l'on considère qu'il est à valeurs dans $\ddot{\mathbb{R}} = (-\infty, 0^-] \cup [0^+, +\infty)$. Soit alors $\mathcal{D}^\gamma$ l'espace des fonctions $f : \ddot{\mathbb{R}} \to \mathbb{R}$ bornées et deux fois continuement dérivables sur chaque demi droite telles que

$$\partial_x f(0^+) = \partial_x f(0^-) = \gamma(f(0^+) - f(0^-)).$$

Le mouvement brownien partiellement réfléchi est alors caractérisé comme l'unique processus de Markov càdlàg à valeurs dans $\ddot{\mathbb{R}}$ tel que pour toute fonction $f \in \mathcal{D}^\gamma$ le processus

$$f(X_t) - \frac{\sigma^2}{2}\int_0^t \partial_{xx} f(X_s) ds$$

soit une martingale locale pour la filtration engendrée par $(X_t)_{t\geq 0}$.

Pour finir, nous donnons dans le Chapitre 4 une formule explicite pour les densités de transition de ce processus. Posons

$$G_t(x) = \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp\left(-\frac{x^2}{2\sigma^2 t}\right).$$

Si $f : \ddot{\mathbb{R}} \to \mathbb{R}$ est continue et bornée alors pour $t > 0$ et $x \in \ddot{\mathbb{R}}$,

$$\mathbb{E}_x\left[f(X_t)\right] = \int_{\ddot{\mathbb{R}}} f(y) g_t(x, y) dy$$

où

$$g_t(x,y) = \begin{cases} G_t(x-y) + G_t(x+y) - 2\gamma \displaystyle\int_0^\infty e^{-2\gamma l} G_t(|x| + |y| + l) dl & \text{si } xy \geq 0 \\ 2\gamma \displaystyle\int_0^\infty e^{-2\gamma l} G_t(|x| + |y| + l) dl & \text{si } xy \leq 0. \end{cases} \tag{1.58}$$

Cette formule apparaît dans [GVNL14] pour modéliser la diffusion à travers une surface poreuse, mais sans mention du processus $(X_t)_{t\geq 0}$ (voir aussi [Nov+11]).

1.6.4 Flux de gênes à travers une barrière

La convergence de la suite de marches aléatoires du Théorème 1.6 permet de retrouver le résultat de Nagylaki [Nag76]. En effet, en l'absence de dérive génétique, la relation de dualité (1.10) devient

$$p(t,x) = \mathbb{E}_x\left[p_0(\xi_t)\right].$$

En passant à la limite après le changement d'échelle (1.57), on obtient

$$p_\infty(t,x) = \mathbb{E}_x\left[p_0(X_t)\right],$$

où $(X_t)_{t\geq 0}$ est le mouvement brownien partiellement réfléchi. Or le semi groupe engendré par ce dernier est donné par la solution de (1.56).

On peut également considérer d'autres régimes, pour lesquels on doit prendre en compte la dérive génétique. En particulier, le Théorème 1.1 déjà adapté au cas de la dispersion hétérogène dans le Chapitre 3 est directement adaptable au cas d'une barrière génétique. On peut également chercher à obtenir un résultat analogue au Théorème 1.2 [EVY14] pour lequel on doit donner un sens à l'équation (1.21) lorsque le Laplacien est remplacé par le générateur du mouvement brownien partiellement réfléchi.

Une autre perspective serait d'étendre ces résultats à des modèles en espace continu comme le SLFV. Cela peut être fait en considérant que les événements de reproduction qui affectent une région donnée autour de l'origine ont un rayon plus faible que dans le reste de l'espace. Plusieurs difficultés surviennent dans ce cas car les différentes excursions des lignées ancestrales hors de cette région ne sont plus indépendantes et la probabilité de sortir d'un côté ou de l'autre de la barrière dépend du point par lequel on y entre.

Mais l'application la plus prometteuse de ces résultats est la détection de barrières au flux de gènes dans des populations qui présentent une forte structure spatiale. La formule (1.58) permet de calculer efficacement certaines fonctionnelles du mouvement brownien partiellement réfléchi, et ainsi d'estimer la perméabilité de la barrière γ. Cela a été fait récemment par Ringbauer et al. [Rin+17] pour tester la présence d'une barrière au flux de gène au niveau d'une zone hybride chez *Antirrhinum majus* dans les Pyrénées.

1.7 Inférence des paramètres démographiques à l'aide de la recombinaison

La plupart des modèles discutés plus haut font l'hypothèse que les individus sont haploides et se reproduisent de manière asexuée. Pour appliquer ces résultats à l'espèce humaine (ou à d'autres espèces de mammifères par exemple), il faut préciser comment ils s'adaptent dans les cas où les individus sont diploïdes (où chaque individu porte deux copies de chaque chromosome, et donc de chaque gène). En général, on se contente de remplacer la taille de la population N par le double du nombre d'individus dans la population, soit $2N$. En effet, si N individus se reproduisent, ce sont $2N$ copies de chaque gène qui évoluent, et les autres propriétés des modèles considérés sont conservées.

Si la population est divisée en deux sexes qui ne peuvent se reproduire qu'avec des individus de l'autre sexe, on peut montrer [Eth11] qu'il suffit alors de remplacer N par

$$\frac{4N_1N_2}{N_1+N_2}$$

où N_1 et N_2 sont les effectifs des deux sexes. Sawyer [Saw77] discute notamment de l'adaptation du résultat de Kimura et Weiss [KW64] au cas d'une population diploïde sexuée.

1.7.1 La recombinaison

Lorsqu'une population se reproduit de manière sexuée, on peut reconstruire la généalogie d'un échantillon d'individus en utilisant la recombinaison. Chaque individu hérite un chromosome de chaque paire de son père et l'autre de sa mère. Ce chromosome, loin d'être une copie de l'un des

deux chromosomes portés par le parent en question, est en réalité une mosaïque formée à partir de ces derniers. Pour obtenir la séquence génétique correspondante, on place côte à côte les deux chromosomes parentaux, et on lit alternativement la séquence de l'un ou de l'autre, en changeant de chromosome en des points pris aléatoirement le long de la séquence. Ces points sont appelés crossovers. Ce mélange des chromosomes parentaux est ce qu'on appelle la recombinaison (on fera attention au fait que les deux chromosomes qui recombinent sont portés par le même parent).

Il s'ensuit que deux frères (non jumeaux) héritent des mosaïques différentes des chromosomes parentaux. Ils partagent donc de longues portions de leur séquence génétique, qui mises bout à bout représentent en moyenne la moitié de leur génome. À leur tour deux cousins partagent une plus faible portion de leur génome (en moyenne un quart) et répartie le long de plus courtes portions continues de chromosome (car ils ont pu recombiner deux fois).

On appelle ces segments continus de génome hérités d'un ancêtre commun via la même lignée (c'est-à-dire sans recombinaison) des blocs d'identité par descendance (blocs d'IBD). Un bloc d'IBD hérité d'un ancêtre commun récent sera en moyenne plus long qu'un bloc hérité d'un ancêtre plus lointain. Si deux individus partagent un ou plusieurs longs blocs d'IBD, cela indique donc qu'ils partagent au moins un ancêtre commun récent.

Si l'on parvient à observer ces blocs d'IBD entre les paires d'individus d'un large échantillon, il est alors possible de reconstruire une partie de l'histoire et de la géographie de la population échantillonnée. Cette approche présente deux intérêts : premièrement elle permet d'accéder à l'histoire récente des populations. Comme la recombinaison agit sur une échelle de temps plus courte que les mutations neutres, le signal lié aux blocs d'IBD provient en majorité de l'histoire récente de la population (typiquement des dernières 50-100 générations). De plus, les méthodes d'inférence fondées sur les corrélations dans les fréquences alléliques (et donc sur la formule de Wright-Malécot [Rou97 ; Bar+13 ; SB89]) ne permettent d'estimer que le produit de la densité de population et du taux de dispersion. Ringbauer et al. [RCB16] ont montré qu'en utilisant les blocs d'IBD, il était possible d'estimer ces deux paramètres séparément (voir aussi [PP13]).

Deux difficultés substantielles apparaissent lorsque l'on tente d'utiliser les blocs d'IBD pour l'inférence démographique. Premièrement, il est impossible en pratique d'observer précisément ces blocs de génome. En effet, certains événements de recombinaison peuvent ne pas laisser de trace dans la séquence génétique si les deux séquences qui recombinent coïncident en plusieurs endroits. De plus, la fréquence des crossovers n'est pas uniforme le long du génome. Il importe donc de connaître précisément les taux de recombinaison le long du génome pour pouvoir utiliser les blocs d'IBD comme une sorte d'horloge génétique. On compte alors la longueur des blocs d'IBD en Morgan (ou en centi-Morgan, cM), où la distance en Morgan entre deux points du génome correspond au nombre moyen de crossovers par génération entre ces deux sites.

Enfin, il reste encore coûteux d'obtenir la séquence génétique complète d'un grand nombre d'individus, et on doit en général se contenter de séquencer un certains nombre de sites bien connus pour leur variabilité au sein de l'espèce étudiée (on parle de SNP - prononcer "snip" - pour *Single Nucleotide Polymorphism* lorsque la variabilité n'affecte qu'une seule base de la séquence génétique). À l'heure actuelle, ces considérations restreignent le champ d'application de ces méthodes à un tout petit nombre d'espèces, dont l'espèce humaine, pour laquelle les taux de recombinaison le long du génome sont connus.

Le second défi est analytique : les processus qui produisent les blocs d'IBD que l'on observe au sein

d'une population sont complexes et la plupart des quantités d'intérêt pour l'inférence des paramètres sont difficiles à exprimer (en particulier le nombre moyen de blocs d'une longueur donnée partagés par deux individus en fonction de la distance géographique les séparant). Pour être en mesure d'estimer ces paramètre, il est nécessaire de trouver de bonnes approximations pour ces quantités qui soient de plus rapides à évaluer numériquement.

1.7.2 Détection des blocs d'IBD

Ralph et Coop [RC13] se sont intéressés au jeu de données appelé *Population Reference Sample* (POPRES) en Europe [Nel+08]. Chaque individu de l'étude, dont la langue maternelle et le pays d'origine ont été consignés, a été génotypé à 500 000 sites (SNPs) le long de son génome. POPRES contient des données relatives à plusieurs milliers d'individus en Europe, toutes collectées à Londres et à Lausanne. Les individus dont les quatre grands-parents ne venaient pas du même pays ont été exclus de l'étude. On obtient ainsi un échantillon de 2 257 individus, regroupés en 40 populations selon leur pays d'origine et leur langue maternelle (voir le tableau 1 dans [RC13]).

Ralph et Coop ont utilisé le programme fastIBD [BB11] pour détecter les blocs d'IBD entre les paires d'individus dans cet échantillon. Ce programme identifie les longs segments de génomes partagés par deux individus qui sont présents à une fréquence très faible dans le reste de l'échantillon. Ces segments sont donc très probablement identiques par descendance, c'est-à-dire que ce sont des blocs d'IBD. La puissance de cette méthode est limitée par la densité d'allèles rares le long du génome. Elle justifie de plus l'exclusion des individus dont les grands-parents ne sont pas nés en Europe car un segment présent chez un petit nombre de migrants récents sera détecté comme un bloc d'IBD.

Dans les régions où peu de marqueurs sont présents, fastIBD peut regrouper deux petits blocs d'IBD consécutifs en un seul bloc plus long. Ces cas sont appelés faux positifs par Ralph et Coop (la méthode détecte un long bloc alors qu'il n'y en a pas). En mélangeant artificiellement les génomes échantillonnés et en y appliquant la méthode de détection des blocs d'IBD, ils sont capables d'estimer le taux de faux positifs dans leur échantillon en fonction de la longueur du bloc. Un bloc de longueur x (en cM) est ainsi un faux positif avec probabilité

$$\exp(-13 - 2x + 4,3\sqrt{x}).$$

Cela correspond à un taux de faux positifs de 10% et une puissance de 85% à 2cM, cette dernière atteignant 95% à 4cM.

1.7.3 Inférence démographique à partir des blocs d'IBD

Deux individus n'héritent leur matériel génétique commun que d'une partie de leurs ancêtres généalogiques, ces ancêtres sont dits génétiques. Les blocks d'IBD sont des traces laisées par ces ancêtres communs génétiques qui nous donnent un aperçu de l'ensemble des ancêtres généalogiques.

Ralph et Coop [RC13] montrent que deux individus en Europe, même vivant à plusieurs milliers de kilomètres l'un de l'autre, ont une chance raisonnable de partager des ancêtres génétiques au cours des 1 000 dernières années, et en partagent de manière quasi-certaine au cours des 2 500 dernières années. Il en résulte que deux européens même très éloignés géographiquement partagent un grand nombre d'ancêtres généalogiques qui ont vécu au cours du dernier millénaire.

Leur analyse met également en évidence une forte structuration spatiale du nombre de blocs d'IBD entre individus. En effet le nombre moyen de blocs d'IBD partagés par deux individus décroît avec la distance géographique les séparant. Cette structuration spatiale invite à une analyse paramétrique pour estimer certains paramètres démographiques, en particulier la dispersion moyenne des individus par génération.

Cette analyse a été menée par Ringbauer, Coop et Barton [RCB16] dans une sous-région de l'Europe de l'Est. Ils maximisent une pseudo-vraisemblance calculée à l'aide d'une approximation similaire à celle de la formule de Wright-Malécot (plus de détails sont donnés dans le Chapitre 5).

Étant données deux populations à une distance géographique r (calculée suivant les géodésiques terrestres), les auteurs calculent le nombre attendu de blocs d'IBD d'une longueur comprise entre L et $L + \Delta L$ pour chaque paire d'individus issus de ces populations. En supposant que les longueurs des différents blocs d'IBD sont indépendantes le long du génome, le nombre de blocs d'IBD d'une longueur donnée peut être modélisée par une loi de Poisson. Les auteurs font également l'hypothèse que le nombre de blocs de différentes longueurs sont indépendants et que les différentes paires d'individus sont indépendantes. On peut alors exprimer la vraisemblance des données en fonction d'un jeu de paramètres (ici la densité de population efficace et le taux de dispersion des individus).

En maximisant cette pseudo-vraisemblance, Ringbauer et al. sont capables d'estimer ces deux paramètres à partir du nombre de blocs d'IBD dans l'échantillon. Les méthodes d'inférence classiques qui reposent sur des comparaisons locus par locus entre individus [Rou97 ; RL07 ; SB89 ; Bar+13] ne permettent pas d'estimer séparément le taux de dispersion σ et la taille de la population N, elles ont seulement accès au produit de ces deux quantités, aussi appelée taille de voisinage (voir [BDE02]). De plus, Ringbauer et al. prennent en compte dans le calcul de la pseudo-vraisemblance le taux d'erreur calculé par Ralph et Coop [RC13].

Leur méthode d'inférence a d'abord été testée sur des données simulées. Elle est capable d'estimer correctement les paramètres des simulations pour des valeurs similaires à celles attendues dans leur jeu de données. Elle peut également prendre en compte une croissance polynomiale de la densité de la population au cours du temps et estimer l'exposant correspondant.

Ringbauer et al. limitent leur analyse à l'Europe de l'Est en raison de l'importante disparité que présentent les données. En effet la décroissance du nombre moyen de blocs d'IBD entre deux populations avec la distance géographique est sensiblement différente en Europe de l'Est et en Europe occidentale [RC13]. L'échantillon de 13 populations retenu par Ringbauer et al. semble relativement homogène et présente une assez bonne résolution spatiale pour envisager d'estimer des paramètres démographiques uniformes dans cette région.

En appliquant leur méthode d'inférence, ils trouvent trois jeux de paramètres en fonction de l'hypothèse retenue pour la croissance de la densité de population. Si la population est supposée constante, le modèle prédit un taux de dispersion de $\sigma = 67,8 \pm 5,0 km/\sqrt{gen}$ et une densité efficace de $D_e = 0,0047$. Si la population croît linéairement dans le temps, alors $\sigma = 62,6 km/\sqrt{gen}$ et $D_e = D/t$ avec $D = 1,71$. Si D_e est de la forme $Dt^{-\beta}$, alors le maximum de vraisemblance est atteint pour $\beta = 1,05$ avec $D = 2,13$ et $\sigma = 63,0 km/\sqrt{gen}$.

Dans le Chapitre 5, nous généralisons la méthode développée par Ringbauer et al. [RCB16] au cas où l'espace est divisé en deux régions au sein desquelles les valeurs du taux de dispersion et de la densité de population diffèrent. Nous adaptons le calul de la vraisemblance en utilisant les résultats du chapitre 3, développant ainsi une méthode d'inférence démographique en environnement

hétérogène capable d'estimer jusqu'à 5 paramètres (la densité de population et le taux de dispersion dans chaque région et l'exposant β).

Nous testons cette nouvelle méthode sur plusieurs jeux de données simulées pour montrer qu'elle est capable d'estimer correctement les paramètres des simulations dans un nombre suffisant de cas. Après ces premiers résultats encourageants, nous comptons tester la robustesse de notre méthode d'inférence sur différents modèles microscopiques avant de l'appliquer aux données étudiées par Ralph et Coop [RC13].

Chapter 2

A central limit theorem for the spatial Λ-Fleming-Viot process with selection

Joint work with Sarah Penington (Oxford University), published in *Electronic Journal of Probability* 22.5 (2017) [FP17].

Introduction

Consider a population distributed across a geographical space (typically of dimension one or two). Suppose that each individual carries one of several possible versions (or *alleles*) of a gene. How do the different allele frequencies evolve with time and how are they shaped by the main evolutionary forces, such as natural selection and migration? To answer this question, early models from population genetics were adapted by G. Malécot [Mal48], S. Wright [Wri43] and M. Kimura [Kim53] to include spatial structure. These spatial models either considered subdivided populations reproducing locally and exchanging migrants at each generation or made inconsistent assumptions about the distribution of individuals across space.

In this chapter, we focus on a mathematical model for populations evolving in a spatial continuum, the spatial Λ-Fleming-Viot process (SLFV for short), originally proposed in [Eth08]. The main feature of this model is that instead of each individual carrying exponential clocks determining its reproduction and death times, reproduction times are specified by a Poisson point process of extinction-recolonization events. At each of these events, some proportion - often denoted u - of the individuals present in the region affected by the event is replaced by the offspring of an individual (the *parent*) chosen within this region. (The proportion u which is replaced is called the *impact parameter*.) We shall only consider cases where the region affected is a (d-dimensional) ball, and the Poisson point process specifies the time, centre and radius of reproduction events. (Since we consider scaling limits, minor changes to this assumption would not change our results.)

We suppose that each individual in the population has a type taken from a compact space K. The state of the SLFV process at time t is then given by a map $\rho_t : \mathbb{R}^d \to M(K)$ defined Lebesgue almost everywhere, where $M(K)$ is the set of probability measures on the type space K. We think of $\rho_t(x)$ as the distribution of the type of an individual sampled from location x at time t. More precisely, the

spatial Λ-Fleming-Viot process can be obtained as the high population density limit of an individual based model (see [BEV13a]) where the sequence of empirical measures of the individuals' location and type converges to the measure $\rho_t(x)dx$. We thus sometimes use heuristics based on the behaviour of individuals in the prelimiting model even though one cannot speak of individuals in the SLFV.

Natural selection can be included in the SLFV by introducing an independent Poisson point process of selective events which give an advantage to a particular type. Multiple *potential parents* are sampled in the region affected by the event and one is chosen to be the parent and have offspring in a biased way depending on their types. The *selection parameter* determines the rate of this Poisson point process.

A comprehensive survey of recent developments related to the SLFV can be found in [BEV13a]. Several works have focussed on characterising the behaviour of this model over large space and time scales, in the special case where only two types (or two alleles) a and A are present in the population. In this case the state of the process is given by a map $q_t : \mathbb{R}^d \to [0,1]$ defined Lebesgue almost everywhere, where $q_t(x)$ denotes the probability that a randomly chosen individual at location x and time t is of type a, or in other words the proportion of type a at location x and at time t. We shall first consider the simplest form of selection when individuals are *haploid*, i.e. each individual has one copy of the gene, and type A is favoured. At selective events, two potential parents are chosen and if their types are different, the parent is the one which has type A. In [EVY14], rescaling limits of this form of the spatial Λ-Fleming-Viot process with selection (SLFVS) have been obtained when both the impact parameter and the selection parameter tend to zero. Earlier results on the large scale behaviour of the SLFV had already been established in [BEV13b] in the neutral case (*i.e.* without selection), but keeping the impact parameter macroscopic. The behaviour of the SLFVS in the corresponding regime is studied in [EFS15] and [Eth+15].

The limiting process obtained by [EVY14] turns out to be deterministic as soon as $d \geq 2$, and, when the reproduction events have bounded radius, it is given by the celebrated Fisher-KPP equation,

$$\frac{\partial f_t}{\partial t} = \frac{1}{2}\Delta f_t - s f_t(1 - f_t). \tag{2.1}$$

This result fits the original interpretation of this equation proposed by R. A. Fisher as a model for the spread of advantageous genes in a spatially distributed population [Fis37]. The spatial Λ-Fleming-Viot process with selection (SLFVS) can thus be thought of as a refinement of the Fisher-KPP equation, combining spatial structure and a random sampling effect at each generation - what biologists call *genetic drift*.

In the present work we prove a slightly stronger form of convergence to this deterministic rescaling limit. We also study the fluctuations of the allele frequency about (an approximation of) $(f_t)_{t\geq 0}$. We find that if the impact parameter is sufficiently small compared to the selection parameter and the fluctuations are rescaled in the right way then in the limit they solve the following stochastic partial differential equation,

$$dz_t = \left[\frac{1}{2}\Delta z_t - s(1 - 2f_t)z_t\right] dt + \sqrt{f_t(1 - f_t)}\, dW_t, \tag{2.2}$$

where W is space-time white noise, and f is the solution of (2.1). More detailed statements with the precise conditions on the parameters of the SLFVS are given in Section 2.2.

A very similar result was proved by F. Norman in the non-spatial setting [Nor75a] (see also [Nor74a], [Nor77] and [Nor75b]). Norman considered the Wright-Fisher model for a population of size N under natural selection (see [Eth11] for an introduction to such models). Let p_n^N denote the proportion of individuals not carrying the favoured allele at generation n, and suppose that the selection parameter is given by $s_N = \varepsilon_N s$, with $\varepsilon_N \to 0$ and $\varepsilon_N N \to \infty$ as $N \to \infty$. (At each generation, individuals choose a parent of the favoured type with probability $\frac{(1+s_N)(1-p_n^N)}{1+s_N(1-p_n^N)}$.) Norman showed that, as $N \to \infty$, $p^N_{\lfloor t/\varepsilon_N \rfloor}$ converges to g_t, which satisfies

$$\frac{dg_t}{dt} = -sg_t(1-g_t).$$

(In the weak selection regime - *i.e.* when $Ns_N = \mathcal{O}(1)$ - one recovers the classical Wright-Fisher diffusion, see [Eth11].) Furthermore, the fluctuations of p^N_{t/ε_N} around g_t are of order $(N\varepsilon_N)^{-1/2}$. More precisely, for $t = n\varepsilon_N$, $n \in \mathbb{N}$, set

$$Z^N(t) = (N\varepsilon_N)^{1/2}\left(p^N_{t/\varepsilon_N} - g_t\right),$$

and define $Z^N(t)$ for all $t \geq 0$ by linear interpolation. Theorem 2 in [Nor75a] states that, as $N \to \infty$, $\left(Z^N(t)\right)_{t\geq 0}$ converges to the solution of the following stochastic differential equation,

$$dz_t = -s(1-2g_t)z_t dt + \sqrt{g_t(1-g_t)}dB_t,$$

where $(B_t)_{t\geq 0}$ is a standard Brownian motion; note that $(z_t)_{t\geq 0}$ is a Gaussian diffusion. A similar regime in the case of a neutral model with mutations was already studied by W. Feller in [Fel51, Section 9], who identified the limiting diffusion for the fluctuations around the equilibrium frequency.

Norman's result can be extended to other classical models from population genetics, and in particular to continuous-time processes such as the Moran model and the (non-spatial) Λ-Fleming-Viot process (introduced in [BLG03]). The necessary tools can be found mainly in [EK86, Chapter 11] (see also Chapter 6 of the same book) and in [Kur71]. In this chapter we adapt these methods to the setting of the spatial Λ-Fleming-Viot process, with the necessary tools for stochastic partial differential equations taken from [Wal86] (see also [MT95] and [DMFL86]).

We also consider a second regime for the SLFVS to allow large scale extinction-recolonization events; we let the radius of reproduction events follow an α-stable distribution truncated at zero. For this regime, as in [EVY14], we find the Fisher-KPP equation with non-local diffusion as a rescaling limit (*i.e.* with a fractional Laplacian instead of the usual Laplacian). The Laplacian is also replaced by a fractional Laplacian in (2.2), the equation satisfied by the limiting fluctuations, and the noise W becomes a coloured noise with spatial correlations of order $|x-y|^{-\alpha}$ (see Subsection 2.2.2).

These results are valid for a general class of selection mechanisms, with modified versions of (2.1) and (2.2) (and our proof will cover the general case). As an application of our results on the fluctuations, we turn to a particular kind of selection mechanism. Suppose a given gene is present in two different forms - denoted A_1 and A_2 - within a population. Suppose also that each individual carries two copies of this gene (each inherited from one of two parents). We say that individuals are *diploid*, and *homozygous* individuals are those who carry two copies of the same type (A_1A_1 or A_2A_2) while *heterozygous* individuals carry one copy of each type (A_1A_2). *Overdominance* occurs when the

relative fitnesses of the three possible genotypes are as follows,

$$\begin{array}{ccc} A_1A_1 & A_1A_2 & A_2A_2 \\ 1-s_1 & 1 & 1-s_2, \end{array}$$

where s_1, $s_2 > 0$. In words, heterozygous individuals produce more offspring than both types of homozygous individuals. In this setting, in an infinite population a stable intermediate allele frequency is expected to be maintained, preventing either type from disappearing. If q is the frequency of type A_1 and $p = 1 - q$ that of type A_2 and if mating is random, the respective proportions of the three genotypes will be q^2, $2qp$, p^2, hence the population cannot remain composed exclusively of heterozygous individuals. As a consequence, even when the stable equilibrium is reached, the mean fitness of the population will not be as high as the highest possible individual fitness (*i.e.* that of heterozygous individuals). This fitness reduction is referred to as the *segregation load*.

In finite populations, because of finite sample size, the allele frequency is never exactly at its optimum. This was the subject of a work by A. Robertson [Rob70] who considered this specific configuration of the relative fitnesses. He argued that the mean fitness in a *panmictic* population (*i.e.* one with no spatial structure) with finite but relatively large size N is reduced by a term of order $(4N)^{-1}$, irrespective of the strength of selection. This is due to a trade-off between genetic drift and natural selection. The stronger selection is, the quicker the allele frequency is pushed back to the equilibrium, but at the same time even a small step away from the optimal frequency is very costly in terms of mean fitness. On the other hand, if natural selection is relatively weak, the allele frequency can wander off more easily, but the mean fitness of the population decreases more slowly. This reduction in the mean fitness due to genetic drift - which is added to the reduction from the segregation load - is called the *drift load*.

Robertson's result can be made rigorous using tools found in [Nor74a] and [Nor74b]. We adapt these to our setting and study the same effect in spatially structured populations. We find that the spatial structure significantly reduces the drift load, in a way that depends crucially on dimension. It turns out that migration prevents the allele frequencies from straying too far away from the equilibrium frequency, because incoming migrants are on average close to this equilibrium.

The chapter is laid out as follows. We define the spatial Λ-Fleming-Viot process for a haploid model with general frequency dependent selection and for a diploid model of overdominance in Section 2.1. In Section 2.2 we state the main convergence results for the SLFVS in the bounded radius and stable radius regimes and we present our estimate of the drift load in spatially structured populations. In Section 2.3, we present the main ingredient of the proof: a martingale problem satisfied by the SLFVS. At the end of Subsections 2.3.2 and 2.3.3, we state more general results on solutions to these martingale problems which imply our convergence results for the SLFVS. Most of the remainder of this chapter is dedicated to the proofs of these results. The central limit theorem in the bounded radius case is proved in Section 2.4, while the stable regime is dealt with in Section 2.5 (the two proofs share the same structure, but differ in the details of the estimates). Finally, the asymptotics of the drift load are derived in Section 2.6.

2.1 Definition of the model

2.1.1 The state space of the spatial Λ-Fleming-Viot process with selection

We now turn to a precise definition of the underlying model, the spatial Λ-Fleming-Viot process with selection on $\mathbb{R}^d$, starting with the state space of the process. At each time $t \geq 0$, $\{q_t(x) : x \in \mathbb{R}^d\}$ is a random function such that

$$q_t(x) := \text{proportion of type } a \text{ alleles at spatial position } x \text{ at time } t, \tag{2.3}$$

which is in fact defined up to a Lebesgue null set of $\mathbb{R}^d$. More precisely, let Ξ be the quotient of the space of Lebesgue-measurable maps $f : \mathbb{R}^d \to [0,1]$ by the equivalence relation

$$f \sim f' \iff Leb(\{x \in \mathbb{R}^d : f(x) \neq f'(x)\}) = 0.$$

We endow Ξ with the topology of vague convergence: letting $\langle f, \phi \rangle = \int_{\mathbb{R}^d} f(x)\phi(x)dx$, a sequence $(f_n)_n$ converges vaguely to $f \in \Xi$ if and only if $\langle f_n, \phi \rangle \underset{n\to\infty}{\longrightarrow} \langle f, \phi \rangle$ for any continuous and compactly supported function $\phi : \mathbb{R}^d \to \mathbb{R}$. A convenient metric for this topology is given by choosing a separating family $(\phi_n)_{n\geq 1}$ of smooth, compactly supported functions which are uniformly bounded in $L^1(\mathbb{R}^d)$. Then for $f, g \in \Xi$,

$$d_\Xi(f,g) = \sum_{n\geq 1} \frac{1}{2^n} |\langle f, \phi_n \rangle - \langle g, \phi_n \rangle| \tag{2.4}$$

defines a metric for the topology of vague convergence on Ξ. The SLFVS up to time T is then going to be a $\mathrm{D}([0,T], \Xi)$-valued random variable: a Ξ-valued process with càdlàg paths.

Definition 2.1.1. *For $T > 0$, let $f, g \in \mathrm{D}([0,T], \Xi)$ be a pair of càdlàg maps $(f_t)_{0\leq t\leq T}$, $(g_t)_{0\leq t\leq T}$ from $[0,T]$ to Ξ. Then*

$$d(f,g) = \sup_{t\in[0,T]} d_\Xi(f_t, g_t)$$

is a metric for the topology of uniform convergence on $\mathrm{D}([0,T], \Xi)$.

For more details, see Section 2.2 of [VW15].

2.1.2 The spatial Λ-Fleming-Viot process with selection

Let us now define the dynamics of the process. Let $u \in (0,1]$ and $s \in [0,1]$, and let $\mu(dr)$ be a finite measure on $(0,\infty)$ satisfying

$$\int_0^\infty r^d \mu(dr) < \infty. \tag{2.5}$$

For $m \in \mathbb{N}$ and $w \in [0,1]$, let $\vec{B}_w^m$ be a vector of m independent random variables taking the value a with probability w and A otherwise. Then let $F : [0,1] \to \mathbb{R}$ be a polynomial such that for some $m \in \mathbb{N}$ and $p : \{a, A\}^m \to [0,1]$, for each $w \in [0,1]$,

$$w - F(w) = \mathbb{E}\left[p(\vec{B}_w^m)\right]. \tag{2.6}$$

(The choice of p and m is not unique, but this will not matter.)

Definition 2.1.2 (SLFVS, haploid case with general frequency dependent selection)**.** *Let Π and Π^S be two independent Poisson point processes on $\mathbb{R}_+ \times \mathbb{R}^d \times (0,\infty)$ with intensity measures $(1-s)\,dt \otimes dx \otimes \mu(dr)$ and $s\,dt \otimes dx \otimes \mu(dr)$ respectively. The spatial Λ-Fleming-Viot process with selection for a haploid population in $\mathbb{R}^d$ with impact parameter u, radius of reproduction events given by $\mu(dr)$, selection parameter s and selection function F is defined as follows. If $(t,x,r) \in \Pi$, a neutral event occurs at time t within the ball $B(x,r)$:*

1. *Choose a location y uniformly at random in $B(x,r)$ and sample a parental type $k \in \{a,A\}$ according to $q_{t^-}(y)$ (i.e. $k = a$ with probability $q_{t^-}(y)$).*
2. *Update q as follows:*

$$\forall z \in \mathbb{R}^d, q_t(z) = q_{t^-}(z) + u\mathbb{1}_{\{|x-z|<r\}}(\mathbb{1}_{\{k=a\}} - q_{t^-}(z)). \tag{2.7}$$

Similarly, if $(t,x,r) \in \Pi^S$, a selective event occurs at time t inside $B(x,r)$:

1. *Choose m locations $y_1,\dots,y_m$ independently uniformly at random in $B(x,r)$, sample a type k_i at each location y_i according to $q_{t^-}(y_i)$ and then let $k = a$ with probability $p(k_1,\dots,k_m)$ and $k = A$ otherwise.*
2. *Update q as in* (2.7)*.*

Note that if we let $w = |B(x,r)|^{-1}\int_{B(x,r)} q_{t^-}(z)\,dz$, then at a neutral reproduction event, $\mathbb{P}(k=a) = w$ and at a selective event, $\mathbb{P}(k=a) = w - F(w)$. This justifies the definition in terms of the selection function F (and the terminology for F), since the law of the process depends only on F, and not on the specific choice of p and m in (2.6).

Remark. *The existence of a unique Ξ-valued process following these dynamics under condition* (2.5) *is proved in [EVY14, Theorem 1.2] in the special case $F(w) = w(1-w)$ (in the neutral case $s = 0$, this was done in [BEV10a]). In our general case, the condition on $w - F(w)$ in* (2.6) *allows one to define a branching and coalescing dual process and hence prove existence and uniqueness in the same way as in [EVY14].*

We shall consider two different distributions μ for the radii of events,

i) the fixed radius case : $\mu(dr) = \delta_R(dr)$ for some $R > 0$,

ii) the stable radius case : $\mu(dr) = \frac{\mathbb{1}_{\{r\geq 1\}}}{r^{d+\alpha+1}}dr$ for a fixed $\alpha \in (0, 2\wedge d)$.

In each case, (2.5) is clearly satisfied.

We give two variants of this definition corresponding to the two selection mechanisms discussed in the introduction. We begin with a model for a selective advantage for A alleles in haploid reproduction.

Definition 2.1.3 (SLFVS, haploid model, genic selection)**.** *The spatial Λ-Fleming-Viot process with genic selection with impact parameter u, radius of reproduction events given by $\mu(dr)$ and selection parameter s is defined as in Definition 2.1.2 with $F(w) = w(1-w)$. In this case, $m = 2$ and the function p equals simply*

$$p(k_1,k_2) = \mathbb{1}_{\{k_1=k_2=a\}}.$$

In other words, during selective reproduction events, two types are sampled in $B(x,r)$ and $k = a$ if and only if both types are a.

2.1.3 The SLFVS with overdominance

We now define a variant of the SLFVS to model overdominance. Individuals are diploid and we study a gene which is present in two different forms within the population, denoted A_1 and A_2. For $t \geq 0$ and $x \in \mathbb{R}^d$, let

$$q_t(x) := \text{ the proportion of the allele type } A_1 \text{ at location } x \text{ at time } t.$$

(If p_1 is the proportion of A_1A_1 indiduals and p_H is the proportion of A_1A_2 heterozygous individuals, then $q = p_1 + \frac{1}{2}p_H$.) We assume that the relative fitnesses of the different genotypes are as follows:

$$\begin{array}{ccc} A_1A_1 & A_1A_2 & A_2A_2 \\ 1-s_1 & 1 & 1-s_2. \end{array}$$

In other words, for an event (t,x,r) in the SLFVS with $w = |B(x,r)|^{-1} \int_{B(x,r)} q_{t^-}(z)\,dz$, we want to choose parental types $(k_1,k_2) \in \{A_1,A_2\}^2$ at random with

$$\begin{aligned} \mathbb{P}(\{k_1,k_2\} = \{A_1,A_1\}) = P_{11} &:= \tfrac{(1-s_1)w^2}{1-s_1w^2-s_2(1-w)^2}, \\ \mathbb{P}(\{k_1,k_2\} = \{A_1,A_2\}) = P_{12} &:= \tfrac{2w(1-w)}{1-s_1w^2-s_2(1-w)^2}, \\ \mathbb{P}(\{k_1,k_2\} = \{A_2,A_2\}) = P_{22} &:= \tfrac{(1-s_2)(1-w)^2}{1-s_1w^2-s_2(1-w)^2}. \end{aligned}$$

We also suppose that, with probability ν_1, the type A_1 alleles produced mutate to type A_2, and that, with probability ν_2, the type A_2 mutate to type A_1 (this is a technical assumption to ensure that $q_t(x) \notin \{0,1\}$; we shall assume that ν_1 and ν_2 are small compared to s_1 and s_2). This gives us the following modified probabilities for the parental types:

$$\begin{aligned} \mathbb{P}(\{k_1,k_2\} = \{A_1,A_1\}) &= (1-\nu_1)P_{11} + \nu_2(1-P_{11}), \\ \mathbb{P}(\{k_1,k_2\} = \{A_1,A_2\}) &= (1-\nu_1-\nu_2)P_{12}, \\ \mathbb{P}(\{k_1,k_2\} = \{A_2,A_2\}) &= (1-\nu_2)P_{22} + \nu_1(1-P_{22}). \end{aligned}$$

We are going to be interested in small values of s_i and ν_i, so we expand:

$$\begin{aligned} &\mathbb{P}(\{k_1,k_2\} = \{A_1,A_1\}) \\ &\quad = (1-\nu_1)(w^2(1-s_1+s_1w^2+s_2(1-w)^2) + \mathcal{O}(s^2)) + \nu_2(1-w^2+\mathcal{O}(s)) \\ &\quad = (1-s_1-s_2-\nu_1-\nu_2)w^2 + s_1w^4 + s_2w^2(1+(1-w)^2) + \nu_2 + \mathcal{O}(s^2+\nu s), \end{aligned} \tag{2.8}$$

where $s = s_1 + s_2$ and $\nu = \nu_1 + \nu_2$. Similarly, we have

$$\begin{aligned}\mathbb{P}\left(\{k_1,k_2\} = \{A_1,A_2\}\right) &= (1 - s_1 - s_2 - \nu_1 - \nu_2)2w(1-w) \\ &\quad + s_1 2w(1-w)(1+w^2) + s_2 2w(1-w)(1+(1-w)^2) + \mathcal{O}\left(s^2 + \nu s\right), \\ \mathbb{P}\left(\{k_1,k_2\} = \{A_2,A_2\}\right) &= (1 - s_1 - s_2 - \nu_1 - \nu_2)(1-w)^2 \\ &\quad + s_1(1-w)^2(1+w^2) + s_2(1-w)^4 + \nu_1 + \mathcal{O}\left(s^2 + \nu s\right).\end{aligned} \tag{2.9}$$

The following model results in the parental type probabilities given in (2.8) and (2.9), neglecting the $\mathcal{O}\left(s^2 + \nu s\right)$ terms.

Definition 2.1.4 (SLFVS, overdominance)**.** *Suppose that $\nu_1 + \nu_2 + s_1 + s_2 < 1$. Let Π, Π^{S_i} and Π^{ν_i}, $i = 1, 2$ be five independent Poisson point processs on $\mathbb{R}_+ \times \mathbb{R}^d \times (0,\infty)$ with respective intensity measures $(1 - s_1 - s_2 - \nu_1 - \nu_2)\, dt \otimes dx \otimes \mu(dr)$, $s_i\, dt \otimes dx \otimes \mu(dr)$ and $\nu_i\, dt \otimes dx \otimes \mu(dr)$. The spatial Λ-Fleming-Viot process with overdominance with impact parameter u, radius of reproduction events given by μ, selection parameters s_1, s_2 and mutation parameters ν_1, ν_2 is defined as follows. If $(t,x,r) \in \Pi$, a neutral event occurs at time t in $B(x,r)$:*

1. *Pick two locations y_1 and y_2 uniformly at random within $B(x,r)$ and sample one parental type $k_i \in \{A_1, A_2\}$ at each location according to $q_{t^-}(y_i)$, independently of each other.*
2. *Update q as follows:*

$$\forall z \in \mathbb{R}^d, \quad q_t(z) = q_{t^-}(z) + u\mathbb{1}_{\{|x-z|<r\}}\left(\tfrac{1}{2}(\mathbb{1}_{\{k_1=A_1\}} + \mathbb{1}_{\{k_2=A_1\}}) - q_{t^-}(z)\right). \tag{2.10}$$

If $(t,x,r) \in \Pi^{S_i}$, a selective event occurs at time t in $B(x,r)$:

1. *Pick four locations uniformly at random within $B(x,r)$ and sample one type at each location, forming two pairs of types. If one pair is $\{A_i, A_i\}$, let $\{k_1,k_2\}$ be the other pair; otherwise pick one pair at random, each with probability $1/2$. (If the two sampled pairs are $\{A_i,A_i\}$, then $\{k_1,k_2\} = \{A_i,A_i\}$.)*
2. *Update q as in (2.10).*

If $(t,x,r) \in \Pi^{\nu_i}$, a mutation event occurs at time t in $B(x,r)$:

1. *Set $\{k_1,k_2\} = \{A_{3-i}, A_{3-i}\}$, irrespective of the state of q_{t^-}. (In other words we suppose that the A_i genes of the offspring mutate to type A_{3-i}.)*
2. *Update q as in (2.10).*

Remark. *Similarly to the haploid case, existence and uniqueness for this process can be proved as in [EVY14] using a dual process.*

We shall see in Section 2.3 that this process satisfies essentially the same martingale problem as the general haploid process in Definition 2.1.2 with

$$F(w) = w(1-w)(w - \frac{s_2}{s_1+s_2}) + \frac{\nu_1}{s_1+s_2}w - \frac{\nu_2}{s_1+s_2}(1-w).$$

2.2 Statement of the results

In this section, we present our main results. We consider the SLFVS as in Definitions 2.1.2 and 2.1.4, and we let the impact parameter and the selection and mutation parameters tend to zero. On a suitable space and time scale (depending on the regime of the radii of reproduction events) the process $\left(q_t^N\right)_{t\geq 0}$ converges to a deterministic process. We also characterise the limiting fluctuations of $\left(q_t^N\right)_{t\geq 0}$ about an approximation to this deterministic process as the solution to a stochastic partial differential equation.

2.2.1 Fixed radius of reproduction events

We begin by considering the regime in which the radii of the regions affected by reproduction events are bounded. We shall only give the proof in the case of fixed radius events; the proof for bounded radius events is the same but notationally awkward. Fix $u, s \in (0,1]$ and $R > 0$, and choose $w_0 : \mathbb{R}^d \to [0,1]$ with uniformly bounded spatial derivatives of up to the fourth order. Take two sequences $(\varepsilon_N)_{N\geq 1}$, $(\delta_N)_{N\geq 1}$ of positive real numbers in $(0,1]$ decreasing to zero, and set

$$s_N = \delta_N^2 s, \qquad u_N = \varepsilon_N u, \qquad r_N = \delta_N R, \qquad q_0^N(x) = w_0(\delta_N x).$$

Let $\mu(dr) = \delta_R$, and let $F : \mathbb{R} \to \mathbb{R}$ be a smooth, bounded function with bounded derivatives of order up to four such that $F|_{[0,1]}$ satisfies (2.6) for some $m \in \mathbb{N}$ and $p : \{a, A\}^m \to [0,1]$. Then for $N \geq 1$, let $\left(q_t^N\right)_{t\geq 0}$ be the spatial Λ-Fleming-Viot process in $\mathbb{R}^d$ with selection following the dynamics of Definition 2.1.2 with impact parameter u_N, radius of reproduction events R, selection parameter s_N and selection function F started from the initial condition q_0^N.

Define the rescaled process $\left(\mathbf{q}_t^N\right)_{t\geq 0}$ by setting:

$$\forall x \in \mathbb{R}^d, t \geq 0,\ \mathbf{q}_t^N(x) = q^N_{t/(\varepsilon_N \delta_N^2)}(x/\delta_N). \tag{2.11}$$

We justify this scaling as follows. Recall from (2.3) that we think of $q_t^N(x)$ as denoting the proportion of the population at location x and time t which is of type a. Consider an individual randomly chosen from the population at location x at time t. It finds itself within a region affected by a reproduction event at rate $|B(0,R)|$. The probability that it dies and is replaced by a new individual is $u_N = \varepsilon_N u$, so, if we rescale time by $1/\varepsilon_N$, this will happen at rate $\mathcal{O}(1)$. Also, we are going to see later (see Section 2.3.2) that the reproduction events act like a discrete heat flow on the allele frequencies. We rescale time further by $1/\delta_N^2$ and space by $1/\delta_N$, which corresponds to the diffusive scaling of this discrete heat flow. Since selective events also take place at rate $\mathcal{O}\left(\delta_N^2\right)$, this is the right scaling to consider in order to observe the effects of both migration and selection in the limit. (Due to this diffusive scaling we shall refer to this regime as the Brownian case.)

We need to introduce some notation. Let $L^{1,\infty}(\mathbb{R}^d)$ denote the space of bounded and integrable real-valued functions on $\mathbb{R}^d$. For $r > 0$, we set $V_r = |B(0,r)|$ and, for $x, y \in \mathbb{R}^d$,

$$V_r(x,y) = |B(x,r) \cap B(y,r)|\,. \tag{2.12}$$

For $\phi : \mathbb{R}^d \to \mathbb{R}$ and $x \in \mathbb{R}^d$, set

$$\overline{\phi}(x,r) = \frac{1}{V_r} \int_{B(x,r)} \phi(y)dy.$$

When there is no ambiguity, we shall not specify the radius r and simply write $\overline{\phi}(x)$. This notation will be used throughout this chapter and formulae will routinely involve averages of averages, *etc.* For example we also write

$$\overline{\overline{\phi}}(x,r) = \frac{1}{V_r^2} \int_{B(x,r)} \int_{B(y,r)} \phi(z)dzdy. \tag{2.13}$$

Let us define a linear operator $\mathcal{L}^{(r)}$ by setting

$$\mathcal{L}^{(r)}\phi(x) = \frac{d+2}{2r^2} \left(\overline{\overline{\phi}}(x,r) - \phi(x) \right). \tag{2.14}$$

Finally let $\mathcal{S}(\mathbb{R}^d)$ denote the Schwartz space of rapidly decreasing smooth functions on $\mathbb{R}^d$, whose derivatives of all orders are also rapidly decreasing. Accordingly, let $\mathcal{S}'(\mathbb{R}^d)$ denote the space of tempered distributions.

Lemma 2.2.1. *If $w_0 : \mathbb{R}^d \to [0,1]$ has uniformly bounded spatial derivatives of order up to four, then*

$$\begin{cases} \dfrac{\partial f_t^N}{\partial t} = uV_R \left[\dfrac{2R^2}{d+2} \mathcal{L}^{(r_N)} f_t^N - s\overline{F(\overline{f_t^N})}(r_N) \right], \\ f_0^N = w_0. \end{cases} \tag{2.15}$$

defines a unique (deterministic) function f^N in $L^\infty([0,T] \times \mathbb{R}^d)$. In addition, it admits spatial derivatives of order up to four which are all in $L^\infty([0,T] \times \mathbb{R}^d)$.

We prove this lemma in Section 2.8 with a Picard iteration.

As stated in the introduction, the spatial Λ-Fleming-Viot process with genic selection with fixed radius of reproduction events converges, under what can be considered a diffusive scaling, to the solution of the Fisher-KPP equation (as in [EVY14] for $d \geq 2$) while the limiting fluctuations are given by the solution to a stochastic partial differential equation which generalises the result obtained in [Nor75a]. We can now give a precise statement of this result for general frequency dependent selection. The same result holds for radius distributions given by a finite measure μ on a bounded interval.

Theorem 2.1 (Central Limit Theorem for the SLFVS in $\mathbb{R}^d$ with fixed radius of reproduction events)**.** *Let $\left(\mathbf{q}_t^N\right)_{t\geq 0}$ be defined as in* (2.11)*. Suppose that $\varepsilon_N = o\left(\delta_N^{d+2}\right)$, then the process $\left(\mathbf{q}_t^N\right)_{t\geq 0}$ converges in L^1 and in probability (for the metric d of Definition 2.1.1) to the deterministic solution $f : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ of the following PDE,*

$$\begin{cases} \dfrac{\partial f_t}{\partial t} = uV_R \left[\dfrac{R^2}{d+2} \Delta f_t - sF(f_t) \right], \\ f_0 = w_0. \end{cases}$$

In addition,

$$Z_t^N = (\varepsilon_N \delta_N^{d-2})^{-1/2}(\mathbf{q}_t^N - f_t^N)$$

defines a sequence of distribution-valued processes converging in distribution in $\mathrm{D}\left([0,T], \mathcal{S}'(\mathbb{R}^d)\right)$ *to the solution of the following stochastic partial differential equation,*

$$\begin{cases} dz_t = uV_R \left[\dfrac{R^2}{d+2}\Delta z_t - sF'(f_t)z_t\right] dt + uV_R\sqrt{f_t(1-f_t)}dW_t, \\ z_0 = 0, \end{cases}$$

where W *is a space-time white noise.*

Remark. *The impact parameter* u_N *is inversely proportional to the neighbourhood size - i.e. the probability that two individuals have a common parent in the previous generation (see Section 3.6 of [BEV13a] for details). Hence, letting* u_N *tend to zero corresponds to letting the neighbourhood size grow to infinity.*

We shall show in Section 2.3 that Theorem 2.1 is a consequence of Theorem 2.4. The latter is a result on sequences of solutions to a martingale problem and is proved in Section 2.4. In [EVY14], the authors already showed that in the special case of genic selection (as in Definition 2.1.3), for $d \geq 2$, the sequence of averages of $\left(\mathbf{q}_t^N\right)_{t\geq 0}$ over balls of radius r_N converges in distribution in $\mathrm{D}([0,\infty), \Xi)$ to the solution of the Fisher-KPP equation.

Remark. *It would have been more natural to consider the fluctuations directly around the deterministic limit* $(f_t)_{t\geq 0}$*, but in fact the difference between* f_N *and* f *is too large (of order* δ_N^2*, see Proposition 2.4.5). We have that* $|Z_t^N - (\varepsilon_N \delta_N^{d-2})^{-1/2}(\mathbf{q}_t^N - f_t)| = \mathcal{O}(\delta_N^2(\varepsilon_N \delta_N^{d-2})^{-1/2})$ *but if* $\varepsilon_N = o\left(\delta_N^{d+2}\right)$ *then* $(\varepsilon_N \delta_N^{d-2})^{1/2}\delta_N^{-2} = o(\delta_N^{d-2})$ *and so* $(\varepsilon_N \delta_N^{d-2})^{-1/2}\delta_N^2 \to \infty$ *as* $N \to \infty$ *as soon as* $d \geq 2$*.*

2.2.2 Stable radii of reproduction events

In the previous subsection, we assumed that the radius of dispersion of the offspring produced at reproduction events was small. We now wish to allow large scale extinction-recolonization events to take place to illustrate the fact that "catastrophic" extinction events can occur, followed by a quick replacement of the dead individuals by the offspring of a small subset (here only one individual) of the survivors. To do so, we suppose that the intensity measure for the radius of reproduction events $\mu(dr)$ has a power law behaviour, following the work in [EVY14]. The corresponding limiting behaviour is described by reaction-diffusion equations with non-local diffusion, studied for example in [Chm13; AK15]. Suppose that the measure $\mu(dr)$ for the radius of reproduction events is given by

$$\mu(dr) = \frac{\mathbb{1}_{\{r\geq 1\}}}{r^{d+\alpha+1}}dr, \tag{2.16}$$

for some $\alpha \in (0, 2 \wedge d)$. Fix $u, s \in (0,1]$ and choose $w_0 : \mathbb{R}^d \to [0,1]$ with uniformly bounded spatial derivatives of up to the second order. Again, take $(\varepsilon_N)_{N\geq 1}$ and $(\delta_N)_{N\geq 1}$ two sequences in $(0,1]$

decreasing to zero, and set

$$s_N = \delta_N^\alpha s, \qquad u_N = \varepsilon_N u, \qquad q_0^N(x) = w_0(\delta_N x).$$

Let $F : \mathbb{R} \to \mathbb{R}$ be a smooth, bounded function with bounded first and second derivatives such that $F|_{[0,1]}$ satisfies (2.6) for some $m \in \mathbb{N}$ and $p : \{a, A\}^m \to [0, 1]$. Then for $N \geq 1$, let $\left(q_t^N\right)_{t\geq 0}$ be the spatial Λ-Fleming-Viot process with selection following the dynamics of Definition 2.1.2 with impact parameter u_N, radius of reproduction events given by $\mu(dr)$ in (2.16), selection parameter s_N and selection function F started from the initial condition q_0^N.

The main difference with the setting of Subsection 2.2.1 is that the flow resulting from the reproduction events is the α-stable version of the heat flow (see Section 2.3.3). Thus we apply a stable scaling of time by $1/\delta_N^\alpha$ and space by $1/\delta_N$ (after rescaling time by $1/\varepsilon_N$ as previously). Since we have chosen $s_N = \delta_N^\alpha s$, this is the right scaling to consider in order to observe both selection and migration in the limit. For all $x \in \mathbb{R}^d$ and $t \geq 0$, set

$$\mathbf{q}_t^N(x) = q^N_{t/(\varepsilon_N \delta_N^\alpha)}(x/\delta_N). \tag{2.17}$$

We need some more notation; recall the notation for double averages in (2.13). The following will take up the role played by $\overline{F(\overline{w})}$ in the fixed radius case. For $H : [0, 1] \to \mathbb{R}$, $\delta > 0$, and $f \in \Xi$, set

$$H^{(\delta)}(f) : x \mapsto \alpha \int_1^\infty \overline{H(\overline{f})}(x, \delta r) \frac{dr}{r^{\alpha+1}}. \tag{2.18}$$

Recalling the notation in (2.12), set, for $x, y \in \mathbb{R}^d$,

$$\Phi(|x-y|) = \int_{\frac{|x-y|}{2}}^\infty \frac{V_r(x,y)}{V_r} \frac{dr}{r^{d+\alpha+1}}, \qquad \Phi^{(\delta)}(|x-y|) = \int_{\frac{|x-y|}{2}\vee\delta}^\infty \frac{V_r(x,y)}{V_r} \frac{dr}{r^{d+\alpha+1}}.$$

For $\phi : \mathbb{R}^d \to \mathbb{R}$ which admits uniformly bounded spatial derivatives of order up to two and $\psi \in L^{1,\infty}(\mathbb{R}^d)$,

$$\mathcal{D}^\alpha \phi(x) = \int_{\mathbb{R}^d} \Phi(|x-y|)(\phi(y) - \phi(x))dy, \quad \mathcal{D}^{\alpha,\delta}\psi(x) = \int_{\mathbb{R}^d} \Phi^{(\delta)}(|x-y|)(\psi(y) - \psi(x))dy. \tag{2.19}$$

Remark. *Up to a multiplicative constant, depending on d and α, $\mathcal{D}^\alpha$ is the fractional Laplacian (this can be seen via the Fourier transform, see [SKM93]).*

We can now formulate our result for the stable radii regime. The main difference from Theorem 2.1 is that the Laplacian has to be replaced by the operator $\mathcal{D}^\alpha$ and that the noise driving the fluctuations is replaced by a coloured noise which is white in time and has spatial correlations which decay like $K_\alpha(z_1, z_2)$ as $|z_1 - z_2| \to \infty$, where

$$K_\alpha(z_1, z_2) = \int_{\frac{|z_1-z_2|}{2}}^\infty V_r(z_1, z_2) \frac{dr}{r^{d+\alpha+1}} = \frac{C_{d,\alpha}}{|z_1 - z_2|^\alpha}. \tag{2.20}$$

We also set the following notation: for $f \in \Xi$,

$$[f]_\alpha(z_1,z_2) = \frac{\int_{\frac{|z_1-z_2|}{2}}^\infty \frac{dr}{r^{d+\alpha+1}} \int_{B(z_1,r)\cap B(z_2,r)} \overline{f}(x,r)dx}{\int_{\frac{|z_1-z_2|}{2}}^\infty V_r(z_1,z_2)\frac{dr}{r^{d+\alpha+1}}}. \tag{2.21}$$

Note that if f denotes the frequency of type a in $\mathbf{q}_t^N$ immediately before a (neutral) reproduction event which hits both z_1 and z_2 with $|z_1 - z_2| \geq 2\delta_N$, then $[f]_\alpha(z_1,z_2)$ is the probability that the offspring produced in this event are of type a.

The following lemma provides a deterministic centering term f^N around which we consider the fluctuations of $\mathbf{q}^N$.

Lemma 2.2.2. *If $w_0 : \mathbb{R}^d \to [0,1]$ has uniformly bounded spatial derivatives of order up to two, then*

$$\begin{cases} \dfrac{\partial f_t^N}{\partial t} = u\left[\mathcal{D}^{\alpha,\delta_N} f_t^N - \dfrac{V_1 s}{\alpha} F^{(\delta_N)}(f_t^N)\right], \\ f_0^N = w_0 \end{cases} \tag{2.22}$$

defines a unique (deterministic) function f^N in $L^\infty([0,T]\times\mathbb{R}^d)$. In addition, it admits spatial derivatives of order up to two which are all in $L^\infty([0,T]\times\mathbb{R}^d)$.

This lemma is proved in Section 2.8.

Theorem 2.2 (Central Limit Theorem for the SLFVS in $\mathbb{R}^d$ with stable radii of reproduction events)**.** *Let $\left(\mathbf{q}_t^N\right)_{t\geq 0}$ be defined as in (2.17). Suppose that $\varepsilon_N = o\left(\delta_N^{2\alpha}\right)$; then $\left(\mathbf{q}_t^N\right)_{t\geq 0}$ converges in L^1 and in probability (for the metric d of Definition 2.1.1) to the deterministic solution $f : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ of the following PDE,*

$$\begin{cases} \dfrac{\partial f_t}{\partial t} = u\left[\mathcal{D}^\alpha f_t - \dfrac{sV_1}{\alpha} F(f_t)\right], \\ f_0 = w_0. \end{cases} \tag{2.23}$$

In addition,

$$Z_t^N = \varepsilon_N^{-1/2}(\mathbf{q}_t^N - f_t^N)$$

defines a sequence of distribution-valued processes, converging in distribution in $\mathrm{D}\left([0,T],\mathcal{S}'(\mathbb{R}^d)\right)$ to the solution of the following stochastic partial differential equation,

$$\begin{cases} dz_t = u\left[\mathcal{D}^\alpha z_t - \dfrac{sV_1}{\alpha} F'(f_t) z_t\right] dt + u\, dW_t^\alpha \\ z_0 = 0, \end{cases}$$

where W^α is a coloured noise with covariation measure given by

$$\begin{aligned} Q^\alpha(dz_1dz_2ds) = K_\alpha(z_1,z_2)\,([f_s]_\alpha(z_1,z_2)(1-f_s(z_1))(1-f_s(z_2)) \\ +(1-[f_s]_\alpha(z_1,z_2))f_s(z_1)f_s(z_2))\,dz_1dz_2ds. \end{aligned} \tag{2.24}$$

Remark. *The fact that the correlations in the noise decay as $|z_1 - z_2|^{-\alpha}$ can be expected from the results in [BEK06] (see also [BEK10]). The authors prove that, if N is a Poisson point process*

on $\mathbb{R}^d \times \mathbb{R}_+$ whose intensity measure is of the form $dxf(r)dr$ with $f(r) \sim \frac{C}{r^{1+\alpha+d}}$, one can define a generalized random field X on the space of signed measures on $\mathbb{R}^d$ with finite total variation by

$$\langle X, \mu \rangle = \int_{\mathbb{R}^d \times \mathbb{R}_+} \mu(B(x,r))N(dx,dr).$$

Under a suitable scaling of the radius and of the intensity measure, it is shown that the fluctuations of X converge (in the sense of finite dimensional distributions) to a centred Gaussian random linear functional W^α with

$$\mathbb{E}\left[W^\alpha(\mu)W^\alpha(\nu)\right] = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |z_1 - z_2|^{-\alpha}\, \mu(dz_1)\nu(dz_2).$$

(The notation has been changed so as to fit that of our setting; in [BEK06], $\beta = \alpha + d$.) The second factor in (2.24) *(besides $K_\alpha(z_1, z_2)$) comes from the expression for the covariation between the jump in the allele frequency during a reproduction event at the two locations z_1 and z_2 (see* (2.37) *and the definition of $[f]_\alpha$ in* (2.21)*).*

We show in Section 2.3 that Theorem 2.2 is a consequence of Theorem 2.5. The latter is a result on sequences of solutions to a martingale problem and is proved in Section 2.5. In [EVY14], the authors showed that in the special case of genic selection (as in Definition 2.1.3), for $d \geq 2$, a sequence of spatially averaged versions of $\left(\mathbf{q}_t^N\right)_{t\geq 0}$ converges in distribution in $\mathrm{D}([0,\infty), \Xi)$ to the solution of (2.23).

2.2.3 Drift load for a spatially structured population

We shall illustrate the application of our results by studying the drift load in the SLFVS with overdominance as in Definition 2.1.4, in the case of bounded radii.

As in Section 2.2.1, fix u, s_1, s_2, ν_1, ν_2 in $(0,1]$ and $R > 0$ such that $s_1 + s_2 + \nu_1 + \nu_2 < 1$, take two sequences $(\varepsilon_N)_{N\geq 1}$, $(\delta_N)_{N\geq 1}$ of positive real numbers in $(0,1]$ decreasing to zero, and set

$$u_N = \varepsilon_N u, \qquad r_N = \delta_N R \qquad s_{i,N} = \delta_N^2 s_i, \qquad \nu_{i,N} = \delta_N^2 \nu_i \tag{2.25}$$

for $i = 1, 2$. Then for $N \geq 1$, let $\left(q_t^N\right)_{t\geq 0}$ be the SLFVS following the dynamics of Definition 2.1.4 with impact parameter u_N, radius of reproduction events R, selection parameters $s_{i,N}$ and mutation parameters $\nu_{i,N}$, started from some initial condition q_0^N.

One thing to note is that for our results to hold, we need to make sure that the allele frequencies do not get "stuck" - even locally - at the boundaries (*i.e.* upon reaching 0 or 1), which could significantly slow down the convergence to the equilibrium frequency. For this reason we choose to assume that during some mutation reproduction events the type of the offspring can differ from that of its parent. This will not affect the results in any other way provided that the mutation parameters are negligible compared to the selection parameters.

Now let

$$F(w) = w(1-w)(w - \frac{s_2}{s_1+s_2}) + \frac{\nu_1}{s_1+s_2}w - \frac{\nu_2}{s_1+s_2}(1-w). \tag{2.26}$$

We shall see in Section 2.3 that this function plays the same role as in the haploid case. Note that F

satisfies the following conditions:

$$\exists \lambda \in [0,1] : F(\lambda) = 0; \tag{2.27}$$

furthermore there is only one such λ and it satisfies

$$0 < \lambda < 1 \text{ and } F'(\lambda) > 0. \tag{2.28}$$

For the function F given in (2.26), λ is given by

$$\lambda = \frac{s_2}{s_1 + s_2} + \mathcal{O}\left(\frac{\nu_1 + \nu_2}{s_1 + s_2}\right) \tag{2.29}$$

(since the first term is a solution of $w(1-w)(w - \frac{s_2}{s_1+s_2}) = 0$).

Let us define $K^N(t,x)$, the local mean fitness at a point $x \in \mathbb{R}^d$, as the expected fitness of an individual formed by fusing two gametes chosen uniformly at random from $B(x,R)$ at time $t \geq 0$. In other words, its two copies of the gene are sampled independently by selecting two parental locations y_1 and y_2 uniformly at random in $B(x,R)$ and then types according to $q_t(y_1)$ and $q_t(y_2)$. Then (see [Rob70]),

$$\begin{aligned}
K^N(t,x) &= \mathbb{E}\left[(1-s_{1,N})\overline{q_t^N}(x,R)^2 + 2\overline{q_t^N}(x,R)(1-\overline{q_t^N}(x,R)) + (1-s_{2,N})(1-\overline{q_t^N}(x,R))^2\right] \\
&= 1 - \mathbb{E}\left[s_{1,N}\overline{q_t^N}(x,R)^2 + s_{2,N}(1-\overline{q_t^N}(x,R))^2\right] \\
&= 1 - \frac{s_{1,N}s_{2,N}}{s_{1,N}+s_{2,N}} - (s_{1,N}+s_{2,N})\mathbb{E}\left[\left(\overline{q_t^N}(x,R) - \frac{s_2}{s_1+s_2}\right)^2\right] \\
&= 1 - \frac{s_{1,N}s_{2,N}}{s_{1,N}+s_{2,N}} - (s_{1,N}+s_{2,N})\mathbb{E}\left[\left(\overline{q_t^N}(x,R) - \lambda\right)^2\right] \\
&\qquad - (s_{1,N}+s_{2,N})\left(\lambda - \frac{s_2}{s_1+s_2}\right)\left(2\mathbb{E}\left[\overline{q_t^N}(x,R)\right] - \lambda - \frac{s_2}{s_1+s_2}\right).
\end{aligned}$$

The first term $\frac{s_{1,N}s_{2,N}}{s_{1,N}+s_{2,N}}$ is the segregation load mentioned in the introduction, and it is of order δ_N^2. The second term is then the local drift load, which we aim to estimate at large times for large N. The last term is an error due to the mutation events; by (2.29), it is $\mathcal{O}\left(\delta_N^2(\nu_1+\nu_2)\right)$, and so negligible compared to the segregation load if $\frac{\nu_1+\nu_2}{s_1+s_2}$ is small. Let us set

$$\Delta^N(t,x) = (s_{1,N}+s_{2,N})\mathbb{E}\left[\left(\overline{q_t^N}(x,R) - \lambda\right)^2\right]. \tag{2.30}$$

The following theorem is proved in Section 2.6 using some of the intermediate results used to prove Theorem 2.1.

Theorem 2.3. *Suppose that $q_0^N(x) = \lambda$ for all x and assume that $\varepsilon_N = o\left(\delta_N^4\right)$. There exists a constant $C > 0$, depending only on the dimension d, such that, for all $x \in \mathbb{R}^d$, as $N, t \to \infty$, if t grows fast enough that $\varepsilon_N t \to \infty$ if $d \geq 3$ and $\varepsilon_N \delta_N^2 t \to \infty$ if $d \leq 2$,*

$$\Delta^N(t,x) \underset{N,t\to\infty}{\sim} C\varepsilon_N \delta_N^2 c_N,$$

where

$$c_N = \begin{cases} 1 & \text{if } d \geq 3, \\ \left|\log \delta_N^2\right| & \text{if } d = 2, \\ \delta_N^{-1} & \text{if } d = 1. \end{cases} \tag{2.31}$$

Assumption (2.27)-(2.28) is crucial in [Nor74a], which serves as a basis for this result. In fact this condition ensures that λ is the only equilibrium point for the allele frequency, and that it is stable.

Remark. *We chose to start the process from the equilibrium frequency λ - i.e. very near stationarity - but we need not do so. The same result can be obtained starting from an arbitrary initial condition, provided we let t grow sufficiently fast that the process reaches stationarity quickly enough. The corresponding centering term f^N is then defined as in* (2.15)*, and* (2.27)-(2.28) *ensures that it converges to λ exponentially quickly. Starting from λ simplifies the proof as in this case, for all $t \geq 0$, $f_t^N = \lambda$.*

In the non-spatial setting of the Λ-Fleming Viot process, a simplified version of the proof of Theorem 2.3 shows that the drift load is asymptotically proportional to u_N. We can see u_N as being inversely proportional to the neighbourhood size, in other words the probability that two individuals had a common parent in the previous generation (see [BEV13a] for details). This agrees with Robertson's estimate [Rob70] of $(4N)^{-1}$, where N is the total population size in a panmictic population. Note that this estimate is independent of the strength of selection. This can be seen as the result of a trade off between selection and genetic drift: if selection is weak, the allele frequency can be far from the equilibrium whereas if selection is stronger, the allele frequency stays nearer to the equilibrium and in both cases the mean fitness of the population is the same.

For spatially structured populations, however, Theorem 2.3 shows that the local drift load is significantly smaller than in the non-spatial setting and does depend on the strength of natural selection. For example, if a population lives in a geographical space of dimension 2, the corresponding drift load will be of order $\varepsilon_N \delta_N^2 \left|\log \delta_N^2\right|$, and since $u_N = u\varepsilon_N$ and $s_N := s_{1,N} + s_{2,N} = \delta_N^2(s_1 + s_2)$, it is of order $u_N s_N \left|\log s_N\right|$. Moreover, we see a strong effect of dimension on this estimate. Populations living in a space with a higher dimension have a reduced drift load compared to populations evolving in smaller dimensions. This result illustrates the fact that, in a higher dimension, migration is more efficient at preventing the allele frequencies from being locally far from the equilibrium frequency. It turns out from the proof that this is linked to the recurrence properties of Brownian motion.

Remark (Drift load in the stable case)**.** *If one considers instead the SLFVS with stable radii of reproduction events, under similar conditions to those in Theorem 2.2, one finds that for all $d \geq 1$ and $\alpha \in (0, 2 \wedge d)$, $\Delta^N(t, x)$ is asymptotically equivalent to a constant times $u_N s_N \left|\log s_N\right|$.*

2.3 Martingale problems for the SLFVS

This section provides the basic ingredients for the proofs of Theorems 2.1 and 2.2. In Subsection 2.3.1, we prove that the SLFVS satisfies a martingale problem. In Subsections 2.3.2 and 2.3.3, we study the martingale problem for the rescaled version of this process, in the fixed radius case and in the stable radii case, and state general convergence results for processes satisfying these martingale problems. Theorems 2.1 and 2.2 are direct consequences of these results.

2.3.1 The martingale problem for the SLFVS

Let $\left(q_t^N\right)_{t\geq 0}$ be defined (as in Sections 2.2.1 and 2.2.2) as the SLFVS as in Definition 2.1.2 with impact parameter u_N, distribution of reproduction event radii given by $\mu(dr)$, selection parameter s_N and selection function F. Let $(\mathcal{F}_t)_{t\geq 0}$ denote the natural filtration of this process.

For $p > 0$ and $\phi : \mathbb{R}^d \to \mathbb{R}$, let $\|\phi\|_p = (\int_{\mathbb{R}^d} |\phi(x)|^p dx)^{1/p}$.

Proposition 2.3.1. *Suppose that $\int_0^\infty V_r^2 \mu(dr) < \infty$. For any $\phi : \mathbb{R}^d \to \mathbb{R}$ in $L^{1,\infty}(\mathbb{R}^d)$,*

$$\left\langle q_t^N, \phi \right\rangle - \left\langle q_0^N, \phi \right\rangle - \int_0^t \int_0^\infty u_N V_r \left\{ \left\langle q_s^N, \overline{\overline{\phi}}(r) - \phi \right\rangle - s_N \left\langle \overline{F(\overline{q_s^N})}(r), \phi \right\rangle \right\} \mu(dr) ds \tag{2.32}$$

defines a (mean zero) square integrable $\mathcal{F}_t$-martingale with (predictable) variation process

$$\int_0^t \int_0^\infty u_N^2 V_r^2 \int_{(\mathbb{R}^d)^2} \phi(z_1)\phi(z_2)\sigma^{(r)}_{z_1,z_2}(q_s^N) dz_1 dz_2 \mu(dr) ds + \mathcal{O}\left(t u_N^2 s_N \|\phi\|_2^2 \right), \tag{2.33}$$

where

$$\sigma^{(r)}_{z_1,z_2}(q) = \frac{1}{V_r^2} \int_{B(z_1,r)\cap B(z_2,r)} [\overline{q}(x,r)(1-q(z_1))(1-q(z_2)) + (1-\overline{q}(x,r))q(z_1)q(z_2)] dx. \tag{2.34}$$

Proposition 2.3.1 can be seen as a way to write q_t as the sum of the effects of the different evolutionary forces at play in this model. The term $\overline{\overline{\phi}} - \phi$ represents migration, while the term involving the function F in (2.32) accounts for the bias introduced during selective events. As for the martingale term, it corresponds to the stochasticity at each reproduction event, which is called *genetic drift*.

Proof of Proposition 2.3.1. We drop the superscript N from q^N in this proof. Let $\mathbb{P}_{t,x,r}$ (resp. $\mathbb{P}^S_{t,x,r}$) denote the distribution of the parental type k at a reproduction event $(t,x,r) \in \Pi$ (resp. in Π^S). Then, from the definition of $(q_t)_{t\geq 0}$,

$$\begin{aligned}\lim_{\delta t \downarrow 0} \frac{1}{\delta t} \mathbb{E}\left[\langle q_{t+\delta t}, \phi\rangle - \langle q_t, \phi\rangle \mid q_t = q \right] = \\ \int_{\mathbb{R}^d} dx \int_0^\infty \mu(dr) \int_{\mathbb{R}^d} \phi(z) u_N \mathbb{1}_{\{|x-z|<r\}} \big\{ (1-s_N)\mathbb{E}_{t,x,r}\left[\mathbb{1}_{\{k=a\}} - q_t(z) \mid q_t = q \right] \\ + s_N \mathbb{E}^S_{t,x,r}\left[\mathbb{1}_{\{k=a\}} - q_t(z) \mid q_t = q \right] \big\} dz. \end{aligned} \tag{2.35}$$

Recall from Definition 2.1.2 that $\mathbb{P}_{t,x,r}(k = a \mid q_t = q) = \overline{q}(x,r)$ and

$$\mathbb{P}^S_{t,x,r}(k = a \mid q_t = q) = \overline{q}(x,r) - F(\overline{q}(x,r)).$$

Integrating with respect to the variable x over $B(z,r)$ then yields

$$\begin{aligned}\lim_{\delta t \downarrow 0} \frac{1}{\delta t} \mathbb{E}\left[\langle q_{t+\delta t}, \phi\rangle - \langle q_t, \phi\rangle \mid q_t = q \right] \\ = \int_0^\infty \mu(dr) u_N V_r \int_{\mathbb{R}^d} \phi(z) \left\{ (\overline{\overline{q}}(z,r) - q(z)) - s_N \overline{F(\overline{q})}(z,r) \right\} dz. \end{aligned}$$

Thus (2.32) indeed defines a martingale - see for example [EK86, Proposition 4.1.7] (we can change the order of integration to do the averaging on ϕ instead of q in the first term). To compute its variation process, write

$$\lim_{\delta t\downarrow 0}\frac{1}{\delta t}\mathbb{E}\left[\left(\langle q_{t+\delta t},\phi\rangle-\langle q_t,\phi\rangle\right)^2\,\middle|\,q_t=q\right]=\int_{\mathbb{R}^d}\int_0^\infty\int_{(\mathbb{R}^d)^2}\phi(z_1)\phi(z_2)u_N^2\mathbb{1}_{\left\{\substack{|z_1-x|<r\\|z_2-x|<r}\right\}}$$
$$\left\{(1-s_N)\mathbb{E}_{t,x,r}\left[(\mathbb{1}_{\{k=a\}}-q_t(z_1))(\mathbb{1}_{\{k=a\}}-q_t(z_2))\,\middle|\,q_t=q\right]\right.$$
$$\left.+\,s_N\,\mathbb{E}^S_{t,x,r}\left[(\mathbb{1}_{\{k=a\}}-q_t(z_1))(\mathbb{1}_{\{k=a\}}-q_t(z_2))\,\middle|\,q_t=q\right]\right\}dz_1dz_2\mu(dr)dx. \quad (2.36)$$

But

$$\mathbb{E}_{t,x,r}\left[(\mathbb{1}_{\{k=a\}}-q_t(z_1))(\mathbb{1}_{\{k=a\}}-q_t(z_2))\,\middle|\,q_t=q\right]$$
$$=\overline{q}(x,r)(1-q(z_1))(1-q(z_2))+(1-\overline{q}(x,r))q(z_1)q(z_2), \quad (2.37)$$

and the other term within the curly brackets is $\mathcal{O}(s_N)$. Thus, integrating with respect to x and using (2.34), we recover

$$\lim_{\delta t\downarrow 0}\frac{1}{\delta t}\mathbb{E}\left[\left(\langle q_{t+\delta t},\phi\rangle-\langle q_t,\phi\rangle\right)^2\,\middle|\,q_t=q\right]=\int_0^\infty\mu(dr)u_N^2V_r^2\int_{(\mathbb{R}^d)^2}\phi(z_1)\phi(z_2)\sigma^{(r)}_{z_1,z_2}(q)dz_1dz_2$$
$$+\,\mathcal{O}(s_N)\int_0^\infty\mu(dr)u_N^2V_r^2\int_{\mathbb{R}^d}\left(\frac{1}{V_r}\int_{B(x,r)}\phi(z)dz\right)^2dx. \quad (2.38)$$

By Jensen's inequality, $\int_{\mathbb{R}^d}\left(\frac{1}{V_r}\int_{B(x,r)}\phi(z)dz\right)^2dx\le\|\phi\|_2^2$ and the result follows from the assumption that $\int_0^\infty V_r^2\mu(dr)<\infty$. □

Now let $\left(q_t^N\right)_{t\ge 0}$ denote the SLFVS with overdominance as defined in Definition 2.1.4 with impact parameter u_N, radius of reproduction events R, selection parameters $s_{i,N}$ and mutation parameters $\nu_{i,N}$ defined in (2.25). Recall the definition of F in (2.26) and let $(\mathcal{F}_t)_{t\ge 0}$ denote the natural filtration of this process.

Proposition 2.3.2. *Let $s=s_1+s_2$ (and $s_N=s_{1,N}+s_{2,N}$). For any $\phi:\mathbb{R}^d\to\mathbb{R}$ in $L^{1,\infty}(\mathbb{R}^d)$,*

$$\left\langle q_t^N,\phi\right\rangle-\left\langle q_0^N,\phi\right\rangle-\int_0^t u_NV_r\left\{\left\langle q_s^N,\overline{\overline{\phi}}(R)-\phi\right\rangle-s_N\left\langle\overline{F(\overline{q_s^N})}(R),\phi\right\rangle\right\}ds \quad (2.39)$$

defines a (mean zero) square integrable $\mathcal{F}_t$-martingale with (predictable) variation process

$$\int_0^t u_N^2V_R^2\int_{(\mathbb{R}^d)^2}\phi(z_1)\phi(z_2)\rho^{(R)}_{z_1,z_2}(q_s^N)dz_1dz_2ds+\mathcal{O}\left(tu_N^2\delta_N^2\left\|\phi\right\|_2^2\right), \quad (2.40)$$

where

$$\rho^{(r)}_{z_1,z_2}(q)=\frac{1}{V_r^2}\int_{B(z_1,r)\cap B(z_2,r)}[\overline{q}(x,r)^2(1-q(z_1))(1-q(z_2))$$
$$+\,2\overline{q}(x,r)(1-\overline{q}(x,r))(\tfrac{1}{2}-q(z_1))(\tfrac{1}{2}-q(z_2))+(1-\overline{q}(x,r))^2q(z_1)q(z_2)]dx. \quad (2.41)$$

Proof. Suppose a reproduction event hits the ball $B(x,r)$ at time t, and let $w = \overline{q_{t^-}^N}(x,r)$. Then,

$$\begin{aligned}
\mathbb{P}\left(\{k_1,k_2\} = \{A_1,A_1\}\right) &= (1 - s_{1,N} - s_{2,N} - \nu_{1,N} - \nu_{2,N})w^2 + s_{1,N}w^4 + s_{2,N}w^2(1 + (1-w)^2) \\
&\quad + \nu_{2,N}, \\
\mathbb{P}\left(\{k_1,k_2\} = \{A_1,A_2\}\right) &= (1 - s_{1,N} - s_{2,N} - \nu_{1,N} - \nu_{2,N})2w(1-w) + s_{1,N}2w(1-w)(1+w^2) \\
&\quad + s_{2,N}2w(1-w)(1+(1-w)^2), \\
\mathbb{P}\left(\{k_1,k_2\} = \{A_2,A_2\}\right) &= (1 - s_{1,N} - s_{2,N} - \nu_{1,N} - \nu_{2,N})(1-w)^2 + s_{1,N}(1-w)^2(1+w^2) \\
&\quad + s_{2,N}(1-w)^4 + \nu_{1,N}.
\end{aligned}$$

Hence

$$\begin{aligned}
&\mathbb{E}\left[\tfrac{1}{2}\left(\mathbb{1}_{\{k_1=A_1\}} + \mathbb{1}_{\{k_2=A_1\}}\right) \,\middle|\, \overline{q_{t^-}}(x,r) = w\right] \\
&\qquad = w + w(1-w)(-s_{1,N}w + s_{2,N}(1-w)) - \nu_{1,N}w + \nu_{2,N}(1-w) \\
&\qquad = w - s_N F(w),
\end{aligned}$$

recalling the definition of F in (2.26) and $s_{i,N}$, $\nu_{i,N}$ in (2.25) and that $s_N = s_{1,N} + s_{2,N}$. Therefore

$$\mathbb{E}_{t,x,r}\left[\tfrac{1}{2}\left(\mathbb{1}_{\{k_1=A_1\}} + \mathbb{1}_{\{k_2=A_1\}}\right) - q_{t^-}^N(z) \,\middle|\, q_{t^-}^N = q\right] = \overline{q}(x,r) - q(z) - s_N F(\overline{q}(x,r)).$$

It follows as in the proof of Proposition 2.3.1 that (2.39) is a martingale. The result for the variation process also follows as in the proof of Proposition 2.3.1 (all terms containing either $s_{i,N}$ or $\nu_{i,N}$ are collected in the error term in (2.40)). Note that $\sigma^{(r)}$ is replaced by $\rho^{(r)}$ in order to account for the fact that (2.7) is replaced by (2.10). □

Remark. *If q were continuous then as $r \to 0$, $\sigma_{z_1,z_2}^{(r)}(q) \to \delta_{z_1=z_2}q(z_1)(1-q(z_1))$ and $\rho_{z_1,z_2}^{(r)}(q) \to \frac{1}{2}\delta_{z_1=z_2}q(z_1)(1-q(z_1))$. The factor of $1/2$ represents the doubling of effective population size for a diploid population compared to a haploid one.*

2.3.2 The rescaled martingale problem - Fixed radius case

As at the start of Subsection 2.2.1, let $(\varepsilon_N)_{N\geq 1}$, $(\delta_N)_{N\geq 1}$ be sequences in $(0,1]$ decreasing towards zero, and let $F : \mathbb{R} \to \mathbb{R}$.

Definition 2.3.3 (Martingale Problem (M1))**.** *Given $(\varepsilon_N)_{N\geq 1}$, $(\delta_N)_{N\geq 1}$ and F, let $\eta_N = \varepsilon_N\delta_N^2$, $\tau_N = \varepsilon_N^2\delta_N^d$ and $r_N = \delta_N R$. Then for $N \geq 1$, we say that a Ξ-valued process $(w_t^N)_{t\geq 0}$ satisfies the martingale problem (M1) if for all ϕ in $L^{1,\infty}(\mathbb{R}^d)$,*

$$\left\langle w_t^N, \phi\right\rangle - \left\langle w_0, \phi\right\rangle - \eta_N u V_R \int_0^t \left\{ \frac{2R^2}{d+2}\left\langle w_s^N, \mathcal{L}^{(r_N)}\phi\right\rangle - s\left\langle \overline{F(\overline{w_s^N})}(r_N), \phi\right\rangle \right\} ds \tag{2.42}$$

defines a (mean zero) square-integrable martingale with (predictable) variation process

$$\tau_N u^2 V_R^2 \int_0^t \int_{(\mathbb{R}^d)^2} \phi(z_1)\phi(z_2)\sigma_{z_1,z_2}^{(r_N)}(w_s^N)dz_1dz_2ds + \mathcal{O}\left(t\tau_N\delta_N^2 \left\|\phi\right\|_2^2\right). \tag{2.43}$$

Remark. *Of course, one cannot expect uniqueness to hold for this martingale problem, due to the unspecified error term in* (2.43)*. In the limit when $N \to \infty$, however, the error terms will vanish.*

Let $\left(q_t^N\right)_{t\geq 0}$ be defined as at the start of Section 2.2.1. Set $w_t^N(x) = q_t^N(x/\delta_N)$.

Proposition 2.3.4. *For each N, the process $\left(w_t^N\right)_{t\geq 0}$ satisfies the martingale problem (M1).*

Proof. From Proposition 2.3.1, we know that, for $\phi \in L^{1,\infty}(\mathbb{R}^d)$,

$$\left\langle q_t^N, \phi\right\rangle = \left\langle q_0^N, \phi\right\rangle + u_N V_R \int_0^t \left\{ \left\langle q_s^N, \overline{\overline{\phi}}(R) - \phi \right\rangle - s_N \left\langle \overline{F(\overline{q_s^N})}(R), \phi \right\rangle \right\} ds + \mathcal{M}_t^N(\phi),$$

where $\mathcal{M}_t^N(\phi)$ is a martingale. By a change of variables,

$$\left\langle w_t^N, \phi\right\rangle = \delta_N^d \left\langle q_t^N, \phi^{(\delta_N)} \right\rangle, \tag{2.44}$$

with $\phi^{(\delta)}(x) = \phi(\delta x)$. Also,

$$\delta_N^d \left\langle q_s^N, \overline{\overline{\phi^{(\delta_N)}}}(R) \right\rangle = \left\langle w_s^N, \overline{\overline{\phi}}(\delta_N R) \right\rangle \quad \text{and} \quad \delta_N^d \left\langle \overline{F(\overline{q_s^N})}(R), \phi^{(\delta_N)} \right\rangle = \left\langle \overline{F(\overline{w_s^N})}(\delta_N R), \phi \right\rangle. \tag{2.45}$$

Thus, recalling the definition of the operator $\mathcal{L}^{(r)}$ in (2.14) and the initial condition $q_0^N(x) = w_0(\delta_N x)$, we have

$$\left\langle w_t^N, \phi\right\rangle = \langle w_0, \phi\rangle + \varepsilon_N \delta_N^2 u V_R \int_0^t \left\{ \frac{2R^2}{d+2} \left\langle w_s^N, \mathcal{L}^{(\delta_N R)}\phi \right\rangle - s \left\langle \overline{F(\overline{w_s^N})}(\delta_N R), \phi \right\rangle \right\} ds + \delta_N^d \mathcal{M}_t^N(\phi^{(\delta_N)}).$$

Moreover, by a change of variables in the variation process given in (2.33),

$$\delta_N^{2d} \left\langle \mathcal{M}^N(\phi^{(\delta_N)}) \right\rangle_t = \varepsilon_N^2 u^2 V_R^2 \int_0^t \int_{(\mathbb{R}^d)^2} \phi(z_1)\phi(z_2)\sigma^{(R)}_{z_1/\delta_N, z_2/\delta_N}(q_s^N) dz_1 dz_2 ds + \mathcal{O}\left(t\varepsilon_N^2 \delta_N^2 \delta_N^d \left\|\phi\right\|_2^2 \right), \tag{2.46}$$

and

$$\sigma^{(R)}_{z_1/\delta_N, z_2/\delta_N}(q_s^N) = \delta_N^d \sigma^{(\delta_N R)}_{z_1, z_2}(w_s^N).$$

Hence w^N satisfies the martingale problem (M1). □

Proposition 2.3.4 is the main ingredient in the proof of Theorem 2.1. In fact we shall now see that under suitable conditions on the parameters $(\varepsilon_N)_{N\geq 1}$ and $(\delta_N)_{N\geq 1}$, the function F and the initial condition w_0, any sequence of processes $\left(w_t^N\right)_{t\geq 0}$ satisfying the martingale problem (M1) in Definition 2.3.3 will also satisfy a result analogous to Theorem 2.1. If τ_N is of a smaller order than η_N, w^N can be expected to be asymptotically deterministic (on a suitable time-scale), and we can study its fluctuations around a deterministic centering term. Define $f^N : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ as in (2.15). Quite naturally, this corresponds to equating (2.42) to zero and making its time-scale fit that of the limiting process.

Since the operator $\mathcal{L}^{(r)}$ approximates the Laplacian as $r \to 0$ (see Proposition 2.7.2), f_t^N converges to $f : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ as $N \to \infty$, where f_t is the solution of the following equation,

$$\begin{cases} \dfrac{\partial f_t}{\partial t} = uV_R \left(\dfrac{R^2}{d+2} \Delta f_t - sF(f_t) \right), \\ f_0 = w_0. \end{cases} \tag{2.47}$$

(See Proposition 2.4.5 for a precise statement.) The following result is proved in Section 2.4.

Theorem 2.4. *Suppose that $\left(w_t^N\right)_{t\geq 0}$ is a Ξ-valued process which satisfies the martingale problem (M1) in Definition 2.3.3 for some smooth, bounded $F : \mathbb{R} \to \mathbb{R}$ with bounded derivatives of order up to four and $(\delta_N)_N$, $(\varepsilon_N)_N$ converging to zero as $N \to \infty$. Moreover, suppose*

$$\tau_N / \eta_N = o\left(\delta_N^{2d}\right). \tag{2.48}$$

Suppose also that w_0 has uniformly bounded derivatives of up to the fourth order and that there exists α_N such that the jumps of $\left(w_t^N\right)_{t\geq 0}$ are (almost surely) dominated by

$$\sup_{t\geq 0} \left| \left\langle w_t^N, \phi \right\rangle - \left\langle w_{t^-}^N, \phi \right\rangle \right| \leq \alpha_N \left\| \phi \right\|_1 \tag{2.49}$$

for every $\phi \in L^{1,\infty}(\mathbb{R}^d)$, with $\alpha_N^2 = o\left(\tau_N / \eta_N\right)$. Then

$$\left(w_{t/\eta_N}^N \right)_{t\geq 0} \underset{N\to\infty}{\overset{L^1, P}{\longrightarrow}} (f_t)_{t\geq 0} \tag{2.50}$$

in $\left(\mathrm{D}\left([0,T], \Xi\right), d\right)$ for every $T > 0$ with d given by Definition 2.1.1. In addition,

$$Z_t^N = (\eta_N / \tau_N)^{1/2} (w_{t/\eta_N}^N - f_t^N)$$

defines a sequence of distribution-valued processes which converges in distribution in $\mathrm{D}\left([0,T], \mathcal{S}'(\mathbb{R}^d)\right)$ to the solution of the following stochastic partial differential equation,

$$\begin{cases} dz_t = uV_R [\dfrac{R^2}{d+2} \Delta z_t - sF'(f_t) z_t] dt + uV_R \sqrt{f_t(1-f_t)} \cdot dW_t, \\ z_0 = 0, \end{cases} \tag{2.51}$$

W being a space-time white noise.

Theorem 2.1 is now a direct consequence.

Proof of Theorem 2.1. Recall that $(\mathbf{q}_t^N)_{t\geq 0}$ is defined in (2.11) as a rescaling of $\left(q_t^N\right)_{t\geq 0}$, and that by Proposition 2.3.4, letting $w_t^N(x) = q_t^N(x/\delta_N)$, $\left(w_t^N\right)_{t\geq 0}$ satisfies the martingale problem (M1). Also $\tau_N / \eta_N = o\left(\delta_N^{2d}\right)$ follows from $\varepsilon_N = o(\delta_N^{d+2})$, and the bound on the jumps (2.49) holds with $\alpha_N = \varepsilon_N u$ by (2.7). Hence Theorem 2.4 applies and the result follows by noting that $w_{t/\eta_N}^N = \mathbf{q}_t^N$. □

The proof of Theorem 2.4 can be found in full detail in Section 2.4, but, in order to shed some light on the limiting equations that we obtain and to identify the difficulties in proving this result, let us outline the first calculations involved in the proof. As in [Kur71], we use bounds on the martingale

(2.42) to show the convergence of $\left(w^N_{t/\eta_N}\right)_{t\geq 0}$. When properly rescaled, this martingale converges to a continuous Gaussian martingale, implying the convergence of the fluctuation process $\left(Z^N_t\right)_{t\geq 0}$.

For ease of notation, we shall set the constants uV_R, $2R^2/(d+2)$ and s to 1 in the definition of (M1). Let $\mathbb{M}^N_t(\phi)$ denote $\tau_N^{-1/2}$ times the martingale defined in (2.42). Formally, we can then write (M1) as

$$dw^N_t = \eta_N \left[\mathcal{L}^{(r_N)} w^N_t - \overline{F(\overline{w^N_t})}(r_N)\right] dt + \tau_N^{1/2} d\mathbb{M}^N_t.$$

Now set

$$M^N_t(\phi) = \eta_N^{1/2} \mathbb{M}^N_{t/\eta_N}(\phi).$$

(This Brownian scaling is not surprising since in the SLFVS case $\mathbb{M}^N$ is essentially an integral against a compensated Poisson point process, and we expect M^N to converge to an integral against white noise.) Replacing t by t/η_N above, we have

$$dw^N_{t/\eta_N} = \left[\mathcal{L}^{(r_N)} w^N_{t/\eta_N} - \overline{F(\overline{w^N_{t/\eta_N}})}(r_N)\right] dt + (\tau_N/\eta_N)^{1/2} dM^N_t.$$

Subtracting the equation

$$df^N_t = \left[\mathcal{L}^{(r_N)} f^N_t - \overline{F(\overline{f^N_t})}(r_N)\right] dt,$$

and multiplying by $(\eta_N/\tau_N)^{1/2}$ on both sides, we obtain

$$dZ^N_t = \left[\mathcal{L}^{(r_N)} Z^N_t - (\eta_N/\tau_N)^{1/2} \overline{\left(F(\overline{w^N_{t/\eta_N}}) - F(\overline{f^N_t})\right)}(r_N)\right] dt + dM^N_t. \tag{2.52}$$

Since the function $F : \mathbb{R} \to \mathbb{R}$ is smooth, for $k \in \{1, 2\}$ and $x, y \in [0, 1]$, we can define the following:

$$R_k(x, y) = \int_0^1 \frac{t^{k-1}}{(k-1)!} F^{(k)}(x + t(y - x))dt. \tag{2.53}$$

Then R_k is continuous and bounded by $\frac{1}{k!} \left\|F^{(k)}\right\|_\infty$. In addition, by Taylor's formula,

$$F(x) = F(y) + (x - y)R_1(x, y), \tag{2.54}$$

$$F(x) = F(y) + (x - y)F'(y) + (x - y)^2 R_2(x, y). \tag{2.55}$$

Substituting the second relation into (2.52) yields

$$dZ^N_t = \left[\mathcal{L}^{(r_N)} Z^N_t - \overline{\overline{Z^N_t} F'(\overline{f^N_t})}(r_N) - (\tau_N/\eta_N)^{1/2} \overline{(\overline{Z^N_t})^2 R_2(\overline{w^N_{t/\eta_N}}, \overline{f^N_t})}(r_N)\right] dt + dM^N_t.$$

In fact, this equality holds in mild form,

$$\begin{aligned}\left\langle Z^N_t, \phi\right\rangle = \int_0^t \left\langle Z^N_s, \mathcal{L}^{(r_N)}\phi - \overline{F'(\overline{f^N_s})\overline{\phi}}(r_N)\right\rangle ds \\ - (\tau_N/\eta_N)^{1/2} \int_0^t \left\langle (\overline{Z^N_s})^2, R_2(\overline{w^N_{s/\eta_N}}, \overline{f^N_s})\overline{\phi}(r_N)\right\rangle ds + M^N_t(\phi).\end{aligned} \tag{2.56}$$

(In other words, every step above can be done using the integral form, yielding (2.56).) We can see M^N as a martingale measure and, from a change of variables in (2.43), its covariation measure is

given by

$$Q^N(dz_1dz_2ds) = \sigma^{(r_N)}_{z_1,z_2}(w^N_{s/\eta_N})dz_1dz_2ds + \mathcal{O}\left(\delta_N^2\right)\delta_{z_1=z_2}(dz_1dz_2)ds. \tag{2.57}$$

Accordingly, we will sometimes write $M^N_t(\phi)$ as a stochastic integral (as defined in [Wal86, Chapter 2]),

$$M^N_t(\phi) = \int_0^t \int_{\mathbb{R}^d} \phi(x) M^N(dxds).$$

Note that we have linearised the drift term in (2.42) around the deterministic centering term, and that the remaining term (where R_2 appears) is the error due to this linearisation. The main difficulty in proving the convergence of Z^N is to control this error. At first sight, it would seem that the factor $(\tau_N/\eta_N)^{1/2}$ in front of it is enough to make it vanish in the limit. However, some care is needed in dealing with the quadratic term in the spatial integral. Since Z^N is going to converge as a distribution-valued process, its square does not make sense in the limit. The control of this term is achieved through Lemma 2.4.4, where we bound the square of the average of Z^N_t over a ball of radius r_N. It is for this purpose that we require that $\tau_N/\eta_N = o\left(r_N^{2d}\right)$.

Once this is done, we will be in a good position to prove the convergence of Z^N. Indeed, as r_N tends to zero, $\mathcal{L}^{(r_N)}\phi - \overline{F'(\overline{f^N_s})}\overline{\phi}(r_N)$ is well approximated by $\frac{1}{2}\Delta\phi - F'(f_s)\phi$ (see Proposition 2.7.2). We also prove that M^N converges to $\sqrt{f_t(1-f_t)}\cdot W_t$ (as defined in [Wal86, Chapter 2]) using the expression (2.57) for its covariance.

The proof of convergence of Z^N follows the classical strategy of proving that the sequence is tight before uniquely characterising its possible limit points. We are outside the safe borders of real-valued processes, but the theory presented in [Wal86] provides the main tools needed for the proof of our result. In particular, the argument relies heavily on Mitoma's Theorem (Theorem 6.13 in [Wal86]), which states that a sequence of processes $(X^n_t)_{t\geq 0}$, $n \geq 1$ with sample paths in $\mathrm{D}\left([0,T], \mathcal{S}'(\mathbb{R}^d)\right)$ a.s. is tight if and only if, for each $\phi \in \mathcal{S}(\mathbb{R}^d)$, the sequence of real-valued processes $(\langle X^n, \phi\rangle)_{n\geq 1}$ is tight in $\mathrm{D}([0,T],\mathbb{R})$ (see also Theorem 2.6).

2.3.3 The rescaled martingale problem - Stable radii case

For $\phi \in L^{1,\infty}(\mathbb{R}^d)$, and $\alpha \in (0, d\wedge 2)$, define the following norm

$$\|\phi\|^2_{(\alpha)} = \int_{(\mathbb{R}^d)^2} \phi(z_1)\phi(z_2)\left|z_1 - z_2\right|^{-\alpha} dz_1dz_2. \tag{2.58}$$

Let $(\varepsilon_N)_{N\geq 1}$, $(\delta_N)_{N\geq 1}$ be sequences in $(0,1]$ decreasing towards zero, and let $F : \mathbb{R} \to \mathbb{R}$.

Definition 2.3.5 (Martingale Problem (M2))**.** *Given $(\varepsilon_N)_{N\geq 1}$, $(\delta_N)_{N\geq 1}$ and F, let $\eta_N = \varepsilon_N\delta_N^\alpha$ and $\tau_N = \varepsilon_N^2\delta_N^\alpha$. Then for $N \geq 1$, we say that a Ξ-valued process $(w^N_t)_{t\geq 0}$ satisfies the martingale problem (M2) if for all ϕ in $L^{1,\infty}(\mathbb{R}^d)$,*

$$\left\langle w^N_t, \phi\right\rangle - \left\langle w_0, \phi\right\rangle - \eta_N u \int_0^t \left\{\left\langle w^N_s, \mathcal{D}^{\alpha,\delta_N}\phi\right\rangle - \frac{sV_1}{\alpha}\left\langle F^{(\delta_N)}(w^N_s), \phi\right\rangle\right\} ds \tag{2.59}$$

defines a (mean zero) square-integrable martingale with (predictable) variation process

$$\tau_N u^2 \int_0^t \int_{(\mathbb{R}^d)^2} \phi(z_1)\phi(z_2)\sigma^{(\alpha,\delta_N)}_{z_1,z_2}(w_s^N)dz_1dz_2ds + \mathcal{O}\left(t\tau_N\delta_N^\alpha \|\phi\|^2_{(\alpha)}\right), \tag{2.60}$$

where, for $\sigma^{(r)}$ defined as in (2.34),

$$\sigma^{(\alpha,\delta)}_{z_1,z_2}(w) = \int_{\frac{|z_1-z_2|}{2}\vee\delta}^\infty V_r^2 \sigma^{(r)}_{z_1,z_2}(w)\frac{dr}{r^{d+\alpha+1}}. \tag{2.61}$$

(Note that the remark about uniqueness made after Definition 2.3.3 also applies to the martingale problem (M2).)

Let $\left(q_t^N\right)_{t\geq 0}$ be defined as at the start of Section 2.2.2. Set $w_t^N(x) = q_t^N(x/\delta_N)$.

Proposition 2.3.6. *For each $N \geq 1$ the process $\left(w_t^N\right)_{t\geq 0}$ satisfies the martingale problem (M2).*

Proof. This is proved in a similar way to Proposition 2.3.4. Note that we cannot apply Proposition 2.3.1 directly, since in the stable case, $\int_0^\infty V_r^2\mu(dr) = \infty$ (recall that $V_r = |B(0,r)|$). However, its proof carries over with small adjustments.

For any $\phi \in L^{1,\infty}(\mathbb{R}^d)$,

$$\left\langle q_t^N, \phi\right\rangle = \left\langle q_0^N, \phi\right\rangle + \int_0^t \int_1^\infty u_N V_r \left\{\left\langle q_s^N, \overline{\overline{\phi}}(r) - \phi\right\rangle - s_N \left\langle \overline{F(\overline{q_s^N})}(r), \phi\right\rangle\right\}\frac{dr}{r^{d+\alpha+1}}ds + \mathcal{M}_t^N(\phi),$$

where $\mathcal{M}_t^N(\phi)$ is a martingale. Using (2.44) and (2.45), it follows that

$$\left\langle w_t^N, \phi\right\rangle = \langle w_0, \phi\rangle + \varepsilon_N u \int_0^t \int_1^\infty V_r\Big\{ \left\langle w_s^N, \overline{\overline{\phi}}(\delta_N r) - \phi\right\rangle - s_N \left\langle \overline{F(\overline{w_s^N})}(\delta_N r), \phi\right\rangle \Big\}\frac{dr}{r^{d+\alpha+1}}ds + \delta_N^d \mathcal{M}_t^N(\phi^{(\delta_N)}).$$

By the definition of $\mathcal{D}^{\alpha,\delta}$ in (2.19) and $V_r(x,y)$ in (2.12),

$$\begin{aligned}\int_1^\infty V_r(\overline{\overline{\phi}}(x,\delta_N r) - \phi(x))\frac{dr}{r^{d+\alpha+1}} &= \delta_N^\alpha \int_{\delta_N}^\infty \int_{\mathbb{R}^d} \frac{V_r(x,y)}{V_r}(\phi(y)-\phi(x))dy\frac{dr}{r^{d+\alpha+1}} \\ &= \delta_N^\alpha \mathcal{D}^{\alpha,\delta_N}\phi(x).\end{aligned}$$

Further, by (2.18),

$$\int_1^\infty V_r \overline{F(\overline{w_s^N})}(\delta_N r)\frac{dr}{r^{d+\alpha+1}} = \frac{V_1}{\alpha}F^{(\delta_N)}(w_s^N).$$

As a result,

$$\left\langle w_t^N, \phi\right\rangle = \langle w_0, \phi\rangle + \varepsilon_N\delta_N^\alpha u \int_0^t \left\{\left\langle w_s^N, \mathcal{D}^{\alpha,\delta_N}\phi\right\rangle - \frac{sV_1}{\alpha}\left\langle F^{(\delta_N)}(w_s^N), \phi\right\rangle\right\}ds + \delta_N^d\mathcal{M}_t^N(\phi^{(\delta_N)}).$$

For the predictable variation process, the term from the second line of (2.36) in the proof of Proposition 2.3.1 can be bounded by

$$\mathcal{O}(s_N)\, u_N^2 \int_{(\mathbb{R}^d)^2}\int_0^\infty \phi(z_1)\phi(z_2)V_r(z_1,z_2)\frac{dr}{r^{1+d+\alpha}}dz_1dz_2.$$

We recover the error term in (2.60) since $V_r(z_1, z_2) \leq r^d \mathbb{1}_{\left\{r \geq \frac{1}{2}|z_1 - z_2|\right\}}$. The first term in (2.60) follows from the definition of $\sigma^{(\alpha,r)}$ in (2.61). □

As in Subsection 2.3.2, we can now state a general result for a sequence of processes satisfying (M2) which implies Theorem 2.2. Let f^N be defined as in (2.22) and define f as the solution to

$$\begin{cases} \dfrac{\partial f_t}{\partial t} = u(\mathcal{D}^\alpha f_t - \frac{sV_1}{\alpha} F(f_t)), \\ f_0 = w_0. \end{cases} \tag{2.62}$$

The following result is proved in Section 2.5.

Theorem 2.5. *Suppose that $\left(w_t^N\right)_{t\geq 0}$ satisfies the martingale problem (M2) in Definition 2.3.5 for some smooth, bounded function $F : \mathbb{R} \to \mathbb{R}$ with bounded first and second derivatives and $(\delta_N)_N$, $(\varepsilon_N)_N$ converging to zero as $N \to \infty$. Moreover, suppose*

$$\tau_N/\eta_N = o\left(\delta_N^{2\alpha}\right). \tag{2.63}$$

Suppose also that w_0 has uniformly bounded derivatives of up to the second order and that there exists α_N such that the jumps of $\left(w_t^N\right)_{t\geq 0}$ are dominated by

$$\sup_{t\geq 0} \left|\left\langle w_t^N, \phi\right\rangle - \left\langle w_{t^-}^N, \phi\right\rangle\right| \leq \alpha_N \left\|\phi\right\|_1$$

for every $\phi \in L^{1,\infty}(\mathbb{R}^d)$, with $\alpha_N^2 = o\left(\tau_N/\eta_N\right)$. Then

$$\left(w_{t/\eta_N}^N\right)_{t\geq 0} \overset{\substack{L^1,P \\ \longrightarrow \\ N\to\infty}}{} (f_t)_{t\geq 0}$$

in $\left(\mathrm{D}\left([0,T], \Xi\right), d\right)$. In addition,

$$Z_t^N = (\eta_N/\tau_N)^{1/2}(w_{t/\eta_N}^N - f_t^N)$$

defines a sequence of distribution-valued processes which converges in distribution in $\mathrm{D}\left([0,T], \mathcal{S}'(\mathbb{R}^d)\right)$ to the solution of the following stochastic partial differential equation,

$$\begin{cases} dz_t = u[\mathcal{D}^\alpha z_t - \frac{sV_1}{\alpha} F'(f_t) z_t]dt + u \cdot dW_t^\alpha \\ z_0 = 0, \end{cases} \tag{2.64}$$

where W^α is a coloured noise with covariation measure given by (2.24).

Theorem 2.2 is now a direct consequence.

Proof of Theorem 2.2. Recall that $(\mathbf{q}_t^N)_{t\geq 0}$ is defined in (2.17) as a rescaling of $\left(q_t^N\right)_{t\geq 0}$, and that by Proposition 2.3.6, letting $w_t^N(x) = q_t^N(x/\delta_N)$, $\left(w_t^N\right)_{t\geq 0}$ satisfies the martingale problem (M2). Also $\tau_N/\eta_N = o\left(\delta_N^{2\alpha}\right)$ follows from $\varepsilon_N = o(\delta_N^{2\alpha})$, and the bound on the jumps (2.49) holds with $\alpha_N = \varepsilon_N u$ by (2.7). We conclude by applying Theorem 2.5 to $w_{t/\eta_N}^N = \mathbf{q}_t^N$. □

The proof of Theorem 2.5 will make use of the same ideas as in the proof of Theorem 2.4 and, to improve readability, the steps of the proof which are most similar to those in the Brownian case will be dealt with more quickly, going into details only when the two arguments differ.

2.4 The Brownian case - proof of Theorem 2.4

As in the sketch of the proof in Subsection 2.3.2, for ease of notation, we shall set the constants uV_R, $2R^2/(d+2)$ and s to 1 in the definition of (M1). Recall the expression for $\langle Z_t^N, \phi\rangle$ in (2.56); the next subsection shows how time-dependent test functions can be used to write $\langle Z_t^N, \phi\rangle$ as the sum of a stochastic integral against a martingale measure and a non-linear term. Subsection 2.4.2 will provide a bound on this quadratic term using a Gronwall estimate. We can then prove the convergence of the process $\left(w^N_{t/\eta_N}\right)_{t\geq 0}$ to $(f_t)_{t\geq 0}$ in Subsection 2.4.3.

The following result is used to reduce the convergence of distribution-valued processes to the convergence of a family of real-valued processes; it is a direct corollary of Mitoma's theorem [Wal86, Theorem 6.13].

Theorem 2.6 ([Wal86, Theorem 6.15])**.** *Let $(X^n)_{n\geq 1}$ be a sequence of processes with sample paths in $\mathrm{D}\left([0,T], \mathcal{S}'(\mathbb{R}^d)\right)$. Suppose*

i) for each $\phi \in \mathcal{S}(\mathbb{R}^d)$, $(\langle X^n, \phi\rangle)_{n\geq 1}$ is tight,

ii) for each $\phi_1, \ldots, \phi_k$ in $\mathcal{S}(\mathbb{R}^d)$ and $t_1, \ldots, t_k$ in $[0,T]$, the distribution of $(\langle X^n_{t_1}, \phi_1\rangle, \ldots, \langle X^n_{t_k}, \phi_k\rangle)$ converges weakly on $\mathbb{R}^k$.

Then there exists a process $(X_t)_{t\geq 0}$ with sample paths in $\mathrm{D}\left([0,T], \mathcal{S}'(\mathbb{R}^d)\right)$ such that X^n converges in distribution to X.

In order to apply this result to the sequence of distribution-valued processes $\left(Z^N\right)_{N\geq 1}$, we need to check that the two conditions (i) and (ii) are satisfied. The first one is proved in Subsection 2.4.4, thus implying the tightness of the sequence by Mitoma's theorem. Subsection 2.4.5 deals with the convergence of the martingale measure M^N (again as a distribution valued process, so this subsection will use Theorem 2.6). Finally condition (ii) is checked in Subsection 2.4.6.

In this section, in order to simplify the notation we often drop the sub- and superscripts N when there is no ambiguity; for instance, $\mathcal{L}$ should always be read $\mathcal{L}^{(r)}$, with $r = r_N$, and $w^N_{t/\eta}$ should be read w^N_{t/η_N}.

2.4.1 Time dependent test functions

Fix $\phi \in \mathcal{S}(\mathbb{R}^d)$. We consider time dependent test functions $\varphi : \mathbb{R}^d \times \{(s,t) : 0 \leq s \leq t \leq T\} \to \mathbb{R}$ such that (with a slight abuse of notation) $\varphi(s,t) \in L^\infty(\mathbb{R}^d)$ for all $0 \leq s \leq t$ and φ is continuously differentiable with respect to the time variables. Let us recall equation (2.56):

$$\langle Z_t^N, \phi\rangle = \int_0^t \left\langle Z_s^N, \mathcal{L}^{(r_N)}\phi - \overline{F'(\overline{f_s^N})\overline{\phi}}(r_N)\right\rangle ds \\ - (\tau_N/\eta_N)^{1/2} \int_0^t \left\langle (\overline{Z_s^N})^2, R_2(\overline{w^N_{s/\eta_N}}, \overline{f_s^N})\overline{\phi}(r_N)\right\rangle ds + M_t^N(\phi).$$

Adapting Exercise 5.1 of [Wal86], we obtain that for any time dependent test function φ,

$$\left\langle Z_t^N, \varphi(t,t)\right\rangle = \int_0^t \left\langle Z_s^N, \partial_s\varphi(s,t) + \mathcal{L}^{(r_N)}\varphi(s,t) - \overline{F'(\overline{f_s^N})\overline{\varphi(s,t)}}(r_N)\right\rangle ds$$
$$- (\tau_N/\eta_N)^{1/2}\int_0^t \left\langle (\overline{Z_s^N})^2, R_2(\overline{w_{s/\eta_N}^N}, \overline{f_s^N})\overline{\varphi(s,t)}(r_N)\right\rangle ds + \int_0^t\int_{\mathbb{R}^d}\varphi(x,s,t)M^N(dxds). \quad (2.65)$$

(To see this, use (2.56) to get an expression for $\left\langle Z_s^N, \partial_s\varphi(s,t)\right\rangle$ and integrate over s, using Fubini's theorem, then apply (2.56) again with $\phi = \varphi(t,t)$; see also Theorem 2.6 in [Wal86].) Suppose then that φ^N solves

$$\begin{cases} \partial_s\varphi^N(x,s,t) + \mathcal{L}^{(r_N)}\varphi^N(x,s,t) - \overline{F'(\overline{f_s^N})\overline{\varphi^N(s,t)}}(x,r_N) = 0, \\ \varphi^N(x,t,t) = \phi(x). \end{cases} \quad (2.66)$$

Equations (2.65) and (2.56) then yield

$$\left\langle Z_t^N, \phi\right\rangle = -(\tau_N/\eta_N)^{1/2}\int_0^t \left\langle (\overline{Z_s^N})^2, R_2(\overline{w_{s/\eta}^N}, \overline{f_s^N})\overline{\varphi^N(s,t)}\right\rangle ds + \int_0^t\int_{\mathbb{R}^d}\varphi^N(x,s,t)M^N(dxds). \quad (2.67)$$

Here we see that in the special case where F is linear, $R_2 = 0$ and it remains to prove the convergence of the stochastic integral of φ^N against the martingale measure M^N. Define φ as the solution to

$$\begin{cases} \partial_s\varphi(x,s,t) + \frac{1}{2}\Delta\varphi(x,s,t) - F'(f_s(x))\varphi(x,s,t) = 0, \\ \varphi(x,t,t) = \phi(x). \end{cases} \quad (2.68)$$

For a multi-index $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{N}_0^d$, let $|\beta| = \beta_1 + \ldots + \beta_d$ and for $g : \mathbb{R}^d \to \mathbb{R}$, let the derivative with respect to β be given by $\partial_\beta g(x) = \frac{\partial^{|\beta|}}{\partial x_1^{\beta_1}\ldots\partial x_d^{\beta_d}} g(x_1, \ldots, x_d)$.

Lemma 2.4.1. *Fix $T > 0$ and take $\phi \in \mathcal{S}(\mathbb{R}^d)$. There exists a unique solution φ^N to (2.66) in $L^\infty(\mathbb{R}^d \times \{(s,t) : 0 \le s \le t \le T\})$ which admits spatial derivatives of order up to four. Moreover, for any multi-index β with $0 \le |\beta| \le 4$,*

$$\sup_{0\le s\le t\le T} \left\|\partial_\beta\varphi^N(s,t)\right\|_q < \infty \qquad \text{and} \qquad \sup_{0\le s\le t\le T} \left\|\varphi(s,t)\right\|_q < \infty \quad (2.69)$$

for all $q \in [1,\infty]$.

A proof of this lemma is given in Section 2.9. The following lemma, whose proof is also given in Section 2.9, shows that φ^N converges to φ as $N \to \infty$ and provides uniform bounds on $\partial_\beta\varphi^N$.

Lemma 2.4.2. *For $T > 0$, $\phi \in \mathcal{S}(\mathbb{R}^d)$, there exists a constant K_1 such that, for all $N \ge 1$ and for $q \in \{1,2\}$,*

$$\sup_{0\le s\le t\le T} \left\|\varphi^N(s,t) - \varphi(s,t)\right\|_q \le K_1 r_N^2.$$

In addition, there exist constants K_2 and K_3 such that, for any multi-index β with $0 < |\beta| \le 4$, for

all $N \geq 1$,

$$\sup_{0\leq s\leq t\leq T} \left\|\varphi^N(s,t)\right\|_q \leq K_2 \left\|\phi\right\|_q, \qquad \text{and} \qquad \sup_{0\leq s\leq t\leq T} \left\|\partial_\beta \varphi^N(s,t)\right\|_q \leq K_3.$$

and K_2 *does not depend on* ϕ.

We shall see in Subsection 2.4.5 that M^N converges weakly in $\mathrm{D}\left([0,T],\mathcal{S}'(\mathbb{R}^d)\right)$, and hence in Subsection 2.4.6 (using results of [Wal86]) that the second term in (2.67) converges. However, in the general case where F is not linear, the first term in (2.67) has to be controlled.

Remark. *Recall the definition of* R_1 *in* (2.54)*; coming back to equation* (2.56)*, and using* (2.54) *instead of* (2.55)*, we write*

$$\left\langle Z_t^N, \phi\right\rangle = \int_0^t \left\{\left\langle \mathcal{L}^{(r_N)} Z_s^N, \phi\right\rangle - \left\langle \overline{\overline{Z_s^N} R_1(\overline{w_{s/\eta}^N}, \overline{f_s^N})}, \phi\right\rangle\right\} ds + \int_0^t \int_{\mathbb{R}^d} \phi(y) M^N(dyds).$$

Then by the same argument as for (2.65)*, for a time dependent test function* φ*,*

$$\left\langle Z_t^N, \varphi(t,t)\right\rangle = \int_0^t \left\langle Z_s^N, \partial_s\varphi(s,t) + \mathcal{L}^{(r_N)}\varphi(s,t) - \overline{R_1(\overline{w_{s/\eta}^N}, \overline{f_s^N})\overline{\varphi(s,t)}}(r_N)\right\rangle ds + \int_0^t \int_{\mathbb{R}^d} \varphi(x,s,t) M^N(dxds). \quad (2.70)$$

It is tempting to try to define φ^N *as the solution to*

$$\begin{cases} \partial_s\varphi^N(x,s,t) + \mathcal{L}^{(r_N)}\varphi^N(x,s,t) - \overline{R_1(\overline{w_{s/\eta}^N}, \overline{f_s^N})\overline{\varphi^N(s,t)}}(x, r_N) = 0, \\ \varphi^N(x,t,t) = \phi(x). \end{cases}$$

In this way, we would get rid of the first integral in (2.70)*. However, in this case,* $s \mapsto \varphi^N(\cdot,s,\cdot)$ *is not adapted to the canonical filtration of our process and the stochastic integral with respect to the martingale measure* M^N *is not well defined.*

2.4.2 Regularity estimate

The following result is an easy consequence of the definition of M^N.

Lemma 2.4.3. *There exists a constant* K_4 *such that for all* $\phi \in L^2((0,t)\times\mathbb{R}^d)$*,*

$$\mathbb{E}\left[\left(\int_0^t \int_{\mathbb{R}^d} \phi_s(x) M^N(dxds)\right)^2\right] \leq K_4 \int_0^t \|\phi_s\|_2^2 \, ds.$$

Proof. From the definition of Q^N in (2.57) and the definition of $\sigma^{(r)}_{z_1,z_2}$ in (2.34), letting $r = r_N$,

$$\begin{aligned}
&\mathbb{E}\left[\left(\int_0^t \int_{\mathbb{R}^d} \phi_s(x) M^N(dxds)\right)^2\right] \\
&\quad = \mathbb{E}\left[\int_0^t \int_{(\mathbb{R}^d)^2} \phi_s(z_1)\phi_s(z_2)\sigma_{z_1,z_2}(w^N_{s/\eta})dz_1dz_2ds\right] + \mathcal{O}\left(\delta_N^2 \int_0^t \|\phi_s\|_2^2\, ds\right) \\
&\quad \le \int_0^t \int_{(\mathbb{R}^d)^3} \frac{1}{V_r^2} \mathbb{1}_{\left\{\substack{|x-z_1|<r \\ |x-z_2|<r}\right\}} |\phi_s(z_1)|\,|\phi_s(z_2)|\, dxdz_1dz_2ds + \mathcal{O}\left(\delta_N^2 \int_0^t \|\phi_s\|_2^2\, ds\right) \\
&\quad = \int_0^t \int_{\mathbb{R}^d} \left(\frac{1}{V_r}\int_{B(x,r)} |\phi_s(z)|\, dz\right)^2 dxds + \mathcal{O}\left(\delta_N^2 \int_0^t \|\phi_s\|_2^2\, ds\right) \\
&\quad \le K_4 \int_0^t \|\phi_s\|_2^2\, ds.
\end{aligned}$$

(We have used Jensen's inequality in the last line.) □

For $t > 0$ and $x \in \mathbb{R}^d$, let

$$G_t(x) = (2\pi t)^{-d/2} \exp\left(-\frac{|x|^2}{2t}\right)$$

be the fundamental solution to the heat equation on $\mathbb{R}^d$; $\phi \mapsto G_t * \phi$ is then the semigroup of standard Brownian motion. Then f_t as defined in (2.47) satisfies

$$f_t(x) = G_t * w_0(x) - \int_0^t G_{t-s} * F(f_s)(x) ds.$$

Likewise, for $r > 0$, recall the definition of $\mathcal{L}^{(r)}$ in (2.14) and let $\left(\xi^{(r)}_t\right)_{t\ge 0}$ be a symmetric Lévy process on $\mathbb{R}^d$ with generator $\phi \mapsto \mathcal{L}^{(r)}\phi$. Let $\phi \mapsto G^{(r)}_t * \phi$ be the corresponding semigroup. Note that since $\xi^{(r)}_t = 0$ with positive probability, $G^{(r)}_t$ is not a well-defined function, but we do have $x \mapsto \overline{G^{(r)}_t}(x,r) \in L^{1,\infty}$. Then f^N as defined in (2.15) satisfies

$$f^N_t(x) = G^{(r_N)}_t * w_0(x) - \int_0^t G^{(r_N)}_{t-s} * \overline{F(\overline{f^N_s})}(x) ds. \tag{2.71}$$

The following provides a bound on the second moment of $\overline{Z^N_t}$, which allows us to control the quadratic term in (2.67). Note that $x \mapsto \overline{Z^N_t}(x, r_N)$ is a well defined function (despite the fact that $w^N_{t/\eta}$ is only defined up to a Lebesgue-null set).

Lemma 2.4.4. *For $T > 0$, there exists a constant $K_5 > 0$, independent of N, such that for $0 \le t \le T$,*

$$\sup_{x\in\mathbb{R}^d} \mathbb{E}\left[\overline{Z^N_t}(x, r_N)^2\right] \le \frac{K_5}{V_{r_N}}.$$

The proof of this result mirrors that of Theorem 1 in [Nor75a], although it is more technical because of the Laplacian and the various spatial averages.

Proof. We take $r = r_N$, $\eta = \eta_N$ and $\mathcal{L} = \mathcal{L}^{(r_N)}$ throughout the proof. Use equation (2.70) with the

time-dependent test function $\varphi(s,t) = G^{(r)}_{t-s} * \phi$, yielding

$$\left\langle Z^N_t, \phi \right\rangle = -\int_0^t \left\langle G^{(r)}_{t-s} * \left(\overline{\overline{Z^N_s} R_1(\overline{w^N_{s/\eta}}, \overline{f^N_s})} \right), \phi \right\rangle ds + \int_0^t \int_{\mathbb{R}^d} G^{(r)}_{t-s} * \phi(y) M^N(dyds).$$

Now take $\phi(y) = \frac{1}{V_r} \mathbb{1}_{\{|x-y|<r\}}$, and use Proposition 2.7.1 in Section 2.7 to obtain

$$\begin{aligned}
\overline{Z^N_t}(x) &= -\int_0^t G^{(r)}_{t-s} * \left(\overline{\overline{Z^N_s} R_1(\overline{w^N_{s/\eta}}, \overline{f^N_s})} \right)(x) ds + \int_0^t \int_{\mathbb{R}^d} \overline{G^{(r)}_{t-s}}(x-y) M^N(dyds) \\
&= -\int_0^t \int_{\mathbb{R}^d} \overline{\overline{G^{(r)}_{t-s}}}(x-y) \overline{Z^N_s}(y) R_1(\overline{w^N_{s/\eta}}(y), \overline{f^N_s}(y)) dyds + \int_0^t \int_{\mathbb{R}^d} \overline{G^{(r)}_{t-s}}(x-y) M^N(dyds).
\end{aligned}$$

We now want to apply Gronwall's lemma, but the last term must be controlled carefully. Taking the square of both sides and using $(a+b)^2 \le 2(a^2+b^2)$, we have

$$\begin{aligned}
\overline{Z^N_t}(x)^2 \le 2 \left(\int_0^t \int_{\mathbb{R}^d} \overline{\overline{G^{(r)}_{t-s}}}(x-y) \overline{Z^N_s}(y) R_1(\overline{w^N_{s/\eta}}(y), \overline{f^N_s}(y)) dyds \right)^2 \\
+ 2 \left(\int_0^t \int_{\mathbb{R}^d} \overline{G^{(r)}_{t-s}}(x-y) M^N(dyds) \right)^2.
\end{aligned}$$

By Jensen's inequality (and noting that $\int_{\mathbb{R}^d} \overline{G^{(r)}_t}(x)dx = 1$) and the bound $\|R_k\|_\infty \le \frac{1}{k!} \left\|F^{(k)}\right\|_\infty$ from (2.53), we have

$$\overline{Z^N_t}(x)^2 \le 2t \int_0^t \int_{\mathbb{R}^d} \overline{\overline{G^{(r)}_{t-s}}}(x-y) \left\|F'\right\|^2_\infty \overline{Z^N_s}(y)^2 dyds + 2 \left(\int_0^t \int_{\mathbb{R}^d} \overline{G^{(r)}_{t-s}}(x-y) M^N(dyds) \right)^2.$$

Taking expectations on both sides and using Fubini's theorem, we obtain

$$\begin{aligned}
\mathbb{E}\left[\overline{Z^N_t}(x)^2 \right] \le 2t \left\|F'\right\|^2_\infty \int_0^t \int_{\mathbb{R}^d} \overline{\overline{G^{(r)}_{t-s}}}(x-y) \mathbb{E}\left[\overline{Z^N_s}(y)^2 \right] dyds \\
+ 2\mathbb{E}\left[\left(\int_0^t \int_{\mathbb{R}^d} \overline{G^{(r)}_{t-s}}(x-y) M^N(dyds) \right)^2 \right].
\end{aligned}$$

From Lemma 2.4.3, we have

$$\begin{aligned}
\mathbb{E}\left[\left(\int_0^t \int_{\mathbb{R}^d} \overline{G^{(r)}_{t-s}}(x-y) M^N(dyds) \right)^2 \right] &\le K_4 \int_0^t \left\| \overline{G^{(r)}_{t-s}}(x-\cdot) \right\|^2_2 ds \\
&= K_4 \int_0^t \int_{\mathbb{R}^d} \mathbb{E}_0 \left[\frac{1}{V_r} \mathbb{1}_{\left\{ \left| \xi^{(r)}_{t-s} - (x-y) \right| < r \right\}} \right]^2 dyds \\
&\le K_4 \int_0^t \mathbb{E}_0 \left[\int_{\mathbb{R}^d} \left(\frac{1}{V_r} \mathbb{1}_{\left\{ \left| \xi^{(r)}_{t-s} - (x-y) \right| < r \right\}} \right)^2 dy \right] ds \\
&= \frac{K_4}{V_r} t. \qquad (2.72)
\end{aligned}$$

($\mathbb{E}_0[\cdot]$ denotes the expectation with respect to the law of $\left(\xi^{(r)}_t\right)_{t\ge 0}$ started from the origin.) In addition,

we note that $\mathbb{E}\left[\overline{Z_s^N}(y)^2\right] \leq \sup_{x\in\mathbb{R}^d} \mathbb{E}\left[\overline{Z_s^N}(x)^2\right]$ and, combined with the fact that $\int_{\mathbb{R}^d} \overline{G_t^{(r)}}(x)dx = 1$, this yields

$$\mathbb{E}\left[\overline{Z_t^N}(x)^2\right] \leq 2t\left\|F'\right\|_\infty^2 \int_0^t \sup_{y\in\mathbb{R}^d} \mathbb{E}\left[\overline{Z_s^N}(y)^2\right] ds + 2\frac{K_4}{V_r}t.$$

The right hand side does not depend on x, so we can take the supremum over $x \in \mathbb{R}^d$ on the left and write for $0 \leq t \leq T$

$$\sup_{x\in\mathbb{R}^d} \mathbb{E}\left[\overline{Z_t^N}(x)^2\right] \leq 2T\left\|F'\right\|_\infty^2 \int_0^t \sup_{x\in\mathbb{R}^d} \mathbb{E}\left[\overline{Z_s^N}(x)^2\right] ds + 2\frac{K_4}{V_r}T.$$

For each $N \geq 1$, for any $t \in [0,T]$ and $x \in \mathbb{R}^d$, $\overline{Z_t^N}(x) \leq (1 + \sup_{t\in[0,T]} \left\|f_t^N\right\|_\infty)(\eta_N/\tau_N)^{1/2} < \infty$ by Lemma 2.2.1. As a result, $t \mapsto \sup_{x\in\mathbb{R}^d} \mathbb{E}\left[\overline{Z_t^N}(x)^2\right]$ is bounded on $[0,T]$. Hence we can apply Gronwall's lemma (see e.g. Theorem 5.1 in [EK86]) to deduce that

$$\sup_{x\in\mathbb{R}^d} \mathbb{E}\left[\overline{Z_t^N}(x)^2\right] \leq 2\frac{K_4}{V_r}Te^{2Tt\left\|F'\right\|^2} \leq \frac{K_5}{V_r}.$$

□

2.4.3 Convergence to the deterministic limit

The following result, proved in Section 2.8, shows that f^N converges to f.

Proposition 2.4.5. *For $T > 0$, there exist constants K_6 and K_7 such that, for all $N \geq 1$,*

$$\sup_{0\leq t\leq T} \left\|f_t^N - f_t\right\|_\infty \leq K_6 r_N^2,$$

and for all multi-indices $\beta \in \mathbb{N}_0^d$ with $0 \leq |\beta| \leq 4$,

$$\sup_{0\leq t\leq T} \left\|\partial_\beta f_t^N\right\|_\infty \leq K_7.$$

We are now in a position to prove the first statement of Theorem 2.4, namely the convergence of the process $(w^N_{t/\eta_N})_{t\geq 0}$. We are going to prove the following lemma.

Lemma 2.4.6. *For $T > 0$, there exists a constant K_8 such that for all $N \geq 1$ and for any function ϕ satisfying $\left\|\phi\right\|_q \leq 1$ and $\max_{|\beta|=2} \left\|\partial_\beta\phi\right\|_q \leq 1$ for $q \in \{1,2\}$,*

$$\mathbb{E}\left[\sup_{0\leq t\leq T} \left|\left\langle Z_t^N, \phi\right\rangle\right|\right] \leq K_8. \tag{2.73}$$

Before we prove Lemma 2.4.6, we show that it implies the convergence of $\left(w^N_{t/\eta}\right)_{t\geq 0}$. We can choose a separating family $(\phi_n)_{n\geq 1}$ of compactly supported smooth functions satisfying $\left\|\phi_n\right\|_q \leq 1$

and $\max_{|\beta|=2} \|\partial_\beta \phi_n\|_q \leq 1$ for $q \in \{1,2\}$, and define d as in (2.4) using this family. Then

$$\begin{aligned}\mathbb{E}\left[\sup_{0\leq t\leq T} d(w^N_{t/\eta}, f_t)\right] &\leq \sum_{n\geq 1}\frac{1}{2^n}\left\{\mathbb{E}\left[\sup_{0\leq t\leq T}\left|\left\langle w^N_{t/\eta},\phi_n\right\rangle - \left\langle f^N_t,\phi_n\right\rangle\right|\right] + \sup_{0\leq t\leq T}\left|\left\langle f^N_t,\phi_n\right\rangle - \left\langle f_t,\phi_n\right\rangle\right|\right\}\\ &\leq \sum_{n\geq 1}\frac{1}{2^n}\left\{(\tau_N/\eta_N)^{1/2}\mathbb{E}\left[\sup_{0\leq t\leq T}\left|\left\langle Z^N_t,\phi_n\right\rangle\right|\right] + \sup_{0\leq t\leq T}\left\|f^N_t - f_t\right\|_\infty \|\phi_n\|_1\right\}\\ &\leq \sum_{n\geq 1}\frac{1}{2^n}\left\{K_8(\tau_N/\eta_N)^{1/2} + K_6 r_N^2\right\} = K_8(\tau_N/\eta_N)^{1/2} + K_6 r_N^2,\end{aligned}$$

where the last line follows by Proposition 2.4.5 and Lemma 2.4.6. The right-hand-side converges to zero as $N\to\infty$, yielding the uniform convergence (on compact time intervals) of $(w^N_{t/\eta_N})_{t\geq 0}$ to $(f_t)_{t\geq 0}$, the solution of equation (2.47). Note that, as soon as $d\geq 2$, r_N^2 is the leading order on the right-hand-side (see (2.48)).

Proof of Lemma 2.4.6. We are going to make use of (2.56) and apply Doob's maximal inequality to the martingale part. Let us first show that there exist two constants K and K' such that, for $t\in[0,T]$,

$$\mathbb{E}\left[\left|\left\langle Z^N_t,\phi\right\rangle\right|\right] \leq K\|\phi\|_2 + K'\frac{(\tau_N/\eta_N)^{1/2}}{V_{r_N}}\|\phi\|_1. \tag{2.74}$$

(From now on we shall write $\tau=\tau_N$, $\eta=\eta_N$ and $r=r_N$). Indeed, taking the expectation of the absolute value of both sides of (2.67) and using Lemma 2.4.3, we have

$$\begin{aligned}\mathbb{E}\left[\left|\left\langle Z^N_t,\phi\right\rangle\right|\right] &\leq (\tau/\eta)^{1/2}\frac{1}{2}\|F''\|_\infty\int_0^t\left\langle\mathbb{E}\left[(\overline{Z^N_s})^2\right],\left|\overline{\varphi^N(s,t)}\right|\right\rangle ds + \left(K_4\int_0^t\left\|\varphi^N(s,t)\right\|_2^2 ds\right)^{1/2}.\\ &\leq \frac{1}{2}\|F''\|_\infty K_5 T\frac{(\tau/\eta)^{1/2}}{V_r}K_2\|\phi\|_1 + K_4^{1/2}T^{1/2}K_2\|\phi\|_2,\end{aligned}$$

where we used Lemmas 2.4.4 and 2.4.2 in the last line. We have thus proved (2.74). Recalling (2.56) and the notation $M^N_t(\phi) = \int_0^t\int_{\mathbb{R}^d}\phi(x)M^N(dxds)$, we write

$$\begin{aligned}\sup_{t\in[0,T]}\left|\left\langle Z^N_t,\phi\right\rangle\right| \leq \int_0^T\left|\left\langle Z^N_s, \mathcal{L}\phi - \overline{F'(\overline{f^N_s})\overline{\phi}}\right\rangle\right| ds + \frac{1}{2}\|F''\|_\infty(\tau/\eta)^{1/2}\int_0^T\left\langle(\overline{Z^N_s})^2,|\overline{\phi}|\right\rangle ds\\ + \sup_{t\in[0,T]}\left|M^N_t(\phi)\right|.\end{aligned}$$

Taking expectations on both sides, we use Lemma 2.4.4 and apply (2.74) with ϕ replaced by $\mathcal{L}\phi - \overline{F'(\overline{f^N_s})\overline{\phi}}$ to write

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in[0,T]}\left|\left\langle Z^N_t,\phi\right\rangle\right|\right] \leq \int_0^T\left\{K(\|\mathcal{L}\phi\|_2 + \|F'\|_\infty\|\phi\|_2) + K'\frac{(\tau/\eta)^{1/2}}{V_r}(\|\mathcal{L}\phi\|_1 + \|F'\|_\infty\|\phi\|_1)\right\}ds\\ + \frac{1}{2}\|F''\|_\infty K_5 T\frac{(\tau/\eta)^{1/2}}{V_r}\|\phi\|_1 + \mathbb{E}\left[\sup_{t\in[0,T]}\left|M^N_t(\phi)\right|^2\right]^{1/2}.\end{aligned} \tag{2.75}$$

By Doob's inequality and Lemma 2.4.3,

$$\mathbb{E}\left[\sup_{t\in[0,T]}\left|M_t^N(\phi)\right|^2\right] \leq 4K_4T\left\|\phi\right\|_2^2.$$

Furthermore, $\left\|\mathcal{L}\phi\right\|_q \leq \frac{d(d+2)}{2}\max_{|\beta|=2}\left\|\partial_\beta\phi\right\|_q$ by Proposition 2.7.2.i in Section 2.7, and $\frac{(\tau/\eta)^{1/2}}{V_r}$ tends to zero as $N\to\infty$ due to assumption (2.48). Hence, if $\left\|\phi\right\|_q \leq 1$ and $\max_{|\beta|=2}\left\|\partial_\beta\phi\right\|_q \leq 1$ for $q\in\{1,2\}$, the right-hand-side of (2.75) is bounded by some constant independent of N and ϕ. □

2.4.4 Tightness

To prove that the sequence $\left(Z^N\right)_{N\geq1}$ is tight in $\mathrm{D}\left([0,T],\mathcal{S}'(\mathbb{R}^d)\right)$, we adapt the argument from the proof of Theorem 7.13 in [Wal86].

Proposition 2.4.7. *For any $\phi\in\mathcal{S}(\mathbb{R}^d)$, for any arbitrary sequence $(T_N,\rho_N)_{N\geq1}$ such that T_N is a stopping time (with respect to the natural filtration of the process $\left(Z_t^N\right)_{t\geq0}$) with values in $[0,T]$ for all N and ρ_N is a deterministic sequence of positive numbers decreasing to zero as $N\to\infty$,*

$$\left\langle Z^N_{T_N+\rho_N},\phi\right\rangle - \left\langle Z^N_{T_N},\phi\right\rangle \to 0 \tag{2.76}$$

in probability as $N\to\infty$.

By Aldous' criterion ([Ald78] and [Wal86, Theorem 6.8]), Proposition 2.4.7 together with Lemma 2.4.6 imply that the sequence of real-valued processes $\left(\left\langle Z^N,\phi\right\rangle\right)_{N\geq1}$ is tight in $\mathrm{D}\left([0,T],\mathbb{R}\right)$ for any $\phi\in\mathcal{S}(\mathbb{R}^d)$. In turn, Mitoma's theorem [Wal86, Theorem 6.13] implies the tightness of $\left(Z^N\right)_{N\geq1}$ in $\mathrm{D}\left([0,T],\mathcal{S}'(\mathbb{R}^d)\right)$.

The proof of Proposition 2.4.7 requires the three following lemmas (the first two are proved in Section 2.9). Extend φ^N to $\mathbb{R}^d\times[0,T]^2$ by setting, for $s,t\in[0,T]$,

$$\varphi^N(x,s,t) := \varphi^N(x,s\wedge t,t). \tag{2.77}$$

In other words, for $s>t$, $\varphi^N(s,t)$ equals ϕ.

Lemma 2.4.8. *For $T>0$, there exists a constant K_9 such that, for all $N\geq1$ and for $q\in\{1,2\}$,*

$$\forall s,t,t'\in[0,T],\quad \left\|\varphi^N(s,t')-\varphi^N(s,t)\right\|_q \leq K_9\left|t'-t\right|.$$

Lemma 2.4.9. *For $T>0$, there exists a constant K_{10} such that, for all $s\in[0,T]$,*

$$\left\|\sup_{t\in[s,T]}\left|\varphi^N(s,t)\right|\right\|_1 \leq K_{10}.$$

Define

$$V_t^N = \int_0^T\int_{\mathbb{R}^d}\varphi^N(x,s,t)M^N(dxds).$$

Lemma 2.4.10. *For $T>0$ and for any $0<\beta<1/2$, there exists a random variable Y_N such that*

$$\forall t,t'\in[0,T],\quad \left|V_{t'}^N-V_t^N\right| \leq Y_N\left|t'-t\right|^\beta, \tag{2.78}$$

almost surely, and $\mathbb{E}\left[Y_N^2\right] \leq C'$ *for all* $N \geq 1$.

Proof. By Lemma 2.4.3 and then Lemma 2.4.8,

$$\begin{aligned}\mathbb{E}\left[\left|V_{t'}^N - V_t^N\right|^2\right] &= \mathbb{E}\left[\left(\int_0^T \int_{\mathbb{R}^d} (\varphi^N(x,s,t') - \varphi^N(x,s,t)) M^N(dxds)\right)^2\right]. \\ &\leq K_4 \int_0^T \left\|\varphi^N(s,t') - \varphi^N(s,t)\right\|_2^2 ds \\ &\leq (K_9)^2 T K_4 \left|t' - t\right|^2.\end{aligned}$$

The result follows by Kolmogorov's continuity theorem [Wal86, Corollary 1.2]. □

Proof of Proposition 2.4.7. We are going to treat each term in (2.67) separately. The first one converges to zero in L^1, uniformly on $[0,T]$, as a consequence of Lemma 2.4.4. The second one is dealt with as in [Wal86, Theorem 7.13]. From (2.67), write

$$\begin{aligned}\left\langle Z_{T_N+\rho_N}^N, \phi\right\rangle - \left\langle Z_{T_N}^N, \phi\right\rangle &= (\tau_N/\eta_N)^{1/2} \int_0^{T_N} \left\langle (\overline{Z_s^N})^2, R_2(\overline{w_{s/\eta}^N}, \overline{f_s^N}) \overline{\varphi^N(s,T_N)} \right\rangle ds \\ &\quad - (\tau_N/\eta_N)^{1/2} \int_0^{T_N+\rho_N} \left\langle (\overline{Z_s^N})^2, R_2(\overline{w_{s/\eta}^N}, \overline{f_s^N}) \overline{\varphi^N(s,T_N+\rho_N)} \right\rangle ds \\ &\quad + \left(V_{T_N+\rho_N}^N - V_{T_N}^N\right) + \int_{T_N}^{T_N+\rho_N} \int_{\mathbb{R}^d} \phi(x) M^N(dxds). \quad (2.79)\end{aligned}$$

Let us deal with each term separately. The first two are similar so we need only consider the first one. Since inside the integral $s \leq T_N \leq T$, $\left|\varphi^N(s,T_N)\right| \leq \sup_{t\in[s,T]} \left|\varphi^N(s,t)\right|$ and we have

$$\left|\int_0^{T_N} \left\langle (\overline{Z_s^N})^2, R_2(\overline{w_{s/\eta}^N}, \overline{f_s^N}) \overline{\varphi^N(s,T_N)} \right\rangle ds\right| \leq \frac{1}{2} \left\|F''\right\|_\infty \int_0^T \left\langle (\overline{Z_s^N})^2, \overline{\sup_{t\in[s,T]} |\varphi^N(s,t)|} \right\rangle ds.$$

Taking the expectation on both sides, we get

$$\begin{aligned}\mathbb{E}\left[\left|\int_0^{T_N} \left\langle (\overline{Z_s^N})^2, R_2(\overline{w_{s/\eta}^N}, \overline{f_s^N}) \overline{\varphi^N(s,T_N)} \right\rangle ds\right|\right] &\leq \frac{1}{2} \left\|F''\right\|_\infty \int_0^T \left\langle \mathbb{E}\left[(\overline{Z_s^N})^2\right], \overline{\sup_{t\in[s,T]} |\varphi^N(s,t)|} \right\rangle ds \\ &\leq \frac{1}{2} \left\|F''\right\|_\infty \frac{K_5}{V_{r_N}} \int_0^T \left\| \sup_{t\in[s,T]} \left|\varphi^N(s,t)\right| \right\|_1 ds \\ &\leq \frac{1}{2} \left\|F''\right\|_\infty \frac{K_5}{V_{r_N}} T K_{10}. \quad (2.80)\end{aligned}$$

where the second line follows by Lemma 2.4.4 and the third line follows by Lemma 2.4.9. Recall that we assumed in (2.48) that $\tau_N/\eta_N = o\left(r_N^{2d}\right)$; hence the first term on the right-hand-side of (2.79) converges to zero in L^1. By Lemma 2.4.10, we have, almost surely,

$$\left|V_{T_N+\rho_N}^N - V_{T_N}^N\right| \leq Y_N \rho_N^{1/4}.$$

Taking the expectation of the square of both sides, we write

$$\mathbb{E}\left[\left|V^N_{T_N+\rho_N} - V^N_{T_N}\right|^2\right] \leq C'\rho_N^{1/2}.$$

Hence the third term converges to zero in L^2 and in probability as $N \to \infty$. Finally, since T_N is a stopping time, we can apply Lemma 2.4.3 to the fourth term,

$$\begin{aligned}\mathbb{E}\left[\left(\int_{T_N}^{T_N+\rho_N}\int_{\mathbb{R}^d}\phi(x)M^N(dxds)\right)^2\right] &\leq K_4\mathbb{E}\left[\int_{T_N}^{T_N+\rho_N}\|\phi\|_2^2\,ds\right]\\ &\leq K_4\|\phi\|_2^2\,\rho_N.\end{aligned}$$

This concludes the proof of Proposition 2.4.7. □

2.4.5 Convergence of the martingale measure M^N

The next step is to show that the martingale measure M^N converges weakly in $\mathrm{D}\left([0,T],\mathcal{S}'(\mathbb{R}^d)\right)$ as $N \to \infty$ to M, where $M_t = \sqrt{f_t(1-f_t)}\cdot W_t$ is a stochastic integral (as defined in [Wal86, Chapter 2]) against the space-time white noise W and f is the solution of (2.47). We will naturally use Theorem 2.6, along with the following result on convergence to Gaussian martingales (which is a consequence of Lévy's characterisation of Brownian motion).

For any $\mathbb{R}^d$-valued càdlàg process $(Y_t)_{t\geq 0}$, define $\Delta Y_t = Y_t - Y_{t^-}$.

Theorem 2.7 ([JS03, Theorem VIII 3.11])**.** *Fix $T > 0$ and suppose $(X_t)_{t\geq 0} = (X_t^1,\ldots,X_t^d)_{t\geq 0}$ is a continuous d-dimensional Gaussian martingale and for each $n \geq 1$, $(X_t^n)_{t\geq 0} = (X_t^{n,1},\ldots,X_t^{n,d})_{t\geq 0}$ is a càdlàg, locally square-integrable martingale such that*

(i) $|\Delta X_t^n|$ is bounded uniformly in n for all t, and $\sup_{t\leq T}|\Delta X_t^n| \xrightarrow[n\to\infty]{P} 0$.

(ii) For each $t \in \mathbb{Q}\cap[0,T]$, $\left\langle X^{n,i}, X^{n,j}\right\rangle_t \xrightarrow[n\to\infty]{P} \left\langle X^i, X^j\right\rangle_t$.

Then X^n converges in distribution to X in $D([0,T],\mathbb{R}^d)$.

In our setting, the limiting process $(M_t(\phi))_{t\geq 0}$ is a continuous martingale with quadratic variation

$$\langle M(\phi)\rangle_t = \int_0^t\int_{\mathbb{R}^d}\phi(x)^2 f_s(x)(1-f_s(x))dxds.$$

(See [Wal86, Theorem 2.5].) Since this quantity is deterministic, $(M_t(\phi))_{t\geq 0}$ is Gaussian, and we can apply the result above. The following lemma is then enough to conclude that M^N converges to M.

Lemma 2.4.11. *For any $\phi \in \mathcal{S}(\mathbb{R}^d)$,*

i) For all $t \geq 0$, $\left|\Delta M_t^N(\phi)\right| \leq K$ for some constant K, and $\sup_{0\leq t\leq T}\left|\Delta M_t^N(\phi)\right| \xrightarrow[N\to\infty]{P} 0$.

ii) For each $t \in [0,T]$, $\left\langle M^N(\phi)\right\rangle_t \xrightarrow[N\to\infty]{P} \langle M(\phi)\rangle_t$.

Indeed, by polarisation, we can recover $\left\langle M^N(\phi_i), M^N(\phi_j)\right\rangle_t$ from $\left\langle M^N(\phi_i+\phi_j)\right\rangle_t$ and $\langle M^N(\phi_i - \phi_j)\rangle_t$, and (ii) of Theorem 2.7 is satisfied by vectors of the form $(M_t^N(\phi_1),\ldots,M_t^N(\phi_p))_{t\geq 0}$.

As a result, for any $\phi_1, \dots, \phi_p$ in $\mathcal{S}(\mathbb{R}^d)$, $\left(M_t^N(\phi_1), \dots, M_t^N(\phi_p)\right)_{t\geq 0}$ converges in distribution to $(M_t(\phi_1), \dots, M_t(\phi_p))_{t\geq 0}$ in $\mathrm{D}\left([0,T], \mathbb{R}^d\right)$. In particular, for any $\phi \in \mathcal{S}(\mathbb{R}^d)$, $\left(M^N(\phi)\right)_{N\geq 1}$ is tight, and M^N satisfies the assumptions of Theorem 2.6, hence M^N converges in distribution to M as $N \to \infty$ in $\mathrm{D}\left([0,T], \mathcal{S}'(\mathbb{R}^d)\right)$.

Proof of Lemma 2.4.11. By the definition of $M^N(\phi)$,

$$\begin{aligned} M_t^N(\phi) - M_{t^-}^N(\phi) &= \left\langle Z_t^N, \phi \right\rangle - \left\langle Z_{t^-}^N, \phi \right\rangle \\ &= (\eta_N/\tau_N)^{1/2} \left(\left\langle w_{t/\eta}^N, \phi \right\rangle - \left\langle w_{t^-/\eta}^N, \phi \right\rangle \right). \end{aligned}$$

The bound on the jumps of $\langle w_t, \phi \rangle$ in (2.49) implies

$$\begin{aligned} \sup_{t\geq 0} \left|\Delta M_t^N(\phi)\right| &\leq (\eta_N/\tau_N)^{1/2} \sup_{t\geq 0} \left|\left\langle w_t^N - w_{t^-}^N, \phi \right\rangle\right| \\ &\leq \alpha_N (\eta_N/\tau_N)^{1/2} \left\|\phi\right\|_1. \end{aligned}$$

But we have assumed that $\alpha_N^2 = o\left(\tau_N/\eta_N\right)$, so (i) is satisfied. To prove (ii), recall from (2.57) that

$$\left\langle M^N(\phi) \right\rangle_t = \int_0^t \int_{(\mathbb{R}^d)^2} \phi(z_1)\phi(z_2)\sigma_{z_1,z_2}^{(r_N)}(w_{s/\eta}^N) dz_1 dz_2 ds + \mathcal{O}\left(\delta_N^2 t \left\|\phi\right\|_2^2\right).$$

The rationale here is to show that the main contribution to this term comes from the diagonal $\{(z_1, z_2) : z_1 = z_2\}$ when $r_N \to 0$. From the definition of $\sigma^{(r_N)}$ in (2.34), letting $r = r_N$,

$$\begin{aligned} \int_{(\mathbb{R}^d)^2} \phi(z_1)\phi(z_2)\sigma_{z_1,z_2}^{(r)}(w_{s/\eta}^N) dz_1 dz_2 = \frac{1}{V_r^2} \int_{(\mathbb{R}^d)^3} \phi(z_1)\phi(z_2) [\overline{w_{s/\eta}^N}(x)(1 - w_{s/\eta}^N(z_1))(1 - w_{s/\eta}^N(z_2)) \\ + (1 - \overline{w_{s/\eta}^N}(x)) w_{s/\eta}^N(z_1) w_{s/\eta}^N(z_2)] \mathbb{1}_{\left\{\substack{|z_1 - x| < r \\ |z_2 - x| < r}\right\}} dx dz_1 dz_2. \end{aligned}$$

Changing the order of integration gives

$$\int_{(\mathbb{R}^d)^2} \phi(z_1)\phi(z_2)\sigma_{z_1,z_2}^{(r)}(w_{s/\eta}^N) dz_1 dz_2 = \left\langle \overline{w_{s/\eta}^N}, \left(\overline{\phi(1 - w_{s/\eta}^N)}\right)^2 \right\rangle + \left\langle 1 - \overline{w_{s/\eta}^N}, \left(\overline{\phi\, w_{s/\eta}^N}\right)^2 \right\rangle. \quad (2.81)$$

We are left with showing that the right-hand-side of (2.81) converges in probability to

$$\left\langle f_s, (\phi(1 - f_s))^2 \right\rangle + \left\langle 1 - f_s, (\phi f_s)^2 \right\rangle = \left\langle f_s(1 - f_s), \phi^2 \right\rangle.$$

To do this, we first justify that ϕ can be let out of the average, we use Lemma 2.4.4 to argue that we can replace $w_{s/\eta}^N$ by f_s^N, then the regularity of f^N allows us to remove the averages and finally we know from Proposition 2.4.5 that f^N converges to f. First note that

$$\overline{\phi w}(x) - \phi(x)\overline{w}(x) = \frac{1}{V_r} \int_{B(x,r)} w(y)(\phi(y) - \phi(x)) dy.$$

Since $0 \leq w(y) \leq 1$ a.e., we have

$$\left|\overline{\phi w}(x) - \phi(x)\overline{w}(x)\right| \leq \frac{1}{V_r}\int_{B(x,r)} |\phi(y) - \phi(x)|\, dy \leq \frac{1}{V_r}\int_{B(x,r)} \sum_{i=1}^{d} \|\partial_i \phi\|_\infty |y-x|_i \, dy.$$

Hence

$$\left\|\overline{\phi w} - \phi\overline{w}\right\|_\infty \leq d\, r_N \max_i \|\partial_i \phi\|_\infty .$$

As a consequence,

$$\begin{aligned}
\left\langle 1-\overline{w}, (\overline{\phi w})^2\right\rangle - \left\langle 1-\overline{w}, \phi^2\overline{w}^2\right\rangle &= \left\langle 1-\overline{w}, (\overline{\phi w} - \phi\overline{w})(\overline{\phi w} + \phi\overline{w})\right\rangle \\
\left|\left\langle 1-\overline{w}, (\overline{\phi w})^2\right\rangle - \left\langle 1-\overline{w}, \phi^2\overline{w}^2\right\rangle\right| &\leq 2\left\|\phi\right\|_1 \left\|\overline{\phi w} - \phi\overline{w}\right\|_\infty \\
&\leq 2d\left\|\phi\right\|_1 r_N \max_i \left\|\partial_i\phi\right\|_\infty .
\end{aligned}$$

By the same argument (replacing w by $1-w$), we can also let ϕ out of the average in the first term on the right-hand-side of (2.81), and the problem reduces to showing the convergence of

$$\left\langle \overline{w^N_{s/\eta}}, \phi^2(1-\overline{w^N_{s/\eta}})^2\right\rangle + \left\langle 1-\overline{w^N_{s/\eta}}, \phi^2\overline{w^N_{s/\eta}}^2\right\rangle = \left\langle \overline{w^N_{s/\eta}}(1-\overline{w^N_{s/\eta}}), \phi^2\right\rangle .$$

We now see that it is enough to show that uniformly for $s \in [0,T]$,

$$\sup_{x\in\mathbb{R}^d} \mathbb{E}\left[\left|\overline{w^N_{s/\eta}}(x) - f_s(x)\right|\right] \underset{N\to\infty}{\longrightarrow} 0.$$

But, by the triangle inequality,

$$\begin{aligned}
\mathbb{E}\left[\left|\overline{w^N_{s/\eta}}(x) - f_s(x)\right|\right] &\leq (\tau_N/\eta_N)^{1/2}\mathbb{E}\left[\overline{Z^N_s}(x)^2\right]^{1/2} + \left\|\overline{f^N_s} - f^N_s\right\|_\infty + \left\|f^N_s - f_s\right\|_\infty \\
&\leq \frac{(\tau_N/\eta_N)^{1/2}}{V^{1/2}_{r_N}} K_5^{1/2} + \frac{d}{2} r_N^2 K_7 + K_6 r_N^2.
\end{aligned}$$

(We have used Lemma 2.4.4, Proposition 2.7.2 in Section 2.7 and Proposition 2.4.5.) Note that this bound is uniform for $s \in [0,T]$. The right-hand-side converges to zero as $N \to \infty$ (due to assumption (2.48)), providing the desired result. From all this we conclude

$$\int_{(\mathbb{R}^d)^2} \phi(z_1)\phi(z_2)\sigma^{(r_N)}_{z_1,z_2}(w^N_{s/\eta})dz_1 dz_2 \underset{N\to\infty}{\overset{L^1}{\longrightarrow}} \int_{\mathbb{R}^d} \phi(x)^2 f_s(x)(1-f_s(x))dx,$$

uniformly for $s \in [0,T]$, which gives us (ii). □

2.4.6 Conclusion of the proof

We are almost done. We have proved that the sequence of processes $\left(Z^N\right)_{N\geq 1}$ is tight, and we need only characterise its potential limit points. Recall the following expression for $\left\langle Z^N_t, \phi\right\rangle$ from (2.67):

$$\left\langle Z^N_t, \phi\right\rangle = -(\tau_N/\eta_N)^{1/2}\int_0^t \left\langle (\overline{Z^N_s})^2, R_2(\overline{w^N_{s/\eta}}, \overline{f^N_s})\overline{\varphi^N(s,t)}\right\rangle ds + \int_0^t \int_{\mathbb{R}^d} \varphi^N(x,s,t) M^N(dxds).$$

The first term converges to zero in L^1 from (2.80). Also,

$$\int_0^t \int_{\mathbb{R}^d} \varphi^N(x,s,t) M^N(dxds) - \int_0^t \int_{\mathbb{R}^d} \varphi(x,s,t) M^N(dxds) \xrightarrow[N\to\infty]{L^2} 0,$$

since, from Lemma 2.4.3 and Lemma 2.4.2,

$$\begin{aligned}\mathbb{E}\left[\left(\int_0^t \int_{\mathbb{R}^d} (\varphi^N(x,s,t) - \varphi(x,s,t)) M^N(dxds)\right)^2\right] &\leq K_4 \int_0^t \left\|\varphi^N(s,t) - \varphi(s,t)\right\|_2^2 ds \\ &\leq K_1^2 T K_4 r_N^4.\end{aligned}$$

For $\phi_1, \ldots, \phi_p$ in $\mathcal{S}(\mathbb{R}^d)$, let $\varphi_1, \ldots, \varphi_p$ be the corresponding solutions of (2.68) with $\phi = \phi_i$. Since we showed in Section 2.4.5 that M^N converges weakly to M, by [Wal86, Proposition 7.12], for $t_1, \ldots, t_p \in [0,T]$,

$$\begin{aligned}&\left(\int_0^{t_1} \int_{\mathbb{R}^d} \varphi_1(x,s,t_1) M^N(dxds), \ldots, \int_0^{t_k} \int_{\mathbb{R}^d} \varphi_k(x,s,t_k) M^N(dxds)\right) \\ &\qquad \xrightarrow[N\to\infty]{d} \left(\int_0^{t_1} \int_{\mathbb{R}^d} \varphi_1(x,s,t_1) M(dxds), \ldots, \int_0^{t_k} \int_{\mathbb{R}^d} \varphi_k(x,s,t_k) M(dxds)\right).\end{aligned}$$

This uniquely characterises the potential limit points of $\left(Z^N\right)_{N\geq 1}$. By Theorem 2.6, $\left(Z_t^N\right)_{t\geq 0}$ converges in distribution to a distribution-valued process $(z_t)_{t\geq 0}$ given by

$$\langle z_t, \phi\rangle = \int_0^t \int_{\mathbb{R}^d} \varphi(x,s,t) M(dxds), \tag{2.82}$$

where φ satisfies the backwards heat equation (2.68) with terminal condition ϕ,

$$\begin{cases} \partial_s \varphi(x,s,t) + \dfrac{1}{2}\Delta\varphi(x,s,t) - F'(f_s(x))\varphi(x,s,t) = 0, \\ \varphi(x,t,t) = \phi(x). \end{cases}$$

It is an easy exercise to prove that z_t satisfies

$$\langle z_t, \phi\rangle = \int_0^t \left\langle z_s, \frac{1}{2}\Delta\phi - F'(f_s)\phi \right\rangle ds + \int_0^t \int_{\mathbb{R}^d} \phi(x) M(dxds). \tag{2.83}$$

(See the proof of [Wal86, Theorem 5.2].) In other words, $(z_t)_{t\geq 0}$ is the (mild) solution of (2.51) (recall that $M_t = \sqrt{f_t(1-f_t)}\cdot W_t$) and Theorem 2.4 is proved.

2.5 The stable case - proof of Theorem 2.5

Turning to the proof of the central limit theorem in the stable case, we warn that its overall structure is the same as that in the Brownian case. Whenever the details of the argument are exactly the same as previously, we simply mention intermediate results without detailing their proof. Some steps need a different treatment however, and we explain those in more detail. To simplify our formulas, we use

the following notation:

$$a_n \lesssim b_n \Leftrightarrow \exists K > 0 : \forall n \geq 1,\ 0 \leq a_n \leq K\, b_n. \tag{2.84}$$

The specific constants can always be retrieved from Section 2.4 or from a trivial calculation. Also as in Section 2.4 we set the constants u and $(sV_1)/\alpha$ to 1 in the martingale problem (M2) defined in Definition 2.3.5. Let us write (M2) as

$$dw_t^N = \eta_N \left[\mathcal{D}^{\alpha,\delta_N} w_t^N - F^{(\delta_N)}(w_t^N)\right] dt + \tau_N^{1/2} d\mathbb{M}_t^N.$$

Recall that $Z_t^N = (\eta_N/\tau_N)^{1/2}(w_{t/\eta}^N - f_t^N)$. Setting $M_t^N(\phi) = \eta_N^{1/2}\mathbb{M}_{t/\eta_N}^N(\phi)$ and using the definition of $F^{(\delta_N)}$ in (2.18), we have, by the same argument as for (2.52),

$$dZ_t^N = \left[\mathcal{D}^{\alpha,\delta_N} Z_t^N - \alpha(\eta_N/\tau_N)^{1/2} \int_1^\infty \overline{\left(F(\overline{w_{t/\eta}^N}) - F(\overline{f_t^N})\right)}(\delta_N r)\frac{dr}{r^{1+\alpha}}\right] dt + dM_t^N. \tag{2.85}$$

Using the definition of R_2 in (2.55) as for (2.56), one obtains (in mild form)

$$\begin{aligned}\left\langle Z_t^N, \phi\right\rangle = &\int_0^t \left\langle Z_s^N, \mathcal{D}^{\alpha,\delta_N}\phi - \alpha\int_1^\infty \overline{F'(\overline{f_s^N})\overline{\phi}}(\delta_N r)\frac{dr}{r^{\alpha+1}}\right\rangle ds \\ &- \left(\frac{\tau_N}{\eta_N}\right)^{\frac{1}{2}} \alpha \int_0^t\int_1^\infty \left\langle (\overline{Z_s^N}(\delta_N r))^2, R_2(\overline{w_{s/\eta}^N}, \overline{f_s^N})\overline{\phi}(\delta_N r)\right\rangle \frac{dr}{r^{\alpha+1}} ds + M_t^N(\phi),\end{aligned} \tag{2.86}$$

and the covariation measure of M^N is given by

$$Q^N(dz_1 dz_2 ds) = (\sigma_{z_1,z_2}^{(\alpha,\delta_N)}(w_{s/\eta}^N) + |z_1 - z_2|^{-\alpha}\,\mathcal{O}\,(\delta_N^\alpha))dz_1 dz_2 ds. \tag{2.87}$$

2.5.1 Time dependent test functions

Note that the analogue of (2.65) holds in the stable setting, that is, for any time dependent test function φ, by (2.86) and the same argument as for (2.65),

$$\begin{aligned}\left\langle Z_t^N, \varphi(t,t)\right\rangle = &\int_0^t \left\langle Z_s^N, \partial_s\varphi(s,t) + \mathcal{D}^{\alpha,\delta_N}\varphi(s,t) - \alpha\int_1^\infty \overline{F'(\overline{f_s^N})\overline{\varphi(s,t)}}(\delta_N r)\frac{dr}{r^{\alpha+1}}\right\rangle ds \\ &- \left(\frac{\tau_N}{\eta_N}\right)^{\frac{1}{2}} \alpha \int_0^t\int_1^\infty \left\langle (\overline{Z_s^N}(\delta_N r))^2, R_2(\overline{w_{s/\eta}^N}, \overline{f_s^N})\overline{\varphi(s,t)}(\delta_N r)\right\rangle \frac{dr}{r^{\alpha+1}} ds \\ &+ \int_0^t\int_{\mathbb{R}^d} \varphi(x,s,t) M^N(dxds).\end{aligned} \tag{2.88}$$

For $\phi \in \mathcal{S}(\mathbb{R}^d)$, suppose φ^N solves

$$\begin{cases} \partial_s\varphi^N(x,s,t) + \mathcal{D}^{\alpha,\delta_N}\varphi^N(x,s,t) - \alpha\displaystyle\int_1^\infty \overline{F'(\overline{f_s^N})\overline{\varphi^N(s,t)}}(x,\delta_N r)\frac{dr}{r^{\alpha+1}} = 0 \\ \varphi^N(x,t,t) = \phi(x). \end{cases} \tag{2.89}$$

By (2.88), we have

$$\left\langle Z_t^N, \phi \right\rangle = -\left(\frac{\tau_N}{\eta_N}\right)^{1/2} \alpha \int_0^t \int_1^\infty \left\langle (\overline{Z_s^N}(\delta_N r))^2, R_2(\overline{w_{s/\eta}^N}, \overline{f_s^N})\overline{\varphi^N(s,t)}(\delta_N r) \right\rangle \frac{dr}{r^{\alpha+1}} ds + \int_0^t \int_{\mathbb{R}^d} \varphi^N(x,s,t) M^N(dxds). \quad (2.90)$$

We are thus left with finding a suitable way to bound the first term above and showing the convergence of the stochastic integral against M^N. The convergence of the martingale measure M^N is going to involve slightly different calculations compared to the previous case as the limiting noise is not a space-time white noise. The convergence of φ^N, however, is proved in a similar way to before. For $\phi \in \mathcal{S}(\mathbb{R}^d)$, define φ as the solution to the following

$$\begin{cases} \partial_s \varphi(x,s,t) + \mathcal{D}^\alpha \varphi(x,s,t) - F'(f_s(x))\varphi(x,s,t) = 0 \\ \varphi(x,t,t) = \phi(x). \end{cases} \quad (2.91)$$

Lemma 2.5.1. *Fix $T > 0$ and take $\phi \in \mathcal{S}(\mathbb{R}^d)$. There exists a unique solution φ^N to (2.89) in $L^\infty(\mathbb{R}^d \times \{(s,t) : 0 \le s \le t \le T\})$ which admits spatial derivatives of order up to two. Moreover, for any multi-index β with $0 \le |\beta| \le 2$,*

$$\sup_{0\le s\le t\le T} \left\|\partial_\beta \varphi^N(s,t)\right\|_q < \infty \qquad \text{and} \qquad \sup_{0\le s\le t\le T} \left\|\varphi(s,t)\right\|_q < \infty \quad (2.92)$$

for $q \in \{1, \infty\}$.

The proof of Lemma 2.5.1 is a straightforward adaptation of that of Lemma 2.4.1. The following lemma, whose proof is given in Section 2.9, provides the convergence of φ^N to φ.

Lemma 2.5.2. *For $T > 0$, $\phi \in \mathcal{S}(\mathbb{R}^d)$ and for $q \in \{1, \infty\}$,*

$$\sup_{0\le s\le t\le T} \left\|\varphi^N(s,t) - \varphi(s,t)\right\|_q \lesssim \delta_N^{\alpha\wedge(2-\alpha)}.$$

In addition, for $0 < |\beta| \le 2$,

$$\sup_{0\le s\le t\le T} \left\|\varphi^N(s,t)\right\|_q \lesssim \|\phi\|_q \qquad \text{and} \qquad \sup_{0\le s\le t\le T} \left\|\partial_\beta \varphi^N(s,t)\right\|_q \lesssim 1.$$

2.5.2 Regularity estimate

Let us first state the following L^2 bound for the stochastic integral.

Lemma 2.5.3. *For $0 \le t \le T$ and $\alpha < d$, suppose $\phi_t : \mathbb{R}^d \to \mathbb{R}$ is in $L^{1,\infty}(\mathbb{R}^d)$; then*

$$\mathbb{E}\left[\left(\int_0^t \int_{\mathbb{R}^d} \phi_s(x) M^N(dxds)\right)^2\right] \lesssim \int_0^t \int_{(\mathbb{R}^d)^2} |\phi_s(z_1)|\,|\phi_s(z_2)|\,(\delta_N \vee \tfrac{|z_1-z_2|}{2})^{-\alpha} dz_1 dz_2 ds \quad (2.93)$$

$$\lesssim \int_0^t \|\phi_s\|_1(\|\phi_s\|_\infty + \|\phi_s\|_1) ds. \quad (2.94)$$

The proof uses the following lemma, which is proved in Section 2.9.

Lemma 2.5.4. *For $\alpha < d$, then for $f, g \in L^{1,\infty}(\mathbb{R}^d)$*

$$\left| \int_{(\mathbb{R}^d)^2} f(z_1) g(z_2) \left| z_1 - z_2 \right|^{-\alpha} dz_1 dz_2 \right| \leq \|f\|_1 \left(\frac{dV_1}{d-\alpha} \|g\|_\infty + \|g\|_1 \right).$$

Proof of Lemma 2.5.3. From the expression for the covariation measure in (2.87),

$$\mathbb{E}\left[\left(\int_0^t \int_{\mathbb{R}^d} \phi_s(x) M^N(dxds) \right)^2 \right] = \mathbb{E}\left[\int_0^t \int_{(\mathbb{R}^d)^2} \phi_s(z_1)\phi_s(z_2) \sigma^{\alpha,\delta_N}_{z_1,z_2}(w^N_{s/\eta}) dz_1 dz_2 ds \right] + \mathcal{O}\left(\delta_N^\alpha\right) \int_0^t \int_{(\mathbb{R}^d)^2} \phi_s(z_1)\phi_s(z_2) |z_1 - z_2|^{-\alpha} dz_1 dz_2 ds.$$

But, by the definition of $\sigma^{(\alpha,\delta)}$ in (2.61),

$$\begin{aligned} \left| \sigma^{\alpha,\delta}_{z_1,z_2}(w) \right| &\leq \int_{\delta \vee \frac{|z_1 - z_2|}{2}}^\infty V_r(z_1, z_2) \frac{dr}{r^{d+\alpha+1}} \\ &\lesssim V_1 \int_{\delta \vee \frac{|z_1 - z_2|}{2}}^\infty \frac{dr}{r^{\alpha+1}} \\ &\lesssim \left(\delta \vee \frac{|z_1 - z_2|}{2} \right)^{-\alpha}, \end{aligned}$$

yielding (2.93). Since $(\delta_N \vee \frac{|z_1-z_2|}{2})^{-\alpha} \leq (\frac{|z_1-z_2|}{2})^{-\alpha}$, inequality (2.94) is obtained from (2.93) and Lemma 2.5.4. □

Let $\mathcal{G}^{(\alpha)}$ (resp. $\mathcal{G}^{(\alpha,\delta)}$) denote the fundamental solution to the fractional heat equation with the operator $\mathcal{D}^\alpha$ (resp. the fractional heat equation with the truncated operator $\mathcal{D}^{\alpha,\delta}$). Then the centering term f^N as defined in (2.22) can be written as

$$f^N_t(x) = \mathcal{G}^{(\alpha,\delta_N)}_t * w_0(x) - \int_0^t \mathcal{G}^{(\alpha,\delta_N)}_{t-s} * F^{(\delta_N)}(f^N_s)(x) ds. \tag{2.95}$$

Likewise, using the definition of f_t in (2.62),

$$f_t(x) = \mathcal{G}^{(\alpha)}_t * w_0(x) - \int_0^t \mathcal{G}^{(\alpha)}_{t-s} * F(f_s)(x) ds.$$

We now prove the following counterpart of the regularity estimate (Lemma 2.4.4), which allows us to bound the quadratic error term in (2.90).

Lemma 2.5.5. *Fix $T > 0$; for $0 \leq t \leq T$,*

$$\int_1^\infty \sup_{x \in \mathbb{R}^d} \mathbb{E}\left[\overline{Z^N_t}(x, \delta_N r)^2 \right] \frac{dr}{r^{\alpha+1}} \lesssim \frac{1}{\delta_N^\alpha}.$$

Proof. From (2.85) and the definition of R_1 in (2.54), we have

$$\left\langle Z^N_t, \phi \right\rangle = \int_0^t \left\langle Z^N_s, \mathcal{D}^{\alpha,\delta_N} \phi - \alpha \int_1^\infty \overline{R_1(\overline{w^N_{s/\eta}}, \overline{f^N_s})\overline{\phi}}(\delta_N r) \frac{dr}{r^{\alpha+1}} \right\rangle ds + \int_0^t \int_{\mathbb{R}^d} \phi(y) M^N(dyds).$$

By the same argument as for (2.65), for a time dependent test function φ,

$$\left\langle Z_t^N, \varphi(t,t)\right\rangle = \int_0^t \left\langle Z_s^N, \partial_s\varphi(s,t) + \mathcal{D}^{\alpha,\delta_N}\varphi(s,t) - \alpha\int_1^\infty \overline{R_1(\overline{w_{s/\eta}^N}, \overline{f_s^N})\overline{\varphi(s,t)}}(\delta_N r)\frac{dr}{r^{\alpha+1}}\right\rangle ds + \int_0^t\int_{\mathbb{R}^d} \varphi(x,s,t)M^N(dxds).$$

Applying this with $\varphi(x,s,t) = \mathcal{G}_{t-s}^{(\alpha,\delta_N)} * \phi(x)$ yields

$$\left\langle Z_t^N, \phi\right\rangle = -\alpha\int_0^t\int_1^\infty \left\langle \mathcal{G}_{t-s}^{(\alpha,\delta_N)} * \overline{\overline{Z_s^N} R_1(\overline{w_{s/\eta}^N}, \overline{f_s^N})}(\delta_N r), \phi\right\rangle \frac{dr}{r^{\alpha+1}}ds + \int_0^t\int_{\mathbb{R}^d} \mathcal{G}_{t-s}^{(\alpha,\delta_N)} * \phi(y) M^N(dyds).$$

Now we take $\phi(y) = \frac{1}{V_R}\mathbb{1}_{\{|x-y|<R\}}$ to obtain

$$\overline{Z_t^N}(x,R) = -\alpha\int_0^t\int_1^\infty \left(\overline{\mathcal{G}_{t-s}^{(\alpha,\delta_N)}}(R) * \overline{\overline{Z_s^N} R_1(\overline{w_{s/\eta}^N}, \overline{f_s^N})}(\delta_N r)\right)(x)\frac{dr}{r^{\alpha+1}}ds + \int_0^t\int_{\mathbb{R}^d} \overline{\mathcal{G}_{t-s}^{(\alpha,\delta_N)}}(x-y,R)M^N(dyds).$$

Repeating the same steps as in the proof of Lemma 2.4.4 and using Jensen's inequality, we get

$$\left(\overline{Z_t^N}(x,R)\right)^2 \lesssim \alpha\int_0^t\int_1^\infty\int_{\mathbb{R}^d} \overline{\overline{\mathcal{G}_{t-s}^{(\alpha,\delta_N)}}}(x-y,R,\delta_N r)\left(\overline{Z_s^N}(y,\delta_N r)\right)^2 dy\frac{dr}{r^{\alpha+1}}ds + \left(\int_0^t\int_{\mathbb{R}^d} \overline{\mathcal{G}_{t-s}^{(\alpha,\delta_N)}}(x-y,R)M^N(dyds)\right)^2.$$

Using the first inequality of Lemma 2.5.3 and bounding $\left(\delta_N \vee \frac{|z_1-z_2|}{2}\right)^{-\alpha}$ by $\delta_N^{-\alpha}$, we have, for $0 \le t \le T$,

$$\begin{aligned}\mathbb{E}\left[\left(\int_0^t\int_{\mathbb{R}^d} \overline{\mathcal{G}_{t-s}^{(\alpha,\delta_N)}}(x-y,R)M^N(dyds)\right)^2\right] &\lesssim \delta_N^{-\alpha}\int_0^t\left(\int_{\mathbb{R}^d} \overline{\mathcal{G}_{t-s}^{(\alpha,\delta_N)}}(x-y,R)dy\right)^2 ds\\ &\lesssim \delta_N^{-\alpha}.\end{aligned}$$

As a result

$$\mathbb{E}\left[\left(\overline{Z_t^N}(x,R)\right)^2\right] \lesssim \int_0^t\int_1^\infty\int_{\mathbb{R}^d} \overline{\overline{\mathcal{G}_{t-s}^{(\alpha,\delta_N)}}}(x-y,R,\delta_N r)\mathbb{E}\left[\left(\overline{Z_s^N}(y,\delta_N r)\right)^2\right] dy\frac{dr}{r^{\alpha+1}}ds + \delta_N^{-\alpha}.$$

Taking the supremum of $\mathbb{E}\left[\overline{Z_s^N}(y,\delta_N r)^2\right]$ over y inside the integral on the right-hand-side, the function $\overline{\mathcal{G}^{(\alpha,\delta)}}$ integrates to 1, yielding

$$\sup_{x\in\mathbb{R}^d}\mathbb{E}\left[\left(\overline{Z_t^N}(x,R)\right)^2\right] \lesssim \int_0^t\int_1^\infty \sup_{x\in\mathbb{R}^d}\mathbb{E}\left[\left(\overline{Z_s^N}(x,\delta_N r)\right)^2\right]\frac{dr}{r^{\alpha+1}}ds + \delta_N^{-\alpha}.$$

Integrating over R, we get

$$\int_1^\infty \sup_{x\in\mathbb{R}^d} \mathbb{E}\left[\left(\overline{Z_t^N}(x,\delta_N r)\right)^2\right]\frac{dr}{r^{\alpha+1}} \lesssim \int_0^t\int_1^\infty \sup_{x\in\mathbb{R}^d} \mathbb{E}\left[\left(\overline{Z_s^N}(x,\delta_N r)\right)^2\right]\frac{dr}{r^{\alpha+1}}ds + \delta_N^{-\alpha}.$$

Moreover, for each $N \geq 1$ and for all $t \in [0,T]$,

$$\int_1^\infty \sup_{x\in\mathbb{R}^d} \mathbb{E}\left[\left(\overline{Z_t^N}(x,\delta_N r)\right)^2\right]\frac{dr}{r^{\alpha+1}} \leq \frac{\eta_N}{\alpha\tau_N}(1+\sup_{t\in[0,T]}\left\|f_t^N\right\|_\infty)^2$$

which is bounded on $[0,T]$ by Lemma 2.2.2. Hence, by Gronwall's inequality, for $0 \leq t \leq T$,

$$\int_1^\infty \sup_{x\in\mathbb{R}^d} \mathbb{E}\left[\left(\overline{Z_t^N}(x,\delta_N r)\right)^2\right]\frac{dr}{r^{\alpha+1}} \lesssim \delta_N^{-\alpha}.$$

□

2.5.3 Convergence to the deterministic limit

The following result is proved in Section 2.8.

Proposition 2.5.6. *For $T > 0$,*

$$\sup_{0\leq t\leq T}\left\|f_t^N - f_t\right\|_\infty \lesssim \delta_N^{\alpha\wedge(2-\alpha)},$$

and for $0 \leq |\beta| \leq 2$,

$$\sup_{0\leq t\leq T}\left\|\partial_\beta f_t^N\right\|_\infty \lesssim 1.$$

By the same argument as in Section 2.4.3, choosing a separating family $(\phi_n)_{n\geq 1}$ of compactly supported smooth functions satisfying $\|\phi_n\|_q \leq 1$ and $\max_{|\beta|=2}\|\partial_\beta\phi_n\|_q \leq 1$ for $q \in \{1,\infty\}$ and using the corresponding metric d on Ξ,

$$\mathbb{E}\left[\sup_{0\leq t\leq T} d(w_{t/\eta}^N, f_t^N)\right] \leq \sum_{n\geq 1}\frac{1}{2^n}\left\{(\tau_N/\eta_N)^{1/2}\mathbb{E}\left[\sup_{0\leq t\leq T}\left|\langle Z_t^N,\phi_n\rangle\right|\right] + \sup_{0\leq t\leq T}\left\|f_t^N - f_t\right\|_\infty\|\phi_n\|_1\right\}.$$

The convergence of $w_{t/\eta}^N$ to f_t in L^1 follows from Proposition 2.5.6 and Lemma 2.5.7 below.

Lemma 2.5.7. *For any function ϕ satisfying $\|\phi\|_q \leq 1$ and $\max_{|\beta|=2}\|\partial_\beta\phi\|_q \leq 1$ for $q \in \{1,\infty\}$,*

$$\mathbb{E}\left[\sup_{0\leq t\leq T}\left|\langle Z_t^N,\phi\rangle\right|\right] \lesssim 1.$$

As a result,

$$\mathbb{E}\left[\sup_{0\leq t\leq T} d(w_{t/\eta}^N, f_t^N)\right] \lesssim (\tau_N/\eta_N)^{1/2} + \delta_N^{\alpha\wedge(2-\alpha)}.$$

From (2.63), it can be seen that the leading term on the right-hand-side is $\delta_N^{\alpha\wedge(2-\alpha)}$, which goes to zero as $N\to\infty$, yielding the convergence of $\left(w_{t/\eta}^N\right)_{t\geq 0}$. The following lemma is needed for the proof of Lemma 2.5.7 and is proved in the same manner as (2.74) in Section 2.4.3.

Lemma 2.5.8. *For $\phi \in L^{1,\infty}(\mathbb{R}^d)$ and $t \in [0,T]$,*

$$\mathbb{E}\left[\left|\left\langle Z_t^N, \phi\right\rangle\right|\right] \lesssim \|\phi\|_1 + \|\phi\|_\infty .$$

Proof. Taking expectations on both sides of (2.90),

$$\mathbb{E}\left[\left|\left\langle Z_t^N, \phi\right\rangle\right|\right] \lesssim (\tau_N/\eta_N)^{1/2} \int_0^t \int_1^\infty \left\langle \mathbb{E}\left[\left(\overline{Z_s^N}(\delta_N r)\right)^2\right], \overline{|\varphi^N(s,t)|}(\delta_N r)\right\rangle \frac{dr}{r^{\alpha+1}} ds + \mathbb{E}\left[\left(\int_0^t \int_{\mathbb{R}^d} \varphi^N(x,s,t) M^N(dxds)\right)^2\right]^{1/2}. \quad (2.96)$$

In the first integral, we have

$$\left\langle \mathbb{E}\left[\left(\overline{Z_s^N}(\delta_N r)\right)^2\right], \overline{|\varphi^N(s,t)|}(\delta_N r)\right\rangle \leq \left\|\varphi^N(s,t)\right\|_1 \sup_{x\in\mathbb{R}^d} \mathbb{E}\left[\left(\overline{Z_s^N}(x,\delta_N r)\right)^2\right].$$

Hence, applying Lemma 2.5.5 to the first term and Lemma 2.5.3 to the second term on the right-hand-side of (2.96) yields

$$\begin{aligned}
\mathbb{E}\left[\left|\left\langle Z_t^N, \phi\right\rangle\right|\right] &\lesssim \frac{(\tau_N/\eta_N)^{1/2}}{\delta_N^\alpha} \int_0^t \left\|\varphi^N(s,t)\right\|_1 ds + \left(\int_0^t \left\|\varphi^N(s,t)\right\|_1 \left(\left\|\varphi^N(s,t)\right\|_\infty + \left\|\varphi^N(s,t)\right\|_1\right) ds\right)^{1/2} \\
&\lesssim \frac{(\tau_N/\eta_N)^{1/2}}{\delta_N^\alpha} \|\phi\|_1 + \left(\|\phi\|_1 \left(\|\phi\|_\infty + \|\phi\|_1\right)\right)^{1/2} \\
&\lesssim \|\phi\|_1 + \|\phi\|_\infty .
\end{aligned}$$

We have used the fact that (by Lemma 2.5.2) $\left\|\varphi^N(s,t)\right\|_q \lesssim \|\phi\|_q$ to pass from the first line to the second. The third line follows since $\tau_N/\eta_N = o\left(\delta_N^{2\alpha}\right)$ by (2.63). □

Proof of Lemma 2.5.7. The proof of Lemma 2.5.7 is similar to the proof of Lemma 2.4.6. Setting $\psi_s = \mathcal{D}^{\alpha,\delta_N}\phi - \alpha \int_1^\infty \overline{F'(\overline{f_s^N})\overline{\phi}}(\delta_N r) \frac{dr}{r^{\alpha+1}}$ and using (2.86),

$$\begin{aligned}
\sup_{0\leq t\leq T} \left|\left\langle Z_t^N, \phi\right\rangle\right| &\lesssim \int_0^T \left|\left\langle Z_s^N, \psi_s\right\rangle\right| ds \\
&\quad + (\tau_N/\eta_N)^{1/2} \|F''\|_\infty \int_0^T \int_1^\infty \left\langle \left(\overline{Z_s^N}(\delta_N r)\right)^2, \overline{|\phi|}(\delta_N r)\right\rangle \frac{dr}{r^{\alpha+1}} ds + \sup_{0\leq t\leq T} \left|M_t^N(\phi)\right|.
\end{aligned}$$

Taking the expectation on both sides, Lemma 2.5.8 can be used in the first term, and Lemma 2.5.5 in the second one, to yield

$$\mathbb{E}\left[\sup_{0\leq t\leq T} \left|\left\langle Z_t^N, \phi\right\rangle\right|\right] \lesssim \int_0^T (\|\psi_s\|_1 + \|\psi_s\|_\infty) ds + \frac{(\tau_N/\eta_N)^{1/2}}{\delta_N^\alpha} \|F''\|_\infty \|\phi\|_1 + \mathbb{E}\left[\sup_{0\leq t\leq T} \left|M_t^N(\phi)\right|^2\right]^{1/2}.$$

But $\|\psi_s\|_q \lesssim \left\|\mathcal{D}^{\alpha,\delta}\phi\right\|_q + \|F'\|_\infty \|\phi\|_q$ and, by Proposition 2.7.3.i in Section 2.7, $\left\|\mathcal{D}^{\alpha,\delta}\phi\right\|_q \lesssim \|\phi\|_q +$

$\max_{|\beta|=2} \|\partial_\beta \phi\|_q$. In addition, by Doob's inequality, and using Lemma 2.5.3,

$$\mathbb{E}\left[\sup_{0\le t\le T} \left|M_t^N(\phi)\right|^2\right] \lesssim \mathbb{E}\left[M_T^N(\phi)^2\right] \lesssim \|\phi\|_1 \left(\|\phi\|_1 + \|\phi\|_\infty\right).$$

As a result, if $\|\phi\|_q \le 1$ and $\max_{|\beta|=2} \|\partial_\beta \phi\|_q \le 1$ for $q \in \{1, \infty\}$,

$$\mathbb{E}\left[\sup_{0\le t\le T} \left|\left\langle Z_t^N, \phi\right\rangle\right|\right] \lesssim 1.$$

□

2.5.4 Tightness

The overall argument for the tightness of the sequence $\left(Z_t^N\right)_{t\ge 0}$ is the same as in Section 2.4.4.

Proposition 2.5.9. *For any $\phi \in \mathcal{S}(\mathbb{R}^d)$ and for any sequence $(T_N, \rho_N)_{N\ge 1}$ such that T_N is a stopping time with values in $[0,T]$ for every $N \ge 1$ and $\rho_N \downarrow 0$ as $N \to \infty$,*

$$\left\langle Z^N_{T_N+\rho_N}, \phi\right\rangle - \left\langle Z^N_{T_N}, \phi\right\rangle \xrightarrow[N\to\infty]{P} 0. \tag{2.97}$$

Tightness of $\left(Z^N\right)_{N\ge 1}$ in $\mathrm{D}\left([0,T], \mathcal{S}'(\mathbb{R}^d)\right)$ then follows from Aldous' criterion [Ald78] and Mitoma's theorem [Wal86, Theorem 6.13].

Extend φ^N to $\mathbb{R}^d \times [0,T]^2$ as in (2.77); we need estimates on φ^N as in Lemmas 2.4.8 and 2.4.9. The proof of the following lemma is in Section 2.9.

Lemma 2.5.10. *For $T > 0$, $q \in \{1, \infty\}$ and for all $s, t, t' \in [0,T]$,*

$$\left\|\varphi^N(s,t') - \varphi^N(s,t)\right\|_q \lesssim |t - t'|.$$

In addition, for all $s \in [0,T]$,

$$\left\|\sup_{t\in[s,T]} \left|\varphi^N(s,t)\right|\right\|_1 \lesssim 1.$$

Proof of Proposition 2.5.9. We only detail how the quadratic part of (2.90) can be bounded using Lemma 2.5.5, and refer to Section 2.4.4 for the rest of the proof of Proposition 2.5.9. For T_N a stopping time with values in $[0,T]$, write

$$\left|\int_0^{T_N}\int_1^\infty \left\langle (\overline{Z_s^N}(\delta_N r))^2, R_2(\overline{w^N_{s/\eta}}, \overline{f_s^N})\overline{\varphi^N(s,T_N)}(\delta_N r)\right\rangle \frac{dr}{r^{\alpha+1}} ds\right| \lesssim \|F''\|_\infty \int_0^T \int_1^\infty \left\langle (\overline{Z_s^N}(\delta_N r))^2, \overline{\sup_{t\in[s,T]} |\varphi^N(s,t)|}(\delta_N r)\right\rangle \frac{dr}{r^{\alpha+1}} ds.$$

Taking the expectation on both sides and the supremum inside the spatial integral against φ^N, we get

$$\begin{aligned}\mathbb{E}\left[\left|\int_0^{T_N}\int_1^\infty \left\langle (\overline{Z_s^N}(\delta_N r))^2, R_2(\overline{w_{s/\eta}^N}, \overline{f_s^N})\overline{\varphi^N(s,T_N)}(\delta_N r)\right\rangle \frac{dr}{r^{\alpha+1}}ds\right|\right] \\ \lesssim \|F''\|_\infty \int_0^T\int_1^\infty \sup_{x\in\mathbb{R}^d}\mathbb{E}\left[\overline{Z_s^N}(x,\delta_N r)^2\right]\left\|\sup_{t\in[s,T]}\left|\varphi^N(s,t)\right|\right\|_1 \frac{dr}{r^{\alpha+1}}ds \\ \lesssim \delta_N^{-\alpha},\end{aligned}$$

by Lemma 2.5.5 and Lemma 2.5.10. The other terms in (2.97) are bounded as in the proof of Proposition 2.4.7 in Section 2.4.4, using Lemmas 2.5.10 and 2.5.3. □

2.5.5 Convergence of the martingale measure M^N

The convergence of M^N relies on applying Theorem 2.7 to vectors of the form $\big(M_t^N(\phi_1),\dots, M_t^N(\phi_p)\big)_{t\geq 0}$, although the details differ from the proof in the Brownian case (in Section 2.4.5). Indeed, M^N no longer converges to a stochastic integral against a space-time white noise, but to W^α, a coloured Gaussian noise such that

$$\langle W^\alpha(\phi)\rangle_t = \int_0^t\int_{(\mathbb{R}^d)^2}\phi(z_1)\phi(z_2)\sigma^\alpha_{z_1,z_2}(f_s)dz_1dz_2ds$$

with

$$\sigma^\alpha_{z_1,z_2}(f) = \int_{\frac{|z_1-z_2|}{2}}^\infty \frac{dr}{r^{d+\alpha+1}}\int_{B(z_1,r)\cap B(z_2,r)}\left[\overline{f}(x,r)(1-f(z_1)-f(z_2))+f(z_1)f(z_2)\right]dx.$$

Hence the weak convergence of M^N to W^α in $\mathrm{D}\left([0,T],\mathcal{S}'(\mathbb{R}^d)\right)$ will follow (as in Section 2.4.5) from the following lemma.

Lemma 2.5.11. *For any $\phi\in\mathcal{S}(\mathbb{R}^d)$,*

i) For all $t\geq 0$, $\left|\Delta M_t^N(\phi)\right|\lesssim 1$, and $\sup_{0\leq t\leq T}\left|\Delta M_t^N(\phi)\right|\xrightarrow[N\to\infty]{P}0$.

ii) For each $t\in[0,T]$, $\left\langle M^N(\phi)\right\rangle_t\xrightarrow[N\to\infty]{P}\langle W^\alpha(\phi)\rangle_t$.

Proof. The proof of the first part is the same as for Lemma 2.4.11:

$$\sup_{t\geq 0}\left|\Delta M_t^N(\phi)\right|\leq \alpha_N(\eta_N/\tau_N)^{1/2}\left\|\phi\right\|_1,$$

which tends to zero since $\alpha_N^2 = o\left(\tau_N/\eta_N\right)$. For the second part of the statement, we first show that

$$\left|\int_{(\mathbb{R}^d)^2}\phi(z_1)\phi(z_2)\sigma^{(\alpha,\delta_N)}_{z_1,z_2}(w_{s/\eta}^N)dz_1dz_2 - \int_{(\mathbb{R}^d)^2}\phi(z_1)\phi(z_2)\sigma^{(\alpha,\delta_N)}_{z_1,z_2}(f_s^N)dz_1dz_2\right|\xrightarrow[N\to\infty]{L^1}0. \tag{2.98}$$

From the definition of $\sigma_{z_1,z_2}^{(\alpha,\delta_N)}$ in (2.61),

$$\sigma_{z_1,z_2}^{(\alpha,\delta_N)}(w) = \int_{\delta_N \vee \frac{|z_1-z_2|}{2}}^{\infty} \Big\{ (1 - w(z_1) - w(z_2)) \int_{B(z_1,r)\cap B(z_2,r)} \overline{w}(x,r)dx \\ + V_r(z_1,z_2)w(z_1)w(z_2) \Big\} \frac{dr}{r^{d+\alpha+1}}.$$

Subtracting the corresponding expressions with $w_{s/\eta}^N$ and f_s^N and reordering terms, we write

$$\begin{aligned}
&\sigma_{z_1,z_2}^{(\alpha,\delta_N)}(w_{s/\eta}^N) - \sigma_{z_1,z_2}^{(\alpha,\delta_N)}(f_s^N) \\
&\quad = \int_{\delta_N \vee \frac{|z_1-z_2|}{2}}^{\infty} \Big\{ (1 - w_{s/\eta}^N(z_1) - w_{s/\eta}^N(z_2)) \int_{B(z_1,r)\cap B(z_2,r)} \Big(\overline{w_{s/\eta}^N}(x,r) - \overline{f_s^N}(x,r) \Big) dx \\
&\qquad + (f_s^N(z_1) - w_{s/\eta}^N(z_1) + f_s^N(z_2) - w_{s/\eta}^N(z_2)) \int_{B(z_1,r)\cap B(z_2,r)} \overline{f_s^N}(x,r)dx \\
&\qquad + V_r(z_1,z_2) \Big(w_{s/\eta}^N(z_1)(w_{s/\eta}^N(z_2) - f_s^N(z_2)) + f_s^N(z_2)(w_{s/\eta}^N(z_1) - f_s^N(z_1)) \Big) \Big\} \frac{dr}{r^{d+\alpha+1}}. \quad (2.99)
\end{aligned}$$

We shall deal with the terms from each of the three lines separately, so let us call them $A(z_1,z_2)$, $B(z_1,z_2)$ and $C(z_1,z_2)$ (they are in fact defined for a.e. z_1 and z_2, and so is all that follows, but this is not a problem since what we really show is (2.98)). For the first term write

$$\begin{aligned}
\mathbb{E}\left[|A(z_1,z_2)|\right] &\le (\tau_N/\eta_N)^{1/2} \int_{\delta_N \vee \frac{|z_1-z_2|}{2}}^{\infty} \int_{B(z_1,r)\cap B(z_2,r)} \mathbb{E}\left[\left| \overline{Z_s^N}(x,r) \right| \right] dx \frac{dr}{r^{d+\alpha+1}} \\
&\le (\tau_N/\eta_N)^{1/2} \int_{\delta_N \vee \frac{|z_1-z_2|}{2}}^{\infty} V_r(z_1,z_2) \sup_{x\in\mathbb{R}^d} \mathbb{E}\left[\left| \overline{Z_s^N}(x,r) \right| \right] \frac{dr}{r^{d+\alpha+1}} \\
&\le (\tau_N/\eta_N)^{1/2} V_1 \int_{\delta_N}^{\infty} \mathbb{1}_{\{2r>|z_1-z_2|\}} \sup_{x\in\mathbb{R}^d} \mathbb{E}\left[\left| \overline{Z_s^N}(x,r) \right|^2 \right]^{1/2} \frac{dr}{r^{\alpha+1}} \\
&\le (\tau_N/\eta_N)^{1/2} \frac{V_1}{\alpha^{1/2}} \left(\delta_N \vee \frac{|z_1-z_2|}{2} \right)^{-\alpha/2} \left(\int_{\delta_N}^{\infty} \sup_{x\in\mathbb{R}^d} \mathbb{E}\left[\left| \overline{Z_s^N}(x,r) \right|^2 \right] \frac{dr}{r^{\alpha+1}} \right)^{1/2}.
\end{aligned}$$

(We have used the Cauchy-Schwartz inequality in the last line.) In addition, by Lemma 2.5.5,

$$\begin{aligned}
\int_{\delta_N}^{\infty} \sup_{x\in\mathbb{R}^d} \mathbb{E}\left[\left| \overline{Z_s^N}(x,r) \right|^2 \right] \frac{dr}{r^{\alpha+1}} &= \delta_N^{-\alpha} \int_1^{\infty} \sup_{x\in\mathbb{R}^d} \mathbb{E}\left[\left| \overline{Z_s^N}(x,\delta_N r) \right|^2 \right] \frac{dr}{r^{\alpha+1}} \\
&\lesssim \delta_N^{-2\alpha}.
\end{aligned}$$

Hence

$$\mathbb{E}\left[|A(z_1,z_2)|\right] \lesssim (\tau_N/\eta_N)^{1/2} \delta_N^{-\alpha} \left|z_1 - z_2\right|^{-\alpha/2},$$

and, using Lemma 2.5.4 and (2.63),

$$\int_{(\mathbb{R}^d)^2} \phi(z_1)\phi(z_2)A(z_1,z_2)dz_1dz_2 \xrightarrow[N\to\infty]{L^1} 0. \quad (2.100)$$

For the second term, by symmetry,

$$\left|\int_{(\mathbb{R}^d)^2}\phi(z_1)\phi(z_2)B(z_1,z_2)dz_1dz_2\right| \le 2(\tau_N/\eta_N)^{1/2}\int_{\mathbb{R}^d}|\phi(z_2)|\left|\left\langle Z_s^N,\psi_{z_2}^N\right\rangle\right|dz_2,$$

where

$$\psi_{z_2}^N(z_1)=\phi(z_1)\int_{\delta_N\vee\frac{|z_1-z_2|}{2}}^{\infty}\int_{B(z_1,r)\cap B(z_2,r)}\overline{f_s^N}(x,r)dx\frac{dr}{r^{d+\alpha+1}}.$$

In particular, by Proposition 2.5.6 $\left\|\psi_{z_2}^N\right\|_q \lesssim \delta_N^{-\alpha}\left\|\phi\right\|_q$ for $q\in\{1,\infty\}$ and, since $\psi_{z_2}^N$ is deterministic, by Lemma 2.5.8

$$\mathbb{E}\left[\left|\int_{(\mathbb{R}^d)^2}\phi(z_1)\phi(z_2)B(z_1,z_2)dz_1dz_2\right|\right] \lesssim (\tau_N/\eta_N)^{1/2}\delta_N^{-\alpha}\left\|\phi\right\|_1\left(\left\|\phi\right\|_1+\left\|\phi\right\|_\infty\right).$$

Hence, by (2.63),

$$\int_{(\mathbb{R}^d)^2}\phi(z_1)\phi(z_2)B(z_1,z_2)dz_1dz_2 \xrightarrow[N\to\infty]{L^1} 0. \tag{2.101}$$

The third term is controlled in a similar way, this time setting

$$\psi_{z_2}^N(z_1)=\phi(z_1)\int_{\delta_N\vee\frac{|z_1-z_2|}{2}}^{\infty}V_r(z_1,z_2)\frac{dr}{r^{d+\alpha+1}},$$

which satisfies the same inequalities as the previous $\psi_{z_2}^N$. The bound on $\left\|f_s^N\right\|_\infty$ from Proposition 2.5.6 yields

$$\int_{(\mathbb{R}^d)^2}\phi(z_1)\phi(z_2)C(z_1,z_2)dz_1dz_2 \xrightarrow[N\to\infty]{L^1} 0. \tag{2.102}$$

The convergence (2.98) follows from (2.100), (2.101), (2.102) and (2.99).

Recall the definition of $\sigma^{(\alpha,\delta)}$ in (2.61) and write

$$\begin{aligned}\left|\sigma_{z_1,z_2}^{(\alpha,\delta_N)}(f_s^N)-\sigma_{z_1,z_2}^{\alpha}(f_s^N)\right| &\lesssim \mathbb{1}_{\{|z_1-z_2|\le 2\delta_N\}}\int_{\frac{|z_1-z_2|}{2}}^{\delta_N}V_r(z_1,z_2)\frac{dr}{r^{d+\alpha+1}}\\ &\lesssim \mathbb{1}_{\{|z_1-z_2|\le 2\delta_N\}}\left|z_1-z_2\right|^{-\alpha}.\end{aligned}$$

Hence

$$\begin{aligned}&\left|\int_{(\mathbb{R}^d)^2}\phi(z_1)\phi(z_2)(\sigma_{z_1,z_2}^{(\alpha,\delta_N)}(f_s^N)-\sigma_{z_1,z_2}^{\alpha}(f_s^N))dz_1dz_2\right|\\ &\qquad\lesssim \int_{(\mathbb{R}^d)^2}|\phi(z_1)|\,|\phi(z_2)|\,\mathbb{1}_{\{|z_1-z_2|\le 2\delta_N\}}\left|z_1-z_2\right|^{-\alpha}dz_1dz_2\\ &\qquad\lesssim \left\|\phi\right\|_\infty\int_{\mathbb{R}^d}|\phi(z_1)|\int_0^{2\delta_N}r^{-\alpha+d-1}drdz_1\\ &\qquad\lesssim \left\|\phi\right\|_\infty\left\|\phi\right\|_1\delta_N^{d-\alpha}\xrightarrow[N\to\infty]{}0.\end{aligned} \tag{2.103}$$

Finally, proceeding as in (2.99) with f_s instead of $w^N_{s/\eta}$ and σ^α instead of $\sigma^{(\alpha,\delta_N)}$, we write

$$\begin{aligned}
&\sigma^\alpha_{z_1,z_2}(f_s) - \sigma^\alpha_{z_1,z_2}(f^N_s) \\
&\qquad = \int_{\frac{|z_1-z_2|}{2}}^\infty \Big\{ (1 - f_s(z_1) - f_s(z_2)) \int_{B(z_1,r)\cap B(z_2,r)} \Big(\overline{f_s}(x,r) - \overline{f^N_s}(x,r) \Big) dx \\
&\qquad\qquad + (f^N_s(z_1) - f_s(z_1) + f^N_s(z_2) - f_s(z_2)) \int_{B(z_1,r)\cap B(z_2,r)} \overline{f^N_s}(x,r) dx \\
&\qquad\qquad\qquad + V_r(z_1,z_2) \left(f_s(z_1)(f_s(z_2) - f^N_s(z_2)) + f^N_s(z_2)(f_s(z_1) - f^N_s(z_1)) \right) \Big\} \frac{dr}{r^{d+\alpha+1}}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\left|\sigma^\alpha_{z_1,z_2}(f^N_s) - \sigma^\alpha_{z_1,z_2}(f_s)\right| &\le (4 + 3 \sup_{s\in[0,T]} \left\|f^N_s\right\|_\infty) \left\|f^N_s - f_s\right\|_\infty \int_{\frac{|z_1-z_2|}{2}}^\infty V_r(z_1,z_2) \frac{dr}{r^{1+d+\alpha}} \\
&\lesssim |z_1 - z_2|^{-\alpha} \, \delta_N^{\alpha\wedge(2-\alpha)}
\end{aligned}$$

by Proposition 2.5.6. It follows from Lemma 2.5.4 that

$$\left| \int_{(\mathbb{R}^d)^2} \phi(z_1)\phi(z_2)(\sigma^\alpha_{z_1,z_2}(f^N_s) - \sigma^\alpha_{z_1,z_2}(f_s)) dz_1 dz_2 \right| \underset{N\to\infty}{\longrightarrow} 0. \tag{2.104}$$

By (2.98), (2.103) and (2.104), we have shown that for all $t \in [0,T]$,

$$\left\langle M^N(\phi) \right\rangle_t \overset{L^1,P}{\underset{N\to\infty}{\longrightarrow}} \left\langle W^\alpha(\phi) \right\rangle_t .$$

□

2.5.6 Conclusion of the proof

We can now conclude the proof of Theorem 2.5. We have proved that the sequence $\left(Z^N\right)_{N\ge1}$ is tight and we can characterise its potential limit points using the convergence of M^N. Recall the following expression for $\left\langle Z^N_t, \phi \right\rangle$ from (2.90) :

$$\begin{aligned}
\left\langle Z^N_t, \phi \right\rangle = -\left(\frac{\tau_N}{\eta_N}\right)^{1/2} \alpha \int_0^t \int_1^\infty \left\langle (\overline{Z^N_s}(\delta_N r))^2, R_2(\overline{w^N_{s/\eta}}, \overline{f^N_s}) \overline{\varphi^N(s,t)}(\delta_N r) \right\rangle \frac{dr}{r^{\alpha+1}} ds \\
+ \int_0^t \int_{\mathbb{R}^d} \varphi^N(x,s,t) M^N(dxds).
\end{aligned}$$

In Section 2.5.4, we showed that the first term converges to zero in L^1. In addition, by Lemmas 2.5.2 and 2.5.3,

$$\int_0^t \int_{\mathbb{R}^d} \varphi^N(x,s,t) M^N(dxds) - \int_0^t \int_{\mathbb{R}^d} \varphi(x,s,t) M^N(dxds) \overset{L^2}{\underset{N\to\infty}{\longrightarrow}} 0.$$

For $\phi_1, \dots, \phi_p$ in $\mathcal{S}(\mathbb{R}^d)$, let $\varphi_1, \dots, \varphi_p$ be the corresponding solutions of (2.91) with $\phi = \phi_i$. Since M^N converges weakly to W^α, by [Wal86, Proposition 7.12], for $t_1, \dots, t_p \in [0, T]$

$$\left(\int_0^{t_1} \int_{\mathbb{R}^d} \varphi_1(x,s,t_1) M^N(dxds), \dots, \int_0^{t_k} \int_{\mathbb{R}^d} \varphi_k(x,s,t_k) M^N(dxds) \right) \xrightarrow[N\to\infty]{d} \left(\int_0^{t_1} \int_{\mathbb{R}^d} \varphi_1(x,s,t_1) W^\alpha(dxds), \dots, \int_0^{t_k} \int_{\mathbb{R}^d} \varphi_k(x,s,t_k) W^\alpha(dxds) \right).$$

Hence the same convergence holds (in distribution) for $\left(\left\langle Z_{t_1}^N, \phi_1 \right\rangle, \dots, \left\langle Z_{t_p}^N, \phi_p \right\rangle \right)$ and this characterises the potential limit points of $\left(Z^N\right)_{N\geq 1}$. By Theorem 2.6, $\left(Z_t^N\right)_{t\geq 0}$ converges in distribution to a distribution-valued process $(z_t)_{t\geq 0}$ which satisfies

$$\langle z_t, \phi \rangle = \int_0^t \int_{\mathbb{R}^d} \varphi(x,s,t) W^\alpha(dxds).$$

By the same argument as in Section 2.4.6, $(z_t)_{t\geq 0}$ solves the stochastic PDE (2.64), which concludes the proof.

2.6 Drift load - proof of Theorem 2.3

Recall the definition of F and $\rho_{z_1,z_2}^{(r_N)}$ in (2.26) and (2.41) respectively.

Definition 2.6.1 (Martingale Problem (M3))**.** *Given $(\varepsilon_N)_{N\geq 1}$, $(\delta_N)_{N\geq 1}$ and $F : \mathbb{R} \to \mathbb{R}$, let $\eta_N = \varepsilon_N \delta_N^2$, $\tau_N = \varepsilon_N^2 \delta_N^d$ and $r_N = \delta_N R$. Then for $N \geq 1$, we say that a Ξ-valued process $(w_t^N)_{t\geq 0}$ satisfies the martingale problem (M3) if for all ϕ in $L^{1,\infty}(\mathbb{R}^d)$,*

$$\left\langle w_t^N, \phi \right\rangle - \langle w_0, \phi \rangle - \eta_N u V_R \int_0^t \left\{ \frac{2R^2}{d+2} \left\langle w_s^N, \mathcal{L}^{(r_N)} \phi \right\rangle - s \left\langle \overline{F(\overline{w_s^N})}(r_N), \phi \right\rangle \right\} ds \tag{2.105}$$

defines a (mean zero) square-integrable martingale with (predictable) variation process

$$\tau_N u^2 V_R^2 \int_0^t \int_{(\mathbb{R}^d)^2} \phi(z_1)\phi(z_2) \rho_{z_1,z_2}^{(r_N)}(w_s^N) dz_1 dz_2 ds + \mathcal{O}\left(t \tau_N \delta_N^2 \left\| \phi \right\|_2^2 \right). \tag{2.106}$$

(Again, uniqueness does not hold for this martingale problem, but we will not require it.)

Let q_t^N denote the SLFVS with overdominance defined in Definition 2.1.4 with parameters as defined in (2.25) in Section 2.2.3. As in Subsection 2.3.2, we consider the rescaled process $w_t^N(x) = q_t^N(x/\delta_N)$. By Proposition 2.3.2, using the same rescaling argument as in Proposition 2.3.4, we have the following result.

Proposition 2.6.2. *The process $\left(w_t^N\right)_{t\geq 0}$ satisfies the martingale problem (M3).*

As in Theorem 2.1, we define the process of rescaled fluctuations by

$$Z_t^N = \left(\eta_N / \tau_N\right)^{1/2} \left(w_{t/\eta}^N - \lambda \right). \tag{2.107}$$

(Recall that since $w_0 = \lambda$, the centering term is constant and equals λ.) Then by the definition of

Δ^N in (2.30),

$$\Delta^N(t,x) = \delta_N^2(s_1+s_2)\varepsilon_N\delta_N^{d-2}\mathbb{E}\left[\overline{Z^N_{\eta_N t}}(\delta_N x, r_N)^2\right].$$

Let us define the following notation for any $\phi \in L^{1,\infty}(\mathbb{R}^d)$,

$$\phi_r(x) = \frac{1}{r^d}\phi(x/r). \tag{2.108}$$

Theorem 2.3 is then a direct consequence of the following theorem.

Theorem 2.8. *Suppose that $\tau_N/\eta_N = o\left(r_N^{d+2}\right)$. Then for all $\phi \in L^{1,\infty}(\mathbb{R}^d)$, there exists a constant $C > 0$ - depending on the dimension d - such that, as $N \to \infty$ with $t \to \infty$ for $d \le 2$ and $t\delta_N^{-2} \to \infty$ for $d \ge 3$,*

$$\mathbb{E}\left[\left\langle Z_t^N, \phi_{r_N}\right\rangle^2\right] \underset{N,t\to\infty}{\sim} C\delta_N^{2-d}c_N.$$

Proof of Theorem 2.3. Setting $\phi = \mathbb{1}_{\{|x|\le 1\}}$ gives the result for $\Delta^N(t,0)$; the general result follows by translation invariance. □

Note that the only difference between the martingale problems (M1) and (M3) in Definitions 2.3.3 and 2.6.1 is that $\sigma^{(r_N)}_{z_1,z_2}$ is replaced by $\rho^{(r_N)}_{z_1,z_2}$. Hence it is easy to see that Lemma 2.4.3 and Lemma 2.4.4 also hold in this case (with different constants). It is also possible to adapt the proofs in Section 2.4.5 to show that on compact time intervals, $\left(Z_t^N\right)_{t\ge 0}$ converges to the solution of the following SPDE,

$$dz_t = \left[\frac{1}{2}\Delta z_t - F'(\lambda)z_t\right]dt + \sqrt{\frac{1}{2}\lambda(1-\lambda)}dW_t.$$

This process admits a stationary distribution, under which $\langle z_t, \phi\rangle$ is a Gaussian random variable with variance

$$\frac{1}{2}\lambda(1-\lambda)\int_0^\infty\int_{\mathbb{R}^d} e^{-2F'(\lambda)t}G_t * \phi(x)^2 dx dt.$$

We can thus hope to extend the convergence of $\left(Z_t^N\right)_{t\ge 0}$ to the whole real line (as in [Nor77]), and use the above expression to estimate the second moment of $\left\langle Z_t^N, \phi_{r_N}\right\rangle$ for large times. Some care is needed though, as we are letting the support of the test function vanish as $N \to \infty$.

Proof of Theorem 2.8. Since $q_0^N = \lambda$, by the same argument as for (2.56),

$$\begin{aligned} dZ_t^N &= \left[\mathcal{L}^{(r_N)}Z_t^N - (\eta_N/\tau_N)^{1/2}\overline{\left(F(\overline{w^N_{t/\eta}}) - F(\lambda)\right)}(r_N)\right]dt + dM_t^N && (2.109)\\ &= \left[\mathcal{L}^{(r_N)}Z_t^N - F'(\lambda)\overline{\overline{Z_t^N}}(r_N) - (\tau_N/\eta_N)^{1/2}\,\overline{(\overline{Z_t^N})^2 R_2(\overline{w^N_{t/\eta}}, \lambda)}(r_N)\right]dt + dM_t^N, && (2.110)\end{aligned}$$

where M^N is a martingale measure with covariation measure Q^N given by

$$Q^N(dz_1dz_2ds) = \rho^{(r_N)}_{z_1,z_2}(w^N_{s/\eta})dz_1dz_2ds + \mathcal{O}\left(\delta_N^2\right)\delta_{z_1=z_2}(dz_1dz_2)ds. \tag{2.111}$$

Consider a time dependent test function φ^N which solves

$$\begin{cases} \partial_s\varphi^N(x,s,t) + \mathcal{L}^{(r_N)}\varphi^N(x,s,t) - F'(\lambda)\overline{\overline{\varphi^N(s,t)}}(x,r_N) = 0,\\ \varphi^N(x,t,t) = \phi(x). \end{cases} \tag{2.112}$$

Then, by (2.110), by the same argument as for (2.65),

$$\left\langle Z_t^N, \phi \right\rangle = -(\tau_N/\eta_N)^{1/2} \int_0^t \left\langle (\overline{Z_s^N})^2, R_2(\overline{w_{s/\eta}^N}, \lambda)\overline{\varphi^N(s,t)}(r_N) \right\rangle ds + \int_0^t \int_{\mathbb{R}^d} \varphi^N(x,s,t) M^N(dxds). \tag{2.113}$$

The remainder of the proof now consists of proving that the main contribution to the variance of $\left\langle Z_t^N, \phi_{r_N} \right\rangle$ is made by the last term on the right-hand-side and then estimating this contribution. Note that φ^N is given explicitly by

$$\varphi^N(x,s,t) = e^{-F'(\lambda)(t-s)} G^{(r_N)}_{D_N(t-s)} * \phi(x), \tag{2.114}$$

with $D_N = 1 - F'(\lambda)\frac{2r_N^2}{d+2}$. To see this, differentiate with respect to s and write

$$\partial_s \varphi^N(x,s,t) = F'(\lambda)\varphi^N(x,s,t) - \left(1 - F'(\lambda)\frac{2r_N^2}{d+2}\right) \mathcal{L}^{(r_N)} \varphi^N(x,s,t).$$

Since by (2.14), $\frac{2r_N^2}{d+2}\mathcal{L}^{(r_N)}\varphi^N = \overline{\overline{\varphi^N}}(r_N) - \varphi^N$, we have a solution to (2.112). In particular,

$$\left\| \varphi^N(s,t) \right\|_q \leq \|\phi\|_q \, e^{-F'(\lambda)(t-s)}. \tag{2.115}$$

The following lemma extends the result of Lemma 2.4.4 to arbitrarily large times, and will be proved in Subsection 2.6.1.

Lemma 2.6.3. *There exist constants K_1' and K_2' such that, for all $x \in \mathbb{R}^d$ and all $t \geq 0$,*

$$\mathbb{E}\left[\overline{Z_t^N}(x, r_N)^2\right] \leq \frac{K_1'}{r_N^d}, \qquad \text{and} \qquad \mathbb{E}\left[\overline{Z_t^N}(x, r_N)^4\right] \leq \frac{K_2'}{r_N^{2d}}.$$

Using the expression for φ^N in (2.114) and then the Cauchy-Schwartz inequality,

$$\begin{aligned}
&\mathbb{E}\left[\left(\int_0^t \left\langle (\overline{Z_s^N})^2, R_2(\overline{w_{s/\eta}^N}, \lambda)\overline{\varphi^N(s,t)} \right\rangle ds\right)^2\right] \\
&\quad \leq \frac{1}{4}\|F''\|_\infty^2 \, \mathbb{E}\left[\left(\int_0^t e^{-F'(\lambda)(t-s)/2}\left(e^{-F'(\lambda)(t-s)/2}\left\langle (\overline{Z_s^N})^2, \left|\overline{G^{(r_N)}_{D_N(t-s)} * \phi}\right| \right\rangle\right) ds\right)^2\right] \\
&\quad \leq \frac{1}{4}\|F''\|_\infty^2 \, \frac{1 - e^{-F'(\lambda)t}}{F'(\lambda)} \mathbb{E}\left[\int_0^t e^{-F'(\lambda)(t-s)} \left\langle (\overline{Z_s^N})^2, \left|\overline{G^{(r_N)}_{D_N(t-s)} * \phi}\right| \right\rangle^2 ds\right].
\end{aligned}$$

Another use of the Cauchy-Schwartz inequality yields

$$\left\langle (\overline{Z_s^N})^2, \left|\overline{G^{(r_N)}_{D_N(t-s)} * \phi}\right| \right\rangle^2 \leq \left\| G^{(r_N)}_{D_N(t-s)} * \phi \right\|_1 \left\langle (\overline{Z_s^N})^4, \left|\overline{G^{(r_N)}_{D_N(t-s)} * \phi}\right| \right\rangle.$$

Hence, using Lemma 2.6.3 and the fact that $\left\| G_t^{(r)} * \phi \right\|_1 \leq \|\phi\|_1$,

$$\mathbb{E}\left[\left\langle (\overline{Z_s^N})^2, \left|\overline{G^{(r_N)}_{D_N(t-s)} * \phi}\right| \right\rangle^2\right] \leq \|\phi\|_1^2 \, \frac{K_2'}{r_N^{2d}}.$$

As a result,

$$\mathbb{E}\left[\left(\int_0^t \left\langle (\overline{Z_s^N})^2, R_2(\overline{w_{s/\eta}^N}, \lambda)\overline{\varphi^N(s,t)}\right\rangle ds\right)^2\right] \lesssim \|\phi\|_1^2 \, r_N^{-2d}, \tag{2.116}$$

uniformly in $t \in \mathbb{R}_+$. We now move on to estimating the contribution of the second term in (2.113). The following lemma will be proved in Subsection 2.6.1.

Lemma 2.6.4. *As $N \to \infty$,*

$$\mathbb{E}\left[\left(\int_0^t \int_{\mathbb{R}^d} \varphi^N(x,s,t) M^N(dxds)\right)^2\right] = \frac{1}{2}\lambda(1-\lambda)\int_0^t \left\|\overline{\varphi^N(s,t)}(r_N)\right\|_2^2 ds + o(r_N)\int_0^t \left\|\varphi^N(s,t)\right\|_2^2 ds.$$

As we shall see in Subsection 2.6.1, this is a consequence of the fact that in the expression for Q^N in (2.111), w^N can be replaced by λ. As a result, using (2.114) and (2.116) in (2.113) and since $\tau_N/\eta_N = o\left(r_N^{d+2}\right)$, we have

$$\begin{aligned}\mathbb{E}\left[\left\langle Z_t^N, \phi_{r_N}\right\rangle^2\right] = \frac{1}{2}\lambda(1-\lambda)\int_0^t e^{-2F'(\lambda)s}\left\|\overline{G_{D_N s}^{(r_N)} * \phi_{r_N}}(r_N)\right\|_2^2 ds \\ + o(r_N)\int_0^t e^{-2F'(\lambda)s}\left\|G_{D_N s}^{(r_N)} * \phi_{r_N}\right\|_2^2 ds + o\left(r_N^{2-d}\right).\end{aligned} \tag{2.117}$$

To study the asymptotic behaviour of the first integral, we use the scaling properties of the function $G^{(r)}$. Recall that $\left(\xi_t^{(r)}\right)_{t\geq 0}$ is a Lévy process with infinitesimal generator $\mathcal{L}^{(r)}$; it is not difficult to show that it satisfies the following scaling property:

$$\forall c > 0, \quad \mathbb{E}_x\left[\phi(\xi_t^{(r)})\right] = \mathbb{E}_{x/c}\left[\phi(c\,\xi_{t/c^2}^{(r/c)})\right]. \tag{2.118}$$

(Simply look at the infinitesimal generator of both processes.) Hence

$$G_t^{(r_N)} * \phi_{r_N}(x) = r_N^{-d} G_{t/r_N^2}^{(1)} * \phi_1(x/r_N).$$

Set $f(t) = \left\|\overline{G_t^{(1)} * \phi}(1)\right\|_2^2$; it follows that

$$\left\|\overline{G_{D_N s}^{(r_N)} * \phi_{r_N}}(r_N)\right\|_2^2 = r_N^{-d} f(D_N s/r_N^2). \tag{2.119}$$

We shall show that, as $N, t \to \infty$, there is a constant $\tilde{C} > 0$ such that

$$\int_0^t e^{-2F'(\lambda)s} f(D_N s/r_N^2) ds \sim \tilde{C} r_N^2 c_N. \tag{2.120}$$

Theorem 2.8 then follows from (2.120) and (2.117). To prove (2.120) we need the following estimate of $f(t)$ when $t \to \infty$.

Lemma 2.6.5. *For $\phi \geq 0$, as $t \to \infty$,*

$$f(t) \sim (4\pi t)^{-d/2} \|\phi\|_1^2. \tag{2.121}$$

For the proof of this estimate we will use the following properties of the semigroup $G^{(r)}$, which will be proved in Section 2.10.

Lemma 2.6.6. *For any $r > 0$ and $t > 0$, the law of $\xi_t^{(r)}$ takes the form*

$$G_t^{(r)}(dx) = e^{-\frac{(d+2)}{2r^2}t}\delta_0(dx) + g_t^{(r)}(x)dx.$$

Furthermore, $g_t^{(r)}$ is continuous on $\mathbb{R}^d$, is invariant under rotations which fix the origin and $g_t^{(r)}(y)$ is a decreasing function of $|y|$.

Proof of Lemma 2.6.5. By the semigroup property of $\phi \mapsto G_t^{(r)} * \phi$, $f(t)$ can also be written $\left\langle G_{2t}^{(1)} * \overline{\phi}(1), \overline{\phi}(1)\right\rangle$. In addition, by the scaling property of $\left(\xi_t^{(r)}\right)_{t\geq 0}$ in (2.118) and using Lemma 2.6.6,

$$\begin{aligned} G_{2t}^{(1)} * \phi(x) &= \mathbb{E}_0\left[\phi(x + \sqrt{t}\xi_2^{(1/\sqrt{t})})\right] \\ &= \phi(x)e^{-(d+2)t} + \int_{\mathbb{R}^d} \phi(x + \sqrt{t}y) g_2^{(1/\sqrt{t})}(y)dy \\ &= \phi(x)e^{-(d+2)t} + t^{-d/2}\int_{\mathbb{R}^d} \phi(x+y) g_2^{(1/\sqrt{t})}(y/\sqrt{t})dy. \end{aligned}$$

By Proposition 2.7.2.ii and Theorem 4.8.2 in [EK86], the finite dimensional distributions of $\left(\xi_t^{(r)}\right)_{t\geq 0}$ converge to those of standard Brownian motion as $r \to 0$. In particular, $\xi_2^{(r)} \xrightarrow[r\to 0]{d} \mathcal{N}(0,2)$, and $g_2^{(r)}(x) \to G_2(x)$ as $r \to 0$ for almost every $x \in \mathbb{R}^d$ (the probability that $\xi_t^{(r)} = 0$ vanishes as $r \to 0$ for any $t > 0$). Since G_2 is continuous on $\mathbb{R}^d$ and $g_2^{(r)}$ is decreasing as a function of the modulus, this convergence takes place uniformly on compact sets by Dini's second theorem. So, fixing $\epsilon > 0$, for any $R > 0$, for r small enough,

$$\sup_{|x|<R}\left|g_2^{(r)}(x) - G_2(x)\right| \leq \epsilon.$$

As a result, using the continuity of G_2, for any y, for t large enough,

$$\left|g_2^{(1/\sqrt{t})}(y/\sqrt{t}) - G_2(0)\right| \leq 2\epsilon.$$

Hence, since $g_t^{(r)}(y) \leq g_t^{(r)}(0)$, by dominated convergence,

$$\int_{\mathbb{R}^d} \phi(x+y) g_2^{(1/\sqrt{t})}(y/\sqrt{t})dy \underset{t\to\infty}{\longrightarrow} (4\pi)^{-d/2}\int_{\mathbb{R}^d}\phi(y)dy. \tag{2.122}$$

From the above expression for f,

$$f(t) = e^{-(d+2)t}\int_{\mathbb{R}^d}\overline{\phi}(x,1)^2dx + t^{-d/2}\int_{(\mathbb{R}^d)^2} g_2^{(1/\sqrt{t})}(y/\sqrt{t})\overline{\phi}(x+y,1)\overline{\phi}(x,1)dydx.$$

Replacing ϕ with $\overline{\phi}(1)$ in (2.122) and letting $t \to \infty$ yields the result. □

Furthermore, $0 \leq f(t) \leq \|\phi\|_2^2$ for all $t \geq 0$, and thus f is integrable on $(0,\infty)$ if and only if $d \geq 3$.

Remark. *This is in fact a consequence of the fact that $\left(\xi_t^{(1)}\right)_{t\geq 0}$ is transient if and only if $d \geq 3$*

(as with Brownian motion). The function f can be expressed in terms of the probability of $\xi_{2t}^{(1)}$ being in a ball of radius 1, which is integrable on $(0,\infty)$ if and only if $\left(\xi_t^{(1)}\right)_{t\geq 0}$ is transient.

We now prove (2.120) separately for each regime.

High dimension If $d \geq 3$, change the variable of integration to write

$$\int_0^t e^{-2F'(\lambda)s} f(D_N s/r_N^2)ds = r_N^2 \int_0^{D_N t/r_N^2} e^{-2F'(\lambda)D_N^{-1}r_N^2 s} f(s)ds.$$

Since f is integrable, by dominated convergence, and since $t\delta_N^{-2} \to \infty$,

$$\int_0^{D_N t/r_N^2} e^{-2F'(\lambda)D_N^{-1}r_N^2 s} f(s)ds \underset{N,t\to\infty}{\longrightarrow} \int_0^\infty f(s)ds.$$

(Also recall that $D_N = 1 + \mathcal{O}\left(r_N^2\right)$.)

Dimension 1 If $d = 1$, however, from (2.121), we see that, as $N \to \infty$, $\frac{1}{r_N} f(s/r_N^2) \to (4\pi s)^{-1/2} \|\phi\|_1^2$, so, by dominated convergence,

$$\int_0^t e^{-2F'(\lambda)s} f(D_N s/r_N^2)ds \underset{N,t\to\infty}{\sim} r_N \|\phi\|_1^2 \hat{C},$$

for some constant $\hat{C} > 0$.

Dimension 2 If $d = 2$, let T_1 and T_2 be two positive constants. For $t \geq T_2$ and N large enough such that $r_N^2 T_1 \leq T_2$, we split the integral as follows :

$$\begin{aligned}\int_0^t e^{-2F'(\lambda)s} f(D_N s/r_N^2)ds = &\int_0^{r_N^2 T_1} e^{-2F'(\lambda)s} f(D_N s/r_N^2)ds \\ &+ \int_{r_N^2 T_1}^{T_2} e^{-2F'(\lambda)s} f(D_N s/r_N^2)ds + \int_{T_2}^t e^{-2F'(\lambda)s} f(D_N s/r_N^2)ds.\end{aligned}$$

We first show that the first and last terms are of order r_N^2. Since $0 \leq f(t) \leq \|\phi\|_2^2$ for all $t \geq 0$,

$$\left|\int_0^{r_N^2 T_1} e^{-2F'(\lambda)s} f(D_N s/r_N^2)ds\right| \leq r_N^2 T_1 \|\phi\|_2^2,$$

and by (2.121)

$$\left|\int_{T_2}^t e^{-2F'(\lambda)s} f(D_N s/r_N^2)ds\right| \lesssim r_N^2 \int_{T_2}^\infty e^{-2F'(\lambda)s} \frac{ds}{s}.$$

For the middle term, by (2.121), $\frac{1}{r_N^2} f(s/r_N^2) \underset{N\to\infty}{\longrightarrow} (4\pi s)^{-1} \|\phi\|_1^2$, and $D_N = 1 + \mathcal{O}\left(r_N^2\right)$ so as $N \to \infty$, by dominated convergence,

$$\int_{r_N^2 T_1}^{T_2} e^{-2F'(\lambda)s} f(D_N s/r_N^2) ds \sim r_N^2 (4\pi)^{-1} \|\phi\|_1^2 \int_{r_N^2 T_1}^{T_2} e^{-2F'(\lambda)s} \frac{ds}{s}.$$

Further

$$\left| \int_{T_1 r_N^2}^{T_2} e^{-2F'(\lambda)s} \frac{ds}{s} - \int_{T_1 r_N^2}^{T_2} \frac{ds}{s} \right| \leq 2F'(\lambda) \int_{T_1 r_N^2}^{T_2} s \frac{ds}{s} \leq 2F'(\lambda) T_2,$$

and

$$\int_{T_1 r_N^2}^{T_2} \frac{ds}{s} = \log\left(\frac{T_2}{T_1 r_N^2}\right) \sim \left|\log r_N^2\right|.$$

As a result

$$\int_0^t e^{-2F'(\lambda)s} f(D_N s/r_N^2) ds \sim \frac{\|\phi\|_1^2}{4\pi} r_N^2 \left|\log r_N^2\right|,$$

as $N, t \to \infty$. We have thus proved (2.120), and the result. □

2.6.1 Proofs of Lemmas 2.6.3 and 2.6.4

The proof of Lemma 2.6.3 requires the following two technical lemmas, which are proved in Section 2.10.

Lemma 2.6.7. *Let $\phi : \mathbb{R}^d \to \mathbb{R}$, $r > 0$ and suppose that $g : \mathbb{R}^d \to \mathbb{R}$ satisfies $0 < \gamma \leq g(x) \leq 1$ for all $x \in \mathbb{R}^d$. Then*

$$2\phi(x)\mathcal{L}^{(r)}\phi(x) - 2\phi(x)\overline{\overline{\phi\, g}}(x,r) \leq \mathcal{L}^{(r)}\phi^2(x) - 2\left(\gamma - \frac{r^2}{d+2}\right)\phi(x)^2.$$

Further, for some constant $c > 0$, for r small enough,

$$4\phi(x)^3\mathcal{L}^{(r)}\phi(x) - 4\phi(x)^3\overline{\overline{\phi\, g}}(x,r) \leq \mathcal{L}^{(r)}\phi^4(x) - 4(\gamma - c\, r^2)\phi(x)^4.$$

Lemma 2.6.8. *Suppose $h : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ is a function that is continuously differentiable with respect to the time variable t and which satisfies the following differential inequality for some positive α :*

$$\partial_t h_t(x) \leq \mathcal{L} h_t(x) - \alpha h_t(x) + g_t(x).$$

Then for all $0 \leq s \leq t$ and for any $1 \leq q \leq \infty$,

$$\|h_t\|_q \leq e^{-\alpha(t-s)} \|h_s\|_q + \frac{1}{\alpha} \sup_{u\in[s,t]} \|g_u\|_q.$$

Proof of Lemma 2.6.3. Set

$$h(t,x) = \mathbb{E}\left[\overline{Z_t^N}(x, r_N)^2\right].$$

We are going to make use of Lemma 2.6.8, so we want to obtain a differential inequality for h. To this end, average (2.109) on $B(x, r_N)$ and use the expression for R_1 in (2.54) to get

$$d\overline{Z_s^N}(x, r_N) = \left[\mathcal{L}^{(r_N)}\overline{Z_s^N}(x, r_N) - \overline{\overline{Z_s^N\, R_1(\overline{w_{s/\eta}^N}, \lambda)}}(x, r_N)\right] ds + \frac{1}{V_{r_N}} dM_s^N(B(x, r_N)).$$

(From now on all averages will be over radius r_N.) By the generalised Itô formula, noting $\Delta Y_s = Y_s - Y_{s^-}$,

$$d\left(\overline{Z_s^N}(x)\right)^2 = 2\overline{Z_s^N}(x)d\overline{Z_s^N}(x) + d\left[\overline{Z_\cdot^N}(x)\right]_s + \left(\overline{Z_{s^-}^N}(x) + \Delta\overline{Z_s^N}(x)\right)^2 - \left(\overline{Z_{s^-}^N}(x)\right)^2 - 2\overline{Z_{s^-}^N}(x)\Delta\overline{Z_s^N}(x) - \left(\Delta\overline{Z_s^N}(x)\right)^2.$$

Expanding the brackets, the terms on the second line cancel and, integrating for $s \in [0,t]$, we have

$$\overline{Z_t^N}(x)^2 = 2\int_0^t \overline{Z_s^N}(x)\left[\mathcal{L}^{(r_N)}\overline{Z_s^N}(x) - \overline{\overline{Z_s^N}\, R_1(\overline{w_{s/\eta}^N}, \lambda)}(x)\right] ds + \frac{2}{V_{r_N}}\int_0^t \overline{Z_s^N}(x) dM_s^N(B(x,r_N)) + \frac{1}{V_{r_N}^2}\left[M^N(B(x,r_N))\right]_t.$$

Taking expectations on both sides, since the second term is a martingale,

$$h(t,x) = 2\int_0^t \mathbb{E}\left[\overline{Z_s^N}(x)\mathcal{L}^{(r_N)}\overline{Z_s^N}(x) - \overline{Z_s^N}(x)\overline{\overline{Z_s^N}\, R_1(\overline{w_{s/\eta}^N}, \lambda)}(x)\right] ds + \frac{1}{V_{r_N}^2}\mathbb{E}\left[\left\langle M^N(B(x,r_N))\right\rangle_t\right].$$

Differentiating yields

$$\frac{\partial h}{\partial t}(t,x) = 2\mathbb{E}\left[\overline{Z_t^N}(x)\mathcal{L}^{(r_N)}\overline{Z_t^N}(x) - \overline{Z_t^N}(x)\overline{\overline{Z_t^N}\, R_1(\overline{w_{t/\eta}^N}, \lambda)}(x)\right] + \frac{1}{V_{r_N}^2}\mathbb{E}\left[\int_{B(x,r_N)^2} \rho_{z_1,z_2}^{(r_N)}(w_{t/\eta}^N) dz_1 dz_2\right] + \mathcal{O}\left(\frac{\delta_N^2}{V_{r_N}}\right).$$

The second term is bounded by $\frac{1}{V_{r_N}}$, and the first one has the same form as the left-hand-side of the first statement of Lemma 2.6.7. In [Nor74b] (at the beginning of the proof of Theorem 3.2), it is proved that the conditions on F in (2.27)-(2.28) imply

$$\inf_{x\in[0,1]} R_1(x,\lambda) =: \gamma > 0. \tag{2.123}$$

Then, taking $\phi = \overline{Z_t^N}$ and $g = R_1(\overline{w_{s/\eta}^N}, \lambda)$, Lemma 2.6.7 implies that, for all $t \geq 0$,

$$\frac{\partial h}{\partial t}(t,x) \leq \mathcal{L}h(t,x) - \alpha_N h(t,x) + \frac{1 + \mathcal{O}\left(\delta_N^2\right)}{V_{r_N}},$$

with $\alpha_N = \gamma + \mathcal{O}\left(r_N^2\right)$. Using Lemma 2.6.8 (with $s = 0$ and $q = \infty$) we can now write, since $Z_0^N = 0$,

$$\mathbb{E}\left[\overline{Z_t^N}(x)^2\right] \leq \frac{1 + \mathcal{O}\left(\delta_N^2\right)}{\alpha_N V_{r_N}} \lesssim \frac{1}{r_N^d}. \tag{2.124}$$

The second inequality is proved in essentially the same way, although the computations become more

involved. We compute the fourth moment of $\overline{Z_t^N}$ with Itô's formula, as before:

$$d\left(\overline{Z_t^N}(x)\right)^4 = 4(\overline{Z_t^N}(x))^3 d\overline{Z_t^N}(x) + \frac{1}{2}4\times 3(\overline{Z_t^N}(x))^2 d\left[\overline{Z_\cdot^N}\right]_t$$
$$+\left(\overline{Z_{t^-}^N}(x)+\Delta\overline{Z_t^N}(x)\right)^4 - \left(\overline{Z_{t^-}^N}(x)\right)^4 - 4(\overline{Z_{t^-}^N}(x))^3\Delta\overline{Z_t^N}(x) - \frac{1}{2}3\times 4(\overline{Z_{t^-}^N}(x))^2(\Delta\overline{Z_t^N}(x))^2.$$

Hence, taking expectations, the martingale terms can be dropped and we write:

$$\mathbb{E}\left[\left(\overline{Z_t^N}(x)\right)^4\right] = 4\int_0^t \mathbb{E}\left[\overline{Z_s^N}(x)^3\mathcal{L}^{(r_N)}\overline{Z_s^N}(x) - \overline{Z_s^N}(x)^3\overline{\overline{Z_s^N}\,R_1(\overline{w_{s/\eta}^N},\lambda)}(x)\right]ds$$
$$+6\frac{1}{V_{r_N}^2}\int_0^t\int_{B(x,r_N)^2}\mathbb{E}\left[\overline{Z_s^N}(x)^2\rho_{z_1,z_2}^{(r_N)}(w_{s/\eta}^N)\right]dz_1dz_2ds + \mathcal{O}\left(\delta_N^2\right)\frac{1}{V_{r_N}}\int_0^t\mathbb{E}\left[\overline{Z_s^N}(x)^2\right]ds$$
$$+\mathbb{E}\left[\sum_{s\le t}\left\{4\overline{Z_{s^-}^N}(x)(\Delta\overline{Z_s^N}(x))^3 + (\Delta\overline{Z_s^N}(x))^4\right\}\right],$$

where the sum is over jump times for the process $(\overline{Z_t^N}(x))_{t\ge 0}$. We can bound the size of the jumps $\Delta\overline{Z_s^N}(x)$ by a deterministic constant. By the definition of the SLFVS with overdominance in Definition 2.1.4,

$$\sup_{t\ge 0}\left|\left\langle q_t^N,\phi\right\rangle - \left\langle q_{t^-}^N,\phi\right\rangle\right| \le u\varepsilon_N\left\|\phi\right\|_1.$$

Hence $\left|\Delta\overline{Z_s^N}(x)\right| \le u\varepsilon_N(\eta_N/\tau_N)^{1/2} = u\varepsilon_N^{1/2}\delta_N^{1-d/2}$. As a result

$$\mathbb{E}\left[\sum_{s\le t}\left\{4\overline{Z_{s^-}^N}(x)(\Delta\overline{Z_s^N}(x))^3 + (\Delta\overline{Z_s^N}(x))^4\right\}\right]$$
$$\le \mathbb{E}\left[\sum_{s\le t}\left\{4(u\varepsilon_N^{1/2}\delta_N^{1-d/2})^3\left|\overline{Z_{s^-}^N}(x)\right| + (u\varepsilon_N^{1/2}\delta_N^{1-d/2})^4\right\}\right],$$

where the sum is still over the jump times of $\overline{Z_s^N}(x)$. These jumps occur according to a Poisson process with rate $V_{2R}\,\eta_N^{-1}$, so, using (2.124) to bound $\mathbb{E}\left[\left|\overline{Z_{s^-}^N}(x)\right|\right]$, we obtain

$$\mathbb{E}\left[\sum_{s\le t}\left\{4\overline{Z_{s^-}^N}(x)(\Delta\overline{Z_s^N}(x))^3 + (\Delta\overline{Z_s^N}(x))^4\right\}\right]$$
$$\le V_{2R}\,\eta_N^{-1}\left\{4(u\varepsilon_N^{1/2}\delta_N^{1-d/2})^3\mathbb{E}\left[\int_0^t\left|\overline{Z_{s^-}^N}(x)\right|ds\right] + t(u\varepsilon_N^{1/2}\delta_N^{1-d/2})^4\right\} = o\left(r_N^{-2d}\right).$$

Now note that

$$\int_{B(x,r_N)^2}\mathbb{E}\left[\overline{Z_s^N}(x)^2\rho_{z_1,z_2}^{(r)}(w_{s/\eta}^N)\right]dz_1dz_2 \lesssim \frac{1}{r_N^d}\int_{B(x,r_N)^2}\frac{V_{r_N}(z_1,z_2)}{V_{r_N}^2}dz_1dz_2$$
$$\lesssim 1.$$

Hence, setting $h(t,x) = \mathbb{E}\left[(\overline{Z_t^N}(x))^4\right]$,

$$\frac{\partial h}{\partial t}(t,x) = 4\mathbb{E}\left[\overline{Z_t^N}(x)^3 \mathcal{L}^{(r_N)}\overline{Z_t^N}(x) - \overline{Z_t^N}(x)^3 \overline{\overline{Z_t^N}\, R_1(\overline{w_{t/\eta}^N}, \lambda)}\right] + \frac{g_t(x)}{r_N^{2d}},$$

where $|g_t(x)| \lesssim 1$. Now the second statement of Lemma 2.6.7 yields :

$$\frac{\partial h}{\partial t}(t,x) \leq \mathcal{L}h(t,x) - 4(\gamma - cr_N^2)h(t,x) + \frac{g_t(x)}{r_N^{2d}},$$

and by Lemma 2.6.8, we have

$$h(t,x) \lesssim \frac{1}{r_N^{2d}},$$

uniformly in $t \geq 0$. □

The following lemma is needed in the proof of Lemma 2.6.4.

Lemma 2.6.9. *The following holds uniformly for all $t \geq 0$:*

$$\mathbb{E}\left[\left|\left\langle Z_t^N, \phi\right\rangle\right|\right] \lesssim r_N^{1-d/2} c_N^{1/2}(\|\phi\|_1 + r_N^{d/2}\|\phi\|_2).$$

Proof. Recall the expression for $\left\langle Z_t^N, \phi\right\rangle$ in (2.113) and the expression for φ^N in (2.114); using Lemma 2.6.3 and Lemma 2.4.3, we can write

$$\mathbb{E}\left[\left|\left\langle Z_t^N, \phi\right\rangle\right|\right] \lesssim \frac{(\tau_N/\eta_N)^{1/2}}{r_N^d}\int_0^t \|\phi\|_1 e^{-F'(\lambda)(t-s)}ds + \left(\int_0^t e^{-2F'(\lambda)(t-s)}\left\|G^{(r_N)}_{D_N(t-s)} * \phi\right\|_2^2 ds\right)^{1/2}.$$

Replacing ϕ by $(\phi_{1/r_N})_{r_N}$ - as defined in (2.108) - to use (2.119) and then looking at the proof of (2.120) in the proof of Theorem 2.8, we see that

$$\int_0^t e^{-2F'(\lambda)(t-s)}\left\|G^{(r_N)}_{D_N(t-s)} * \phi\right\|_2^2 ds \lesssim r_N^{2-d} c_N(\left\|\phi_{1/r_N}\right\|_1^2 + \left\|\phi_{1/r_N}\right\|_2^2).$$

But $\left\|\phi_{1/r_N}\right\|_1 = \|\phi\|_1$ and $\left\|\phi_{1/r_N}\right\|_2 = r_N^{d/2}\|\phi\|_2$, hence

$$\mathbb{E}\left[\left|\left\langle Z_t^N, \phi\right\rangle\right|\right] \lesssim \|\phi\|_1 \frac{(\tau_N/\eta_N)^{1/2}}{r_N^d} + r_N^{1-d/2} c_N^{1/2}(\|\phi\|_1 + r_N^{d/2}\|\phi\|_2),$$

and we have the required result since $\tau_N/\eta_N = o\left(r_N^{d+2}\right)$. □

Proof of Lemma 2.6.4. We drop the superscript N from φ^N throughout the proof and take averages over the radius $r := r_N$. Recall from the expressions for Q^N in (2.111) and $\rho^{(r)}$ in (2.41) that the

variance of the stochastic integral $\int_0^t \int_{\mathbb{R}^d} \varphi(x,s,t)\ M^N(dxds)$ is given by

$$\int_0^t \int_{(\mathbb{R}^d)^3} \frac{1}{V_r^2} \mathbb{1}_{\left\{\substack{|x-z_1|<r\\|x-z_2|<r}\right\}} \varphi(z_1,s,t)\varphi(z_2,s,t) \mathbb{E}\Big[\overline{w^N_{s/\eta}}(x,r_N)^2(1-w^N_{s/\eta}(z_1))(1-w^N_{s/\eta}(z_2)) \\ + 2\overline{w^N_{s/\eta}}(x,r_N)(1-\overline{w^N_{s/\eta}}(x,r_N))(\tfrac{1}{2}-w^N_{s/\eta}(z_1))(\tfrac{1}{2}-w^N_{s/\eta}(z_2)) \\ + (1-\overline{w^N_{s/\eta}}(x,r_N))^2 w^N_{s/\eta}(z_1)w^N_{s/\eta}(z_2)\Big] dxdz_1dz_2ds + \mathcal{O}\left(\delta_N^2\right) \int_0^t \|\varphi(s,t)\|_2^2\, ds,$$

which can also be written

$$\int_0^t \mathbb{E}\left[\left\langle (\overline{w^N_{s/\eta}})^2, \left(\overline{(1-w^N_{s/\eta})\varphi(s,t)}\right)^2\right\rangle + \left\langle 2\overline{w^N_{s/\eta}}(1-\overline{w^N_{s/\eta}}), \left(\overline{(\tfrac{1}{2}-w^N_{s/\eta})\varphi(s,t)}\right)^2\right\rangle \right. \\ \left. + \left\langle (1-\overline{w^N_{s/\eta}})^2, \left(\overline{w^N_{s/\eta}\varphi(s,t)}\right)^2\right\rangle + \mathcal{O}\left(\delta_N^2 \|\varphi(s,t)\|_2^2\right)\right] ds. \tag{2.125}$$

We want to show that in this expression, $w^N_{s/\eta}$ can (asymptotically) be replaced by λ, hence we write

$$\left\langle (\overline{w^N_{t/\eta}})^2, \left(\overline{(1-w^N_{t/\eta})\varphi}\right)^2\right\rangle - \langle \lambda^2, (1-\lambda)^2\overline{\varphi}^2\rangle \\ = \left\langle (\overline{w^N_{t/\eta}})^2 - \lambda^2, \left(\overline{(1-w^N_{t/\eta})\varphi}\right)^2\right\rangle + \left\langle \lambda^2, \left(\overline{(1-w^N_{t/\eta})\varphi}\right)^2 - (1-\lambda)^2\overline{\varphi}^2\right\rangle.$$

Since $(\overline{w^N_{t/\eta}})^2 - \lambda^2 = (\tau_N/\eta_N)^{1/2}\overline{Z^N_t}(\overline{w^N_{t/\eta}} + \lambda)$, using Lemma 2.6.3,

$$\mathbb{E}\left[\left|\left\langle (\overline{w^N_{t/\eta}})^2 - \lambda^2, \left(\overline{(1-w^N_{t/\eta})\varphi}\right)^2\right\rangle\right|\right] \le 2(\tau_N/\eta_N)^{1/2}\left\langle \mathbb{E}\left[(\overline{Z^N_t})^2\right]^{1/2}, \overline{|\varphi|}^2\right\rangle \\ \lesssim \frac{(\tau_N/\eta_N)^{1/2}}{r_N^{d/2}}\|\varphi\|_2^2 = o\left(r_N\|\varphi\|_2^2\right).$$

In addition,

$$\left\langle \lambda^2, \left(\overline{(1-w^N_{t/\eta})\varphi}\right)^2 - (1-\lambda)^2\overline{\varphi}^2\right\rangle \\ = \lambda^2 \int_{(\mathbb{R}^d)^3} \frac{1}{V_r^2}\mathbb{1}_{\left\{\substack{|x-z_1|<r\\|x-z_2|<r}\right\}} \varphi(z_1)\varphi(z_2)\Big\{(1-w^N_{t/\eta}(z_1))(1-w^N_{t/\eta}(z_2)) - (1-\lambda)^2 \\ + (1-\lambda)(1-w^N_{t/\eta}(z_1)) - (1-\lambda)(1-w^N_{t/\eta}(z_2))\Big\} dxdz_1dz_2 \\ = \lambda^2 \int_{(\mathbb{R}^d)^3} \frac{1}{V_r^2}\mathbb{1}_{\left\{\substack{|x-z_1|<r\\|x-z_2|<r}\right\}} \varphi(z_1)\varphi(z_2)((1-w^N_{t/\eta}(z_1)) - (1-\lambda))(1-w^N_{t/\eta}(z_2)+1-\lambda) dxdz_1dz_2.$$

(The last two terms inside the curly braces cancel out by permuting z_1 and z_2.) Thus,

$$\left|\left\langle \lambda^2, \left(\overline{(1-w^N_{t/\eta})\varphi}\right)^2 - (1-\lambda)^2\overline{\varphi}^2\right\rangle\right| \le 2\lambda^2(\tau_N/\eta_N)^{1/2}\int_{\mathbb{R}^d} |\varphi(z_2)|\left|\left\langle Z^N_t, \psi^N_{z_2}\right\rangle\right| dz_2,$$

where $\psi^N_{z_2}(z_1) = \frac{V_r(z_1,z_2)}{V_r^2}\varphi(z_1)$. In particular,

$$\left\|\psi^N_{z_2}\right\|_1 = \overline{\overline{|\varphi|}}(z_2, r_N), \qquad \text{and} \qquad \left\|\psi^N_{z_2}\right\|_2^2 \le \frac{1}{V_{r_N}}\overline{\overline{|\varphi|^2}}(z_2, r_N). \tag{2.126}$$

By Lemma 2.6.9, we get

$$\begin{aligned}\mathbb{E}\left[\int_{\mathbb{R}^d} |\varphi(z_2)| \left|\left\langle Z^N_t, \psi_{z_2}\right\rangle\right| dz_2\right] &\lesssim r_N^{1-d/2} c_N^{1/2} \int_{\mathbb{R}^d} |\varphi(z_2)| \left(\left\|\psi^N_{z_2}\right\|_1 + r_N^{d/2}\left\|\psi^N_{z_2}\right\|_2\right) dz_2 \\ &\lesssim r_N^{1-d/2} c_N^{1/2} \|\varphi\|_2 \left(\int_{\mathbb{R}^d} (\left\|\psi^N_{z_2}\right\|_1^2 + r_N^d \left\|\psi^N_{z_2}\right\|_2^2) dz_2\right)^{1/2},\end{aligned}$$

using the Cauchy-Schwartz inequality in the second line. By (2.126),

$$\left(\int_{\mathbb{R}^d} (\left\|\psi^N_{z_2}\right\|_1^2 + r_N^d \left\|\psi^N_{z_2}\right\|_2^2) dz_2\right)^{1/2} \le \left(\int_{\mathbb{R}^d} \left(\overline{\overline{|\phi|^2}}(z_2, r_N) + \frac{1}{V_1}\overline{\overline{|\phi|^2}}(z_2, r_N)\right) dz_2\right)^{1/2} \lesssim \|\varphi\|_2 .$$

Since $\tau_N/\eta_N = o\left(r_N^{d+2}\right)$,

$$\mathbb{E}\left[\left|\left\langle \lambda^2, \left(\overline{(1 - w^N_{t/\eta})\varphi}\right)^2 - (1-\lambda)^2\overline{\varphi}^2\right\rangle\right|\right] = o\left(r_N^2 c_N^{1/2} \|\varphi\|_2^2\right).$$

We use a similar argument for the other terms in (2.125) to show that replacing $w^N_{s/\eta}$ by λ makes a difference of $o\left(r_N^2 c_N^{1/2} \|\varphi\|_2^2\right)$. We have thus shown that, since $r_N c_N^{1/2} \underset{N\to\infty}{\longrightarrow} 0$,

$$\begin{aligned}\mathbb{E}\Bigg[\left\langle (\overline{w^N_{s/\eta}})^2, \left(\overline{(1 - w^N_{s/\eta})\varphi}\right)^2\right\rangle + \left\langle 2\overline{w^N_{s/\eta}}(1 - \overline{w^N_{s/\eta}}), \left(\overline{(\tfrac{1}{2} - w^N_{s/\eta})\varphi}\right)^2\right\rangle \\ + \left\langle (1 - \overline{w^N_{s/\eta}})^2, \left(\overline{w^N_{s/\eta}\varphi}\right)^2\right\rangle\Bigg] = \tfrac{1}{2}\lambda(1-\lambda)\|\overline{\varphi}\|_2^2 + o\left(r_N \|\varphi\|_2^2\right),\end{aligned}$$

uniformly in $s \ge 0$. The result follows from the bound on $\left\|\varphi^N\right\|_q$ in (2.115). □

2.7 Approximating the (fractional) Laplacian

We start by stating some basic properties of averaged functions which are used throughout the chapter.

Proposition 2.7.1. *Let $\phi : \mathbb{R}^d \to \mathbb{R}$ and $\psi : \mathbb{R}^d \to \mathbb{R}$ be in $L^{1,\infty}(\mathbb{R}^d)$. Then*

i) $\overline{\phi * \psi} = \overline{\phi} * \psi = \phi * \overline{\psi}$,

ii) $\left\langle \overline{\phi}, \psi\right\rangle = \left\langle \phi, \overline{\psi}\right\rangle$.

iii) If in addition $\phi \in L^q(\mathbb{R}^d)$, by Jensen's inequality, $\left\|\overline{\phi}\right\|_q \le \|\phi\|_q$.

iv) If β is a multi-index in $\mathbb{N}_0^d$, and ϕ is differentiable enough that $\partial_\beta\phi$ is well defined on $\mathbb{R}^d$, then $\partial_\beta\overline{\phi} = \overline{\partial_\beta\phi}$.

v) Also, $\partial_\beta(\psi * \phi) = \psi * \partial_\beta\phi$.

We use here the notation $\lesssim$ defined in (2.84).

Proposition 2.7.2. *Let* $\phi : \mathbb{R}^d \to \mathbb{R}$ *be twice continuously differentiable and suppose that* $\|\partial_\beta\phi\|_q < \infty$ *for* $0 \le |\beta| \le 2$ *and* $1 \le q \le \infty$. *Then*

i) $\left\|\overline{\phi}(r) - \phi\right\|_q \le \frac{d}{2} r^2 \max_{|\beta|=2} \|\partial_\beta\phi\|_q$.

If in addition, ϕ *admits* $\|\cdot\|_q$*-bounded derivatives of up to the fourth order,*

ii) $\left\|\overline{\overline{\phi}}(r) - \phi - \frac{r^2}{d+2}\Delta\phi\right\|_q \le \frac{d^3}{3} r^4 \max_{|\beta|=4} \|\partial_\beta\phi\|_q$.

Proof of Proposition 2.7.2. By Taylor's theorem,

$$\phi(y) = \phi(x) + \sum_{i=1}^{d} \partial_i\phi(x)(y-x)_i + \sum_{i,j} R_{ij}(y)(y-x)_{ij},$$

where $R_{ij}(y) = \int_0^1 (1-t)\partial_{ij}\phi(x + t(y-x))dt$ (we use the notation $x_{i_1\dots i_k} = x_{i_1}\dots x_{i_k}$). By symmetry, the average of the first sum over a ball of centre x and radius r vanishes, and

$$\left|\overline{\phi}(x,r) - \phi(x)\right| \le \sum_{i,j} \frac{1}{V_r} \int_{B(x,r)} |R_{ij}(y)|\,|y-x|_{ij}\,dy. \tag{2.127}$$

If $q = \infty$, then $|R_{ij}(y)| \le \frac{1}{2}\|\partial_{ij}\phi\|_\infty$ and we write

$$\left\|\overline{\phi}(r) - \phi\right\|_\infty \le \frac{1}{2} d \max_{|\beta|=2} \|\partial_\beta\phi\|_\infty \frac{1}{V_r}\int_{B(0,r)} |y|^2\,dy = \frac{d^2}{2(d+2)} r^2 \max_{|\beta|=2} \|\partial_\beta\phi\|_\infty .$$

If instead $1 \le q < \infty$, write

$$\begin{aligned}\left\|\overline{\phi}(r) - \phi\right\|_q &\le \sum_{i,j} \left(\int_{\mathbb{R}^d} \left(\frac{1}{V_r}\int_{B(0,r)} |R_{ij}(x+y)|\,|y|_{ij}\,dy\right)^q dx\right)^{1/q} \\ &\le \sum_{i,j} \left(\int_{\mathbb{R}^d} \left(\frac{1}{V_r}\int_{B(0,r)} |y|_{ij}\,dy\right)^{q-1} \frac{1}{V_r}\int_{B(0,r)} |R_{ij}(x+y)|^q\,|y|_{ij}\,dydx\right)^{1/q},\end{aligned}$$

by Hölder's inequality. But, by the definition of R_{ij}

$$\begin{aligned}\int_{\mathbb{R}^d} |R_{ij}(x+y)|^q\,dx &\le \frac{1}{2^{q-1}}\int_0^1 (1-t)\int_{\mathbb{R}^d} |\partial_{ij}\phi(x+ty)|^q\,dxdt \\ &= \frac{1}{2^q}\|\partial_{ij}\phi\|_q^q .\end{aligned}$$

Plugging this into the previous inequality, we get

$$\begin{aligned}\left\|\overline{\phi}(r)-\phi\right\|_q &\leq \sum_{i,j}\frac{1}{2}\left\|\partial_{ij}\phi\right\|_q\left(\left(\frac{1}{V_r}\int_{B(0,r)}|y|_{ij}\,dy\right)^{q-1}\frac{1}{V_r}\int_{B(0,r)}|y|_{ij}\,dy\right)^{1/q}\\ &\leq \frac{1}{2}d\max_{|\beta|=2}\left\|\partial_\beta\phi\right\|_q\frac{1}{V_r}\int_{B(0,r)}|y|^2\,dy\\ &\leq \frac{d}{2}r^2\max_{|\beta|=2}\left\|\partial_\beta\phi\right\|_q.\end{aligned}$$

The second inequality is proved in essentially the same way. We expand ϕ according to Taylor's theorem to the fourth order:

$$\begin{aligned}\phi(y)=\phi(x)+\sum_i\partial_i\phi(x)(y-x)_i+\frac{1}{2}\sum_{i,j}\partial_{ij}\phi(x)(y-x)_{ij}\\ +\frac{1}{3!}\sum_{i,j,k}\partial_{ijk}\phi(x)(y-x)_{ijk}+\sum_{ijkl}R_{ijkl}(y)(y-x)_{ijkl},\end{aligned}$$

where $R_{ijkl}(y)=\frac{1}{3!}\int_0^1(1-t)^3\partial_{ijkl}\phi(x+t(y-x))dt$. Integrating, all the antisymmetric terms vanish and we obtain

$$\begin{aligned}\overline{\overline{\phi}}(x,r)-\phi(x)=\frac{1}{2}\sum_i\partial_{ii}\phi(x)\frac{1}{V_r^2}\int_{(\mathbb{R}^d)^2}(y-x)_{ii}\mathbb{1}_{\left\{\substack{|x-z|<r\\|y-z|<r}\right\}}dzdy\\ +\sum_{ijkl}\frac{1}{V_r^2}\int_{(\mathbb{R}^d)^2}R_{ijkl}(y)(y-x)_{ijkl}\mathbb{1}_{\left\{\substack{|x-z|<r\\|y-z|<r}\right\}}dzdy.\end{aligned}$$

We begin by calculating the first term before bounding the second one. Note that, by symmetry, the integral of $(y-x)_{ii}$ does not depend on i, so the first sum above can be written as

$$\frac{1}{2}\Delta\phi(x)\frac{1}{dV_r^2}\int_{(\mathbb{R}^d)^2}|y-x|^2\mathbb{1}_{\left\{\substack{|x-z|<r\\|y-z|<r}\right\}}dzdy.$$

By the parallelogram identity, $|y-x|^2=2(|x-z|^2+|y-z|^2)-|2z-(x+y)|^2$. Integrating, we see that

$$\begin{aligned}\frac{1}{V_r^2}\int_{(\mathbb{R}^d)^2}|y-x|^2\mathbb{1}_{\left\{\substack{|x-z|<r\\|y-z|<r}\right\}}dzdy=4\frac{1}{V_r}\int_{B(0,r)}|y|^2\,dy\\ -\frac{1}{V_r^2}\int_{(\mathbb{R}^d)^2}|(2z-y)-x|^2\mathbb{1}_{\left\{\substack{|x-z|<r\\|(2z-y)-z|<r}\right\}}dzdy.\end{aligned}$$

Setting $y'=2z-y$ in the rightmost integral and moving this term to the left-hand side, we obtain

$$\begin{aligned}\frac{1}{V_r^2}\int_{(\mathbb{R}^d)^2}|y-x|^2\mathbb{1}_{\left\{\substack{|x-z|<r\\|y-z|<r}\right\}}dzdy&=\frac{2}{V_r}\int_{B(0,r)}|y|^2\,dy\\ &=\frac{2d}{d+2}r^2.\end{aligned}$$

Replacing this term in the equation above, we can write

$$\left|\overline{\overline{\phi}}(x,r) - \phi(x) - \frac{r^2}{d+2}\Delta\phi(x)\right| \leq \sum_{ijkl} \frac{1}{V_r^2} \int_{(\mathbb{R}^d)^2} |R_{ijkl}(y)|\, |y-x|_{ijkl} \mathbb{1}_{\left\{\substack{|x-z|<r \\ |y-z|<r}\right\}} dzdy.$$

Proceeding exactly as before and writing $|y|_{ijkl} \leq \frac{1}{4}(|y_i|^4 + |y_j|^4 + |y_k|^4 + |y_l|^4)$, one shows that

$$\left\|\overline{\overline{\phi}}(r) - \phi - \frac{r^2}{d+2}\Delta\phi\right\|_q \leq \frac{d^3}{4!} \max_{|\beta|=4} \|\partial_\beta \phi\|_q \sum_i \frac{1}{V_r^2} \int_{(\mathbb{R}^d)^2} |y_i|^4 \mathbb{1}_{\left\{\substack{|-z|<r \\ |y-z|<r}\right\}} dzdy.$$

Note that $\sum_i |y_i|^4 \leq |y|^4$, and by the parallelogram identity, $|y|^4 + |2z-y|^4 \leq 8(|z|^4 + |z-y|^4)$. As before, we can integrate on both sides:

$$\frac{1}{V_r^2} \int_{(\mathbb{R}^d)^2} |y|^4 \mathbb{1}_{\left\{\substack{|-z|<r \\ |y-z|<r}\right\}} dzdy + \frac{1}{V_r^2} \int_{(\mathbb{R}^d)^2} |2z-y|^4 \mathbb{1}_{\left\{\substack{|-z|<r \\ |(2z-y)-z|<r}\right\}} dzdy \leq 16 \frac{1}{V_r} \int_{B(0,r)} |y|^4 \, dy.$$

Hence,

$$\frac{1}{V_r^2} \int_{(\mathbb{R}^d)^2} |y|^4 \mathbb{1}_{\left\{\substack{|-z|<r \\ |y-z|<r}\right\}} dzdy \leq 8 \frac{1}{V_r} \int_{B(0,r)} |y|^4 \, dy = \frac{8d}{d+4} r^4.$$

□

Proposition 2.7.3. *Take $\phi : \mathbb{R}^d \to \mathbb{R}$ to be twice continuously differentiable and suppose that $\|\partial_\beta \phi\|_q < \infty$ for $0 \leq |\beta| \leq 2$ and $q \in \{1, \infty\}$. Then*

i) $\left\|\mathcal{D}^{\alpha,\delta}\phi\right\|_q \lesssim \|\phi\|_q + \max_{|\beta|=2} \|\partial_\beta \phi\|_q$,

ii) $\left\|\mathcal{D}^{\alpha,\delta}\phi - \mathcal{D}^{\alpha}\phi\right\|_q \lesssim \delta^{2-\alpha} \max_{|\beta|=2} \|\partial_\beta \phi\|_q$.

Further if $0 \leq \phi \leq 1$,

iii) $\left\|F^{(\delta)}(\phi) - F(\phi)\right\|_\infty \lesssim \delta^\alpha \left(1 + \max_{|\beta|=1} \|\partial_\beta \phi\|_\infty^2 + \max_{|\beta|=2} \|\partial_\beta \phi\|_\infty\right)$.

Proof of Proposition 2.7.3. From the definition of $\mathcal{D}^{\alpha,\delta}$ and $\Phi^{(\delta)}$ in (2.19), and then changing the order of integration,

$$\begin{aligned}
\mathcal{D}^{\alpha,\delta}\phi(x) &= \int_{\mathbb{R}^d} \int_{\frac{|x-y|}{2} \vee \delta}^{\infty} \frac{V_r(x,y)}{V_r} (\phi(y) - \phi(x)) \frac{dr}{r^{d+\alpha+1}} dy \\
&= \int_\delta^\infty \int_{\mathbb{R}^d} \mathbb{1}_{\{|x-y|<2r\}} \frac{1}{V_r r^d} \int_{\mathbb{R}^d} \mathbb{1}_{\left\{\substack{|z-x|<r \\ |z-y|<r}\right\}} (\phi(y) - \phi(x)) dzdy \frac{dr}{r^{\alpha+1}} \\
&= V_1 \int_\delta^\infty \left(\overline{\overline{\phi}}(x,r) - \phi(x)\right) \frac{dr}{r^{\alpha+1}}.
\end{aligned}$$

The last line follows by noting that $\mathbb{1}_{\{|x-y|<2r\}} \mathbb{1}_{\{|x-z|<r,\, |y-z|<r\}} = \mathbb{1}_{\{|x-z|<r,\, |y-z|<r\}}$, and then changing the order of integration again to integrate first with respect to y and then with respect to z. Assume that $\delta < 1$ (otherwise simply ignore the first term below). Then for $q \in \{1, \infty\}$, by Proposition 2.7.1.iii,

$$\left\|\mathcal{D}^{\alpha,\delta}\phi\right\|_q \leq V_1 \int_\delta^1 \left\|\overline{\overline{\phi}}(r) - \phi\right\|_q \frac{dr}{r^{\alpha+1}} + 2V_1 \|\phi\|_q \int_1^\infty \frac{dr}{r^{\alpha+1}}.$$

By the triangular inequality and Proposition 2.7.1.iii,

$$\begin{aligned}\left\|\overline{\overline{\phi}}-\phi\right\|_q &\le \left\|\overline{\overline{\phi}-\phi}\right\|_q + \left\|\overline{\phi}-\phi\right\|_q \\ &\le 2\left\|\overline{\phi}-\phi\right\|_q.\end{aligned}$$

Hence by Proposition 2.7.2,

$$\begin{aligned}\left\|\mathcal{D}^{\alpha,\delta}\phi\right\|_q &\lesssim \max_{|\beta|=2}\left\|\partial_\beta\phi\right\|_q \int_\delta^1 r^2\frac{dr}{r^{\alpha+1}} + \left\|\phi\right\|_q\int_1^\infty \frac{dr}{r^{\alpha+1}} \\ &\lesssim \max_{|\beta|=2}\left\|\partial_\beta\phi\right\|_q + \left\|\phi\right\|_q.\end{aligned}$$

Likewise, we have

$$\mathcal{D}^\alpha\phi(x) - \mathcal{D}^{\alpha,\delta}\phi(x) = V_1\int_0^\delta\left(\overline{\overline{\phi}}(x,r)-\phi(x)\right)\frac{dr}{r^{\alpha+1}}.$$

By Proposition 2.7.2, we then write

$$\begin{aligned}\left\|\mathcal{D}^\alpha\phi - \mathcal{D}^{\alpha,\delta}\phi\right\|_q &\le V_1\int_0^\delta\left\|\overline{\overline{\phi}}(r)-\phi\right\|_q\frac{dr}{r^{\alpha+1}} \\ &\lesssim \max_{|\beta|=2}\left\|\partial_\beta\phi\right\|_q\int_0^\delta r^2\frac{dr}{r^{\alpha+1}} \\ &\lesssim \delta^{2-\alpha}\max_{|\beta|=2}\left\|\partial_\beta\phi\right\|_q.\end{aligned}$$

The third statement is a rewording of the first one in a slightly different setting. Indeed by (2.18),

$$F^{(\delta)}(\phi)(x) - F(\phi(x)) = \alpha\delta^\alpha\int_\delta^\infty\left(\overline{F(\overline{\phi})}(x,r) - F(\phi(x))\right)\frac{dr}{r^{\alpha+1}}.$$

Hence as in the proof of (i)

$$\left\|F^{(\delta)}(\phi)-F(\phi)\right\|_\infty \lesssim \delta^\alpha\left(\left\|F(\phi)\right\|_\infty + \max_{|\beta|=2}\left\|\partial_\beta F(\overline{\phi})\right\|_\infty + \left\|F'\right\|_\infty\max_{|\beta|=2}\left\|\partial_\beta\phi\right\|_\infty\right).$$

The last term appears because there is an average inside the function F. The result then follows from the fact that $\partial_{ij}F(\phi) = \partial_{ij}\phi\,F'(\phi) + \partial_i\phi\,\partial_j\phi\,F''(\phi)$. □

2.8 The centering term

2.8.1 The Brownian case

Proof of Lemma 2.2.1. Fix $N \ge 1$ and let $r = r_N$. Define an operator $S : L^\infty([0,T]\times\mathbb{R}^d) \to L^\infty([0,T]\times\mathbb{R}^d)$ by

$$S(g)(t,x) = G_t^{(r)} * w_0(x) - \int_0^t G_{t-s}^{(r)} * \overline{F(\overline{g(s)})}(x,r)ds.$$

Define the norm $\|g\|_{[0,T]} = \sup_{t\in[0,T]} \sup_{x\in\mathbb{R}^d} |g(t,x)|$; then S is Lipschitz on $L^\infty([0,T]\times\mathbb{R}^d)$ with respect to this norm, since (as $\left\|G^{(r)}_{t-s} * \phi\right\|_\infty \leq \|\phi\|_\infty$ and by Proposition 2.7.1.iii)

$$\|S(f) - S(g)\|_{|0,T]} \leq T \|F'\|_\infty \|f-g\|_{[0,T]}. \tag{2.128}$$

Let us choose for now $T > 0$ small enough that $k = T\|F'\|_\infty < 1$. The results can be extended to arbitrarily large time intervals by iterating the argument on $[T,2T]$, $[2T,3T]$ and so on. The operator S is then a contraction in $L^\infty([0,T]\times\mathbb{R}^d)$ and admits a unique fixed point in this space. This fixed point is precisely f^N (see (2.71)). Define a sequence $(g_n)_{n\geq 0}$ of functions in $L^\infty([0,T]\times\mathbb{R}^d)$ by

$$\begin{cases} g_{n+1} = S(g_n), \\ g_0(t,x) = w_0(x). \end{cases}$$

Note that since w_0 admits spatial derivatives of order up to four, so does g_n for each n. A Picard iteration argument then yields the convergence of g_n to f^N in $L^\infty([0,T]\times\mathbb{R}^d)$. More precisely,

$$\left\|g_n - f^N\right\|_{[0,T]} \leq k \left\|g_{n-1} - f^N\right\|_{[0,T]} \leq \ldots \leq k^n \left\|w_0 - f^N\right\|_{[0,T]}. \tag{2.129}$$

Fix $1 \leq i \leq d$ and set, for g, h in $L^\infty([0,T],\mathbb{R}^d)$,

$$S^1(g,h)(t,x) = G^{(r)}_t * \partial_i w_0(x) - \int_0^t G^{(r)}_{t-s} * \overline{F'(\overline{g(s)})\overline{h(s)}}(x) ds.$$

Then by Proposition 2.7.1.v, $\partial_i g_{n+1} = S^1(g_n, \partial_i g_n)$. In addition, for h_1, h_2, h_3 in $L^\infty([0,T]\times\mathbb{R}^d)$,

$$\left\|S^1(h_1,h_2) - S^1(h_1,h_3)\right\|_{[0,T]} \leq T\|F'\|_\infty \|h_2 - h_3\|_{[0,T]}, \tag{2.130}$$

$$\left\|S^1(h_1,h_2) - S^1(h_3,h_2)\right\|_{[0,T]} \leq T\|F''\|_\infty \|h_2\|_{[0,T]} \|h_1 - h_3\|_{[0,T]}. \tag{2.131}$$

Hence $S^1(g,\cdot)$ is a contraction in $L^\infty([0,T]\times\mathbb{R}^d)$ for any g in $L^\infty([0,T]\times\mathbb{R}^d)$. Let us call $\tilde{g}$ the unique fixed point of $S^1(f^N,\cdot)$ in this space. We shall now show that $\partial_i f^N$ exists and is equal to $\tilde{g} \in L^\infty([0,T]\times\mathbb{R}^d)$. Adapting the argument of the Picard iteration and using the inequalities (2.130) and (2.131) above, we write for $n \geq 1$,

$$\begin{aligned} \|\partial_i g_n - \tilde{g}\|_{[0,T]} &= \left\|S^1(g_{n-1}, \partial_i g_{n-1}) - S^1(f^N, \tilde{g})\right\|_{[0,T]} \\ &\leq T\|F'\|_\infty \|\partial_i g_{n-1} - \tilde{g}\|_{[0,T]} + T\|F''\|_\infty \|\tilde{g}\|_{[0,T]} \left\|g_{n-1} - f^N\right\|_{[0,T]} \\ &\leq k\|\partial_i g_{n-1} - \tilde{g}\|_{[0,T]} + T\|F''\|_\infty \|\tilde{g}\|_{[0,T]} k^{n-1} \left\|w_0 - f^N\right\|_{[0,T]}, \end{aligned}$$

where $k = T\|F'\|_\infty < 1$ and the last term is bounded using (2.129). Iterating yields

$$\|\partial_i g_n - \tilde{g}\|_{[0,T]} \leq k^n \|\partial_i w_0 - \tilde{g}\|_{[0,T]} + nk^{n-1} T \|F''\|_\infty \|\tilde{g}\|_{[0,T]} \left\|w_0 - f^N\right\|_{[0,T]}.$$

Hence $\partial_i g_n$ converges to $\tilde{g}$ uniformly on $[0,T]\times\mathbb{R}^d$ (recall that we assumed $\partial_i w_0 \in L^\infty(\mathbb{R}^d)$). Since we already showed in (2.129) that g_n converges uniformly to f^N, this implies that $\partial_i f^N = \tilde{g} \in L^\infty([0,T]\times\mathbb{R}^d)$. The proof for higher order derivatives of f^N is similar and we omit the details. □

Proof of Proposition 2.4.5. Recall the following expression for f^N from (2.71),

$$f_t^N(x) = G_t^{(r)} * w_0(x) - \int_0^t G_{t-s}^{(r)} * \overline{F(\overline{f_s^N})}(x)ds. \tag{2.132}$$

Since $\left\|G_t^{(r)} * \phi\right\|_\infty \leq \|\phi\|_\infty$, it follows that $\left\|f_t^N\right\|_\infty \leq \|w_0\|_\infty + T\|F\|_\infty$ for $t \leq T$. We can now prove the second part of the statement by induction on $|\beta|$. (Recall that β is a multi-index $(\beta_1, \ldots, \beta_d)$ in $\mathbb{N}_0^d$ and that $|\beta| = \beta_1 + \ldots + \beta_d$.) Suppose that for every β' with $0 \leq |\beta'| < k \leq 4$, there exists a constant $K_{\beta'} < \infty$ independent of N such that $\sup_{0\leq t\leq T} \left\|\partial_{\beta'} f_t^N\right\|_\infty \leq K_{\beta'}$; take β such that $|\beta| = k$. (From now on we omit the superscript N in the induction proof.) Note that for some constants $C_{\alpha_1,\ldots,\alpha_i} \in \mathbb{N}$,

$$\partial_\beta F(f) = \sum_{i\geq 1} F^{(i)}(f)\left(\sum_{\alpha_1+\ldots+\alpha_i=\beta} C_{\alpha_1,\ldots,\alpha_i}\partial_{\alpha_1} f \ldots \partial_{\alpha_i} f\right)$$

where the second sum is over all possible multisets with i elements of non-zero multi-indices $(\alpha_1, \ldots, \alpha_i)$ in $\mathbb{N}^d \setminus \{(0, \ldots, 0)\}$ such that $\alpha_1 + \ldots + \alpha_i = \beta$ (computing the sum coordinate by coordinate). Also, w_0 is assumed to have uniformly bounded derivatives of up to the fourth order. Using Proposition 2.7.1 (iv and v), we can differentiate on both sides of (2.132) and we obtain

$$\begin{aligned}\partial_\beta f_t(x) = {} & G_t^{(r)} * \partial_\beta w_0(x) \\ & - \int_0^t G_{t-s}^{(r)} * \left(C_\beta \overline{F'(\overline{f_s})\overline{\partial_\beta f_s}} + \sum_{\substack{\alpha_1+\ldots+\alpha_i=\beta \\ i\geq 2}} C_{\alpha_1,\ldots,\alpha_i} \overline{F^{(i)}(\overline{f_s})\overline{\partial_{\alpha_1} f_s} \ldots \overline{\partial_{\alpha_i} f_s}}\right)(x)ds.\end{aligned}$$

The sum is uniformly bounded by a constant K by the induction hypothesis, and so, using the fact that $\left\|G_t^{(r)} * \phi\right\|_\infty \leq \|\phi\|_\infty$,

$$\|\partial_\beta f_t\|_\infty \leq \|\partial_\beta w_0\|_\infty + TK + C_\beta \|F'\|_\infty \int_0^t \|\partial_\beta f_s\|_\infty ds.$$

The function $t \mapsto \|\partial_\beta f_t\|_\infty$ is bounded on $[0, T]$ by Lemma 2.2.1. We can therefore apply Gronwall's inequality to conclude

$$\|\partial_\beta f_t\|_\infty \leq \left(\|\partial_\beta w_0\|_\infty + TK\right) e^{C_\beta \|F'\| T},$$

where the right hand side is independent of both $t \in [0, T]$ and $N \geq 1$. We can now prove the first statement using Gronwall's inequality again, together with Proposition 2.7.2 and the first part of the proof.

Recall that G_t denotes the fundamental solution to the heat equation. Recalling that we set the constants uV_R, $2R^2/(d+2)$ and s to 1, equations (2.15) and (2.47) can be written as

$$f_t^N(x) = G_t * w_0(x) + \int_0^t G_{t-s} * \left(\mathcal{L} f_s^N - \frac{1}{2}\Delta f_s^N - \overline{F(\overline{f_s^N})}\right)(x)ds,$$

and

$$f_t(x) = G_t * w_0(x) + \int_0^t G_{t-s} * F(f_s)ds.$$

Recall the definition of $\mathcal{L}^{(r)}$ in (2.14); by Proposition 2.7.2,

$$\left\| \mathcal{L} f_s^N - \frac{1}{2}\Delta f_s^N \right\|_\infty \leq \frac{d^3(d+2)}{6} r_N^2 \max_{|\beta|=4} \left\| \partial_\beta f_s^N \right\|_\infty \lesssim r_N^2,$$

since $\max_{|\beta|=4} \left\| \partial_\beta f_s^N \right\|_\infty$ is uniformly bounded from the previous argument. Also by Proposition 2.7.2,

$$\left\| \overline{F(\overline{f_s^N})} - F(f_s^N) \right\|_\infty \leq \frac{d}{2} r_N^2 \left(\max_{|\beta|=2} \left\| \partial_\beta F(\overline{f_s^N}) \right\|_\infty + \|F'\|_\infty \max_{|\beta|=2} \left\| \partial_\beta f_s^N \right\|_\infty \right) \lesssim r_N^2.$$

(The term within brackets is uniformly bounded from the first part of the proof.) Finally, we also have

$$\left\| F(f_s^N) - F(f_s) \right\|_\infty \leq \|F'\|_\infty \left\| f_s^N - f_s \right\|_\infty .$$

Hence, using the fact that $\|G_t * \phi\|_\infty \leq \|\phi\|_\infty$, there exists a constant $C > 0$ such that, for $t \in [0, T]$,

$$\left\| f_t^N - f_t \right\|_\infty \leq C r_N^2 + \|F'\|_\infty \int_0^t \left\| f_s^N - f_s \right\|_\infty ds.$$

Applying Gronwall's inequality (the function $t \mapsto \left\| f_t^N - f_t \right\|_\infty$ is bounded on $[0, T]$ by Lemma 2.2.1),

$$\left\| f_t^N - f_t \right\|_\infty \leq C e^{\|F'\| T} r_N^2.$$

□

2.8.2 The stable case

Proof of Lemma 2.2.2. Lemma 2.2.2 is proved exactly as Lemma 2.2.1 in the Brownian case. The corresponding operator $S : L^\infty([0, T] \times \mathbb{R}^d) \to L^\infty([0, T] \times \mathbb{R}^d)$ is given by

$$S(g)(t, x) = \mathcal{G}_t^{(\alpha, \delta_N)} * w_0(x) - \int_0^t \mathcal{G}_{t-s}^{(\alpha, \delta_N)} * F^{(\delta_N)}(f_s^N)(x) ds.$$

It satisfies the same contraction property as (2.128), yielding the existence of a solution to (2.22) in $L^\infty([0, T] \times \mathbb{R}^d)$. The argument for the existence of spatial derivatives in $L^\infty([0, T] \times \mathbb{R}^d)$ is the same as in the Brownian case, with S^1 given by

$$S^1(g, h)(t, x) = \mathcal{G}_t^{(\alpha, \delta_N)} * \partial_i w_0(x) - \alpha \int_0^t \int_1^\infty \mathcal{G}_{t-s}^{(\alpha, \delta_N)} * \overline{F'(\overline{g(s)})\overline{h(s)}}(x, \delta_N r) \frac{dr}{r^{1+\alpha}} ds.$$

□

Proof of Proposition 2.5.6. The proof of the convergence of the centering term in the stable case goes along the same lines as in the Brownian case of Proposition 2.4.5. Differentiating (2.95) yields

(dropping superscripts N)

$$\partial_\beta f_t(x) = \mathcal{G}_t^{(\alpha,\delta)} * \partial_\beta w_0(x) - \alpha \int_0^t \int_1^\infty \mathcal{G}_{t-s}^{(\alpha,\delta)} * \Bigg(C_\beta \overline{F'(\overline{f_s})\overline{\partial_\beta f_s}}(\delta r) + \sum_{\substack{\alpha_1+\ldots+\alpha_k=\beta \\ k\geq 2}} C_{\alpha_1,\ldots,\alpha_k} \overline{F^{(k)}(\overline{f_s})\overline{\partial_{\alpha_1} f_s}\ldots\overline{\partial_{\alpha_k} f_s}}(\delta r) \Bigg)(x) \frac{dr}{r^{\alpha+1}} ds.$$

One can then proceed by induction as previously to show

$$\|\partial_\beta f_t\|_\infty \lesssim \|\partial_\beta w_0\|_\infty + T + \|F'\|_\infty \int_0^t \|\partial_\beta f_s\|_\infty \, ds,$$

and Gronwall's inequality (using Lemma 2.2.2) yields the second part of the statement. For the first part, the proof is identical to that in the Brownian case, one simply has to replace the operators $\frac{1}{2}\Delta$ and $\mathcal{L}^{(r)}$ by $\mathcal{D}^\alpha$ and $\mathcal{D}^{\alpha,\delta}$, respectively, and likewise replace $\overline{F(\overline{f_t})}$ by $F^{(\delta)}(f_t)$. Proposition 2.7.3 then yields the correct estimates on the corresponding error terms. □

2.9 Time dependent test functions

2.9.1 The Brownian case

Proof of Lemma 2.4.1. In the spirit of the proof of Lemma 2.2.1, we characterize φ^N as the fixed point of a contraction in $L^\infty(\mathbb{R}^d \times \{(s,t) : 0 \leq s \leq t \leq T\})$. By the definition of φ^N in (2.66),

$$\varphi^N(x,s,t) = G_{t-s}^{(r_N)} * \phi(x) - \int_s^t G_{u-s}^{(r_N)} * \overline{F'(\overline{f_u^N})\overline{\varphi^N(u,t)}}(x) du. \tag{2.133}$$

In other words, φ^N is a fixed point of the following operator,

$$S(g)(x,s,t) = G_{t-s}^{(r_N)} * \phi(x) - \int_s^t G_{u-s}^{(r_N)} * \overline{F'(\overline{f_u^N})\overline{g(u,t)}}(x) du.$$

For $q \in [1,\infty]$, define the norm $\|g\|_{q,[0,T]} = \sup_{0\leq s\leq t\leq T} \|g(s,t)\|_q$; then since $G_{u-s}^{(r)}$ is a contraction in L^q and by Proposition 2.7.1.iii,

$$\|S(h) - S(g)\|_{q,[0,T]} \leq T \|F'\|_\infty \|h-g\|_{q,[0,T]}$$

so for T small enough, S is a contraction. Note that the space of (equivalence classes of) measurable functions $g : \{(s,t) : 0 \leq s \leq t \leq T\} \to L^q(\mathbb{R}^d)$ such that $\|g\|_{q,[0,T]} < \infty$ is a Bochner space (and therefore a Banach space). Hence for each $q \in [1,\infty]$, there exists a unique fixed point of S which is uniformly bounded in $L^q(\mathbb{R}^d)$, obtained as the limit of the sequence

$$\begin{cases} g_{n+1} = S(g_n), \\ g_0(x,s,t) = \phi(x). \end{cases}$$

Since this sequence does not depend on q, the fixed point is the same for all q. This fixed point is φ^N. Proceeding as in the proof of Lemma 2.2.1, one shows that the spatial derivatives of g_n (of order up to four) converge uniformly to some function which is uniformly bounded in $L^q(\mathbb{R}^d)$ for all $q \in [1, \infty]$. As a result φ^N admits spatial derivatives of order up to four which are uniformly bounded in $L^q(\mathbb{R}^d)$ for $q \in [1, \infty]$. □

Proof of Lemma 2.4.2. The proof of Lemma 2.4.2 is similar in spirit to that of Proposition 2.4.5. We start by proving the bound on the derivatives of φ^N. Using the fact that $G_t^{(r)}$ is a contraction in L^q, we have, using (2.133), for $q = 1, 2$,

$$\begin{aligned}\left\|\varphi^N(s,t)\right\|_q^q &\leq 2^{q-1}\left\|\phi\right\|_q^q + (2(t-s))^{q-1}\int_s^t \left\|\overline{F'(\overline{f_u^N})\overline{\varphi^N(u,t)}}\right\|_q^q du \\ &\leq 2^{q-1}\left\|\phi\right\|_q^q + (2(t-s))^{q-1}\left\|F'\right\|_\infty^q \int_s^t \left\|\varphi^N(u,t)\right\|_q^q du,\end{aligned}$$

by Proposition 2.7.1.iii. In addition, by Lemma 2.4.1, the function $s \mapsto \left\|\varphi^N(s,t)\right\|_q$ is bounded on $[0,t]$. By Gronwall's inequality, we conclude that

$$\left\|\varphi^N(s,t)\right\|_q \leq 2^{(q-1)/q}\left\|\phi\right\|_q e^{\frac{2^{q-1}}{q}T^q\left\|F'\right\|^q}.$$

Thus the statement holds for $\beta = 0$. We can then proceed by induction on $|\beta|$ as in the proof of Proposition 2.4.5 to show that the same holds for every $0 \leq |\beta| \leq 4$ (making use of the fact that by Proposition 2.4.5, f^N has uniformly bounded derivatives). We omit the details.

We are left with proving the convergence estimate for φ^N which is again a Gronwall estimate. As in the proof of Proposition 2.4.5, write (2.66) and (2.68) as

$$\varphi^N(x,s,t) = G_{t-s} * \phi(x) + \int_s^t G_{u-s} * \left(\mathcal{L}^{(r_N)}\varphi^N(u,t) - \frac{1}{2}\Delta\varphi^N(u,t) - \overline{F'(\overline{f_u^N})\overline{\varphi^N(u,t)}}\right)(x)du, \tag{2.134}$$

and

$$\varphi(x,s,t) = G_{t-s} * \phi(x) - \int_s^t G_{u-s} * \left(F'(f_u)\varphi(u,t)\right)(x)du. \tag{2.135}$$

By Proposition 2.7.2 and the bound on the spatial derivatives of φ^N,

$$\left\|\mathcal{L}^{(r_N)}\varphi^N(u,t) - \frac{1}{2}\Delta\varphi^N(u,t)\right\|_q \lesssim r_N^2.$$

Still by Proposition 2.7.2, (omitting superscripts N and time variables)

$$\begin{aligned}\left\|\overline{F'(\overline{f})\overline{\varphi}} - F'(f)\varphi\right\|_q \leq \frac{d}{2}r_N^2 \Bigg(&\max_{|\beta|=2}\left\|\partial_\beta(F'(\overline{f})\overline{\varphi})\right\|_q + \left\|F'\right\|_\infty \max_{|\beta|=2}\left\|\partial_\beta\varphi\right\|_q \\ &+ \left\|\varphi\right\|_q \left\|F''\right\|_\infty \max_{|\beta|=2}\left\|\partial_\beta f\right\|_\infty\Bigg).\end{aligned}$$

The last term inside the brackets is uniformly bounded by Proposition 2.4.5 and the second to last is bounded as a consequence of the first part of the proof. Also, $\partial_{ij}(F'(\overline{f})\overline{\varphi})$ is dominated by a linear combination of (averages of) derivatives of both f and φ. The latter are bounded in L^q while the

former are bounded in L^∞, hence the first term within the brackets is also uniformly bounded. To sum up,

$$\left\|\overline{F'(\overline{f_u^N})\overline{\varphi^N(u,t)}} - F'(f_u^N)\varphi^N(u,t)\right\|_q \lesssim r_N^2. \tag{2.136}$$

Finally, by Proposition 2.4.5,

$$\left\|F'(f_u^N) - F'(f_u)\right\|_q \lesssim r_N^2.$$

Hence, subtracting (2.135) from (2.134) and using Jensen's inequality as above with the L^q-contraction property of G_t, we have, for $t \in [0,T]$,

$$\left\|\varphi^N(s,t) - \varphi(s,t)\right\|_q^q \lesssim r_N^{2q} + \int_s^t \left\|\varphi^N(u,t) - \varphi(u,t)\right\|_q^q du.$$

Also, by Lemma 2.4.1, the function $s \mapsto \left\|\varphi^N(s,t) - \varphi(s,t)\right\|_q$ is bounded on $[0,t]$. We conclude with Gronwall's inequality, yielding the first statement of Lemma 2.4.2. □

Proof of Lemma 2.4.8. We can assume that $t' > t \geq s$ (if $t' \geq s \geq t$, then $\varphi^N(s,t) = \phi = \varphi^N(s,s)$ and the problem reduces to bounding $\varphi^N(s,t') - \varphi^N(s,s)$). Using (2.133), we write

$$\begin{aligned}\varphi^N(x,s,t') - \varphi^N(x,s,t) = G^{(r_N)}_{t'-s} * \phi(x) - G^{(r_N)}_{t-s} * \phi(x) - \int_t^{t'} G^{(r_N)}_{u-s} * \overline{F'(\overline{f_u^N})\overline{\varphi^N(u,t')}}(x)du \\ - \int_s^t G^{(r_N)}_{u-s} * \left(\overline{F'(\overline{f_u^N})(\overline{\varphi^N(u,t')} - \overline{\varphi^N(u,t)})}\right)(x)du.\end{aligned}$$

From the way we extended φ^N in (2.77), we see that for $u \geq t'$, $\varphi^N(u,t') - \varphi^N(u,t) = 0$ and for $t \leq u \leq t'$, $\varphi^N(u,t') - \varphi^N(u,t) = \varphi^N(u,t') - \phi$, so (omitting superscripts N)

$$\begin{aligned}\varphi(x,s,t') - \varphi(x,s,t) = G^{(r)}_{t'-s} * \phi(x) - G^{(r)}_{t-s} * \phi(x) - \int_t^{t'} G^{(r)}_{u-s} * \overline{F'(\overline{f_u})\overline{\phi}}(x)du \\ - \int_s^T G^{(r)}_{u-s} * \left(\overline{F'(\overline{f_u})(\overline{\varphi(u,t')} - \overline{\varphi(u,t)})}\right)(x)du.\end{aligned}$$

Again, we use the L^q-contraction property of $G_t^{(r)}$ to write

$$\begin{aligned}\|\varphi(s,t') - \varphi(s,t)\|_q^q \leq 3^{q-1}\left\|G^{(r)}_{t'-s} * \phi - G^{(r)}_{t-s} * \phi\right\|_q^q + 3^{q-1}|t'-t|^q \|F'\|_\infty^q \|\phi\|_q^q \\ + (3(T-s))^{q-1}\|F'\|_\infty^q \int_s^T \|\varphi(u,t') - \varphi(u,t)\|_q^q du.\end{aligned} \tag{2.137}$$

We need a bound on the first term; recalling the definition of $G^{(r)}$ in Subsection 2.4.2, we have

$$G^{(r)}_{t'-s} * \phi(x) - G^{(r)}_{t-s} * \phi(x) = \int_t^{t'} G^{(r)}_{u-s} * \mathcal{L}^{(r)}\phi(x)du.$$

By Jensen's inequality,

$$\left\|G^{(r)}_{t'-s} * \phi - G^{(r)}_{t-s} * \phi\right\|_q^q \leq |t'-t|^{q-1} \int_t^{t'} \left\|\mathcal{L}^{(r)}\phi\right\|_q^q du \lesssim |t'-t|^q,$$

by Proposition 2.7.2. Hence, returning to (2.137),

$$\|\varphi(s,t') - \varphi(s,t)\|_q^q \lesssim |t'-t|^q + \int_s^T \|\varphi(u,t') - \varphi(u,t)\|_q^q \, du.$$

Noting that from Lemma 2.4.1, we know that $s \mapsto \|\varphi(s,t') - \varphi(s,t)\|_q$ is bounded on $[0,T]$, Gronwall's inequality yields the result. □

Proof of Lemma 2.4.9. By the definition of $G^{(r)}$,

$$G^{(r_N)}_{t-s} * \phi(x) = \phi(x) + \int_s^t G^{(r_N)}_{u-s} * \mathcal{L}^{(r_N)}\phi(x) du.$$

Hence

$$\left\| \sup_{t\in[s,T]} G^{(r_N)}_{t-s} * \phi \right\|_1 \leq \|\phi\|_1 + (T-s)\left\|\mathcal{L}^{(r_N)}\phi\right\|_1 \leq \|\phi\|_1 + T\frac{d(d+2)}{2} \max_{|\beta|=2} \|\partial_\beta \phi\|_1 \tag{2.138}$$

(we have used Proposition 2.7.2.i to bound $\left\|\mathcal{L}^{(r_N)}\phi\right\|_1$ independently of N). Recall from (2.133) that

$$\varphi^N(x,s,t) = G^{(r_N)}_{t-s} * \phi(x) - \int_s^t G^{(r_N)}_{u-s} * \overline{F'(\overline{f^N_u})\overline{\varphi^N(u,t)}}(x) du.$$

Within the second integral, $u \leq t$, so we can write $\left|\varphi^N(u,t)\right| \leq \sup_{t'\in[u,T]} \left|\varphi^N(u,t')\right|$. Thus (omitting superscripts and subscripts) by Proposition 2.7.1.i,

$$\sup_{t\in[s,T]} |\varphi(x,s,t)| \leq \sup_{t\in[s,T]} |G^{(r)}_{t-s} * \phi(x)| + \|F'\|_\infty \int_s^T \overline{\overline{G^{(r)}_{u-s}}} * \sup_{t\in[u,T]} |\varphi(u,t)|\,(x) du.$$

Integrating with respect to the variable $x \in \mathbb{R}^d$ yields

$$\left\| \sup_{t\in[s,T]} |\varphi(s,t)| \right\|_1 \leq \|\phi\|_1 + T\frac{d(d+2)}{2} \max_{|\beta|=2} \|\partial_\beta\phi\|_1 + \|F'\|_\infty \int_s^T \left\| \sup_{t\in[u,T]} |\varphi(u,t)| \right\|_1 du.$$

Consider the space X of functions $g : \mathbb{R}^d \times [0,T]^2 \to \mathbb{R}$ such that $g(x,s,t) = g(x,t,t)$ for $s \geq t$ and the norm

$$\sup_{s\in[0,T]} \left\| \sup_{t\in[0,T]} |g(s,t)| \right\|_1$$

is finite. (As a closed subspace of a Bochner space, X is complete with respect to this norm.) Looking at the proof of Lemma 2.4.1, extend S to an operator on X by setting $S(g)(x,s,t) = \phi(x)$ for $s \geq t$. Then $S : X \to X$ (using (2.138)) and by the same argument as in the proof of Lemma 2.4.1, for T

sufficiently small, S is a contraction on X. As a result we obtain that $s \mapsto \left\| \sup_{t\in[s,T]} |\varphi(s,t)| \right\|_1$ is bounded on $[0,T]$. Hence, by Gronwall's inequality,

$$\left\| \sup_{t\in[s,T]} |\varphi(s,t)| \right\|_1 \leq \left(\|\phi\|_1 + T\frac{d(d+2)}{2} \max_{|\beta|=2} \|\partial_\beta \phi\|_1 \right) e^{\|F'\|(T-s)}.$$

□

2.9.2 The stable case

Proof of Lemma 2.5.2. By the definition of φ^N in (2.89),

$$\varphi^N(x,s,t) = \mathcal{G}_{t-s}^{(\alpha,\delta_N)} * \phi(x) - \alpha \int_s^t \int_1^\infty \mathcal{G}_{u-s}^{(\alpha,\delta_N)} * \overline{F'(\overline{f_u^N})\overline{\varphi^N(u,t)}}(\delta_N r)(x) \frac{dr}{r^{\alpha+1}} du. \tag{2.139}$$

Note that since we are only considering $q \in \{1,\infty\}$, we have $\left\| \int_1^\infty f(\cdot,r)dr \right\|_q \leq \int_1^\infty \|f(\cdot,r)\|_q \, dr$. Hence the bound on the derivatives of φ^N can be proved following the same argument as in the proof of Lemma 2.4.2 in the Brownian case, using Lemma 2.5.1 in place of Lemma 2.4.1. By the definition of φ,

$$\varphi(x,s,t) = \mathcal{G}_{t-s}^{(\alpha)} * \phi(x) - \int_s^t \mathcal{G}_{u-s}^{(\alpha)} * \left(F'(f_u)\varphi(u,t)\right)(x) du.$$

By Proposition 2.7.3 and by the bound on the spatial derivatives of φ^N,

$$\left\| \mathcal{D}^{\alpha,\delta_N}\varphi^N(u,t) - \mathcal{D}^\alpha \varphi^N(u,t) \right\|_q \lesssim \delta_N^{2-\alpha}.$$

Using (2.136) (which is still true in this case by the bound on the derivatives of φ^N), we have

$$\begin{aligned} \int_1^\infty \left\| \overline{F'(\overline{f_u^N})\overline{\varphi^N(u,t)}}(\delta_N r) - F'(f_u^N)\varphi^N(u,t) \right\|_q \frac{dr}{r^{\alpha+1}} &\lesssim \delta_N^2 \int_1^{\delta^{-1}} r^2 \frac{dr}{r^{\alpha+1}} + \int_{\delta_N^{-1}}^\infty \frac{dr}{r^{\alpha+1}} \\ &\lesssim \delta_N^\alpha. \end{aligned}$$

Finally, by Proposition 2.5.6,

$$\left\| F'(f_u^N) - F'(f_u) \right\|_q \lesssim \delta_N^{\alpha\wedge(2-\alpha)}.$$

As a result, by the same argument as in the proof of Lemma 2.4.2, by Gronwall's inequality,

$$\left\| \varphi^N(s,t) - \varphi(s,t) \right\|_q \lesssim \delta_N^{\alpha\wedge(2-\alpha)}.$$

□

Proof of Lemma 2.5.10. The argument for the continuity estimate is the same as in the proof of Lemma 2.4.8, using Proposition 2.7.3. For the second bound, we use the same argument as in Lemma 2.4.9, again using Proposition 2.7.3. □

Proof of Lemma 2.5.4. Splitting the integral with respect to z_2, we have

$$\int_{(\mathbb{R}^d)^2} |f(z_1)|\,|g(z_2)|\,|z_1-z_2|^{-\alpha}dz_1dz_2 \le \|g\|_\infty \int_{\mathbb{R}^d} |f(z_1)| \int_{B(z_1,1)} |z_1-z_2|^{-\alpha}dz_2dz_1 + \int_{\mathbb{R}^d} |f(z_1)| \int_{\mathbb{R}^d\setminus B(z_1,1)} |g(z_2)|dz_2dz_1$$

But $\int_{B(0,1)} |y|^{-\alpha}\,dy = \frac{dV_1}{d-\alpha}$ and we have :

$$\int_{(\mathbb{R}^d)^2} |f(z_1)|\,|g(z_2)|\,|z_1-z_2|^{-\alpha}dz_1dz_2 \le \|g\|_\infty \frac{dV_1}{d-\alpha}\int_{\mathbb{R}^d} |f(z_1)|\,dz_1 + \|g\|_1 \int_{\mathbb{R}^d} |f(z_1)|\,dz_1.$$

□

2.10 Estimates for drift load proofs

Proof of Lemma 2.6.6. For all $t > 0$, $\xi_t^{(r)}$ can be written as $\xi_t^{(r)} = \sum_{k=1}^{N_t} Y_k$, where $(N_t)_{t\ge 0}$ is a Poisson process with intensity $\frac{(d+2)}{2r^2}$ and $(Y_k)_{k\ge 1}$ is a sequence of independent and identically distributed random variables with density $\psi(y) = \frac{V_r(0,y)}{V_r^2}$. As a result, the law of $\xi_t^{(r)}$ can be written

$$G_t^{(r)}(dx) = e^{-\frac{d+2}{2r^2}t}\delta_0(dx) + e^{-\frac{d+2}{2r^2}t}\sum_{n\ge 1}\frac{\left(\frac{d+2}{2r^2}t\right)^n}{n!}\psi^{*n}(x)dx.$$

Since ψ is continuous on $\mathbb{R}^d$, so is ψ^{*n} for any $n \ge 1$. In addition, $\psi(y)$ is decreasing as a function of $|y|$, and $\phi * \psi(x) = \overline{\overline{\phi}}(x,r)$ so, by induction it follows that $\psi^{*n}(y)$ is also decreasing as a function of $|y|$. Since the above sum converges uniformly, we can conclude that $g_t^{(r)}$ is continuous on $\mathbb{R}^d$ and that $g_t^{(r)}(y)$ is rotation invariant and is a decreasing function of $|y|$. □

Proof of Lemma 2.6.7. By some elementary algebra,

$$\begin{aligned}\phi(y)^2 - \phi(x)^2 - 2\phi(x)(\phi(y)-\phi(x)) + 2\frac{2r^2}{d+2}\phi(x)(\phi(y)-\phi(x))g(y) &= \left(\phi(y)-\phi(x)+\frac{2r^2}{d+2}\phi(x)g(y)\right)^2 - \left(\frac{2r^2}{d+2}\right)^2\phi(x)^2g(y)^2\\ &\ge -\left(\frac{2r^2}{d+2}\right)^2\phi(x)^2,\end{aligned}$$

since $g(y)^2 \le 1$. Averaging the above inequality in y twice around x and multiplying by $\frac{d+2}{2r^2}$ yields

$$\mathcal{L}^{(r)}\phi^2(x) - 2\phi(x)\mathcal{L}^{(r)}\phi(x) + 2\phi(x)\overline{\overline{\phi\, g}}(x,r) - 2\phi(x)^2\overline{\overline{g}}(x,r) \ge -\frac{2r^2}{d+2}\phi(x)^2.$$

The first result then follows from the fact that $\gamma \le g$. For the second inequality, set $a = \phi(y)$, $\epsilon = \frac{2r^2}{d+2}g(y)$ and $b = (1-\epsilon)^{1/3}\phi(x)$; then

$$\phi(y)^4 - \phi(x)^4 - 4\left(1-\frac{2r^2}{d+2}g(y)\right)\phi(x)^3(\phi(y)-\phi(x)) = a^4 - b^4 - 4b^3(a-b) + b^4 - \phi(x)^4 - 4b^3(b-\phi(x)).$$

By convexity of the function $x \mapsto x^4$, $a^4 - b^4 - 4b^3(a-b) \geq 0$, so the above expression is greater than

$$\phi(x)^4 \left[(1-\epsilon)^{4/3} - 1 - 4(1-\epsilon)((1-\epsilon)^{1/3} - 1)\right] \underset{\epsilon \to 0}{\sim} -\frac{2}{3}\phi(x)^4\epsilon^2.$$

Hence there exists c such that, for r small enough,

$$\phi(y)^4 - \phi(x)^4 - 4\left(1 - \frac{2r^2}{d+2}g(y)\right)\phi(x)^3(\phi(y) - \phi(x)) \geq -4c\,r^4\phi(x)^4.$$

Averaging in y twice around x as above yields the second statement. □

Proof of Lemma 2.6.8. We define the following :

$$\mathcal{H}(x,u,t) = e^{-\alpha(t-u)}G^{(r)}_{t-u} * h_u(x).$$

Differentiating with respect to u yields

$$\begin{aligned}\frac{\partial \mathcal{H}}{\partial u}(x,u,t) &= e^{-\alpha(t-u)}G^{(r)}_{t-u} * \left(\partial_u h_u - \mathcal{L}h_u + \alpha h_u\right)(x) \\ &\leq e^{-\alpha(t-u)}G^{(r)}_{t-u} * g_u(x). \end{aligned} \tag{2.140}$$

Integrating (2.140) over $u \in [s,t]$, we have

$$h_t(x) \leq e^{-\alpha(t-s)}G^{(r)}_{t-s} * h_s(x) + \int_s^t e^{-\alpha(t-u)}G^{(r)}_{t-u} * g_u(x)du. \tag{2.141}$$

By Jensen's inequality,

$$\begin{aligned}\left(\int_s^t e^{-\alpha(t-u)}G^{(r)}_{t-u} * g_u(x)du\right)^q &\leq \left(\int_s^t e^{-\alpha(t-u)}du\right)^{q-1}\int_s^t e^{-\alpha(t-u)}\left(G^{(r)}_{t-u} * g_u(x)\right)^q du \\ &\leq \frac{1}{\alpha^{q-1}}\int_s^t e^{-\alpha(t-u)}G^{(r)}_{t-u} * g^q_u(x)du.\end{aligned}$$

The result follows by taking $\|\cdot\|_q$ norms on each side of (2.141). □

Chapter 3

Dispersal heterogeneity in the spatial Λ-Fleming-Viot process

Introduction

Landscape genetics studies the influence of geographical features of the environment on evolutionary processes and on the genetic composition of populations. Habitat fragmentation and ecological interfaces play a significant role in this field [MH13]. Scientists strive to detect, map and quantify the long term effects on genetic diversity of spatial heterogeneities by observing the genetic patterns that they have produced through evolution [Sla87]. For example, genetic differentiation between two subpopulations separated by a physical obstacle can be used to measure the reduction in gene flow caused by the obstacle [Su+03; Ril+06; Gay+07].

Our focus in this chapter is the special case in which individuals spread their offspring farther from themselves in some parts of space than in others. By comparing the genomes of individuals and the frequencies of different genetic types at different locations, one tries to infer the strength of dispersal in these regions and to measure the effect of the interface.

Simple models for the evolution of gene frequencies are then required which can be fitted to field data with reasonable computational power. That is why mathematicians in the field of population genetics establish large scale approximations of microscopic models which take into account the interaction between geographical features and evolutionary forces [Mal48; KW64; BDE02].

Nagylaki [Nag76] studied the effect of a discontinuity in the migration rate in the linear stepping stone model. He considered colonies located at the points $k/\sqrt{n}$, $k \in \mathbb{Z}$, which evolve in discrete generations spanning $1/n$ units of time. At each generation, adjacent colonies to the left of the origin exchange a proportion $m/2$ of migrants while adjacent colonies to the right exchange a proportion $v^2 m/2$, as depicted in Figure 3.1.

Letting $n \to \infty$ and considering that the number of individuals in each colony is so large that genetic drift (*i.e.* fluctuations due to random sampling of individuals at each generation) can be ignored, Nagylaki showed that the proportion of individuals of a given type at location $x \in \mathbb{R}$ at time

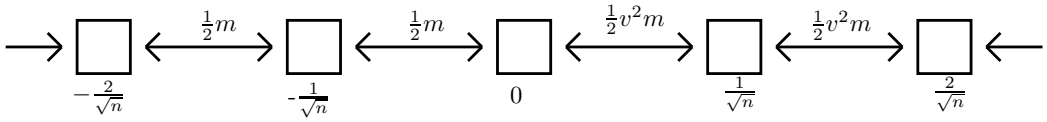

Figure 3.1: Discrete model with a discontinuity in the migration rate

$t \geq 0$, denoted by $p(t,x)$, is well approximated by the solution to the following equation

$$\begin{cases} \dfrac{\partial p}{\partial t}(t,x) = \dfrac{m}{2}\dfrac{\partial^2 p}{\partial x^2}(t,x) & \text{if } x < 0 \\ \dfrac{\partial p}{\partial t}(t,x) = \dfrac{v^2 m}{2}\dfrac{\partial^2 p}{\partial x^2}(t,x) & \text{if } x > 0 \end{cases}$$

and, for $t > 0$,

$$p(t,0^+) = p(t,0^-), \qquad \frac{\partial p}{\partial x}(t,0^-) = v^2 \frac{\partial p}{\partial x}(t,0^+).$$

In words, allele frequencies must be continuous at zero but their first spatial derivative has a discontinuity which is given as a simple function of the ratio of the migration rates on each side of the habitat (see Figure 3.3). He extended this result [NB88] to the probability of identity by descent, *i.e.* the probability that two uniformly sampled individuals have inherited the same allele from a common ancestor without mutation as a function of the distance between the sampling locations. Nagylaki found similar conditions for the first derivative of the probability of identity as for the allele frequencies. Along with Ayati and Dupont [ADN99], he further investigated the qualitative properties of the probability of identity in this setting and provided numerical approximations.

In parallel to these developments, a diffusion process has been introduced [IM63; Wal78; HS81] and used to study diffusion in physical systems presenting an interface between different media [App+11]. The so-called *skew Brownian motion* with parameter $\alpha \in [0,1]$ can be described as an $\mathbb{R}$-valued stochastic process which performs Brownian excursions from the origin, on the positive half line with probability α and on the negative half line with probability $1-\alpha$. See [Lej06] for a review of the definition and properties of skew Brownian motion.

In this chapter, we study the genealogy of a sample of individuals in the presence of heterogeneous dispersal. This genealogy is described by a system of ancestral lineages which at time t correspond to the positions of the ancestors of the sample t generations in the past. We find that, in the diffusion limit, those ancestral lineages follow skew Brownian motions with different diffusion coefficients on each side of the interface (Proposition 3.3.3 below). The genealogy of a sample of individuals is then given by a system of skew Brownian motions which coalesce upon meeting in one dimension but never coalesce in higher dimensions (Theorem 3.2). As a consequence, allele frequencies follow a deterministic partial differential equation in dimensions two and higher while in one dimension, patches of different types form and evolve randomly (Theorem 3.1). Our method allows for more general assumptions on the microscopic model than [Nag76; Nag88] (*e.g.* continuous spatial structure and non-nearest neighbour migration).

We use the spatial Λ-Fleming-Viot process framework introduced in [BEV10a] and [Eth08] to model the evolution of allele frequencies in a continuous space (see [BEV13a] for a review on this

process). In this model, reproduction events occur according to a Poisson point process on $\mathbb{R}_+ \times \mathbb{R}^d$ which specifies their time and location. During these reproduction events, a proportion u - called the *impact parameter* - of individuals in a ball of radius r is replaced by the offspring of a uniformly sampled individual in this ball. To model heterogeneous dispersal, we assume that the radius of the reproduction event depends on the halfspace in which its centre falls, as illustrated in Figure 3.2. We study the large scale behaviour of the spatial Λ-Fleming-Viot process (SLFV) under a diffusive rescaling similar to the one considered in the homogeneous setting in [BEV13b]. In particular, the impact parameter is kept constant as we rescale space and time.

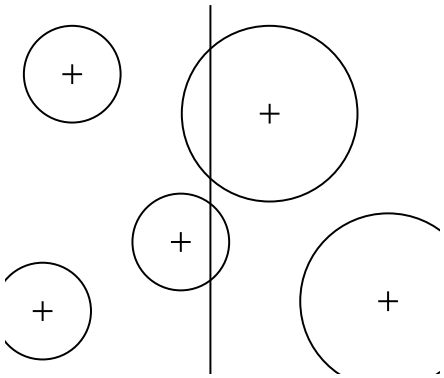

Figure 3.2: Size of reproduction events
The size of the region affected by a reproduction event depends on the halfspace in which its centre falls ($x_1 > 0$ or $x_1 < 0$).

Our results and their proofs are similar in spirit to those in [BEV13b]. We use the fact that the SLFV has a dual in the form of a system of coalescing particles moving in $\mathbb{R}^d$ (interpreted as the locations in the past of the ancestors of a random sample of individuals). We show (Theorem 3.2) that the rescaled dual converges to a system of skew Brownian motions which evolve independently of each other until they meet, and then coalesce instantaneously upon meeting. In particular, when $d \geq 2$, the particles never meet and evolve independently of each other. Our approach improves on [BEV13b] as our proof covers any configuration where ancetral lineages converge to Markov processes with continuous paths.

As a consequence, we obtain a scaling limit of the process describing the evolution of allele frequencies across space (Theorem 3.1). The limit is deterministic as soon as $d \geq 2$ and solves a heat equation on each halfspace. The fact that ancestral lineages follow skew Brownian motions translates into a discontinuity of the first spatial derivative along the normal of the interface, in agreement with Nagylaki's result. When $d = 1$, each site is occupied by only one type of individuals at any positive time, and the boundaries between patches of different types evolve according to a system of annihilating skew Brownian motions.

The proof of the convergence of the motion of lineages to skew Brownian motion is adapted from the work of A. Iksanov and A. Pilipenko [IP16], where skew Brownian motion is obtained as a scaling limit of a Markov chain on $\mathbb{Z}$ which behaves like simple random walk outside a bounded region around the origin. The difficulty in proving convergence to skew Brownian motion comes from the fact that martingale problem characterizations of the limiting process are ill suited to this setting. (In particular, scale functions of the limiting process do not turn the random walk into a martingale.) Following [IP16], we circumvent this by studying the positive and negative parts of the process separately, and then linking the two by their respective local times at the origin. This method turns out to be readily applicable to more general migration patterns than originally studied in [Nag76], as we show here by dealing with a continuous spatial structure.

The chapter is laid out as follows. We define the SLFV with heterogeneous dispersal in Section 3.1 and we state our main result (Theorem 3.1) in Section 3.2. Section 3.3 gives a description of the dual of the SLFV and states its convergence under the diffusive rescaling (Theorem 3.2). The latter is

proved in Section 3.4 and implies Theorem 3.1. Finally, the convergence of the motion of an ancestral lineage to skew Brownian motion is proved in Section 3.5, following the arguments of [IP16].

3.1 Definition of the model

Consider a model where individuals are scattered in a continuous space of dimension d and can be of two types, denoted by 0 or 1. We suppose that the density of individuals is constant in space. The population is represented by a random function $\{w(t,x), t \geq 0, x \in \mathbb{R}^d\}$, where $w(t,x) \in [0,1]$ is interpreted as the proportion of type 1 individuals at location x at time t. Define the two halfspaces $\mathbb{H}^+$, $\mathbb{H}^-$ by

$$\mathbb{H}^{\pm} = \left\{x \in \mathbb{R}^d : \pm x_1 > 0\right\}.$$

Take $u \in (0,1]$ and $0 < r_- \leq r_+ < +\infty$. We denote the volume of the ball of radius r in $\mathbb{R}^d$ by V_r. The SLFV with heterogeneous dispersal is defined as follows.

Definition 3.1.1 (SLFV with heterogeneous dispersal)**.** *Let $\Pi^{\pm}$ be a Poisson point process on $\mathbb{H}^{\pm} \times \mathbb{R}_+$ with intensity $\frac{1}{V_{r_\pm}} dxdt$. For each point (x,t) in $\Pi^{\pm}$, a reproduction event takes place in $B(x, r_\pm)$ at time t:*

1) Pick a location y uniformly at random in $B(x, r_\pm)$ and sample a type $k \in \{0,1\}$ from the types present at y (i.e. $k = 1$ with probability $\frac{1}{V_{r_\pm}} \int_{B(x,r_\pm)} w(t_-, y)dy$).

2) Update $w(t,z)$ for $z \in B(x, r_\pm)$ as follows:

$$w(t,z) = (1-u)w(t_-, z) + u\mathbb{1}_{\{k=1\}}.$$

In other words, a proportion u of individuals in the ball of centre x and radius $r_\pm$ dies and is replaced by the offspring of an individual sampled uniformly from this ball.

Remark. *The factor $\frac{1}{V_{r_\pm}}$ in the rate of the Poisson point process ensures that the mean lifetime of individuals is the same in both halfspaces (far enough from the interface).*

Theorem 4.2 in [BEV10a] can be adapted without difficulty to show that there exists a unique càdlàg Markov process $(w(t,\cdot))_{t \geq 0}$ satisfying this definition and taking values in the quotient space Ξ of Lebesgue-measurable maps from $\mathbb{R}^d$ to $[0,1]$ that are identified when they coincide up to a Lebesgue-null set. This space can be identified with (a subset of) the space of measures on $\mathbb{R}^d$ that are absolutely continuous with respect to Lebesgue measure. It is endowed with the following metric d which induces the topology of vague convergence of measures on $\mathbb{R}^d$. Let $(f_n)_{n \geq 1}$ be a separating family of uniformly bounded and compactly supported real-valued functions on $\mathbb{R}^d$, then

$$d(w, w') = \sum_{n=1}^{\infty} \frac{1}{2^n} \left|\langle w, f_n\rangle - \langle w', f_n\rangle\right|, \qquad w, w' \in \Xi.$$

3.2 Large scale behaviour of the SLFV with heterogeneous dispersal

Fix $w_0 : \mathbb{R}^d \to [0,1]$. For $n \geq 1$, set $w^n(t,x) = w(nt, \sqrt{n}x)$ and assume that $w^n(0,x) = w_0(x)$ for all $n \geq 1$. For $\beta \in (-1,1)$, let $\mathcal{D}^\beta$ denote the set of all continuous functions $\phi : \mathbb{R}^d \to \mathbb{R}$, twice continuously differentiable on each halfspace $\mathbb{H}^\pm$, such that

$$(1+\beta)\left.\frac{\partial \phi}{\partial x_1}\right|_{x_1=0^+} = (1-\beta)\left.\frac{\partial \phi}{\partial x_1}\right|_{x_1=0^-}.$$

Theorem 3.1. *As $n \to \infty$, the sequence of Ξ-valued processes $\{w^n(t,\cdot), t \geq 0\}$ converges in the sense of finite dimensional distributions in the vague topology to a process $\{p(t,\cdot), t \geq 0\}$. In dimension one, $p(t,x)$ is a Bernoulli random variable with parameter $\rho(t,x)$ and the correlations between the values of $p(t,\cdot)$ at distinct sites are non trivial and are given in* (3.8) *(see also Figure 3.4). In dimensions two and higher, $p(t,x)$ is deterministic and equals $\rho(t,x)$. In both cases, there exists $\beta \in (0,1)$ such that $t \mapsto \rho(t,\cdot)$ is the $\mathcal{D}^\beta$ valued solution to the following equation*

$$\begin{cases} \dfrac{\partial \rho}{\partial t}(t,x) = \dfrac{u r_\pm^2}{d+2}\Delta\rho(t,x) & \text{if } x \in \mathbb{H}^\pm, \\ \rho(0,x) = w_0(x) & x \in \mathbb{R}^d. \end{cases} \tag{3.1}$$

Finding solutions to (3.1) in $\mathcal{D}^\beta$ can be reduced to finding classical solutions to the heat equation with discontinuous coefficients by a change of variables as shown in [Nag76]. Existence and uniqueness of the solution in $\mathcal{D}^\beta$ to (3.1) was also proved in [Por79a] and [Por79b], see also Proposition 1 in [Lej06]. We prove Theorem 3.1 by studying the dual of the SLFV with heterogeneous dispersal.

The fact that the solution to (3.1) has to be found in $\mathcal{D}^\beta$ with $\beta \geq 0$ agrees with the findings of Nagylaki [Nag76] (equations 8 and 9). This transmission condition reflects the fact that individuals living near the frontier between the two halfspaces are more likely to have ancestors coming from $\mathbb{H}^+$ than from $\mathbb{H}^-$ (recall that we take $r_- \leq r_+$), see Figure 3.3.

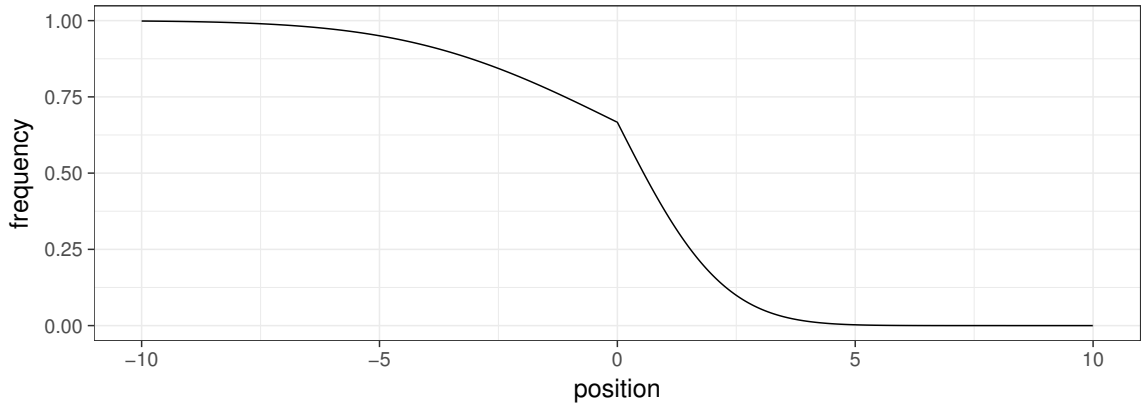

Figure 3.3: Diffusion of an allele with heterogeneous dispersal
Graphical representation of $x \mapsto \rho(t,x)$ started from a Heavyside initial condition $\mathbb{1}_{\{x<0\}}$ at time $t = 12$ with parameters: $\sigma_+ = 0.5$, $\sigma_- = 1$, $\beta = -0.6$. Note the discontinuity in the first spatial derivative at $x = 0$.

As already noted by Nagylaki [Nag76], β depends on the microscopic model in a rather intricate way. We give an expression for β in (3.26) when the microscopic model is the SLFV. This dependence on the choice of the model is a potential issue when trying to infer demographic parameters from genetic data. Inferring β as an independent parameter would reduce the power of such an inference scheme, so one would like to choose a particular model and make β a function of the other parameters in the model. However it isn't clear how one should choose among the great variety of possible microscopic models.

3.3 Duality

3.3.1 The dual of the SLFV with heterogeneous dispersal

We now define a system of coalescing particles whose displacements are driven by the same Poisson point process of reproduction events as the SLFV. The particles at time t describe the positions of the set of ancestors at time $-t$ of a sample of individuals alive at time 0. Since the Poisson point processes $\Pi^{\pm}$ are reversible with respect to time, the reproduction events which took place in the past have the same distribution as those which occur forwards in time.

Definition 3.3.1 (Dual of the SLFV with heterogeneous dispersal)**.** *Let $\Pi^{\pm}$ be Poisson point processes in $\mathbb{H}^{\pm} \times \mathbb{R}_+$ with intensity $\frac{1}{V_{r_\pm}} dxdt$. Let $(\mathcal{A}_t)_{t\geq 0}$ be a system of finitely many particles whose dynamics are as follows. For each point (x,t) in $\Pi^{\pm}$, a reproduction event takes place in $B(x, r_\pm)$ at time t:*

1) Pick a location y uniformly at random in $B(x, r_\pm)$.

2) Each particle sitting inside $B(x, r_\pm)$ at time t_- is marked with probability u, independently of each other.

3) All marked particles coalesce and move to y.

We denote the number of particles present at time t by N_t and their spatial locations by $\xi_t^1, \dots, \xi_t^{N_t}$, so that $\mathcal{A}_t = \{\xi_t^1, \dots, \xi_t^{N_t}\}$.

Let $B^{\pm}(x,r)$ denote the intersection of $B(x,r)$ and $\mathbb{H}^{\pm}$. The motion of one particle is a Markov process on $\mathbb{R}^d$ with infinitesimal generator

$$\mathcal{L}f(x) = u \int_{\mathbb{R}^d} \Phi(x,y)(f(y) - f(x))dy \tag{3.2}$$

with

$$\Phi(x,y) = \frac{|B^+(x,r_+) \cap B^+(y,r_+)|}{V_{r_+}^2} + \frac{|B^-(x,r_-) \cap B^-(y,r_-)|}{V_{r_-}^2}. \tag{3.3}$$

This is seen by noting that a particle located at x finds itself in the region of a reproduction event of $\Pi^{\pm}$ at rate

$$\frac{|B^{\pm}(x,r_\pm)|}{V_{r_\pm}}.$$

It is further affected by such an event with probability u and moves to a location y chosen uniformly in the ball of radius $r_\pm$ affected by the event. See [BEV13a] (paragraph 3.5) for a more detailed

justification in the homogeneous case. The law of $(\mathcal{A}_t)_{t\geq 0}$ started from j lineages at locations $\underline{x} = (x_1, \ldots, x_j)$ is denoted by $\mathbb{P}_{\underline{x}}(\cdot)$.

Let us now give the (weak) duality relation between $(w_t)_{t\geq 0}$ and $(\mathcal{A}_t)_{t\geq 0}$. Let $C_c(\mathbb{R}^d)$ be the space of compactly supported real valued functions on $\mathbb{R}^d$. For $\psi : (\mathbb{R}^d)^j \to \mathbb{R}_+$ in $C_c((\mathbb{R}^d)^j)$ and $w \in \Xi$, set

$$I(w, \psi) = \int_{(\mathbb{R}^d)^j} \prod_{i=1}^{j} w(x_i)\psi(x_1, \ldots, x_j)dx_1 \ldots dx_j.$$

Also set

$$\langle w, \mathcal{A}_t \rangle = \prod_{i=1}^{N_t} w(\xi_t^i).$$

Then, for any $j \in \mathbb{N}$, for $\psi \in C_c((\mathbb{R}^d)^j)$, [BEV10a]

$$\mathbb{E}_{w_0}\left[I(w_t, \psi)\right] = \int_{(\mathbb{R}^d)^j} \mathbb{E}_{\underline{x}}\left[\langle w_0, \mathcal{A}_t\rangle\right] \psi(\underline{x})d\underline{x}. \tag{3.4}$$

Since the linear span of functions of the form $I(\cdot, \psi)$ and constant functions is dense in $C(\Xi)$ (Lemma 4.1 in [BEV10a]), one can prove Theorem 3.1 by showing that, for any $0 \leq t_1 < \ldots < t_k$ and $\psi_1, \ldots, \psi_k$ in $C_c((\mathbb{R}^d)^j)$,

$$\lim_{n\to\infty} \mathbb{E}\left[\prod_{i=1}^{k} I(w_{t_i}^n, \psi_i)\right] = \mathbb{E}\left[\prod_{i=1}^{k} I(p_{t_i}, \psi_i)\right]. \tag{3.5}$$

We shall do this using the duality relation (3.4) above. For $n \geq 1$, define the rescaled dual process $(\mathcal{A}_t^n)_{t\geq 0}$ by

$$\mathbb{E}_{\underline{x}}\left[f(\mathcal{A}_t^n)\right] = \mathbb{E}_{\sqrt{n}\underline{x}}\left[f\left(\tfrac{1}{\sqrt{n}}\xi_{nt}^1, \ldots, \tfrac{1}{\sqrt{n}}\xi_{nt}^{N_{nt}}\right)\right].$$

Then $(\mathcal{A}_t^n)_{t\geq 0}$ is dual to $(w_t^n)_{t\geq 0}$ in the sense that

$$\mathbb{E}_{w_0}\left[I(w_t^n, \psi)\right] = \int_{(\mathbb{R}^d)^j} \mathbb{E}_{\underline{x}}\left[\langle w_0, \mathcal{A}_t^n\rangle\right] \psi(\underline{x})d\underline{x}.$$

In Section 3.4, we prove the convergence of $(\mathcal{A}_t^n)_{t\geq 0}$ to a system of coalescing skew Brownian motions. Note that in dimensions two and higher, skew Brownian motions never meet and the dual of the SLFV with heterogeneous dispersal thus converges to a system of independent skew Brownian motions. This is the reason why the SLFV converges to a deterministic process when $d \geq 2$ in Theorem 3.1.

3.3.2 Skew Brownian motion

In [HS81] (see also [Wal78], [LG84b] and [Lej06]) it is shown that for $\beta \in [-1, 1]$, there exists a unique solution to the equation

$$X_t = B_t + \beta L_t^0(X),$$

where B is standard Brownian motion and $L_t^0(X)$ is the local time at 0 of X. This process is called skew Brownian motion with parameter $\alpha = \frac{\beta+1}{2}$. (For $\beta = 1$, $(X_t)_{t\geq 0}$ is reflected Brownian motion.) This result can be extended to the d-dimensional case where the first coordinate of the process follows

skew Brownian motion.

Proposition 3.3.2. *Let $B = \left(B_t^1, \dots, B_t^d\right)_{t\geq 0}$ be standard (d dimensional) Brownian motion. Let $\sigma : \mathbb{R}^d \to (0,\infty)$ be defined by $\sigma^2(x) = \sigma_\pm^2 \mathbb{1}_{\{x\in\mathbb{H}^\pm\}}$ with $\sigma_\pm^2 > 0$ and take $x_0 = (x_0^1, \dots, x_0^d) \in \mathbb{R}^d$. Then, for $\beta \in [-1,1]$, there exists a unique $\mathbb{R}^d$-valued Markov process $(X_t)_{t\geq 0}$ satisfying*

$$\begin{aligned} X_t^1 &= x_0^1 + \int_0^t \sigma(X_s) dB_s^1 + \beta L_t^0(X^1) \\ X_t^i &= x_0^i + \int_0^t \sigma(X_s) dB_s^i \qquad\qquad \text{for } 2 \leq i \leq d. \end{aligned} \tag{3.6}$$

Furthermore, the law of $(X_t)_{t\geq 0}$ is the unique solution to the (hence well posed) martingale problem associated with the generator L, defined on the domain $\mathcal{D}^\beta$ by

$$L\phi(x) = \frac{1}{2}\sigma^2(x)\Delta\phi(x), \qquad \forall \phi \in \mathcal{D}^\beta.$$

This result is proved in [Lej06] (Proposition 10) in the case $d = 1$ and $\sigma_+ = \sigma_-$. The extension to higher dimensions is straightforward and the case $\sigma_+ \neq \sigma_-$ can be treated with the help of [BP87]. In [Por79a], [Por79b], it is proved that L generates a Feller semigroup. Part of the work in showing Theorem 3.1 is the proof that the motion of particles in $\mathcal{A}^n$ converges to a solution to (3.6), as stated in the following Proposition. Its proof is given in Section 3.5.

Proposition 3.3.3 (Convergence to skew Brownian motion)**.** *Let $(\xi_t)_{t\geq 0}$ be an $\mathbb{R}^d$-valued Markov process with infinitesimal generator $\mathcal{L}$ given in* (3.2)*. For $n \geq 1$, set $\xi_t^n = \frac{1}{\sqrt{n}}\xi_{nt}$ and suppose ξ_0^n is deterministic and converges to $x_0 \in \mathbb{R}$ as $n \to \infty$. Fix $T > 0$. Then, as $n \to \infty$, $(X_t^n)_{t\in[0,T]}$ converges in distribution in the Skorokhod space $\mathrm{D}\left([0,T], \mathbb{R}^d\right)$ to $(X_t)_{t\in[0,T]}$, a solution to* (3.6) *with $\sigma_\pm^2 = u\frac{2r_\pm^2}{d+2}$, and $\beta \in (0,1)$.*

The parameter β is given as a (complicated) function of the law of $(\xi_t)_{t\geq 0}$ in (3.26). Note however that $\beta \geq 0$ as soon as $r_+ \geq r_-$.

3.3.3 Large scale behaviour of the dual process

Let $(\mathcal{A}_t^\infty)_{t\geq 0}$ be a system of particles moving in $\mathbb{R}^d$ according to independent skew Brownian motions (*i.e.* solutions to (3.6)) with $\sigma_\pm^2 = u\frac{2r_\pm^2}{d+2}$ and with the same parameter β which coalesce instantaneously upon meeting. In particular, in dimension 2 and higher, the particles never meet and $(\mathcal{A}_t^\infty)_{t\geq 0}$ is a system of independent skew Brownian motions. We denote the locations of the particles at time t by $\{X_t^1, \dots, X_t^{N_t}\}$.

From [Eva97], we know that there exists a Ξ-valued process $\{p(t,x), t \geq 0, x \in \mathbb{R}^d\}$ which is dual to $\mathcal{A}^\infty$ in the sense that, for $\psi \in C_c((\mathbb{R}^d)^j)$,

$$\mathbb{E}_{w_0}\left[I(p_t,\psi)\right] = \int_{(\mathbb{R}^d)^j} \mathbb{E}_{\underline{x}}\left[\langle \mathcal{A}_t^\infty, w_0\rangle\right] \psi(\underline{x}) d\underline{x}. \tag{3.7}$$

Furthermore, by Lemma 3.2 in [BEV13b], in dimension one, $p(t,x)$ is a Bernoulli random variable with parameter $\rho(t,x) = \mathbb{E}_x\left[w_0(Z_t)\right]$ while in dimensions two and higher, $p(t,x)$ is deterministic and

equals $\rho(t,x)$. The fact that ρ can be characterized as the solution to (3.1) is a direct consequence of operator semigroup theory (see [EK86, Proposition 1.1.5] and recall that L generates a Feller semigroup). In [BEV13b], it is shown that the following theorem implies (3.5) and hence Theorem 3.1 (see their proof of Theorem 1.1).

Theorem 3.2. *As $n \to \infty$, $(\mathcal{A}_t^n)_{t\geq 0}$ converges in the sense of finite dimensional distributions to $(\mathcal{A}_t^\infty)_{t\geq 0}$.*

Moreover, for $k \in \mathbb{N}$ and $0 \leq t_1 < \ldots < t_k$, suppose that we start $\mathcal{A}^n$ with j_0 particles at locations $\underline{x}_0$, let the process evolve until time t_1, add j_1 lineages at locations $\underline{x}_1$, let the process evolve until time t_2 and so on. Call the resulting process $\hat{\mathcal{A}}^n$ and define $\hat{\mathcal{A}}^\infty$ analogously. Then for any $t \geq 0$, $\hat{\mathcal{A}}_t^n$ converges in distribution to $\hat{\mathcal{A}}_t^\infty$ as $n \to \infty$.

From (3.7), we obtain that for Lebesgue almost every $(x_1, \ldots, x_j) \in (\mathbb{R}^d)^j$,

$$\mathbb{E}_{w_0}\left[\prod_{i=1}^{j} p(t,x_i)\right] = \mathbb{E}_{x_1,\ldots,x_j}\left[\prod_{i=1}^{N_t} w_0(X_t^i)\right] \tag{3.8}$$

yielding the correlations between the values of $p(t,\cdot)$ at different sites.

In dimensions two and higher, lineages never coalesce and evolve independently of each other. As a result, one can show (see [BEV13b])

$$\mathbb{E}_{w_0}\left[p(t,x)^2\right] = \left(\mathbb{E}_x\left[w_0(X_t)\right]\right)^2 = \left(\mathbb{E}_{w_0}\left[p(t,x)\right]\right)^2,$$

which is only possible if p is deterministic.

In dimension one, since lineages coalesce when they meet, at any positive time each location is occupied by only one type of individuals. Small patches of type 1 and type 0 individuals then form, whose borders can be shown to follow anihilating skew Brownian motions. Neighbouring patches of the same type thus merge whenever their borders meet, as illustrated in Figure 3.4.

Remark. *Lineages coalesce instantaneously upon meeting because the impact parameter u (which should be interpreted as the inverse of the effective population size) is kept constant as we rescale time and space. Other scalings would result in different limiting behaviours. If u is of order $1/\sqrt{n}$, then we expect that, in the limit, lineages coalesce when they accumulate a local time together equal to an independent exponential random variable, as in [DR08]. The evolution of allele frequencies is then described by a stochastic partial differential equation in one spatial dimension (but remains deterministic in higher dimensions as skew Brownian motions never meet), as in [EVY14]. Moreover, if $u = o\left(1/\sqrt{n}\right)$, lineages never coalesce in the limit, even in one dimension, and the evolution of allele frequencies is deterministic (and equal to ρ).*

3.4 Proof of Theorem 3.2

Proposition 3.3.3 gives the convergence of the law of the motion of each particle in $\mathcal{A}^n$ to skew Brownian motion. To show Theorem 3.2, we thus need to control the coalescence of the particles. The following proposition helps fulfill this goal.

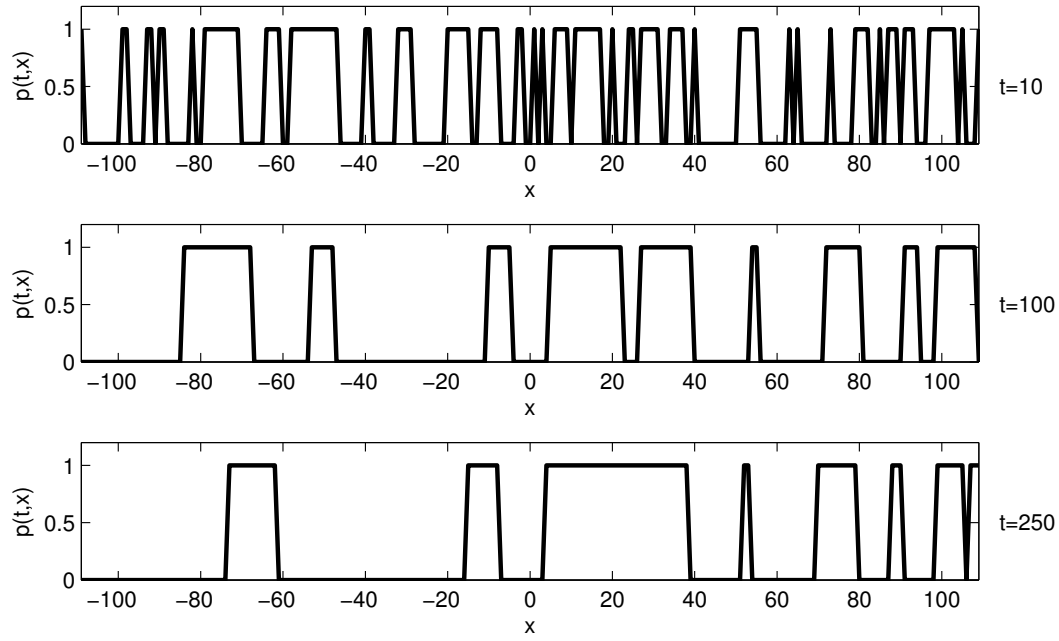

Figure 3.4: The limiting process in dimension one

Numerical simulation of $(p(t,\cdot))_{t\geq 0}$ in a one dimensional space of length 220 with $\sigma_-^2 = 0.2$, $\sigma_+^2 = 0.06$ and $\beta = 7/13$, started from the initial condition $w_0(x) \equiv 0.5$, shown at time $t = 10$, $t = 100$ and $t = 250$. Notice how the number of patches decreases with time as their interfaces meet and annihilate each other. Patches on the right are smaller and more numerous than patches on the left because diffusion is stronger on the left than on the right of the origin.

Proposition 3.4.1. *Let $O \subset \mathbb{R}^d$ be an open set and let $F \subset \mathbb{R}^d$ be a closed set. Suppose that a sequence of functions (or processes) $f_n : \mathbb{R}_+ \to \mathbb{R}^d$ converges uniformly on every compact interval to a continuous function $f : \mathbb{R}_+ \to \mathbb{R}^d$. Define $T_O^n = \inf\{t \geq 0 : f_n(t) \in O\}$ and T_F^n, T_O and T_F accordingly. Then*

$$T_F \leq \liminf_{n\to\infty} T_F^n, \qquad \limsup_{n\to\infty} T_O^n \leq T_O.$$

This Proposition is proved in Section 3.6. An immediate consequence is that if a sequence of processes $\{(X_t^n)_{t\geq 0}, n \geq 1\}$ converges in distribution in $\mathrm{D}\left([0,T], \mathbb{R}^d\right)$ to a continuous process $(X_t)_{t\geq 0}$, and if $T_O = T_F$ a.s. when F is the closure of O (defining T_F, T_O, T_F^n and T_O^n as the hitting times of these sets by the processes $(X_t)_{t\geq 0}$ and $(X_t^n)_{t\geq 0}$ respectively), then, by the Skorokhod representation theorem, both T_O^n and T_F^n converge in distribution to $T_O = T_F$.

Proof of Theorem 3.2. We prove the first part of the result when starting from two particles; the proof is easily extended to a larger sample (see [BEV13b]). The two particles in $\mathcal{A}^n$ evolve independently of each other until they come within a distance $2r_+/\sqrt{n}$ of each other (since $r_- \leq r_+$). Let us then

define T_n as the first time at which the two particles come close to each other in the rescaled setting

$$T_n = \inf\left\{t \geq 0 : \left|\xi_t^{n,1} - \xi_t^{n,2}\right| \leq \frac{2r_+}{\sqrt{n}}\right\}. \tag{3.9}$$

When $d \geq 2$, we show that $\mathbb{P}_{x_1,x_2}(T_n \leq t) \to 0$ as $n \to \infty$ for all $t > 0$. For $\varepsilon > 0$, define

$$T_n^\varepsilon = \inf\left\{t \geq 0 : \left|\xi_t^{n,1} - \xi_t^{n,2}\right| \leq 2r_+\varepsilon\right\}.$$

This is the hitting time of the closed set $\{(x,y) : |x-y| \leq 2r_+\varepsilon\}$ by the process $(\xi_t^{n,1}, \xi_t^{n,2})_{t\geq 0}$. Since $\xi^{n,1}$ and $\xi^{n,2}$ are independent up to time T_n and, for n large enough, $T_n \geq T_n^\varepsilon$, by Proposition 3.3.3 and Proposition 3.4.1, T_n^ε converges in distribution to T^ε, defined as the hitting time of $\{(x,y) : |x-y| \leq 2r_+\varepsilon\}$ by two independent solutions to (3.6) started from x_1 and x_2. As a result, since $T_n \geq T_n^\varepsilon$ a.s. for n large enough,

$$\limsup_{n\to\infty} \mathbb{P}_{x_1,x_2}(T_n \leq t) \leq \mathbb{P}_{x_1,x_2}(T^\varepsilon \leq t).$$

The right-hand-side vanishes as $\varepsilon \downarrow 0$ when $d \geq 2$, yielding the result in this case.

We treat the case $d = 1$ in two steps. First we prove that the trajectory of the two particles up to time T_n converges in distribution to the motion of two independent skew Brownian motions up to their meeting time. Then we argue that the coalescence happens soon enough once the two particles are close to each other that the delay between T_n and the coalescence time (denoted by T_n^c) vanishes in the limit.

By the Skorokhod representation theorem and by Proposition 3.3.3, there exist sequences of processes $(\tilde\xi_t^{n,1}, \tilde\xi_t^{n,2})_{t\geq 0}$ and $(\tilde X_t^1, \tilde X_t^2)_{t\geq 0}$ defined on some probability space such that

i) $(\tilde\xi_t^{n,1})_{t\geq 0}$ and $(\tilde\xi_t^{n,2})_{t\geq 0}$ are independent Markov processes with infinitesimal generator $\mathcal{L}$,

ii) $(\tilde X_t^1)_{t\geq 0}$ and $(\tilde X_t^2)_{t\geq 0}$ are independent solutions to (3.6),

iii) $(\tilde\xi_t^{n,i})_{t\geq 0}$ converges uniformly on compact time intervals to $(\tilde X_t^i)_{t\geq 0}$ almost surely for $i \in \{1,2\}$.

Defining $\tilde T_n$ analogously to (3.9), $(\tilde\xi_t^{n,1}, \tilde\xi_t^{n,2})_{t\leq \tilde T_n}$ has the same distribution as $(\xi_t^{n,1}, \xi_t^{n,2})_{t\leq T_n}$. Suppose that $\tilde X_0^1 > \tilde X_0^2$ and define the hitting time of the diagonal by $(\tilde X_t^1, \tilde X_t^2)_{t\geq 0}$ as

$$\tilde T^\Delta = \inf\{t \geq 0 : \tilde X_t^1 \leq \tilde X_t^2\}.$$

Let us show that $\tilde T_n \underset{n\to\infty}{\longrightarrow} \tilde T^\Delta$ almost surely. Set

$$\tilde T_n^\Delta = \inf\{t \geq 0 : \tilde\xi_t^{n,1} \leq \tilde\xi_t^{n,2}\}$$

and note that since the jumps of $\tilde\xi^{n,i}$ are of size at most $2r_+/\sqrt{n}$, the two lineages cannot jump over one another without coming within a distance $2r_+/\sqrt{n}$ of each other, *i.e.* $\tilde T_n \leq \tilde T_n^\Delta$ almost surely. Moreover, define $\tilde T_n^\varepsilon$ and $\tilde T^\varepsilon$ as the hitting times of $\{(x,y) : |x-y| \leq 2r_+\varepsilon\}$ by $(\tilde\xi_t^{n,1}, \tilde\xi_t^{n,2})_{t\geq 0}$ and $(\tilde X_t^1, \tilde X_t^2)_{t\geq 0}$ respectively. By Proposition 3.4.1, $\tilde T_n^\Delta \underset{n\to\infty}{\longrightarrow} \tilde T^\Delta$ a.s. and $\tilde T_n^\varepsilon \underset{n\to\infty}{\longrightarrow} \tilde T^\varepsilon$ a.s. As a result, for all $\varepsilon > 0$,

$$\tilde T^\varepsilon \leq \liminf_{n\to\infty} \tilde T_n \leq \limsup_{n\to\infty} \tilde T_n \leq \tilde T^\Delta \qquad a.s.$$

By the continuity of $t \mapsto (\tilde{X}_t^1, \tilde{X}_t^2)$, $\tilde{T}^\varepsilon \to \tilde{T}^\Delta$ almost surely as $\varepsilon \downarrow 0$, yielding the almost sure convergence of $\tilde{T}_n$ to $\tilde{T}^\Delta$. As a result, $(\tilde{\xi}_t^{n,1}, \tilde{\xi}_t^{n,2})_{t \le \tilde{T}_n}$ converges almost surely to $(\tilde{X}_t^1, \tilde{X}_t^2)_{t \le \tilde{T}^\Delta}$. In other words, $(\xi_t^{n,1}, \xi_t^{n,2})_{t \le T_n}$ converges in distribution to $(X_t^1, X_t^2)_{t \le T^\Delta}$, the trajectory of two independent skew Brownian motions stopped at the time when they hit each other.

We now show that the two particles coalesce quickly once they come within a distance $2r_+/\sqrt{n}$ of each other. This is a consequence of the following result, which is proved as in [BEV10a], Proposition 6.4.

Lemma 3.4.2. *Let T^c denote the coalescence time of the two particles ξ_t^1, ξ_t^2 in $(\mathcal{A}_t)_{t \ge 0}$ (i.e. in the original time scale). Then*

$$\lim_{t \to \infty} \sup_{|y_1 - y_2| \le 2r_+} \mathbb{P}_{y_1, y_2} \left(T^c > t \right) = 0.$$

By the strong Markov property,

$$\mathbb{P}_{x_1, x_2} \left(T_n^c - T_n > t \right) = \mathbb{E}_{x_1, x_2} \left[\mathbb{P}_{\sqrt{n}\xi_{T_n}^{n,1}, \sqrt{n}\xi_{T_n}^{n,2}} \left(T^c > nt \right) \right]. \tag{3.10}$$

The term inside the expectation on the right-hand-side is bounded by $\sup_{|y_1 - y_2| \le 2r_+} \mathbb{P}_{y_1, y_2} \left(T^c > nt \right)$, which converges to zero as $n \to \infty$ by Lemma 3.4.2. In addition, the distance covered by $\xi^{n,i}$ between T_n and T_n^c vanishes as $n \to \infty$. Indeed, in Section 3.5, we prove the following.

Lemma 3.4.3. *For any $\varepsilon > 0$ and $T > 0$,*

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} \mathbb{P} \left(\sup_{\substack{s,t \in [0, nT] \\ |s-t| \le \delta n}} |\xi_s - \xi_t| > \varepsilon \sqrt{n} \right) = 0.$$

Write

$$\mathbb{P} \left(\left| \xi_{T_n^c}^{n,i} - \xi_{T_n}^{n,i} \right| > \varepsilon \right) \le \mathbb{P} \left(\sup_{\substack{s,t \in [0, nT] \\ |s-t| \le \delta n}} |\xi_s - \xi_t| > \varepsilon \sqrt{n} \right) + \mathbb{P} \left(|T_n^c - T_n| > \delta \right) + \mathbb{P} \left(T_n > nT \right) + \mathbb{P} \left(T_n^c > nT \right).$$

Letting $n \to \infty$, the second term on the right-hand-side converges to zero by (3.10). So do the last two terms since both T_n and T_n^c converge in distribution as $n \to \infty$. Then letting $\delta \downarrow 0$, the first term vanishes by Lemma 3.4.3. As a consequence, $(\xi_{T_n^c}^{n,1}, T_n^c)$ converges in distribution (and even in probability) to $(X_{T^\Delta}^1, T^\Delta)$. Since the remaining particle after the coalescence event follows a Markov process with infinitesimal generator $\mathcal{L}$, we know by Proposition 3.3.3 that $(\xi_{T_n^c + t}^{n,1})_{t \ge 0}$ converges in distribution to skew Brownian motion started at $X_{T^\Delta}^1$.

This proves the convergence in distribution of $\mathcal{A}_t^n$ to $\mathcal{A}_t^\infty$ when started from two particles. For larger samples, it is enough to note that three particles (or more) almost never simultaneously come within a distance $2r_+/\sqrt{n}$ of each other. The proof of the convergence of the finite dimensional distributions and that of the second part of the statement follow the same lines, using the Markov property at suitable times. Details can be found in [BEV13b]. □

3.5 Convergence to skew Brownian motion

We now give the proof of Proposition 3.3.3. The arguments are adapted from the work of Iksanov and Pilipenko [IP16]. We limit ourselves to the one dimensional case for the proof, but the generalisation to higher dimensions is straightforward. Iksanov and Pilipenko treat the case of a discrete time Markov chain on $\mathbb{Z}$ which behaves like a simple random walk outside a bounded region centered at the origin. We extend their proof to continuous time jump Markov processes with continuous state space.

3.5.1 Proof of Proposition 3.3.3

Recall that $(\xi_t)_{t\geq 0}$ is a Markov process with generator $\mathcal{L}$ given by (3.2) and $\xi^n(t) = \frac{1}{\sqrt{n}}\xi_{nt}$. As announced above, we restrict ourselves to $d = 1$ and we follow the lines of [IP16]. Set

$$\tilde{X}^{\pm}(t) = \pm\xi_t \mathbb{1}_{\{\pm\xi_t > r_+\}}$$

and

$$\begin{aligned}
\tau_0^{\pm} &= \inf\{t > 0 : |\xi_t| \leq r_+\}, \\
\sigma_k^{\pm} &= \inf\left\{t > \tau_k^{\pm} : \pm\xi_t > r_+\right\}, && k \geq 0, \\
\tau_{k+1}^{\pm} &= \inf\left\{t > \sigma_k^{\pm} : \pm\xi_t \leq r_+\right\}, && k \geq 0.
\end{aligned}$$

One can then write the decomposition (see formula 2.1 in [IP16])

$$\tilde{X}^{\pm}(t) = \tilde{X}^{\pm}(0) + M^{\pm}(t) + L^{\pm}(t) \mp \sum_{i\geq 0} \xi(\tau_i^{\pm}) \mathbb{1}_{\left\{\tau_i^{\pm} \leq t < \sigma_i^{\pm}\right\}} \tag{3.11}$$

with

$$\begin{aligned}
M^{\pm}(t) &= \pm \int_0^t \mathbb{1}_{\{\pm\xi(s^-) > r_+\}} d\xi_s, \\
L^{\pm}(t) &= \pm \sum_{i\geq 0} \left(\xi(\sigma_i^{\pm}) - \xi(\tau_i^{\pm})\right) \mathbb{1}_{\left\{\sigma_i^{\pm} \leq t\right\}}.
\end{aligned}$$

Also set

$$M_n^{\pm}(t) = \frac{1}{\sqrt{n}} M^{\pm}(nt), \qquad L_n^{\pm}(t) = \frac{1}{\sqrt{n}} L^{\pm}(nt).$$

Let $\xi_t^+ = \xi_t \vee 0$ and $\xi_t^- = (-\xi_t) \vee 0$. The following now holds.

Lemma 3.5.1. *For any fixed $T > 0$, the sequence of random variables $(\xi_n^{\pm}, M_n^{\pm}, L_n^{\pm})_{n\geq 1}$ is tight in $\mathrm{D}\left([0,T], \mathbb{R}^6\right)$. Furthermore, any limit point $(X_\infty^{\pm}, M_\infty^{\pm}, L_\infty^{\pm})$ of the sequence is a continuous process satisfying*

$$\int_0^T \mathbb{1}_{\left\{X_\infty^{\pm}(t) = 0\right\}} dt = 0, \quad \text{a.s.} \tag{3.12}$$

Lemma 3.5.2. *Let $(X_\infty^{\pm}, M_\infty^{\pm}, L_\infty^{\pm})$ be the limit point of a converging subsequence of $(\xi_n^{\pm}, M_n^{\pm}, L_n^{\pm})$ in $\mathrm{D}\left([0,T], \mathbb{R}^6\right)$. Then*

i) the processes $L_\infty^\pm$ *are non-decreasing almost surely and satisfy*

$$\int_0^T \mathbb{1}_{\left\{X_\infty^\pm(t)>0\right\}} dL_\infty^\pm(t) = 0 \quad a.s.$$

ii) the processes $M_\infty^\pm$ *are continuous* $\mathcal{F}_t$*-martingales with* $\mathcal{F}_t = \sigma(X_\infty^\pm(s),\, L_\infty^\pm(s),\, M_\infty^\pm(s),\, s \in [0,t])$ *with predictable quadratic variation*

$$\left\langle M_\infty^\pm \right\rangle_t = \sigma_\pm^2 \int_0^t \mathbb{1}_{\left\{X_\infty^\pm(s)>0\right\}} ds$$

where $\sigma_\pm^2 = u\frac{2r_\pm^2}{d+2}$.

Lemma 3.5.3. *There exists* $\beta \in [-1,1]$ *such that, for* $t \geq 0$,

$$L_\infty^+(t) = \frac{1+\beta}{1-\beta} L_\infty^-(t)$$

almost surely.

Proposition 3.3.3 follows from the above lemmas and Proposition 2.1 in [IP16]. Lemma 3.5.1 is proved in Subsection 3.5.3. The proof of Lemma 3.5.2 does not differ from the one given for Lemma 2.2 in [IP16] and we omit the details. The proof of Lemma 3.5.3 is given in Subsection 3.5.4.

3.5.2 Occupation time of the boundary

We begin with the following result controlling the time spent by $(\xi_t)_{t\geq 0}$ in the region $[-r_+, r_+]$.

Lemma 3.5.4. *For* $t \geq 0$, *define* $\nu(t) = \int_0^t \mathbb{1}_{\{|\xi_s| \leq r_+\}} ds$. *Then*

i) $\lim_{t\to\infty} \nu(t) = +\infty$ *almost surely,*

ii) $\sup_{x\in\mathbb{R}} \mathbb{E}_x\left[\nu(t)\right] = \mathcal{O}\left(\sqrt{t}\right)$ *a.s. as* $t \to \infty$.

Proof. The fact that $\nu(t) \to \infty$ as $t \to \infty$ is well known. Set $\zeta_0 = 0$ and

$$\begin{aligned} \varsigma_i &= \inf\left\{t > \zeta_{i-1} : |\xi_t| \leq r_+\right\}, && i \geq 1, \\ \zeta_i &= \inf\left\{t > \varsigma_i : |\xi_t| > r_+\right\}, && i \geq 1. \end{aligned}$$

Then $\nu(t)$ can be written as the sum of the lengths of the excursions inside $[-r_+, r_+]$ up to time t,

$$\nu(t) = \sum_{i\geq 1} \left(\zeta_i \wedge t - \varsigma_i \wedge t\right).$$

Hence

$$\mathbb{E}_x\left[\nu(t)\right] \leq \mathbb{E}_x\left[\sum_{i\geq 1} \mathbb{E}\left[\zeta_i - \varsigma_i \mid \mathcal{F}_{\varsigma_i}\right] \mathbb{1}_{\{\varsigma_i \leq t\}}\right].$$

Noting that there exists $\varepsilon > 0$ such that $\mathbb{P}\left(|\xi(t+dt)| > r_+ \mid \xi_t = x\right) \geq \varepsilon dt$ for all $|x| \leq r_+$, we see that $\zeta_i - \varsigma_i$ is stochastically dominated by an exponential random variable with parameter ε. Hence

$$\mathbb{E}_x\left[\nu(t)\right] \leq \frac{1}{\varepsilon}\mathbb{E}_x\left[\sum_{i\geq 1}\mathbb{1}_{\{\varsigma_i \leq t\}}\right].$$

In addition, the number of visits to $[-r_+, r_+]$ before time t is less than the number of visits to this set before the first excursion longer than t, *i.e.*

$$\sum_{i\geq 1}\mathbb{1}_{\{\varsigma_i \leq t\}} \leq m(t) := \inf\{i \geq 1 : \varsigma_{i+1} - \zeta_i > t\}.$$

Let $(W_t)_{t\geq 0}$ be a continuous time random walk on $\mathbb{R}$ with jump rate u and independent increments distributed according to

$$\frac{|B(0,1) \cap B(y,1)|}{V_1^2} dy.$$

Then for any $x > r_+$,

$$\mathbb{P}_{\pm x}\left(\varsigma_1 - \zeta_0 > t\right) \geq \mathbb{P}_0\left(\inf_{0\leq s\leq t} W_s \geq 0\right).$$

(Notice that the right-hand-side isn't changed if W is replaced by $r_\pm W$.) As a result $m(t)$ is stochastically dominated by a geometric random variable with parameter

$$p(t) = \mathbb{P}_0\left(\inf_{0\leq s\leq t} W_s \geq 0\right).$$

Furthermore, there exists $\eta > 0$ such that, for all $t \geq 0$, $p(t) \geq \frac{\eta}{\sqrt{t}}$, (see pp. 381-382 in [BGT89] or equations (3.4) and (3.5) in [IP16]). As a result,

$$\mathbb{E}_x\left[\nu(t)\right] \leq \frac{1}{\varepsilon p(t)} \leq \frac{\sqrt{t}}{\varepsilon\eta}.$$

□

3.5.3 Tightness of $(\xi_n^\pm,\ M_n^\pm,\ L_n^\pm)_{n\geq 1}$

Let us now give the proof of Lemma 3.5.1. To prove that the sequence $(\xi_n^\pm,\ M_n^\pm,\ L_n^\pm)_{n\geq 1}$ is tight in $\mathrm{D}\left([0,T], \mathbb{R}^6\right)$, we use the following criterion proved by Aldous [Ald78].

Theorem 3.3 (Aldous [Ald78])**.** *Suppose $(X_n, n \geq 0)$ is a sequence of random variables taking values in $\mathrm{D}\left([0,T], \mathbb{R}\right)$ such that*

i) $(X_n(0), n \geq 0)$ and $\left(\sup_{t\geq 0}|X_n(t) - X_n(t^-)|, n \geq 0\right)$ are tight in $\mathbb{R}$,

ii) for any sequence $\{\tau_n, \delta_n\}$ such that τ_n is a stopping time with respect to the natural filtration of X_n and $\delta_n \in [0,1]$ is a constant such that $\delta_n \to 0$ as $n \to \infty$,

$$X_n(\tau_n + \delta_n) - X_n(\tau_n) \xrightarrow[n\to\infty]{\mathcal{P}} 0.$$

Then $(X_n, n \geq 0)$ is tight in $\mathrm{D}\left([0,T], \mathbb{R}\right)$.

Proof of Lemma 3.5.1. From (3.11), and the fact that

$$\left|\sum_{i\geq 0}\xi(\tau_i^{\pm})\mathbb{1}_{\left\{\tau_i^{\pm}\leq t<\sigma_i^{\pm}\right\}}\right|\leq r_+,$$

it is enough to prove the tightness of $\xi_n^{\pm}$ and $M_n^{\pm}$. We use Aldous' criterion to prove that $M_n^{\pm}$ is tight and then we use the fact that the increments of ξ are bounded by those of $M := M^+ - M^-$ (equation (3.13) below) to show that ξ_n is tight.

From the definition of ξ, we have $M_n^{\pm}(0) = 0$ and

$$\sup_{t\geq 0}\left|M_n^{\pm}(t) - M_n^{\pm}(t_-)\right| \leq \frac{2r_+}{\sqrt{n}}.$$

Moreover, for any stopping time T and $\delta > 0$, since outside $[-r_+, r_+]$, ξ behaves as a simple random walk,

$$\mathbb{E}\left[\left(M_n^{\pm}(T+\delta) - M_n^{\pm}(T)\right)^2\right] \leq \sigma_{\pm}^2\delta.$$

The assumptions of Theorem 3.3 are thus satisfied, proving the tightness of $\left(M_n^{\pm}\right)_n$.

Now take $0 \leq s \leq t$. If ξ does not visit $[-r_+, r_+]$ between time s and time t, then $\xi_t - \xi_s = M(t) - M(s)$. If it does visit this set, then let α be the first time ξ enters $[-r_+, r_+]$ after time s and θ the last time ξ leaves this set before time t. Then

$$\begin{aligned}|\xi_t - \xi_s| &\leq |\xi_t - \xi_\theta| + |\xi_\theta - \xi_\alpha| + |\xi_\alpha - \xi_s| \\ &\leq 2r_+ + |M(t) - M(\theta)| + |M(\alpha) - M(s)|.\end{aligned}$$

As a result, for $\delta > 0$,

$$\sup_{|s-t|\leq\delta n}|\xi_s - \xi_t| \leq 2r_+ + 2\sup_{|s-t|\leq\delta n}|M(s) - M(t)|. \tag{3.13}$$

This bound is proved in [IP16] (equation (3.10)).

The tightness of $\left(\xi_n\right)_n$ then follows from that of $\left(M_n^{\pm}\right)_n$ by writing

$$\begin{aligned}\lim_{\delta\downarrow 0}\limsup_{n\to\infty}\mathbb{P}\Big(\sup_{\substack{|s-t|\leq\delta n\\ s,t\in[0,nT]}}|\xi_s - \xi_t| > \varepsilon\sqrt{n}\Big)& \\ \leq \lim_{\delta\downarrow 0}\limsup_{n\to\infty}\mathbb{P}\Big(2r_+ + 2\sup_{\substack{|s-t|\leq\delta n\\ s,t\in[0,nT]}}|M(s) - M(t)| > \varepsilon\sqrt{n}\Big) = 0.&\end{aligned} \tag{3.14}$$

It remains to prove (3.12). Note that any limit point $(X_\infty^{\pm}, M_\infty^{\pm}, L_\infty^{\pm})$ satisfies

$$X_\infty(t) = X_\infty^+(t) - X_\infty^-(t) = M_\infty^+(t) - M_\infty^-(t) + L_\infty^+(t) - L_\infty^-(t) = M_\infty(t) + L_\infty(t).$$

From the definition of $M_n^{\pm}$ and Lemma 3.5.4, one shows, as in [IP16], that M_∞ is a stochastic integral

with respect to standard Brownian motion $(B_t)_{t\geq 0}$

$$M_\infty(t) = \int_0^t \sigma(X_\infty(s))dB_s.$$

In addition, $L_\infty^\pm$ is a continuous process with locally bounded variation. As a result $\langle X_\infty\rangle_t = \langle M_\infty\rangle_t$ and (3.12) follows from the occupation density formula. □

Note that (3.14) proves Lemma 3.4.3.

3.5.4 The left and right local time at zero of $(\xi_t)_{t\geq 0}$

The proof of Lemma 3.5.3 is adapted from that of Lemma 2.3 in [IP16]. Recall the expression for the left and right local time of $(\xi_t)_{t\geq 0}$,

$$L^\pm(t) = \pm\sum_{i\geq 0}(\xi(\sigma_i^\pm) - \xi(\tau_i^\pm))\mathbb{1}_{\{\sigma_i^\pm\leq t\}}.$$

For any particular visit of ξ to $[-r_+, r_+]$, the value of $\xi(\sigma_i^\pm) - \xi(\tau_i^\pm)$ depends on the value of ξ when it enters this set. However, over many visits to $[-r_+, r_+]$, $L^\pm(t)$ only records an average of these values. The "typical" value of $\xi(\sigma_i^\pm) - \xi(\tau_i^\pm)$ can thus be expressed with the help of the stationary distribution of the process describing the visits of ξ to $[-r_+, r_+]$ (Y below). The left and right local time of ξ then become asymptotically proportional to the occupation time of the boundary $\nu(t)$, with different coefficients whose expressions can be found below.

Recall from Lemma 3.5.4 that $\nu(t) = \int_0^t \mathbb{1}_{\{|\xi_s|\leq r_+\}}ds$ and set, for $t \geq 0$,

$$\alpha(t) = \inf\{\alpha > 0 : \nu(\alpha) > t\}.$$

Define $Y(t) = \xi(\alpha(t))$ for $t \geq 0$. The process $(Y(t))_{t\geq 0}$ is a jump Markov process taking values in $[-r_+, r_+]$, describing the values taken by ξ inside this region. Let $\bar\alpha$ denote the left-continuous version of α, *i.e.* for $t \geq 0$,

$$\bar\alpha(t) = \sup\{\alpha \geq 0 : \nu(\alpha) < t\}.$$

If $t \geq 0$ is such that $\bar\alpha(t) \neq \alpha(t)$, then ξ makes an excursion outside $[-r_+, r_+]$ between time $\bar\alpha(t)$ and time $\alpha(t)$.

Let $V^\pm$ be defined by

$$V^\pm(t) = \pm\sum_{0<s\leq t}(Y(s) - Y(s^-)) \pm \sum_{0<s\leq t}(\xi(\bar\alpha(s)) - \xi(\alpha(s)))\mathbb{1}_{\{\pm\xi(\bar\alpha(s))>r_+\}}.$$

Lemma 3.5.5. *There exist $C > 0$ and $\beta \in [-1, 1]$ such that, as $t \to \infty$,*

$$\frac{1}{t}V^\pm(t) \to C(1 \pm \beta)$$

almost surely.

To prove Lemma 3.5.5, we use the Nummelin splitting technique [Num78] to turn Y into a renewal process. We can then build its stationary probability distribution (see Subsection 3.5.5), following

Chapter 6.8 in [Dur10]. Lemma 3.5.5 is then reduced to the strong law of large numbers for renewal processes. The detailed argument is given in Subsection 3.5.6.

Proof of Lemma 3.5.3. We first show that $V^{\pm}(\nu(t))$ provides a good approximation of $L^{\pm}(t)$ and then conclude with the help of Lemma 3.5.5. Note that $\pm\xi(\bar{\alpha}(s)) > r_+$ with $s > 0$ if and only if $\bar{\alpha}(s) = \sigma_i^{\pm}$ for some $i \geq 1$, and in this case, $\alpha(s) = \tau_{i+1}^{\pm}$. In addition, $s \leq \nu(t)$ if and only if $\bar{\alpha}(s) \leq t$, as a result

$$V^{\pm}(\nu(t)) = \pm(Y(\nu(t)) - Y(0)) \pm \sum_{i\geq 1}(\xi(\sigma_i^{\pm}) - \xi(\tau_{i+1}^{\pm}))\mathbb{1}_{\{\sigma_i^{\pm}\leq t\}}.$$

Hence

$$\left|V^{\pm}(\nu(t)) - L^{\pm}(t)\right| \leq |Y(\nu(t))| + |Y(0)| + \left|\xi(\sigma_0^{\pm})\right| + \left|\xi(\tau_0^{\pm})\right| + \left|\xi(\tau_1^{\pm})\right| + \sum_{i\geq 2}\left|\xi(\tau_i^{\pm})\right|\mathbb{1}_{\{\sigma_{i-1}^{\pm}\leq t<\sigma_i^{\pm}\}}.$$

Since $|Y(t)| \leq r_+$, $\left|\xi(\tau_i^{\pm})\right| \leq r_+$ and $\left|\xi(\sigma_i^{\pm})\right| \leq 3r_+$,

$$\left|V^{\pm}(\nu(t)) - L^{\pm}(t)\right| \leq 8r_+.$$

From this, Lemma 3.5.5, and using Lemma 3.5.4.i, we obtain

$$\lim_{t\to\infty}\frac{L^+(t)}{L^-(t)} = \frac{1+\beta}{1-\beta}. \tag{3.15}$$

□

3.5.5 The Stationary distribution of Y

Let $\Phi^Y : [-r_+, r_+]^2 \to \mathbb{R}_+$ be such that

$$\mathcal{L}^Y f(x) = u\int_{[-r_+,r_+]}\Phi^Y(x,y)(f(y) - f(x))dy$$

is the infinitesimal generator of $(Y(t))_{t\geq 0}$. Clearly, from (3.3), for $x, y \in [-r_+, r_+]$,

$$\Phi^Y(x,y) \geq \Phi(x,y).$$

Note that Φ is continuous on the compact set $[-r_+, r_+]^2$ and that it stays strictly positive on sets of the form $U_{a,b} = [-r_+, r_+] \times [a, b]$ with $-r_+ < a < b < -r_+ + 2r_-$. Fix one such set $U_{a,b}$ and set $\Phi_{min} = \inf_{U_{a,b}}\Phi > 0$. As a result

$$\Phi^{\varepsilon}(x,y) := \Phi^Y(x,y) - \frac{\varepsilon}{b-a}\mathbb{1}_{\{[a,b]\}}(y) \geq 0$$

for $\varepsilon = (b-a)\Phi_{min} > 0$.

We now follow Chapter 6.8 of [Dur10] to build the (unique) stationary probability measure of Y.

Define an operator $\mathcal{L}^Z$ on real-valued functions f on $[-r_+, r_+] \cup \{\partial\}$ by

$$\mathcal{L}^Z f(x) = \begin{cases} u \int_{[-r_+,r_+]} \Phi^\varepsilon(x,y)(f(y) - f(x))dy + u\varepsilon(f(\partial) - f(x)) & \text{if } x \in [-r_+, r_+], \\ \dfrac{1}{b-a} \int_a^b (f(y) - f(\partial))dy & \text{if } x = \partial, \end{cases} \tag{3.16}$$

and let $(Z(t))_{t\geq 0}$ be a Markov process on $[-r_+, r_+] \cup \{\partial\}$ with generator $\mathcal{L}^Z$. Let

$$\lambda(t) = \inf\left\{\lambda > 0 : \int_0^\lambda \mathbb{1}_{\{Z(s)\neq\partial\}} ds > t\right\}, \tag{3.17}$$

then

$$(Z(\lambda(t)), t \geq 0) \overset{d}{=} (Y(t), t \geq 0)\,.$$

Set $E_0 = 0$ and, for $k \geq 0$,

$$\begin{aligned} R_k &= \inf\{t \geq E_k : Z(t) = \partial\}, \\ E_{k+1} &= \inf\{t \geq R_k : Z(t) \neq \partial\}. \end{aligned}$$

Then $R_k - E_k$ is an exponential random variable with parameter $u\varepsilon$ for all $k \geq 1$ and ∂ is a positive recurrent state for Z. We can then use this fact to build a stationary probability measure for Y. Let $\mathbb{E}_\partial$ denote the expectation with respect to $\mathbb{P}(\,\cdot \mid Z(0) = \partial)$.

Lemma 3.5.6. *The measure π defined by*

$$\int_{[-r_+,r_+]} f(x)\pi(dx) = u\varepsilon \mathbb{E}_\partial \left[\int_{E_1}^{R_1} f(Z(s))ds\right] \tag{3.18}$$

is an invariant probability measure for $(Y(t))_{t\geq 0}$.

Since Y is irreducible with respect to the Lebesgue measure on $[-r_+, r_+]$, *i.e.* any two sets of positive Lebesgue measure communicate with each other (see [Dob40] or [Num04]), π is unique.

Proof. Let $f : [-r_+, r_+] \cup \{\partial\} \to \mathbb{R}$ be bounded and measurable. Since $\mathcal{L}^Z$ is the generator of Z, by the optional stopping time theorem,

$$\mathbb{E}_\partial \left[f(Z(R_1)) - f(Z(E_1)) - \int_{E_1}^{R_1} \mathcal{L}^Z f(Z(s))ds\right] = 0.$$

By the definition of R_1 and E_1,

$$f(Z(R_1)) = f(\partial), \qquad \mathbb{E}_\partial\left[f(Z(E_1))\right] = \frac{1}{b-a}\int_a^b f(y)dy.$$

And by the definition of $\mathcal{L}^Z$ in (3.16),

$$\mathcal{L}^Z f(x) = \mathcal{L}^Y f(x) + u\varepsilon\left(f(\partial) - \frac{1}{b-a}\int_a^b f(y)dy\right).$$

Combining these equalities with the fact that $\mathbb{E}_\partial [R_1 - E_1] = \dfrac{1}{u\varepsilon}$, we obtain

$$\int_{[-r_+,r_+]} \mathcal{L}^Y f \, d\pi = u\varepsilon \mathbb{E}_\partial \left[\int_{E_1}^{R_1} \mathcal{L}^Y f(Z(s)) ds \right] = 0.$$

□

Furthermore, using the fact that $\Phi(x,y) = \Phi(y,x)$, we are able to identify π.

Lemma 3.5.7. *The measure π is the uniform probability distribution on $[-r_+, r_+]$.*

Proof. For f and g two bounded and measurable functions on $[-r_+, r_+]$, let

$$\langle f, g \rangle_\pi = \int_{-r_+}^{r_+} f(x) g(x) dx.$$

We want to show

$$\left\langle \mathcal{L}^Y f, g \right\rangle_\pi = \left\langle f, \mathcal{L}^Y g \right\rangle_\pi . \tag{3.19}$$

For $f : [-r_+, r_+] \to \mathbb{R}$ and $x \in \mathbb{R}$, let

$$Ef(x) := \mathbb{E}_x \left[f(Y(0)) \right] \tag{3.20}$$

and note that $\mathcal{L}^Y f(x) = \mathcal{L} E f(x)$. In addition, since $\Phi(x,y) = \Phi(y,x)$, for any $f, g \in L^2(\mathbb{R})$,

$$\langle \mathcal{L} f, g \rangle_\mathbb{R} = \langle f, \mathcal{L} g \rangle_\mathbb{R} .$$

However, $Ef \notin L^2(\mathbb{R})$. For $n \geq 1$, define

$$\mathcal{T}^n = \inf\{t > 0 : |\xi(t)| \leq r_+ \text{ or } |\xi(t)| \geq n\}$$

and, for $f : [-r_+, r_+] \to \mathbb{R}$ bounded,

$$E^n f(x) = \mathbb{E}_x \left[f(\xi(\mathcal{T}^n)) \mathbb{1}_{\{|\xi(\mathcal{T}^n)| \leq r_+\}} \right].$$

Then

$$E^n f(x) = \begin{cases} f(x) & \text{if } |x| \leq r_+ \\ 0 & \text{if } |x| \geq n, \end{cases} \tag{3.21}$$

$$\mathcal{L} E^n f(x) = 0 \qquad \text{if } r_+ < |x| < n. \tag{3.22}$$

In particular, $E^n f \in L^2(\mathbb{R})$. As a result,

$$\langle \mathcal{L} E^n f, E^n g \rangle_\mathbb{R} = \langle E^n f, \mathcal{L} E^n g \rangle_\mathbb{R} . \tag{3.23}$$

In addition, from (3.21) and (3.22),

$$\begin{aligned}\langle \mathcal{L}E^n f, E^n g\rangle_{\mathbb{R}} &= \langle \mathcal{L}E^n f, E^n g\rangle_{\pi} \\ &= \langle \mathcal{L}E^n f, g\rangle_{\pi}\,.\end{aligned}$$

Finally, for any $x \in \mathbb{R}$, $\mathcal{T}^n \underset{n\to\infty}{\longrightarrow} Y(0) = \inf\{t > 0 : |\xi(t)| \leq r_+\}$ almost surely. By dominated convergence, for $x \in \mathbb{R}$, $E^n f(x) \underset{n\to\infty}{\longrightarrow} Ef(x)$ and, using dominated convergence once more, we obtain

$$\langle \mathcal{L}E^n f, g\rangle_{\pi} \underset{n\to\infty}{\longrightarrow} \langle \mathcal{L}Ef, g\rangle_{\pi}\,.$$

Applying the same argument to the right-hand-side of (3.23), we obtain (3.19). As a result the uniform measure on $[-r_+, r_+]$ is invariant for Y. Since Y is irreducible with respect to the Lebesgue measure and π defined in (3.18) is absolutely continuous with respect to the Lebesgue measure, π is the uniform probability measure on $[-r_+, r_+]$. □

3.5.6 Proof of Lemma 3.5.5

Now that we have built the stationary probability measure for Y, we can prove Lemma 3.5.5, adapting the arguments of [IP16, Lemma 2.3]. The proof is an application of the law of large numbers to the renewal process $(Z(t))_{t\geq 0}$.

Recall that $Y(t) = Z(\lambda(t))$ with λ defined in (3.17). From the definition of λ, for $t \geq 0$,

$$\lambda^{-1}(t) = \int_0^t \mathbb{1}_{\{Z_s \neq \partial\}} ds.$$

For $k \geq 0$, set

$$\tilde{R}_k = \lambda^{-1}(R_k) = \lambda^{-1}(E_{k+1}),$$

and

$$V_k^{\pm} = V^{\pm}(\tilde{R}_{k+1}) - V^{\pm}(\tilde{R}_k).$$

Then $V_0^{\pm}, V_1^{\pm}, \ldots$ are independent and for all $k \geq 1$, $V_k^{\pm}$ is distributed as $V_1^{\pm}$ under $\mathbb{E}_\partial$. Recall the definition of the operator E in (3.20) and set $\iota(x) = x$ for $x \in \mathbb{R}$. We prove the following lemma at the end of this subsection.

Lemma 3.5.8.

$$\mathbb{E}_\partial\left[V_1^{\pm}\right] = \frac{1}{2r_+\varepsilon}\int_{-r_+}^{r_+}\int_{\mathbb{R}} \Phi(x,y)(y - E\iota(y))^{\pm} dy dx,$$

where $(\cdot)^+$ (resp $(\cdot)^-$) denotes the positive (resp. negative) part.

Proof of Lemma 3.5.5. Setting $N(t) = \max\{k \geq 0 : \tilde{R}_{k+1} \leq t\}$, by the strong law of large numbers for renewal processes, we have

$$\lim_{t\to\infty} \frac{1}{t}\sum_{k=1}^{N(t)} V_k^{\pm} = \lim_{t\to\infty} \frac{N(t)}{t}\frac{1}{N(t)}\sum_{k=1}^{N(t)} V_k^{\pm} = u\varepsilon\mathbb{E}_\partial\left[V_1^{\pm}\right] \quad \text{a.s.} \tag{3.24}$$

Moreover,

$$V^{\pm}(t) - \sum_{k=0}^{N(t)} V_k^{\pm} = V^{\pm}(t) - V^{\pm}(\tilde{R}_{N(t)+1}) \\ = \pm(Y(t) - Y(\tilde{R}_{N(t)+1})) \pm \sum_{\tilde{R}_{N(t)+1} \leq s < t} (\xi(\bar{\alpha}(s)) - \xi(\alpha(s)))\mathbb{1}_{\{\pm\xi(\bar{\alpha}(s))>r_+\}}.$$

Taking absolute values on both sides, we have

$$\left| V^{\pm}(t) - \sum_{k=0}^{N(t)} V_k^{\pm} \right| \leq \left| Y(t) + Y(\tilde{R}_{N(t)+1}) \right| + \sum_{\tilde{R}_{N(t)+1} \leq s < t} |\xi(\bar{\alpha}(s)) - \xi(\alpha(s))|\, \mathbb{1}_{\{\pm\xi(\bar{\alpha}(s))>r_+\}}$$

Since when $\pm\xi(\bar{\alpha}(s)) > r_+$, $\pm(\xi(\bar{\alpha}(s)) - \xi(\alpha(s))) \geq 0$, we can add the terms for which $t \leq s < \tilde{R}_{N(t)+2}$ on the right-hand-side,

$$\left| V^{\pm}(t) - \sum_{k=0}^{N(t)} V_k^{\pm} \right| \leq \left| Y(t) + Y(\tilde{R}_{N(t)+1}) \right| + \left| \sum_{\tilde{R}_{N(t)+1} \leq s < \tilde{R}_{N(t)+2}} (\xi(\bar{\alpha}(s)) - \xi(\alpha(s)))\mathbb{1}_{\{\pm\xi(\bar{\alpha}(s))>r_+\}} \right|$$

Adding and subtracting $Y(\tilde{R}_{N(t)+2}) - Y(\tilde{R}_{N(t)+1})$ inside the absolute value, we obtain,

$$\left| V^{\pm}(t) - \sum_{k=0}^{N(t)} V_k^{\pm} \right| \leq 4r_+ + \left| V_{N(t)+1}^{\pm} \right|.$$

Hence, since the $V_k^{\pm}$ are identically distributed for $k \geq 1$,

$$\left| \frac{1}{t} V^{\pm}(t) - \frac{1}{t} \sum_{k=0}^{N(t)} V_k^{\pm} \right| \leq \frac{4r_+}{t} + \frac{1}{t} \left| V_{N(t)+1}^{\pm} \right|.$$

The right-hand-side converges to zero almost surely as $t \to \infty$ since the $V_k^{\pm}$ are identically distributed for $k \geq 1$. As a result, from (3.24)

$$\lim_{t\to\infty} \frac{1}{t} V^{\pm}(t) = u\varepsilon \mathbb{E}_{\partial} \left[V_1^{\pm} \right]. \tag{3.25}$$

The statement of Lemma 3.5.5 now follows from Lemma 3.5.8 by taking

$$\beta = \frac{\int_{-r_+}^{r_+} \int_{\mathbb{R}} \Phi(x,y)(y - E\iota(y))dydx}{\int_{-r_+}^{r_+} \int_{\mathbb{R}} \Phi(x,y) \left| y - E\iota(y) \right| dydx}. \tag{3.26}$$

□

We now prove Lemma 3.5.8.

Proof of Lemma 3.5.8. Define

$$h^{\pm}(x) = \pm u \int_{\mathbb{R}} \Phi(x,y) \mathbb{1}_{\{\pm y \leq r_+\}} (E\iota(y) - x) dy \pm u \int_{\mathbb{R}} \Phi(x,y) \mathbb{1}_{\{\pm y > r_+\}} (y - x) dy. \tag{3.27}$$

Writing

$$V^{\pm}(t) = \pm \sum_{0<s\leq t} (Y(s) - Y(s^-)) \mathbb{1}_{\{\pm\xi(\bar{\alpha}(s))\leq r_+\}} \pm \sum_{0<s\leq t} (\xi(\bar{\alpha}(s)) - Y(s^-)) \mathbb{1}_{\{\pm\xi(\bar{\alpha}(s))> r_+\}},$$

it follows that

$$V^{\pm}(t) - \int_0^t h^{\pm}(Y(s)) ds$$

is a martingale with respect to the filtration associated with $(Y(t))_{t\geq 0}$. As a result,

$$\begin{aligned} u\varepsilon \mathbb{E}_{\partial}\left[V_1^{\pm}\right] &= u\varepsilon \mathbb{E}_{\partial}\left[\int_{\lambda^{-1}(E_1)}^{\lambda^{-1}(R_1)} h^{\pm}(Y(s)) ds\right] \\ &= u\varepsilon \mathbb{E}_{\partial}\left[\int_{E_1}^{R_1} h^{\pm}(Z(s)) ds\right] \\ &= \left\langle h^{\pm}, \pi \right\rangle, \end{aligned} \tag{3.28}$$

by (3.18). Note that since $E\iota(y) = y$ when $|y| \leq r_+$, $h^{\pm}$ can be written as

$$\begin{aligned} h^{\pm}(x) &= \pm u \int_{\mathbb{R}} \Phi(x,y)(E\iota(y) - E\iota(x)) dy \\ &\qquad \pm u \int_{\mathbb{R}} \Phi(x,y) \mathbb{1}_{\{\pm y > r_+\}} (y - E\iota(y)) dy \\ &= \pm \mathcal{L} E\iota(x) + u \int_{\mathbb{R}} \Phi(x,y)(y - E\iota(y))^{\pm} dy, \end{aligned}$$

Besides, we noted above that $\mathcal{L}Ef = \mathcal{L}^Y f$, hence $\langle \mathcal{L}E\iota, \pi \rangle = 0$. Furthermore, from Lemma 3.5.7,

$$\left\langle h^{\pm}, \pi \right\rangle = \frac{u}{2r_+} \int_{-r_+}^{r_+} \int_{\mathbb{R}} \Phi(x,y)(y - E\iota(y))^{\pm} dy dx.$$

This, together with (3.28) concludes the proof of Lemma 3.5.8. □

3.6 Inequalities for hitting times

Proof of Proposition 3.4.1. We first prove the inequality for T_O^n. Suppose that $\limsup T_O^n > T_O$ and fix $\varepsilon > 0$ such that $T_O + \varepsilon \leq \limsup T_O^n$. There exists a subsequence $(n_k)_k$ such that for all $k \in \mathbb{N}$, $T_O^{n_k} \geq T_O + \varepsilon$. By the definition of T_O, there exists $t \in [T_O, T_O + \varepsilon)$ such that $f(t) \in O$. By the convergence of f_n to f, $f_{n_k}(t)$ converges to $f(t)$ as $k \to \infty$. Since $f(t) \in O$ which is open, for k large enough, $f_{n_k}(t) \in O$ and $T_O^{n_k} \leq t$, leading to a contradiction.

For the second inequality, suppose that $\liminf T_F^n < T_F$ and take $\varepsilon > 0$ such that $\liminf T_F^n \leq T_F - 2\varepsilon$. There exists a subsequence $(n_k)_k$ such that for all $k \in \mathbb{N}$, $T_F^{n_k} \leq T_F - 2\varepsilon$. Since f is continuous, the image of $[0, T_F - \varepsilon]$ by f is a compact set which does not intersect F, hence there

exists $\eta > 0$ such that its η-neighbourhood is in $\mathbb{R}^d \setminus F$. By the locally uniform convergence of f_n to f, $\sup\{|f_{n_k}(t) - f(t)| : t \in [0, T_F - \varepsilon]\}$ converges to zero as $k \to \infty$. Taking k large enough that this quantity is smaller that η, we have that $f_{n_k}(t) \notin F$ for $t \in [0, T_F - \varepsilon]$. Hence $T_F^{n_k} \geq T_F - \varepsilon$, which is a contradiction. □

Chapter 4

Gene flow accross geographical barriers - scaling limits of random walks with obstacles

Introduction

Barriers to gene flow are physical obstacles to migration. Examples include mountain ranges, highways, political borders and the Great Wall of China [Su+03]. All these geographical features leave traces in the genetic composition of populations living on both sides of the barrier. Geneticists try to use these traces to detect barriers to gene flow and to quantify their effect on migration.

A naive approach to this problem would be to compute a measure of genetic differentiation (*e.g.* Fst) between the two populations on each side of the candidate barrier, and to say that the latter acts as a barrier to gene flow if two individuals living on the same side are more related to each other on average than two indivudals living on different sides of the barrier.

This method assumes that the two subpopulations on each side of the obstacle are well mixed. This may not always be a reasonable assumption and in some cases it is preferable to take into account the finer scale geographic structure of the sampled population.

Mathematical models for spatially extended populations with barriers to gene flow already exist in the literature [Nag76; Sla73], but most assume a discrete space and finding analytical formulae in this framework is challenging at best. Such formulae are particularly useful for inference purposes, where computational power is limiting. This chapter is a step towards a rigorous mathematical framework to model genetic isolation by distance with barriers to gene flow in a continuous space.

Stepping stone model with a cline Nagylaki and his co-authors proposed the following model [Nag76; Nag12a; NZ16]. Consider a population living in a discrete linear space, with colonies (or demes) at locations $\{\ldots,-2,-1,1,2,\ldots\}$. Each deme contains N individuals, and at each generation, those individuals are replaced by the offspring of the previous generation. Migration takes place between nearest neighbours and demes sitting at $\{\ldots-3,-2,2,3,\ldots\}$ exchange a proportion $m/2$ of

individuals with each of their neighbours at each generation, for some $m \in (0,1)$. The two demes sitting at $\{-1,1\}$ exchange a proportion $cm/2$ of their individuals with each other at each generation, with $c \in (0,1)$. Migration probabilities are depicted in Figure 4.1. Properties of this model and applications to various settings were studied in a sequence of papers [Nag76; NB88; NKD93; Nag12b; Nag16].

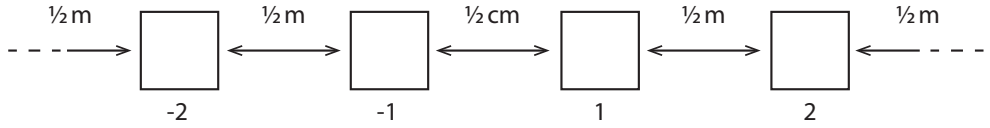

Figure 4.1: Stepping stone model with a barrier to gene flow

In this model, two individuals sampled at a given distance from each other will be more related if they are sampled on the same side of the origin than if they are not. This can be seen by assuming that each new individual mutates to a type never seen before with some probability $\mu \in (0,1)$. Relatedness between individuals can then be measured by the probability that two sampled individuals are of the same type. This probability is called the probability of identity by descent, and it is given by the probability generating function of the age of the most recent common ancestor of these individuals. Properties of this function were studied in this setting in [Bar08].

Also of interest is the evolution of the frequency of a given type (or allele) in the population. Ignoring mutations and assuming that N is large enough, this frequency solves a simple difference equation. Nagylaki [Nag76] showed that, over large spatial and temporal scales, it can be approximated by the solution to the following equation

$$\begin{cases} \partial_t p(t,x) = \frac{\sigma^2}{2} \partial_{xx} p(t,x) & \text{for } x \in \mathbb{R} \setminus \{0\}, \\ \partial_x p(t,0^-) = \partial_x p(t,0^+) = \gamma(p(t,0^+) - p(t,0^-)) \end{cases} \tag{4.1}$$

where $\sigma^2 = m$ and $\gamma = c/\varepsilon$, ε being the distance between neighbouring colonies, see Figure 4.2. In [NB88] (see also [Bar08]), Nagylaki showed a similar approximation for the probability of identity by descent.

Duality An alternative way to study this model from the forwards in time evolution of types is to look back in time for the position of one's ancestor some number of generations in the past. If ξ_t^x denotes the position of the ancestor t generations ago of an individual sampled at $x \in \mathbb{Z} \setminus \{0\}$, then $(\xi_t^x, t \geq 0)$ is a random walk with transition probabilities given by the migration rates in Figure 4.1. In the absence of mutations, the proportion of individuals carrying a given allele at location x is the proportion of those individuals whose ancestor t generations ago carried the same allele. As a result, for N large enough,

$$p(t,x) = \mathbb{E}\left[p(0,\xi_t^x)\right].$$

Likewise, the probability of identity by descent can be expressed with the help of the coalescence time of two random walks ξ^x, ξ^y, *i.e.* the first time that the two ancestors have the same parent.

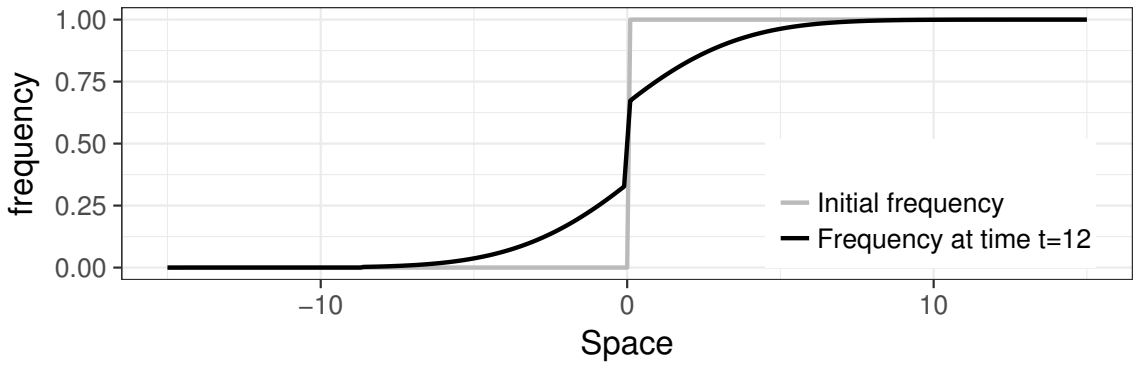

Figure 4.2: Evolution of allele frequency
Frequency of a type initially present at the right of the origin after a few generations.

Scaling limits of random walks with obstacles In this chapter, we present a result on the scaling limits of a class of random walks with obstacles which includes $(\xi_t^x, t \geq 0)$. For $n \geq 1$, if we set

$$X_n(t) = \frac{1}{\sqrt{n}} \xi_{nt},$$

and if c is of order $n^{-1/2}$, we show that X_n converges in distribution to a continuous stochastic process. This process resembles Brownian motion everywhere except near the origin where it has a singular behaviour. More precisely, this process behaves like reflected Brownian motion until its local time at the origin reaches an exponential random variable, after which it becomes reflected Brownian motion on the other side of the origin, until its local time reaches another exponential variable, and so on. We call this process partially reflected Brownian motion. It generalises elastic Brownian motion considered for example in [Gre06].

The same process was obtained as a limit of one dimensional diffusions in [MP16]. For $\varepsilon > 0$, they consider $(X_\varepsilon(t), t \geq 0)$, solution to

$$dX_\varepsilon(t) = \frac{L_\varepsilon}{\varepsilon} a\left(\frac{1}{\varepsilon} X_\varepsilon(t)\right) dt + dB_t,$$

where B is standard Brownian motion and $\text{sign}(x)a(x) > 0$, $supp(a) \subset [-1, 1]$, and they give conditions under which X_ε converges to partially reflected Brownian motion (which they call Brownian motion with a hard membrane).

In addition, we give a different construction of partially reflected Brownian motion inspired by the speed and scale construction of one dimensional diffusions. Starting with standard Brownian motion, we glue together its excursions above level $x > 0$ and below level $-x$ and we show that the result is the same process as the one described above.

Moreover, we provide a martingale problem characterisation of partially reflected Brownian motion, where equation (4.1) can be seen as the action of the semigroup of partially reflected Brownian motion on the initial allele frequency. In particular, the domain of the infinitesimal generator associated to partially reflected Brownian motion is precisely the space of twice continuously differentiable

functions on $\mathbb{R} \setminus \{0\}$ satisfying

$$\partial_x p(0^+) = \partial_x p(0^-) = \gamma(p(0^+) - p(0^-))$$

for some $\gamma > 0$.

We also provide an explicit formula for the transition density of partially reflected Brownian motion in Corollary 4.1.6 below. It turns out that this transition density has already been in use in the field of diffusion in porous media [Nov+11; GVNL14], but without mention of the underlying stochastic process.

This chapter is laid out as follows. In Section 4.1, we present our main results: partially reflected Brownian motion is defined as the solution to a martingale problem and two constructions of this process are given, we also state the convergence of a class of random walks to partially reflected Brownian motion. In Section 4.2, we prove that the martingale problem which charaterizes partially reflected Brownian motion is well posed and we show that the two constructions in Section 4.1 provide solutions to this martingale problem. Finally in Section 4.3, we prove the convergence in distribution of a sequence of random walks to partially reflected Brownian motion.

4.1 Main results

4.1.1 Definition and constructions of partially reflected Brownian motion

We first give a definition of partially reflected Brownian motion as a solution to a martingale problem on a space consisting of the disjoint union of the positive and negative half lines. We will show in Section 4.2 that this martingale problem is well posed. We then give two constructions of this process.

Definition Let $\ddot{\mathbb{R}}$ be the disjoint union of the positive and negative half lines,

$$\ddot{\mathbb{R}} = (-\infty, 0^-] \cup [0^+, +\infty).$$

It is endowed with the metric d defined by

$$\forall x, y \in \ddot{\mathbb{R}}, \quad d(x, y) = |x - y| + \mathbb{1}_{\{xy \leq 0\}}.$$

Let $\hat{C}(\ddot{\mathbb{R}})$ be the set of continuous real-valued functions on $\ddot{\mathbb{R}}$ which vanish at infinity. For $\gamma \in [0, +\infty]$, let $\mathcal{D}^\gamma$ be the subspace of functions $f \in \hat{C}(\ddot{\mathbb{R}})$ which are twice continuously differentiable on each half line and satisfy

$$\partial_x f(0^-) = \partial_x f(0^+) = \gamma(f(0^+) - f(0^-)). \tag{4.2}$$

(For $\gamma = +\infty$, (4.2) becomes $\partial_x f(0^-) = \partial_x f(0^+)$ and $f(0^-) = f(0^+)$.) Fix $\sigma > 0$ and let us define a linear operator L^γ on $\mathcal{D}^\gamma$ by

$$L^\gamma f = \frac{\sigma^2}{2} \partial_{xx} f, \quad \forall f \in \mathcal{D}^\gamma. \tag{4.3}$$

The operator L^γ is the generator of partially reflected Brownian motion. Let $D(\mathbb{R}_+, \ddot{\mathbb{R}})$ denote the space of càdlàg functions from $\mathbb{R}_+$ to $\ddot{\mathbb{R}}$.

Definition 4.1.1 (partially reflected Brownian motion)**.** *Let $(X_t)_{t\geq 0}$ be a càdlàg, $\ddot{\mathbb{R}}$-valued Markov process on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and call $\mathbb{P}^X$ its law on $D(\mathbb{R}_+, \ddot{\mathbb{R}})$. The process $(X_t)_{t\geq 0}$ (resp. its law $\mathbb{P}^X$) is said to be (the law of) partially reflected Brownian motion if it is a solution to the martingale problem associated with L^γ for some $\gamma \in [0, +\infty]$, i.e. if for any $f \in \mathcal{D}^\gamma$, the process*

$$f(X_t) - \int_0^t L^\gamma f(X_s) ds \tag{4.4}$$

is a martingale with respect to the filtration generated by $(X_t)_{t\geq 0}$. We call γ the permeability of the barrier.

Naturally, we say that $(X_t)_{t\geq 0}$ is partially reflected Brownian motion with initial distribution μ if it is a solution to the martingale problem associated with (L^γ, μ), *i.e.* if (4.4) is a martingale for all $f \in \mathcal{D}^\gamma$ and if $\mathbb{P}^X(X_0 \in \cdot) = \mu(\cdot)$.

This definition does not seem to give much information about possible solutions to this martingale problem. It does not even tell us if such solutions exist or if they are unique (in distribution). This is the object of the next proposition. Below, we also give two ways to construct solutions to this martingale problem.

It should be noted that for $\gamma = 0$ (impermeable barrier), the operator L^γ is the generator of reflected Brownian motion (see for example Exercice VII.1.23 in [RY13] in the case $\alpha = 1$), while for $\gamma = +\infty$ (completely permeable barrier), L^γ is the generator of Brownian motion.

Proposition 4.1.2. *For any $\gamma \in [0, +\infty]$, the martingale problem associated with L^γ has at most one $D(\mathbb{R}_+, \ddot{\mathbb{R}})$ valued solution.*

Proof. The operator L^γ satisfies the positive maximum principle on $\mathcal{D}^\gamma$, *i.e.* whenever $f \in \mathcal{D}^\gamma$ and $\sup_{x\in\ddot{\mathbb{R}}} f(x) = f(x_0) \geq 0$, we have $L^\gamma f(x_0) \leq 0$. By Lemma 4.2.1 of [EK86], L^γ is thus dissipative on $\mathcal{D}^\gamma$ (recall that we ask the functions in $\mathcal{D}^\gamma$ to vanish at infinity).

Let us now show that for any positive λ, the range of $\lambda - L^\gamma$ contains the space $\hat{C}(\ddot{\mathbb{R}})$ of continuous functions vanishing at infinity. We do it in the case $\sigma^2 = 2$, but the general case is similar. Let $f \in \hat{C}(\ddot{\mathbb{R}})$ be such a function and define

$$g(x) = \begin{cases} e^{-\sqrt{\lambda}x} \displaystyle\int_0^x \frac{e^{\sqrt{\lambda}y}}{2\sqrt{\lambda}} f(y)dy + e^{\sqrt{\lambda}x} \int_x^{+\infty} \frac{e^{-\sqrt{\lambda}y}}{2\sqrt{\lambda}} f(y)dy + Ae^{-\sqrt{\lambda}x} & \text{if } x \geq 0^+, \\ e^{-\sqrt{\lambda}x} \displaystyle\int_{-\infty}^x \frac{e^{\sqrt{\lambda}y}}{2\sqrt{\lambda}} f(y)dy + e^{\sqrt{\lambda}x} \int_x^{0} \frac{e^{-\sqrt{\lambda}y}}{2\sqrt{\lambda}} f(y)dy + Be^{\sqrt{\lambda}x} & \text{if } x \leq 0^-, \end{cases}$$

for some $A, B \in \mathbb{R}$. Then g is twice continuously differentiable on $\ddot{\mathbb{R}}$, vanishes at infinity and satisfies

$$\partial_{xx} g(x) = \lambda g(x) - f(x)$$

for all $x \in \ddot{\mathbb{R}}$. The constants A and B can then be chosen so that g also satisfies (4.2). As a result we have found a function g in $\mathcal{D}^\gamma$ such that $\lambda g - L^\gamma g = f$ for any $f \in \hat{C}(\ddot{\mathbb{R}})$.

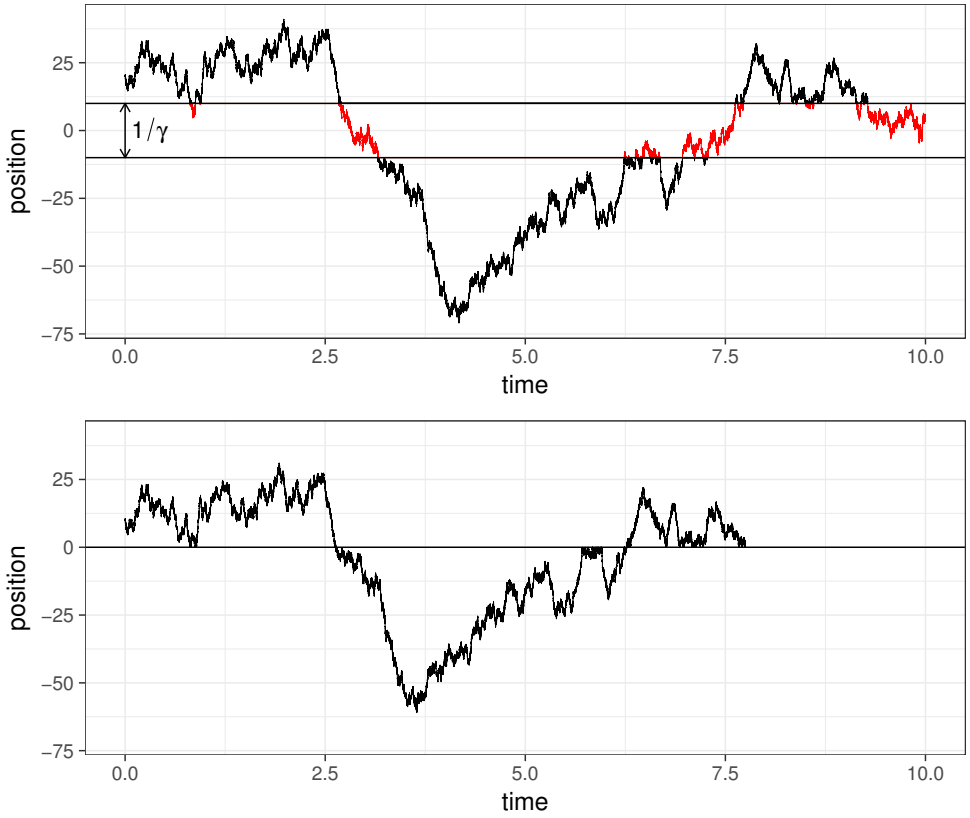

Figure 4.3: Speed and scale construction of partially reflected Brownian motion
Construction of partially reflected Brownian motion as a time-changed Brownian motion, with $\sigma^2 = 400$, $x = 20$ and $\gamma = 0.05$.

In particular, since $\mathcal{D}^\gamma$ is a subset of $\hat{C}(\ddot{\mathbb{R}})$, it is in the range of $\lambda - L^\gamma$ for any $\lambda > 0$. Furthermore $\hat{C}(\ddot{\mathbb{R}})$ is separating in the sense of Section 3.4 in [EK86]. Proposition 4.1.2 then follows from Corollary 4.4.4 in [EK86]. □

"Speed and scale" construction of partially reflected Brownian motion We now present a way to construct partially reflected Brownian motion from Brownian motion, via an analogy with the speed and scale construction of one dimensional diffusions. This will give us a better sense of what "typical" trajectories of this process look like. Indeed, we show that the excursions of partially reflected Brownian motion outside the origin are given by the sequence of excursions of a Brownian motion outside a macroscopic region of length $\frac{1}{\gamma}$, as illustrated in Figure 4.3.

Fix $\gamma \in (0, +\infty)$ and suppose for simplicity that $\sigma^2 = 1$. Define $r : \mathbb{R} \to \ddot{\mathbb{R}}$ by

$$r(x) = \begin{cases} x - \frac{1}{2\gamma} & \text{if } x > \frac{1}{2\gamma}, \\ x + \frac{1}{2\gamma} & \text{if } x < -\frac{1}{2\gamma}, \\ 0^+ & \text{if } 0 \leq x \leq \frac{1}{2\gamma}, \\ 0^- & \text{if } -\frac{1}{2\gamma} \leq x < 0, \end{cases}$$

(see Figure 4.4). Further define $r^{-1} : \ddot{\mathbb{R}} \to \mathbb{R}$ by

$$r^{-1}(x) = x + \text{sign}(x)\frac{1}{2\gamma}.$$

(Note that r^{-1} is only the right inverse of r, *i.e.* $r \circ r^{-1} = Id_{\ddot{\mathbb{R}}}$ but $r^{-1} \circ r \neq Id_{\mathbb{R}}$.) Now fix $x \in \ddot{\mathbb{R}}$ and let $(B_t)_{t\geq 0}$ be standard Brownian motion started from $r^{-1}(x)$. Also set, for $t \geq 0$,

$$\tau(t) = \inf\left\{\tau > 0 : \int_0^\tau \mathbb{1}_{\{|B_s| > \frac{1}{2\gamma}\}} ds > t\right\}. \tag{4.5}$$

Finally, let $X_t = r(B_{\tau(t)})$.

Proposition 4.1.3. *The process $(X_t)_{t\geq 0}$ is partially reflected Brownian motion started from x, i.e. it is a solution to the martingale problem associated with (L^γ, δ_x).*

We prove this result in Subsection 4.2.1. The construction is illustrated in Figure 4.3. In words, we map the two intervals $(-\infty, -\frac{1}{2\gamma}]$ and $[\frac{1}{2\gamma}, +\infty)$ onto $(-\infty, 0^-]$ and $[0^+, +\infty)$, and we change time in order to drop the time intervals where $|B_s| \leq \frac{1}{2\gamma}$.

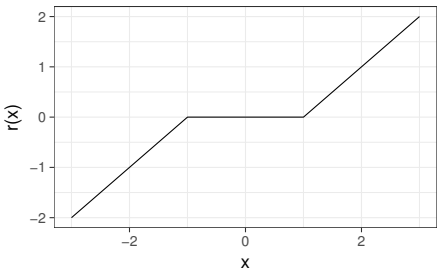

Figure 4.4: Graph of the function $r : \mathbb{R} \to \ddot{\mathbb{R}}$ for $\gamma = 0.5$.

Remark. *From this construction, we recover the property stated by Nagylaki [Nag88, Equation 56] that, for $0 < x < y$,*

$$\mathbb{P}_x\left(X_t \text{ reaches } 0^- \text{ before } y\right) = \frac{y - x}{y + \frac{1}{\gamma}}.$$

Corollary 4.1.4. *For any $\gamma \in [0, +\infty]$, the martingale problem associated to L^γ is well posed, i.e. it has a unique solution.*

Let $\ddot{\pi} : \ddot{\mathbb{R}} \to \mathbb{R}$ be the natural projection of $\ddot{\mathbb{R}}$ onto $\mathbb{R}$ (*i.e.* mapping both 0^+ and 0^- onto 0). We sometimes also call the projection $(\ddot{\pi}(X_t))_{t\geq 0}$ partially reflected Brownian motion, even though the latter isn't a Markov process. (For example, as we shall see below, the sequence of random walks considered in Subsection 4.1.2 converges to the projection of partially reflected Brownian motion.)

Construction involving the local time at the origin From the previous construction, one is led to think that $(|X_t|)_{t\geq 0}$ has the law of reflected Brownian motion. It is then natural to ask if partially reflected Brownian motion can be constructed by randomly "flipping" the excursions of reflected Brownian motion. The next proposition provides such a construction. It turns out that the crossing times of the origin are the times at which the local time at the origin of the process reaches the levels of an independent Poisson process with parameter γ, see Figure 4.5.

Fix $x \in \ddot{\mathbb{R}}$ and let $(W_t)_{t\geq 0}$ be reflected Brownian motion on $\mathbb{R}_+$ started from $|x|$. Let $(N(t), t \geq 0)$ be a Poisson process with rate $\gamma \in (0, \infty)$, independent of $(W_t)_{t\geq 0}$. Let $L_t^0(W)$ denote the local time accumulated at the origin up to time t by W. Set

$$X_t = \text{sign}(x)(-1)^{N(L_t^0(W))} W_t,$$

where $\pm 1 \times 0 = 0^\pm$ (see Figure 4.5).

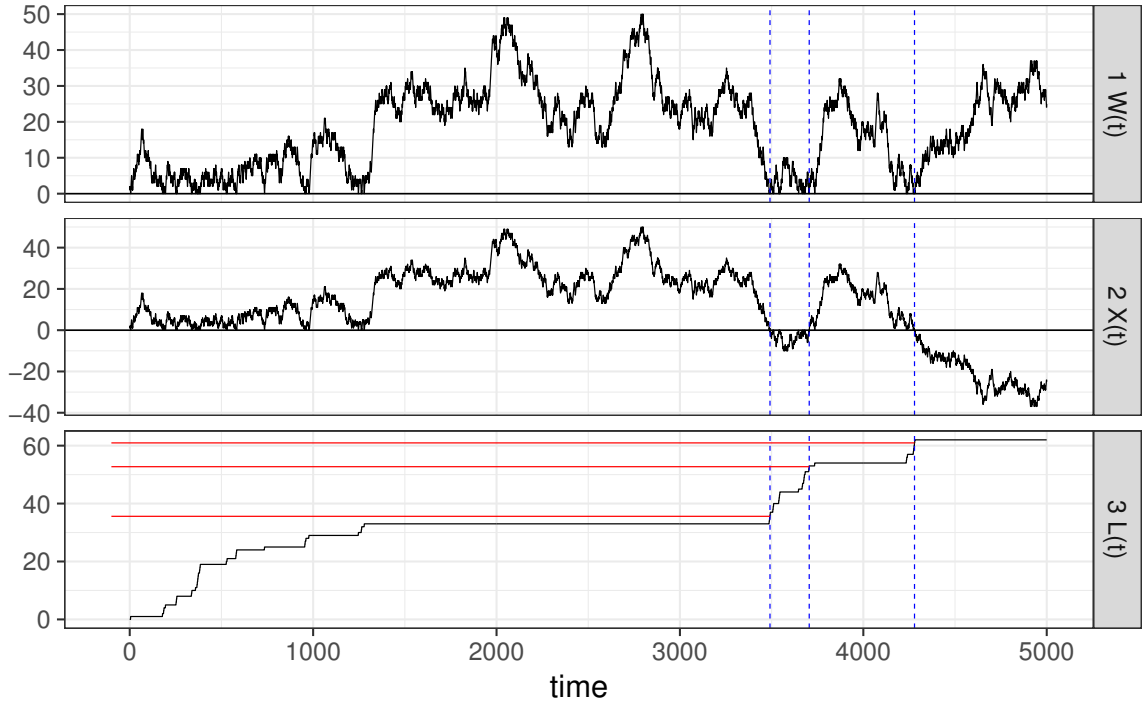

Figure 4.5: Construction of partially reflected Brownian motion involving the local time at the origin Top graphic shows a realisation of reflected Brownian motion W_t. Bottom graphic shows its local time accumulated at the origin $L(t)$. The heights of horizontal red lines are drawn according to a Poisson process on the y axis. The graphic in the middle is obtained by "flipping" W_t at the times when $L(t)$ reaches the red lines, and is distributed as the projection of partially reflected Brownian motion.

Proposition 4.1.5. *The process $(X_t)_{t\geq 0}$ is partially reflected Brownian motion started from x.*

We prove this result in Subsection 4.2.2 using the previous construction and a Ray-Knight theorem [SK91, Theorem 6.4.7], which states that the local time accumulated by Brownian motion at $\frac{1}{2\gamma}$ before reaching $-\frac{1}{2\gamma}$ is an exponential random variable with parameter γ.

Proposition 4.1.5 yields an explicit formula for the transition density of partially reflected Brownian motion. For $t > 0$ and $x \in \mathbb{R}$, set

$$G_t(x) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{x^2}{2t}\right).$$

Corollary 4.1.6. *If $(X_t)_{t\geq 0}$ is partially reflected Brownian motion with permeability $\gamma \in (0,\infty)$ started from $x \in \ddot{\mathbb{R}}$, then $\mathbb{P}_x(X_t \in dy) = g_t(x,y)dy$ with*

$$g_t(x,y) = \begin{cases} G_t(x-y) + G_t(x+y) - 2\gamma \int_0^{+\infty} e^{-2\gamma l} G_t(|x|+|y|+l)dl & \text{if } xy \geq 0^+, \\ 2\gamma \int_0^{+\infty} e^{-2\gamma l} G_t(|x|+|y|+l)dl & \text{if } xy \leq 0^-. \end{cases}$$

We derive this formula in Subsection 4.2.4.

4.1.2 Scaling limits of a class of random walks

Let us now state the main convergence result, namely that partially reflected Brownian motion is the scaling limit of a class of random walks with an obstacle. We consider a more general case than [Nag76], where the barrier to gene flow has width $K \in \mathbb{N}^*$ (K being the number of edges with reduced migration rate). The cases $K = 1$ (the one considered in [Nag76]) and $K = 2$ are illustrated in Figure 4.6. We define the process describing the motion of an ancestral lineage as follows.

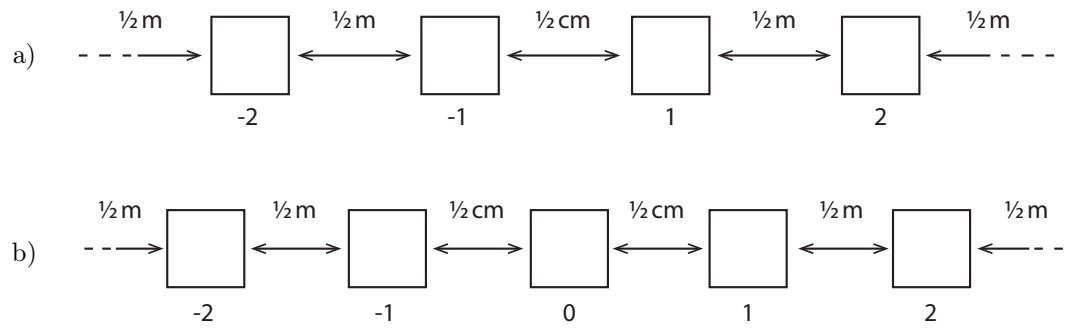

Figure 4.6: Jump rates of random walks with an obstacle
Transition rates of the random walk $(\xi_n(t), t \geq 0)$ in a) case $K = 1$ and b) case $K = 2$.

Definition 4.1.7 (Random walk with an obstacle)**.** *Let $(c_n)_{n\geq 1}$ be a sequence of positive real numbers, and fix $m > 0$. Suppose that $\left(x_n^0\right)_{n\geq 1}$ is a sequence of elements of $\mathbb{N}^*$.*

If K is even, *let $E = \mathbb{Z}$ and define, for $i, j \in E$,*

$$q_n(i,j) = \begin{cases} \frac{m}{2} & \text{if } |i-j| = 1 \text{ and } |i| \vee |j| > \frac{K}{2}, \\ c_n \frac{m}{2} & \text{if } |i-j| = 1 \text{ and } |i| \vee |j| \leq \frac{K}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.6}$$

If K is odd, *let $E = \mathbb{Z} \setminus \{0\}$ and set*

$$q_n(i,j) = \begin{cases} \frac{m}{2} & \text{if } |i-j| = 1 \text{ and } |i| \vee |j| > \frac{K+1}{2}, \\ c_n \frac{m}{2} & \text{if } |i-j| = 1 \text{ and } |i| \vee |j| \leq \frac{K+1}{2}, \\ & \text{or if } \{i,j\} = \{+1,-1\}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.7}$$

Then let $(\xi_n(t), t \geq 0)$ be a continuous time random walk on E started from x_n^0 with jump rates $q_n(\cdot,\cdot)$.

For $n \geq 1$, set $X_n(t) = \frac{1}{\sqrt{n}}\xi_n(nt)$. We now state conditions under which the rescaled random walk X_n converges to partially reflected Brownian motion. We equip $\mathrm{D}(\mathbb{R}_+, \mathbb{R})$ with the topology of Skorokhod convergence on compact time intervals. If $d_{sko}^T(\cdot,\cdot)$ is a metric for the Skorokhod

convergence on $\mathrm{D}([0,T],\mathbb{R})$ for $T>0$, then

$$d(f,g)=\int_0^{+\infty} e^{-t}(d^t_{sko}(f,g)\wedge 1)dt \tag{4.8}$$

is a metric for Skorokhod convergence on compact time intervals.

Theorem 4.1. *Suppose* $\frac{1}{\sqrt{n}}x^0_n \underset{n\to\infty}{\longrightarrow} x^0$ *with* $x^0\neq 0$ *and* $\lim_{n\to\infty}\frac{\sqrt{n}}{K}c_n=\gamma\in[0,+\infty]$. *Then as* $n\to\infty$, *the sequence of real-valued processes* $(X_n(t),t\geq 0)$ *converges in distribution in* $(\mathrm{D}(\mathbb{R}_+,\mathbb{R}),d)$ *to a continuous real-valued process* $(X_t)_{t\geq 0}$ *which is (a projection of) a solution to the martingale problem associated with* (L^γ,δ_{x^0}), *with* $\sigma^2=m$.

In other words, if $\sqrt{n}c_n\to+\infty$, X_n converges to Brownian motion, if $\sqrt{n}c_n\to 0$, X_n converges to reflected Brownian motion, while if $\frac{\sqrt{n}}{K}c_n\to\gamma\in(0,\infty)$, X_n converges to (the projection of) partially reflected Brownian motion.

Remark. *In the case* $x^0=0$, *the convergence still holds provided the probability of first exiting the set* $[-K/2,K/2]$ *on the right converges as* $n\to\infty$. *The initial distribution is then a convex combination of* δ_{0^+} *and* δ_{0^-}, *given by the exit probabilities.*

Theorem 4.1 is proved in Section 4.3 in the case $K=2$. The generalisation to other values of K is straighforward, and the case $K=1$ introduces some simplifications, which makes the case $K=2$ more representative of the general case.

Note that in Nagylaki's model presented in Figure 4.1, ancestral lineages are distributed as the random walk of Definition 4.1.7 with $K=1$.

4.2 Constructions of partially reflected Brownian motion

4.2.1 Speed and scale construction

Here we prove that the process $X_t=r(B_{\tau(t)})$ defined in Subsection 4.1.1 is a solution to the martingale problem associated with L^γ. This proof will require the following lemma, proved in Subsection 4.2.3.

Lemma 4.2.1. *Set* $W_t=|X_t|$. *Then* $(W_t)_{t\geq 0}$ *is distributed as reflected Brownian motion.*

Proof of Proposition 4.1.3. Recall that B is standard Brownian motion started at $r^{-1}(x)$, hence $X_0=x$ almost surely. Let $\left(\mathcal{F}^B_t\right)_{t\geq 0}$ denote the natural filtration of $(B_t)_{t\geq 0}$, and let $\mathcal{F}_t=\mathcal{F}^B_{\tau(t)}$. Then $(\mathcal{F}_t)_{t\geq 0}$ is a filtration, $(X_t)_{t\geq 0}$ is $(\mathcal{F}_t)_{t\geq 0}$ adapted and, for $s,t\geq 0$ and $f:\ddot{\mathbb{R}}\to\mathbb{R}$ bounded and continuous,

$$\mathbb{E}\left[f(X_{t+s})\mid\mathcal{F}_t\right]=\mathbb{E}_{r^{-1}(X_t)}\left[f(r(B_{\tau(s)}))\right].$$

Now let $\left(\mathcal{F}^X_t\right)_{t\geq 0}$ be the filtration generated by $(X_t)_{t\geq 0}$. Since $\mathcal{F}^X_t\subset\mathcal{F}_t$ for $t\geq 0$,

$$\begin{aligned}\mathbb{E}\left[f(X_{t+s})\mid\mathcal{F}^X_t\right]&=\mathbb{E}\left[\mathbb{E}\left[f(X_{t+s})\mid\mathcal{F}_t\right]\mid\mathcal{F}^X_t\right]\\&=\mathbb{E}\left[\mathbb{E}_{r^{-1}(X_t)}\left[f(r(B_{\tau(s)}))\right]\mid\mathcal{F}^X_t\right]\\&=\mathbb{E}_{r^{-1}(X_t)}\left[f(r(B_{\tau(s)}))\right].\end{aligned}$$

In other words, $(X_t)_{t\geq 0}$ is a Markov process.

Suppose now that for any $x \in \ddot{\mathbb{R}}$ and $f \in \mathcal{D}^\gamma$,

$$\lim_{t\downarrow 0} \frac{1}{t}\mathbb{E}_x\left[f(X_t) - f(X_0)\right] = \frac{1}{2}\partial_{xx} f(x). \tag{4.9}$$

(Recall that we assumed $\sigma^2 = 1$ for simplicity.) Then, by Proposition 4.1.7 in [EK86], (4.4) is an $\mathcal{F}^X$-martingale for all $f \in \mathcal{D}^\gamma$ (X is progressive since it is right-continuous). It follows that $(X_t)_{t\geq 0}$ is a solution to the martingale problem associated with L^γ.

Let us now show (4.9). Since X behaves as standard Brownian motion until the first time it hits the origin, (4.9) clearly holds for all $x \in \ddot{\mathbb{R}} \setminus \{0^+, 0^-\}$. By symmetry, we can restrict the proof to $x = 0^+$. For any $t \geq 0$,

$$\begin{aligned}\mathbb{E}_{0^+}\left[f(X_t)\right] = \mathbb{E}_{0^+}\left[\left(f(X_t) - f(0^+)\right)\mathbb{1}_{\{X_t\geq 0^+\}} + \left(f(X_t) - f(0^-)\right)\mathbb{1}_{\{X_t\leq 0^-\}}\right] \\ + \mathbb{E}_{0^+}\left[f(0^+)\mathbb{1}_{\{X_t\geq 0^+\}} + f(0^-)\mathbb{1}_{\{X_t\leq 0^-\}}\right].\end{aligned}$$

Subtracting $f(0^+)$ on both sides we obtain

$$\begin{aligned}\mathbb{E}_{0^+}\left[f(X_t) - f(0^+)\right] = \mathbb{E}_{0^+}\left[\left(f(X_t) - f(0^+)\right)\mathbb{1}_{\{X_t\geq 0^+\}} + \left(f(X_t) - f(0^-)\right)\mathbb{1}_{\{X_t\leq 0^-\}}\right] \\ + \mathbb{E}_{0^+}\left[\left(f(0^-) - f(0^+)\right)\mathbb{1}_{\{X_t\leq 0^-\}}\right].\end{aligned} \tag{4.10}$$

Since f is twice continuously differentiable on $[0^+, +\infty)$, for any y in $[0^+, +\infty)$, there exists $h(y) \in [0^+, y]$ such that

$$f(y) - f(0^+) = \partial_x f(0^+)y + \frac{1}{2}\partial_{xx} f(h(y))y^2.$$

Replacing y by X_t, we write, for any $r > 0$,

$$\begin{aligned}(f(X_t) - f(0^+))\mathbb{1}_{\{X_t\geq 0^+\}} = \left(\partial_x f(0^+)X_t + \frac{1}{2}\partial_{xx} f(h(X_t))X_t^2\right)\mathbb{1}_{\{0^+\leq X_t\leq r\}} \\ + (f(X_t) - f(0^+))\mathbb{1}_{\{X_t > r\}}.\end{aligned}$$

By the Markov inequality and Lemma 4.2.1,

$$\mathbb{P}\left(|X_t| > r\right) \leq 3\frac{t^2}{r^4}. \tag{4.11}$$

As a result, since f is bounded,

$$\mathbb{E}_{0^+}\left[(f(X_t) - f(0^+))\mathbb{1}_{\{X_t > r\}}\right] \leq 6\left\|f\right\|_\infty \frac{t^2}{r^4}. \tag{4.12}$$

In addition,

$$\mathbb{E}_{0^+}\left[\partial_x f(0^+)X_t\mathbb{1}_{\{0^+\leq X_t\leq r\}}\right] = \partial_x f(0^+)\mathbb{E}_{0^+}\left[X_t\mathbb{1}_{\{X_t\geq 0^+\}}\right] - \partial_x f(0^+)\mathbb{E}_{0^+}\left[X_t\mathbb{1}_{\{X_t > r\}}\right], \tag{4.13}$$

and by Cauchy-Schwartz inequality, Lemma 4.2.1 and (4.11),

$$\begin{aligned}\left|\mathbb{E}_{0^+}\left[X_t\mathbb{1}_{\{X_t>r\}}\right]\right| &\leq \mathbb{E}_{0^+}\left[X_t^2\right]^{1/2}\mathbb{P}_{0^+}\left(X_t>r\right)^{1/2}\\ &\leq t^{1/2}\sqrt{3}\frac{t}{r^2}.\end{aligned} \tag{4.14}$$

Moreover, since $\partial_{xx}f$ is continuous on $[0^+,+\infty)$, it is uniformly continuous on compact sets and there exists $C_r>0$ such that

$$\forall x,y\in[0^+,r],\qquad \left|\partial_{xx}f(y)-\partial_{xx}f(x)\right|\leq C_r\left|x-y\right|.$$

As a result,

$$\left|\mathbb{E}_{0^+}\left[\partial_{xx}f(h(X_t))X_t^2\mathbb{1}_{\{0^+\leq X_t\leq r\}}\right]-\mathbb{E}_{0^+}\left[\partial_{xx}f(0^+)X_t^2\mathbb{1}_{\{0\leq X_t\leq r\}}\right]\right|\leq C_r\mathbb{E}_{0^+}\left[\left|X_t\right|^3\right], \tag{4.15}$$

and by Lemma 4.2.1, $\mathbb{E}_{0^+}\left[\left|X_t\right|^3\right]=\mathcal{O}\left(t^{3/2}\right)$. Proceeding as for (4.14), we also have

$$\mathbb{E}_{0^+}\left[\frac{1}{2}\partial_{xx}f(0^+)X_t^2\mathbb{1}_{\{0^+\leq X_t\leq r\}}\right]=\frac{1}{2}\partial_{xx}f(0^+)\mathbb{E}_{0^+}\left[X_t^2\mathbb{1}_{\{X_t\geq 0^+\}}\right]+\mathcal{O}\left(t^{3/2}\right). \tag{4.16}$$

Putting together (4.13), (4.15) and (4.16), we obtain

$$\mathbb{E}_{0^+}\left[(f(X_t)-f(0^+))\mathbb{1}_{\{X_t\geq 0^+\}}\right]=\partial_x f(0^+)\mathbb{E}_{0^+}\left[X_t\mathbb{1}_{\{X_t\geq 0^+\}}\right]+\frac{1}{2}\partial_{xx}f(0^+)\mathbb{E}_{0^+}\left[X_t^2\mathbb{1}_{\{X_t\geq 0^+\}}\right]+o\left(t\right).$$

Likewise, we have

$$\mathbb{E}_{0^+}\left[(f(X_t)-f(0^-))\mathbb{1}_{\{X_t\leq 0^-\}}\right]=\partial_x f(0^-)\mathbb{E}_{0^+}\left[X_t\mathbb{1}_{\{X_t\leq 0^-\}}\right]+\frac{1}{2}\partial_{xx}f(0^-)\mathbb{E}_{0^+}\left[X_t^2\mathbb{1}_{\{X_t\leq 0^-\}}\right]+o\left(t\right).$$

Plugging these two equations in (4.10) and using the fact that $\partial_x f(0^-)=\partial_x f(0^+)$, we obtain

$$\begin{aligned}\mathbb{E}_{0^+}\left[f(X_t)-f(X_0)\right]=\partial_x f(0^\pm)\mathbb{E}_{0^+}\left[X_t\right]&+\frac{1}{2}\partial_{xx}f(0^+)\mathbb{E}_{0^+}\left[X_t^2\right]\\ &+\frac{1}{2}\left(\partial_{xx}f(0^-)-\partial_{xx}f(0^+)\right)\mathbb{E}_{0^+}\left[X_t^2\mathbb{1}_{\{X_t\leq 0^-\}}\right]\\ &+\left(f(0^-)-f(0^+)\right)\mathbb{P}_{0^+}\left(X_t\leq 0^-\right)+o\left(t\right).\end{aligned} \tag{4.17}$$

Moreover, by the construction of X_t,

$$\begin{aligned}\mathbb{E}_{0^+}\left[X_t\right]&=\mathbb{E}_{\frac{1}{2\gamma}}\left[r(B_{\tau(t)})\right]\\ &=\mathbb{E}_{\frac{1}{2\gamma}}\left[\left(B_{\tau(t)}-\frac{1}{2\gamma}\right)\mathbb{1}_{\left\{B_{\tau(t)}\geq\frac{1}{2\gamma}\right\}}+\left(B_{\tau(t)}+\frac{1}{2\gamma}\right)\mathbb{1}_{\left\{B_{\tau(t)}\leq-\frac{1}{2\gamma}\right\}}\right]\\ &=\mathbb{E}_{\frac{1}{2\gamma}}\left[B_{\tau(t)}\right]+\frac{1}{2\gamma}\mathbb{E}_{\frac{1}{2\gamma}}\left[\mathbb{1}_{\left\{B_{\tau(t)}\leq-\frac{1}{2\gamma}\right\}}-\mathbb{1}_{\left\{B_{\tau(t)}\geq\frac{1}{2\gamma}\right\}}\right].\end{aligned} \tag{4.18}$$

Note that $\tau(t)$ is an $\mathcal{F}_t^B$-stopping time. Furthermore, for any given $t\geq 0$, the martingale $(B_{s\wedge\tau(t)},s\geq$

0) is uniformly integrable. To see this, write

$$\sup_{s\geq 0}\left|B_{s\wedge\tau(t)}\right| \leq \frac{1}{2\gamma} + \sup_{0\leq s\leq t} W_s,$$

and note that the right-hand-side is integrable by Lemma 4.2.1 and Doob's maximal inequality. Hence, by the Optional Stopping Theorem, $\mathbb{E}_{\frac{1}{2\gamma}}\left[B_{\tau(t)}\right] = \frac{1}{2\gamma}$. As a result, returning to (4.18),

$$\mathbb{E}_{0^+}\left[X_t\right] = \frac{1}{\gamma}\mathbb{P}_{0^+}\left(X_t \leq 0^-\right).$$

Since $f \in \mathcal{D}^\gamma$, the first term in (4.17) cancels with the last one. By Lemma 4.2.1, $\mathbb{E}_{0^+}\left[X_t^2\right] = t$. Also note that by the Cauchy-Schwarz inequality,

$$\mathbb{E}_{0^+}\left[X_t^2 \mathbb{1}_{\{X_t\leq 0^-\}}\right] \leq \sqrt{3}\, t\, \mathbb{P}_{0^+}\left(X_t \leq 0^-\right)^{1/2}.$$

(We have used Lemma 4.2.1 to compute the fourth moment of X_t.) Furthermore, $\mathbb{P}_{0^+}\left(X_t \leq 0^-\right) = \mathbb{P}_{\frac{1}{2\gamma}}\left(B_{\tau(t)} \leq -\frac{1}{2\gamma}\right) \underset{t\to 0}{\longrightarrow} 0$.

Coming back to (4.17), dividing both sides by t and letting $t \downarrow 0$, we obtain

$$\lim_{t\downarrow 0}\frac{1}{t}\mathbb{E}_{0^+}\left[f(X_t) - f(X_0)\right] = \frac{1}{2}\partial_{xx} f(0^+).$$

The proof of Proposition 4.1.3 is now complete. □

4.2.2 Construction involving the local time at the origin

Proof of Proposition 4.1.5. Let $(B_t)_{t\geq 0}$ be standard Brownian motion and let $X_t = r(B_{\tau(t)})$ be partially reflected Brownian motion constructed as before. Set $W_t = |X_t|$ and

$$T_0 = 0, \qquad T_{i+1} = \inf\{t > T_i : X_{T_i} X_t < 0\}, \quad i \geq 0.$$

For $i \geq 1$ set

$$E_i = L^0_{T_i}(X) - L^0_{T_{i-1}}(X)$$

and for $t \geq 0$,

$$N(t) = \max\left\{n \in \mathbb{N} : \sum_{i=0}^{n} E_i \leq t\right\}.$$

Then, for all $t \geq 0$,

$$X_t = \operatorname{sign}(X_0)(-1)^{N(L^0_t(W))} W_t.$$

We know from Lemma 4.2.1 that $(W_t)_{t\geq 0}$ is distributed as reflected Brownian motion. Proposition 4.1.5 will be proven if we show that $N(t)$ is a Poisson process with rate γ and that it is independent of $(W_t)_{t\geq 0}$.

Note that the left (resp. right) local time accumulated by X at the origin up to time t is the local time accumulated by B at $-\frac{1}{2\gamma}$ (resp. $\frac{1}{2\gamma}$) up to time $\tau(t)$. Indeed, by the Tanaka formula [RY13,

Theorem VI.1.2] (see also (1.51)), letting $x^+ = \max(x, 0)$,

$$\frac{1}{2}L_t^{0^+}(X) = X_t^+ - X_0^+ - \int_0^t \mathbb{1}_{\{X_s>0\}} dX_s$$

and

$$\frac{1}{2}L_{\tau(t)}^{1/2\gamma}(B) = (B_{\tau(t)} - \frac{1}{2\gamma})^+ - (B_0 - \frac{1}{2\gamma})^+ - \int_0^{\tau(t)} \mathbb{1}_{\left\{B_s>\frac{1}{2\gamma}\right\}} dB_s.$$

(For Brownian motion, considering the right, the left or the symmetric local time makes no difference.) By the construction of X, $X_t^+ = (B_{\tau(t)} - \frac{1}{2\gamma})^+$ and, since $\mathbb{1}_{\left\{B_s>\frac{1}{2\gamma}\right\}} = 0$ when $s \in (\tau(t^-), \tau(t))$,

$$\int_0^{\tau(t)} \mathbb{1}_{\left\{B_s>\frac{1}{2\gamma}\right\}} dB_s = \int_0^t \mathbb{1}_{\{X_s>0\}} dX_s.$$

As a result,

$$L_t^{0^+}(X) = L_{\tau(t)}^{1/2\gamma}(B) \tag{4.19}$$

and likewise, $L_t^{0^-}(X) = L_{\tau(t)}^{-1/2\gamma}(B)$.

For $a \in \mathbb{R}$, set $\mathbb{T}_a = \inf\{t > 0 : B_t = a\}$. Assuming without loss of generality that $X_0 > 0$, $\tau(T_1) = \mathbb{T}_{-1/2\gamma}$. Then,

$$\begin{aligned} E_1 = L_{T_1}^0(X) &= \frac{1}{2}\left(L_{T_1}^{0^+}(X) + L_{T_1}^{0^-}(X)\right) \\ &= \frac{1}{2}L_{\mathbb{T}_{-1/2\gamma}}^{1/2\gamma}(B). \end{aligned}$$

By the Ray-Knight theorem [SK91, Theorem 6.4.7],

$$L_{\mathbb{T}_{-1/2\gamma}}^{1/2\gamma}(B)$$

is an exponential random variable with parameter $\frac{\gamma}{2}$. Hence E_1 is exponential with parameter γ.

Further, the strong Markov property of $(B_t)_{t\geq 0}$ and its symmetry imply that the E_i are independent and identically distributed. As a result $(N(t), t \geq 0)$ is a Poisson process with rate γ. It remains to show that it is independent of $(W_t)_{t\geq 0}$.

Set

$$\theta(t) = \inf\left\{\theta > 0 : \int_0^\theta \mathbb{1}_{\left\{|B_s|\leq\frac{1}{2\gamma}\right\}} ds > t\right\}$$

and define

$$S_0 = 0, \qquad S_i = \inf\left\{t > S_{i-1} : B_{\theta(t)} = \frac{(-1)^i}{2\gamma}\right\}.$$

By the same argument as above,

$$L_{T_i}^0(W) = L_{S_i}^{1/2\gamma}(B_\theta) + L_{S_i}^{-1/2\gamma}(B_\theta).$$

As a result, the E_i, and $(N(t), t \geq 0)$, are measurable with respect to the sigma field generated by $(B_\theta(t), t \geq 0)$. We prove the following in Subsection 4.2.3.

Lemma 4.2.2. *The processes* $(\left|B_{\tau(t)}\right|, t \geq 0)$ *and* $(B_{\theta(t)}, t \geq 0)$ *are independent.*

Since $W_t = \left|B_{\tau(t)}\right| - \frac{1}{2\gamma}$, this concludes the proof of Proposition 4.1.5. □

4.2.3 The absolute value of partially reflected Brownian motion

For $t \geq 0$, set

$$I^1(t) = \int_0^t \mathbb{1}_{\{B_s > \frac{1}{2\gamma}\}} dB_s - \int_0^t \mathbb{1}_{\{B_s < -\frac{1}{2\gamma}\}} dB_s,$$
$$I^2(t) = \int_0^t \mathbb{1}_{\{|B_s| \leq \frac{1}{2\gamma}\}} dB_s.$$

Both I^1 and I^2 are continuous $\mathcal{F}_t^B$ martingales with

$$\left\langle I^1 \right\rangle_t = \int_0^t \mathbb{1}_{\{|B_s| > \frac{1}{2\gamma}\}} ds$$
$$\left\langle I^2 \right\rangle_t = \int_0^t \mathbb{1}_{\{|B_s| \leq \frac{1}{2\gamma}\}} ds$$
$$\left\langle I^1, I^2 \right\rangle_t = 0.$$

By F. B. Knight's theorem [Kni71] (also Theorem 3.4.13 in [SK91]), the processes

$$\tilde{B}_t^1 = W_0 + I^1(\tau(t)), \qquad \tilde{B}_t^2 = B_{\theta(0)} + I^2(\theta(t)),$$

are independent standard Brownian motions.

Proof of Lemma 4.2.1. By the Tanaka formula [RY13, Theorem VI.1.2],

$$\frac{1}{2} L_t^{1/2\gamma}(B) = \left(B_t - \tfrac{1}{2\gamma}\right)^+ - \left(B_0 - \tfrac{1}{2\gamma}\right)^+ - \int_0^t \mathbb{1}_{\{B_s > \frac{1}{2\gamma}\}} dB_s, \tag{4.20}$$
$$\frac{1}{2} L_t^{-1/2\gamma}(B) = \left(B_t + \tfrac{1}{2\gamma}\right)^- - \left(B_0 + \tfrac{1}{2\gamma}\right)^- + \int_0^t \mathbb{1}_{\{B_s < -\frac{1}{2\gamma}\}} dB_s. \tag{4.21}$$

On the other hand, from the construction of X_t,

$$W_t = |X_t| = (B_{\tau(t)} - \frac{1}{2\gamma})^+ + (B_{\tau(t)} + \frac{1}{2\gamma})^-$$

and from (4.19),

$$L_t^0(W) = L_t^0(X) = \frac{1}{2}\left(L_{\tau(t)}^{1/2\gamma}(B) + L_{\tau(t)}^{-1/2\gamma}(B)\right).$$

Adding (4.20) and (4.21) and replacing t by $\tau(t)$, we obtain

$$\tilde{B}_t^1 = W_t - L_t^0(W).$$

Since $\tilde{B}^1$ is standard Brownian motion, W is reflected Brownian motion [RY13, p. VI.2]. □

Proof of Lemma 4.2.2. Since $\tilde{B}^1_t = W_t - L^0_t(W)$, $t \mapsto L^0_t(W)$ is a solution of the Skorokhod problem for $t \mapsto \tilde{B}^1_t$ (see [SK91, Lemma 3.6.14] and Lemma 1.5.2), and

$$W_t = \tilde{B}^1_t + \inf_{s \le t} (\tilde{B}^1_s)^-.$$

On the other hand, $B_{\theta(t)}$ is a function of $(\tilde{B}^2_t, t \ge 0)$. To see this, note that since $B_{\theta(t)} \in [-1/2\gamma, 1/2\gamma]$,

$$B_{\theta(t)} = \left(B_{\theta(t)} + \frac{1}{2\gamma}\right)^+ - \left(B_{\theta(t)} - \frac{1}{2\gamma}\right)^+ - \frac{1}{2\gamma}.$$

By the Tanaka formula,

$$\left(B_t + \frac{1}{2\gamma}\right)^+ = \left(B_0 + \frac{1}{2\gamma}\right)^+ + \int_0^t \mathbb{1}_{\{B_s > -\frac{1}{2\gamma}\}} dB_s + L_t^{-\frac{1}{2\gamma}}(B),$$
$$\left(B_t - \frac{1}{2\gamma}\right)^+ = \left(B_0 - \frac{1}{2\gamma}\right)^+ + \int_0^t \mathbb{1}_{\{B_s > \frac{1}{2\gamma}\}} dB_s + L_t^{\frac{1}{2\gamma}}(B).$$

Subtracting these equations with t replaced by $\theta(t)$, and noting that $\mathbb{1}_{\{B_s \ge -1/2\gamma\}} - \mathbb{1}_{\{B_s > 1/2\gamma\}} = \mathbb{1}_{\{|B_s| \le 1/2\gamma\}}$, we obtain

$$B_{\theta(t)} = \tilde{B}^2_t + L^{-\frac{1}{2\gamma}}_{\theta(t)}(B) - L^{\frac{1}{2\gamma}}_{\theta(t)}(B).$$

From this equation, we see that $(L^{-\frac{1}{2\gamma}}_{\theta(\cdot)}(B), L^{\frac{1}{2\gamma}}_{\theta(\cdot)}(B))$ is a solution of the two-sided Skorokhod equation for $\tilde{B}^2$ with reflection at $\pm 1/2\gamma$. As a result, $(B_{\theta(t)}, t \ge 0)$ is determined by $(\tilde{B}^2_t, t \ge 0)$ [Har85, p. 2.4.6] (the pair $(L^{-\frac{1}{2\gamma}}_{\theta(\cdot)}(B), L^{\frac{1}{2\gamma}}_{\theta(\cdot)}(B))$ is called the two-sided regulator of $\tilde{B}^2$ in this book).

Since $\left|B_{\tau(t)}\right| = W_t + \frac{1}{2\gamma}$ is a function of $\tilde{B}^1$, $B_{\theta(t)}$ is a function of $\tilde{B}^2$, and $\tilde{B}^1$ is independent of $\tilde{B}^2$, $\left(\left|B_{\tau(t)}\right|, t \ge 0\right)$ and $\left(B_{\theta(t)}, t \ge 0\right)$ are independent. □

4.2.4 Transition density of partially reflected Brownian motion

Proof of Corollary 4.1.6. Recall that X_t was defined as

$$X_t = \operatorname{sign}(x)(-1)^{N(L^0_t(W))} W_t,$$

where W is reflected Brownian motion started from $|x|$ and $(N_t)_{t \ge 0}$ is an independent Poisson process with rate γ. Hence, summing over all possible values of $L^0_t(W)$,

$$\mathbb{P}_x\left(X_t \in dy\right) = \int_0^\infty \mathbb{P}\left(N(l) \overset{2}{=} \operatorname{sign}(x) - \operatorname{sign}(y)\right) \mathbb{P}_{|x|}\left(W_t \in d\left|y\right|, L^0_t(W) \in dl\right),$$

where $x \overset{2}{=} y$ means that x and y have the same parity. Since $N(l)$ is a Poisson random variable with parameter γl,

$$\mathbb{P}\left(N(l) \overset{2}{=} 0\right) = \frac{1 + e^{-2\gamma l}}{2}, \qquad \mathbb{P}\left(N(l) \overset{2}{=} 1\right) = \frac{1 - e^{-2\gamma l}}{2}.$$

In addition [SK91, Problem 6.3.4], for $x, y \geq 0$,

$$\mathbb{P}_x\left(W_t \in dy, L_t^0(W) \in dl\right) = \left(G_t(x-y) - G_t(x+y)\right) dy\delta_0(dl) - 2\partial_x G_t(x+y+l)dydl.$$

As a result, if $xy \geq 0^+$,

$$\mathbb{P}_x\left(X_t \in dy\right) = \left(G_t(x-y) - G_t(x+y)\right) dy - 2\int_0^\infty \frac{1+e^{-2\gamma l}}{2}\partial_x G_t(|x|+|y|+l)dldy.$$

Integrating by parts yields

$$\frac{\mathbb{P}_x\left(X_t \in dy\right)}{dy} = G_t(x-y) + G_t(x+y) - 2\gamma\int_0^\infty e^{-2\gamma l} G_t(|x|+|y|+l)dl.$$

Likewise if $xy \leq 0^-$,

$$\begin{aligned}\frac{\mathbb{P}_x\left(X_t \in dy\right)}{dy} &= -2\int_0^\infty \frac{1-e^{-2\gamma l}}{2}\partial_x G_t(|x|+|y|+l)dl \\ &= 2\gamma\int_0^\infty e^{-2\gamma l} G_t(|x|+|y|+l)dl.\end{aligned}$$

The proof of Corollary 4.1.6 is now complete. □

4.3 Scaling limit of random walks with a barrier

Here, we prove the convergence of the sequence of random walks defined in Subsection 4.1.2 to partially reflected Brownian motion (Theorem 4.1), in the case $K = 2$ and $\gamma \in (0, \infty)$ (the general case is treated similarly).

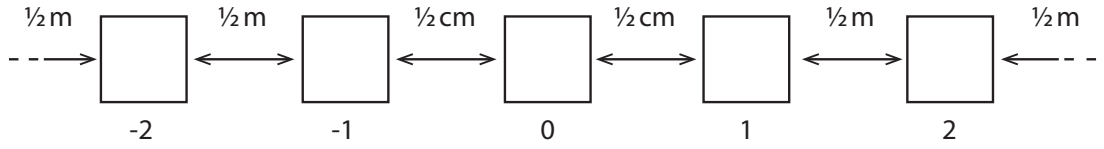

Figure 4.7: Jump rates of the random walk with an obstacle for $K = 2$

Proof of Theorem 4.1. Recall that $(\xi_n(t), t \geq 0)$ is a random walk on E with jump rates given in (4.6), (4.7) (Figure 4.7) and that $X_n(t) = \frac{1}{\sqrt{n}}\xi_n(nt)$. Also recall that d is a metric for Skorokhod convergence on compact time intervals (4.8).

Lemma 4.3.1. *The sequence* $\{(X_n(t))_{t\geq 0}, n \geq 1\}$ *is tight in* $\left(\mathrm{D}\left(\mathbb{R}_+, \mathbb{R}\right), d\right)$.

Let $(X_\infty(t))_{t\geq 0}$ be an arbitrary limit point of this sequence (*i.e.* the limit of a converging subsequence).

Lemma 4.3.2. $|X_\infty|$ *is distributed as reflected Brownian motion with diffusion coefficient* m.

Let $T_0 = 0$ and for $i \geq 0$, $T_{i+1} = \inf\{t > T_i : X_\infty(T_i)X_\infty(t) < 0\}$.

Lemma 4.3.3. *$(L^0_{T_{i+1}}(X_\infty) - L^0_{T_i}(X_\infty))_{i\geq 0}$ is a sequence of independent exponential random variables with parameter γ. This sequence is independent of $(|X_\infty(t)|)_{t\geq 0}$.*

By Proposition 4.1.5, X_∞ is characterized as (the projection on $\mathbb{R}$ of) partially reflected Brownian motion. Since the sequence X_n is tight and has only one possible limit point in $\mathrm{D}(\mathbb{R}_+, \mathbb{R})$, it converges in distribution to partially reflected Brownian motion. □

The rest of this section is devoted to the proof of Lemmas 4.3.2, 4.3.3 and 4.3.1, in that order. In what follows, we assume, with a slight abuse of notation, that $(X_n, n \geq 1)$ is a subsequence of the original sequence of processes which converges in distribution to X_∞.

4.3.1 The absolute value of X_∞

Proof of Lemma 4.3.2. To prove that the absolute value of any possible limit point of X_n is reflected Brownian motion, we write $|X_n|$ as the sum of a martingale term and a non-decreasing term. We then show that the martingale term converges to Brownian motion while the non-decreasing term converges to the opposite of the running minimum of this Brownian motion. The conclusion follows from a classical result on reflected Brownian motion [RY13, p. VI.2].

Set

$$\tilde{X}_n(t) = |X_n(t)| \mathbb{1}_{\left\{|X_n(t)| \geq \frac{2}{\sqrt{n}}\right\}}$$

and, for $i \geq 0$,

$$\sigma_0^n = 0, \qquad \tau_i^n = \inf\{t > \sigma_i^n : |X_n(t)| \leq \tfrac{1}{\sqrt{n}}\},$$
$$\sigma_{i+1}^n = \inf\{t > \tau_i^n : |X_n(t)| > \tfrac{1}{\sqrt{n}}\}.$$

The process $\tilde{X}_n$ can then be decomposed as follows [IP16]

$$\tilde{X}_n(t) = M_n(t) + L_n(t) - \sum_{i\geq 0} |X_n(\tau_i^n)| \mathbb{1}_{\{\tau_i^n \leq t < \sigma_{i+1}^n\}}, \tag{4.22}$$

with

$$M_n(t) = |X_n(0)| + \int_0^t \mathbb{1}_{\left\{|X_n(s)| > \frac{1}{\sqrt{n}}\right\}} d\,|X_n|\,(s)$$

and

$$\begin{aligned} L_n(t) &= \sum_{i\geq 0} \left(\left|X_n(\sigma_{i+1}^n)\right| - |X_n(\tau_i^n)|\right) \mathbb{1}_{\left\{\sigma_{i+1}^n \leq t\right\}} \\ &= \frac{1}{\sqrt{n}} \sum_{i\geq 0} \mathbb{1}_{\{\sigma_{i+1}^n \leq t\}}. \end{aligned} \tag{4.23}$$

The term M_n is a martingale, while L_n counts the number of visits (in fact of exits) of $[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}]$.

Define the running minimum $V_n(t)$ of the martingale part as

$$V_n(t) = \sup_{s\leq t} \left(\tfrac{2}{\sqrt{n}} - M_n(s)\right)^+ \tag{4.24}$$

and note that V_n first becomes positive when M_n first reaches $\frac{1}{\sqrt{n}}$, *i.e.*

$$\inf\{t \geq 0 : V_n(t) \geq \tfrac{1}{\sqrt{n}}\} = \inf\{t \geq 0 : M_n(t) \leq \tfrac{1}{\sqrt{n}}\}.$$

Since up to that time the other terms on the right-hand-side of (4.22) are zero, we get

$$\inf\{t \geq 0 : V_n(t) \geq \tfrac{1}{\sqrt{n}}\} = \tau_0^n.$$

The next time V_n increases is

$$\inf\{t \geq 0 : V_n(t) \geq \tfrac{2}{\sqrt{n}}\} = \inf\{t \geq 0 : M_n(t) \leq 0\}.$$

By (4.22), this is also τ_1^n. By induction,

$$V_n(t) = \frac{1}{\sqrt{n}} \sum_{i \geq 0} \mathbb{1}_{\{\tau_i^n \leq t\}}. \tag{4.25}$$

This translates the fact that the excursions of M_n above its running minimum are given by the excursions of $|X_n|$ above $\frac{1}{\sqrt{n}}$, see also Figure 4.8. Returning to (4.22), we have shown

$$\tilde{X}_n(t) = M_n(t) + V_n(t). \tag{4.26}$$

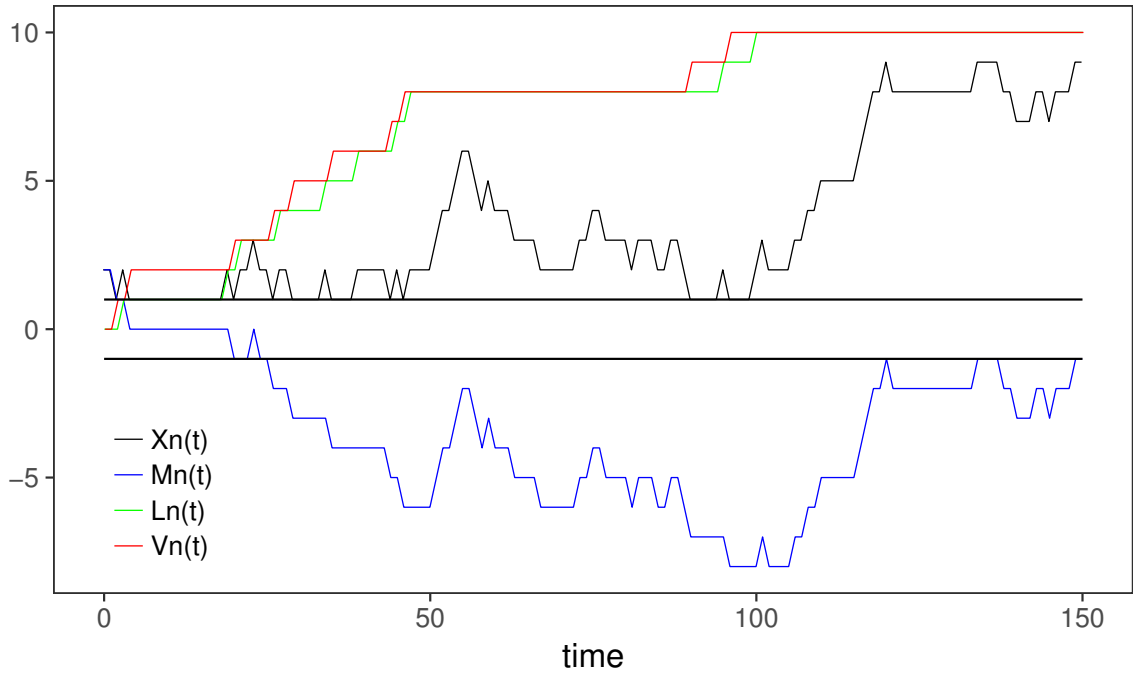

Figure 4.8: Decomposition of X_n
The black line shows a sample path of X_n for $k = 2$, $m = 0.4$ and $c_n = 0.1$. The blue line is M_n while the green (resp. red) lines show L_n (resp. V_n). We see that the excursions of M_n above its running minimum are given by the excursions of X_n outside $\{-1, 1\}$.

We show below that M_n converges in distribution in $(\mathrm{D}(\mathbb{R}_+,\mathbb{R}),d)$ to M_∞, a Brownian motion with variance parameter m (started from $|X_\infty(0)|$). Recall that we are already considering a subsequence along which X_n converges to X_∞. Passing to the limit in (4.26), we obtain

$$|X_\infty(t)| = M_\infty(t) + \sup_{s\le t}\left(-M_\infty(s)\right)^+.$$

This equation implies [RY13, p. VI.2] that $|X_\infty|$ is reflected Brownian motion (with variance parameter m) and that

$$L_t^0(X_\infty) = \sup_{s\le t}(-M_\infty(s))^+. \tag{4.27}$$

To show that M_n converges to Brownian motion, we note that M_n is a square integrable martingale with predictable variation

$$\langle M_n\rangle_t = m(t-\nu^n(t))$$

where

$$\nu^n(t) = \int_0^t \mathbb{1}_{\left\{|X_n(s)|\le\frac{1}{\sqrt{n}}\right\}}ds.$$

We prove the following in Subsection 4.3.4.

Lemma 4.3.4. *For any $t\ge 0$, $\mathbb{E}[\nu^n(t)] = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$.*

As a result $\langle M_n\rangle_t \to mt$ in probability as $n\to\infty$. Moreover,

$$\sup_{t\ge 0}\left|M_n(t)-M_n(t^-)\right| \le \frac{1}{\sqrt{n}}$$

almost surely. Hence, for example from [Reb80, Proposition II.1], M_n converges to Brownian motion in distribution in $\mathrm{D}([0,T],\mathbb{R})$ for all $T>0$. The proof of Lemma 4.3.2 is now complete. □

In passing, we have proved the following lemma.

Lemma 4.3.5. $(X_n, L_n) \xrightarrow[n\to\infty]{d} \left(X_\infty, L^0(X_\infty)\right)$

Proof. From (4.25) and (4.23), we have for $t\ge 0$

$$|L_n(t)-V_n(t)| \le \frac{1}{\sqrt{n}}.$$

Passing to limit $n\to\infty$ in (4.24), L_n converges in $\mathrm{D}([0,T],\mathbb{R})$ to L_∞ where

$$L_\infty(t) = \sup_{s\le t}(-M_\infty(s))^+.$$

We conclude the proof with (4.27). □

4.3.2 Local time accumulated between crossings

Proof of Lemma 4.3.3. To prove that the local time accumulated by X_∞ at the origin between crossings is a sequence of exponential variables, we show that the number of visits of the random walk X_n to $[-\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}]$ before the first time it reaches $-\frac{2}{\sqrt{n}}$ is a geometric random variable.

Let $(T_i^n, i \geq 0)$ be the sequence of crossing times of $[-1/\sqrt{n}, 1/\sqrt{n}]$ by X_n, *i.e.* for $n \geq 0$, set $T_0^n = 0$ and

$$T_{i+1}^n = \inf\{t > T_i^n : \mathrm{sign}(X_n(T_i^n))X_n(t) < -\tfrac{1}{\sqrt{n}}\}.$$

Let Y_n be the number of visits to $[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}]$ up to the first crossing time,

$$Y_n = \sum_{i \geq 1} \mathbb{1}_{\{\sigma_i^n \leq T_1^n\}}.$$

By the Markov property, Y_n is a geometric random variable with parameter

$$p_n = \mathbb{P}_{\frac{1}{\sqrt{n}}}\left(X_n(\sigma_1^n) = -\tfrac{2}{\sqrt{n}}\right).$$

For $K = 2$, $p_n = \frac{c_n}{2(1+c_n)}$ (and in the general case, $p_n \sim \frac{c_n}{K}$ as $n \to \infty$). Since $\frac{\sqrt{n}}{2} c_n \to \gamma \in (0, \infty)$,

$$L_n(T_1^n) = \frac{1}{\sqrt{n}} Y_n$$

converges in distribution to an exponential random variable with parameter γ. Set $E_i^n = L_n(T_{i+1}^n) - L_n(T_i^n)$. The random variables $E_0^n, E_1^n, \ldots$ are independent and identically distributed by the strong Markov property and by symmetry. As a result, $(E_i^n)_{i \geq 0}$ converges in distribution as n tends to infinity to a sequence $(E_i)_{i \geq 0}$ of independent and identically distributed exponential random variables with parameter γ. To show that this limit coincides with $(L^0_{T_{i+1}}(X_\infty) - L^0_{T_i}(X_\infty))_{i \geq 0}$, consider the following lemma.

Lemma 4.3.6. *As n tends to infinity,*

$$\left(X_n, L_n, (T_i^n)_{i \geq 0}\right) \xrightarrow[n \to \infty]{d} \left(X_\infty, L^0(X_\infty), (T_i)_{i \geq 0}\right)$$

in $\mathrm{D}\left([0,T]^2, \mathbb{R}^2\right) \times \mathbb{R}^{\mathbb{N}}$.

Since $t \mapsto L_t^0(X_\infty)$ is continuous almost surely, it follows that

$$(L_n(T_{i+1}^n) - L_n(T_i^n))_{i \geq 0} \xrightarrow[n \to \infty]{d} (L^0_{T_{i+1}}(X_\infty) - L^0_{T_i}(X_\infty))_{i \geq 0}.$$

The proof of Lemma 4.3.6 is given in Subsection 4.3.5.

We would like to show that the sequence $(E_i^n)_{i \geq 0}$ is independent of $\tilde{X}_n$, but this fails when $K \geq 2$. To circumvent this issue, we tweak $\tilde{X}_n$ so that it "forgets" the amount of time X_n spends in $[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}]$. We do this via a time change. Set

$$\delta_i^n = \inf\{t > \tau_i^n : X_n(t) \neq X_n(t^-)\}$$

and

$$\theta^n(t) = \inf\left\{\theta > 0 : \int_0^\theta \sum_{i \geq 0} \mathbb{1}_{\{s \notin [\delta_i^n, \sigma_{i+1}^n[\}} ds > t\right\}.$$

Then $(\tilde{X}_n(\theta^n(t)), t \geq 0)$ and $(L_n(T^n_{i+1}) - L_n(T^n_i))_{i\geq 0}$ are independent. Furthermore, for $t \geq 0$,

$$\left| \int_0^t \sum_{i\geq 0} \mathbb{1}_{\left\{s \notin [\delta^n_i, \sigma^n_{i+1}[\right\}} ds - t \right| \leq \nu^n(t).$$

Hence by Lemma 4.3.4, $\theta^n(t) \to t$ as $n \to \infty$ uniformly on compact sets. As a result, $\tilde{X}_n \circ \theta^n$ converges in the Skorokhod topolgy to $|X_\infty|$. We can thus conclude that $(L^0_{T_{i+1}}(X_\infty) - L^0_{T_i}(X_\infty))_{i\geq 0}$ is independent of $|X_\infty|$. □

4.3.3 Tightness

Proof of Lemma 4.3.1. Tightness of the sequence X_n follows from the convergence in distribution of M_n (recall the decomposition (4.22)). Reasoning as in [IP16] (Proof of Lemma 2.1), we show below that for any $\delta > 0$,

$$\sup_{|s-t|<\delta} |X_n(t) - X_n(s)| \leq \tfrac{3}{\sqrt{n}} + 2 \sup_{|s-t|<\delta} |M_n(t) - M_n(s)|. \tag{4.28}$$

We can thus write, for $T > 0$ and $\varepsilon > 0$

$$\lim_{\delta\downarrow 0} \limsup_{n\to\infty} \mathbb{P}\left(\sup_{\substack{|t-s|<\delta \\ s,t\in[0,T]}} |X_n(s) - X_n(t)| > \varepsilon \right)$$
$$\leq \lim_{\delta\downarrow 0} \limsup_{n\to\infty} \mathbb{P}\left(\tfrac{3}{\sqrt{n}} + 2 \sup_{\substack{|t-s|<\delta \\ s,t\in[0,T]}} |M_n(t) - M_n(s)| > \varepsilon \right).$$

The right-hand-side is zero because the sequence M_n converges in distribution in $\mathrm{D}([0,T],\mathbb{R})$, and tightness of X_n in $\mathrm{D}([0,T],\mathbb{R})$ follows [Bil99, Theorem 7.3]. Since X_n is tight in $\mathrm{D}([0,T],\mathbb{R})$ for all $T > 0$, it is tight in $(\mathrm{D}(\mathbb{R}_+,\mathbb{R}), d)$.

Let us now prove (4.28). Fix $0 \leq s \leq t$. If $|X_n(u)| > \frac{1}{\sqrt{n}}$ for all $u \in [s,t]$, then

$$X_n(t) - X_n(s) = M_n(t) - M_n(s).$$

Otherwise, let

$$\alpha = \inf\{u > s : |X_n(u)| \leq \tfrac{1}{\sqrt{n}}\},$$
$$\beta = \sup\{u < t : |X_n(u)| \leq \tfrac{1}{\sqrt{n}}\},$$

and note that

$$|X_n(t) - X_n(s)| \leq |X_n(t) - X_n(\beta)| + |X_n(\beta) - X_n(\alpha)| + |X_n(\alpha) - X_n(s)|$$
$$\leq \tfrac{3}{\sqrt{n}} + |M_n(t) - M_n(\beta)| + |M_n(s) - M_n(\alpha)|.$$

Inequality (4.28) thus holds and the proof of Lemma 4.3.1 is complete. □

4.3.4 Occupation time of the barrier

Proof of Lemma 4.3.4. The bound on the expected time spent inside $[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}]$ follows after showing that the expected length of a visit in this set is of order $\frac{1}{n}$ while the expected number of those visits is of order $\sqrt{n}$. By the definition of $\nu^n(t)$,

$$\begin{aligned}\nu^n(t) &= \sum_{i\geq 0} \left(\sigma_{i+1}^n \wedge t - \tau_i^n \wedge t\right) \\ &\leq \sum_{i\geq 0} \left(\sigma_{i+1}^n - \tau_i^n\right) \mathbb{1}_{\left\{\tau_i^n \leq t\right\}}.\end{aligned}$$

By the strong Markov property,

$$\mathbb{E}\left[\nu^n(t)\right] \leq \mathbb{E}\left[\sum_{i\geq 0} h^n(X_n(\tau_i^n))\mathbb{1}_{\left\{\tau_i^n \leq t\right\}}\right]$$

where $h^n(x) = \mathbb{E}_x\left[\inf\{t>0 : |X_n(t)| > \frac{1}{\sqrt{n}}\}\right]$. By the Markov property, for $i \in \{-1, 0, 1\}$,

$$n\sum_{j\in E} q_n(i,j)\left(h^n(j/\sqrt{n}) - h^n(i/\sqrt{n})\right) = -1.$$

Also $h^n(x) = 0$ when $|x| > \frac{1}{\sqrt{n}}$. Solving these equations for $K = 2$ yields

$$h^n\left(\pm\tfrac{1}{\sqrt{n}}\right) = \frac{3}{mn}, \qquad\qquad h^n(0) = \frac{3}{mn} + \frac{1}{c_n mn}.$$

(In the general case, $h^n\left(\lfloor\frac{K+1}{2}\rfloor/\sqrt{n}\right) = \frac{K+1}{mn}$.) For $i \geq 1$, $X_n(\tau_i^n) = \pm\frac{1}{\sqrt{n}}$, hence

$$\mathbb{E}\left[\nu^n(t)\right] \leq \frac{1}{c_n mn} + \frac{3}{mn}\mathbb{E}\left[\sum_{i\geq 0} \mathbb{1}_{\left\{\tau_i^n \leq t\right\}}\right].$$

But the number of visits of X_n to $[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}]$ before time t is less than the number of excursions outside $[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}]$ before the first excursion of length longer than t. By the Markov property, the latter is a geometric random variable with parameter

$$\mathbb{P}_{\frac{2}{\sqrt{n}}}\left(\tau_0^n > t\right).$$

But, for $t > 0$, there exists $c \in (0, \infty)$ such that [LL10, Proposition 4.2.4]

$$\lim_{n\to\infty} \sqrt{n}\mathbb{P}_{\frac{2}{\sqrt{n}}}\left(\tau_0^n > t\right) = c.$$

Hence, since $\sqrt{n}c_n \to K\gamma \in (0, \infty)$,

$$\mathbb{E}\left[\nu^n(t)\right] \leq \frac{1}{c_n nm} + \frac{3}{m\sqrt{n}}\left(\sqrt{n}\mathbb{P}_{\frac{2}{\sqrt{n}}}\left(\tau_0^n > t\right)\right)^{-1} = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

This concludes the proof of Lemma 4.3.4. □

4.3.5 Convergence of the crossing times

Proof of Lemma 4.3.6. From Lemma 4.3.5, we already know that

$$(X_n, L_n) \xrightarrow[n\to\infty]{d} (X_\infty, L^0(X_\infty)).$$

Furthermore, for all $i \geq 0$,

$$T_i^n = L_n^{-1}\left(\sum_{k=1}^{i} E_k^n\right),$$

where $t \mapsto L_n^{-1}(t)$ is the right continuous inverse of L_n. Since $(E_i^n)_{i\geq 0}$ converges in distribution to $(E_i)_{i\geq 0}$ and L_n converges in distribution to $L^0(X_\infty)$, the sequence $(T_i^n)_{n\geq 1}$ is tight in $\mathbb{R}$ for all $i \geq 0$.

As a result the sequence of random variables $(X_n, L_n, (T_i^n)_{i\geq 0})$ is tight in $\mathrm{D}\left([0,T]^2, \mathbb{R}^2\right) \times \mathbb{R}^{\mathbb{N}}$, where this space is endowed with the product topology. Let $(X_\infty, L^0(X_\infty), (\tilde{T}_i)_{i\geq 0})$ be a possible limit point of this subsequence. By the Skorokhod embedding theorem, we can assume that there exists (a version of) a subsequence which converges to (a version of) this limit point almost surely. For ease of notation we denote this subsequence by $(X_n, L_n, (T_i^n)_{i\geq 0})$.

Let $\mathcal{N} \subset \Omega$ be the negligible set on which this convergence fails, and suppose that there exists $\omega \in \Omega \setminus \mathcal{N}$ such that $\tilde{T}_1(\omega) < T_1(\omega)$. We show that for this to happen, one of two very improbable things must occur: either $\tilde{T}_1(\omega) = \tilde{T}_2(\omega)$ (but remember that $L_n(T_2^n) - L_n(T_1^n)$ is asymptotically exponentially distributed) or X_∞ must remain equal to zero for a positive amount of time after $\tilde{T}_1$.

Assume without loss of generality that $X_\infty(0) > 0$ and that $X_n(0) > 0$ for all $n \geq 1$. Then take $\varepsilon > 0$ such that $\tilde{T}_1 + \varepsilon < T_1$ (ω is kept fixed in the remainder of the proof). Since $X_n \Rightarrow X_\infty$,

$$\inf\{X_n(s), T_1^n \leq s \leq T_1^n + \varepsilon\} \underset{n\to\infty}{\longrightarrow} \inf\{X_\infty(s), \tilde{T}_1 \leq s \leq \tilde{T}_1 + \varepsilon\}.$$

Since $X_\infty(s) \geq 0$ for $s < T_1$, the right-hand-side is non-negative while the left-hand-side is non-positive because $X_n(T_1^n) = -\frac{2}{\sqrt{n}}$. As a result

$$\lim_{n\to\infty} \inf\{X_n(s), T_1^n \leq s \leq T_1^n + \varepsilon\} = 0.$$

Also note that

$$\sup\{|X_n(s)|, T_1^n \leq s \leq T_2^n \wedge (T_1^n + \varepsilon)\} \leq |\inf\{X_n(s), T_1^n \leq s \leq T_1^n + \varepsilon\}|.$$

Moreover the left-hand-side converges to

$$\sup\{|X_\infty(s)|, \tilde{T}_1 \leq s \leq \tilde{T}_2 \wedge (\tilde{T}_1 + \varepsilon)\}.$$

The latter must then be zero. Hence either $\tilde{T}_1 = \tilde{T}_2$ or there exists $\eta > 0$ such that $|X_\infty(s)| = 0$ for all $\tilde{T}_1 \leq s \leq \tilde{T}_1 + \eta$. Since $L_n(T_1^n) - L_n(T_2^n)$ converges to an exponential random variable with parameter $\gamma \in (0, \infty)$ and $|X_\infty|$ is distributed as reflected Brownian motion, both these events have probability zero.

Suppose now that $\tilde{T}_1(\omega) > T_1(\omega)$ for some $\omega \in \Omega \setminus \mathcal{N}$. By the definition of T_1, there exists $t \in (T_1, \tilde{T}_1)$ such that $X_\infty(t) < 0$. Since $T_1 \to \tilde{T}_1 > t$, there exists n_0 large enough that $T_1^n > t$ for all $n \geq n_0$. Then for all $n \geq n_0$, $X_n(t) \geq -\frac{1}{\sqrt{n}}$, but at the same time $X_n(t) \to X_\infty(t) < 0$, leading to a contradiction.

We have thus shown that $\tilde{T}_1 = T_1$ almost surely. By induction one shows that $\tilde{T}_i = T_i$ almost surely for all $i \geq 0$. It follows that $(X_n, L_n, (T_i)_{i\geq 0})$ is the only possible limit point of the sequence $(X_n, L_n, (T_i^n)_{i\geq 0})$. Together with the tightness of this sequence, this concludes the proof of Lemma 4.3.6. □

Chapter 5

Inference of demographic parameters in heterogeneous environment using long shared sequence blocks

Work in progress with Harald Ringbauer (IST Austria) and Graham Coop (UC Davis, California).

Introduction

The genetic composition of a population is the result of a long history but also reflects its more recent structure. Assuming that the population follows a simple dynamics, one can sum up this structure by a handful of parameters. Demographic inference aims to recover these parameters from a genetic sample in the present population. But for example in human populations, large scale migration events can only be neglected for up to a few hundred generations at most. Hence for such an inference method to make sense, one should rely on a signal arising from the recent history of the sampled population.

Recently, the use of long shared sequences of genome has emerged as a fruitful tool to study recent ancestry in human populations [Pal+12; Bah+16]. Since recombination breaks up the genome at each generation, thus decoupling the genealogies at different locii, long (*i.e.* $>$ 2-4 cM) continuous blocks of genome shared identical by descent between individuals are the result of very recent coancestry (typically less than 60 generations ago). Hence the geographic structure of these long bocks (called IBD blocks) carries information about the recent demography of the sample.

Detecting IBD blocks is by no means an easy task. One often has to guess their length and position from relatively dense SNP data. This was done by Ralph and Coop [RC13] on the European POPRES dataset using the fastIBD method implemented in BEAGLE v3.3 [BB11]. The POPRES dataset contains language and country of origin for several thousand Europeans genotyped at 500,000 SNPs [Nel+08]. Ralph and Coop found a total of 1.9 million IBD segments in a sample of 2,257 individuals. They were also able to estimate the rate of false positives (detected segments which

are not true IBD blocks) and the bias in the inferred length of the blocks. They found that most Europeans, even separated by several thousand kilometers, share genetic and genealogic ancestors who lived less than 3,000 years ago. More importantly, they found that the number of shared IBD blocks between individuals decreases with the geographic distance between them.

One can thus hope to use this decrease to infer the strength of dispersal in Europe, *i.e.* the mean distance individuals travel between birth and reproduction. This was done in the Eastern European subregion by Ringbauer et al. [RCB16]. They computed the expected number of IBD blocks of a given length in a pair of individuals as a function of their geographic separation under a simple isolation by distance model. Assuming independence between different blocks along the genome and between different pairs of individuals, they obtained an approximate likelihood function for the three parameters of their model: strength of dispersal (σ) and local effective population density t generations ago ($N(t) = Dt^{-\gamma}$). Numerically maximising the likelihood of the observed data, they obtained the following estimates for these parameters in Eastern Europe: $\sigma = 63 \pm 6.8$ km$/\sqrt{\text{gen}}$, $D = 2.13 \pm 0.735$ and $\gamma = 1.05 \pm 0.075$.

The reason they limited their study to Eastern Europe is that the whole European sample presents significant heterogeneities in the decrease in the number of IBD blocks with geographic distance. This heterogeneity might be due to various factors but one possible explanation is that Western and Eastern Europe have different dispersal rates and / or population densities. In order to test this hypothesis and to infer the corresponding parameters, one must change the underlying model to compute the expected number of IBD blocks under dispersal heterogeneity.

This is the aim of the work presented in this chapter. According to Theorem 3.2 in Chapter 3, the genealogy of a sample can be approximated by a system of coalescing skew Brownian motions. The discrepancy in the population density (which was not studied in Chapter 3) impacts the rate of coalescence in each halfspace and the transmission parameter $\beta \in (-1, 1)$ in equation (3.6). We then adapt the formulas used by Ringbauer et al. to our setting.

The main difficulty when doing this is finding an expression for the transition density of skew Brownian motion in a two dimensional space with different diffusion coefficients in each halfspace. We achieve this using previous work by Appuhamillage et al. [App+11]. Unfortunately, the resulting formula isn't suited to numerical computations (see Section 5.7 below). Instead, we develop an alternative approach where we replace the model in continuous space by a stepping stone model with variable migration proposed by Nagylaki [NB88]. The coalescence probabilities in this model can be computed efficiently and they agree with those of the model in continuous space.

Before applying our inference scheme to the POPRES dataset, we tested it on simulated data. We simulated structured coalescents with recombination on a large grid (see Definition 5.1.1 below) with 9 different sets of parameters and used the resulting IBD blocks as an input for our inference method. In most runs, it was able to estimate the original parameters used for the simulations with satisfying accuracy (see Figure 5.2).

This chapter is laid out as follows. We start with a definition of the model for isolation by distance with recombination in Section 5.1, and we give a definition of IBD blocks in this model. We then explain how to compute the expected number of such blocks in a pair of individuals as a function of their separation. We give the fulll expression for the transition density of skew Brownian motion in two dimensions in Section 5.2. In Section 5.3, we explain how we replace the continuous model by a stepping stone model for numerical computations. Section 5.4 details how one passes from the

expected number of IBD blocks to the full likelihood function, and how one can account for block detection errors in the computations. In Section 5.5, we present the tests of the inference scheme on simulated datasets. These results and their limits are discussed in Section 5.6.

5.1 The ancestral recombination graph

The models discussed in the previous chapters of this thesis all kept track of a single locus and ignored the effects of recombination. Here, we wish to study long blocks of genome which share the same ancestry at several close by locii, we thus need to extend the spatial coalescent to keep track of the genealogy of the whole genome of a set of individuals. This is done with the help of the ancestral recombination graph.

Definition of the model We define a system of moving and coalescing marked particles, where the markers are subsets of $[0, G]$ (G is the length of the genome, in Morgan units).

Definition 5.1.1 (Ancestral recombination graph with spatial structure)**.** *Fix $x_0 = (x_0^1, x_0^2, \dots, x_0^n)$ in $(\mathbb{R}^d)^n$. We let $(\mathcal{A}_t, t \geq 0)$ be a process such that, at time t, its state is given by*

$$\left\{(\xi_t^1, \mathcal{T}_t^1), (\xi_t^2, \mathcal{T}_t^2), \dots, (\xi_t^{N_t}, \mathcal{T}_t^{N_t})\right\},$$

where $N_t \in \mathbb{N}$ is the number of particles in $\mathcal{A}_t$ at time t, $\{\xi_t^1, \dots, \xi_t^{N_t}\}$ are their spatial locations in $\mathbb{R}^d$ and $\mathcal{T}_t^i \subset [0, G]$ is the marker of the i-th particle. The process is started from

$$\mathcal{A}_0 = \left\{(x_0^1, [0, G]), \dots, (x_0^2, [0, G])\right\}.$$

Each particle undergoes a recombination event at rate G, independently of each other. When a particle $(\xi_t^i, \mathcal{T}_t^i)$ recombines, a location on the genome v is sampled uniformly in $[0, G]$. The particle then splits into two particles with spatial location ξ_t^i and markers $\mathcal{T}_t^i \cap [0, v[$, $\mathcal{T}_t^i \cap [v, G]$. If the marker of a particle is $\emptyset$, we remove this particle.

Between recombination events, $\mathcal{A}_t$ evolves according to the spatial Λ-coalescent (Definition 1.3.4) and when particles coalesce, the marker of the parent particle is the union of those of all the daughter particles.

We also define the ancestral recombination graph (ARG) in the presence of heterogeneous dispersal.

Definition 5.1.2 (Ancestral recombination graph with heterogeneous dispersal)**.** *The ARG with heterogeneous dispersal is defined as the ARG in Definition 5.1.1, but between recombination events, $\mathcal{A}_t$ evolves according to the SLFV with heterogeneous dispersal (Definition 3.1.1).*

For $v \in [0, G]$, we define the genealogy at the locus v, noted $(\mathcal{A}_t^v, t \geq 0)$, as the subset of $(\mathcal{A}_t, t \geq 0)$ of all particles whose marker contains v:

$$\mathcal{A}_t^v = \{\xi_t^i : v \in \mathcal{T}_t^i\}.$$

Note that $(\mathcal{A}_t^v, t \geq 0)$ is a spatial Λ-coalescent (with heteregeneous dispersal in the case of Definition 5.1.2).

IBD blocs If we start from two individuals, the genealogy at each locus is always a system of two random walks which coalesce after some time. The two individuals thus share a common ancestor at every locus. For $v \in [0, G]$, let T_v denote the coalescence time of the two particles in $\mathcal{A}^v$. Following [RC13], we then define a block of identity by descent (IBD block) as a maximal continuous segment of genome where the two individuals share the same ancestry.

Definition 5.1.3 (IBD block [RC13])**.** *A segment $[a, b[\subset [0, G]$ belongs to an IBD block if for all $0 \leq t < max_{v\in[a,b[}T_v$, there exist $i(t) \neq j(t)$ in $\{1, \ldots, N_t\}$ such that $[a, b[\subset \mathcal{T}_t^{i(t)}$ and $[a, b[\subset \mathcal{T}_t^{j(t)}$. In words, the segment $[a, b]$ is not broken up by recombination in either lineage before they coalesce.*

A segment $[a, b[\subset [0, G]$ is called an IBD block if it belongs to an IBD block and if it is maximal, i.e. any segment strictly containing $[a, b[$ does not belong to an IBD block.

Note that our model doesn't include any mutation mechanism, so that, given the usual definition of identity by descent (see Section 1.3.2), everyone is identical by descent everywhere on their genome. IBD blocks are simply continuous streches of genome which are identical by descent *through the same ancestry*. On the other hand, one needs to consider mutations when trying to *observe* those blocks: long streches of genome with very few mutations between two individuals will stand out as good candidates for IBD blocks [BB11].

With this definition, any locus belongs to an IBD block, but some blocks are longer than others. IBD blocks inherited from very ancient common ancestors will be much shorter than blocks arising from very recent coalescence events. We can thus hope to learn about the recent demography of populations from the distribution of long IBD blocks in a sample of individuals.

For $L \in [0, G]$, let N_L be the number of IBD blocks of length at least L in $(\mathcal{A}_t, t \geq 0)$. We can then express the expected number of IBD blocks longer that L as a function of the distribution of the coalescence time at a single locus.

Proposition 5.1.4. *For $L \in [0, G]$,*

$$\mathbb{E}[N_L] = \mathbb{E}\left[2(G - L)T_0 e^{-2LT_0}\right].$$

Proof. Let $\mathcal{R} \subset [0, G]$ be the set of locii delimiting IBD blocs, *i.e.* $r \in \mathcal{R}$ if and only if either $r \in \{0, G\}$ or their exists $\varepsilon > 0$ such that $[r - \varepsilon, r[$ and $[r, r + \varepsilon[$ both belong to IBD blocks but $[r - \varepsilon, r + \varepsilon]$ does not (it is not hard to show that the points in $\mathcal{R}$ are isolated). Then N_L can be expressed as

$$N_L = \sum_{r\in\mathcal{R}} \mathbb{1}_{\{]r,r+L[\cap\mathcal{R}=\emptyset\}} \mathbb{1}_{\{r\leq G-L\}}.$$

Define a filtration $(\mathcal{F}_v)_{v\in[0,G]}$ by

$$\mathcal{F}_v = \sigma\left(\mathcal{A}_t^u, t \geq 0, u \leq v\right)$$

and note that $\{r \in \mathcal{R}\} \in \mathcal{F}_r$. Let

$$\mathcal{R} = \{r_1 < r_2 < \ldots < r_{|\mathcal{R}|}\}$$

and for $i > |\mathcal{R}|$, set $r_i = +\infty$. Then r_i is a stopping time for the filtration $(\mathcal{F}_v)_{v\in[0,G]}$ for all $i \geq 1$, and we define the filtration $\mathcal{F}_{r_i}$ in the usual way. The number of IBD blocks of length at least L can

then be written

$$N_L = \sum_{i\geq 1} \mathbb{1}_{\{r_i \leq G-L\}} \mathbb{1}_{\{]r_i, r_i+L[\cap\mathcal{R}=\emptyset\}}.$$

Taking expectations, we write

$$\mathbb{E}\left[N_L\right] = \sum_{i\geq 1} \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{r_i \leq G-L\}} \mathbb{1}_{\{]r_i, r_i+L[\cap\mathcal{R}=\emptyset\}} \;\middle|\; \mathcal{F}_{r_i}\right]\right].$$

But $\{r_i \leq G - L\} \in \mathcal{F}_{r_i}$ so

$$\mathbb{E}\left[N_L\right] = \sum_{i\geq 1} \mathbb{E}\left[\mathbb{1}_{\{r_i \leq G-L\}} \mathbb{E}\left[\mathbb{1}_{\{]r_i, r_i+L[\cap\mathcal{R}=\emptyset\}} \;\middle|\; \mathcal{F}_{r_i}\right]\right].$$

Note that T_{r_i} is $\mathcal{F}_{r_i}$-measurable, and that recombination events fall in $]r_i, r_i + L[$ at rate L on each lineage, so the probability that none of the lineages recombines in this region before time T_{r_i} is $e^{-2LT_{r_i}}$. As a result

$$\mathbb{E}\left[N_L\right] = \mathbb{E}\left[\sum_{r\in\mathcal{R}} \mathbb{1}_{\{0\leq r\leq G-L\}} e^{-2LT_r}\right].$$

Setting $r_{max} = \inf\{r \in \mathcal{R} : r > G - L\}$, this is also

$$\mathbb{E}\left[N_L\right] = \mathbb{E}\left[\sum_{r\in\mathcal{R}} \mathbb{1}_{\{0<r\leq r_{max}\}} e^{-2LT_{r^-}}\right].$$

Since recombination events fall on $[r, r + dr]$ at rate $2T_{r^-}$ and $r \mapsto e^{-2LT_{r^-}}$ is predictable,

$$\sum_{r\in\mathcal{R}} \mathbb{1}_{\{0<r\leq t\}} e^{-2LT_{r^-}} - \int_0^t e^{-2LT_{r^-}} 2T_{r^-} dr$$

is an $\mathcal{F}_t$-martingale. Since r_{max} is a bounded $\mathcal{F}_t$-stopping time, we have

$$\mathbb{E}\left[N_L\right] = \int_0^{G-L} \mathbb{E}\left[2T_{r^-} e^{-2LT_{r^-}}\right] dr.$$

Since the $(T_r)_{r\in[0,G]}$ are identically distributed, the term inside the integral is constant and

$$\mathbb{E}\left[N_L\right] = \mathbb{E}\left[2(G-L)T_0 e^{-2LT_0}\right].$$

□

To compute $\mathbb{E}\left[N_L\right]$, we thus need to find the probability density function of T_0. Since T_0 is the coalescence time at a single locus, we can use the same approximation as in the Wright-Malécot formula (see also [BDE02]). Let σ^2 be the variance of the motion of a single lineage and denote the local population density t generations in the past by $N(t)$ (recall that the population we consider is diploid). In the homogeneous case, following [RCB16, Equation 5] (see also [BDE02]), we can

approximate the density of T_0 if the two lineages are started at positions x and y by

$$\mathbb{P}\left(T_0 \in dt\right) \simeq \frac{1}{2N(t)} \frac{1}{4\pi\sigma^2 t} \exp\left(-\frac{|x-y|^2}{4\sigma^2 t}\right) dt.$$

This approximation is valid when t is not too large and N is large enough. In the following, we assume $N(t) = Dt^{-\gamma}$ ($\gamma = 0$ corresponds to a constant population density). Combined with Proposition 5.1.4, this yields the following approximation [RCB16] (where we assume $G \gg L$)

$$\mathbb{E}\left[N_L\right] = 2^{-(5+3\gamma)/2} \frac{G}{\pi D\sigma^2} \left(\frac{|x-y|}{\sqrt{L}\sigma}\right)^{1+\gamma} K_{1+\gamma}\left(\sqrt{2L}\frac{|x-y|}{\sigma}\right)$$

where K_α is the modified Bessel function of the second kind of degree α.

In the case of the ARG with heterogeneous dispersal, we let $g_t(x,y)$ denote the transition density of skew Brownian motion. The probability density of the coalescence time can then be approximated by

$$\mathbb{P}\left(T_0 \in dt\right) \simeq \frac{1}{2N(t)} \int_{\mathbb{R}^d} g_t(x,z) g_t(y,z) dz dt. \tag{5.1}$$

We are now left with finding an expression for $g_t(x,y)$.

5.2 Transition density of skew Brownian motion

The transition density of standard skew Brownian motion (*i.e.* with unit diffusion coefficient on both sides of the origin) is well known [Lej06]. For $t > 0$ and $x \in \mathbb{R}$ let

$$G_t(x) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{x^2}{2t}\right).$$

If $(X_t, t \geq 0)$ solves

$$X_t = x + B_t + \beta L_t^0(X) \tag{5.2}$$

for $\beta \in [-1,1]$ and $(B_t, t \geq 0)$ standard Brownian motion, then

$$\mathbb{P}_x\left(X_t \in dy\right) = \left(G_t(x-y) + \beta\,\mathrm{sign}(y) G_t(|x| + |y|)\right) dy.$$

This is easily extended to solutions to

$$X_t = x + \int_0^t \sigma(X_s) dB_s + \beta L_t^0(X),$$

with $\sigma(x) = \sigma_\pm \mathbb{1}_{\{\pm x > 0\}}$. In this case,

$$\mathbb{P}\left(X_t \in dy\right) = \frac{1}{\sigma(y)} \left(G_t\left(\frac{x}{\sigma(x)} - \frac{y}{\sigma(y)}\right) + \gamma\,\mathrm{sign}(y) G_t\left(\frac{|x|}{\sigma(x)} + \frac{|y|}{\sigma(y)}\right)\right) dy,$$

with

$$\gamma = \frac{\sigma_-(1+\beta) - \sigma_+(1-\beta)}{\sigma_-(1+\beta) + \sigma_+(1-\beta)}.$$

However, we are interested in applications to populations living in a two dimensional space. The corresponding stochastic differential equation is (3.6)

$$\begin{cases} X_t^1 = x_1 + \int_0^t \sigma(X_s^1) dB_s^1 + \beta L_t^0(X^1) \\ X_t^2 = x_2 + \int_0^t \sigma(X_s^1) dB_s^2. \end{cases} \tag{5.3}$$

The transition density for two dimensional skew Brownian motion turns out to be considerably more intricate than in one dimension. The difficulty comes from the fact that the distribution of the position of the second coordinate depends on the amount of time that the first coordinate spends on either side of the origin.

Let $\mathcal{O}_+(t)$ (resp. $\mathcal{O}_-(t)$) denote the occupation time of the positive (resp. negative) half line by X^1 up to time t, *i.e.*

$$\mathcal{O}_\pm(t) = \int_0^t \mathbb{1}_{\{\pm X_s^1 > 0\}} ds.$$

Conditional on $(X_s^1, 0 \leq s \leq t)$, X_t^2 is then a Gaussian random variable

$$X_t^2 \sim \mathcal{N}\left(x_2, \mathcal{O}_+(t)\sigma_+^2 + \mathcal{O}_-(t)\sigma_-^2\right). \tag{5.4}$$

Determining the joint distribution of X_t^1 and $\mathcal{O}_+(t)$ is then enough to obtain that of (X_t^1, X_t^2). This is done in the following proposition.

For $t > 0$ and $x \in \mathbb{R}$ let

$$h(t,x) = \frac{|x|}{2\pi t^{3/2}} \exp\left(-\frac{x^2}{2t}\right).$$

Also set $x^+ = \max(x, 0)$ and $x^- = \max(-x, 0)$.

Proposition 5.2.1. *For $y \in \mathbb{R}$, $l \geq 0$ and $0 \leq \tau \leq t$,*

$$\begin{aligned} &\mathbb{P}_x\left(X_t^1 \in dy, L_t^0(X^1) \in dl, \mathcal{O}_+(t) \in d\tau\right) \\ &\quad = \frac{1}{\sigma(y)}\left(G_t\left(\frac{x}{\sigma(x)} - \frac{y}{\sigma(y)}\right) - G_t\left(\frac{x}{\sigma(x)} + \frac{y}{\sigma(y)}\right)\right) \mathbb{1}_{\{xy>0\}} dy \delta_0(dl) \delta_{t\mathbb{1}_{\{x>0\}}}(d\tau) \\ &\quad + 2\frac{1 + \operatorname{sign}(y)\beta}{\sigma(y)^2} h\left(\tau, \frac{x^+ + y^+}{\sigma_+} + (1+\beta)\frac{l}{\sigma_+}\right) h\left(t-\tau, \frac{x^- + y^-}{\sigma_-} + (1-\beta)\frac{l}{\sigma_-}\right) dy dl d\tau. \end{aligned}$$

This proposition is a generalisation of a previous result [SK91] on the joint density of standard Brownian motion, its local time at the origin and the occupation time of the positive half line. This result was generalised to solutions of (5.2) in [App+11], using similar arguments. We give a self-contained proof of Proposition 5.2.1 in Section 5.7.

Using Proposition 5.2.1 and (5.4), we obtain the transition density for two dimensional skew

Brownian motion:

$$\begin{aligned}\mathbb{P}_x\left(X_t^1 \in dy_1, X_t^2 \in dy_2\right) = \\ &\left(G_{\sigma(y)^2 t}\left(x_1 - y_1\right) - G_{\sigma(y)^2 t}\left(x_1 + y_1\right)\right) G_{\sigma(y)^2 t}(y_2 - x_2) \mathbb{1}_{\{x_1 y_1 > 0\}} dy_1 dy_2 \\ &+ 2\frac{1 + \operatorname{sign}(y_1)\beta}{\sigma(y)^2} \int_0^t \int_0^\infty h\left(\tau, \frac{x_1^+ + y_1^+}{\sigma_+} + (1+\beta)\frac{l}{\sigma_+}\right) \\ &h\left(t - \tau, \frac{x_1^- + y_1^-}{\sigma_-} + (1-\beta)\frac{l}{\sigma_-}\right) G_{\sigma_+^2 \tau + \sigma_-^2 (t-\tau)}\left(y_2 - x_2\right) dl d\tau dy_1 dy_2.\end{aligned}$$

Plugging this formula into (5.1), and integrating this density against $2(G-L)te^{-2Lt}$, we obtain an analytical formula for the expected number of IBD blocks of length at least L in the ARG with heterogeneous dispersal. Unfortunately, this expression contains too many integrals to be computed numerically in a reasonably short time and cannot be used in a maximum likelihood approach.

5.3 Numerical approximation

To approximate the probability density function of the coalescence time of two lineages, we approximate skew Brownian motion with a random walk on a large grid. We choose the transition probabilities of the random walk following a stepping stone model with heterogeneous dispersal proposed by Nagylaki [NB88]. (Nagylaki's model is one dimensional but its extension to a two dimensional grid is straightforward.) Over large spatial and temporal scales, ancestral lineages in this model approach skew Brownian motions (*i.e.* solutions to (5.3)). Numerically computing the distribution of the coalescence time of two lineages in this discrete model provides a way to approximate (5.1).

Stepping stone model with heterogeneous dispersal Fix $m_+, m_- \in (0, 1/2)$ and $N_+, N_- > 0$ such that

$$\frac{3}{2} m_\pm + \frac{1}{2}\frac{N_\mp}{N_\pm} m_\mp < 1. \tag{5.5}$$

Consider a population living in discrete demes at the vertices of $\mathbb{Z}^2$. Each deme left (resp. right) of the origin contains N_- (resp. N_+) individuals. At each generation, each deme on the left (resp. on the right) sends $\frac{1}{2}N_- m_-$ (resp $\frac{1}{2}N_+ m_+$) migrants to each of its neighbours. This corresponds to the discrete stepping-stone model described in Definition 1.3.1 with forwards migration rates as in Figure 5.1.

The probability that an individual in deme i has an ancestor in a neighbouring deme j in the previous generation is then given by

$$m_{ij} = \frac{1}{2}\frac{N_j}{N_i} m_j,$$

where $N_i = N_+$, $m_i = m_+$ on the right of the origin and $N_i = m_-$, $m_i = m_-$ on the left. Condition (5.5) ensures that the sum of these probabilities is less than one, so that lineages have a positive probability of having an ancestor in the same deme in the previous generation. Let $M = (m_{ij})_{i,j \in \mathbb{Z}^2}$ be the (backwards) migration matrix. Ancestral lineages in this model follow discrete time random walks on $\mathbb{Z}^2$ with transition matrix M.

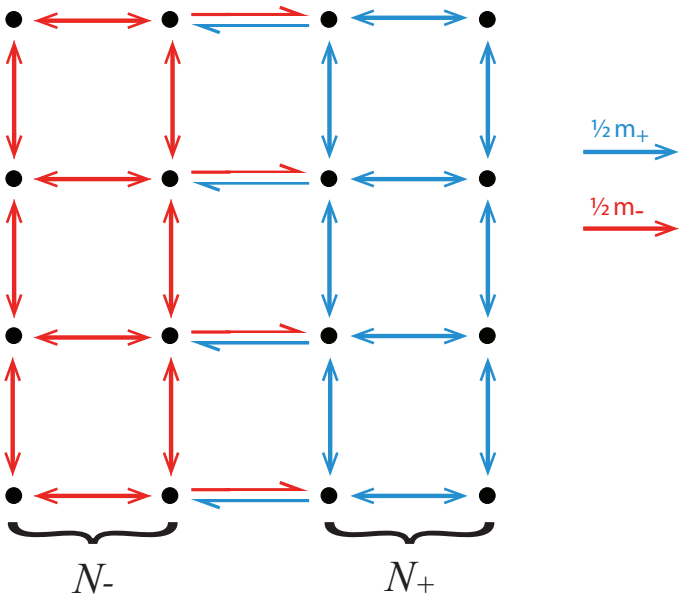

Figure 5.1: Stepping stone model with heterogeneous dispersal in two dimensions
Islands left of the origin send $\frac{1}{2}N_- m_-$ migrants per generation along blue arrows while islands right of the origin send $\frac{1}{2}N_+ m_+$ migrants per generation along red arrows.

Theorem 5.1. *Let $(\xi_k, k \in \mathbb{N})$ be a random walk on $\mathbb{Z}^2$ with transition matrix M^k for some fixed $k \geq 1$. For $n \geq 1$ and $t \geq 0$, set*

$$\xi_t^n = \frac{1}{\sqrt{n}} \xi_{\lfloor nt \rfloor}.$$

Then, as $n \to \infty$, $(\xi_t^n, t \geq 0)$ converges in distribution to skew Brownian motion (i.e. to a solution of (5.3)) with

$$\sigma_\pm^2 = k\, m_\pm, \qquad \beta = \frac{(N_+ m_+)^2 - (N_- m_-)^2}{(N_+ m_+)^2 + (N_- m_-)^2}. \tag{5.6}$$

Theorem 5.1 is a particular case of Theorem 1.1 in [IP16], and (5.6) can be seen from equation 19a in [NB88]. The exponent k is introduced because $m_\pm$ cannot be too large in order to satisfy (5.5), but we want to compute the distribution of skew Brownian motion for any $\sigma_\pm \in (0, \infty)$.

Let C be a diagonal matrix with coefficients

$$\frac{1}{2N_i} \delta_{ij}, \quad i, j \in \mathbb{Z}^2.$$

If $N_\pm$ is large enough and t isn't too large, one can then approximate the probability density function of the coalescence time of a pair of lineages started at $(i, j) \in (\mathbb{Z}^2)^2$ by

$$\mathbb{P}\left(T_0 \in [t, t+1[\right) \simeq \left((M^{\lfloor t \rfloor})^T C M^{\lfloor t \rfloor} \right)_{ij}. \tag{5.7}$$

(This is the analogue of (5.1) in discrete time and discrete space.) In view of Proposition 5.1.4, if we sum this against $2(G-L)te^{-2Lt}$, we obtain the expected number of IBD blocks of length at least L

shared between two individuals. For numerical computations, we replace $\mathbb{Z}^2$ by a large finite grid and we impose reflecting boundary conditions at the edges of the grid. In order to include varying population size, we replace C in (5.7) by $C(t)$ where

$$N_{\pm}(t) = D_{\pm} t^{-\gamma}.$$

Since the migration matrix M only depends on the ratio N_+/N_-, it isn't affected by this assumption, however the parameter γ needs to be identical in both halfspaces otherwise migration isn't homogeneous in time and the above formulas break down.

In fact, we are more interested in the expected number of IBD blocks whose length falls in a particular interval $[L, L+\Delta L[$, which we denote by $\lambda(L, x, y)$, where x and y are the positions of the sampled individuals. Using the fact that $L \ll G$, this quantity is approximated by

$$\begin{aligned}\lambda(L,x,y) &\simeq \mathbb{E}\left[4GLT_0^2 e^{-2LT_0}\right]\Delta L \\ &\simeq \sum_{t\in\mathbb{N}} 4GLt^2 e^{-2Lt}\Delta L \left((M^{\lfloor t\rfloor})^T C(t) M^{\lfloor t\rfloor}\right)_{\lfloor x\rfloor \lfloor y\rfloor}.\end{aligned}$$

Each successive term in the sum can be computed from the previous one in a reasonable number of steps, making this approximation scheme available for numerical computations. In the following, we say that a block is of length L if its length falls in $[L, L+\Delta L[$, where ΔL is kept fixed throughout (in [RCB16], $\Delta L = 0.1$ cM).

Remark. *By choosing to approximate skew Brownian motion with this particular random walk, we have constrained our model somewhat. Indeed, the parameter β is now a function of the other parameters of the model, given by* (5.6)*. Different microscopic models would have resulted in different functions, as pointed out by Nagylaki [NB88]. Further analysis would require to investigate if the estimates obtained with this method are robust to changes in the microscopic model.*

5.4 Composite Likelihood estimation

We are now able to compute the expected number of IBD blocks of length L shared by two individuals sampled at locations x and y. Let us see how we can use it to compute the approximate likelihood of a set of parameters. To compute a real likelihood, one would need to know the whole distribution of the number of IBD blocks of different lengths shared between different pairs of individuals. This distribution is quite complex and possibly untractable without further assumptions. However, since inheriting one block from a recent common ancester is a rare event, inheriting two blocks from the same ancestor is even more unlikely. Different IBD blocks should thus mostly arise from distinct coalescence events, and we expect the corresponding genealogies to be almost independent.

Following [RCB16] and [RC13], we assume that block-sharing is not correlated along the chromosome or among different pairs of individuals. This is equivalent to assuming that the number of IBD blocks of a given length shared between pairs of individuals are independent Poisson variables. The expected value of the number of IBD blocks shared between two individuals is then enough to compute the corresponding likelihood. This procedure of assuming independent observations to compute the likelihood is called a composite likelihood approach.

Let N be the number of sampled populations and $\{x_1, \dots, x_N\}$ their spatial locations in $\mathbb{R}^2$. Let $\mathbb{L}$ be the set of block lengths used for inference, *i.e.* for every pair of populations, we suppose that we know how many pairs of individuals share IBD blocks of length L for all $L \in \mathbb{L}$. Let $N(L, i, j)$ be the number of blocks of length L shared among pairs of individuals in populations i and j, and let $\{D_1, \dots, D_N\}$ be the number of sampled individuals in each population.

The likelihood of a set of parameters $(\sigma_+, \sigma_-, N_+, N_-, \gamma)$ is then given by

$$\mathcal{L} = C \prod_{L \in \mathbb{L}} \prod_{i<j} \lambda(L, x_i, x_j)^{N(L,i,j)} e^{-D_i D_j \lambda(L, x_i, x_j)}, \tag{5.8}$$

where C is a constant which doesn't depend on the parameters of the model.

Detection errors Detecting IBD blocks from SNP data is prone to errors. Some segments of genome might be called as an IBD block without having been inherited from a common ancestor, and some true IBD blocks might go unnoticed by the detection algorithm. In addition there is a relative uncertainty concerning the length of IBD blocks, especially in portions of the genome where there are too few SNPs.

Fortunately, Ralph and Coop [RC13] were able to estimate the rate of these errors by running the IBD detection scheme on a modified dataset. The expected number of observed IBD blocks $\tilde{\lambda}(L, x, y)$ given the expected number of true IBD blocks $\lambda(L, x, y)$ takes the form

$$\tilde{\lambda}(L, x, y) = f(L) + \int_0^G R(L, L') c(L') \lambda(L', x, y) dL', \tag{5.9}$$

where $f(L)$ is the false positive rate, $c(L)$ is the power to detect a block of length L and $R(L, L')dL'$ is the probability of detecting a block of true length L' as a block of length L. These three functions were estimated in [RC13].

To account for these detection errors in our inference scheme, we replace λ by $\tilde{\lambda}$ in (5.8). The corrected likelihood becomes

$$\mathcal{L} = C \prod_{L \in \mathbb{L}} \prod_{i<j} \tilde{\lambda}(L, x_i, x_j)^{N(L,i,j)} e^{-D_i D_j \tilde{\lambda}(L, x_i, x_j)}. \tag{5.10}$$

Uncertainty of estimates We obtain confidence intervals around our estimates with the help of the curvature of the likelihood surface at its maximum, using the asymptotic normality of the maximum likelihood estimate [Lin88]. Let $(\hat{\sigma}_+, \hat{\sigma}_-, \hat{N}_+, \hat{N}_-, \hat{\gamma})$ be the maximum likelihood estimate of the parameters of the model. The 95% confidence interval around $\hat{\sigma}_+$ is then given by

$$\hat{\sigma}_+ \pm 1.96 \left(\frac{\partial^2 \log(\mathcal{L})}{\partial \sigma_+^2} (\hat{\sigma}_+, \hat{\sigma}_-, \dots, \hat{\gamma}) \right)^{-1/2} \tag{5.11}$$

and the confidence intervals around the other estimates are obtained in the same way. This is based on a number of asumptions, namely that the maximum likelihood estimator is asymptotically Gaussian with variance given by the Fisher-Information matrix and that the curvature of the log-likelihood approaches this matrix. This will be the case if all observations are independent replicates, when observations are correlated, these confidence intervals will be too tight [Lin88; Cof+15] and they can

be adjusted by bootstrapping over the data (see [RCB16]). Since we believe that the correlations between IBD blocks are weak, we use this methode as a first guess for the uncertainety of our estimates and we plan to use bootstrapping to compute more accurate confidence intervals.

5.5 Tests on simulated data

To test our inference method, we simulated the ancestral recombination graph on a square grid of size 200, with 6 populations on each side of the interface (starting with 10 lineages in each population). We used 9 different sets of parameters (see Table 5.1) and each of these scenarii was simulated 20 times. We only recorded IBD blocks longer than 2cM in our simulations, and the migration scheme was chosen to fit the one used in the numerical approximation of skew Brownian motion in Section 5.3. The output of the simulation gives the number of IBD blocks of each length shared between different populations.

We ran the maximum likelihood estimation on this data (without including the error model from Ralph and Coop [RC13]) and it was able to correctly estimate the parameters used for each scenario in most cases (see Figure 5.2), even without knowing the exact spatial discretisation used in the simulation. Some systematic deviations seem to occur (*e.g.* in scenarii 1 and 2), but at this stage we cannot say if this is due to the low spatial discretisation used in the inference scheme or to the fact that our confidence intervals are too tight because of the lack of independence between different oberved blocks. One can also note that errors on the estimation of the population densities are almost always accompanied by errors on the growth parameter γ, which seems to indicates that the inference schemes is able to infer the population densities at some point in the past, from which most IBD blocks arise.

The code used for the simulations and the inference methode is available freely at `https://github.com/hringbauer/IBD-Analysis`.

5.6 Discussion

The demographic history of real populations is likely to be much more complex than we seem to assume in our model. Local heterogeneities and ancient population structure both affect the genetic composition of populations and can potentially confound or bias our estimates. Ringbauer et al.

Scenario	1	2	3	4	5	6	7	8	9
σ_-	0.8	0.4	0.5	0.5	0.4	0.4	0.4	0.4	0.8
σ_+	0.4	0.8	0.5	0.5	0.8	0.8	0.8	0.8	0.8
D_-	500	1000	40	2000	40	1500	20	100	100
D_+	1000	500	20	1000	20	1000	40	200	100
γ	1	1	0	1	0	1	0	0.5	0.5

Table 5.1: Parameters used in the simulations

For each scenario we simulated the ARG with the parameters shown above (with $N(t) = Dt^{-\gamma}$). Scenarii 1 and 2 mirror each other and are used to spot potential bugs in our implementation; scenarii 3 and 4 test the ability to detect a change in the population density when $\gamma = 0$ and $\gamma = 1$; scenarii 5, 6, 7 and 8 test various combinations of high *vs.* low population density and high *vs.* low dispersal rate; finally scenario 9 tests the inference method when space is homogeneous.

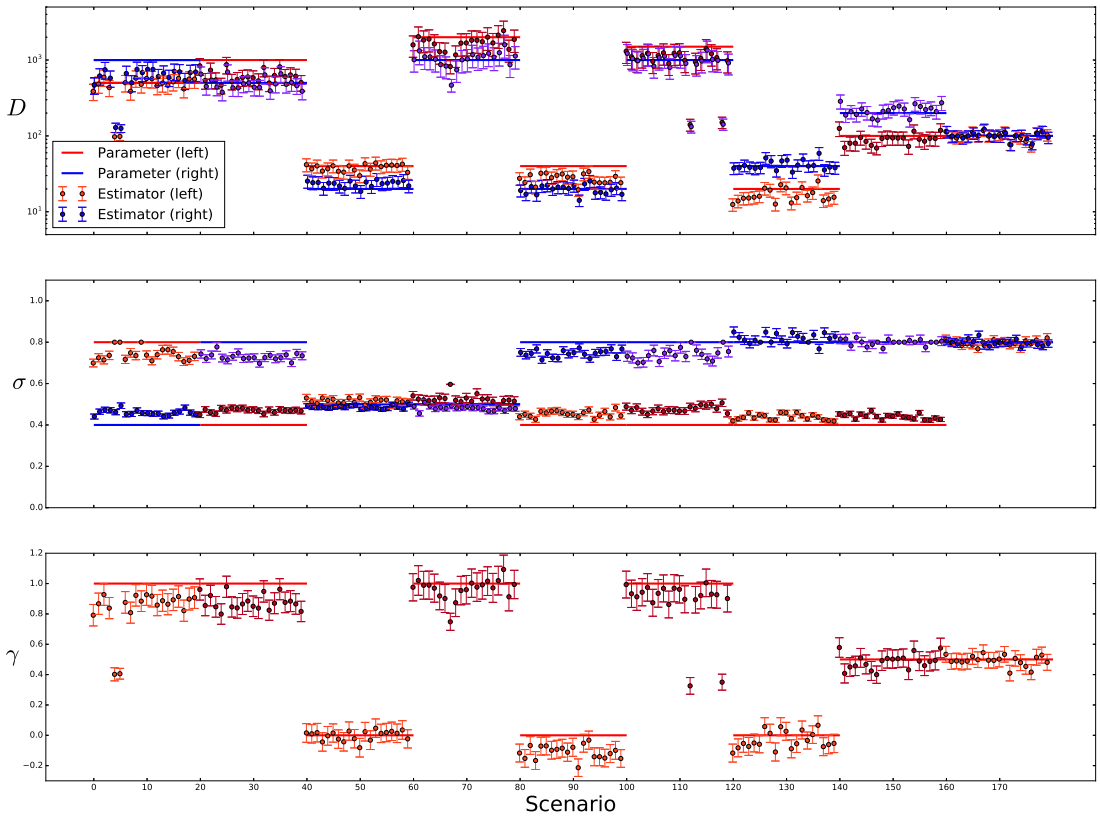

Figure 5.2: First results on simulated data

Results of the inference method on the data simulated for the 9 scenarii described in Table 5.1. For each run, the solid lines give the parameters used in the simulation while the dots with error bars represent the maximum likelihood estimators with 95% confidence interval computed with (5.11). Blue lines and dots stand for parameters and estimates on the right while red lines and dots are used for the parameters on the left (see Figure 5.1).

[RCB16] have already shown that using long IBD blocks is robust to confounding by ancestral structure as most of those blocks result from very recent common ancestry (typically less than 50 generations ago for blocks longer than 4cM).

The geography of the European continent is also more complex than in our model, for example some countries are separated from each other by seas, rivers or mountain ranges. Moreover, current population densities clearly do not fit our hypotheses. Nevertheless our method allows one to test whether current levels of genetic relatedness accross Europe can roughly be explained by a discontinuity in both dispersal rate and population density. More complex scenarii can readily be implemented in our framework (one only needs to specify where an individual's potential ancestors come from at each generation), but given the spatial resolution of the POPRES dataset (either country of origin or spoken language), testing this crude scenario is probably the best one can hope for at this stage. More intensive sampling would open the way for more tailored inference methods and better constrain our estimates.

Our inference scheme also assumes that we are able to accurately observe recombination events

along the genome. This shortcoming is partly addressed by the error model of Ralph and Coop [RC13]. Lineages that have recently recombined can potentially coalesce back too quickly to accumulate enough mutations, leading to so-called ineffective recombinations (and overestimation of the length of IBD blocks). Barton et al. [Bar+13] show that this can be accounted for by replacing the theoretical recombination rate by an effective recombination rate. In practice, if the population density is large enough, ineffective recombinations are unlikely and one can use the true recombination rate for demographic inference.

Likewise, correlations between the lengths of different blocks are very weak if population densities are reasonably large. Moreover, as already noted in [RCB16], we seldom find more than one IBD block in any given pair of individuals in the European POPRES dataset. Consequently, ignoring correlations along the genome should not significantly affect our estimates.

The fact that the inference scheme is able to correctly estimate the parameters from simulated datasets gives further credit to this approximation. It also provides a practical justification for our use of the approximation (5.7). In particular, since the spatial discretisation used in the inference scheme is different from the one used in the simulations, we conclude that the diffusion approximation for the motion of lineages is valid in this range of parameters.

From the results in Chapter 3 and the remarks made by Nagylaki [NB88], we note that the parameter $\beta \in (-1, 1)$ depends on the details of the microscopic model. As a result it cannot be deduced from the other parameters of the model and should be estimated alongside them. However, our method prevents this since by approximating skew Brownian motion with the random walk with transition rates m_{ij}, we force β to be a function of the dispersal rates and populations sizes (see (5.6)). Hence we plan to simulate the ARG with one migration scheme and to run the inference method assuming a different microscopic model. If it is still able to recover the other parameters of the model, we shall conclude that the choice of the microscopic migration model does not affect our estimates.

5.7 Joint density of skew Brownian motion, its local time at the origin and the occupation time of the positive half line

In this section, we prove Proposition 5.2.1. Let $(B_t, t \geq 0)$ be standard Brownian motion, $\sigma(x) = \sigma_{\pm}\mathbb{1}_{\{\pm x>0\}}$ and $\beta \in (-1, 1)$. Suppose that $(X_t, t \geq 0)$ solves

$$X_t = X_0 + \int_0^t \sigma(X_s)dB_s + \beta L_t^0(X).$$

We begin by proving the formula when $X_0 = 0$. In this case, Proposition 5.2.1 reduces to

$$\mathbb{P}_0\left(X_t \in dy, L_t^0(X) \in dl, \mathcal{O}_+(t) \in d\tau\right) = 2\frac{1+\text{sign}(y)\beta}{\sigma(y)^2} h\left(\tau, \frac{y^+}{\sigma_+} + (1+\beta)\frac{l}{\sigma_+}\right) h\left(t-\tau, \frac{y^-}{\sigma_-} + (1-\beta)\frac{l}{\sigma_-}\right) dydld\tau, \quad (5.12)$$

for $y \in \mathbb{R}$, $l \geq 0$ and $0 \leq \tau \leq t$. To prove this, we adapt the arguments of Chapter 6.3 in [SK91]. The main idea is to construct two reflected Brownian motions, one from the positive excursions of X, and one from its negative excursions. These two reflected Brownian motions are independent, but one can recover the value of X_t from them if the occupation time of the positive half line is known, because the left and right local time at the origin of X_t satisfy a linear relation (see Lemma 3.5.3).

Two independent reflected Brownian motion We start by decomposing X into two independent reflected Brownian motions. For $t \geq 0$, set

$$I_{\pm}(t) = \mp \int_0^t \mathbb{1}_{\{\pm X_s > 0\}} dX_s$$

and

$$\Gamma_{\pm}(t) = \langle I_{\pm} \rangle_t = \sigma_{\pm}^2 \int_0^t \mathbb{1}_{\{\pm X_s > 0\}} ds.$$

Define the right-continuous inverse of $\Gamma_{\pm}$:

$$\Gamma_{\pm}^{-1}(\tau) = \inf\{t > 0 : \Gamma_{\pm}(t) > \tau\}.$$

Lemma 5.7.1. *There exist two independent standard Brownian motions* $(B_+(\tau), \tau \geq 0)$ *and* $(B_-(\tau), \tau \geq 0)$ *such that, for* $\tau \geq 0$*,*

$$B_{\pm}(\tau) = I_{\pm}(\Gamma_{\pm}^{-1}(\tau)).$$

Proof. Let $(\mathcal{F}_t)_{t \geq 0}$ denote the filtration associated to $(X_t)_{t \geq 0}$. Since

$$\int_0^t \mathbb{1}_{\{\pm X_s > 0\}} dL_s^0(X) = 0 \text{ a.s.}, \tag{5.13}$$

we can write

$$I_{\pm}(t) = \mp \int_0^t \mathbb{1}_{\{\pm X_s > 0\}} \sigma(X_s) dB_s.$$

The process $(I_{\pm}(t))_{t \geq 0}$ is thus an $\mathcal{F}_t$-martingale with quadratic variation

$$\langle I_{\pm} \rangle_t = \sigma_{\pm}^2 \int_0^t \mathbb{1}_{\{\pm X_s > 0\}} ds = \Gamma_{\pm}(t).$$

Furthermore

$$\langle I_+, I_- \rangle_t = \sigma_+ \sigma_- \int_0^t \mathbb{1}_{\{X_s > 0\}} \mathbb{1}_{\{X_s < 0\}} ds = 0.$$

By F. B. Knight's theorem [Kni71] (also Theorem 3.4.13 in [SK91]), there exists two independent Brownian motions $(B_+(\tau), \tau \geq 0)$ and $(B_-(\tau), \tau \geq 0)$ such that, for $\tau \geq 0$,

$$B_{\pm}(\tau) = I_{\pm}(\Gamma_{\pm}^{-1}(\tau)).$$

□

For $\tau \geq 0$, set

$$X_\pm(\tau) = \pm X_{\Gamma_\pm^{-1}(\tau)}, \qquad L_\pm(\tau) = L^{0\pm}_{\Gamma_\pm^{-1}(\tau)}(X).$$

Lemma 5.7.2. *For all $\tau \geq 0$,*

$$X_\pm(\tau) = \max_{0\leq u\leq \tau} B_\pm(u) - B_\pm(\tau), \tag{5.14}$$

$$\frac{1}{2}L_\pm(\tau) = \max_{0\leq u\leq \tau} B_\pm(u). \tag{5.15}$$

As a consequence, $(X_+(\tau), \tau \geq 0)$ and $(X_-(\tau), \tau \geq 0)$ are two independent reflected Brownian motions.

Proof. By the definition of left and right local time (see Section 1.5.3),

$$X_t^\pm = -I_\pm(t) + \frac{1}{2}L_t^{0\pm}(X).$$

Together with (5.13), this implies that $\frac{1}{2}L^{0\pm}(X)$ is a solution of the Skorokhod problem for $I_\pm$. It follows from Lemma 1.5.2 (see [SK91, Lemma 3.6.14]) that

$$\frac{1}{2}L_t^{0\pm}(X) = \max_{0\leq s\leq t} I_\pm(s)$$
$$X_t^\pm = \max_{0\leq s\leq t} I_\pm(s) - I_\pm(t).$$

Lemma 5.7.2 follows by replacing t with $\Gamma_\pm^{-1}(\tau)$. □

Williams formula for skew Brownian motion The following Lemma is a generalisation of the first formula of D. Williams [SK91, Theorem 6.3.6].

Lemma 5.7.3. *For $\tau \geq 0$,*

$$L_-^{-1}\left(\frac{1-\beta}{1+\beta}L_+(\tau)\right) = \sigma_-^2\Gamma_+^{-1}(\tau) - \left(\frac{\sigma_-}{\sigma_+}\right)^2\tau.$$

Proof. By the definition of $L_-^{-1}(t)$, for $\tau \geq 0$,

$$L_-^{-1}\left(\frac{1-\beta}{1+\beta}L_+(\tau)\right) = \inf\left\{t > 0 : L_-(t) > \frac{1-\beta}{1+\beta}L_+(\tau)\right\}.$$

By the definition of $L_\pm$,

$$L_-(t) > \frac{1-\beta}{1+\beta}L_+(\tau) \quad \Leftrightarrow \quad L^{0-}_{\Gamma_-^{-1}(t)}(X) > \frac{1-\beta}{1+\beta}L^{0+}_{\Gamma_+^{-1}(\tau)}(X).$$

But (see Lemma 3.5.3), for all $t \geq 0$,

$$L_t^{0-}(X) = \frac{1-\beta}{1+\beta}L_t^{0+}(X).$$

In addition,

$$L_t^{0+}(X) > L_s^{0+}(X) \quad \Leftrightarrow \quad t > s \text{ and } \exists u \in (s,t) : X_u = 0.$$

Since $X(\Gamma_+^{-1}(t)) \geq 0$ and $X(\Gamma_-^{-1}(t)) \leq 0$, for all $t, \tau > 0$, there exists u in the open interval delimited by $\Gamma_+^{-1}(t)$ and $\Gamma_-^{-1}(\tau)$ such that $X_u = 0$. As a result,

$$L_-(t) > \frac{1-\beta}{1+\beta} L_+(\tau) \quad \Leftrightarrow \quad \Gamma_-^{-1}(t) > \Gamma_+^{-1}(\tau).$$

But

$$\inf\{t > 0 : \Gamma_-^{-1}(t) > \Gamma_+^{-1}(\tau)\} = \Gamma_- \left(\Gamma_+^{-1}(\tau)\right).$$

Furthermore,

$$\frac{1}{\sigma_+^2}\Gamma_+(t) + \frac{1}{\sigma_-^2}\Gamma_-(t) = t.$$

replacing t by $\Gamma_+^{-1}(\tau)$, we obtain

$$L_-^{-1}\left(\frac{1-\beta}{1+\beta}L_+(\tau)\right) = \sigma_-^2 \Gamma_+^{-1}(\tau) - \left(\frac{\sigma_-}{\sigma_+}\right)^2 \tau.$$

This concludes the proof of Lemma 5.7.3. □

Note that (5.15) implies, for $b \geq 0$,

$$L_\pm^{-1}(b) := \inf\{t > 0 : L_\pm(t) > b\} = \inf\{t > 0 : B_\pm(t) > b\} =: T_b^\pm.$$

From this and Lemma 5.7.3, we have

$$\Gamma_+^{-1}(\tau) = \frac{\tau}{\sigma_+^2} + \frac{1}{\sigma_-^2} T^-_{\frac{1-\beta}{1+\beta}L_+(\tau)}.$$

From the independence of B^+ and B^- and the Laplace transform of the hiting time of Brownian motion, we have

$$\mathbb{E}\left[e^{-\alpha\Gamma_+^{-1}(\tau)} \,\middle|\, X_+(u), 0 \leq u \leq +\infty\right] = \exp\left(-\frac{\alpha\tau}{\sigma_+^2} - \frac{\sqrt{2\alpha}}{\sigma_-}\frac{1-\beta}{1+\beta}L_+(\tau)\right).$$

This is equivalent to

$$\mathbb{P}\left(\Gamma_+^{-1}(\tau) \in dt \,\middle|\, X_+(\tau) = a, L_+(\tau) = l\right) = h\left(t - \frac{\tau}{\sigma_+^2}, \frac{1-\beta}{1+\beta}\frac{l}{\sigma_-}\right) dt.$$

Joint density started from the origin In addition, since $(X_+(\tau), \tau \geq 0)$ is reflected Brownian motion [SK91, Problem 6.3.4],

$$\mathbb{P}_0\left(X_+(\tau) \in da, L_+(\tau) \in dl\right) = 2h(\tau, l+a)dadl.$$

As a result,

$$\mathbb{P}_0\left(X_+(\tau)\in da, L_+(\tau)\in dl, \Gamma_+^{-1}(\tau)\in dt\right) = 2h(\tau, l+a)h\left(t-\frac{\tau}{\sigma_+^2}, \frac{1-\beta}{1+\beta}\frac{l}{\sigma_-}\right)dadldt.$$

Recall that $X_+(\tau) = X_{\Gamma_+^{-1}(\tau)}$ and $L_+(\tau) = L^{0+}_{\Gamma_+^{-1}(\tau)}(X)$. Reasonning as in [SK91, Proposition 6.3.9], we obtain, for $y \geq 0$ and $0 \leq \tau \leq t$,

$$\mathbb{P}_0\left(X_t \in dy, L_t^{0+}(X)\in dl, \Gamma_+(t)\in d\tau\right) = \frac{2}{\sigma_+^2}h(\tau, l+y)h\left(t-\frac{\tau}{\sigma_+^2}, \frac{1-\beta}{1+\beta}\frac{l}{\sigma_-}\right)dydld\tau.$$

Furthermore, $L_t^0(X) = \frac{1}{1+\beta}L_t^{0+}(X)$ and $\mathcal{O}_+(t) = \frac{1}{\sigma_+^2}\Gamma_+(t)$, hence for $y \geq 0$,

$$\begin{aligned}\mathbb{P}_0\left(X_t \in dy, L_t^0(X)\in dl, \mathcal{O}_+(t)\in d\tau\right)&\\ = 2\frac{1+\beta}{\sigma_+^2}h\left(\tau, \frac{y}{\sigma_+}+(1+\beta)\frac{l}{\sigma_+}\right)h\left(t-\tau, (1-\beta)\frac{l}{\sigma_-}\right)dydld\tau.&\end{aligned}$$

Yielding (5.12) when $y \geq 0$. The corresponding formula for $y \leq 0$ is obtained by replacing β by $-\beta$, exchanging σ_+ and σ_- and replacing τ by $t-\tau$.

Joint density with arbitrary initial condition Now start X_t from an arbitrary point $x \in \mathbb{R}$ and set

$$T_0 = \inf\{t > 0 : X_t = 0\}.$$

We assume for now that $x \geq 0$, the same arguments apply for $x < 0$. Then, since up to time T_0, X is Brownian motion with diffusion coefficient σ_+, [App+11],

$$\mathbb{P}_x\left(T_0 \in ds\right) = h\left(s, \frac{x}{\sigma_+}\right)ds.$$

By the strong Markov property,

$$\begin{aligned}\mathbb{P}_x\left(X_t \in dy, L_t^0(X)\in dl, \mathcal{O}_+(t)\in d\tau\right)&\\ = \mathbb{E}_x\left[\mathbb{1}_{\{T_0<t\}}\mathbb{P}_0\left(X_{t-T_0}\in dy, L^0_{t-T_0}(X)\in dl, \mathcal{O}_+(t-T_0)\in d\tau - T_0\right)\right]&\\ + \delta_0(dl)\delta_t(d\tau)\mathbb{P}_x\left(X_t\in dy, T_0 \geq t\right).&\quad (5.16)\end{aligned}$$

Using the reflection principle, we obtain

$$\mathbb{P}_x\left(X_t\in dy, T_0\geq t\right) = \mathbb{1}_{\{xy>0\}}\left(G_{\sigma_+^2 t}(x-y) - G_{\sigma_+^2 t}(x+y)\right). \quad (5.17)$$

On the other hand,

$$\begin{aligned}&\mathbb{E}_x\left[\mathbb{1}_{\{T_0<t\}}\mathbb{P}_0\left(X_{t-T_0}\in dy, L^0_{t-T_0}(X)\in dl, \mathcal{O}_+(t-T_0)\in d\tau - T_0\right)\right]\\ &= 2\frac{1+\mathrm{sign}(y)\beta}{\sigma(y)^2}\int_0^\tau h\left(s, \frac{x}{\sigma_+}\right)h\left(\tau-s, \frac{y^+}{\sigma_+}(1+\beta)\frac{l}{\sigma_+}\right)h\left(t-\tau, \frac{y^-}{\sigma_-}+(1-\beta)\frac{l}{\sigma_-}\right)dsdydld\tau.\end{aligned}$$

Since $h(\cdot, x) \star h(\cdot, y)(t) = h(t, x+y)$, the right-hand-side becomes

$$2\frac{1+\text{sign}(y)\beta}{\sigma(y)^2} h\left(\tau, \frac{x+y^+}{\sigma_+} + (1+\beta)\frac{l}{\sigma_+}\right) h\left(t-\tau, \frac{y^-}{\sigma_-} + (1-\beta)\frac{l}{\sigma_-}\right) dydld\tau.$$

Combined with (5.16) and (5.17), this yields Proposition 5.2.1 for $x \geq 0$. The case $x < 0$ is similar.

Bibliography

[ADN99] Bruce P Ayati, Todd F Dupont, and Thomas Nagylaki. "The Influence of Spatial Inhomogeneities on Neutral Models of Geographical Variation IV. Discontinuities in the Population Density and Migration Rate". In: *Theoretical Population Biology* 56.3 (1999), pp. 337–347. ISSN: 0040-5809.

[Agu+17] Stepfanie M. Aguillon et al. "Deconstructing isolation-by-distance: The genomic consequences of limited dispersal". In: *PLOS Genetics* 13.8 (Aug. 2017), e1006911. ISSN: 1553-7404.

[AK15] Franz Achleitner and Christian Kuehn. "Traveling waves for a bistable equation with nonlocal diffusion". EN. In: *Advances in Differential Equations* 20.9/10 (Sept. 2015), pp. 887–936. ISSN: 1079-9389.

[Ald78] David Aldous. "Stopping times and tightness". In: *The Annals of Probability* 6.2 (1978), pp. 335–340.

[App+11] Thilanka Appuhamillage et al. "Occupation and local times for skew Brownian motion with applications to dispersion across an interface". In: *The Annals of Applied Probability* 21.1 (2011), pp. 183–214. ISSN: 1050-5164.

[AS64] Milton Abramowitz and Irene A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Vol. 55. Courier Corporation, 1964.

[Bah+16] Soheil Baharian et al. "The great migration and African-American genomic diversity". In: *PLoS genetics* 12.5 (2016), e1006059.

[Bar+13] Nick H. Barton et al. "Inference in two dimensions: allele frequencies versus lengths of shared sequence blocks". In: *Theoretical population biology* 87 (2013), pp. 105–119.

[Bar08] N. H. Barton. "The effect of a barrier to gene flow on patterns of geographic variation". In: *Genetics research* 90.01 (2008), pp. 139–149.

[BB11] Brian L. Browning and Sharon R. Browning. "A fast, powerful method for detecting identity by descent". In: *The American Journal of Human Genetics* 88.2 (2011), pp. 173–182.

[BDE02] Nick H. Barton, Frantz Depaulis, and Alison M. Etheridge. "Neutral evolution in spatially continuous populations". In: *Theoretical population biology* 61.1 (2002), pp. 31–48.

[BEK06] Hermine Biermé, Anne Estrade, and Ingemar Kaj. "About scaling behavior of random balls models". In: *Proceed. 6th Int. Conf. on Stereology, Spatial Statistics and Stochastic Geometry, Prague*. 2006.

[BEK10] Hermine Biermé, Anne Estrade, and Ingemar Kaj. "Self-similar Random Fields and Rescaled Random Balls Models". en. In: *Journal of Theoretical Probability* 23.4 (Dec. 2010), pp. 1110–1141. ISSN: 0894-9840, 1572-9230.

[BEV10a] Nick H. Barton, Alison M. Etheridge, and Amandine Véber. "A new model for evolution in a spatial continuum". In: *Electronic Journal of Probability* 15.7 (2010), pp. 162–216. ISSN: 1083-6489.

[BEV10b] Nick H. Barton, Alison M. Etheridge, and Amandine Véber. "A new model for evolution in a spatial continuum". In: *Electronic Journal of Probability* 15.7 (2010), pp. 162–216. ISSN: 1083-6489.

[BEV13a] Nick H. Barton, Alison M. Etheridge, and Amandine Véber. "Modeling evolution in a spatial continuum". In: *Journal of Statistical Mechanics: Theory and Experiment* 2013.01 (2013), P01002.

[BEV13b] Nathanaël Berestycki, Alison M. Etheridge, and Amandine Véber. "Large scale behaviour of the spatial Lambda-Fleming–Viot process". In: *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 49.2 (2013), pp. 374–401. ISSN: 0246-0203.

[BGT89] Nicholas H. Bingham, Charles M. Goldie, and Jef L. Teugels. *Regular variation*. Vol. 27. Encyclopedia of Mathematics and its Applications. Cambridge university press, Cambridge, 1989. ISBN: 0-521-37943-1.

[Bil99] Patrick Billingsley. *Convergence of probability measures*. Second. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., New York, 1999. ISBN: 0-471-19745-9.

[BLG03] Jean Bertoin and Jean-François Le Gall. "Stochastic flows associated to coalescent processes". In: *Probability Theory and Related Fields* 126.2 (2003), pp. 261–288. ISSN: 0178-8051.

[BP87] Richard F. Bass and Etienne Pardoux. "Uniqueness for diffusions with piecewise constant coefficients". In: *Probability Theory and Related Fields* 76.4 (1987), pp. 557–572. ISSN: 0178-8051.

[Cas+00] V. Castella et al. "Is the Gibraltar Strait a barrier to gene flow for the bat Myotis myotis (Chiroptera: Vespertilionidae)?" In: *Molecular ecology* 9.11 (2000), pp. 1761–1772.

[Chm13] Adam Chmaj. "Existence of traveling waves in the fractional bistable equation". In: *Archiv der Mathematik* 100.5 (2013), pp. 473–480. ISSN: 0003-889X.

[Cof+15] Alec J. Coffman et al. "Computationally efficient composite likelihood statistics for demographic inference". In: *Molecular biology and evolution* 33.2 (2015), pp. 591–593.

[DMFL86] A. De Masi, P. A. Ferrari, and J. L. Lebowitz. "Reaction-diffusion equations for interacting particle systems". In: *Journal of Statistical Physics* 44.3-4 (1986), pp. 589–644. ISSN: 0022-4715.

[Dob40] Wolfgang Doblin. "Éléments d'une théorie générale des chaînes simples constantes de Markoff". In: *Annales scientifiques de l'École Normale Supérieure* 57 (1940), pp. 61–111.

[DR08] Richard Durrett and Mateo Restrepo. "One-dimensional stepping stone models, sardine genetics and Brownian local time". In: *The Annals of Applied Probability* 18.1 (2008), pp. 334–358.

[Dur08] Richard Durrett. *Probability models for DNA sequence evolution*. Springer, 2008.

[Dur10] Richard Durrett. *Probability: theory and examples*. Fourth. Vol. 31. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge university press, Cambridge, 2010. ISBN: 978-0-521-76539-8.

[EFS15] Alison Etheridge, Nic Freeman, and Daniel Straulino. "The Brownian Net and Selection in the Spatial Lambda-Fleming-Viot Process". In: *arXiv preprint arXiv:1506.01158* (2015).

[EK86] Stewart N. Ethier and Thomas G. Kurtz. *Markov processes: characterization and convergence*. John Wiley & Sons, Inc., New York, 1986. ISBN: 0-471-08186-8.

[Eth+15] Alison Etheridge et al. "Branching Brownian motion and Selection in the Spatial Lambda-Fleming-Viot Process". In: *arXiv preprint arXiv:1512.03766* (2015).

[Eth08] Alison M. Etheridge. "Drift, draft and structure: some mathematical models of evolution". In: *Stochastic models in biological sciences*. Vol. 80. Banach Center Publ. Polish Acad. Sci. Inst. Math., Warsaw, 2008, pp. 121–144.

[Eth11] Alison Etheridge. *Some mathematical models from population genetics*. Vol. 2012. Lecture Notes in Mathematics. Lectures from the 39th Probability Summer School held in Saint-Flour, 2009, Ecole d'Eté de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School]. Springer, Heidelberg, 2011. ISBN: 978-3-642-16631-0.

[Eva97] Steven N. Evans. "Coalescing Markov labelled partitions and a continuous sites genetics model with infinitely many types". In: *Annales de l'Institut Henri Poincaré. Probabilités et Statistiques* 33.3 (1997), pp. 339–358. ISSN: 0246-0203.

[EVY14] Alison Etheridge, Amandine Veber, and Feng Yu. "Rescaling limits of the spatial Lambda-Fleming-Viot process with selection". In: *arXiv preprint arXiv:1406.5884* (2014).

[Fel51] William Feller. "Diffusion processes in genetics". In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*. University of California Press, Berkeley and Los Angeles, 1951, pp. 227–246.

[Fel75] Joseph Felsenstein. "A pain in the torus: some difficulties with models of isolation by distance". In: *American Naturalist* (1975), pp. 359–368.

[Fis37] Ronald Aylmer Fisher. "The wave of advance of advantageous genes". In: *Annals of Eugenics* 7.4 (1937), pp. 355–369.

[FP17] Raphaël Forien and Sarah Penington. "A central limit theorem for the spatial Lambda Fleming-Viot process with selection". In: *Electronic Journal of Probability* 22.5 (2017), pp. 1–68. ISSN: 1083-6489.

[Gay+07] Tenzin Gayden et al. "The Himalayas as a Directional Barrier to Gene Flow". In: *The American Journal of Human Genetics* 80.5 (2007), pp. 884–894. ISSN: 0002-9297.

[Gre06] Denis S. Grebenkov. "Partially reflected Brownian motion: a stochastic approach to transport phenomena". In: *Focus on Probability Theory*. Nova Sci. Publ., New York, 2006, pp. 135–169.

[GVNL14] Denis S. Grebenkov, Dang Van Nguyen, and Jing-Rebecca Li. "Exploring diffusion across permeable barriers at high gradients. I. Narrow pulse approximation". In: *Journal of Magnetic Resonance* 248 (Nov. 2014), pp. 153–163. ISSN: 1090-7807.

[Har85] J. Harrison. "Brownian motion and stochastic flow systems". In: (1985).

[HS81] John Michael Harrison and Lawrence A. Shepp. "On skew Brownian motion". In: *The Annals of Probability* 9.2 (1981), pp. 309–313. ISSN: 0091-1798.

[IM63] K. Itô and H. P. McKean. "Brownian motions on a half line". In: *Illinois Journal of Mathematics* 7 (1963), pp. 181–231. ISSN: 0019-2082.

[IP16] Alexander Iksanov and Andrey Pilipenko. "A functional limit theorem for locally perturbed random walks". In: *Probability and Mathematical Statistics* 36.2 (2016), pp. 353–368.

[JS03] Jean Jacod and Albert N. Shiryaev. *Limit theorems for stochastic processes*. Second. Vol. 288. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 2003. ISBN: 3-540-43932-3.

[Kim53] M. Kimura. "Stepping-stone model of population". In: *Annual Report of the National Institute of Genetics* 3 (1953), pp. 62–63.

[Kim64] Motoo Kimura. "Diffusion models in population genetics". In: *Journal of Applied Probability* 1.2 (1964), pp. 177–232.

[Kim68] Motoo Kimura. "Evolutionary rate at the molecular level". In: *Nature* 217.5129 (1968), pp. 624–626.

[Kim84] Motoo Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1984.

[Kin77] J. F. C. Kingman. "Remarks on the spatial distribution of a reproducing population". In: *Journal of applied probability* (1977), pp. 577–583.

[Kin82] John Frank Charles Kingman. "The coalescent". In: *Stochastic processes and their applications* 13.3 (1982), pp. 235–248.

[Kin93] J. F. C. Kingman. *Poisson processes*. Vol. 3. Oxford Studies in Probability. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, 1993. ISBN: 0-19-853693-3.

[Kni71] Frank B. Knight. "A reduction of continuous square-integrable martingales to Brownian motion". In: *Lecture notes in Math* 190 (1971), pp. 19–31.

[Kur71] Thomas G. Kurtz. "Limit theorems for sequences of jump Markov processes approximating ordinary differential processes". In: *Journal of Applied Probability* 8.2 (1971), pp. 344–356. ISSN: 0021-9002.

[Kur81] Thomas G. Kurtz. *Approximation of population processes*. Vol. 36. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa., 1981. ISBN: 0-89871-169-X.

[KW64] Motoo Kimura and George H. Weiss. "The stepping stone model of population structure and the decrease of genetic correlation with distance". In: *Genetics* 49.4 (1964), p. 561.

[LD11] Heng Li and Richard Durbin. "Inference of human population history from individual whole-genome sequences". In: *Nature* 475.7357 (2011), pp. 493–496.

[Lej06] Antoine Lejay. "On the constructions of the skew Brownian motion". In: *Probability Surveys* 3 (2006), pp. 413–466. ISSN: 1549-5787.

[LG84a] Jean-François Le Gall. "One-dimensional stochastic differential equations involving the local times of the unknown process". In: *Stochastic analysis and applications*. Vol. 1095. Lecture Notes in Math. Springer, Berlin, 1984, pp. 51–82.

[LG84b] Jean-François Le Gall. "One-dimensional stochastic differential equations involving the local times of the unknown process". In: *Stochastic analysis and applications*. Vol. 1095. Lecture Notes in Math. Springer, Berlin, 1984, pp. 51–82.

[Lia09] Richard Hwa Liang. "Two continuum-sites stepping stone models in population genetics with delayed coalescence". PhD thesis. UNIVERSITY OF CALIFORNIA, BERKELEY, 2009.

[Lin88] Bruce G. Lindsay. "Composite likelihood methods". In: *Contemporary mathematics* 80.1 (1988), pp. 221–239.

[LL10] Gregory F. Lawler and Vlada Limic. *Random walk: a modern introduction*. Vol. 123. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 2010. ISBN: 978-0-521-51918-2.

[Mal48] Gustave Malécot. *Les Mathématiques de l'Hérédité*. Masson et Cie., Paris, 1948.

[MH13] Stéphanie Manel and Rolf Holderegger. "Ten years of landscape genetics". In: *Trends in Ecology & Evolution* 28.10 (2013), pp. 614–621.

[MP16] Vidyadhar Mandrekar and Andrey Pilipenko. "On a Brownian motion with a hard membrane". In: *Statistics & Probability Letters* 113 (June 2016). arXiv: 1511.01043, pp. 62–70. ISSN: 01677152.

[MT95] Carl Müeller and Roger Tribe. "Stochastic pde's arising from the long range contact and long range voter processes". In: *Probability theory and related fields* 102.4 (1995), pp. 519–545. ISSN: 0178-8051.

[Nag12a] Thomas Nagylaki. "Clines with partial panmixia". In: *Theoretical population biology* 81.1 (2012), pp. 45–68.

[Nag12b] Thomas Nagylaki. "Clines with partial panmixia in an unbounded unidimensional habitat". In: *Theoretical population biology* 82.1 (2012), pp. 22–28.

[Nag16] Thomas Nagylaki. "Clines with partial panmixia across a geographical barrier". In: *Theoretical Population Biology* 109 (June 2016), pp. 28–43. ISSN: 0040-5809.

[Nag76] Thomas Nagylaki. "Clines with Variable Migration". In: *Genetics* 83.4 (1976), pp. 867–886. ISSN: 0016-6731, 1943-2631.

[Nag78] Thomas Nagylaki. "Clines with Asymmetric Migration". In: *Genetics* 88.4 (1978), pp. 813–827. ISSN: 0016-6731, 1943-2631.

[Nag88] Thomas Nagylaki. "The influence of spatial inhomogeneities on neutral models of geographical variation. I. Formulation". In: *Theoretical Population Biology* 33.3 (1988), pp. 291–310. ISSN: 0040-5809.

[NB88] Thomas Nagylaki and Victor Barcilon. "The influence of spatial inhomogeneities on neutral models of geographical variation. II. The semi-infinite linear habitat". In: *Theoretical Population Biology* 33.3 (1988), pp. 311–343. ISSN: 0040-5809.

[NC00] Michael W. Nachman and Susan L. Crowell. "Estimate of the mutation rate per nucleotide in humans". In: *Genetics* 156.1 (2000), pp. 297–304.

[Nel+08] Matthew R. Nelson et al. "The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research". In: *The American Journal of Human Genetics* 83.3 (2008), pp. 347–358.

[NKD93] T. Nagylaki, P. T. Keenan, and T. F. Dupont. "The Influence of Spatial Inhomogeneities on Neutral Models of Geographical Variation III. Migration across a Geographical Barrier". In: *Theoretical Population Biology* 43.2 (Apr. 1993), pp. 217–249. ISSN: 0040-5809.

[Nor74a] M. Frank Norman. "A central limit theorem for Markov processes that move by small steps". In: *The Annals of Probability* 2 (1974), pp. 1065–1074.

[Nor74b] M. Frank Norman. "Markovian learning processes". In: *SIAM Review* 16.2 (1974), pp. 143–162. ISSN: 0036-1445.

[Nor75a] M. Frank Norman. "Approximation of stochastic processes by Gaussian diffusions, and applications to Wright-Fisher genetic models". In: *SIAM Journal on Applied Mathematics* 29.2 (1975). Special issue on mathematics and the social and biological sciences, pp. 225–242. ISSN: 0036-1399.

[Nor75b] M. Frank Norman. "Limit theorems for stationary distributions". In: *Advances in Applied Probability* 7.3 (1975), pp. 561–575. ISSN: 0001-8678.

[Nor77] M. Frank Norman. "Ergodicity of diffusion and temporal uniformity of diffusion approximation". In: *Journal of Applied Probability* 14.2 (1977), pp. 399–404. ISSN: 0021-9002.

[Nov+11] Dmitry S. Novikov et al. "Random walks with barriers". In: *Nature physics* 7.6 (2011), pp. 508–514.

[Num04] Esa Nummelin. *General irreducible Markov chains and non-negative operators*. Vol. 83. Cambridge University Press, 2004.

[Num78] Esa Nummelin. "A splitting technique for Harris recurrent Markov chains". In: *Probability Theory and Related Fields* 43.4 (1978), pp. 309–318. ISSN: 0178-8051.

[NZ16] Thomas Nagylaki and Kai Zeng. "Clines with partial panmixia across a geographical barrier in an environmental pocket". In: *Theoretical Population Biology* 110 (Aug. 2016), pp. 1–11. ISSN: 0040-5809.

[Pal+12] Pier Francesco Palamara et al. “Length distributions of identity by descent reveal fine-scale demographic history”. In: *The American Journal of Human Genetics* 91.5 (2012), pp. 809–822.

[Por79a] N. I. Portenko. “Diffusion processes with generalized drift coefficients”. In: *Theory of Probability & Its Applications* 24.1 (1979), pp. 62–78.

[Por79b] N. I. Portenko. “Stochastic differential equations with generalized drift vector”. In: *Theory of Probability & Its Applications* 24.2 (1979), pp. 332–347.

[PP13] Pier Francesco Palamara and Itsik Pe’er. “Inference of historical migration rates via haplotype sharing”. In: *Bioinformatics* 29.13 (July 2013), pp. i180–i188. ISSN: 1367-4803.

[PZ14] Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic Equations in Infinite Dimensions*. en. Cambridge University Press, Apr. 2014. ISBN: 978-1-107-05584-1.

[RC13] Peter Ralph and Graham Coop. “The geography of recent genetic ancestry across Europe”. In: *PLoS Biol* 11.5 (2013), e1001555.

[RCB16] Harald Ringbauer, Graham Coop, and Nick Hamilton Barton. “Inferring recent demography from isolation by distance of long shared sequence blocks”. en. In: *bioRxiv* (Sept. 2016), p. 076810.

[Reb80] Rolando Rebolledo. “Central limit theorems for local martingales”. In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 51.3 (1980), pp. 269–286. ISSN: 0044-3719.

[Ril+06] Seth PD Riley et al. “FAST-TRACK: A southern California freeway is a physical and social barrier to gene flow in carnivores”. In: *Molecular Ecology* 15.7 (2006), pp. 1733–1741.

[Rin+17] Harald Ringbauer et al. “Estimating barriers to gene flow from distorted isolation by distance patterns”. In: (2017). in preparation.

[RL07] François Rousset and Raphaël Leblois. “Likelihood and approximate likelihood analyses of genetic structure in a linear habitat: performance and robustness to model misspecification”. In: *Molecular biology and evolution* 24.12 (2007), pp. 2730–2745.

[Roa+10] Jared C. Roach et al. “Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing”. en. In: *Science* 328.5978 (Apr. 2010), pp. 636–639. ISSN: 0036-8075, 1095-9203.

[Rob70] Alan Robertson. “The reduction in fitness from genetic drift at heterotic loci in small populations”. In: *Genetical research* 15.02 (1970), pp. 257–259.

[Rou97] François Rousset. “Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance”. In: *Genetics* 145.4 (1997), pp. 1219–1228.

[RY13] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*. Vol. 293. Springer Science & Business Media, 2013.

[Saw76] Stanley Sawyer. “Results for the Stepping Stone Model for Migration in Population Genetics”. In: *The Annals of Probability* 4.5 (1976), pp. 699–728. ISSN: 0091-1798.

[Saw77] Stanley Sawyer. "Asymptotic properties of the equilibrium probability of identity in a geographically structured population". In: *Advances in Applied Probability* 9.2 (June 1977), pp. 268–282. ISSN: 0001-8678, 1475-6064.

[SB89] Montgomery Slatkin and Nicholas H. Barton. "A Comparison of Three Indirect Methods for Estimating Average Levels of Gene Flow". In: *Evolution* 43.7 (1989), pp. 1349–1368. ISSN: 0014-3820.

[SK91] S. E. Shreve and I. Karatzas. *Brownian motion and stochastic calculus*. Vol. 113. 1991.

[SKM93] Stefan G. Samko, Anatoly A. Kilbas, and Oleg I. Marichev. *Fractional integrals and derivatives*. Theory and applications, Edited and with a foreword by S. M. Nikol'skii, Translated from the 1987 Russian original, Revised by the authors. Gordon and Breach Science Publishers, Yverdon, 1993. ISBN: 2-88124-864-0.

[Sko61] A. Skorokhod. "Stochastic Equations for Diffusion Processes in a Bounded Region". In: *Theory of Probability & Its Applications* 6.3 (Jan. 1961), pp. 264–274. ISSN: 0040-585X.

[Sla73] Montgomery Slatkin. "Gene flow and selection in a cline". In: *Genetics* 75.4 (1973), pp. 733–756.

[Sla87] Montgomery Slatkin. "Gene flow and the geographic structure of natural populations". In: *Science* 236.4803 (1987), pp. 787–792. ISSN: 0036-8075, 1095-9203.

[Su+03] H. Su et al. "The Great Wall of China: a physical barrier to gene flow?" In: *Heredity* 90.3 (2003), pp. 212–219.

[VW15] Amandine Veber and Anton Wakolbinger. "The spatial Lambda-Fleming–Viot process: An event-based construction and a lookdown representation". In: *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 51.2 (2015), pp. 570–598. ISSN: 0246-0203.

[Wal78] John B. Walsh. "A diffusion with a discontinuous local time". In: *Astérisque* 52.53 (1978), pp. 37–45.

[Wal86] John B. Walsh. "An introduction to stochastic partial differential equations". In: *Ecole d'été de probabilités de Saint-Flour, XIV—1984*. Vol. 1180. Lecture Notes in Math. Springer, Berlin, 1986, pp. 265–439.

[Wri43] Sewall Wright. "Isolation by distance". In: *Genetics* 28.2 (1943), p. 114.

[Zal+09] Andrzej Zalewski et al. "Landscape barriers reduce gene flow in an invasive carnivore: geographical and local genetic structure of American mink in Scotland". In: *Molecular Ecology* 18.8 (2009), pp. 1601–1615.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Structure spatiale de la diversité génétique : influence de la sélection naturelle et d'un environnement hétérogène

Mots clés : génétique des populations, processus à valeurs mesure, coalescent spatial

Résumé : Cette thèse porte sur la structure spatiale de la diversité génétique. Dans un premier temps, nous étudions un processus à valeurs mesure décrivant l'évolution de la composition génétique d'une population soumise à la sélection naturelle. Nous montrons que ce processus satisfait un théorème de la limite centrale, et que ses fluctuations sont données par la solution d'une équation aux dérivées partielles stochastique. Nous utilisons ce résultat pour donner une estimation du fardeau de dérive au sein d'une population structurée en espace.
Dans un deuxième temps, nous nous intéressons à la composition génétique d'une population lorsque les individus se déplacent plus facilement dans une région de l'espace que dans l'autre (on parle alors de dispersion hétérogène). Nous démontrons dans ce cas la convergence des fréquences alléliques via la convergence des lignées ancestrales vers un système de mouvements browniens de Walsh.
Nous détaillons également l'impact d'une barrière géographique traversant l'habitat d'une population sur sa diversité génétique. Nous montrons que les lignées ancestrales décrivent dans ce cas des mouvements browniens partiellement réfléchis, dont nous donnons plusieurs constructions.
Dans le but d'appliquer ces travaux, nous adaptons une méthode d'inférence démographique au cas de la dispersion hétérogène. Cette méthode utilise les blocs continus de génome hérités d'un même ancêtre entre les paires d'individus dans l'échantillon et permet d'estimer les caractéristiques démographiques d'une population lorsque celles-ci varient dans l'espace. Pour terminer nous démontrons l'efficacité de notre méthode sur des données simulées.

Title : The Spatial structure of genetic diversity under natural selection and in heterogeneous environments

Keywords : population genetics, measure-valued processes, spatial coalescents

Abstract: This thesis deals with the spatial structure of genetic diversity. We first study a measure-valued process describing the evolution of the genetic composition of a population subject to natural selection. We show that this process satisfies a central limit theorem and that its fluctuations are given by the solution to a stochastic partial differential equation. We then use this result to obtain an estimate of the drift load in spatially structured populations.
Next we investigate the genetic composition of a populations whose individuals move more freely in one part of space than in the other (a situation called dispersal heterogeneity). We show in this case the convergence of allele frequencies via the convergence of ancestral lineages to a system of skew Brownian motions.

We then detail the effect of a barrier to gene flow dividing the habitat of a population. We show that ancestral lineages follow partially reflected Brownian motions, of whom we give several constructions.

To apply these results, we adapt a method for demographic inference to the setting of dispersal heterogeneity. This method makes use of long blocks of genome along which pairs of individuals share a common ancestry, and allows to estimate several demographic parameters when they vary across space. To conclude, we demonstrate the accuracy of our method on simulated datasets.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France