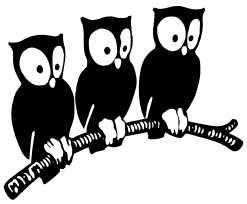
Survie et généalogies dans quelques modèles de dynamique des populations

Damien Simon

École normale supérieure
Département de Physique
Laboratoire de Physique Statistique



Thèse de doctorat
de l'Université Paris-Diderot (Paris 7)

présentée par

Damien Simon

pour obtenir le titre de
Docteur de l'Université Paris-Diderot (Paris 7)
Spécialité « physique théorique »

Survie et généalogies dans quelques modèles de dynamique des populations

soutenue le vendredi 23 mai 2008 devant le jury composé de :

Frédéric Austerlitz	**Examinateur**
Henri Berestycki	**Rapporteur**
Pierre Collet	**Rapporteur**
Francis Comets	**Examinateur**
Bernard Derrida	**Directeur**
Hendrik-Jan Hilhorst	**Examinateur**
Alan McKane	**Examinateur**

Remerciements

Cette thèse s'est déroulée de 2004 à 2008 au laboratoire de physique statistique de l'École normale, dont je tiens à remercier les directeurs successifs, Jacques Meunier et Éric Pérez, pour leur accueil. La variété des thématiques présentes dans ce laboratoire et sa proximité géographique avec les autres laboratoires de physique et de mathématiques de la Montagne Sainte-Geneviève en font un endroit exceptionnel pour s'ouvrir à tout le champ de la recherche actuelle.

Je tiens à remercier chaleureusement Bernard Derrida, qui m'a guidé dans le domaine de la physique statistique pendant ces quelques années de thèse, pour sa disponibilité, son calme et sa patience. En tant qu'enseignant à l'École normale, il m'a donné connaissances de base en physique statistique, en tant que directeur de thèse, il m'a présenté les méthodes et les approches nécessaires pour mener à bien des travaux de recherche. J'espère que ses qualités — que ce soit sa capacité à aborder un même problème de multiples points de vue ou sa rigueur dans la rédaction — auront un peu déteint sur moi. Je conserverai longtemps l'excellent souvenir de toutes les heures passées dans son bureau, à discuter de sujets scientifiques, ou de tout autre chose.

Je voudrais également remercier Henri Berestycki et Pierre Collet, qui m'ont fait l'honneur d'accepter la lourde tâche d'être rapporteurs de la présente thèse, mais aussi tous les autres membres du jury — Frédéric Austerlitz, Francis Comets, Henk Hilhorst et Alan McKane — pour l'intérêt qu'ils ont manifesté pour cette thèse.

Je remercie aussi tout spécialement Aurélie et Joachim pour leurs relectures efficaces lors de la fin de la rédaction de ce mémoire.

Au sein du LPS, j'ai côtoyé de nombreux chercheurs avec lesquels j'ai interagi avec plaisir pendant ces quelques années, autour d'un café ou d'un sandwich, dans leur bureau ou en conseil de laboratoire : je tiens à les remercier tous — à défaut de pouvoir citer chacun d'entre eux — pour leur accueil et leur sympathie. Pour que le travail scientifique progresse, il faut que la logistique suive : j'exprime donc toute ma reconnaissance aux secrétaires du LPS, M. Gefflot, A. Ribaudeau et N. Sadaoui, pour leur aide dans les formalités administratives et aussi à F. Ayrault, Z. Dissi et R. Portier pour leurs compétences en informatique. Un avantage de plus à être thésard dans un lieu comme le département de physique de l'É.N.S. est de côtoyer tant d'autres étudiants avec qui j'ai pu partager mes questions, qu'elles soient existentielles ou purement techniques, aussi bien que mes éclats de rire : je les remercie pour toute la bonne humeur que nous avons

partagée. Je pense en particulier aux « bosons » avec qui j'ai cohabité en DC21 : je remercie les anciens, Camille, Chi-Tuong, Paul et Éric pour leurs conseils, ceux avec qui une bonne partie de la route a été faite, Laurent, Hervé, Grégoire, Olaf, Carlos, Stéphane, Serena et Nestor pour la bonne ambiance qu'ils ont su instaurer, et je souhaite bonne chance à ceux qui prennent la relève, Sébastien, Laetitia et Miguel. Bonne route aussi à Antoine, que j'ai accueilli comme élève au début de ma thèse et qui prend la relève avec Bernard.

Ces années de thèse furent aussi des années d'enseignement. Je tiens ainsi à remercier toute l'équipe enseignante et les « caïmans » du département de physique, et tout particulièrement Vincent Rivasseau et Armel Martin, Stefan Fauve et Christophe Coste, et Noëlle Pottier et Frédéric van Wijland. Enseigner, c'est avant tout se confronter à un public : je remercie tous les élèves qui ont rendu cette tâche si agréable et je leur souhaite une bonne continuation. Enfin, la saison estivale fut aussi celle du concours d'entrée de l'É.N.S. : très loin d'être une corvée, le secrétariat du concours a pour moi été un bol d'air de sociabilité chaque mois de juillet. Merci à tous ceux que j'ai côtoyés dans ce contexte et qui me laissent d'excellents souvenirs.

Si le contexte professionnel est important dans la réalisation d'une thèse, la sphère personnelle l'est tout autant. Mes premiers remerciements vont à mes parents pour tout ce qu'ils m'ont apporté et continuent de m'apporter : ce sont eux qui ont le plus contribué à ma réussite et je leur en suis extrêmement reconnaissant. Je tiens aussi à remercier toute ma famille pour les encouragements que j'ai reçus d'elle depuis tant d'années. Parce que la thèse est aussi l'achèvement d'un cursus scolaire, j'adresse mes plus vifs remerciements et toute ma gratitude à tous les professeurs qui m'ont fait confiance, qui m'ont poussé à donner toujours plus et m'ont donné le goût des études et de la culture ; j'espère, comme enseignant, être digne d'eux et réussir à transmettre autant de choses qu'eux ont su le faire. Enfin, mes derniers remerciements, et non les moindres, vont à tous mes amis, pour tous les bons moments que j'ai partagés avec eux, en Auvergne, à Paris et ailleurs, pour tout le soutien qu'ils m'ont apporté et la source constante de motivation qu'ils ont constituée. Si je ne donne pas de noms, c'est pour ne pas faire de jaloux, ni de favorisé : les intéressés se reconnaîtront. Enfin, à toutes les autres personnes qui, même si elles ne sont pas nécessairement conscientes du rôle qu'elles ont joué, ont contribué à ce que, au final, je puisse écrire ces remerciements, merci. Pour finir, je remercie de tout mon cœur Aurélie pour sa présence durant ces années.

Table des matières

Introduction

Cette thèse est consacrée à l'étude de différentes propriétés de modèles de populations biologiques qui évoluent selon des dynamiques simples. Traditionnellement, la physique statistique décrit des systèmes composés d'un grand nombre d'objets en interaction, comme les molécules d'un gaz : étant donné une description microscopique des dynamiques individuelles, la physique statistique a pour but d'étudier les comportements collectifs qui émergent des « règles du jeu » microscopiques lorsque la taille du système est grande. Plus récemment, la physique statistique s'est progressivement étendue à l'étude d'autres systèmes en-dehors de la physique, en particulier de nombreux modèles récents sont inspirés par la biologie.

La physique statistique a ses origines dans l'étude des gaz et de la chaleur. Au XIX^{e} siècle s'est imposée au fur et à mesure l'idée que la matière est constituée d'entités élémentaires appelées « atomes » et que la température correspond à leur agitation. La théorie cinétique des gaz, à laquelle ont contribué aussi bien Bernoulli et Joule que Clausius, Maxwell, Boltzmann et Gibbs, a permis progressivement de faire le lien entre les quantités macroscopiques auxquelles étaient habitués les physiciens de l'époque, telles que pression et température, et une description microscopique, newtonienne, des mouvements des particules. La deuxième moitié du XIX^{e} siècle a ainsi vu émerger la description probabiliste, utilisée depuis, ainsi que l'introduction du concept d'entropie. Durant tout le XX^{e} siècle, la physique statistique a pu être appliquée avec succès à tous les types de systèmes physiques rencontrés, quantiques ou classiques et devenir ainsi la seule branche de la physique sans objet d'étude unique mais avec un outil spécifique : les probabilités.

Si les systèmes étudiés initialement avaient une dynamique newtonienne avec une énergie et des constantes du mouvement, très vite, le formalisme probabiliste a pu être appliqué aux systèmes dont la dynamique n'était plus déterministe mais aléatoire elle-même, comme en témoigne l'étude du mouvement brownien par Einstein en 1905. Les physiciens purent ainsi aborder tous les systèmes pour lesquels seule une information partielle sur leur dynamique est disponible, quitte à décrire la partie inconnue de la dynamique par des variables aléatoires. La pertinence de la description par des variables aléatoires de l'information manquante n'est pas absolue mais donne de très bons résultats pour de nombreux systèmes. Les processus stochastiques ont alors pris leur essor en mathématiques, en particulier depuis l'introduction du formalisme des chaînes de Markov en 1906.

Physique statistique et processus stochastiques ont donc été utilisés pour décrire tout

type de situations et on peut s'émerveiller du fait que ces domaines soient encore très vivants. Si d'un côté les divers domaines de la physique continuent de fournir des problèmes toujours plus complexes aux physiciens statisticiens, un engouement supplémentaire est venu de la biologie. Celle-ci, dans les quarante dernières années, a connu un développement rapide et les biologistes sont passés d'une connaissance qualitative des êtres vivants à une connaissance plus quantitative. Tout d'abord, la biologie moléculaire a rapproché des chimistes et physiciens les objets d'étude des biologistes et la précision requise expérimentalement nécessite l'adaptation de techniques issues de la physique (marquage de molécules, déplacement d'objets microscopiques, etc.). Dans une seconde phase, les résultats expérimentaux ont mis en évidence l'extraordinaire complexité des mécanismes microscopiques à la base des fonctions vitales des individus : les outils statistiques deviennent alors un moyen privilégié pour comprendre comment toutes les briques élémentaires se combinent pour produire ce que nous observons à l'échelle de l'individu entier.

Un second tournant, et c'est celui sur lequel se base en partie cette thèse, est la formulation de la théorie de l'Évolution par C. Darwin dans la deuxième moitié du XIX^{e} siècle : sous sa forme initiale, cette théorie permet seulement de comprendre qualitativement pourquoi et comment de nouvelles espèces apparaissent au cours du temps alors que d'autres s'éteignent. En se basant sur des critères anatomiques, les scientifiques ont alors esquissé des arbres phylogénétiques entre les espèces afin de comprendre leur relation de parenté. Les premiers modèles de population datent de la même époque mais n'intéressaient alors majoritairement que les mathématiciens. Ce n'est que depuis une trentaine d'années et le séquençage de l'ADN que des données quantitatives existent sur les relations entre espèces : la masse de données non seulement nécessite des outils informatiques sophistiqués mais aussi permet de confronter à l'expérience de nombreux modèles. En particulier, depuis très récemment, des biologistes [BLQ^{+}08] pensent pouvoir lire les effets de la pression de sélection dans les mutations observées sur certains gènes.

Longtemps, l'intérêt pour ces problèmes d'inspiration biologique est resté l'apanage des biologistes et des mathématiciens. La physique statistique s'est tournée vers les modèles d'évolution au moment où les physiciens ont pris conscience des nombreuses similarités de ces modèles avec des systèmes de physique étudiés à la même époque, comme nous le verrons au chapitre 1. De nombreux outils étaient déjà prêts à l'emploi et ont donné lieu à de nombreux résultats sur des systèmes très variés [Pel97, SH05]. Nous pouvons citer par exemple les processus de réaction-diffusion, largement étudiés en physique, qui peuvent tout aussi bien décrire la prolifération de bactéries ou la propagation d'épidémie (avec ou sans guérison, avec ou sans immunisation). Dans de tels modèles, les atomes et leurs niveaux d'excitation sont remplacés par les individus et leur état de santé. Du point de vue statistique, la seule connaissance nécessaire est la connaissance des règles microscopiques d'évolution des entités, à partir desquelles sont déduits les comportements collectifs. La première partie de cette thèse s'inscrit dans cette optique et étudie l'évolution d'une population biologique dont on a modélisé la reproduction et la mort des individus par des réactions de type $A \to 2A$ ou $A \to \emptyset$; la question de la survie ou de l'extinction se transpose alors dans le langage de la physique statistique sous le nom de « transition de phase vers un état absorbant », alors que les individus

et leurs enfants deviennent dans ce langage des marches aléatoires avec branchement. Les quantités qui intéressent traditionnellement les physiciens sont les échelles typiques, spatiales et temporelles, qui apparaissent dans ces systèmes, de même que l'étude des transitions de phase qui sont des seuils qui séparent des comportements macroscopiques distincts.

Dans une première partie, bien que ce que ne fut pas la démarche strictement chronologique de nos travaux, nous nous sommes intéressés aux probabilités de survie d'une population en présence de sélection. Si les propriétés des populations qui évoluent sans sélection (neutres) sont bien comprises, les difficultés liées à la prise en compte de la sélection font que l'étude des population sous sélection est bien plus récente et constitue, aujourd'hui, un domaine actif de la dynamique des populations. Nous nous sommes penchés sur la question de la survie d'une population lorsque la pression de la sélection peut être modélisée par un seuil d'adaptabilité qui augmente avec le temps et l'effet des mutations génétiques par une marche aléatoire de l'adaptabilité d'un individu. Au-delà de la modélisation par des marches aléatoires, le problème peut se ramener à l'étude d'équations (non-linéaires) de propagation de fronts. Ceux-ci apparaissent de manière récurrente en physique, que ce soit pour décrire la propagation d'une combustion ou la manière dont une instabilité se déplace dans un milieu et nous avons adapté un certain nombre de ces outils à la description, nécessairement simpliste, de populations en présence de sélection.

Dans un second temps, ce sont les propriétés des généalogies des individus qui ont attiré notre attention. Non seulement la question des généalogies est inhérente à toute étude d'individus qui se reproduisent mais, de plus, de récents travaux ont montré des connexions avec une certaine classe de systèmes désordonnés traditionnels de la physique. D'un point de vue purement conceptuel, ces systèmes physiques (polymères dirigés, croissance de surface, etc.) peuvent être reformulés en termes de sélection de configurations qui maximisent certains critères. Parmi les similarités entre populations biologiques et certains systèmes désordonnés que nous avons explorées, nous nous sommes plus particulièrement focalisés sur les âges des ancêtres communs les plus récents de groupes d'individus. Plus précisément, à cause de la reproduction et de l'extinction de lignées, tous les individus descendent d'un même ancêtre à condition de remonter suffisamment loin dans le temps : nous nous sommes attachés à étudier quelques aspects statistiques de la durée qui sépare les individus d'une génération de leur dernier branchement commun.

Ce mémoire est divisé en deux parties, chacune traitant d'un aspect spécifique du travail effectué : la survie d'une population sous sélection et sa description par des équations de front d'une part (chapitres 1 à 7), les généalogies et les temps de coalescence avec et sans sélection d'autre part (chapitres 8 à 11).

Les deux premiers chapitres sont deux introductions aux domaines de la physique statistique sur lesquels sont basés les chapitres ultérieurs et se veulent être indépendants. Le chapitre 1 présente les modèles de population qui sont parmi les plus communs et que nous avons utilisés. Autant que possible, les motivations biologiques sous-jacentes sont mises en parallèle avec les concepts de physique statistique correspondants, afin de se familiariser avec le vocabulaire, les notations et le formalisme. Le chapitre 2 est une introduction à la physique de la propagation des fronts : nous y présentons le type d'équations

non-linéaires et le type de solutions utilisées par la suite ainsi que les résultats existants. En particulier, nous nous sommes intéressés plus précisément à la détermination de la vitesse de propagation d'un front ainsi qu'à ses temps caractéristiques de relaxation après une perturbation localisée. Le point de vue est celui du physicien bien que de nombreux résultats mathématiques existent aujourd'hui : la présentation de ces derniers se réduit ici à une compréhension physique de leur signification. Ces deux premiers chapitres ne comportent donc pas de résultats originaux, si ce n'est la méthode de calcul perturbative des temps de relaxations d'un front en présence de conditions aux bords présentée en section 2.4.2 : cette méthode a été développée pour les besoins spécifiques de notre modèle et sa portée plus large justifie son inclusion dans un chapitre général traitant de la propagation de fronts.

Les chapitres 3, 4, 5, 6 et 7 contiennent la partie originale de notre travail. Le chapitre 3 commence par définir le modèle de marches aléatoires avec branchements (la reproduction) et bords absorbants (la sélection) que nous avons étudié et se focalise ensuite sur la transition de phase vers l'extinction que subit la population lorsque la sélection devient trop forte. Les chapitres 4 et 5 étudient le même modèle lorsque la population est conditionnée à avoir une taille finale finie donnée. La dynamique « libre » conduit soit à une extinction, soit à une croissance exponentielle de la population : seul un petit nombre d'évolutions conduisent à une taille finale finie aux temps longs. Nous nous sommes attachés à étudier ce régime conditionné, en lien avec d'autres modèles où la taille est fixée. Le chapitre 4, après avoir rappelé un certain nombre de résultats existants, montre comment ces résultats peuvent être généralisés au modèle que nous avons étudié et montre qu'un régime quasi-stationnaire émerge d'un côté du point critique de la transition de phase étudiée au chapitre 3. Le chapitre 5 montre comment construire un processus biaisé qui est strictement équivalent au modèle de départ conditionné par une taille finale donnée. Ce processus biaisé permet essentiellement, d'une part, de produire des simulations numériques qui évitent l'état absorbant pendant la durée de la simulation, et, d'autre part, d'étudier plus facilement le régime quasi-stationnaire. L'étude de ce dernier autour du point critique mis en évidence au chapitre 3 fait l'objet du chapitre 6, dans lequel l'accent est mis sur le caractère universel d'un certain nombre de quantités. Enfin, le chapitre 7 étudie un modèle particulier, le modèle exponentiel, qui se révèle être exactement soluble et qui permet ainsi d'aller plus loin que les chapitres précédents dans l'étude du régime quasi-stationnaire.

La seconde partie s'articule autour de la caractérisation des généalogies dans divers modèles de population. Le chapitre 8 est une introduction générale aux processus de coalescence. Les cas particuliers des coalescents de Kingman et Bolthausen-Sznitman sont étudiés plus minutieusement et une revue est faite des situations connues où ils apparaissent naturellement. Ce chapitre ne contient aucune contribution originale et vise juste à remettre en perspective les chapitres suivants.

Le chapitre 9 est une étude, principalement numérique, des modifications qui apparaissent dans les généalogies lorsque l'on introduit une structure spatiale : les individus se meuvent sur un réseau mais ne peuvent subir de coalescence que s'ils sont sur le même site. Nous passons en revue les résultats connus sans sélection puis nous montrons dans quelle mesure les modèles de population en présence de sélection sont similaires aux polymères dirigés ou aux processus de croissance de surface. Enfin, nous étudions

les généalogies dans ces modèles issus de la physique statistique et montrons comment ils s'insèrent dans les modèles de coalescence introduits précédemment. Ces derniers résultats font partie d'un travail en cours de rédaction.

Les deux derniers chapitres 10 et 11 reviennent sur des modèles de population sans sélection, dont les propriétés stationnaires des généalogies sont bien connues. Le chapitre 10 décrit notre travail sur les propriétés *dynamiques* des généalogies dans le modèle de Wright-Fisher et présente la construction d'un processus de Markov sur les temps de coalescence que nous utilisons pour caractériser les fluctuations temporelles de l'âge de l'ancêtre commun d'une population. Enfin, le chapitre 11 fait le lien entre généalogies et diversité génétique : après une brève introduction sur la notion de diversité génétique, nous montrons comment la connaissance des génomes de quelques individus dans une population peut modifier l'estimation bayesienne de l'âge de l'ancêtre commun le plus récent de celle-ci.

Première partie

Modèles de populations en évolution

Chapitre 1

Populations et sélection

Ce chapitre introduit les notions et les motivations biologiques sous-jacentes au reste de l'exposé ainsi que quelques modèles parmi les plus simples et les plus utilisés dans les approches physiques et mathématiques de l'évolution des populations. La première section décrit le contexte biologique et la théorie de l'Évolution ; la seconde section décrit les modèles de reproduction couramment utilisés (Galton-Watson, Wright-Fisher, etc.) et la troisième décrit différentes manières de modéliser simplement l'effet de la sélection sur une population. La quatrième section montre comment les idées précédentes ont ensemencé d'autres domaines, comme l'informatique et certaines techniques numériques en physique statistique. Ce chapitre ne contient aucun résultat original.

1.1 Cadre biologique

1.1.1 La théorie de l'Évolution

Les bases de la théorie de l'Évolution des espèces biologiques telle qu'on la connaît aujourd'hui remonte à 1859, lorsque Charles Darwin (voir figure 1.1) publia son célèbre ouvrage *L'origine des espèces* suite à son voyage autour du monde à bord du *Beagle* en tant que naturaliste. Il énonce dans cet ouvrage les lois de la sélection naturelle qui, selon lui, explique la variété d'espèces biologiques, tant celles que l'on observe actuellement que celles dont l'étude des fossiles nous apprend l'existence par le passé. Ces lois sont aujourd'hui largement acceptées. La sélection naturelle explique comment les espèces se créent et disparaissent en tentant de s'adapter à leur milieu environnant. Plus précisément, l'un de ses mécanismes principaux est la *lutte pour la survie* qui émerge entre les individus d'une même espèce ou d'espèces différentes dans des milieux où les ressources sont rares. Cette lutte pour la survie a pour conséquence de sélectionner les individus qui sont les mieux adaptés à l'environnement du moment. D'autre part, l'accumulation, par transmission héréditaire, de transformations favorables différentes d'un individu à

un autre, a pour conséquence l'émergence de nouvelles espèces à partir d'une même population initiale.

Cependant, l'origine des variations entre individus et leur transmission héréditaire étaient encore inconnues au XIX^{e} siècle. Ainsi, Darwin les considère comme spontanées et s'appuie sur les idées de Lamarck pour tenter d'expliquer l'apparition et la disparation d'organes devenus utiles ou inutiles. Ce n'est qu'au XX^{e} siècle, avec l'avènement de la génétique et la découverte de l'acide désoxyribonucléique (ADN) et de ses propriétés, que les variations entre individus sont expliquées par l'apparition de mutations sur l'ADN d'un individu. Depuis, la théorie de l'évolution darwinienne a été affinée et développée bien au-delà de l'ouvrage fondateur de Darwin et on pourra se reporter à [Gou02] pour une description actuelle détaillée de la théorie.

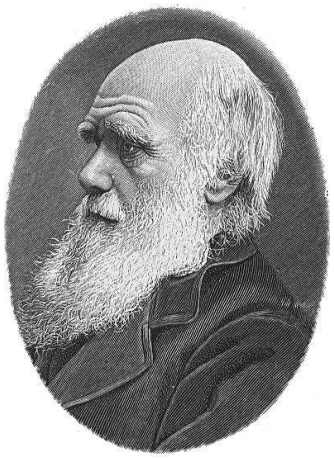


FIG. 1.1: Charles Darwin (1809–1882), biologiste et naturaliste britannique, auteur de *l'Origine des espèces*.

La révolution qui a suivi la découverte de l'ADN a permis non seulement d'élucider les mécanismes à la base de la théorie de Darwin mais aussi de *quantifier* les différences entre espèces. En effet, chaque individu est caractérisé par sa séquence ADN, c'est-à-dire par une séquence de quatre lettres A, G, C, T. Chaque gène peut apparaître sous la forme de différents allèles, correspondant à des séquences ADN légèrement différentes. Une *distance* entre allèles (et donc entre individus) peut alors être définie en comptant le nombre de différences entre les deux séquences. Ces différences étant liées aux mutations qui se sont produites dans les lignées, elles sont un vestige quantitatif de l'évolution des lignées depuis un ancêtre commun ancien et permettent, idéalement, de combler certaines lacunes dans les fossiles biologiques. En particulier, elles permettent une reconstruction d'arbres phylogénétiques qui relient les espèces les unes aux autres et estiment les âges de leurs ancêtres communs les plus récents.

Les briques microscopiques du modèle (reproduction, mutations de différents types, etc.) étant à présent relativement bien connues, de multiples modèles ont été développés pour reproduire la diversité des espèces observées et leurs origines. Cependant, la grande variété de mécanismes et de situations possibles rend l'analyse ardue et on a recours à des modèles simplifiés pour expliquer et reproduire les différentes observations des biologistes et naturalistes. De là sont nées les idées de *populations structurées* et de *fitness* décrites dans la section suivante et le problème est ainsi passé entre les mains des physiciens et mathématiciens [Pel97] qui ont pu mener des calculs élaborés sur les modèles les plus simples présentés ci-dessous.

1.1.2 Notion de population structurée

Même si l'ADN des individus peut être aujourd'hui séquencé, le lien entre les séquences ADN et les fonctions des individus n'est pas si simple. Or ce qui contrôle l'évolution et la survie d'un individu dans son milieu est l'ensemble des fonctions qu'il peut y effectuer. De plus, à séquences ADN identiques, une multiplicité d'autres facteurs peut affecter les fonctions d'un individu, comme l'âge ou les maladies. Loin des mécanismes moléculaires de l'ADN, il faut donc adopter une description heuristique de l'évolution des fonctions d'un individu. Ainsi, dans un souci de simplification, un individu ne sera pas caractérisé par une séquence de millions de bases A, G, C et T mais par un *état interne* plus simple à décrire. D'autre part, l'environnement géographique de l'individu influe à la fois sur sa survie et sur sa proximité avec les autres individus. Une population structurée [WH98, MD86] est une population où les individus sont caractérisés par un jeu de paramètres internes (âge, emplacement, capacité à accomplir une action) pertinents dans la situation que l'on cherche à décrire.

Un exemple simple de population structurée est la répartition des individus en classes d'âge : à chaque âge sont associées certaines qualités, telles la capacité de reproduction et la capacité à survivre jusqu'à atteindre la classe d'âge supérieure. Une modélisation commune en écologie est celle des *modèles de Leslie* [Les45] où à chaque classe d'âge sont attribués deux nombres, le taux de fécondité et le taux de survie. La description du nombre moyen d'individus dans chaque classe aux temps longs s'obtient par un simple exercice d'algèbre linéaire.

Le deuxième type de population structurée très répandu est le modèle de *métapopulations* et schématise la répartition spatiale des individus : les individus sont regroupés par *dèmes* [WA01] ou *métapopulations* de telle sorte qu'à l'intérieur d'un même groupe, ils interagissent tous les uns avec les autres et que, d'un groupe à l'autre, ils n'interagissent pas. Les seuls contacts entre métapopulations correspondent aux individus qui voyagent d'un groupe à un autre. Cette approche n'est qu'une approximation mais elle permet, par exemple, de décrire de manière satisfaisante certaines situations de propagation d'épidémies [CBBV07], dès lors qu'une approche de type *champ moyen* suffit pour décrire les interactions entre individus d'une même ville ou d'un même pays. D'autre part, dans un contexte génétique, la séparation géographique peut mener à l'émergence de nouvelles espèces locales, comme l'a observé Darwin chez certains oiseaux des îles Galapagos : dans le cas de faibles flux de migrations, la variabilité génétique est plus faible à l'intérieur d'un dème qu'entre deux dèmes distincts, si bien que l'accumulation locale de mutations peut mener à l'apparition de nouvelles espèces locales.

Mathématiquement, les populations structurées sont décrites par des graphes dont les sommets correspondent aux états des individus (localisation géographique, classes d'âge, etc.) et dont les arêtes décrivent l'évolution des individus d'un état à un autre. C'est pour cette raison qu'un langage *géographique* sera utilisé par la suite et qu'un certain nombre de résultats dérivés pour des géométries particulières pourront être adaptés à de tels modèles.

Malheureusement, si elle est formulée trop simplement, cette approche ne donne que des valeurs moyennes pour les tailles de population et suppose que les individus sont indépendants : il n'y a donc ni saturation de la population, ni compétition entre espèces.

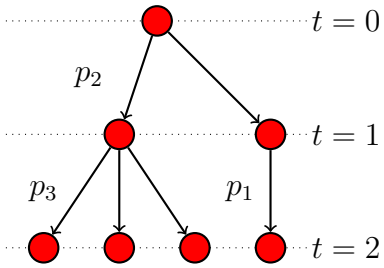


FIG. 1.2: Reproduction dans le modèle de Galton-Watson : à chaque pas de temps, chaque individu se divise indépendamment des autres en k individus avec une probabilité p_k.

Elle est cependant largement utilisée en écologie pour prédire les répartitions en classes d'âge et les densités de peuplement, mais aussi l'effet d'une modification du milieu. En cas de compétition (systèmes de prédateur-proie), les équations d'évolution deviennent non-linéaires (couplages entre les espèces) et l'évolution du système ne se réduit plus à de l'algèbre linéaire mais à un système dynamique plus complexe, avec potentiellement des points fixes et des cycles limites (cf. les équations de Lotka-Volterra par exemple).

1.2 Modèles simples de reproduction sans sélection

Nous présentons ici les modèles les plus répandus de reproduction de population que nous utiliserons par la suite et en profitons pour expliquer les quantités pertinentes pour chacun d'entre eux.

1.2.1 Individus indépendants : le processus de Galton-Watson

Le processus de Galton-Watson est l'un des processus stochastiques les plus élémentaires pour décrire la reproduction des individus d'une population. Il fut introduit pour la première fois par Francis Galton en 1873 pour expliquer l'extinction des surnoms dans l'aristocratie victorienne puis il fut résolu, presque immédiatement, par lui-même et le révérend H. R. Watson [GW74]. L'aspect stochastique permet de prendre en compte la variabilité du nombre d'enfants d'un individu et, ainsi, d'étudier l'évolution moyenne de la taille d'une population et ses fluctuations.

Dans ce modèle de reproduction, les individus sont supposés indépendants et ont les mêmes propriétés. À chaque génération, chaque individu disparaît et laisse place à k descendants, où k est une variable aléatoire à valeurs entières distribuée selon une loi p_k. Le cas $k = 0$ correspond à la mort de l'individu et à l'extinction de sa lignée. Pour $k = 1$, il n'y a aucun changement dans la taille de la population : l'individu survit sans se diviser. La simplicité du processus fait que de nombreux résultats sont connus pour ce modèle et nous nous proposons d'en faire une brève revue.

L'évolution de la taille moyenne N_t de la population est donnée par :

$$\langle N_{t+1}\rangle = \left(\sum_k k p_k\right)\langle N_t\rangle = \bar{k}\langle N_t\rangle \tag{1.1}$$

Aux temps longs, deux comportements sont donc possibles selon la valeur du nombre moyen $\bar{k}$ d'enfants par individu : si $\bar{k} < 1$, la taille moyenne décroît exponentiellement vers 0 alors que pour $\bar{k} > 1$, elle croît exponentiellement. Cela met déjà en évidence l'existence de deux régimes possibles séparés par la valeur seuil $\bar{k} = 1$.

Cette transition s'observe aussi sur une quantité qui nous intéressera par la suite : la probabilité d'extinction $Q_e(t)$ d'un individu c'est-à-dire la probabilité que toute la descendance d'un individu initial à $t = 0$ se soit éteinte avant l'instant t. Autrement dit, $Q_e(t)$ est la probabilité que $N_t = 0$ sachant que $N_0 = 1$. Dans le cas d'individus indépendants, nous pouvons écrire une équation très simple sur l'évolution de $Q_e(t)$. Pour cela, séparons la durée t entre la première génération et les $t-1$ suivantes : durant la première génération, l'individu initial se divise en k individus avec une probabilité p_k. La probabilité que l'individu initial se soit éteint au temps t est donc égale à la probabilité que les lignées de ses k enfants se soient éteintes pendant les $t-1$ générations suivantes. Durant celles-ci, les k individus vont évoluer indépendamment et la probabilité que toutes leurs lignées se soient éteintes est donc égale au produit des probabilités d'extinction de chacune des lignées. Mathématiquement, nous avons l'équation de récurrence suivante pour $Q_e(t)$:

$$Q_e(t) = \sum_k p_k Q_e(t-1)^k \tag{1.2}$$

avec la condition initiale $Q_e(0) = 0$, puisque nous commençons avec une population de taille $N_0 = 1 \neq 0$.

Cette récurrence est du type : $Q_e(t) = F_e\left(Q_e(t-1)\right)$ avec une fonction d'itération F_e donnée par :

$$F_e(Q) = \sum_k p_k Q^k \tag{1.3}$$

qui n'est autre que la fonction génératrice des nombres p_k. Elle satisfait, en particulier, les propriétés suivantes :

$$\begin{cases} F_e(0) = p_0, & \text{(probabilité de mort)} \\ F_e(1) = 1, & \text{(normalisation des probabilités)} \\ F_e'(1) = \bar{k}, & \text{(nombre moyen d'enfants)} \\ F_e''(Q) \geq 0. & \text{(convexité)} \end{cases} \tag{1.4}$$

L'étude de la probabilité de survie se ramène ainsi à l'étude du système dynamique spécifié par (1.2). Le comportement aux temps longs de $Q_e(t)$ découle des propriétés de la fonction F_e : pour $t \to \infty$, $Q_e(t)$ tend vers le seul point fixe stable Q_e^* de F_e dans $[0, 1]$. Celui-ci dépend de la valeur de $\bar{k}$:

- pour $\bar{k} \leq 1$, $Q_e^* = 1$: la lignée de l'individu initiale est sûre de s'éteindre aux temps longs et la probabilité de tomber dans l'état absorbant est donc 1 ;

– pour $\bar{k} > 1$, on a $0 \leq Q_e^* < 1$ et le système a une probabilité non nulle d'échapper à l'état absorbant $N = 0$ aux temps longs : si elle survie, la population croît alors exponentiellement.

Cet exemple simple montre comment l'étude de la transition vers la phase inactive (population éteinte) se ramène à l'étude d'un système dynamique. Nous verrons dans le chapitre 3 comment cette approche se généralise en présence d'une structure spatiale.

Dans la version du modèle en temps continu, les probabilités p_k sont remplacées par les taux β_k définis de la manière suivante : pendant un intervalle dt, chaque individu a une probabilité $\beta_k dt$ de disparaître et de laisser place à k enfants. Cela se reformule en termes de processus de réaction-diffusion [Hin00, Odo04] en disant qu'un individu se divise selon l'une des réactions :

$$A \xrightarrow{\beta_k} kA$$

et la population vide $N_t = 0$ est le seul état absorbant. La probabilité d'extinction $Q_e(t)$ satisfait à présent une équation différentielle $\partial_t Q_e = f_e(Q_e)$ avec $f_e(Q) = \sum_k \beta_k (Q^k - Q)$ et tend vers Q_e^* tel que $f_e(Q_e^*) = 0$.

1.2.2 Saturation et taille constante : les modèles de Wright-Fisher et Moran

Dans le processus de Galton-Watson, les individus sont indépendants et la population peut croître indéfiniment : cette modélisation ne s'applique qu'aux milieux dans lesquels la taille n'est pas limitée par les ressources. Pour prendre en compte cette limitation, nous supposerons que la population est arrivée à saturation et la taille de la population est maintenue constante par un processus externe que nous ne nous attacherons pas à décrire. Une description parmi les plus simples de la reproduction dans une telle population est le modèle de Wright-Fisher [Fis30, Wri31].

Dans ce modèle, les générations ne se recouvrent pas et la population a une taille fixée N. De la génération t à $t+1$, tous les individus sont renouvelés. Chaque individu à $t+1$ a un parent tiré aléatoirement dans la génération précédente.

Dans la limite de grande taille N, un individu au temps t a ainsi un nombre d'enfants dans la génération $t+1$ distribué selon une loi poissonienne. En particulier, le nombre moyen d'enfants par individu est 1 (taille constante) et la fraction moyenne d'individus n'ayant pas d'enfants dans la génération suivante est égale à e^{-1} dans la limite de grande population.

L'aspect non réaliste des générations qui ne se recouvrent pas disparaît dans le modèle de Moran [Mor58]. Là encore, la taille de la population reste fixée égale à N. À chaque pas de temps, une paire d'individus est tirée aléatoirement uniformément dans la population : le premier d'entre eux se divise en deux individus et le deuxième meurt (si le même individu est choisi lors des deux tirages, alors rien ne change). Le modèle de Moran présente plusieurs avantages. Dans certains cas, il permet d'écrire des récurrences plus simples à résoudre et permet de passer plus facilement à la limite de temps continu[1].

[1] Pour cela, il suffit de tirer une paire d'individus avec une probabilité αdt pendant un intervalle de temps infinitésimal dt.

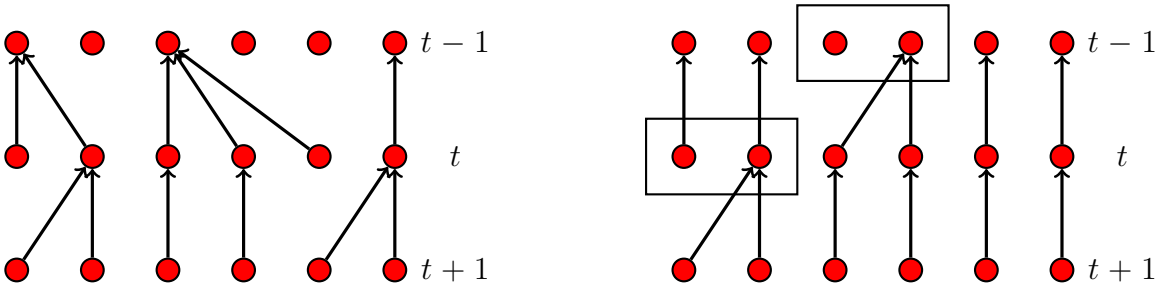


FIG. 1.3: Reproduction dans les modèles de Wright-Fisher (gauche) et Moran (droite) pour une taille de population $N = 6$.

Il est possible d'introduire facilement de la sélection dans ces deux modèles en brisant l'uniformité du tirage aléatoire du parent d'un individu. Pour cela il suffit d'affecter un coefficient w_i à chaque individu d'une génération et il est choisi comme parent d'un individu de la génération suivante avec une probabilité $w_i/(\sum_j w_j)$. En particulier, cela permet de décrire comment un allèle bénéfique A (de poids w_A) porté initialement par un unique individu qui vient de subir une mutation envahit une population d'individus portant l'allèle B tel que $w_B < w_A$. Le *temps de fixation* τ_{fix} est alors le temps moyen au bout duquel toute la population porte l'allèle A. De nombreuses variantes découlent de ses modèles et permettent d'étudier la manière dont certaines caractéristiques sont répandues dans la population (cf. chapitre 11).

Ces deux types de dynamique sont généralement choisis dans les modèles de populations structurées où les métapopulations sont supposées de taille constante au cours du temps. Les échanges entre métapopulations correspondent alors à l'échange des positions de deux individus. Bien qu'ils ne permettent pas de décrire la reproduction sexuée, ils s'appliquent cependant aux populations sexuées lorsque l'on ne considère que les caractères ne provenant que d'un seul des deux parents (chromosome Y et noms de famille chez les hommes, ADN mitochondrial chez les femmes, etc.). La reproduction sexuée complique l'étude des parentés : lorsque l'on remonte les lignées, le nombre d'ancêtres d'un groupe d'individus en reproduction sexuée ne décroît pas strictement vers 1, alors qu'en reproduction asexuée, le nombre d'ancêtres d'un groupe est décroissant dans le temps et tend vers 1 loin dans le passé.

1.3 Sélection et *fitness*

1.3.1 Évolution et physique statistique

Dans les modèles précédents, la sélection naturelle a été négligée. Celle-ci revient à classer les individus selon leur niveau d'adaptation à leur environnement et à les faire évoluer selon ce classement. En particulier, on s'attend à ce qu'un individu génère une proportion future de la population d'autant plus grande qu'il était adapté à son milieu. De plus, l'apparition de mutations favorables met le système hors d'équilibre : des

individus de plus en plus adaptés apparaissent continuellement éliminant par compétition les moins adaptés et la population *progresse* en moyenne au cours du temps.

Le moyen le plus simple est d'attribuer à chaque *phénotype*, c'est-à-dire à l'ensemble des traits d'un individu, un nombre, appelé *fitness* ou « trait » dans la littérature, qui caractérise son adaptation au milieu environnant. Ce trait structure alors la population en classes, des individus les plus adaptés à leur milieu et qui ont le plus de chances de survie aux individus les moins adaptés voués à une disparition rapide sous l'effet de la compétition. Communément, le *fitness* d'un phénotype est défini comme le nombre moyen attendu[2] d'enfants d'un individu ayant un tel phénotype. Pour un individu décrit par un état interne $\mathcal{C}$ (son phénotype), le *fitness* peut être représenté par une fonction $f(\mathcal{C})$. Il est réducteur de décrire la capacité de survie par un simple nombre mais cette image simpliste permet d'aller loin dans la compréhension des mécanismes de sélection. Bien qu'il ne corresponde pas directement à un paramètre mesurable, le *fitness* capture l'essentiel de la dynamique de sélection. Au cours de la reproduction, des mutations peuvent apparaître et l'enfant d'un individu de phénotype $\mathcal{C}$ peut avoir un phénotype $\mathcal{C}'$ différent avec une certaine probabilité de mutation $W(\mathcal{C} \to \mathcal{C}')$.

La sélection est imposée, par exemple, en limitant les ressources du milieu ou en introduisant des prédateurs. Quelle qu'en soit la cause, l'un des effets de la sélection est de réduire la taille de la population par rapport à ce que serait son évolution libre. À chaque génération on élimine ainsi des individus aléatoirement dans la population avec un biais dépendant de leur *fitness*. En particulier, ce sont les individus avec les *fitnesses* les plus bas qui verront leurs lignées s'éteindre le plus rapidement et on observe une progression globale du *fitness* moyen d'une population.

Plusieurs grandes classes de progression du *fitness* au cours du temps ont été proposées. Dans le premier cas, le *fitness* moyen augmente régulièrement au cours du temps et finit par trouver son maximum, autour duquel il se stabilise. Cela correspond à une exploration *régulière* de l'espace des configurations phénotypiques $\mathcal{C}$ où les individus trouvent facilement des *fitnesses* meilleurs. Dans le deuxième cas, l'évolution est caractérisée par l'alternance de deux phases, l'une où le système tend vers un équilibre autour d'un maximum local de *fitness*, l'autre où se produit, suite à une séquence de mutations favorables chez un individu, un saut relativement rapide vers un autre maximum local de *fitness* (*quasispecies model*). Cette dernière dynamique n'est pas sans rappeler l'évolution de systèmes désordonnés en physique statistique où le système reste bloqué dans des minimaux locaux d'énergie et saute parfois d'un minimum local à un nouveau minimum. L'analogie peut être poussée plus loin en posant pour la fonction de *fitness* $f(\mathcal{C}) = e^{\beta F(\mathcal{C})}$, où $F(\mathcal{C})$ est l'équivalent d'une énergie pour la configuration $\mathcal{C}$ et β l'équivalent d'une température inverse de sélection. De la même manière qu'on caractérise un système physique par son *paysage énergétique*, on peut définir un « *paysage de* fitness » pour ce type de modèle. Plusieurs types de paysages ont été étudiés dans la littérature (voir [Pel97, JK06] pour de nombreuses références bibliographiques), dont les cas extrêmes sont les suivants :

1. le cas particulier sans sélection où $F(\mathcal{C})$ ne dépend pas de la configuration (*flat*

[2]Le nombre moyen *attendu* d'enfants est le nombre obtenu en moyennant le nombre d'enfants d'un individu sur un grand nombre de réalisations de l'événement de reproduction pour un individu de *fitness* donné : pour une réalisation donnée, le nombre réel d'enfants peut être différent.

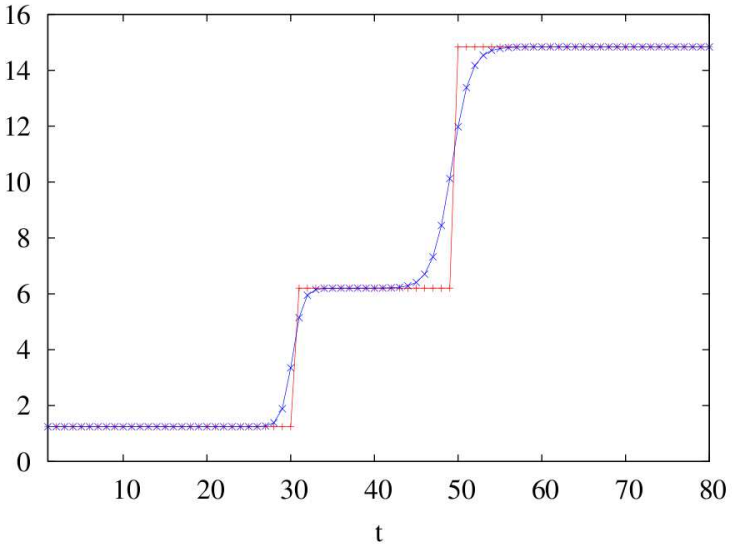


FIG. 1.4: Évolution du *fitness* d'une population dans un paysage de *fitness* rugueux (image reprise de [Jai07]) : le *fitness* de la population est représenté par $\times$ et celui du génotype le plus représenté à chaque pas de temps par $+$. Chaque saut de *fitness* global est précédé de l'apparition dans la population d'un individu meilleur que les autres qui va proliférer très rapidement.

fitness landscape [DP82], cf. par exemple le processus de Galton-Watson de la section précédente) ;

2. le cas d'un paysage très lisse où $F(\mathcal{C})$ varie lentement et n'a au plus qu'un extremum local qui est alors un extremum global (*smooth fitness landscape*) : dans ce cas-là, le *fitness* augmente de manière régulière jusqu'à ce que la population atteigne le phénotype optimal[3] ou bien diverge avec le temps si $F(\mathcal{C})$ n'est pas bornée ;

3. le cas d'un paysage rugueux où $F(\mathcal{C})$ est très irrégulier avec un grand nombre de maxima locaux : le système a une évolution intermittente alternant les phases où il se stabilise pendant une durée assez longue sur un maximum local et celles où il trouve un autre maximum local et se déplace rapidement vers celui-ci [Jai07] (voir figure 1.4).

L'analogie de ce dernier cas avec les milieux désordonnés fait que les outils développés pour étudier ceux-ci peuvent être adaptés à l'étude de ces modèles d'évolution. Cependant, même si ce dernier cas semble plus proche de l'évolution aux temps géologiques des espèces biologiques [Gou02], il ne nous intéressera pas directement dans cet exposé bien qu'il serait pertinent d'y prolonger certains résultats, en particulier ceux de la partie II sur les généalogies. Nous nous focaliserons ici sur les cas où la sélection introduit une compétition *permanente* entre les individus.

[3] cf. l'exemple particulier du *Fujiyama landscape* [Pel97].

1.3.2 Différents régimes de compétitions entre mutations

Dans les modèles d'évolution précédents, les mutations favorables sont celles qui ont la probabilité la plus grande de se répandre dans toute la population. Cependant, si elles sont fréquentes, il faut s'attendre à ce que seules certaines d'entre elles parviennent à éradiquer les autres et à se fixer dans la population [GL98].

Si le délai entre les apparitions de deux mutations successives est trop long, alors la population passe d'un stade d'équilibre à un autre en un temps correspondant à la prolifération de la mutation favorable dans toute la population (cas **a** de la figure 1.5). Cependant, si ces deux échelles de temps deviennent comparables, alors les mutations favorables vont entrer en compétition entre elles. La figure 1.5, extraite de [DF07], explore les différents scénarios possibles de compétition entre mutations et exhibe deux scénarios possibles selon les échelles de temps d'apparition de mutations et de prolifération dans la population :

- l'effet de mutations multiples (cas **d**) correspond à une situation où c'est le cumul de mutations favorables sur un même allèle qui lui permet d'éliminer les autres et de se fixer. Pour cela, il faut non seulement les mutations favorables soient fréquentes mais aussi que le temps de fixation soit du même ordre que le délai entre deux mutations favorables.
- l'effet d'*interférence clonale* (cas **c**), pour reprendre le vocabulaire de [GL98], au contraire, survient lorsque les deux temps (apparition d'une mutation et fixation dans la population) sont comparables et qu'ils sont suffisamment longs : dans ce régime, les mutations sont suffisamment fréquentes pour entrer en compétition mais encore trop rares pour se cumuler pendant la phase de compétition. La fixation d'un allèle et l'élimination des autres sont alors dues à des fluctuations statistiques durant le processus de reproduction.

Ces modèles contiennent différents paramètres (fréquence des mutations, avantage reproductif des mutations, taille de la population) ajustables, dont les variations produisent les différents régimes d'évolution ci-dessus. Néanmoins, pour toute une gamme de valeurs de ces paramètres, la comparaison [DF07] des différentes échelles de temps qui apparaissent dans le problème semble indiquer des liens avec des propagations de fronts : la répartition des *fitnesses* des individus entre le meilleur individu et le dernier, ainsi que la dynamique de ces *fitnesses*, ressemble qualitativement à un front qui se dirige vers les *fitnesses* les meilleurs. De plus, les ordres de grandeur des différents temps caractéristiques obtenus par une étude de ces régimes de sélection exhibent des similarités avec des quantités similaires qui apparaissent dans l'étude de fronts.

De tels régimes peuvent être décrits par la méthode proposée dans [PK07] obtenue en incluant un *fitness* multiplicatif dans le modèle de Wright-Fisher introduit en section 1.2.2 :

1. la reproduction est décrite par le modèle de Wright-Fisher où un poids w_i est affecté à chaque individu (son parent est choisi avec un poids $w_i / \sum_j w_j$) ;
2. lors de la reproduction, chaque enfant hérite du poids w_i de son parent à une mutation près : une mutation peut survenir avec une probabilité μ ; lorsqu'elle survient, son *fitness* est multiplié par $1 + s$ où s est un nombre aléatoire positif (bénéfice de la mutation).

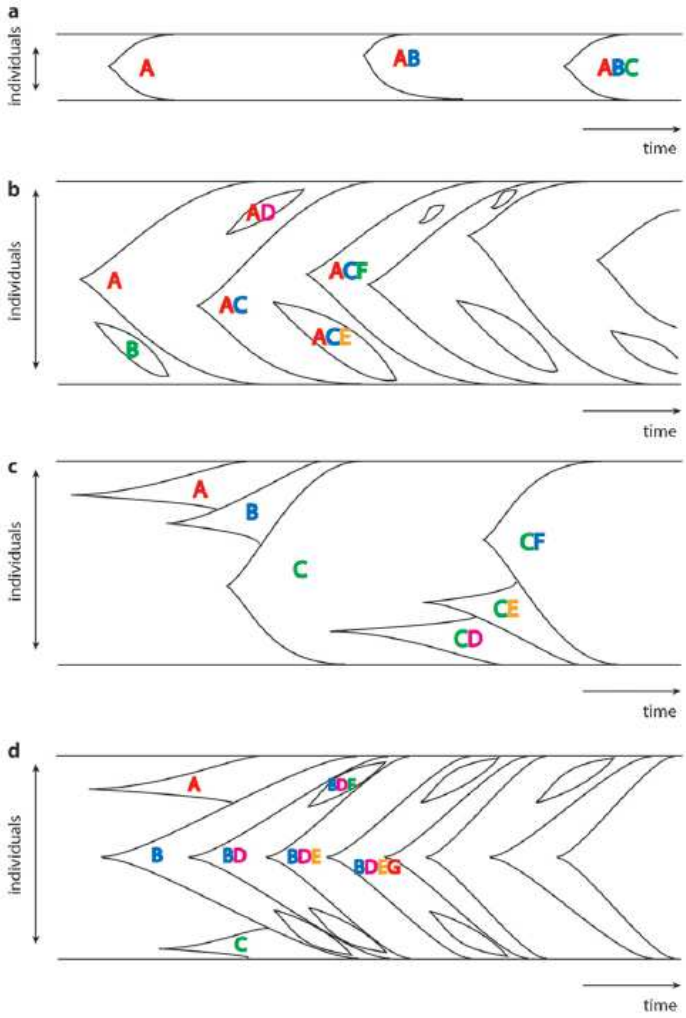


FIG. 1.5: Compétition entre différentes mutations favorables selon les échelles de temps des différents phénomènes (repris de [DF07]). Pour se fixer, une mutation favorable doit apparaître puis s'étendre à la population. Si les mutations favorables sont trop rares (cas **a**), les mutations se succèdent et se fixent les unes après les autres. Pour une population plus grande ou des mutations plus fréquentes (cas **b**), des mutations concurrentes (B par rapport à A, D par rapport à C, etc.) surgissent avant la fixation de la première et entrent en compétition : ce cas regroupe des caractéristiques communes aux deux cas suivants **c** et **d**. Le cas **c** correspond à celui de l'interférence clonale : avant que A n'ait pu se fixer, une nouvelle mutation B est survenue et a éliminé A puis a été éliminée à son tour par C qui s'est finalement fixée. Le cycle recommence ensuite. Le dernier cas, **d**, correspond à une domination par mutations multiples : A, B et C sont en compétition mais la domination de B n'est plus assurée par des fluctuations statistiques de la fréquence de chacune mais par l'apparition d'une nouvelle mutation D qui, cumulée, donne l'avantage décisif à B par rapport à A et C. Dans ce cas-là, sont éliminés ceux qui ne réussissent pas à accumuler suffisamment de mutations favorables.

Ce modèle prédit (voir [PK07]) une croissance linéaire pour $\langle \ln w_i \rangle$ aux temps longs pour une population de taille N. Comme annoncé ci-dessus, un certain nombre de caractéristiques de ce modèle semblent être semblables aux modèles que nous avons considérés (cf. chapitres suivants) et pose la question de l'universalité des résultats obtenus. De même l'interprétation en termes de records de [PK07] n'est pas sans rappeler les arguments phénoménologiques de [BDMM06b]. Enfin, la formulation précédente du modèle décrit en [PK07] se rapproche de la dynamique des polymères dirigés en dimension $1+\infty$ que nous étudierons plus précisément au chapitre 9.

1.3.3 Modélisation par des marches aléatoires avec branchement

Dans la suite, nous dévierons un peu de la description précédente en repérant les individus non pas par le nombre moyen de leurs enfants qui survivent, le *fitness* au sens strict, mais par un nombre réel x que nous appellerons l'*adéquation* ou encore une fois le *fitness* par abus de langage, qui mesurera la capacité d'un individu à se développer dans son environnement. Cette quantité sera liée à sa survie à long terme mais ne sera pas, à strictement parler, reliée à son nombre d'enfants à la génération suivante.

Chaque individu est ainsi représenté par un nombre réel x, d'autant plus grand que l'individu est bien adapté. L'hérédité correspond à la transmission de ce nombre du parent à l'enfant et les mutations sont représentées par un bruit lors de cette transmission. Mathématiquement parlant, lors de la reproduction d'un individu dont le *trait* (ou *fitness* par abus de langage) est x, chacun de ses n enfants naît avec une adéquation $x+\epsilon_i$ où les ϵ_i sont des variables aléatoires non corrélées. Une mutation est ainsi d'autant plus favorable que le bruit ϵ_i associé prend une grande valeur. Dans la description précédente, on remarque que ce *trait* se comporte comme une coordonnée spatiale où la diffusion (biaisée ou non) serait l'analogue des mutations. Tout un langage *spatial* (« position », « coordonnées », « diffusion ») sera utilisé par extension pour décrire le *trait* (ou *fitness*).

Au niveau de l'évolution et de la transmission du *fitness* x, les modèles utilisés par la suite sont les suivants :

temps et espace discrets : l'*adéquation* est une variable entière (sur réseau) et à chaque pas de temps, un individu se reproduit selon le processus de Galton-Watson défini précédemment et les enfants sont positionnés aléatoirement sur les sites voisins de celui du parent (diffusion à partir de la coordonnée du parent). Ce modèle a l'avantage de donner lieu à des simulations numériques aisées.

temps et espace continus : le *fitness* d'un individu exécute une marche aléatoire au cours du temps et, pendant une durée dt, chaque individu a une probabilité $\beta_k dt$ de se diviser en k individus avec le même *fitness* que le parent. Ici, les mutations ne se produisent plus nécessairement au moment de la division mais les résultats restent les mêmes. Ce modèle donne lieu à un traitement analytique plus simple, dans la mesure où il utilise des équations différentielles plutôt que des équations aux différences finies mais il pose des problèmes numériques de discrétisation.

Ces deux derniers modèles correspondent à des marches aléatoires sur réseau ou sur $\mathbb{R}$ avec branchements en k nouvelles marches aléatoires.

Un dernier modèle [BDMM06a], peu réaliste du point de vue de la biologie, peut être introduit en liaison avec d'autres domaines de la physique statistique, comme par

exemple avec les polymères dirigés [BD04] : le temps est discret et, à chaque pas de temps, un individu de *fitness* x produit dans chaque intervalle $[y, y + dy]$ un enfant avec une probabilité $\psi(y - x)dy$ (*Poisson point process*). On supposera que la fonction ψ décroît suffisamment vite en $+\infty$ pour éviter qu'il n'y ait trop d'enfants vers $+\infty$ (afin d'éviter tout problème de divergence avec les modes de sélection considérés par la suite) ; en $-\infty$, le nombre d'enfants est régulé par la sélection décrite ci-dessous. Un cas particulier de ce modèle fait l'objet du chapitre 7.

Ces trois types de modèles sont ainsi reliés au problème physique de marches aléatoires avec branchements dont nous étudierons certaines propriétés dans les chapitres ultérieurs. Contrairement aux modèles de la section 1.3.1, les individus ne sont pas caractérisés par un état interne sous-jacent (comme leur génome ou leur phénotype) dont dépendrait le *fitness* mais directement par celui-ci.

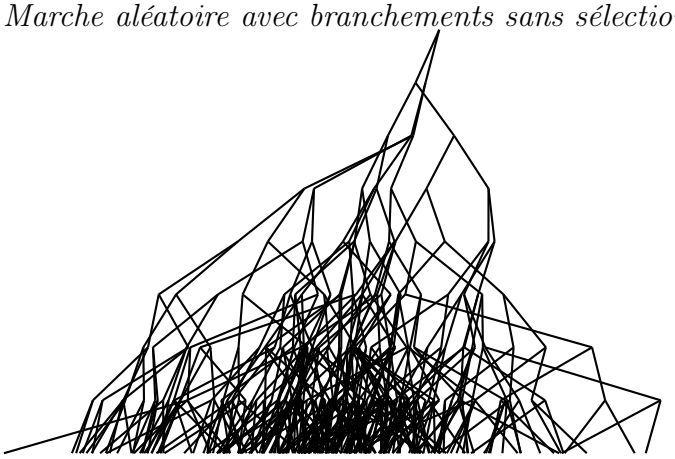
Au niveau de la sélection, le but est d'éliminer les individus avec les *fitnesses* les plus bas, c'est-à-dire les individus les moins adaptés à leur environnement. Nous distinguerons deux principaux types de sélection (cf. figure 1.6) :

la sélection interne : si les ressources du milieu sont insuffisantes, la population ne peut excéder une certaine taille N et les individus entrent en compétition. À chaque pas de temps, on sélectionnera donc les N meilleurs individus et à chaque création d'un individu supplémentaire dans la population, on éliminera celui qui a le *fitness* x le plus bas. Nous qualifions cette sélection d'*interne* car elle induit une interaction entre les individus via leurs *fitnesses* alors que les propriétés du milieu restent constantes au cours du temps. L'influence de la sélection sur la vitesse d'évolution et sur les généalogies des individus a déjà été étudiée dans [BDMM06a, BDMM07]. Cette sélection peut être « adoucie » en prenant non pas les N meilleurs mais N aléatoirement parmi les $2N$ meilleurs ou en utilisant une distribution de Fermi-Dirac pour sélectionner les survivants mais on notera que la sélection reste indépendante du milieu et ne dépend que des *fitnesses* relatifs des individus.

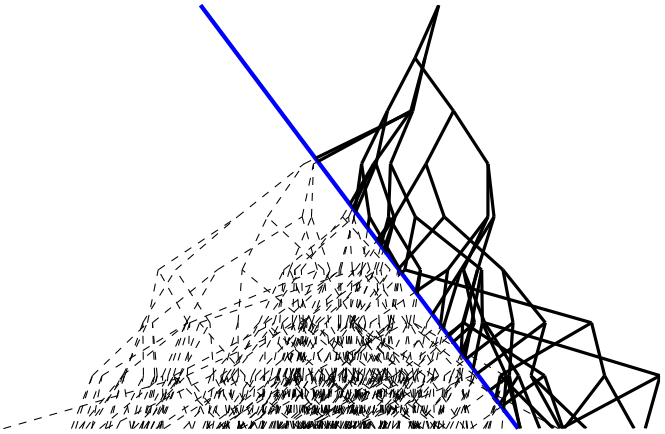
la sélection externe : on suppose ici que la sélection est due à un mécanisme externe qui élimine les individus selon leur *fitness* et que les individus ont des évolutions indépendantes. Le cas le plus abrupt correspond à un seuil avançant avec une vitesse constante v tel que tous les individus dont le *fitness* x passe sous le seuil sont immédiatement éliminés. Là encore, la sélection peut être adoucie en supposant que les individus sont tués avec un taux $\beta_0(x - vt)$ dépendant de leur *fitness* x à l'instant t et tel que $\beta_0(z) \to 0$ lorsque $z \to +\infty$ et $\beta_0(z) \to +\infty$ lorsque $z \to -\infty$. En présence de sélection externe, les individus sont indépendants et leur survie n'est liée qu'à leur interaction avec l'environnement. On notera le parallèle avec des modèles purement géographiques (chaque individu est repéré par sa position et diffuse) où le taux d'extinction dépend de la position : cela explique pourquoi nous considérerons également des modèles où la position est multi-dimensionnelle.

Une différence majeure entre ces deux méthodes réside dans la question de la survie de la population : dans la sélection interne, comme la taille de la population est maintenue constante, la question de la survie ne se pose pas, alors qu'en sélection externe, si le seuil avance trop vite, la population a une probabilité non nulle de s'éteindre. La question

– *Marche aléatoire avec branchements sans sélection :*



– *Sélection externe par un seuil de* fitness *augmentant à la vitesse v (tout individu croisant le mur meurt instantanément) :*



– *Sélection interne ($N = 6$ meilleurs à chaque génération) :*

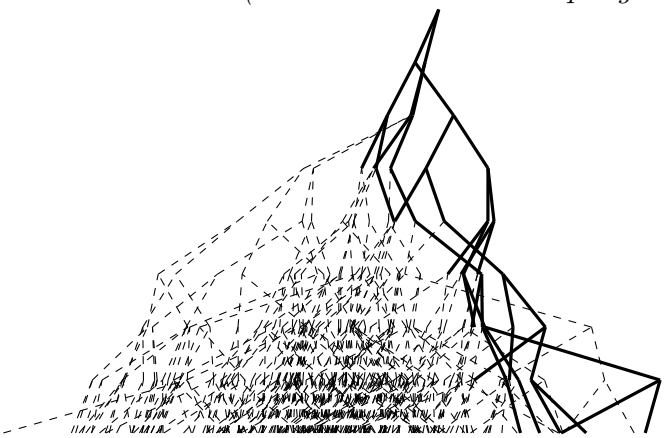


FIG. 1.6: Effet des sélections interne (taille constante) et externe (augmentation du *fitness* seuil) pour une même marche aléatoire avec branchements.

de l'extinction en sélection externe fera l'objet du chapitre 3. L'allure du profil de la population et sa dynamique conditionnée sur la survie de la population feront l'objet des chapitres 5 et 6.

L'un des résultats de nos travaux est la mise en évidence de caractéristiques communes entre ces deux types de sélection. Ces résultats sont présentés au chapitre 6, où nous montrons que la relation entre la taille et la vitesse est la même dans les deux types de sélection.

Mathématiquement, l'évolution de la répartition des *fitnesses* d'une population se réduit à l'étude d'équations de propagation de front, comme nous le verrons dans les chapitres ultérieurs : la population est constamment régénérée à l'avant et éliminée à l'arrière par sélection. L'idée de la description d'une population par un front est déjà présente dans la littérature [TLK96, RWC03, DF07, RBW07] et exhibe le même type de propriétés : en particulier, les effets de taille finie sont importants et, souvent, même à des tailles de l'ordre de 10^{50}, les corrections à la limite de taille infinie ne sont pas négligeables. Le lien explicite entre nos modèles est les propagations de front sera présenté à partir du chapitre 3.

1.4 Quelques modèles actuels de population en dehors de la biologie

La modélisation de la sélection naturelle et de la dynamique des populations a eu des applications dans des domaines autres que la biologie. Nous présentons ici deux exemples de modèles de populations abstraites utilisés en informatique et en physique statistique hors d'équilibre. La caractéristique commune de ces deux applications est d'utiliser la sélection et la reproduction pour renforcer l'exploration numérique de domaines de l'espace des configurations difficiles d'accès par une dynamique « non biaisée ».

1.4.1 Les algorithmes génétiques en informatique

De nombreux problèmes d'informatique consistent à trouver des configurations satisfaisant à une liste de contraintes. La méthode la plus brutale consiste à énumérer toutes les configurations possibles jusqu'à trouver une solution. Cependant, lorsque l'espace des combinaisons est trop vaste, cette méthode s'avère inefficace. Il faut alors développer des méthodes, plus ou moins justifiées mathématiquement, pour accélérer la recherche de solution, comme par exemple des *algorithmes génétiques*.

Considérons l'exemple d'un problème d'optimisation d'une fonction F suffisamment compliquée pour que les méthodes standard échouent. Pour reprendre l'exemple de [FH04], nous cherchons par exemple un réseau de plusieurs espèces chimiques en interaction qui produisent des oscillations de période donnée (origine du rythme circadien) : une configuration est spécifiée par la donnée du nombre d'agents, des réactions dans lesquelles ils interviennent, des taux de réaction et des conditions initiales. La fonction à optimiser est la distance (par exemple l'écart maximal dans le temps) à une sinusoïde de référence. La recherche d'une solution par algorithme génétique consiste à considérer N systèmes avec des configurations différentes puis à itérer la procédure suivante :

1. on regarde le résultat produit par chacun des N systèmes, *i.e.* la valeur de la fonction F à optimiser pour chacun d'eux ;
2. on élimine la fraction d'entre eux qui correspond aux plus mauvais résultats (valeurs de F les plus éloignées de l'extremum visé) ;
3. on duplique les meilleurs d'entre eux ;
4. on ajoute des mutations aléatoires aux systèmes, *i.e.* on ajoute un bruit sur les taux de réaction ou bien une espèce ou une réaction supplémentaire, et/ou on effectue des croisements, *i.e.* on permute des valeurs de paramètres ou des réactions chimiques entre les individus ;
5. on recommence à l'étape 1 avec la nouvelle population.

Cet algorithme est directement inspiré des modèles d'évolution : le génotype est remplacé par des valeurs de paramètres d'un système, le *fitness* est remplacé par la fonction F à optimiser, la sélection correspond à l'étape 2 ci-dessus, la reproduction à l'étape 3 et les mutations correspondent à l'étape 4.

L'efficacité des algorithmes repose sur le fait que ceux-ci vont explorer l'espace des configurations en ne s'éloignant jamais des solutions potentielles grâce à l'étape de sélection. Leur difficulté d'utilisation réside dans le choix des taux de mutations : trop de mutations violentes dispersent la population dans l'espace des phases, trop peu de mutations empêchent une convergence rapide de l'algorithme. De plus, cette méthode ne garantit pas de trouver la solution optimale et peut être bloquée dans des minima locaux de F et ne convient que si l'on se contente de solutions approchées. Par sa procédure de tâtonnement dans l'espace des phases, elle se rapproche, dans le cas de problèmes biologiques, de la véritable évolution biologique.

1.4.2 Évaluation numérique de fonctions de grandes déviations hors d'équilibre

Un certain nombre de techniques de simulations numériques en physique statistique se heurtent au problème de l'échantillonnage de l'espace des phases avec les bonnes probabilités. Il existe un certain nombre de méthodes (Monte-Carlo) pour produire des configurations arbitraires avec les bons poids à l'équilibre. Néanmoins, hors équilibre, la détermination de fonctions de grandes déviations nécessite, par définition, de produire des événements rares, qui sont eux aussi, par définition, difficiles à produire par une dynamique naturelle (*i.e.* suivant la dynamique markovienne qui définit le modèle). Nous présentons ici un exemple simple d'algorithme utilisé dans [GKP06] pour mesurer des fonctions de grandes déviations de courant dans un système hors équilibre (*totally asymmetric exclusion process*). Soit $\mathcal{C}$ une configuration du système et notons $M(\mathcal{C} \to \mathcal{C}')$ le taux de transition d'une configuration $\mathcal{C}$ vers une configuration $\mathcal{C}'$. Soit $J(\mathcal{C}, \mathcal{C}')$ la quantité échangée lors du passage de $\mathcal{C}$ à $\mathcal{C}'$ qui nous intéresse. Pour une séquence de configurations $(\mathcal{C}_0, \mathcal{C}_1, \ldots, \mathcal{C}_T)$, nous nous intéressons à la distribution de $Q_T = J(\mathcal{C}_0, \mathcal{C}_1) + \ldots + J(\mathcal{C}_{T-1}, \mathcal{C}_T)$ au bout d'un temps long. La fonction génératrice de

Q_t s'écrit donc :

$$\begin{aligned}\langle e^{-\lambda Q_T}\rangle &= \sum_{\mathcal{C}_1,\ldots,\mathcal{C}_T} M(\mathcal{C}_0\to\mathcal{C}_1)M(\mathcal{C}_1\to\mathcal{C}_2)\ldots M(\mathcal{C}_{T-1}\to\mathcal{C}_T)e^{-\lambda(J(\mathcal{C}_0,\mathcal{C}_1)+\ldots+J(\mathcal{C}_{T-1},\mathcal{C}_T))}\\ &= \sum_{\mathcal{C}_1,\ldots,\mathcal{C}_t} \widehat{M}_\lambda(\mathcal{C}_0\to\mathcal{C}_1)\widehat{M}_\lambda(\mathcal{C}_1\to\mathcal{C}_2)\ldots\widehat{M}_\lambda(\mathcal{C}_{T-1}\to\mathcal{C}_T)\end{aligned}$$

où la matrice $\widehat{M}_\lambda$ est définie par :

$$\widehat{M}_\lambda(\mathcal{C}\to\mathcal{C}') = M(\mathcal{C}\to\mathcal{C}')e^{-\lambda J(\mathcal{C},\mathcal{C}')}. \tag{1.5}$$

Aux temps longs, $\langle -\lambda Q_T\rangle$ est dominée par la contribution de la valeur propre $\mu(\lambda)$ avec la partie réelle la plus grande de la matrice $\widehat{M}_\lambda$:

$$\langle e^{-\lambda Q_T}\rangle \propto e^{T\mu(\lambda)}$$

Pour déterminer $\mu(\lambda)$, il faut *a priori* diagonaliser la matrice $\widehat{M}_\lambda$ qui est de grande taille. Numériquement, itérer cette matrice est aussi difficile à cause de sa taille. De plus, simuler directement l'évolution induite par $M(\mathcal{C}\to\mathcal{C}')$ nous prive des événements trop rares. L'idée présentée dans [GKP06] est alors d'interpréter $\widehat{M}_\lambda$, qui n'est plus markovienne car la normalisation des colonnes est perdue, comme une matrice qui décrit l'évolution d'une population. Plus précisément, réécrivons $\widehat{M}_\lambda$ sous la forme suivante :

$$\begin{aligned}\widehat{M}_\lambda(\mathcal{C}\to\mathcal{C}') &= D_\lambda(\mathcal{C}\to\mathcal{C}')K_\lambda(\mathcal{C})\\ K_\lambda(\mathcal{C}) &= \sum_{\mathcal{C}'}\widehat{M}_\lambda(\mathcal{C}\to\mathcal{C}')\neq 0\end{aligned}$$

de telle sorte que la matrice D_λ soit à nouveau markovienne.

Les coordonnées $P_\mathcal{C}(t)$, au lieu d'être considérées comme des probabilités (normalisées à 1), peuvent à présent être considérées comme étant les nombres moyens d'*individus* habitant sur les configurations $\mathcal{C}$. Faire évoluer les populations $P_\mathcal{C}(t)$ selon la matrice $\widehat{M}_\lambda$ revient alors à faire évoluer une population avec des taux de branchements par individu contenus dans $K_\lambda(\mathcal{C})$ et des diffusions[4] d'individus contenues dans la matrice $D_\lambda(\mathcal{C}\to\mathcal{C}')$.

La perte de normalisation de $\widehat{M}_\lambda$ fait que la population va soit exploser, soit s'éteindre : il faut alors biaiser l'évolution de la population de telle sorte qu'elle reste constante. Plus précisément, si nous partons de N individus distribués aléatoirement dans l'espace des configurations (N est arbitraire et est choisi aussi grand que possible suivant la capacité de calcul disponible), il faut, à chaque pas de temps, sélectionner un individu et suivre la procédure suivante :

1. l'individu dans la configuration $\mathcal{C}$ produit G individus identiques sur la même configuration $\mathcal{C}$ avec G donné par :

$$G = \begin{cases}[K_\lambda(\mathcal{C})]+1 & \text{avec probabilité } K_\lambda(\mathcal{C})-[K_\lambda(\mathcal{C})]\\ [K_\lambda(\mathcal{C})] & \text{sinon}\end{cases}$$

[4]Le fait que D_λ soit markovienne signifie que la population est en moyenne constante.

où $[x]$ désigne la partie entière de x. Si $[K_\lambda(\mathcal{C})] = 0$, alors cela signifie que l'individu ne laisse aucune progéniture supplémentaire ;

2. tous les individus $(L+G)$ diffusent sur une configuration adjacente $\mathcal{C}'$ avec un taux de diffusion $D_\lambda(\mathcal{C} \to \mathcal{C}')$;
3. on normalise la population par un facteur $A_t = L/(L+G)$ en éliminant au hasard G individus.
4. on stocke dans la mémoire la valeur de A_t.

Alors, aux temps longs, les A_t stockés nous donnent (cf. [GKP06]) la valeur de $\mu(\lambda)$ par la formule :

$$\mu(\lambda) \simeq -\frac{1}{T}\sum_{t=1}^{T} \ln A_t$$

Dans cette procédure, le rôle du *fitness* est joué par le coefficient $K_\lambda(\mathcal{C})$ induit par la perte de normalisation de $\widehat{M}_\lambda$ par rapport à M et les mutations par les sauts vers des configurations adjacentes $\mathcal{C}'$. Nous voyons ainsi comment l'introduction d'une population permet d'étendre les méthodes Monte-Carlo à l'équilibre à la mesure de fonctions de grandes déviations hors d'équilibre. À la lumière des chapitres ultérieurs, il serait intéressant d'explorer les liens entre les différentes quantités étudiées dans les modèles de population et les modèles hors d'équilibre.

Chapitre 2

Propagation de fronts

Ce chapitre est une introduction à l'étude de la propagation de fronts en physique. La présentation est restreinte aux fronts de type *pulled*. La première section introduit le type de problèmes et le type d'équations où apparaissent des fronts. La deuxième section est consacrée à la détermination de la vitesse d'un front et présente les résultats existants. La troisième section montre ce que deviennent ces résultats lorsque le front est affecté par un bruit. La dernière section, plus technique, présente la méthode à suivre pour déterminer les temps de relations d'un front lorsque celui-ci est perturbé localement : cette méthode sera utilisée dans les trois chapitres suivants. Elle reprend un calcul publié dans [SD08].

2.1 Généralités

De nombreux systèmes peuvent être décrits comme des fronts se propageant dans un domaine. Par exemple, pour reprendre un modèle simple de réaction-diffusion, une réaction de combustion d'un combustible A en présence d'un comburant B dans un milieu donne lieu à la propagation d'un front séparant une zone riche en combustible A et une zone où A a été majoritairement consommé. De tels fronts apparaissent aussi dans des modèles de contagion ($A + B \to 2B$ où A désigne un individu sain et B un individu infecté), dans l'étude de marches aléatoires avec branchements [Bra83, McK75] ou de la propagation d'une mutation génétique favorable. Dans ce dernier cas, une mutation favorable se produit sur un gène d'un individu puis, à travers la reproduction et la diffusion, se répand dans la population environnante. De manière assez générale, cela conduit à un front que l'on peut décomposer en trois zones : une zone (en expansion) autour de la mutation initiale où les individus sont majoritairement porteurs de la mutation, une zone (en récession) où la mutation est très minoritaire, voire encore absente, et une zone intermédiaire où coexistent le gène initial et le gène muté. Aux temps longs, le front ainsi produit se déplace sans se déformer et sa vitesse n'est pas, *a priori*, celle des individus

se déplaçant dans le milieu.

Un modèle simple de front décrivant la prolifération d'un gène favorable dans une population a été introduit en 1937 simultanément par Fisher [Fis37] d'une part et Kolmogorov, Petrovsky et Piscounov [KPP37] d'autre part. Dans un milieu unidimensionnel et dans la limite des populations de grandes tailles, le modèle est décrit par l'équation différentielle

$$\partial_t Q = \partial_x^2 Q + \beta(Q - Q^2), \tag{2.1}$$

où $Q(x,t)$ désigne la proportion d'individus portant le gène favorable au point x et à l'instant t. Le laplacien $\partial_x^2 Q$ correspond à la diffusion du gène dans le milieu et le terme non-linéaire $Q-Q^2$ traduit l'augmentation locale de la proportion du gène favorable (on a $Q - Q^2 > 0$ pour $0 < Q < 1$) ainsi que la saturation à $Q = 1$ (tous les individus sont porteurs du gène muté favorable). Le terme non-linéaire est parfois pris égal à $\beta(Q-Q^3)$ mais les propriétés du front sont essentiellement inchangées ; de manière générale, nous considérerons des non-linéarités $f(Q)$ telles que

$$\partial_t Q = \partial_x^2 Q + f(Q) \tag{2.2}$$

et

$$\begin{cases} f(0) = 0, f'(0) = \beta > 0 & (Q = 0 \text{ point fixe instable}) \\ f(1) = 0, f'(1) < 0 & (Q = 1 \text{ point fixe stable}) \\ f(Q) > 0 & \text{pour } 0 < Q < 1 \end{cases} \tag{2.3}$$

pour décrire l'invasion d'une phase instable $Q = 0$ par une phase stable $Q = 1$. L'une des propriétés principales de cette équation est l'existence de solutions de type *front* [Bra83] se propageant à une vitesse v, *i.e.* l'existence de solutions du type :

$$Q(x,t) = Q_v(x + vt). \tag{2.4}$$

La détermination de la vitesse v fait l'objet de la section 2.2. Les propriétés générales de tels fronts déterministes sont aujourd'hui bien connues (voir [van03] pour une revue assez complète des propagations de fronts). La propagation est choisie comme allant vers les x négatifs (voir figure 2.1) afin de faciliter l'adaptation aux calculs du chapitre 3.

L'équation (2.2) est bien souvent le fruit d'une idéalisation du problème de départ. Dans le cadre de la prolifération d'un gène favorable, (2.2) n'est valable que pour une population infinie. Pour prendre en compte les effets de taille finie et l'aspect stochastique des processus, il est nécessaire d'introduire une équation bruitée (cf. [DMS03, MMQ08]) ou avec *cut-off* (cf. [BD97]). Dans le premier cas, (2.2) devient :

$$\partial_t Q = \partial_x^2 Q + f(Q) + a(Q(x,t))\xi(x,t) \tag{2.5}$$

où $a(Q(x,t))$ dépend de la valeur de Q au point x et où $\xi(x,t)$ est un bruit aléatoire. Dans le cas avec *cut-off*, la définition de la non-linéarité $f(Q)$ change lorsque Q est plus petit qu'une certaine quantité ϵ. On peut montrer dans ce dernier cas (cf. [BD97, MMQ08]) que, lorsque ϵ tend vers 0, la convergence vers la solution sans *cut-off* est très lente (en $1/\log(\epsilon)^2$ pour la vitesse du front). Les propriétés de l'équation avec bruit ont suscité un assez vif intérêt récemment (cf. [BDMM06b, DMS03, MMQ08] entre autres).

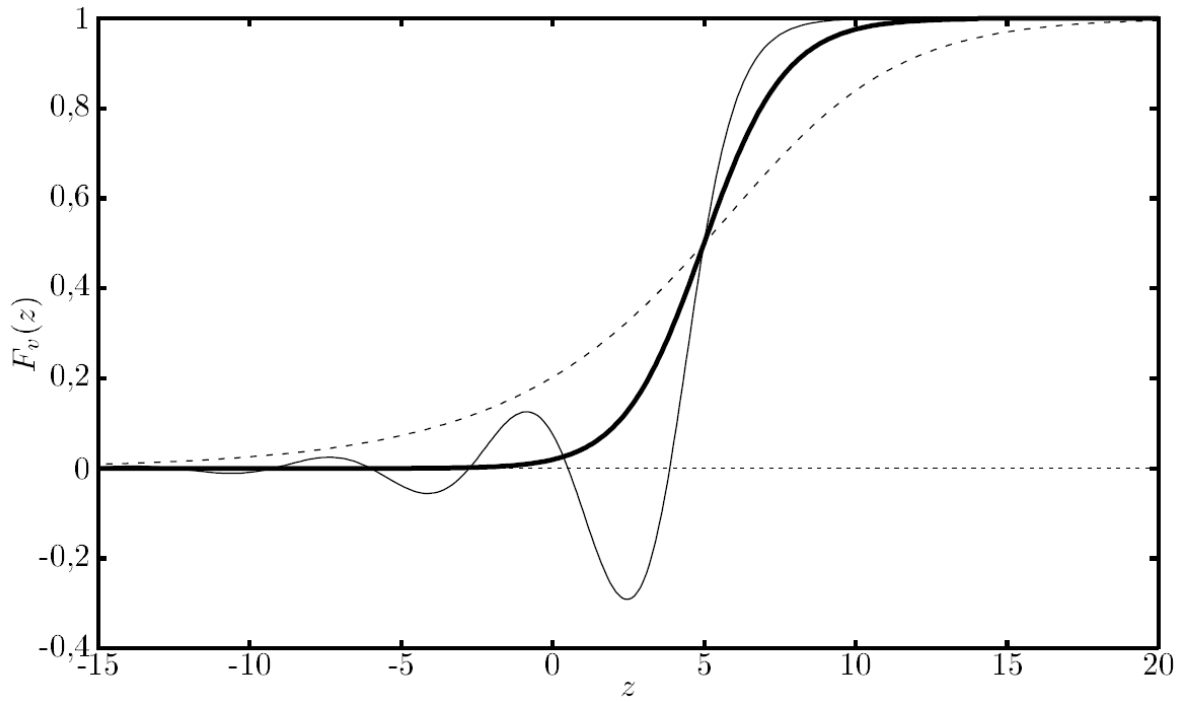


FIG. 2.1: Allure d'un front se propageant vers les x négatifs avec des vitesses $v = v_c = 2$ (en gras), $v = 0.5$ (ligne simple) et $v = 5$ (en pointillés).

L'équation de Fisher-Kolmogorov-Petrovsky-Piscounov (F-KPP) et ses variantes sont en fait présentes dans des domaines nombreux et variés, des problèmes de combustion et réaction-diffusion [KR85] jusqu'à la physique nucléaire [IMM05, MPS05]. Nous montrerons dans les chapitres ultérieurs que l'équation F-KPP est aussi reliée aux questions de survie de marches aléatoires avec branchements.

2.2 Vitesse d'un front

Outre les solutions triviales $Q = 0$ (instable) et $Q = 1$ (stable), (2.2) admet des solution de type (2.4) se propageant avec une vitesse $-v$ et des conditions aux limites

$$\begin{cases} \lim_{z\to-\infty} Q_v(z) = 0, \\ \lim_{z\to+\infty} Q_v(z) = 1. \end{cases} \tag{2.6}$$

La fonction Q_v satisfait alors l'équation

$$\partial_z^2 Q_v - v\partial_z Q_v + f(Q_v) = 0. \tag{2.7}$$

Peu de solutions exactes de cette équation sont connues mais on sait cependant relier la vitesse v d'une solution Q_v aux propriétés de celle-ci lorsque $z \to -\infty$.

2.2.1 Analyse linéaire

Loin dans la phase instable ($z \to -\infty$), Q_v est proche de zéro et on peut linéariser (2.7) :

$$\partial_z^2 Q_v - v\partial_z Q_v + \beta Q_v \simeq 0 \tag{2.8}$$

avec $\beta = f'(0)$. Puisque l'on impose à Q_v de s'annuler en $-\infty$, Q_v prend la forme $Q_v(z) \propto e^{\gamma z}$ dans la phase instable où le coefficient γ est relié à la vitesse v par la relation

$$v = \gamma + \frac{\beta}{\gamma}. \tag{2.9}$$

Pour garantir la convergence vers 0 quand $z \to -\infty$ et la positivité de Q_v (imposée par la nature physique de Q_v dans les problèmes qui nous concernent ici), il faut se restreindre aux γ réels strictement positifs. On trouve alors (cf. figure 2.2.1) que les fronts admissibles doivent avoir une vitesse supérieure à la vitesse critique donnée par

$$\begin{cases} v_c = 2\sqrt{\beta} \\ \gamma_c = \sqrt{\beta} \end{cases} \tag{2.10}$$

et ces fronts ont ainsi une queue exponentielle dans la phase instable. Si l'on autorise les solutions Q_v à devenir négatives, alors γ peut prendre des valeurs complexes et les solutions se comportent comme des oscillations amorties pour les grandes valeurs négatives de z ; dans ce cas, la vitesse est alors plus petite que la vitesse critique v_c. Exactement à la vitesse critique v_c, Q_{v_c} prend la forme asymptotique

$$Q_{v_c}(z) \simeq -(A_c z + B_c)e^{\gamma_c z} \tag{2.11}$$

loin dans la phase instable $z \to -\infty$. L'invariance par translation implique que l'un des deux coefficients A_c et B_c soit libre mais, une fois l'un fixé, l'autre est fixé par les non-linéarités du système et le comportement $Q_v(z) \to 1$ en $+\infty$. Néanmoins, la détermination analytique complète des amplitudes n'est pas possible sauf dans des cas isolés.

2.2.2 Sélection de la vitesse de propagation

L'étude précédente a montré quelle était la forme de $Q_v(z)$ dans la phase instable pour une vitesse v donnée mais il reste à déterminer laquelle de ces vitesses est sélectionnée pour une condition initiale donnée. Aronson et Weinberger ont montré [AW75, AW78] le résultat très général suivant pour des équations de type (2.2) avec les hypothèses (2.3) :

- il existe une solution de type *front* (comprise entre 0 et 1) en translation uniforme à la vitesse v pour tout $v \geq 2\sqrt{f'(0)}$,
- une telle solution est stable si et seulement si $v \geq v_c$ pour une certaine vitesse critique v_c
- on a les bornes suivantes sur la vitesse v_c :

$$2\sqrt{f'(0)} \leq v_c \leq 2 \sup_{0 \leq Q \leq 1} \sqrt{f(Q)/Q} \tag{2.12}$$

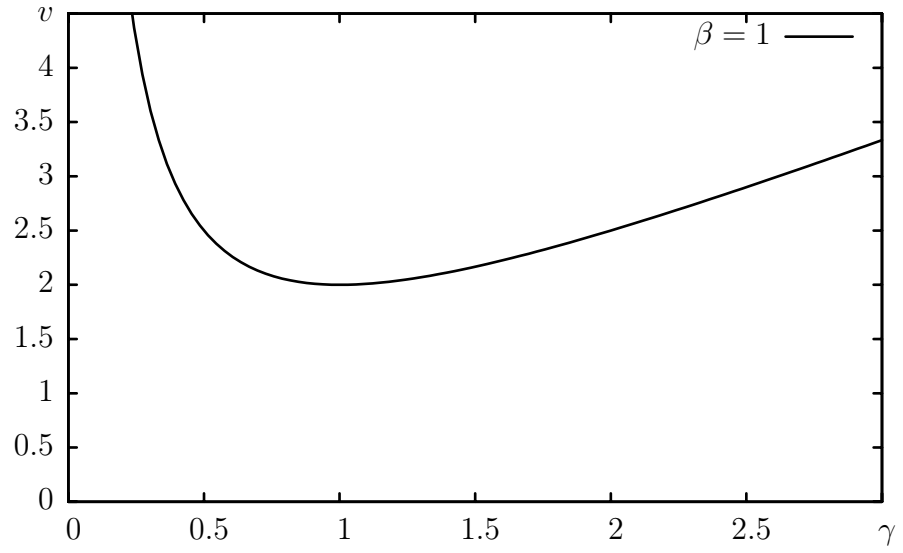


FIG. 2.2: Vitesse de propagation d'un front $Q_v(z)$ en fonction de son taux de décroissance γ loin dans la phase instable $Q \simeq 0$. On voit que les solutions sans oscillations amorties obtenues pour γ réel positif ont une vitesse minimale $v_c = 2$ pour $\beta = 1$. Pour $\gamma > 1$, les solutions sont instables et adoptent la vitesse v_c aux temps longs.

– pour une large catégorie de conditions initiales, la vitesse du front tend vers la vitesse marginalement stable v_c (cela est vrai en particulier[1] pour toutes les conditions initiales $Q(x,0)$ décroissant plus vite que $e^{\gamma_c x}$ pour $x \to -\infty$ pour un certain γ_c dépendant de l'équation de propagation).

La notion de stabilité, introduite ci-dessus, sera discutée en section 2.4 et doit être prise dans le sens suivant : si nous ajoutons à une solution $Q_v(z)$ une perturbation localisée $\epsilon(z)$ qui est suffisamment petite pour que $Q_v(z)+\epsilon(z)$ reste dans l'intervalle $[0,1]$ et qui décroît à l'infini plus rapidement que $Q_v(z)$, alors aux temps longs, la solution de 2.3 converge (à une translation près) à nouveau vers $Q_v(z)$. Comme nous le verrons plus bas, cette stabilité pour $v \geq v_c$ est liée à la monotonie du front $Q_v(z)$.

Pour la plupart des cas que nous étudierons, nous aurons $\sup_{0\leq Q\leq 1}\sqrt{f(Q)/Q} = \sqrt{f'(0)}$ et v_c et γ_c seront donc donnés par (2.10). Dans ces cas-là, toute condition initiale suffisamment abrupte donne un front se propageant à v_c. Cela signifie donc qu'une étude de l'équation F-KPP *linéarisée* suffit dans ces situations à donner les propriétés critiques des fronts. Par ailleurs, cela montre aussi que le seul effet des non-linéarités contenues dans $f(Q)$ consiste à saturer Q à la valeur 1 sans modifier la propagation du front. De tels fronts sont appelés *pulled* d'après la terminologie inventée par Stokes [Sto76], par opposition aux fronts *pushed* pour lequel les non-linéarités influent sur la vitesse de propagation et la stabilité des solutions.

[1]Pour une condition initiale décroissant comme $e^{\gamma x}$ en $-\infty$ avec $0 < \gamma < \gamma_c$, alors la vitesse du front est donnée par $v(\gamma)$.

Les bornes (2.12) sur la vitesse du front trouvées par Aronson et Weinberger sont des conséquences de principes variationnels plus généraux [HR75, BD96b, BD96a] démontrés depuis, dont nous donnons ici une illustration. Ainsi, le principe variationnel de Hadeler et Rothe [HR75] donne, pour une non-linéarité du type (2.3), la vitesse v_c du *front de vitesse minimale* comme la borne inférieure suivante :

$$v_c = \inf_{g\in\mathcal{U}} \sup_{0<u<1} \left(g'(u) + \frac{f(u)}{g(u)} \right) \tag{2.13}$$

où $\mathcal{U} = \{g \in C^1([0,1]) | g(0) = 0, g'(0) > 0, g(Q) > 0, \forall Q \in]0,1[\}$ est l'ensemble des fonctions g de classe C^1 sur $[0,1]$, strictement positive sur $]0,1[$, s'annulant en zéro et de dérivée strictement positive en 0. Un moyen simple de visualiser cette formule est présenté en figure 2.3. Si nous nous plaçons dans l'espace des phases $(q,p) = (Q_v, \partial_x Q_v)$, l'équation stationnaire (2.7) se réécrit :

$$\begin{cases} q' &= F_1(q,p) = p, \\ p' &= F_2(q,p) = vp - f(q). \end{cases} \tag{2.14}$$

Cette dynamique admet les deux points fixes $(q,p) = (0,0)$ et $(q,p) = (1,0)$. Une étude de stabilité linéaire montre que $(0,0)$ est un point fixe instable, comme attendu, et que $(1,0)$ est un point-col. Pour que le front soit monotone (donc positif), il faut que la trajectoire ne forme pas de spirale autour de $(0,0)$: les valeurs propres ne sont réelles que si $v^2 > 4f'(0)$ qui est la première borne de (2.12). D'autre part, la seule trajectoire aboutissant en 1 pour $x \to \infty$ correspond à la partie de la variété instable du point 1 dans le domaine $q < 1$, $p > 0$: il suffit alors de montrer que cette trajectoire provient nécessairement du point $(0,0)$. Pour cela, nous remarquons d'après (2.14) qu'aucune trajectoire ne peut entrer dans le domaine $0 < q < 1$, $p > 0$ en passant par les frontières $(0 < q < 1, p = 0)$ et $(q = 1, p > 0)$. Il suffit alors de trouver une fonction $g(q)$ satisfiant $g(0) = 0$, $g'(0) > 0$, $g(Q) > 0$ telle que le vecteur tangent $(1, g'(q))$ et le champ $(F_1(q,g(q)), F_2(q,g(q)))$ soient orientés dans le sens direct pour être sûr (cf. figure 2.3) que la variété instable arrivant en 1 provienne nécessairement du point $(0,0)$. La vitesse critique v_c est alors obtenue en considérant la borne inférieure sur les fonction $g \in \mathcal{U}$ pour laquelle la condition $F_2(q,g(q)) - g'(q)F_1(q,g(q)) > 0$ est satisfaite.

Le fait d'avoir un *infimum* dans (2.13) implique que des choix appropriés de la fonction g permettent d'obtenir des majorations rapides de la vitesse du front marginalement stable. En particulier, si la non-linéarité f est telle que $f(Q) \leq f'(0)Q$ pour toute valeur de Q dans $[0,1]$ alors le choix $g(u) = \sqrt{f'(0)}u$ donne $v_c < 2\sqrt{f'(0)} = 2\sqrt{\beta}$. Par les bornes (2.12), nous obtenons $v_c = 2\sqrt{\beta}$ et le front est donc de type *pulled*. La condition $f(Q) \leq f'(0)Q$ n'est pas restrictive pour les cas que nous étudierons (f sera convexe) et les fronts que nous considérerons seront de ce type.

2.3 Influence du bruit et d'un *cut-off* sur la vitesse

Un exemple simple d'équation de front en présence de bruit est donné par [MS95, MMQ08] :

$$\partial_t Q = \partial_x^2 Q + f(Q) + \sqrt{Q(1-Q)/N}\xi(x,t) \tag{2.15}$$

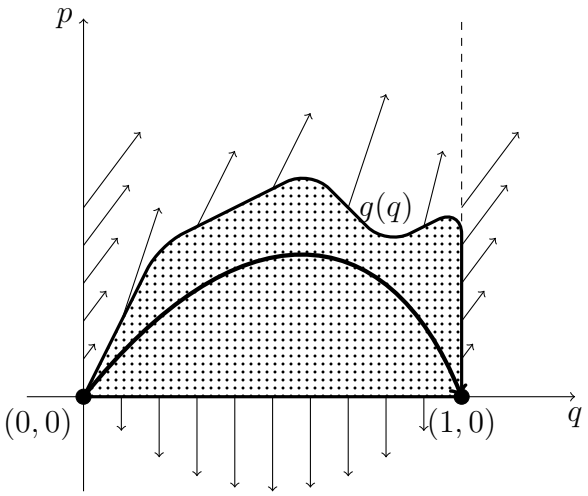


FIG. 2.3: Existence d'un front de vitesse donnée. Si, sur la courbe d'une fonction $g(q)$ telle que $g(0) = 0$, $g'(0) > 0$ et $g(q) > 0$ pour $0 < q < 1$, le champ de vecteur (F_1, F_2) est toujours dirigé vers l'extérieur du domaine hachuré, alors il est impossible d'entrer dans le domaine hachuré sauf en passant par $(0, 0)$. D'autre part, le point $(1, 0)$ est un point col dont la variété instable entre dans le domaine hachuré : celle-ci vient donc nécessairement de 0 et correspond au front Q_v qui satisfait $Q_v(x) \to 0$ pour $x \to -\infty$ et $Q_v(x) \to 1$ pour $x \to +\infty$.

où ξ est un bruit blanc gaussien normalisé. Cet exemple modélise, entre autres, la prolifération d'une mutation favorable dans une population de taille N. La nature multiplicative du bruit en $\sqrt{Q(1-Q)}$ fait que les points fixes $Q = 0$ et $Q = 1$ restent des points fixes en présence de bruit. L'amplitude du bruit est prise en $1/\sqrt{N}$ pour reproduire les effets de taille finie d'une population si Q désigne la proportion d'individus mutés (et donc $1 - Q$ donne la proportion d'individus non affectés par la mutation). Le bruit fait que la position du front n'est plus une quantité déterministe mais diffuse autour d'une valeur moyenne.

L'amplitude du bruit en $\sqrt{Q(1-Q)/N}$ fait que celui-ci est non négligeable dès lors que Q est d'ordre $1/N$: cela correspond aux endroits où le nombre d'invidus mutés est de l'ordre de l'unité et où les effets de taille finie sont donc les plus visibles. Une approximation supplémentaire consiste à résumer les effets de taille finie en un paramètre $\epsilon = 1/N$ de telle sorte que $Q(x, t)$ satisfasse :

$$\partial_t Q = \partial_x^2 Q + f(Q)a(Q) \tag{2.16}$$

avec $a(Q) = 1$ si $Q > \epsilon$ et $a(Q) \ll 1$ si $Q \ll \epsilon$. L'équation ainsi obtenue est déterministe et permet une étude plus simple. On peut montrer [BD97] que la vitesse du front en présence du *cut-off* $\epsilon = 1/N$ est donnée par :

$$v_{\textit{cut-off}} \simeq v_c - \frac{\pi^2 \gamma_c^2 v''(\gamma_c)}{2 \log^2 N} \tag{2.17}$$

où $v''(\gamma_c)$ est donnée par (2.9). La convergence lente en $1/\log^2 N$ montre que l'effet d'un

cut-off loin dans la phase instable $Q \simeq 0$ est loin d'être négligeable, même pour des valeurs de N relativement grandes. Pour établir (2.17), il faut partir de l'allure du front dans le domaine linéaire $Q \ll 1$ et effectuer les raccordements *ad hoc* entre les trois régimes $Q \ll \epsilon \ll 1$, $\epsilon < Q \ll 1$ et $Q \simeq 1$ comme exposé dans [BD97] (voir aussi section 2.4.2 de ce même chapitre).

L'équation (2.17) montre que, pour un front de type *pulled*, une légère fluctuation dans le domaine $Q \ll 1$ suffit à perturber le front de manière significative. Dans le cas d'un bruit stochastique et non d'un *cut-off* déterministe, en plus des résultats rigoureux [MMQ08], il existe des approches phénoménologiques [BDMM06b] donnant des prédictions pour la vitesse du front en accord avec les résultats numériques. Elles sont basées sur le même type de raisonnement que précédemment : tout en restant dans le régime linéaire (ce qui permet l'analyse), on considère l'effet à long terme sur la position du front de fluctuations suffisamment rares pour que le système ait le temps de relaxer entre deux fluctuations successives. Une telle analyse phénoménologique prévoit la correction suivante à la vitesse moyenne du front :

$$v \simeq v_{\textit{cut-off}} + \pi^2 \gamma_c^3 v''(\gamma_c) \frac{3 \log\log N}{\gamma_c \log^3 N} + \dots \tag{2.18}$$

De plus, la nature stochastique de (2.15) fait que la position du front fluctue et la même approche prévoit un coefficient de diffusion du front donné par :

$$D \simeq \pi^2 \gamma_c^3 v''(\gamma_c) \frac{\pi^2/3}{\gamma_c^2 \log^3 N}. \tag{2.19}$$

Il existe plus généralement des prédictions pour les moments supérieurs de la position [BDMM06b] et il serait intéressant de connaître la portée de ces résultats phénoménologiques en leur apportant une démonstration mathématique complète, au-delà de [MMQ08].

2.4 Temps de relaxation et stabilité des solutions

2.4.1 Stabilité et linéarisation autour du front

Pour avoir les solutions $Q_v(z)$ effectivement sélectionnées par un système physique, il convient d'étudier la stabilité des solutions. Des travaux rigoureux [AW75, AW78, BJBD+85, van98] montrent que, si l'on perturbe le front localement[2], seules les solutions monotones (donc nécessairement avec $v \geq v_c$) sont stables dans le référentiel du front. Cette dernière distinction tient au fait que les solutions satisfaisant (2.6) ne sont déterminées qu'à une constante de translation près.

Une manière d'aborder ce résultat qui nous sera utile par la suite consiste à linéariser l'équation (2.7) et à étudier les temps de relaxation, i.e. les valeurs propres de l'opérateur linéarisé :

$$\mathcal{L}[\phi] = \partial_z^2 \phi - v \partial_z \phi + f'(Q_v(z))\phi. \tag{2.20}$$

[2]Comme expliqué plus haut, les perturbations locales considérées ici sont suffisamment petites pour que Q_v reste dans l'intervalle $[0, 1]$ *et* ont une convergence vers 0 en $\pm\infty$ au moins aussi rapide que la convergence de Q_v vers ses limites 0 et 1.

Cet opérateur $\mathcal{L}$ décrit l'évolution de faibles perturbations locales $\eta(z,t)$ autour de la solution de front $Q_v(z)$ dans le référentiel de celui-ci. Plus précisément, il suffit que l'une des valeurs propres λ de $\mathcal{L}$ ait une valeur propre de partie réelle positive pour que la perturbation croisse et déstabilise le front Q_v.

Après un changement de variable $\phi(z) = e^{vz/2}\psi(z)$ et $E = -\lambda$, l'équation aux valeurs propres devient :

$$E\psi(z) = \Big[-\partial_z^2 + \left(\frac{v^2}{4} - f'(Q_v(z))\right)\Big]\psi(z), \tag{2.21}$$

qui n'est autre que l'équation de Schrödinger dans un potentiel $V(z) = v^2/4 - f'(Q_v(z))$ relié à l'allure du front stationnaire. Le problème de stabilité du front se ramène alors à un problème de spectre d'hamiltoniens de Schrödinger : une solution $Q_v(z)$ est stable si et seulement si l'hamiltonien ainsi construit a toutes ses valeurs propres à parties réelles positives.

Cette approche simpliste cache cependant le problème épineux des conditions aux limites. En effet, les vecteurs propres admissibles doivent vérifier les bonnes conditions aux limites, déterminées par la physique du problème. Pour un front infini, on supposera que les perturbations sont localisées (mutation, étincelle, ...) et on considèrera donc des vecteurs propres $\phi(z)$ qui tendent vers 0 en $-\infty$ et en $+\infty$ avec une convergence au moins aussi rapide que la convergence de Q_v vers sa limite. On constate alors que le vecteur propre $\phi(z) = \partial_z Q_v(z)$ satisfait les bonnes conditions aux limites et $\mathcal{L}[\phi] = 0$ et correspond donc au vecteur propre associé à la valeur propre nulle $\lambda = 0$. Ce vecteur propre correspond à une translation du front sans déformation. Dans le cas d'un front monotone, $\partial_z Q_v(z)$ ne s'annule pas et on s'attend à ce qu'il corresponde au fondamental, montrant ainsi que les autres vecteurs propres correspondent à des valeurs propres négatives et que le front doit être stable.

Cette approche, bien que non rigoureuse, est compatible avec le résultat exact disant qu'un front est stable si et seulement si il est monotone. Une étude mathématique détaillée [AW75, AW78] traitant rigoureusement ces comportements à l'infini permet alors de montrer que seuls les fronts monotones sont stables vis-à-vis de ces perturbations.

Dans le cas d'un front avec *cut-off*, les conditions aux limites changent et le raccordement autour du point z_ϵ tel que $Q_v(z_\epsilon) = \epsilon$ modifie les vecteurs propres admissibles [Pv02]. Le spectre est donc à discuter au cas par cas (cf. ci-dessous).

2.4.2 Vecteurs propres sur la ligne infinie dans la limite $v \to v_c^-$

Les chapitres ultérieurs nécessitent une étude précise des vecteurs propres et des temps de relaxation associés d'un front lorsque v est proche de v_c. Nous avons développé à cet effet une méthode perturbative qui permet d'obtenir, *a priori* à tous les ordres, la forme d'un vecteur propre ϕ_λ associé à la valeur propre λ près de la vitesse v_c et pour $\lambda \to 0$. Bien que cette démarche, reproduite dans la deuxième annexe de [SD08], ait été développée initialement afin d'obtenir les résultats exposés dans les chapitres 3 et 6, elle se généralise néanmoins à tout front de type *pulled* et permet d'étudier les temps de relaxation de fronts avec divers types de conditions aux limites loin dans la phase instable. La section 2.4.3 est une application particulière de cette méthode aux cas de fronts avec *cut-off*.

Le but de cette approche est de montrer qu'aux ordres dominants, les propriétés des vecteurs propres au voisinage de v_c loin dans la phase instable sont indépendantes du détail des non-linéarités $f(Q)$ et ne nécessitent que la connaissance des constantes A_c et B_c introduites en (2.11). Plus précisément, loin dans la phase instable $z \to -\infty$, les équations satisfaites par Q_v et ϕ_λ sont données sous forme approchée par :

$$\partial_x^2 Q_v(x) - v\partial_x Q_v(x) + \beta Q_v(x) \simeq 0, \tag{2.22}$$

$$\partial_x^2 \phi_\lambda(x) - v\partial_x \phi_\lambda(x) + \beta \phi_\lambda(x) \simeq \lambda \phi_\lambda(x), \tag{2.23}$$

où le coefficient est défini par $\beta = f'(0)$. Les équations différentielles précédentes sont des équations différentielles linéaires d'ordre 2 homogènes dont les solutions les plus générales sont données, pour $v < v_c$, par :

$$Q_v(x) = U_v \sin\left(\frac{\pi(x+\Phi_v)}{L}\right) e^{vx/2} + O\left(e^{vx}\right), \tag{2.24}$$

$$\phi_\lambda(x) = V_{v,\lambda} \sin\left(\frac{\pi(x+\Psi_{v,\lambda})}{L_\lambda}\right) e^{vx/2} + O\left(e^{vx}\right). \tag{2.25}$$

Les longueurs L et L_λ des arches de sinus sont données par :

$$L = \frac{\pi}{\sqrt{\beta - v^2/4}} \propto (v_c - v)^{-1/2}, \tag{2.26}$$

$$L_\lambda = \frac{\pi}{\sqrt{\beta - v^2/4 - \lambda}}. \tag{2.27}$$

Lorsque $v \to v_c$, la longueur L de l'arche de sinus diverge et on observe alors que le premier zéro de $Q_v(x)$ s'éloigne de la partie non-linéaire près de la vitesse critique.

Les constantes d'intégration U_v et Φ_v sont déterminées de la manière suivante :

- l'une provient de l'invariance par translation des solutions de l'équation de front et sera donc déterminée en fixant arbitrairement une coordonnée x_0 telle que $Q_v(x_0) = 1/2$ par exemple[3] ;
- l'autre est fixée par les non-linéarités et la limite $Q_v(x) \to 1$ lorsque $x \to +\infty$.

La constante $V_{v,\lambda}$ du vecteur propre ϕ est elle aussi arbitraire alors que la phase $\Psi_{v,\lambda}$ est fixée par la limite $\phi(x) \to 0$ imposée en $+\infty$.

A priori, les constantes U_v, Φ_v, $V_{v,\lambda}$ et $\Psi_{v,\lambda}$ de (2.24) et (2.25) dépendent des non-linéarités. Le but de ce qui suit est de montrer qu'au voisinage de v_c et pour λ petit, elles n'en dépendent que faiblement. Pour cela, nous développons une théorie de perturbation en $v_c - v$ pour $Q_v(x)$ et un double développement en $v_c - v$ et λ pour les vecteurs propres ϕ_λ. Dans un souci de commodité pour la suite, le développement autour de la vitesse critique sera fait en $1/L^2$ où L est défini en (3.24) plutôt qu'en $v_c - v$. Posons en effet :

$$Q_v(x) = Q_{v_c}(x) + \sum_{m\geq 1} \frac{q_m(x)}{L^{2m}}$$

$$\phi_\lambda(x) = \partial_x Q_{v_c} + \sum_{m,n\geq 1} s_{mn}(x) \frac{\lambda^n}{L^{2m}}$$

[3] On remarquera alors qu'un certain nombre de propriétés, comme les temps de relaxation, ne doivent pas dépendre de ce x_0.

puisque $\mathcal{L}[\partial_x Q_v] = 0$ est le vecteur propre correspondant à $\lambda = 0$. Toutes les corrections q_m et s_{mn} peuvent être obtenues récursivement par la même équation différentielle homogène de degré 2. En effet, aux ordres les plus bas, les équations (2.2,2.20) donnent :

$$\partial_x^2 q_1 - v_c \partial_x q_1 + f'(Q_{v_c}) q_1 = -\frac{2\pi^2}{v_c} \partial_x Q_{v_c}, \tag{2.28a}$$

$$\partial_x^2 s_{01} - v_c \partial_x s_{01} + f'(Q_{v_c}) s_{01} = \partial_x Q_{v_c}, \tag{2.28b}$$

$$\partial_x^2 s_{10} - v_c \partial_x s_{10} + f'(Q_{v_c}) s_{10} = -\frac{2\pi^2}{v_c} \partial_x^2 Q_{v_c} - f''(Q_{v_c}) q_1 \partial_x Q_{v_c}. \tag{2.28c}$$

Connaissant une solution particulière de l'équation homogène sans second membre (la dérivée $\partial_x Q_{v_c}(x)$), la solution s'annulant en x_0 et $+\infty$ de

$$\partial_x^2 u - v_c \partial_x u + f'(Q_{v_c}) u = K(x)$$

avec un second membre $K(x) = O(Q_{v_c}(x)) = O(e^{-rx})$ exponentiellement décroissant en $+\infty$ est donnée par :

$$u(x) = \partial_x Q_{v_c}(x) \int_{x_0}^{x} \frac{dy}{[\partial_x Q_{v_c}(y)]^2 e^{-vy}} \int_{y}^{+\infty} \partial_x Q_{v_c}(z) e^{-vz} K(z) dz. \tag{2.29}$$

Cela montre ainsi que toutes les corrections q_m et s_{mn} peuvent être obtenues perturbativement dès que $Q_{v_c}(x)$ est connu. Néanmoins, aux ordres les plus bas, les intégrales ci-dessus n'ont pas besoin d'être évaluées pour obtenir les constantes d'intégration des équations (2.24, 2.25). En effet, en revenant à l'expression (2.11), nous nous rendons compte que Q_{v_c} prend la forme d'une exponentielle multipliée par un polynôme et que les seconds membres de (2.28) sont de la même forme pour $x \to -\infty$. Ainsi, les solutions q_1, s_{01}, s_{10} de (2.28) sont aussi de cette forme. La démarche adoptée dans [SD08] consiste alors à identifier ces polynômes avec le développement de Taylor des sinus des équations (2.24, 2.25) dans le domaine de recouvrement $1 \ll -x \ll L$.

Les calculs détaillés présentés dans la deuxième annexe de [SD08] permettent d'aboutir aux développements suivants :

$$\Phi_v = \frac{B_c}{A_c} + O\left(\frac{1}{L^2}\right) \tag{2.30}$$

$$U_v = \frac{A_c}{\pi} L + O\left(\frac{1}{L}\right) \tag{2.31}$$

$$\Psi_{v,\lambda} = \frac{B_c}{A_c} + \frac{2}{v_c} + O\left(\frac{1}{L^2}\right) + O\left(\frac{1}{L_\lambda^2}\right) \tag{2.32}$$

$$V_{v,\lambda} = \frac{A_c v_c}{2\pi} L_\lambda + O\left(\frac{1}{L^2}\right) + O\left(\frac{1}{L_\lambda^2}\right) \tag{2.33}$$

Les termes correctifs peuvent être obtenus à partir de la seule connaissance de Q_{v_c} par l'intermédiaire des constantes A_c et B_c introduites dans (2.11). Nous allons voir que les développements précédents suffisent à donner les temps de relaxations les plus lents dans le cas avec *cut-off* (cf. ci-dessous) et dans le problème de survie de marches aléatoires avec branchements du chapitre 3.

2.4.3 Temps de relaxation en présence d'un *cut-off*

Le but de cette section est de montrer sur l'exemple simple du front avec *cut-off* comment la connaissance des vecteurs propres sur la ligne infinie obtenue dans la section précédente peut être utilisée pour sélectionner les valeurs propres qui correspondent à un front avec un cut-off à travers de simples conditions de raccordement.

Sur la ligne infinie, les conditions aux bords sont simplement $\phi_\lambda(x) \to 0$ si $x \to \pm\infty$. Dans le cas d'un front avec *cut-off*, d'après (2.16), les vecteurs propres satisfont :

$$\partial_x^2\phi_\lambda - v\phi\partial_x\phi_\lambda + f'(Q_v)a(Q_v)\phi_\lambda + f(Q_v)a'(Q_v)\phi_\lambda = \lambda\phi_\lambda. \tag{2.34}$$

Pour une fonction de *cut-off* quelconque $a(Q)$ continue et infiniment dérivable, le problème est bien défini. Nous nous placerons ici dans la limite où le domaine de variation de $a(Q)$ est de taille infinement petite de telle sorte que $a(Q) \to \Theta(Q-\epsilon)$ avec $\epsilon \ll 1$ et $\Theta(x) = 1$ si $x > 0$ et $\Theta(x) = 0$ sinon. Bien que le problème dynamique de départ (2.16) soit mal défini lorsque $a(Q)$ est discontinue, nous supposerons ici que l'emploi de la forme $a(Q) = \Theta(Q-\epsilon)$ donne néanmoins une bonne approximation des temps de relaxations lorsque $a(Q)$ a des variations suffisamment abruptes (cf. [Pv02] pour une démarche similaire).

Soit X_ϵ le point tel que $Q_v(X_\epsilon) = \epsilon$. Alors $a'(Q_v) = \delta(Q_v - \epsilon) = \delta(x - X_\epsilon)/\partial_x Q_v(X_\epsilon)$. Sur $[X_\epsilon, +\infty]$, les vecteurs propres satisfont la même équation différentielle que sur la ligne infinie et donc pour ϵ petit, leurs formes asymptotiques autour de X_ϵ sont bien données par (2.25) à des corrections exponentiellement petites près. Pour $x < X_\epsilon$ alors l'équation différentielle ci-dessus devient linéaire et nous obtenons :

$$\phi_\lambda(x) \propto \exp\left[\frac{1}{2}\left(v + \sqrt{v^2+4\lambda}\right)x\right] \tag{2.35}$$

Enfin, la présence d'un pic $\delta(x - X_\epsilon)$ dans l'opérateur linéaire fait que la condition de raccordement en X_ϵ devient :

$$\lim_{\delta\to 0^+}\left[\frac{\partial_x\phi_\lambda(X_\epsilon+\delta)}{\phi_\lambda(X_\epsilon+\delta)} - \frac{\partial_x\phi_\lambda(X_\epsilon-\delta)}{\phi_\lambda(X_\epsilon-\delta)}\right] = -\frac{f'(\epsilon)}{\partial_x Q_v(X_\epsilon)}. \tag{2.36}$$

En utilisant les expressions (2.24, 2.25), on obtient l'équation suivante :

$$\frac{\pi}{L_\lambda}\cot\left[\frac{\pi}{L_\lambda}\left(X_\epsilon + \Psi_{v,\lambda}\right)\right] - \frac{1}{2}\sqrt{v^2+4\lambda} = -\frac{\beta}{v} \tag{2.37}$$

D'autre part, Q_v et $\partial_x Q_v$ doivent être continues en X_ϵ et la forme du *cut-off* fixe ainsi X_ϵ et la vitesse du front par les équations $Q(X_\epsilon) = \epsilon$ et $Q'(X_\epsilon) = \epsilon v$. Dans la limite $\epsilon \to 0$, on doit avoir :

$$L \simeq (2/v_c)|\ln\epsilon| \tag{2.38}$$

et ainsi retrouver la formule (2.17) pour la vitesse v. De plus on obtient l'équation suivante pour X_ϵ :

$$\frac{\pi}{L}\cot\left[\frac{\pi}{L}\left(X_\epsilon + \Phi_v\right)\right] = \frac{v}{2} \tag{2.39}$$

Les équations (2.37,2.39) permettent ainsi de développer les valeurs propres en puissance de $1/L$, et donc en puissance de $\propto 1/|\ln\epsilon|$. En particulier, il devient possible de retrouver et de prolonger, par un calcul perturbatif plus complet, les résultats de [Pv02] qui donnent les valeurs propres au premier ordre.

En effet, les conditions (2.37, 2.39) impliquent que, pour $\epsilon \to 0$, nous ayons asymptotiquement $\lambda_n \simeq -(n^2-1)\pi^2/L^2$. L'équation (2.39) implique alors le développement suivant pour X_ϵ en puissance de $1/L$:

$$X_\epsilon = -L + \left(\frac{2}{v_c} - \frac{B_c}{A_c}\right) + O\left(\frac{1}{L^2}\right) \tag{2.40}$$

(le terme en $-L$ provient du fait que X_ϵ s'éloigne de la région $Q_v \simeq 1/2$ lorsque $\epsilon \to 0$). Ensuite, en injectant un développement du type

$$\lambda_n = -\frac{(n^2-1)\pi^2}{L^2} - \frac{n^2\pi^2 a_1}{L^3} - \frac{n^2\pi^2 a_2}{L^4} + O(1/L^5) \tag{2.41}$$

dans l'équation (2.37), nous obtenons les développements suivants[4] pour les deux membres de (2.37) :

$$\begin{aligned}
\frac{\pi}{L_\lambda}\cot\left[\frac{\pi}{L_\lambda}\left(X_\epsilon + \Psi_{v,\lambda}\right)\right] &= \frac{1}{-\frac{a_1}{2}+\frac{4}{v_c}} + \frac{1}{L}\left[\frac{a_1}{-\frac{a_1}{2}+\frac{4}{v_c}} + \frac{\frac{a_2}{2} - \frac{a_1^2}{8} - \frac{2a_1}{v_c}}{\left(-\frac{a_1}{2}+\frac{4}{v_c}\right)^2}\right] + O\left(\frac{1}{L^2}\right) \\
-\frac{\beta}{v} + \frac{1}{2}\sqrt{v^2+4\lambda} &= \frac{v_c}{4} + O\left(\frac{1}{L^2}\right)
\end{aligned}$$

Cela montre ainsi que a_1 et a_2 doivent s'annuler et nous obtenons les temps de relaxation suivants :

$$\tau_n^{-1} = -\lambda_n \simeq \frac{[(n+1)^2-1]\pi^2}{L^2} + O\left(\frac{1}{L^5}\right), \qquad n = 1, 2, \ldots \tag{2.42}$$

De plus, le fait que pour $n = 0$, nous ayons $\lambda = 0$ à tous les ordres et $\partial_x Q(x)$ comme vecteur propre associé est compatible avec la liberté par translation sans déformation du front. La correction en $1/L^5$ dans les temps de relaxation fait intervenir la connaissance de B_c et A_c ainsi que le détail des non-linéarités du front : la seule partie universelle (*i.e.* celle qui ne dépend pas du détail de la fonction $f(Q)$) semble donc être celle en $1/L^2$.

[4] On utilise ici la formule asymptotique $\cot(x) \simeq 1/x - x/3 + O(x^3)$ pour la cotangente au voisinage de zéro.

Chapitre 3

La marche aléatoire avec branchements en présence d'un mur absorbant : probabilité de survie

Ce chapitre présente les résultats que nous avons obtenus sur la probabilité de survie d'une marche aléatoire avec branchements dans un domaine unidimensionnel avec conditions aux bords absorbantes. Les six premières sections présentent les résultats publiés dans [DS07] pour un unique mur absorbant se déplaçant à une vitesse constante v. Les trois premières sections sont une introduction au formalisme utilisé ainsi qu'aux méthodes numériques ; les quatrième et cinquième sections étudient les propriétés de la probabilité de survie au voisinage du point critique ; la sixième étudie la divergence des temps de relaxation au voisinage du point critique et présente les résultats complémentaires que nous avons publiés dans [SD08]. La dernière section présente les résultats, non publiés, pour la survie dans une boîte de taille donnée et permet de comparer les différences avec la géométrie semi-infinie précédente.

3.1 Origine et buts du modèle

Dans le cas de la sélection interne (taille maintenue constante égale à N), l'évolution de la population peut être décrite par un front de type (2.15) où le bruit est relié aux effets de taille finie (cf. [BDMM06b, BDMM07, DMS03]). Les difficultés qui surgissent dans l'étude du front avec bruit nous ont poussés à considérer un modèle légèrement différent, dans l'espoir qu'il partage des caractéristiques communes avec la population de taille constante. Nous considérons ici une population d'individus dans un modèle simple de sélection externe, déjà introduit par Harris et Harris [HH07] : les individus évoluent sur la ligne réelle en présence d'un mur absorbant qui se déplace à une vitesse

v vers les x positifs qui les éliminent dès le premier contact.

Les deux modèles de sélection par taille constante (interne) ou par mur (externe) ne peuvent cependant pas être équivalents dans la limite de grande taille pour la raison simple suivante : en présence d'un mur, la taille peut fluctuer, voire s'annuler, et la population peut s'éteindre. Elle peut aussi se mettre à diverger. Puisque rien ne stabilise la taille, la dynamique aux temps longs ne peut être que différente de celle avec taille constante.

Néanmoins, il est possible de restreindre, sous certaines conditions que nous avons explorées, cette taille de manière artificielle : il existe des événements, rares bien sûr, où la taille est finie (quelques unités) même au bout d'un temps T très long. L'observation d'une taille finie au bout d'un temps T très long nécessite non seulement que la population ne se soit pas éteinte entre 0 et T mais aussi que la taille de la population n'ait pas divergé sur ce même intervalle. Si nous nous intéressons aux histoires de la population sur l'intervalle $[0, T]$ qui mènent à une taille finie, nous observons, *sous certaines conditions*, qu'elles correspondent à une population dont la taille fluctue autour d'une même valeur N' sur l'intervalle de temps $[0, T]$. Cette taille N' est reliée à la vitesse du mur v. Il est alors légitime de se poser la question de l'équivalence, dans la limite des grandes tailles, entre ce régime conditionnée de la population en présence d'un mur de sélection avec le cas de la sélection interne par taille constante.

Le plan des chapitres suivants s'organise de la manière suivante : ce chapitre s'attache à décrire la population en présence d'un mur absorbant et en particulier sa dynamique aux temps longs (survie ou extinction) en absence de conditionnement. Les deux chapitres suivants (4 et 5) s'intéressent de plus près au régime conditionné de la population en présence d'un mur absorbant. Le chapitre 6, finalement, étudie le voisinage du point critique et la limite de grande taille et explore les similitudes attendues entre les deux modèles de sélection.

Indépendamment de ces considérations, un modèle très similaire au nôtre a été introduit par Antal *et al.* [ABTR07] pour décrire le processus de sénescence dans une assemblée de cellules en prolifération. Le paramètre x de l'axe unidimensionnel de *fitness* est remplacé par la longueur des télomères des chromosomes d'une cellule. Les télomères sont des morceaux d'ADN présents aux extrêmités des chromosomes dont la longueur est modifiée lors de la division cellulaire (mitose). En effet, le mécanisme enzymatique de réplication de l'ADN ne permet pas de répliquer la toute fin d'un brin d'ADN et ainsi, à chaque réplication, la longueur des télomères décroît un peu. Lorsque les télomères ont disparu, les erreurs de réplication affectent directement l'ADN *utile* de la cellule et perturbent suffisamment le fonctionnement de la cellule (sénescence) pour conduire à la mort de celle-ci. Néanmoins, il existe d'autres mécanismes auxiliaires qui font fluctuer la longueur des télomères : en particulier, des recombinaisons d'ADN aléatoires peuvent se produire et ainsi faire décroître un télomère tout en allongeant l'autre. Ces recombinaisons peuvent ainsi, dans un premier temps, être décrites par un bruit aléatoire alors que le raccourcissement systématique lors d'une réplication peut être modélisé par une dérive constante vers les longueurs décroissantes. Le seuil de longueur nulle correspond ainsi à un changement d'état irréversible (sénescence) de la cellule. La question posée par les auteurs de [ABTR07] est la caractérisation du destin à long terme (mort ou croissance)

d'une colonie en prolifération.

3.2 Description

3.2.1 La marche aléatoire avec branchements

La taille de la population au temps t est notée N_t. Les individus sont repérés par leur adéquation à leur environnement (*fitness*) ou position, *i.e.* un nombre réel x_i. Les positions des individus diffusent au cours du temps et leur diffusion est décrite de la manière suivante : pendant un intervalle de temps infinitésimal dt, on ajoute $\eta\sqrt{dt}$ à la position de l'individu où η est une variable gaussienne de moyenne nulle et de variance[1] 2. Parallèlement, un individu a une probabilité $\beta_k dt$ de se diviser en k descendants : la reproduction est donc semblable à celle du modèle de Galton-Watson en temps continu. La sélection est représentée par un mur absorbant avançant avec une vitesse constante v : tous les individus atteints par le mur sont instantanément éliminés. On remarquera qu'à un changement de référentiel près, cette dynamique est identique à celle obtenue avec un mur immobile et des particules diffusant avec une dérive $-v$ les poussant sur le mur et, dorénavant, toutes les particules seront repérées par leur position par rapport au mur.

On supposera que les particules ne peuvent mourir que sur le mur et le taux de mort spontanée β_0 sera pris égal à 0. Le coefficient β_1 sera pris nul aussi puisqu'il ne correspond à aucun changement dans le système. On introduira aussi le coefficient réduit β :

$$\beta = \sum_{k\geq 2} \beta_k (k-1) \tag{3.1}$$

donnant le taux de croissance de la population en absence de mur. En effet, sur la ligne infinie, le profil moyen $\rho_{\text{libre}}(X)$ de la population, *i.e.* la densité d'individus à la position X, satisfait l'équation différentielle suivante :

$$\partial_t \rho_{\text{libre}} = \Delta \rho_{\text{libre}} + \beta \rho_{\text{libre}} \tag{3.2}$$

où $\Delta = \partial_X^2$ désigne le laplacien unidimensionnel. En partant d'un unique individu à la position x à $t = 0$, on obtient :

$$\rho_{\text{libre}}(X,t) = \frac{1}{\sqrt{4\pi t}} e^{-(X-x)^2/(4t)} e^{\beta t}, \tag{3.3}$$

qui est le produit d'un terme de diffusion et d'un terme de croissance. On vérifie facilement que la taille moyenne est alors donnée par $N_t = e^{\beta t}$. Les points de densité constante ρ évoluent aux temps longs comme

$$x_{\text{const}}(t) = (2\sqrt{\beta})t + o(t)$$

[1] On se place dans l'échelle d'espace où le coefficient de diffusion vaut 1.

Ainsi, grâce au terme croissance en $\beta\rho$, le domaine occupé par la population s'étend linéairement en temps à une vitesse

$$\boxed{v_c = 2\sqrt{\beta}} \tag{3.4}$$

lorsque β est un taux de croissance positif. Ainsi, en présence d'un mur absorbant se déplaçant à la vitesse v, on s'attend à l'existence de deux types de comportement : si $v > v_c$, le mur réussit toujours rattraper tous les individus et la population finit par s'éteindre ; si $v < v_c$, le mur ne rattrape pas nécessairement les individus à l'avant du front et la population a une chance non nulle de survivre et de croître exponentiellement.

En présence d'un mur absorbant se déplaçant à une vitesse v, l'équation (3.2) devient dans le référentiel du mur :

$$\partial_t \rho_{\text{mur}} = \Delta\rho_{\text{mur}} + v\partial_x\rho_{\text{mur}} + \beta\rho_{\text{mur}} \tag{3.5}$$

avec les conditions aux limites $\rho_{\text{mur}}(X = 0) = 0$ et $\lim_{X\to+\infty}\rho_{\text{mur}}(X) = 0$ dans le référentiel du mur. Une simple méthode des images donne alors :

$$\rho_{\text{mur}}(X) = \frac{1}{\sqrt{4\pi t}} e^{(\beta - v^2/4)t} e^{-v(X-x)} \left[e^{-(X-x)^2/(4t)} - e^{-(X+x)^2/(4t)} \right] \tag{3.6}$$

et on retrouve le comportement attendu selon le signe de $v - 2\sqrt{\beta}$.

Ces résultats sont similaires à ceux obtenus dans [ABTR07]. Néanmoins, ce profil moyen, qui découle de l'étude d'une équation linéaire, ne permet pas d'extraire la probabilité d'extinction d'un individu démarrant à une distance x du front.

3.2.2 La probabilité d'extinction

Une quantité-clef de la dynamique du système est la probabilité d'extinction $Q_e(x,t)$ d'un individu à la position x à l'instant $t = 0$, c'est-à-dire la probabilité que tous ses descendants soient éteints à l'instant t. Comme pour le processus de Galton-Watson (voir section 1.2.1), on établira l'équation différentielle satisfaite par $Q_e(x,t)$ en divisant l'intervalle de temps $[0, t+dt]$ en deux intervalles $[0, dt]$ et $[dt, t+dt]$. Pendant le premier instant dt, l'invididu initial diffuse et se divise en k individus avec une probabilité $\beta_k dt$: l'indépendance des individus implique que la probabilité d'extinction du premier individu est alors le produit des probabilités d'extinction de ses enfants et l'on a dans le référentiel du mur :

$$Q_e(x + vdt, t + dt) = \int \frac{e^{-\eta^2/4} d\eta}{\sqrt{4\pi}} Q_e(x + \eta\sqrt{dt}, t) + \sum_k \beta_k dt (Q_e^k - Q_e) \tag{3.7}$$

qui donne l'équation différentielle suivante dans la limite $dt \to 0$:

$$\partial_t Q_e = \partial_x^2 Q_e - v\partial_x Q_e + f_e(Q_e), \tag{3.8}$$

$$f_e(Q) \hat{=} \sum_{k\geq 2} \beta_k (Q^k - Q). \tag{3.9}$$

L'équation précédente est *exacte* et la factorisation en Q_e^k est possible grâce à l'indépendance des lignées. Afin d'établir une analogie plus complète avec le chapitre précédent, nous considérerons plutôt la probabilité de survie définie par :

$$Q_s(x,t) = 1 - Q_e(x,t) \tag{3.10}$$

qui satisfait l'équation :

$$\begin{aligned} \partial_t Q_s &= \partial_x^2 Q_s - v\partial_x Q_s + f_s(Q_s), & (3.11)\\ f_s(Q) &\hat{=} -f_e(1-Q) = \sum_{k\geq 2} \beta_k \left[(1-Q) - (1-Q)^k\right]. & (3.12) \end{aligned}$$

La non-linéarité f_s satisfait les conditions suivantes :

$$\begin{cases} f_s(0) = 0, & f_s'(0) = \beta > 0, \\ f_s(1) = 0, & f_s'(1) < 0, \\ f_s(Q) > 0, & 0 < Q < 1. \end{cases}$$

D'autre part, la présence d'un mur absorbant précise les conditions aux limites pour $Q_s(x,t)$. L'élimination instantanée sur le mur absorbant implique pour $t > 0$:

$$Q_s(0,t) = 0, \tag{3.13}$$

et l'existence d'un individu initial en $x > 0$ donne :

$$Q_s(x,0) = 1. \tag{3.14}$$

L'interprétation de $Q_s(x,t)$ comme une probabilité de survie fait que, à tout temps t, $Q_s(x,t)$ est une fonction croissante de x (la particule part plus loin du mur absorbant) et à x fixé, $Q_s(x,t)$ est une fonction décroissante du temps (il est plus difficile d'échapper au mur). Ainsi, aux temps longs, $Q_s(x,t)$ tend vers une fonction limite $Q_s^*(x)$ qui croît (au sens large) et satisfait :

$$\begin{aligned} \partial_x^2 Q_s^* - v\partial_x Q_s^* + f_s(Q_s^*) = 0, & (3.15)\\ Q_s^*(0) = 0. & (3.16) \end{aligned}$$

En $+\infty$, l'étude de l'équation précédente montre que $Q_s^*(x)$ ne peut tendre que vers $\pm\infty$, 1 ou 0. Puisque c'est une probabilité, le cas $\pm\infty$ est à exclure. D'autre part, puisque $Q_s^*(x)$ est croissante, on a :

$$Q_s^*(x) = 0 \,\text{pour tout } x, \textit{ ou bien } \lim_{x\to+\infty} Q_s^*(x) = 1 \tag{3.17}$$

Dès que $Q_s^*(x)$ ne s'annule pas, on reconnaît donc en (3.11) une équation de front de type F-KPP (2.2) se propageant vers les x négatifs, correspondant à une phase $Q_s = 1$ en $+\infty$ envahissant une phase instable $Q_s = 0$ en $-\infty$, aux conditions aux bords près (mur).

Nous avons vu au chapitre 2 que les solutions de (2.7) sur la ligne infinie pour $v < v_c$ présentent des oscillations amorties dans la phase instable $Q_v \simeq 0$ pour $x \to -\infty$.

On peut donc trouver, grâce à la liberté de translation, une solution $Q_v(x)$ telle que $Q_v(0) = 0$ et $Q_v(x) > 0$ pour $x > 0$: le zéro le plus à droite de Q_v doit donc coïncider avec le mur $x = 0$. Ainsi, $Q_s^*(x)$ est donnée **au-dessous de** v_c par la partie strictement positive d'une front se propageant sur la ligne infinie vers les x négatifs à une vitesse v. Le théorème d'unicité de la solution qui permet d'assimiler la solution sur la ligne semi-infinie à celle sur la ligne infinie a été prouvée mathématiquement par Pinsky [Pin95] qui cite lui-même [AW78] pour des non-linéarités $f_e(Q) = \beta_2(Q^2 - Q)$. Pour une preuve plus récente, on consultera Harris, Harris et Kyprianou [HHK06]. Nous supposerons ici que ce théorème d'unicité reste valable lorsque les taux de division β_k sont quelconques.

Au-dessus de v_c, le mur avance trop rapidement pour que la population puisse survivre : on a donc $Q_s^*(x) = 0$ pour tout x. Les conditions au bord font qu'aucun raccordement n'est possible avec un front monotone sur la ligne infinie pour $v > v_c$ car un tel front ne s'annule pas en un x fini.

Cette marche aléatoire avec branchements en présence d'un mur absorbant présente donc une transition de phase vers un état absorbant selon la valeur de la vitesse du mur v et les calculs précédents ramènent l'étude de la probabilité de survie à un problème de propagation de fronts.

Dans la suite de ce chapitre, nous nous attacherons à étudier $Q_s^*(x)$ au voisinage de la vitesse critique ainsi que la relaxation vers ce point fixe.

3.3 Simulations numériques

Pour les simulations numériques, nous considérerons la version discrétisée suivante :
- les individus se déplacent sur $\mathbb{Z}$;
- à chaque pas de temps, chaque individu se divise en k enfants ;
- pour un parent sur un site x, chaque enfant est positionné au site $x + 1$ avec une probabilité p, au site $x - 1$ avec une probabilité q ou sur le même site x avec la probabilité $r = 1 - p - q$.

La prise en compte du mur absorbant est plus complexe que dans le cas continu. Pour une vitesse entière, cela revient à éliminer les individus sur les v premiers sites à chaque génération. Pour une vitesse non-entière, le nombre de sites à supprimer dépend du pas de temps. Par exemple, pour une vitesse fractionnaire $v = m/n$ avec $m < n$, on éliminera les individus sur le premier site m pas de temps sur n. Plus généralement, pour une vitesse réelle, auront été éliminés au temps t tous les sites x tels $x \leq \lfloor vt \rfloor$ où $\lfloor \cdot \rfloor$ désigne la partie entière. Le cas des vitesses non entières pose un problème de convergence pour $Q_s(x, t)$ puisque les règles d'évolution dépendent du temps. En particulier, pour des vitesses fractionnaires $v = m/n$, la probabilité de survie tend vers un cycle de période n, et non plus vers un point fixe $Q_s^*(x)$ (voir figure 3.1). Nous ne nous préoccuperons pas ici de décrire de tels cycles limites et nous nous limiterons aux vitesses entières.

Nous travaillerons à vitesse nulle[2] $v = 0$. Ainsi, en partant d'un individu initial en $x > 0$ et en séparant le premier pas de temps du reste de l'évolution, on montre que la

[2]De manière strictement équivalente, nous pourrions prendre $v = 1$ avec p, q, r probabilités de sauter du site x vers $x + 2$, x et $x + 1$.

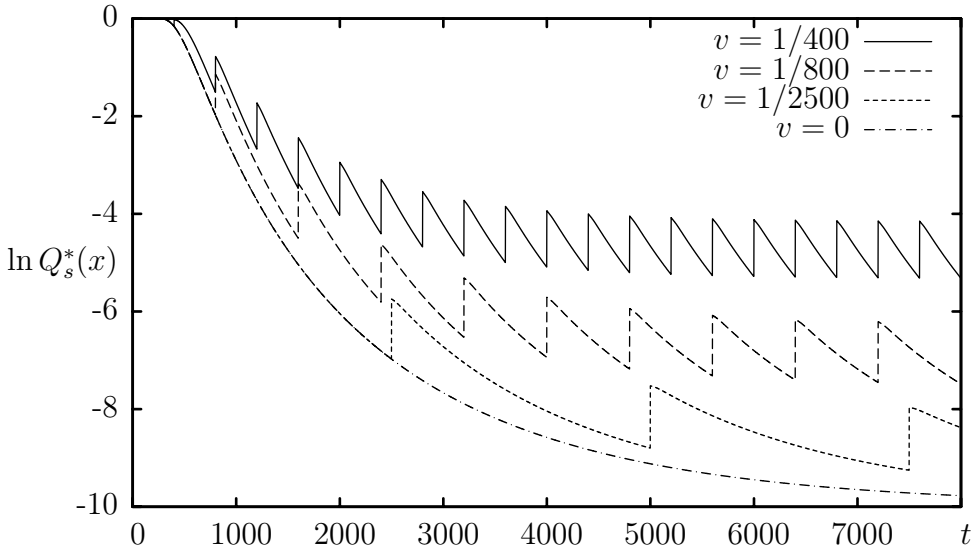


FIG. 3.1: Comportement de $Q_s(x,t)$ en $x = 15$ dans le modèle sur réseau discret de la section 3.3 pour différentes vitesses fractionnaires et la vitesse $v = 0$. Des cycles limites apparaissent lorsque la vitesse est fractionnaire. On remarquera de plus le retard à la décroissance aux temps courts ($t \leq 400$: cela correspond au temps nécessaire à la population initiale avant de percevoir les effets du mur (cf. section 3.5).

probabilité d'extinction est donnée par :

$$Q_e(x,t+1) = \left(pQ_e(x+1,t) + qQ_e(x-1,t) + rQ_e(x,t)\right)^k \tag{3.18}$$

avec la condition au bord $Q_e(0,t) = 1$. La probabilité de survie suit donc l'évolution suivante :

$$Q_s(x,t+1) = 1 - \left(1 - pQ_s(x+1,t) - qQ_s(x-1,t) - rQ_s(x,t)\right)^k \tag{3.19}$$

avec la condition au bord $Q_s(0,t) = 0$.

Pour explorer le voisinage de la vitesse critique, nous conserverons une vitesse $v = 0$ nulle mais les paramètres p, q et r seront ajustés pour ajuster v_c et obtenir l'écart au point critique $\epsilon = v - v_c$ voulu. La vitesse critique v_c est obtenue en se plaçant à l'avant du front ($x \to -\infty$) et où Q_s prend la forme $Q_s(x,t) \sim e^{\gamma(x+vt)}$ (cf. chapitre 2). La relation reliant la vitesse v et la décroissance γ est alors obtenue en linéarisant (3.19) :

$$e^{\gamma v} = k\left(pe^{\gamma} + qe^{-\gamma} + r\right). \tag{3.20}$$

Nous avons ainsi :

$$v_c = \inf_{\gamma} \frac{\ln\left[k(pe^{\gamma} + qe^{-\gamma} + r)\right]}{\gamma}. \tag{3.21}$$

Pour $k = 2$ et le triplet particulier $(p_c, q_c, r_c) = (1/18, 16/18, 1/18)$, nous obtenons $\gamma_c = \ln 4$ et une vitesse critique nulle $v_c = 0$. Pour se décaler du point critique, nous

modifierons ainsi (p, q, r) de la manière suivante : $p = p_c - \delta$, $q = q_c + \delta$ et $r = r_c$. Au premier ordre, δ est reliée à l'écart à la vitesse critique $\epsilon = v - v_c$ par :

$$\delta \simeq \frac{15}{2 \ln 4} \epsilon \tag{3.22}$$

Tous les calculs numériques seront faits avec ce modèle et ces valeurs de paramètres sauf mention contraire.

3.4 Allure de Q_s^* près de la vitesse critique

Pour toute vitesse $v < v_c$, nous avons vu que la probabilité $Q_s^*(x)$ est donnée par la partie droite d'un front $Q_v(x)$ se déplaçant vers les x négatifs. Dans la région $Q_v \simeq 0$, l'équation F-KPP (2.2) peut être linéarisée et (2.24, 2.30, 2.31) impliquent que, dans le référentiel tel que $Q_v(x_0) = 1/2$, $Q_v(x)$ prend la forme $Q_v(x) \simeq -(A_c/\pi) L \sin(\pi(x + B_c/A_c)/L) e^{v_c x/2}$. Le zéro le plus à droite de $Q_v(x)$ se situe donc en $-L - B_c/A_c$. Pour revenir à la probabilité de survie en présence du mur, il suffit de changer de référentiel pour ramener le zéro le plus à droite de la solution sur la ligne infinie Q_v sur le mur en $x = 0$.

Ainsi, nous obtenons pour $Q_s^*(x)$ la forme suivante (voir figures 3.2 et 3.3) :

$$\boxed{Q_s^*(x) \underset{\text{région 1}}{\simeq} \frac{A_c L}{\pi} \sin\left(\frac{\pi x}{L}\right) e^{v_c(x - L - B_c/A_c)/2}} \tag{3.23}$$

dans la région où l'approximation linéaire est valable, c'est-à-dire la région $L - x \gg 1$ (région I). Cette région est de longueur L, définie par :

$$\boxed{L = \frac{\pi}{\sqrt{\beta - v^2/4}} \propto (v_c - v)^{-1/2}} \tag{3.24}$$

Au-delà de $x = L$, $Q_s^*(x)$ ressemble à la solution critique à des corrections en $(v_c - v) \sim 1/L^2$ près :

$$Q_s^*(x) \underset{\text{région 2}}{=} Q_{v_c}(x - L - B_c/A_c) + O(L^{-2}) \tag{3.25}$$

On notera l'analogie entre la forme de $Q_s^*(x)$ sous la vitesse critique et celle d'un front avec *cut-off* : dans les deux cas, la forme en arche de sinus est imposée par une condition de raccordement à une valeur petite (le *cut-off*) ou nulle (le mur) loin dans le domaine $Q_s \ll 1$. Dans les deux cas, les propriétés peuvent être déduites de l'approche perturbative décrite en section 2.4.2.

Les deux régions $x < L$ et $x > L$ correspondent ainsi aux deux destinées suivantes de la particule initiale :

1. un individu situé initialement dans la région I ($x < L$) se fait rattraper, avec une probabilité proche de 1, par le mur avant d'avoir pu proliférer et est presque sûr de disparaître,

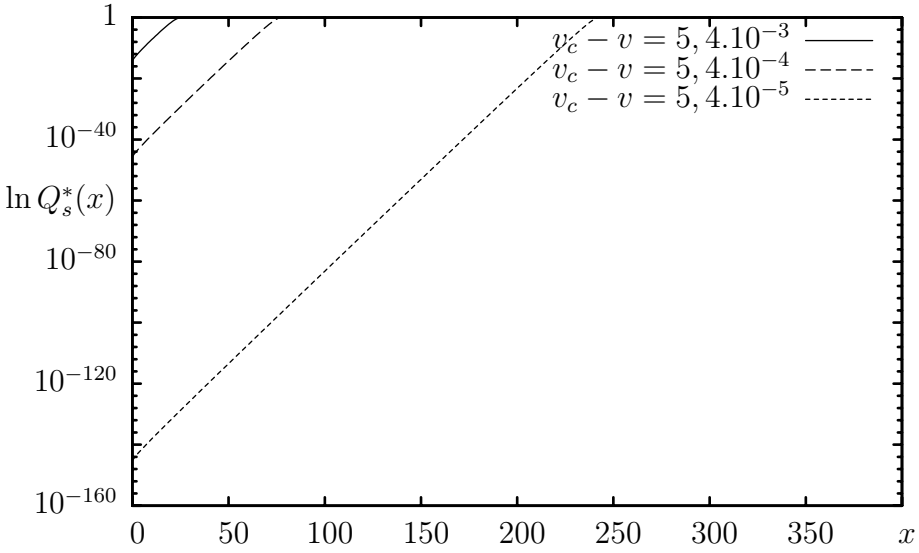


FIG. 3.2: Calcul numérique de la probabilité de survie éternelle $Q_s^*(x)$ pour différentes vitesses proches de la vitesse critique. Les courbes mettent en évidence pour chaque vitesse l'existence de deux régions : l'une de taille L où $Q_s^*(x) \ll 1$, l'autre au-delà de L avec $Q_s^*(x) \simeq 1$. On vérifie que $L \propto (v_c - v)^{1/2}$.

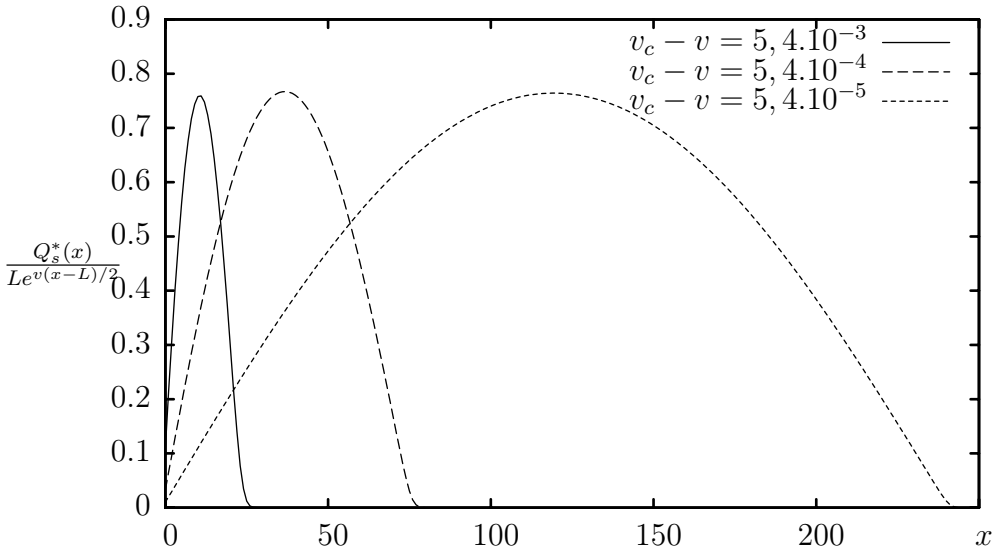


FIG. 3.3: Calcul numérique de $\frac{1}{L}Q_s^*(x)e^{-v(x-L)/2}$ pour différentes vitesses proches de la vitesse critique. Pour déterminer L, nous avons utilisé l'expression attendue (3.24). Les données sont les mêmes que celles utilisées pour la figure précédente.

2. un individu situé initialement dans la région II ($x > L$) a une probabilité proche de 1 de se développer suffisamment avant de percevoir l'existence du mur, de telle sorte qu'il pourra entamer une croissance exponentielle et survivre aux temps longs.

Nous notons en particulier que, près de la vitesse critique et à x fixé, $Q_s^*(x)$ ne tend pas vers 0 comme une loi de puissance mais comme une exponentielle étirée :

$$\boxed{Q_s^*(x) \simeq e^{-C(v_c-v)^{-1/2}}} \tag{3.26}$$

3.5 Comportement de la probabilité de survie aux temps longs : une première approche

La description complète de la dynamique aux temps longs pour $v < v_c$ passe par l'étude des temps de relaxation faite dans la section suivante. On peut cependant se faire une première idée grâce à un *ansatz* assez simple (que nous avons présenté dans [DS07]), qui décrit aussi bien les cas $v < v_c$, $v = v_c$ et $v > v_c$. Il est basé sur l'observation suivante : à tout temps t, la demi-droite $x > 0$ se décompose en deux régions, la première, de 0 à L_t, où l'approximation linéaire dans (2.2) peut être faite, la seconde correspondant à $x > L_t$ où se manifestent les non-linéarités. Cette division est la généralisation à un temps quelconque t de la forme (3.23, 3.25) de la probabilité $Q_s^*(x)$ et de son interprétation en termes de survie d'un individu.

Puisque le front est de type *pulled*, il faut s'attendre à ce que la dynamique la plus lente se manifeste sur la partie linéaire [van03, Pv02] et nous allons nous focaliser sur celle-ci. Nous supposerons que $Q_s(x,t)$ prend l'allure suivante dans la partie linéaire $0 \leq x < L_t$:

$$Q_s(x,t) = \frac{A}{\pi} L_t \sin\left(\frac{\pi x}{L_t}\right) e^{v(x-L_t)/2} \tag{3.27}$$

où le sinus continue à assurer la nullité en zéro. Le comportement pour $x \simeq L_t$ assure le raccordement à la partie non-linéaire (qui est censée relaxer plus rapidement et donc avoir déjà pris sa forme finale). Cet *ansatz* a pour but de capturer la dynamique lente de la fonction $Q_s(x,t)$ en un seul paramètre L_t, la taille du domaine linéaire. Le coefficient A aux temps longs doit tendre vers A_c : à t fini néanmoins, nous supposerons qu'il a une variation lente de telle sorte qu'on puisse négliger les contributions de ses variations devant celle de L_t. Pour la suite du calcul, nous prendrons donc $A = A_c$.

L'interprétation du paramètre L_t est la suivante : la lignée d'un individu situé initialement en x est déjà éteinte au temps t si $x > L_t$ (avec une probabilité proche de 1) et a une forte probabilité d'être encore active si $x > L_t$. Ainsi, nous nous attendons aux temps longs à ce que $L_t \to +\infty$ si $v \geq v_c$ et $L_t \to L$ si $v < v_c$ où L est la longueur (3.24).

En injectant (3.27) dans l'équation (2.2) linéarisée, nous obtenons

$$\begin{aligned}\partial_t Q_s(x,t) &= \partial_t L_t \left[\left(-\frac{vA_cL_t}{2\pi} + \frac{A_c}{\pi}\right)\sin\left(\frac{\pi x}{L_t}\right) - \frac{A_c\pi x}{\pi L_t}\cos\left(\frac{\pi x}{L_t}\right)\right] e^{v(x-L_t)/2} \\ &\simeq -\partial_t L_t \frac{vA_c}{2\pi}\sin\left(\frac{\pi x}{L_t}\right) e^{v(x-L_t)/2},\end{aligned}$$

et, d'autre part,

$$\partial_x^2 Q_s - v\partial_x Q_s + \beta Q_s(x,t) = \frac{A_c}{\pi} L_t \left[-\frac{\pi^2}{L_t^2} - \frac{v^2}{4} + \beta \right] \sin\left(\frac{\pi x}{L_t}\right) e^{v(x-L_t)/2}.$$

L'*ansatz* (3.27) est valable à l'ordre dominant en L_t^{-1} et $v - v_c$ pourvu que L_t ait la dynamique suivante à l'ordre dominant en $(v - v_c)$:

$$\partial_t L_t = (v - v_c) + \frac{2\pi^2}{v_c L_t^2} \tag{3.28}$$

Les deux termes du membre de droite sont de même ordre à condition que L_t soit d'ordre $|v - v_c|^{-1/2}$, comme attendu en (3.24). De plus, on s'attend à ce que L_t s'annule aux temps courts puisque le domaine linéaire n'a pas eu le temps de croître à partir de $Q_s(x,0) = 1$.

Le comportement de la solution de l'équation précédente dépend alors du signe de $v - v_c$. Dans le cas $v < v_c$, elle prend la forme d'échelle en $v - v_c$:

$$L_t = \frac{\pi}{\sqrt{v_c(v_c - v)/2}} \xi_1 \left(\frac{\sqrt{v_c/2}}{\pi} (v_c - v)^{-3/2} t \right) \tag{3.29}$$

où la fonction $\xi_1(z)$ satisfait $\xi_1' = -1 + 1/\xi_1^2$. Après intégration, la solution correspondant à $\xi_1(0) = 0$ satisfait :

$$-2\xi_1(z) + \ln \frac{1 + \xi_1(z)}{1 - \xi_1(z)} = 2z, \quad i.e. \quad \xi_1(z) = \tanh\left(\xi_1(z) + z\right) \tag{3.30}$$

d'où on extrait les comportements suivants :

$$\begin{aligned} \xi_1(z) &= (3z)^{1/3} - 3z/5 + O(z^{5/3}), \quad \text{pour } z \to 0 \\ \xi_1(z) &= 1 - 2e^{-2z-2} + o(e^{-2z-2}), \quad \text{pour } z \to +\infty \end{aligned}$$

Nous lisons sur ce résultat la dynamique aux temps longs de la taille du domaine linéaire L_t : $Q_s(x,t)$ converge bien vers $Q_s^*(x)$ puisque $L_t \to L$ et le temps de relaxation associé est :

$$\boxed{\tau_1 \simeq \frac{\pi}{\sqrt{2v_c}} (v_c - v)^{-3/2} \qquad (v < v_c)} \tag{3.31}$$

à l'ordre dominant en $v_c - v$.

Au-delà de v_c, la longueur L_t diverge puisque la population ne peut pas survivre et $Q_s(x,t) \to 0$ lorsque $t \to \infty$. De fait, L_t prend aussi une forme d'échelle lorsque v tend vers v_c :

$$L_t = \frac{\pi}{\sqrt{v_c(v - v_c)/2}} \xi_2 \left(\frac{\sqrt{v_c/2}}{\pi} (v - v_c)^{-3/2} t \right) \tag{3.32}$$

avec $\xi_2(z)$ satisfaisant cette fois-ci $\xi_2' = 1 + 1/\xi_2^2$ et nous obtenons après intégration :

$$\xi_2(z) = \tan\left(\xi_2(z) - z\right) \tag{3.33}$$

Le comportement aux temps courts est inchangé alors que pour z grand, le développement asymptotique est le suivant :

$$\xi_2(z) = z + \frac{\pi}{2} - \frac{1}{z} + O(1/z), \quad \text{pour } z \to +\infty$$

La longueur de domaine linéaire croît alors linéairement aux temps longs comme :

$$L_t \simeq (v - v_c)t \tag{3.34}$$

Ainsi, à x fixé, la probabilité de survie décroît exponentiellement à l'ordre dominant :

$$Q(x,t) \sim \exp\left[\frac{v_c}{2}\left(x - (v - v_c)t\right)\right], \qquad v > v_c \tag{3.35}$$

dès l'instant où l'abscisse x appartient au domaine linéaire $x < L_t$.

Enfin, à la vitesse critique $v = v_c$, nous avons simplement :

$$L_t \simeq \left(\frac{6\pi^2}{v_c}t\right)^{1/3} \tag{3.36}$$

et la probabilité de survie décroît comme une exponentielle étirée :

$$\boxed{Q_s(x,t) \sim \exp\left[-\frac{v_c}{2}\left(\frac{6\pi^2}{v_c}t\right)^{1/3}\right], \quad (v = v_c)} \tag{3.37}$$

Cette approche, bien que loin d'être rigoureuse, permet de retrouver de manière simple les comportements (3.35) et (3.37) démontrés rigoureusement par Kesten [Kes78] et Harris et Harris [HH07]. De plus, elle prévoit la forme de $Q_s(x,t)$ et son évolution également à des temps plus courts (dès que la partie non-linéaire a convergé vers sa forme finale, de telle sorte que l'*ansatz* (3.27) est justifié). Les comparaisons avec les résultats numériques, tant pour la forme en arche de sinus que pour les convergences (3.35) et (3.37), sont présentées dans les figures 3.4 et 3.5. Enfin, l'*ansatz* donne pour $v < v_c$ une estimation du temps de relaxation le plus lent que nous allons retrouver ci-dessous, par un calcul plus détaillé.

3.6 Temps de relaxation sous la vitesse critique

L'approche précédente permet de comprendre le temps de relaxation le plus lent sous la vitesse critique. Pour les autres temps de relaxation, il convient de procéder comme au chapitre 2 et de linéariser l'équation (3.11) autour de $Q_s^*(x)$ pour $v < v_c$. Nous allons donc chercher les vecteurs et valeurs propres de l'opérateur linéaire défini par :

$$\mathcal{L}[\phi] \hat{=} \partial_x^2\phi - v\partial_x\phi + f_s'(Q_s^*)\phi. \tag{3.38}$$

Puisque $Q_s(0,t) = 0$ pour tout t, les vecteurs propres doivent aussi satisfaire $\phi(0) = 0$. De même, ils doivent s'annuler à l'infini au moins aussi vite que $Q_s(x,t)$ (donc au moins exponentiellement).

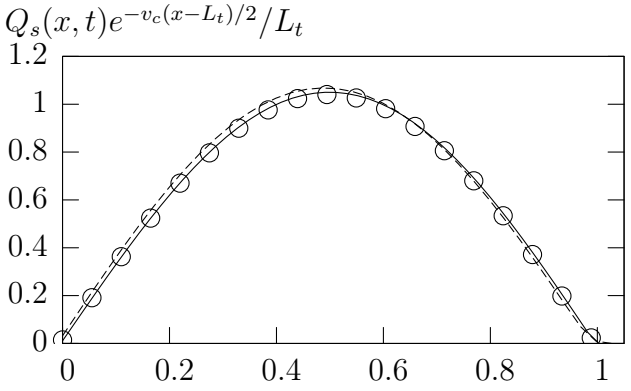


FIG. 3.4: Allure (numérique) de la probabilité de survie normalisée $Q_s(x,t)e^{-v_c(x-L_t)/2}/L_t$ en fonction de x/L_t dans la région I où $Q_s(x,t) \ll 1$ à différents temps $t = 10^5$ (pointillés), $t = 5.10^5$ (cercles), $t = 11.10^5$ (ligne simple) pour un écart à la vitesse critique $v - v_c \simeq 10^{-4} > 0$. Les longueurs L_t sont mesurées directement comme le point tel que $Q_s(L_t,t) = 0.5$ (interpolation entre les points du réseau discret) et valent aux différents temps $L_t \simeq 1,05.10^2$, $2,01.10^2$, $2.95.10^2$. L'accord avec l'*ansatz* (3.27) est excellent.

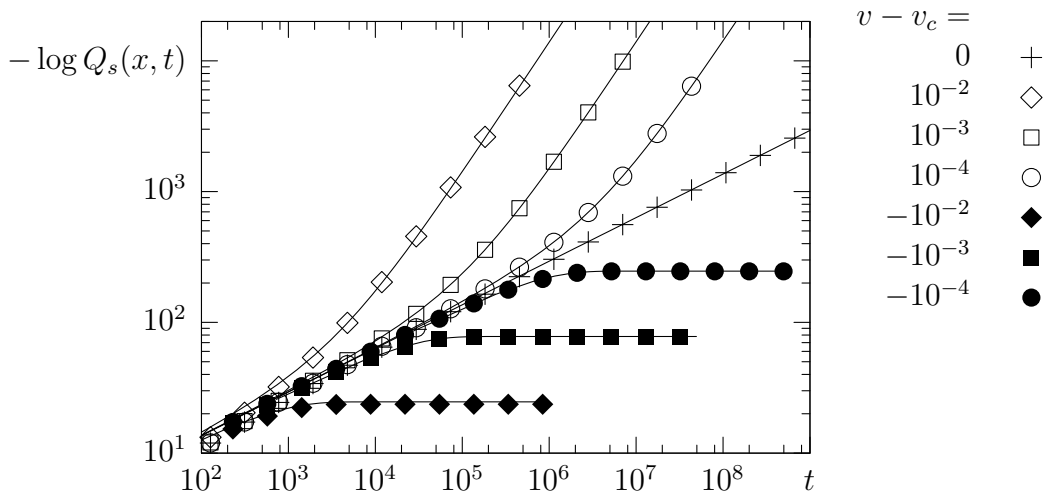


FIG. 3.5: Évolution (numérique) de $-\ln Q_s(x,t)$ en $x = 1$ pour différents écarts $v - v_c$ à la vitesse critique obtenue en itérant (3.19). Les lignes correspondent aux prédictions de la section 3.5 et sont en accord avec les résultats numériques aux temps longs des deux côtés du point critique.

Si nous comparons les expressions (3.38) et (2.20) et si nous nous souvenons que Q_s^* est donné par la partie droite d'une solution de front $Q_v(x)$, alors le problème est presque déjà entièrement résolu. En effet, dans la section 2.4.2, nous avons considéré des vecteurs propres sur la ligne infinie satisfaisant l'équation $\mathcal{L}[\phi_\lambda] = \lambda\phi_\lambda$ avec la même condition aux limites en $+\infty$. De même que Q_s^* est obtenu en restreignant une solution de front $Q_v(x)$, les vecteurs propres en présence du mur sont obtenus en sélectionnant les vecteurs propres sur la ligne infinie qui satisfont les bonnes conditions aux limites puis en les restreignant sur le même domaine que $Q_v(x)$. De fait, la condition aux bords sur le mur conduit ainsi à un ensemble discret de valeurs propres. La position du mur correspond à la position du zéro le plus à droite de $Q_v(x)$: il faut donc que les vecteurs propres ϕ_λ s'annulent au même endroit. Bien évidemment, le zéro de ϕ_λ coïncidant avec le zéro le plus à droite de $Q_v(x)$ n'est pas nécessairement le zéro le plus à droite de ϕ_λ et on obtient ainsi les temps de relaxation successifs.

Dans la démarche du chapitre 2, nous avons obtenu la forme des vecteurs propres ϕ_λ perturbativement pour λ et $v_c - v$ petits et nous allons utiliser ci-dessous la périodicité du sinus de (2.25) pour obtenir les valeurs propres discrètes lorsque λ est petit. Néanmoins, pour λ d'ordre 1, le développement (2.25) n'est plus valable et, par analogie avec l'équation de Schrödinger pour un puits fini, il faut s'attendre à la présence d'un continuum lorsque λ est d'ordre 1 : nous ne préoccuperons pas de la forme des vecteurs propres dans le continuum puisque seules les premiers temps de relaxation nous seront utiles par la suite.

D'après (2.24) et (2.25), le zéro le plus à droite de $Q_v(x)$ est situé en $-L - \Phi_v$ à des corrections exponentiellement petites près lorsque la longueur L diverge près de v_c et le n-ème zéro de ϕ_λ est situé en $-nL_\lambda - \Psi_{v,\lambda}$. Leur coïncidence donne alors :

$$nL_\lambda - L = \Phi_v - \Psi_{v,\lambda} = -\frac{2}{v_c} + O\left(\frac{1}{L^2}\right) + O\left(\frac{1}{L_\lambda^2}\right) \tag{3.39}$$

Le terme de gauche dans cette équation est d'ordre L alors que la première correction dépendant du détail des non-linéarités dans le dernier terme à droite n'est que d'ordre $1/L^2$: ainsi, les trois premiers termes du développement de λ en $1/L$ ne dépendent pas des non-linéarités. Étant donné les définitions (2.26,2.27) de L_λ et L, la n-ème valeur propre λ_n a donc le développement suivant :

$$\lambda_n = -\frac{(n^2-1)\pi^2}{L^2} - \frac{n^2 v_c \pi^2}{L^3} - \frac{3n^2 v_c^2 \pi^2}{L^4} + O\left(\frac{1}{L^5}\right) \tag{3.40}$$

Les vecteurs propres ϕ_n associés à chaque valeur propre λ_n sont ainsi donnés, dans la région I de taille L à partir du mur par la translation des vecteurs ϕ_{λ_n} obtenus en (2.25), par :

$$\phi_n(x) = (-1)^{n-1}\frac{A_c v_c L}{n\pi}\left[\sin\left(\frac{n\pi x}{L}\right) + O(1/L)\right]\exp\left[-\frac{v_c}{2}\left(x - L - \frac{B_c}{A_c}\right)\right] + O(e^{v_c(x-L)}) \tag{3.41}$$

puis, dans la région II où les non-linéarités de Q_s^* se manifestent, par :

$$\phi_n(x) \simeq \partial_x Q_{v_c}(x - L - B_c/A_c). \tag{3.42}$$

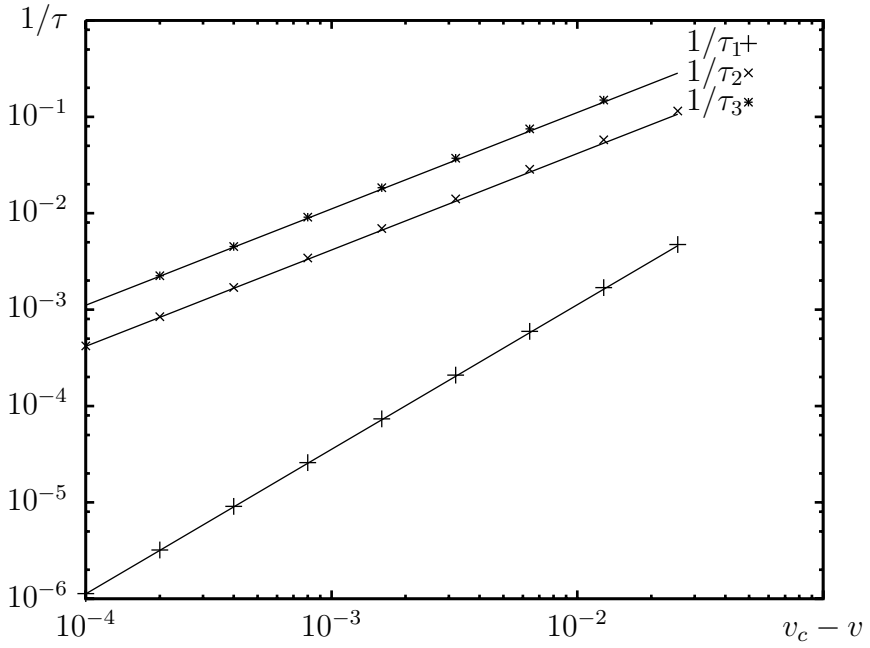


FIG. 3.6: Temps de relaxation τ_1, τ_2, τ_3 en fonction de $v_c - v$ pour $v < v_c$. Les données numériques (points) sont obtenues dans le cadre du modèle discret de la section 3.3. Les lignes correspondent aux prédictions théoriques (3.40) sur les valeurs propres $\lambda_n = -1/\tau_n$ et sont en parfait accord avec les mesures numériques.

Les valeurs propres à l'ordre le plus bas sont semblables aux énergies d'une particule quantique dans un puits de potentiel de largeur L. En effet, si l'on transforme l'équation aux valeurs propres en équation de Schrödinger comme en (2.21), alors le potentiel est relié à la forme de $Q_s^*(x)$: il possède une partie basse de largeur L correspondant à la région I où $Q_s^*(x) \ll 1$ et une partie haute correspondant à la région II où $Q_s^*(x) \simeq 1$. La hauteur du puits est d'ordre $O(1)$, si bien que le développement précédent des valeurs propres n'est plus valable lorsque l'énergie de la particule est de l'ordre de la barrière d'énergie du puits, *i.e.* lorsque n est d'ordre L. Ce résultat était attendu puisque le développement perturbatif des vecteurs propres supposait que λ était proche de 0. L'analogie avec un puits de potentiel nous dit aussi que le spectre de $\mathcal{L}$ devrait avoir une partie continue à partir de $\lambda = O(1)$.

Pour $v > v_c$ cependant, l'approche linéaire ne suffit pas pour caractériser l'évolution du système aux temps longs. L'image de l'arche de sinus développée en section 3.5 a montré en effet qu'aux temps longs, $Q_s(x,t)$ prend une allure de type front dans la région $x > L_t$ et ne tend donc pas uniformément vers 0 : il est alors vain de linéariser l'équation F-KPP autour du point fixe $Q_s^*(x) = 0$. Cette première mise en garde pour $v > v_c$ est corroborée par le fait que l'analogie avec une équation de Schrödinger conduit à un potentiel constant et donc à un spectre entièrement continu : le comportement aux temps longs n'est alors plus dominé par la valeur propre la plus proche de zéro comme en-dessous de la vitesse critique.

3.7 Généralisation à des géométries plus complexes

3.7.1 Détermination du point critique

Les équations écrites pour la marche aléatoire unidimensionnelle se généralisent facilement à des géométries quelconques. Cela concerne la majorité des modèles géographiques de population, à la géométrie généralement plus complexe qu'une demi-droite. La détermination des temps de relaxation sera certes spécifique à chaque domaine mais un certain nombre de constructions, dont celle du chapitre 5, restent valables.

Les équations de type F-KPP en milieu d-dimensionnel sur des domaines plus généraux ont déjà fait l'objet de nombreux travaux mathématiques [BHR05a, BHR05b, BHR04].

En dimension d, nous considérerons un domaine Ω et sa frontière notée $\partial\Omega$. La diffusion d'une particule sera ainsi représentée par un coefficient de diffusion D, supposé ici constant, et un champ de dérive $\vec{v}(\vec{x})$ qui est un vecteur d-dimensionnel dépendant de la position. La reproduction des individus sera caractérisée par des taux de division $\beta_k(\vec{x})$ qui peuvent à leur tour ne pas être constants. Les individus peuvent être éliminés des deux manières suivantes :

- soit en ajoutant un taux de mort spontanée $\beta_0(\vec{x})$, qui s'inclura dans la définition de la non-linéarité $f_s(Q)$;
- soit en prenant des bords absorbants qui imposent la condition aux bords $Q_s(\vec{x},t) = 0$ si $\vec{x} \in \partial\Omega$ sur la probabilité de survie $Q_s(\vec{x},t)$.

La probabilité de survie d'un individu initial placé en $\vec{x}$ satisfait l'équation aux dérivées partielles suivante :

$$\partial_t Q_s = D\vec{\nabla}^2 Q_s + \vec{v}\cdot\vec{\nabla}Q_s + f_s(\vec{x},Q_s) \tag{3.43}$$

où la non-linéarité $f_s(\vec{x},Q)$ est identique à (3.12). La probabilité de survie $Q_s(\vec{x},t)$ est une fonction décroissante du temps, minorée par 0 et qui tend donc vers une fonction limite $Q_s^*(\vec{x})$. La question est alors de savoir en fonction du champ de dérive $\vec{v}(\vec{x})$ et des taux $\beta_k(\vec{x})$ s'il existe une solution stable $Q_s^*(\vec{x})$ non nulle (phase active) ou bien si la population finit toujours par disparaître (phase inactive).

Le seuil de transition entre phases active et inactive pour un domaine Ω est plus facile à obtenir que la détermination de la probabilité de survie $Q_s^*(\vec{x})$. Une première approche consiste à étudier la taille moyenne $\mathcal{N}(\vec{x},t)$ de la population à l'instant t générée par un individu initial en $\vec{x}$ à $t=0$. Le même raisonnement que pour la probabilité d'extinction sur le premier intervalle de temps dt conduit à l'équation différentielle :

$$\partial_t\mathcal{N}(\vec{x},t) = D\Delta\mathcal{N}(\vec{x},t) - \vec{v}(\vec{x})\cdot\vec{\nabla}\mathcal{N}(\vec{x},t) + \sum_k \beta_k(\vec{x})(k-1)\mathcal{N}(\vec{x},t) = \mathcal{K}[\mathcal{N}] \tag{3.44}$$

où $\mathcal{K}$ est un opérateur linéaire. Asymptotiquement, nous avons $\mathcal{N}(\vec{x},t) \propto e^{\mu_1 t}$ où μ_1 est la plus grande valeur propre de l'opérateur $\mathcal{K}$. Il *suffit* donc que la plus grande valeur propre μ_1 de $\mathcal{K}$ ait sa partie réelle négative pour que l'extinction soit certaine :

$$\mathrm{Re}(\mu_1) < 0. \tag{3.45}$$

Des majorations mathématiquement rigoureuses montrent que cette condition est aussi *nécessaire* pour une classe assez vaste de tels modèles [BHR05a, BHR05b, BHR04], en

particulier dès que le domaine Ω est borné. Un moyen de comprendre cela consiste à s'approcher du point critique dans la phase active : la probabilité de survie doit tendre vers 0 sur tout le domaine Ω et l'approximation linéaire de l'équation F-KPP est de plus en plus justifiée. Cette approximation linéaire est exactement l'opérateur $\mathcal{K}$ gouvernant la dynamique de $\mathcal{N}$. Ainsi, on a $\mathcal{K}[Q_s^*] \simeq 0$ et $Q_s^*(x)$ garde un signe constant sur tout Ω : au point critique, $Q_s^*(x)$ ressemble, à une amplitude près, au vecteur propre de $\mathcal{K}$ correspondant à la plus grande valeur propre $\mu_1 = 0$. La véritable preuve mathématique [BHR05a, BHR05b, BHR04] se fait en écrivant des bornes rigoureuses inspirées par ce raisonnement mais nous n'entrerons pas dans ces détails ici. Pour des domaines non-compacts, des difficultés liées à la définition des valeurs propres surgissent mais les mêmes techniques s'appliquent dans de nombreux cas. On remarquera de plus que ces idées sont à rapprocher de l'approximation linéaire donnant la vitesse de fronts sur la ligne infinie.

Lorsque le coefficient de diffusion D est constant et $\vec{v}(\vec{x}) = -\vec{\nabla}U(\vec{x})$, l'opérateur linéaire $\mathcal{K}$ est un objet familier et un changement de variable linéaire sur $\mathcal{N}$ ramène l'étude des valeurs propres de $\mathcal{K}$ à celles des énergies propres d'une particule quantique en présence d'un potentiel qui dépend des taux $\beta_k(\vec{x})$ et de la dérive $\vec{v}(\vec{x})$. L'analogie est complète si le domaine Ω est borné, alors que si Ω n'est pas borné, la condition sur les vecteurs propres de notre problème n'est pas l'intégrabilité du carré de la fonction d'onde. Dans le cas où l'analogie est possible, la condition (3.45) se transforme alors en une condition sur le signe de l'énergie du fondamental de cette particule.

3.7.2 Cas particulier de la boîte en dimension 1

Un exemple simple est celui d'une marche aléatoire avec branchements de coefficient de diffusion $D = 1$, de dérive $-v$ et avec des taux de division β_k constants sur une boîte $\Omega = [0, l]$. La vitesse critique séparant phases active et inactive est donnée par :

$$\beta = \frac{v_c^2}{4} + \frac{\pi^2}{l^2} \tag{3.46}$$

On voit ainsi que la boîte[3] doit être suffisamment large pour qu'une marche avec branchements puisse survivre. Cette condition peut se réécrire :

$$L = l \tag{3.47}$$

où l est la dimension de la boîte et L la longueur définie en (3.24). La longueur L générée par la dynamique de branchement-diffusion s'interprète ainsi comme la longueur minimale dont le système a besoin pour se développer et survivre.

Juste en-dessous de la vitesse critique, $Q_s^*(x)$ peut ainsi être développé en puissance de $(v - v_c)$:

$$Q_s^*(x) = \tilde{Q}_0(x) + (v_c - v)\tilde{Q}_1(x) + \dots$$

[3]En dimension d, le point critique est donné par

$$\beta = \frac{\vec{v}^2}{4} + \pi^2 \left(\frac{1}{l_1^2} + \dots \frac{1}{l_d^2} \right)$$

où l_1,...,l_d désignent les dimensions de la boîte.

À l'ordre le plus bas dans l'équation F-KPP, nous obtenons $\tilde{Q}_0(x) \simeq A_0 \sin(\pi x/l) e^{v_c x/2}$. La constante A_0 n'est fixée qu'à l'ordre suivant en imposant l'annulation de $\tilde{Q}_1(x)$ à la fois en 0 et l. Un rapide calcul montre alors que A_0 doit se comporter comme $(v_c - v)$. Ainsi, près du point critique, nous avons le comportement suivant de $Q_s^*(x)$:

$$Q_s^*(x) \propto (v_c - v) \sin\left(\frac{\pi x}{l}\right) e^{v_c x/2} \tag{3.48}$$

Contrairement à (3.26), la probabilité de survie s'annule ici avec un exposant critique égal à 1.

Une seconde différence significative sépare cette géométrie de la ligne infinie étudiée précédemment : les temps de relaxation sont bien définis des deux côtés de la transition. Les vecteurs propres $\phi_n(\vec{x})$ satisfont :

$$\mathcal{L}[\phi_n] = \partial_x^2 \phi_n - v \partial_x \phi_n + f_s'(Q_s^*)\phi_n = \lambda_n \phi_n \tag{3.49}$$

Dans la phase inactive, l'opérateur $\mathcal{L}$ est identique à l'opérateur $\mathcal{K}$, puisque $Q_s^* = 0$ et le calcul se ramène à celui des énergies d'une particule dans une boîte Ω avec une énergie potentielle égale à $V = -\beta_c + \frac{v^2}{4}$. Les temps de relaxation dans la phase inactive sont alors donnés exactement par :

$$\frac{1}{\tau_n} = \frac{\pi^2 n^2}{l^2} + \frac{v^2}{4} - \beta \tag{3.50}$$

avec les vecteurs propres associés :

$$\phi_n(x) = \sin\left(\frac{n\pi x}{l}\right) e^{vx/2}. \tag{3.51}$$

Dans la phase active $v < v_c$, nous n'avons pas d'expression pour $Q_s^*(\vec{x})$ mais dans la limite $\beta \simeq \beta_c$, $Q_s^*(\vec{x})$ tend uniformément vers 0 et l'opérateur $\mathcal{L}$ tend, à un changement de variable près, vers $\mathcal{K}$ et on observe les mêmes temps de relaxation dans la limite $v \to v_c$ qu'au-dessus de la vitesse critique. Pour le montrer, il suffit de procéder au changement de variable :

$$\phi_n(x) = \frac{\partial_x Q_s^*(x)}{\sin\left(\frac{\pi x}{l} + \Phi\right)} \widetilde{\phi}_n(x) \tag{3.52}$$

où Φ est défini de telle sorte que numérateur et dénominateur s'annule simultanément. Dans la limite $v \to v_c$, $Q_s^*(x)$ prend la forme $e^{vx/2} \sin(\pi x/l)$: dans ce cas, Φ est donné par l'équation implicite $\tan \Phi = 2\pi/(vl)$ et l'équation aux valeurs propres sur $\widetilde{\phi}_n(x)$ devient identique à celle de $\phi_n(x)$ pour $v > v_c$.

Cet exemple simple de domaine borné montre des différences notables avec le cas de la ligne semi-infinie : les temps de relaxation sont définis à la fois au-dessous et au-dessus de la vitesse critique et se comportent de manière identique près de celle-ci. On retrouve ici des caractéristiques classiques des transitions de phase. De plus, à $v \simeq v_c$, seul le premier temps de relaxation diverge comme $|v - v_c|^{-1}$ alors que les autres ont une limite finie. Ces caractéristiques surgiront à nouveau au chapitre 6 dans l'étude du régime quasi-stationnaire.

Chapitre 4

Conditionnement et régime quasi-stationnaire : fonctions génératrices

Ce chapitre commence par une revue de quelques résultats généraux sur les processus conditionnés et quasi-stationnaires. Il se poursuit par l'étude classique du processus de Galton-Watson où nous montrerons comment l'étude des fonctions génératrices permet d'avoir accès aux propriétés du régime quasi-stationnaire. Ensuite, dans une troisième partie, nous présentons la manière dont nous avons généralisé (cf. [DS07, SD08]) cette approche au cas de la marche aléatoire avec branchements avec une taille finale ou une date d'extinction fixées. Enfin, dans une dernière partie, nous décrivons quelques résultats complémentaires utiles pour visualiser les propriétés des régimes conditionnés (calcul numérique de moments, condition suffisante d'existence d'un régime quasi-stationnaire, etc.).

4.1 État absorbant et conditionnement

Dans les modèles de populations, la configuration correspondant à une population vide est un état *absorbant* : un système qui entre dans cette configuration ne peut plus en sortir. En général, de tels systèmes présentent une transition de phase entre une phase active (probabilité non nulle d'échapper aux temps longs à l'état absorbant) et une phase inactive (la population finit toujours par disparaître). Une fois le système tombé dans l'état absorbant, ses propriétés sont triviales puisque la population s'est éteinte. Ainsi, aux temps longs, la partie intéressante de la dynamique correspondra aux événements où la population a survécu. De plus, une simulation numérique « standard » produit beaucoup de séquences menant à une extinction plus rapide que la fenêtre de temps à laquelle on s'intéresse. Il devient donc nécessaire de trouver une méthode de simulation ne produisant que des séquences *actives* sur toute la fenêtre de temps voulue.

Enfin, dans le cas d'une population qui s'éteint à un temps T donné, on peut être

tenté de reconstituer certaines caractéristiques de son histoire entre 0 et T, comme, par exemple, son profil et sa dynamique. Toutes ces questions entrent dans le cadre des probabilités *conditionnelles* : on veut connaître l'évolution du système en prenant en compte notre connaissance de son état final. Étant donné un événement A à l'instant t et un événement B à un instant ultérieur $T > t$, la formule des probabilités conditionnelles donne :

$$\mathrm{Prob}\,(A,t|B,T) = \frac{\mathrm{Prob}\,(B,T|A,t)\,\mathrm{Prob}\,(A,t)}{\mathrm{Prob}\,(B,t)}. \tag{4.1}$$

Ainsi, pour reconstituer l'histoire du système, il faut connaître sa dynamique directe, décrite par le terme $\mathrm{Prob}(B,T|A,t)$, ainsi que les probabilités $\mathrm{Prob}(A,t)$ et $\mathrm{Prob}(B,t)$, qui sont à déduire de la dynamique et des conditions initiales. De telles études [FMM97] ont déjà été menées, par exemple, pour étudier la position initiale d'un mouvement brownien conditionné à passer par 0 à une date t donnée.

Dans les processus stochastiques, l'une des issues possibles aux temps longs est l'existence d'un régime stationnaire, où les propriétés du système ne dépendent plus du temps. Il est défini comme la limite[1], si elle existe, de $\mathrm{Prob}(A,t)$ lorsque $t \to +\infty$. Dans le cas d'un processus conditionné par son état final, une notion analogue est celle de *régime quasi-stationnaire* défini comme la limite, si elle existe, de $\mathrm{Prob}\,(A,t|B,T)$ lorsque $T-t \to +\infty$ et $t \to +\infty$. Une formulation, plus pertinente physiquement, consiste à dire qu'un système conditionné par son état à T atteint un régime quasi-stationnaire pour $t \in [0,T]$ si ses propriétés ne dépendent plus du temps t lorsque l'on est suffisamment loin des bords de la fenêtre $[0,T]$. Dans les données numériques, cela se manifeste par l'apparition de plateaux dans les quantités moyennes lorsque l'on allonge la fenêtre d'observation et de conditionnement $[0,T]$ (voir figure 4.1). Ce régime s'interprète alors comme un régime métastable qui permet au système de retarder l'issue finale du processus (état final de conditionnement) et en particulier d'éviter les configurations absorbantes.

Il nous faut préciser dès à présent que l'existence d'un régime quasi-stationnaire, au sens où nous l'entendons, n'est ni une évidence, ni une caractéristique générale des processus stochastiques avec état absorbant, comme en témoignera l'exemple de la marche aléatoire avec branchements en présence d'un mur absorbant (voir ci-dessous) au-dessus de la vitesse critique. De plus l'existence d'oscillations (valeurs propres complexes) peut aussi venir compliquer la définition d'un régime quasi-stationnaire (cf. section 4.4.2 pour une discussion plus approfondie). Ce chapitre s'attachera en particulier à énoncer quelques critères suffisants pour observer un régime quasi-stationnaire.

Afin d'éviter tout malentendu, il faut préciser que le mot *quasi-stationnaire* désigne parfois dans la littérature mathématique [FMP91, FMP92, CCL$^+$07, SE05, SE04] le conditionnement $\mathrm{Prob}\,(A,t|A \neq \mathcal{C}_{\mathrm{absorbant}},t)$. Cette dernière mesure de probabilité revient à observer le système à un temps t sachant qu'il n'est pas encore tombé dans la configuration absorbante $\mathcal{C}_{\mathrm{absorbant}}$: cela ne présage pas de son comportement ultérieur et, si l'on est dans la phase absorbante, les configurations typiques sont proches de $\mathcal{C}_{\mathrm{absorbant}}$ car il est peu probable que le système survive encore longtemps. Au contraire, la mesure $\lim_{T\to\infty} \mathrm{Prob}\,(A,t|B,T)$ que nous considérons et que nous appelons *quasi-stationnaire* décrit les trajectoires qui ne tomberont jamais dans l'état absorbant : ainsi, les configura-

[1]Dans un souci de simplicité, nous excluons ici le cas des oscillations et des cycles limites.

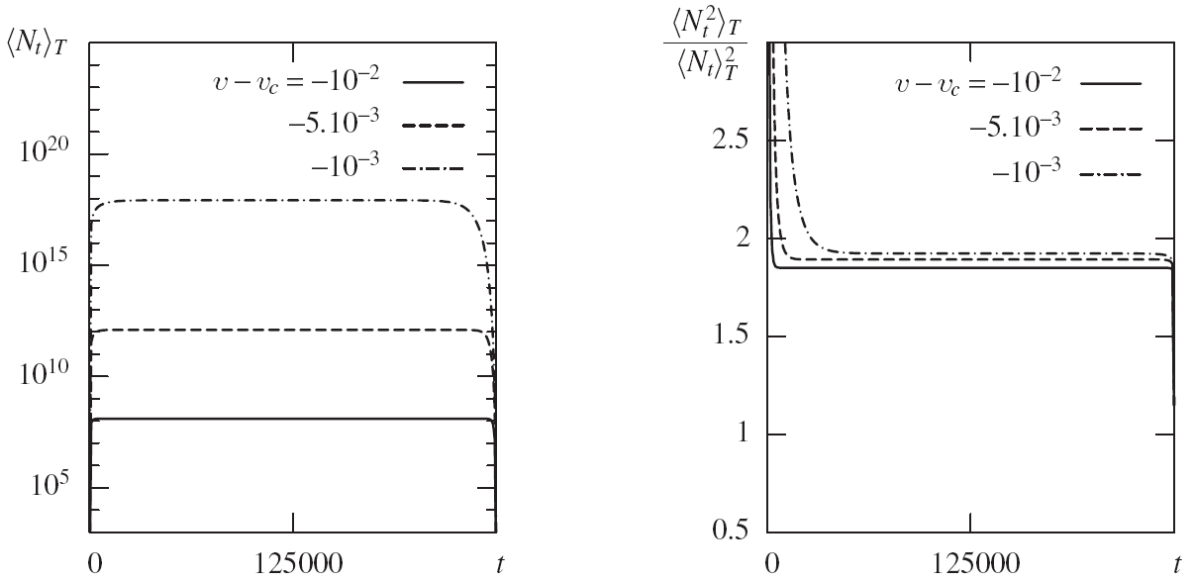


FIG. 4.1: Évolution, pour $v < v_c$ de la marche aléatoire avec branchements en présence d'un mur absorbant lorsqu'elle est conditionnée par une taille finale $N_T = 1$. On constatera l'existence de plateaux pour $t \gg \tau_1$ et $T - t \gg \tau_1$ où τ_1 désigne le temps de relaxation le plus lent du système.

tions typiques n'ont aucune raison d'être proches de la configuration absorbante. Notre définition, plus courante en physique, s'apparente alors aux Q-processus considérés par les mathématiciens [CCL$^+$07]. Le lien entre ces différentes notions est précisé en section 4.4.2.

Les sections suivantes présentent comment, dans des cas simples, certaines quantités comme la taille moyenne de la population peuvent être étudiées dans le régime quasi-stationnaire. En particulier en section 4.2, nous verrons comment, dans le processus de Galton-Watson, les probabilités conditionnelles peuvent être lues dans les fonctions génératrices du système et comment on peut se ramener à l'étude d'un système dynamique. Dans la section 4.4.2, nous énoncerons un critère suffisant pour l'observation d'un régime quasi-stationnaire.

4.2 Processus de Galton-Watson et systèmes dynamiques

4.2.1 Fonctions génératrices et taille finale

Dans le processus de Galton-Watson défini au chapitre 1, la population est entièrement caractérisée par sa taille N_t au temps t. Introduisons les deux fonctions génératrices à

un et deux temps suivantes :

$$G_1(\mu,t) = \left\langle e^{-\mu N_t} \right\rangle \tag{4.2a}$$

$$G_2(\mu,t,\nu,t') = \left\langle e^{-\mu N_t - \nu N_{t+t'}} \right\rangle \tag{4.2b}$$

où μ et ν sont des réels positifs et N_t est la taille à l'instant t de la population générée par un unique individu initial. La fonction G_2 est celle qui permet l'étude du régime conditionné. En effet, si on explicite la moyenne sur la taille $N_{t+t'} = m$, on obtient :

$$G_2(\mu,t,\nu,t') = \sum_{m=0}^{\infty} e^{-\nu m} P_m(t+t') \left\langle e^{-\mu N_t} \middle| N_{t+t'} = m \right\rangle \tag{4.3}$$

où $P_m(t+t')$ est la probabilité d'avoir une taille m à l'instant final $T = t+t'$. La moyenne dans le terme de droite est donc la fonction génératrice de la taille N_t conditionnée sur la taille ultérieure $N_T = m$. Il s'agit alors d'extraire cette information à partir de G_2. Pour cela il suffit de décomposer G_2 en série sur $e^{-\nu}$. Le terme $P_m(t+t')$ peut être obtenu de la même manière si l'on se rend compte que le cas $\mu = 0$ montre que $G_2(0,t,\nu,t') = G_1(\nu, t+t')$. Les propriétés précédentes peuvent ainsi se récrire sous la forme suivante :

$$G_1(\nu, t+t') = \sum_{m=0}^{\infty} e^{-m\nu} P_m(t+t') \tag{4.4}$$

$$G_2(\mu,t,\nu,t') \mathrel{\hat{=}} \sum_{m=0}^{\infty} e^{-m\nu} R_m(\mu,t,t') \tag{4.5}$$

$$\left\langle e^{-\mu N_t} \middle| N_{t+t'} = m \right\rangle = \frac{R_m(\mu,t,t')}{P_m(t+t')} \tag{4.6}$$

et le régime conditionné par la taille finale m est contenu dans le rapport de R_m et P_m.

Les fonctions génératrices G_1 et G_2, quant à elles, évoluent selon la même équation de récurrence (1.2) que la probabilité d'extinction $Q_e(t)$. En effet, de la même manière que la probabilité d'extinction de l'individu initial était égale au produit des probabilités d'extinction de chacun de ses enfants, la taille de la population générée par l'individu initial est la somme des tailles des populations générées par chacun de ses enfants. Dans le cas d'individus indépendants, les évolutions des enfants sont décorrélées et les fonctions génératrices se factorisent. Ainsi, en séparant le premier pas de temps du reste, on obtient les équations de récurrence exactes suivantes :

$$G_1(\mu, t+1) = F_e\left(G_1(\mu,t)\right) \tag{4.7a}$$

$$G_2(\mu, t+1, \nu, t') = F_e\left(G_2(\mu,t,\nu,t')\right) \tag{4.7b}$$

où $F_e(Q) = \sum_k p_k Q^k$ est la même fonction qu'en (1.3). On remarquera que la durée t' ne varie pas puisqu'elle est fixée par le délai entre le temps d'observation t et le temps de conditionnement T et qu'ici nous ôtons le premier pas de temps à la fois à t et T. Les fonctions génératrices sont alors complètement déterminées une fois les conditions initiales précisées : $G_1(\mu,0) = e^{-\mu}$ et $G_2(\mu,0,\nu,t') = e^{-\mu} G_1(\nu,t')$.

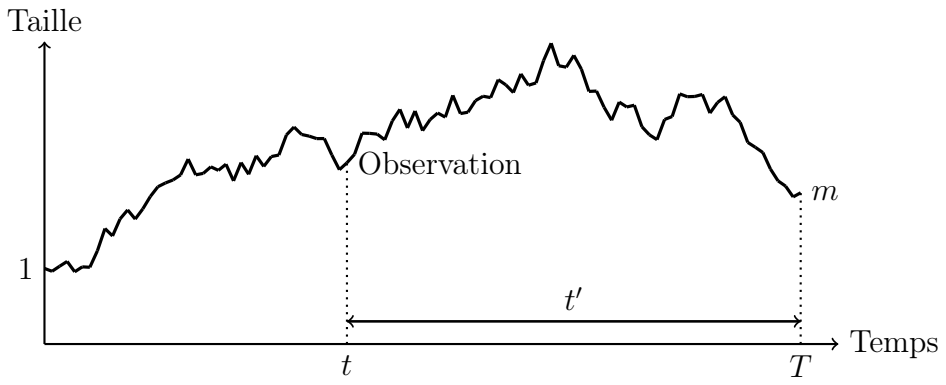


FIG. 4.2: Décomposition de l'intervalle $[0, T]$ en $[0, t]$ et $[t, t + t']$ avec $t' = T - t$. T est le temps final au bout duquel la population est conditionnée à avoir une taille $N_T = m$ finie non nulle.

Les équations (4.7) permettent ainsi d'étudier G_1 et G_2 à la fois analytiquement et numériquement. On peut ainsi mesurer numériquement les fonctions génératrices dans le régime conditionné en décomposant les équations (4.7) à tous les ordres en $e^{-\nu}$ et en utilisant les évolutions induites pour les fonctions P_m et R_m. Cette méthode, généralisée au cas avec espace, est à la base de ce chapitre et de notre travail [DS07]. Le système dynamique défini par $u_{t+1} = F_e(u_t)$ permet donc de décrire tout le régime conditionné en prenant les bonnes conditions initiales.

Le système dynamique (4.7) permet de montrer qu'il existe un régime quasi-stationnaire et d'obtenir la distribution de la taille N_t dans ce régime. La probabilité d'extinction aux temps longs $Q_e^* = 1 - Q_s^*$ est le point fixe stable de (4.7). La dynamique (4.7a) induit donc une convergence de G_1 aux temps longs vers Q_e^* avec un temps de relaxation $\tau_1 = -\ln F_e'(Q_e^*)$:

$$G_1(\mu, t) \underset{t\to+\infty}{\simeq} Q_e^* + \mathcal{A}(e^{-\mu})e^{-t/\tau_1} + o\left(e^{-t/\tau_1}\right) \tag{4.8}$$

où $\mathcal{A}(e^{-\mu})$ est une amplitude ne dépendant que de la condition initiale $G(\mu, 0) = e^{-\mu}$. Pour un système dynamique générique $u_{t+1} = F_e(u_t)$, $\mathcal{A}$ est une fonction de la condition initiale u_0 définie par $\mathcal{A}(u_0) = \lim_{t\to\infty}(u_t - Q_e^*)/e^{-t/\tau_1}$. En particulier, on vérifiera que cette définition conduit à

$$\mathcal{A}'(Q_e^*) = 1 \tag{4.9}$$

et à l'équation fonctionnelle :

$$e^{-1/\tau_1}\mathcal{A}(Q) = \mathcal{A}\left(F_e(Q)\right). \tag{4.10}$$

Lorsque t et t' sont tous deux longs, alors $G_2(\mu, t, \nu, t')$ se décompose de la manière

suivante aux ordres dominants :

$$\begin{aligned} G_2(\mu,t,\nu,t') &= Q_e^* + \mathcal{A}\left(e^{-\mu}G_1(\nu,t')\right)e^{-t/\tau_1} + o\left(e^{-t/\tau_1}\right) \\ &= Q_e^* + \mathcal{A}\left(e^{-\mu}Q_e^* + e^{-\mu}\mathcal{A}(e^{-\nu})e^{-t'/\tau_1} + \ldots\right)e^{-t/\tau_1} + \ldots \\ &= Q_e^* + \mathcal{A}\left(e^{-\mu}Q_e^*\right)e^{-t/\tau_1} \\ &\quad + e^{-\mu}\mathcal{A}'\left(e^{-\mu}Q_e^*\right)\mathcal{A}(e^{-\nu})e^{-(t+t')/\tau_1} + \ldots \end{aligned}$$

En décomposant formellement $\mathcal{A}(e^{-\nu}) = \sum_m a_m e^{-m\nu}$ en série, on se rend compte que, pour $m \geq 1$, les quantités P_m et R_m sont données à l'ordre dominant en $e^{-(t+t')/\tau_1}$ par :

$$\begin{aligned} P_m(t+t') &\sim a_m\mathcal{A}'(Q_e^*)e^{-(t+t')/\tau_1} \simeq a_m e^{-(t+t')/\tau_1}, \\ R_m(\mu,t,t') &\sim a_m e^{-\mu}\mathcal{A}'(e^{-\mu}Q_e^*)e^{-(t+t')/\tau_1}. \end{aligned}$$

Dans la limite $t \to +\infty$ et $t' \to +\infty$, on a ainsi la limite finie suivante :

$$\left\langle e^{-\mu N_t}\middle| N_{t+t'} = m\right\rangle = \frac{R_m(\mu,t,t')}{P_m(t+t')} \to e^{-\mu}\mathcal{A}'(e^{-\mu}Q_e^*). \tag{4.11}$$

Le dernier terme ne dépend alors pas de t et montre bien que l'on a atteint un régime quasi-stationnaire dans la limite $t \gg \tau_1$ et $T - t \gg \tau_1$. On remarquera aussi que cette expression ne dépend pas de la taille finale m de conditionnement : la stratégie de survie pour avoir un nombre fini m de survivants à la date $T = t + t'$ est donc de se placer dans l'état quasi-stationnaire et de « choisir » la taille finale au dernier moment (à partir de $T - t \sim \tau_1$).

Il convient ainsi de définir la distribution de probabilité de taille $\langle e^{-\mu N}\rangle_{\rm qs}$ dans le régime quasi-stationnaire sans préciser ni t, ni t', ni la taille finale m, de la manière suivante :

$$\boxed{\left\langle e^{-\mu N}\right\rangle_{\rm qs} = \lim_{t,t'\to+\infty}\left\langle e^{-\mu N_t}\middle| N_{t+t'} = m\right\rangle = e^{-\mu}\mathcal{A}'(e^{-\mu}Q_e^*)}. \tag{4.12}$$

4.2.2 En temps continu

L'équation (4.12) montre, pour le processus de Galton-Watson, que les propriétés de l'état quasi-stationnaire sont entièrement contenues dans la fonction $\mathcal{A}$. La définition dynamique (4.8) conduit à l'équation (4.10) qui doit permettre de déterminer $\mathcal{A}$ entièrement à partir de F_e. Dans le cas d'un temps discret, cette détermination n'est pas aisée (sauf cas particuliers). Dans le cas d'un temps continu, la démarche devient plus facile puisque (4.10) devient une équation différentielle.

En temps continu, les équations d'évolution de la probabilité d'extinction $Q_e(t)$ et des fonctions génératrices G_1 et G_2 sont des équations différentielles. Aux temps longs, la relaxation vers le point fixe Q_e^* est encore exponentielle et toute la démarche précédente reste valide. Seule l'équation (4.10) satisfaite par $\mathcal{A}$ change et devient une équation différentielle. Le processus de Galton-Watson en temps continu est obtenu (cf. section

1.2.1) en considérant des taux de division infinitésimaux $p_k = \beta_k dt$ pour $k \neq 1$ et $p_1 = 1 - \sum_{k\neq 1} \beta_k dt$. On a alors l'équation suivante pour $\mathcal{A}$ [SD08] :

$$\lambda \mathcal{A}(Q) = \mathcal{A}'(Q) f_e(Q) \tag{4.13}$$

avec les définitions suivantes :

$$f_e(Q) \hat{=} \sum_{k\neq 1} \beta_k (Q^k - Q), \quad \lambda \hat{=} f_e'(Q_e^*). \tag{4.14}$$

L'intégration (4.13) est aisée et conduit, pour un processus de Galton-Watson, à la forme explicite suivante :

$$\mathcal{A}(Q) = (Q - Q_e^*) \exp\left[\int_{Q_e^*}^{Q} \left(\frac{\lambda}{f_e(y)} - \frac{1}{y - Q_e^*}\right) dy\right]. \tag{4.15}$$

On en déduit alors par (4.12) la fonction génératrice dans le régime quasi-stationnaire en temps continu suivante :

$$\left\langle e^{-\mu N} \right\rangle_{\text{qs}} = \lambda Q_e^* e^{-\mu} \frac{e^{-\mu} - 1}{f_e(Q_e^* e^{-\mu})} \exp\left[\int_{Q_e^*}^{Q_e^* e^{-\mu}} \left(\frac{\lambda}{f_e(y)} - \frac{1}{y - Q_e^*}\right) dy\right]. \tag{4.16}$$

Cela permet ainsi de calculer toute la distribution de la taille N dans le régime quasi-stationnaire.

4.2.3 Universalité dans le modèle de Galton-Watson

La formule (4.16) permet d'étudier le voisinage du point critique dans le cas continu. Le point critique est obtenu lorsque la probabilité d'extinction aux temps longs Q_e^* atteint la valeur 1 (phase inactive). Cela correspond à un taux de croissance $\beta = \sum_k \beta_k (k-1)$ nul. La probabilité d'extinction Q_e^* satisfait :

$$f_e(Q_e^*) = 0 \tag{4.17}$$

Ainsi, pour connaître Q_e^* pour β proche de zéro, il convient de développer f_e autour de 1 en allant jusqu'au deuxième terme dominant. On voit ainsi apparaître deux cas distincts selon les taux de division β_k :

1. soit la variance $\sum_k \beta_k k^2$ est finie (les β_k décroissent plus vite que $1/k^2$) et $f_e(Q)$ se développe à l'ordre 2 :

$$f_e(Q) = \beta(Q-1) + B(Q-1)^2 + o\left((Q-1)^2\right); \tag{4.18}$$

   On obtient ainsi $Q_e^* \simeq 1 - \beta/B$ au voisinage de $\beta = 0$.

2. soit la variance $\sum_k \beta_k k^2$ est infinie (les β_k décroissent algébriquement en $k^{-1-\eta}$ avec $1 < \eta < 2$) et $f_e(Q)$ a le développement singulier suivant :

$$f_e(Q) = \beta(Q-1) + C(1-Q)^\eta + o\left((1-Q)^\eta\right). \tag{4.19}$$

   Cela conduit au comportement critique de la probabilité d'extinction $Q_e^* \simeq 1 - \beta^{1/(\eta-1)}/C$.

Ces deux types de comportement conduisent à des classes d'universalité différentes au voisinage de la vitesse critique (cf. première annexe de [SD08] pour le détail des calculs). En dérivant (4.16) par rapport à μ en $\mu = 0$, on obtient, dans le cas de la variance finie, que le nombre moyen diverge comme :

$$\langle N \rangle_{\text{qs}} \simeq -\frac{1}{|\beta|} \left(\sum_k \beta_k k(k-1) \right) \qquad \text{(variance finie)} \tag{4.20}$$

et la fonction génératrice de la taille normalisée $x = N/\langle N \rangle_{\text{qs}}$ devient universelle lorsque $\beta \to 0$:

$$\langle e^{-\mu x} \rangle_{\text{qs}} \simeq \frac{1}{(1+\mu/2)^2} \qquad \text{(variance finie).} \tag{4.21}$$

Cette fonction génératrice est la même quel que soit le signe de β et correspond à une distribution de x en $4xe^{-2x}$. On retrouve ainsi les résultats déjà présents dans [Kha73] pour le cas critique.

Dans le cas de la variance $\sum_k \beta_k k^2$ infinie [Sla68], le taille moyenne n'est plus définie (distribution de la taille à décroissance lente) mais la taille typique diverge au point critique comme :

$$N_{\text{typique}} \propto |\beta|^{-\frac{1}{\eta-1}} \qquad \text{(variance infinie)} \tag{4.22}$$

et la fonction génératrice de la taille normalisée $x = N/N_{\text{typique}}$ tend vers deux formes universelles différentes selon que l'on se place en-deça ou au-delà du point critique :

$$\langle e^{-\mu x} \rangle_{\text{qs}} \underset{\beta<0}{=} \left(\frac{1}{1+\mu^{\eta-1}} \right)^{\frac{\eta}{\eta-1}}, \tag{4.23a}$$

$$\langle e^{-\mu x} \rangle_{\text{qs}} \underset{\beta>0}{=} \left(\frac{1}{1+\mu} \right)^{\eta} \qquad \text{(variance infinie).} \tag{4.23b}$$

Ces différences de comportements sont à mettre en regard, d'une part avec le cas spatial du chapitre 6, et d'autre part avec les différences observées dans les temps de coalescence selon que l'on a une variance finie ou non dans un modèle de type Wright-Fisher (cf. section 8.2.3).

4.3 Généralisation de l'approche par les fonctions génératrices à d'autres modèles

4.3.1 Conditionnement sur la date d'extinction dans le processus de Galton-Watson

La formule (4.12) peut également être obtenue pour un conditionnement sur la date d'extinction finale, plutôt que sur une taille finale égale à 1. Supposons en effet que le dernier individu meure à $T = t + t'$ et que l'on veuille connaître la taille au temps t. La probabilité que la lignée d'un individu meure exactement à t' est $\partial_{t'} Q_e(t')$. Donc la probabilité que les N_t individus à t s'éteignent à $T = t + t'$ est donnée par :

$$\text{Prob}\left(\text{extinction à } t+t' \text{ à } dt' \text{ près} | N_t\right) = N_t Q_e(t')^{N_t - 1} \partial_{t'} Q_e(t') dt'. \tag{4.24}$$

La formule des probabilités conditionnelles (4.1) donne alors :

$$\text{Prob}\left(N_t\middle|\text{extinction à } t+t' \text{ à } dt' \text{ près}\right) = \text{Prob}\left(N_t\right)\frac{N_t Q_e(t')^{N_t-1}\partial_{t'}Q_e(t')}{\partial_{t'}Q_e(t+t')}. \tag{4.25}$$

En sommant sur N_t l'expression précédente multipliée par $e^{-\mu t}$, on obtient la fonction génératrice suivante :

$$\begin{aligned}\left\langle e^{-\mu N_t}\middle|\text{extinction à } t+t' \text{ à } dt' \text{ près}\right\rangle &= \frac{\partial_{t'}\langle (e^{-\mu}Q_e(t'))^{N_t}\rangle}{\partial_{t'}Q_e(t+t')} \\ &= \frac{1}{\partial_{t'}Q_e(t+t')}\partial_{t'}G_1(\mu+\ln Q_e(t'),t).\end{aligned}$$

Le comportement à t grand du membre de droite est à nouveau donné par (4.8) puis en prenant t' grand, on retrouve l'expression (4.12) :

$$\left\langle e^{-\mu N_t}\middle|\text{extinction à } t+t' \text{ à } dt' \text{ près}\right\rangle \underset{t,t'\to+\infty}{\simeq} e^{-\mu}\mathcal{A}'(e^{-\mu}Q_e^*) \tag{4.26}$$

Cela confirme que le régime quasi-stationnaire est commun à différents états finaux, en particulier à tous ceux qui ne sont ni l'état absorbant, ni des tailles macroscopiques.

4.3.2 Marche aléatoire avec branchement

Nous avons étudié au chapitre 3 la marche aléatoire avec branchements en présence d'un mur absorbant se déplaçant à une vitesse v. L'évolution de la probabilité de survie de cette marche aléatoire est donnée par l'équation F-KPP (3.11). L'analyse de cette dynamique a montré que, pour $v < v_c$, $Q_s(x,t)$ tend vers $Q_s^*(x)$ avec un temps de relaxation τ_1 donné par (3.40). Dans ce cas-là, toute la démarche précédente, dans laquelle nous avons introduit les fonctions génératrices à un et deux temps ainsi que l'amplitude $\mathcal{A}$ associée au temps de relaxation le plus lent, reste valable (cf. [SD08]).

Cependant, puisque $Q_e^*(x) = 1 - Q_s^*(x)$ n'est plus un nombre mais une fonction, l'amplitude $\mathcal{A}$ n'est plus une fonction mais une fonctionnelle. Comme précédemment, on peut la relier à la fonctionnelle d'évolution de $Q_e(x,t)$. Formellement, l'équation F-KPP (3.8) peut être récrite sous la forme :

$$\partial_t Q_e(x,t) = \mathcal{F}(Q_e(x,t)) \tag{4.27a}$$

$$\mathcal{F}(Q(x)) \hat{=} \partial_x^2 Q(x) - v\partial_v Q(x) + f_e(Q(x)). \tag{4.27b}$$

où $\mathcal{F}(Q)$ est une fonctionnelle. Une adaptation directe du calcul réalisé pour le processus de Galton-Watson en temps continu (cf. notre travail [SD08]) donne pour $\mathcal{A}$:

$$\boxed{\lambda_1\mathcal{A}(Q) = \frac{d}{ds}\mathcal{A}(Q+s\mathcal{F}(Q))\Big|_{s=0}} \tag{4.28}$$

La probabilité d'extinction Q_e étant à présent une fonction, la dérivée apparaissant dans (4.13) devient une dérivée directionnelle dans (4.28). De plus, la population n'est plus

caractérisée par sa seule taille mais par la position de tous ses individus. Les fonctions génératrices à considérer sont alors du type :

$$G_1(x, g_1; t) = \left\langle \exp\left(-\sum_{i=1}^{N_t} g_1(x_i^{(t)}) \right) \right\rangle \tag{4.29a}$$

$$G_2(x, g_1, g_2; t, t') = \left\langle \exp\left(-\sum_{i=1}^{N_t} g_1(x_i^{(t)}) - \sum_{i=1}^{N_{t+t'}} g_2(x_i^{(t+t')}) \right) \right\rangle \tag{4.29b}$$

où les $x_i^{(t)}$ désignent les positions des N_t individus à l'instant t et g_1 et g_2 sont des fonctions positives. Le cas $g_1(x) = \mu$ constant donne la fonction génératrice de la taille de la population alors que

$$g_1(x) = \begin{cases} \mu & \text{si } x > X \\ 0 & \text{si } x < X \end{cases}$$

permet de compter uniquement les particules à une distance supérieure à X du mur.

Comme nous l'avons expliqué dans [SD08] en prolongeant directement l'approche suivie pour le processus de Galton-Watson, la généralisation de (4.12) est possible et les fonctions génératrices dans le régime quasi-stationnaire de marche aléatoire avec branchements sont données formellement par :

$$\boxed{\left\langle \exp\left(-\sum_{i=1}^{N} g(x_i) \right) \right\rangle_{\text{qs}} = \frac{d}{ds} \mathcal{A}\left((Q_e^* + s\phi_1) e^{-g} \right) \Big|_{s=0}} \tag{4.30}$$

où $\phi_1(x)$ est le vecteur propre associé au temps de relaxation le plus lent τ_1 (cf. équations (3.41, 3.42)) et les dérivées sont prises dans la direction ϕ_1.

A priori, les deux équations (4.28, 4.30) suffisent à décrire les propriétés du régime quasi-stationnaire à partir de celle de la fonctionnelle $\mathcal{F}$ définie en (4.27). Par exemple, un développement de Taylor autour de Q_s^* des fonctionnelles $\mathcal{A}$ et $\mathcal{F}$ tel que

$$\begin{aligned}
\mathcal{A}(Q_e^* + \epsilon\phi) &= \mathcal{A}(Q_e^*) + \epsilon \mathcal{A}^{(1)}[\phi] + \frac{\epsilon^2}{2!} \mathcal{A}^{(2)}[\phi, \phi] + \dots \\
\mathcal{F}(Q_e^* + \epsilon\phi) &= \mathcal{F}(Q_e^*) + \epsilon \mathcal{L}[\phi] + \frac{\epsilon^2}{2!} \mathcal{F}^{(2)}[\phi, \phi] + \dots \\
\mathcal{F}^{(n)}[\psi_1, \dots, \psi_n] &= \frac{d^n}{ds_1 \dots ds_n} \mathcal{F}(Q_e^* + s_1\psi_1 + \dots + s_n\psi_n) \Big|_{s_1 = \dots = s_n = 0}
\end{aligned}$$

suffit au calcul des moments de $\sum_{i=1}^{N} g(x_i)$ puisque celui de $\mathcal{A}$ peut être obtenu récursivement à partir de celui de $\mathcal{F}$ à travers l'équation (4.28). En particulier, nous retiendrons l'expression suivante de la valeur moyenne de la taille :

$$\boxed{\langle N \rangle_{\text{qs}} = -\frac{d^2}{ds d\mu} \mathcal{A}(Q_e^* e^{-\mu} + s\phi_1 e^{-\mu}) \Big|_{s=\mu=0} = 1 + \mathcal{A}^{(2)}[\phi_1, Q_e^*]}. \tag{4.31}$$

Cependant, des obstacles mathématiques rendent la poursuite de cette démarche difficile : le nombre infini de degrés de libertés et de vecteurs propres dans le cas fonctionnel rendent les décompositions sur les vecteurs propres et les sommations infinies difficiles à contrôler. C'est pourquoi le chapitre suivant propose une alternative à cette approche en remplaçant ces sommations infinies par l'étude de solutions d'équations différentielles de type F-KPP.

Néanmoins, une meilleure compréhension de la fonctionnelle $\mathcal{A}$ permettrait d'étudier d'éventuelles propriétés d'universalité des marches aléatoires avec branchements au voisinage du point critique, par une démarche similaire à celle de la section 4.2.3. De plus, la définition de $\mathcal{A}$ n'est pas sans rappeler celle des champs d'échelle non-linéaires introduites dans la théorie de la renormalisation (cf. les *non-linear scaling fields* décrits dans [Weg76]) qui satisfont des équations semblables à (4.28) et décrivent eux aussi des comportements au voisinage de points fixes du flot de renormalisation.

4.4 Résultats complémentaires utiles

4.4.1 Calcul numérique des fonctions génératrices

Au-delà de l'intérêt analytique, les résultats précédents peuvent être utilisés numériquement pour étudier des régimes conditionnés. En effet, les fonctions génératrices satisfont la même équation d'évolution que la probabilité d'extinction et il est possible d'utiliser numériquement cette équation pour mesurer les propriétés statistiques d'une population conditionnée sur sa taille finale. Des techniques de simulations existent déjà pour étudier un système conditionné à ne pas être encore dans l'état absorbant [Dic02] et nous fournissons dans cette section et le chapitre suivant une méthode pour étudier un conditionnement sur une date *ultérieure* à la date d'observation.

De manière générale, dès que les individus sont indépendants, Q_e, G_1 et G_2 évoluent selon

$$\partial_t Q_e = \mathcal{F}(Q_e) \tag{4.32}$$

où $\mathcal{F}$ est une fonction(nelle) non-linéaire.

Les définitions (4.4, 4.5, 4.6) peuvent être directement prolongées aux fonctions génératrices (4.29) en substituant g_1 à μ et en prenant $g_2 = \nu$. L'équation (4.32) induit alors des équations différentielles couplées pour les fonctions R_m et P_m qu'il est facile de programmer. On obtient alors n'importe quelle fonction génératrice conditionnée sur la taille finale en calculant le quotient (4.6).

Pour illustrer le propos, considérons la taille de la population et prenons $(g_1, g_2) = (\mu, \nu)$. Alors il est possible de décomposer $G_2(x, \mu, \nu, t, t')$ aux premiers ordres de la manière suivante :

$$G_2(x, \mu, \nu, t, t') = H_{00} + \mu H_{01} + e^{-\nu} H_{10} + \mu e^{-\nu} H_{11} + \dots \tag{4.33}$$

La dynamique des H_{ij} est obtenue en injectant cette décomposition dans (4.32). La taille moyenne à l'instant t conditionnée par une taille unité à l'instant $t + t'$ est ainsi donnée

par :

$$\langle N_t \rangle_{1,t+t'} = -\frac{H_{11}(x,t,t')}{H_{10}(x,t,t')} \tag{4.34}$$

Pour le second moment de N_t, il suffit de considérer H_{12} au lieu de H_{11} dans le numérateur. Pour un conditionnement par une taille finale égale à m, il suffit de remplacer H_{11} et H_{10} par H_{m1} et H_{m0}.

Nous avons utilisé cette méthode numérique dans [DS07] pour explorer les propriétés de la marche aléatoire avec branchements en présence d'un mur absorbant. Les résultats ont été présentés dans les figures 4.1 et 4.3 et dans l'article [DS07].

Les premières observations sont les suivantes :

- pour $v < v_c$, un plateau apparaît dans les trois premiers moments de la taille dès que t et t' sont grands devant le temps de relaxation le plus lent τ_1 du système : il correspond à l'apparition d'un régime quasi-stationnaire.
- pour $v > v_c$, la taille moyenne $\langle N_t \rangle_{1 \text{ à } T}$ à l'instant t conditionnée par une taille $N_T = 1$ à $T > t$ prend la forme d'une fonction de grande déviation aux temps longs :

$$\ln \langle N_t \rangle_{1 \text{ à } T} \simeq Th\left(\frac{t}{T}\right) \tag{4.35}$$

  Aucun plateau quasi-stationnaire ne peut donc apparaître de ce côté de la vitesse critique (cf. la discussion au chapitre 6).

Cette méthode numérique permet de calculer des valeurs moyennes dans le régime quasi-stationnaire[2] mais ne permet pas d'échantillonner la mesure quasi-stationnaire, ni d'étudier les relations de parenté entre individus. Le chapitre suivant s'attache à combler cette lacune.

4.4.2 Condition suffisante d'existence d'un régime quasi-stationnaire ; lien avec d'autres types de conditionnement

Dans cette section, nous énonçons une condition suffisante pour observer un régime quasi-stationnaire et donnons l'expression de la probabilité quasi-stationnaire d'une configuration formellement en fonction des vecteurs propres de la matrice associée au processus de Markov considéré. Par ailleurs, cela donne l'occasion de faire le lien entre notre définition de *processus quasi-stationnaire*, celle utilisée par les mathématiciens et la notion de Q-processus.

Considérons un système dont la configuration $\mathcal{C}_t$ évolue dans le temps selon un processus markovien dont les probabilités de transition sont données par une matrice $W(\mathcal{C} \to \mathcal{C}')$. Appelons $\mathcal{C}_\emptyset$ la configuration absorbante. Si nous désignons par $P_t(\mathcal{C})$ la probabilité pour le système d'être dans une configuration $\mathcal{C}$ à l'instant t, cette probabilité satisfait

[2] Une généralisation assez simple permet aussi de mesurer des corrélations temporelles en introduisant la fonction génératrice à trois temps.

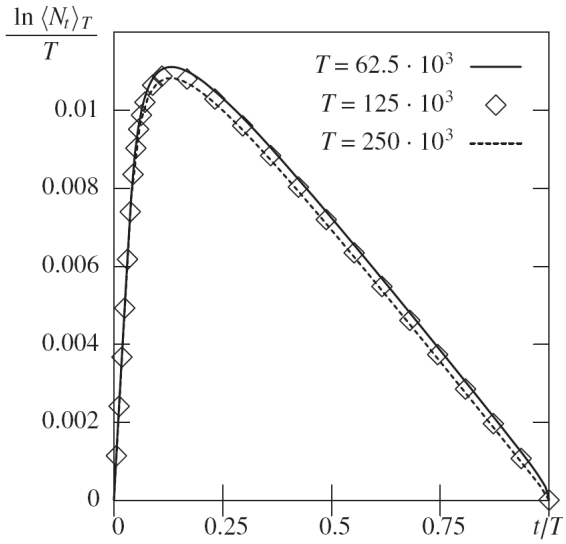


FIG. 4.3: Absence de régime quasi-stationnaire au-delà de la vitesse critique. En ordonnée : la taille moyenne à t conditionnée par une taille finale unité à $T = t + t'$; en abscisse, le temps mis à l'échelle t/T. Le calcul numérique a été effectué pour le modèle discret décrit au chapitre 3 en section 3.3 pour un écart à la vitesse critique $v - v_c = 10^{-2} > 0$ pour différents temps finaux $T = 62500$, 125000 et 250000.

l'équation différentielle :

$$\partial_t P_t(\mathcal{C}) = \sum_{\mathcal{C}' \neq \mathcal{C}} W(\mathcal{C}' \to \mathcal{C}) P_t(\mathcal{C}') - \left(\sum_{\mathcal{C}' \neq \mathcal{C}} W(\mathcal{C} \to \mathcal{C}') \right) P_t(\mathcal{C}) \tag{4.36}$$

L'existence d'un état absorbant[3] implique que la distribution stationnaire est donnée par $P_{\rm st}(\mathcal{C}) = \delta(\mathcal{C}, \mathcal{C}_\emptyset)$. La relaxation vers cet état absorbant est dominée par les valeurs propres dont les parties réelles sont les plus petites en valeur absolue[4].

La première valeur propre est donc $\lambda_0 = 0$ et correspond à la distribution stationnaire $P_{\rm st}(\mathcal{C}) = \delta(\mathcal{C}, \mathcal{C}_\emptyset)$. Le vecteur propre à gauche qui correspond à cette valeur propre est tout simplement le vecteur de composante 1 sur toutes les configurations (conservation de la normalisation des probabilités). Appelons λ_1 la valeur propre suivante, classée par partie réelle décroissante et désignons respectivement par $a_1(\mathcal{C})$ et $b_1(\mathcal{C})$ les vecteurs propres à droite et à gauche de la matrice $W(\mathcal{C} \to \mathcal{C}')$ pour la valeur propre λ_1 :

$$\begin{aligned} \lambda_1 a_1(\mathcal{C}) &= \sum_{\mathcal{C}'} W(\mathcal{C}' \to \mathcal{C}) a_1(\mathcal{C}') \\ \lambda_1 b_1(\mathcal{C}) &= \sum_{\mathcal{C}'} b_1(\mathcal{C}') W(\mathcal{C} \to \mathcal{C}') \end{aligned}$$

[3]Nous supposerons que le processus est irréductible et que toute configuration initiale mène à l'état absorbant.

[4]En temps discret, les valeurs propres doivent être ordonnées selon leur module.

où nous avons pris par convention $W(\mathcal{C} \to \mathcal{C}) = -\sum_{\mathcal{C}' \neq \mathcal{C}} W(\mathcal{C} \to \mathcal{C}')$. Nous supposerons donc tout d'abord qu'il y a une unique valeur propre de partie réelle $\mathrm{Re}(\lambda_1)$.

Aux temps longs, nous avons ainsi :

$$P_t(\mathcal{C}) = P_{\mathrm{st}}(\mathcal{C}) + Ae^{\lambda_1 t} a_1(\mathcal{C}) + O(e^{\lambda_2 t}) \tag{4.37}$$

où A est une constante qui dépend de la projection de l'état initial P_0 sur le vecteur propre à gauche $b_1(\mathcal{C})$. Considérons alors la probabilité conditionnelle $P(\mathcal{C}, t|\mathcal{C} \neq \mathcal{C}_\emptyset)$ d'être dans la configuration $\mathcal{C}$ sachant que le système n'est pas encore tombé dans l'état absorbant. Nous avons ainsi :

$$P(\mathcal{C}, t|\mathcal{C} \neq \mathcal{C}_\emptyset) \underset{t\to\infty}{\simeq} \frac{a_1(\mathcal{C})}{\sum_{\mathcal{C}' \neq \mathcal{C}_\emptyset} a_1(\mathcal{C}')} \tag{4.38}$$

La partie de a_1 orthogonale à $P_{\mathrm{st}}(\mathcal{C})$, *i.e.* tous les coefficients $a_1(\mathcal{C})$ avec $\mathcal{C} \neq \mathcal{C}_\emptyset$, sont alors positifs[5] et $\lim_{t\to\infty} P(\mathcal{C}, t|\mathcal{C} \neq \mathcal{C}_\emptyset)$ définit bien une mesure de probabilité qui est un cas particulier de ce que les mathématiciens appellent distribution *quasi-stationnaire*. La raison du nom *quasi-stationnaire* dans ce cas est essentiellement liée au fait que cette distribution est reliée à l'un des vecteurs propres suivants de la matrice de transition, alors que le premier vecteur propre correspond à la distribution stationnaire.

Dans la littérature mathématique, le terme $\lim_{t\to\infty} P(\mathcal{C}, t|\mathcal{C} \neq \mathcal{C}_\emptyset)$ dans (4.38) s'appelle la limite de Yaglom [Yag47]. Dans un cadre tout-à-fait général, il n'est pas évident que cette limite, définie uniquement en termes de probabilité conditionnelle, coïncide avec la distribution obtenue en considérant le premier vecteur propre a_1 (définition de pure algèbre linéaire) ; parmi les conditions nécessaires, il faut aussi que la constante A dans (4.37), qui dépend de la condition initiale, soit aussi non nulle. Dans le cas contraire, il faudrait considérer le vecteur propre suivant $a_2(\mathcal{C})$, si tant est que les coefficients $a_2(\mathcal{C})$ soient positifs dans les configurations accessibles.

S'il est facile de montrer le type d'égalité (4.38) entre limite de Yaglom et premier vecteur propre différent de la probabilité stationnaire dans le cas de systèmes à nombre fini de configurations tels que $\mathrm{Re}(\lambda_1)$ est isolée, cela est beaucoup plus difficile dans les systèmes à nombre infini d'états [SV66, DV03, CMS95, SE05, SE04, CCL+07, FMP91, FMP92]. Le cas du processus de Galton-Watson est un cas particulier assez simple des systèmes à nombre d'états dénombrables [Yag47, Kha73, Har62] et l'étude présentée en section 4.2 utilise des raisonnements proches de ceux menés ci-dessus pour des systèmes à nombre fini d'états et la limite (4.38) reste bien définie.

Intéressons-nous à présent au cas où $\mathrm{Re}(\lambda_1)$ n'est pas isolée. On pourra considérer par exemple le système à trois états de matrice de transition *en temps discret*

$$W = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{pmatrix} \tag{4.39}$$

[5] Lorsque le nombre de configurations est fini, cela est vrai d'après le théorème de Perron-Frobenius dans la restriction dès que la restriction de $W(\mathcal{C} \to \mathcal{C}')$ au sous-espace supplémentaire de $\delta(\mathcal{C}, \mathcal{C}_\emptyset)$ est primitive (*i.e.* dès lors qu'il existe un temps t au bout duquel n'importe quelle configuration peut être atteinte à partir de n'importe quelle autre configuration de départ. Lorsque la matrice de transition n'est pas primitive, alors le terme sous-dominant dans (4.37) dépend de la condition initiale et peut présenter des cycles.

dont les valeurs propres sont données par 1, 1/2 et $-1/2$. L'état 1 est l'état absorbant ; si le système part à $t = 0$ dans l'état 2, alors aux temps pairs, le système est soit dans 1, soit dans 2, alors qu'aux temps impairs il est soit dans 1, soit dans 3. Nous avons ici une période $T = 2$ et il existe plusieurs mesures conditionnées $P(\mathcal{C}, t|\mathcal{C} \neq \mathcal{C}_\emptyset)$ formant un cycle selon le *déphasage* de t par rapport à la période. Pour une période T générale, la limite de Yaglom n'est plus définie aussi simplement : il faut remplacer la limite $t \to \infty$ par $t = nT + \phi$ où $n \to \infty$ avec ϕ fixé dans $\{0, 1, \ldots, T-1\}$ pour obtenir les différentes mesures limites.

Considérons à présent la probabilité conditionnelle $P(\mathcal{C}, t|\mathcal{C}_{t+t'} \neq \mathcal{C}_\emptyset)$ où $t' > 0$, qui et donc la probabilité d'observer la configuration $\mathcal{C}$ au temps t sachant que le système ne sera pas encore tombé dans l'état absorbant au temps $t + t'$. La formule des probabilités conditionnelles (4.1) donne alors :

$$P(\mathcal{C}, t|\mathcal{C}_{t+t'} \neq \mathcal{C}_\emptyset) = \frac{P(\mathcal{C}_{t+t'} \neq \mathcal{C}_\emptyset|\mathcal{C}, t)P_t(\mathcal{C})}{\text{Prob}\left(\mathcal{C}_{t+t'} \neq \mathcal{C}_\emptyset\right)}$$

Pour t' grand, la probabilité $P(\mathcal{C}_{t+t'} \neq \mathcal{C}_\emptyset|\mathcal{C}, t)$ est dominée, d'après (4.37), par le comportement de la valeur propre λ_1 (et non λ_0 puisque l'on évite la configuration absorbante). Dans le cas présent, la constante A de (4.37) est donnée par la projection de la condition initiale à t sur le vecteur propre à gauche et on a $A = b_1(\mathcal{C})$. De la même manière, les autres probabilités $P_t(\mathcal{C})$ et $\text{Prob}\left(\mathcal{C}_{t+t'} \neq \mathcal{C}_\emptyset\right)$ peuvent être obtenues de manière similaire à (4.37). Dans la limite $t, t' \to \infty$, nous obtenons ainsi :

$$P(\mathcal{C}, t|\mathcal{C}_{t+t'} \neq \mathcal{C}_\emptyset) \underset{t,t'\to\infty}{\simeq} \frac{b_1(\mathcal{C})a_1(\mathcal{C})}{\sum_{\mathcal{C}' \neq \mathcal{C}} b_1(\mathcal{C}')a_1(\mathcal{C}')} \tag{4.40}$$

C'est cette limite que nous appelons dans nos travaux *distribution quasi-stationnaire* et les mathématiciens Q-processus : cette distribution décrit les trajectoires du système dans l'espace des configurations qui ne tomberont jamais dans l'état absorbant. Comme précédemment, la formule précédente est valable dès lors que la valeur propre λ_1 est isolée des suivantes et qu'elle est la seule à avoir la partie réelle $\text{Re}(\lambda_1)$.

Il est intéressant de noter que, dans les deux types de conditionnement (4.38) et (4.40), les distributions ne dépendent que des propriétés de la valeur propre non-triviale et de ses vecteurs propres associés. Alors que la limite de Yaglom ne fait intervenir que le vecteur propre à droite, la distribution quasi-stationnaire au sens où nous l'entendons fait aussi apparaître le vecteur propre à gauche. Cela se comprend aisément lorsque l'on considère le comportement asymptotique de la probabilité $Q_s(\mathcal{C}_0, t)$ que le système dans la configuration $\mathcal{C}_0$ ne soit pas encore entré dans l'état absorbant au temps t. Pour cela il suffit d'écrire explicitement la constante A dans (4.37) et nous obtenons :

$$Q_s(\mathcal{C}_0, t) \propto b_1(\mathcal{C}_0)e^{\lambda_1 t}$$

Les deux facteurs dans (4.40) correspondent ainsi d'une part au fait d'avoir survécu jusqu'au temps $t \gg \tau_1$ et d'autre part de devoir encore survivre jusqu'au temps $t + t'$ avec $t' \gg \tau_1$. Ce produit de deux contributions réapparaîtra naturellement au chapitre suivant dans l'expression (5.20) des densités quasi-stationnaires.

Enfin, le lien ci-dessus entre vecteurs propres, limite de Yaglom et conditionnement à un temps ultérieur laisse présager des liens entre les différentes distributions conditionnées dès lors qu'existent des relations entre les vecteurs propres à gauche et à droite. Le processus de Galton-Watson en est un exemple : la limite de Yaglom dans le domaine $\beta < 0$ (extinction certaine) correspond à une distribution exponentielle e^{-x} de la taille de la population normalisée, de fonction génératrice $1/(1+s)$, alors que la distribution quasi-stationnaire (4.21) donne une distribution en xe^{-x}, de fonction génératrice $1/(1+s)^2$ qui n'est autre que la dérivée de la première fonction génératrice.

L'approche précédente reste valable dès lors qu'existe une première valeur propre non triviale λ_1 de partie réelle isolée avec des vecteurs propres $a_1(\mathcal{C})$ et $b_1(\mathcal{C})$ dont les composantes sont strictement positives[6] dès que $\mathcal{C}$ n'est pas la configuration absorbante. Dans les chapitres suivants, nous travaillerons avec des marches aléatoires dans le continu et, puisque l'étude complète des vecteurs propres dans ce cas soulève des difficultés, nous ne chercherons pas à dériver les résultats à partir de la formule générale (4.40).

[6]Si une composante de $b_1(\mathcal{C})$ s'annule, cela signifie que la configuration initiale est incompatible avec la distribution. Nous supposerons que ce n'est pas le cas ici.

Chapitre 5

Un processus biaisé équivalent au conditionnement

Ce chapitre décrit la construction d'un processus biaisé pour étudier une marche aléatoire avec branchements conditionnée par une taille finale égale à 1. Il débute par une revue des résultats mathématiques préexistants dont nous nous sommes inspirés. Nous présentons ensuite la manière dont nous avons utilisé le formalisme des fonctions génératrices développé au chapitre précédent pour construire le processus biaisé [SD08] dans un cadre général et aller ainsi au-delà du simple calcul numérique de valeurs moyennes. En particulier, cela nous permet de calculer les densités moyennes d'individus dans le régime quasi-stationnaire.

5.1 Introduction

5.1.1 L'approche de *l'épine dorsale* : revue

Dans le cas particulier d'un modèle de populations (*i.e.* sans recombinaison $A + A \to A$), des outils supplémentaires par rapport à ceux du chapitre précédent sont disponibles pour étudier les régimes conditionnés. L'un d'eux, que nous utiliserons abondamment dans le chapitre suivant, est l'approche de *l'épine dorsale* (« *spine* » dans la littérature anglophone).

Supposons par exemple que le système parte d'une configuration initiale à un unique individu et finisse à un temps T avec un seul individu aussi. Parmi toutes les lignées générées pendant cet intervalle, l'une d'elles est privilégiée par rapport aux autres : celle qui relie le survivant à T à l'individu initial. En effet, celle-ci aura généré toutes les autres particules existantes, d'où le nom d'*épine dorsale* : toutes les autres lignées ne sont que des branchements de celle-ci au cours du temps (voir figure 5.1).

Ainsi, dans le régime conditionné, il devient préférable de séparer la population de

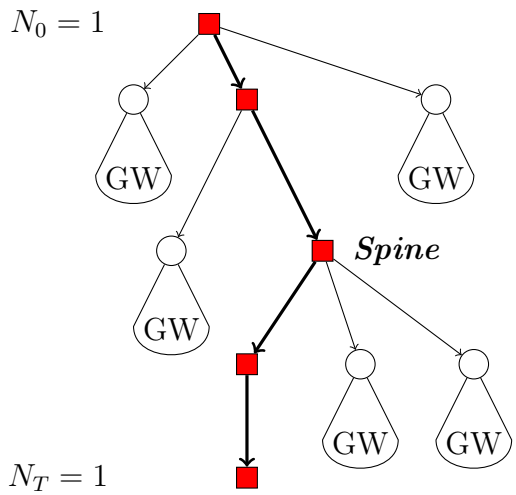


FIG. 5.1: Approche de l'*épine dorsale*. Si la population commence et finit avec une taille $N_0 = N_T = 1$, alors il existe une lignée particulière : celle qui aura survécue tout au long du processus, appelée « épine dorsale » ou « *spine* » dans la littérature anglophone. L'approche de l'épine dorsale consiste alors à séparer l'étude de la dynamique de la population entre 0 et T en deux : une dynamique particulière pour l'« épine dorsale » (carrés) qui condamne celle-ci à survivre et une autre dynamique particulière pour les autres individus (ronds et sous-arbres de Galton-Watson) produits par l'« épine dorsale » qui sont condamnés à s'éteindre avant le temps final T.

taille N_t en une particule spéciale (l'*épine dorsale*) destinée à survivre *durant tout l'intervalle* $[0, T]$ et $N_t - 1$ particules condamnées à disparaître *à la fin*.

La construction mathématique rigoureuse sous-jacente est présente dans la littérature pour la démonstration de propriétés sur les marches aléatoires avec branchements (l'un des articles fondateurs est [LPP95] et d'autres utilisations courantes sont présentées dans [HH04, HH07, Kyp04, CR88]). En particulier, un certain nombre de résultats du chapitre précédent (existence et unicité des solutions pour $Q_e^*(x)$ selon que $v < v_c$ ou $v > v_c$, comportement aux temps longs de $Q_e(x,t)$ pour $v > v_c$) peuvent être prouvés mathématiquement grâce à cette approche [HH07, HHK06].

L'idée principale à retenir est le *many-to-one lemma* [HH04] qui consiste à ramener le calcul d'une quantité $\sum_{i=1}^{N_t} g(x_i(t))$ où les $x_i(t)$ sont les positions de tous les individus à un instant t à une quantité ne dépendant que de la position de l'*épine dorsale* à un facteur près qui résume l'expansion et/ou la disparition des autres particules. Ce facteur correspond alors à un poids effectif sur la distribution de la position l'*épine dorsale* prenant en compte l'évolution de tous les autres individus. L'expression mathématique rigoureuse de ce lemme en termes de filtrations ne sera pas détaillée ici (cf. [HH04] pour un exposé rigoureux) mais guide toute la démarche de ce chapitre.

5.1.2 Validité de la construction

Le but de ce chapitre est de construire un processus stochastique modifié qui décrit l'évolution d'une marche aléatoire avec branchements dans un domaine à bords absorbants conditionnée par l'existence d'un unique survivant au temps final T. Formulé autrement, le processus modifié permet de produire directement (sans test *a posteriori* sur la configuration finale[1]) des configurations avec les bons poids statistiques.

Cette construction, inspirée de l'approche de l'*épine dorsale*, consiste à séparer l'évolution de la lignée qui va survivre du reste de la population. Elle est l'extension directe de notre travail [SD08]. La seule nouveauté est la généralité de la construction puisque celle-ci reste valide :

1. quel que soit le domaine $\mathcal{D}$ où évoluent les individus (discret ou continu : la construction fonctionne aussi sur graphe) ;
2. quels que soient les taux de branchements $\beta_k(\vec{x})$ et la dérive $\vec{v}(\vec{x})$ des marches aléatoires qui peuvent éventuellement dépendre de la position $\vec{x}$;
3. qu'un régime quasi-stationnaire existe ou non (si le domaine $\mathcal{D}$ n'est pas borné par exemple).

La contre-partie est qu'elle nécessite, comme cela sera montré ci-dessous, la connaissance de deux quantités dépendant de la position et du temps :

1. la probabilité $Q_e(\vec{x}, t)$ que la lignée d'un individu initialement situé à la position $\vec{x}$ soit encore vivante au temps t ; elle est donnée par une équation de type KPP (2.2) qu'il faut intégrer (éventuellement numériquement) ;
2. la probabilité $P_1(\vec{x}, t)$ qu'un individu initial à x ait un unique survivant au temps t ; elle est obtenue en linéarisant l'équation différentielle de type KPP sur $Q_e(\vec{x}, t)$ et en l'intégrant jusqu'à t.

On pourrait penser *a priori* que la détermination de $Q_e(\vec{x}, t)$ et $P_1(\vec{x}, t)$ est aussi complexe que le problème de départ qui consiste à conditionner la marche aléatoire avec branchements mais ce n'est effectivement pas le cas. En effet, la simulation simpliste du régime conditionné consisterait à faire des simulations directes du processus et ne conserver que les réalisations vérifiant la bonne condition finale à T : pour T grand, le rendement est très faible (exponentiellement décroissant en T) et il devient très vite impossible d'accumuler des statistiques sur le régime conditionné. Au contraire, la détermination numérique de $Q_e(\vec{x}, t)$ et $P_1(\vec{x}, t)$ consiste à intégrer une seule fois deux équations aux dérivées partielles entre $[0, T]$ et à utiliser le résultat stocké dans la mémoire pour produire autant de réalisations que nous le souhaitons.

L'avantage de produire des réalisations du régime conditionné permet de mesurer aussi bien des propriétés statiques (profil de densité de la population à un instant donné) que des propriétés dynamiques telles que les arbres généalogiques et les temps de coalescence des individus.

[1] C'est-à-dire sans produire des séquences de configurations sur $[0, T]$ qui ne vérifieraient pas la bonne condition finale et seraient donc à éliminer.

5.2 Construction du processus modifié

Le modèle générique de branchement-diffusion considéré ici est le suivant :
- les individus sont repérés par leur position $\vec{x}$ à l'intérieur d'une domaine $\mathcal{D}$ à d dimensions ;
- chaque individu diffuse avec une constante de diffusion D et une dérive $\vec{v}(\vec{x})$;
- en outre, pendant un temps infinitésimal dt, il se divise en k individus avec une probabilité $\beta_k(\vec{x})dt$ ou meurt avec une probabilité $\beta_0(\vec{x})dt$;
- tout individu touchant le bord du domaine $\mathcal{D}$ disparaît instantanément.

Pour des modèles sur réseau et/ou à temps discret, toutes les formules ci-dessous s'adaptent aisément, en remplaçant les dérivées spatiales par des différences finies entre les noeuds du réseau. Les taux de saut et de division similaires à ceux présentés dans le tableau 5.1 peuvent être calculés de la même manière que celle présentée ci-dessous.

5.2.1 Lien avec les fonctions génératrices

Rappelons les définitions des fonctions génératrices G_1 et G_2 pour une géométrie quelconque :

$$\begin{aligned} G_1(\vec{x}, g_1; t) &= \Big\langle \exp\left(-\sum_{i=1}^{N_t} g_1(\vec{x}_i(t)) \right) \Big\rangle \\ G_2(\vec{x}, g_1, g_2; t, t') &= \Big\langle \exp\left(-\sum_{i=1}^{N_t} g_1(\vec{x}_i(t)) - \sum_{j=1}^{N_{t+t'}} g_2(\vec{x}_j(t')) \right) \Big\rangle \end{aligned}$$

où N_t désigne la taille de la population au temps t et les $x_i(t)$ les positions des individus.

Le cas $g_2 = \nu$ constant donne la fonction génératrice de la taille à l'instant $t+t'$. Ainsi, de manière similaire à (4.5), détailler la moyenne sur la taille à l'instant $t + t'$ permet d'écrire G_2 sous la forme suivante :

$$\begin{aligned} G_2(\vec{x}, g_1, \nu; t, t') &= \sum_{m=0}^{\infty} e^{-\nu m} R_m(\vec{x}, g_1; t, t') && (5.1) \\ &= \sum_{m=0}^{\infty} e^{-\nu m} P_m(\vec{x}, t+t') \Big\langle \exp\left(-\sum_{i=1}^{N_t} g_1(\vec{x}_i(t)) \right) \Big| N_{t+t'} = m \Big\rangle && (5.2) \end{aligned}$$

où $P_m(\vec{x}, t+t')$ est la probabilité d'observer une taille m à l'instant $t+t'$ pour un individu initial en $\vec{x}$. On a ainsi en particulier $P_0 = Q_e$. La fonction génératrice de g_1 à l'instant t connaissant la taille m ultérieure est donnée par le rapport (cf. (4.6)) :

$$\boxed{\Big\langle \exp\left(-\sum_{i=1}^{N_t} g_1(\vec{x}_i(t)) \right) \Big| N_{t+t'} = m \Big\rangle = \frac{R_m(\vec{x}, g_1; t, t')}{P_m(\vec{x}, t+t')} \underset{\text{déf.}}{=} \widetilde{R}_m(x, g_1; t, t')} \qquad (5.3)$$

où R_m est donné par le développement en série de $G_2(\vec{x}, g_1, \nu; t, t')$ sur $e^{-\nu}$.

D'autre part, la dynamique de G_1 et G_2 est connue, puisque ces fonctions génératrices satisfont la même équation que la probabilité de survie $Q_e(\vec{x},t)$ (cf. chapitre 4). Nous avons ainsi ainsi :

$$\begin{aligned}\partial_t G_1(\vec{x},g_1;t) &= D\vec{\nabla}^2 G_1(\vec{x},g_1;t) + \vec{v}(\vec{x})\cdot\vec{\nabla}G_1(\vec{x},g_1;t) \\ &\quad + \sum_{k=0}^{\infty}\beta_k(\vec{x})\left(G_1(\vec{x},g_1;t)^k - G_1(\vec{x},g_1;t)\right) \\ &= D\vec{\nabla}^2 G_1(\vec{x},g_1;t) + \vec{v}(\vec{x})\cdot\vec{\nabla}G_1(\vec{x},g_1;t) + f_e\left(G_1(\vec{x},g_1;t)\right)\end{aligned} \tag{5.4}$$

et la même équation pour $G_2(\vec{x},g_1,\nu;t,t')$ avec t' fixé. Afin de faciliter la lecture, nous omettrons la liste des arguments $(x,g_1;t,t')$ des fonctions R_0 et R_1 sauf s'ils sont différents du cas générique. En insérant dans l'équation d'évolution de $G_2(\vec{x},g_1,\nu;t,t')$ le développement (5.1), on obtient aux ordres les plus bas pour $R_m(\vec{x},g_1;t,t')$:

$$\partial_t R_0 = D\vec{\nabla}^2 R_0 + \vec{v}(\vec{x})\cdot\vec{\nabla}R_0 + f_e\left(R_0\right) \tag{5.5a}$$

$$\partial_t R_1 = D\vec{\nabla}^2 R_1 + \vec{v}(\vec{x})\cdot\vec{\nabla}R_1 + f_e'\left(R_0\right)R_1 \tag{5.5b}$$

et les mêmes équations pour P_0 et P_1 puisque celles-ci peuvent être obtenues à partir de R_0 et R_1 en prenant $g_1 = 0$.

L'équation (5.3) montre que les quantités pertinentes pour étudier le régime conditionné sont les rapports $\widetilde{R}_m = R_m/P_m$. Les équations (5.5) sont constituées d'un opérateur différentiel spatial de degré 2 et d'une fonction non-linéaire. Cette structure permet de montrer, en substituant $R_m = \widetilde{R}_m P_m$ dans (5.5), que les rapports $\widetilde{R}_m$ satisfont aussi le même type d'équation à condition de redéfinir convenablement la dérive et les taux de division [SD08] :

$$\partial_t \widetilde{R}_0 = D\vec{\nabla}^2\widetilde{R}_0 + \left(\vec{v}(\vec{x}) + 2D\frac{\vec{\nabla}Q_e}{Q_e}\right)\cdot\vec{\nabla}\widetilde{R}_0 + \sum_{k\geq 0}\beta_k Q_e^{k-1}(\widetilde{R}_0^k - \widetilde{R}_0) \tag{5.6a}$$

$$\begin{aligned}\partial_t \widetilde{R}_1 &= D\vec{\nabla}^2\widetilde{R}_1 + \left(\vec{v}(\vec{x}) + 2D\frac{\vec{\nabla}P_1}{P_1}\right)\cdot\vec{\nabla}\widetilde{R}_1 \\ &\quad + \sum_{k\geq 2} k\beta_k Q_e^{k-1}\left(\widetilde{R}_1\widetilde{R}_0^{k-1} - \widetilde{R}_1\right)\end{aligned} \tag{5.6b}$$

qui ont une structure similaire à (5.4).

Pour caractériser complètement les G_i, R_m, P_m et $\widetilde{R}_m$, il faut étudier leurs conditions aux bords. Pour $x = 0$, l'individu initial est tué instantanément et donc nous avons

$$G_1(0,g_1;t) = G_2(0,g_1,g_2;t,t') = 0.$$

D'autre part, à $t = 0$, la condition initiale correspond à un unique individu en x et nous avons :

$$G_1(x,g_1;0) = e^{-g_1(x)}, \qquad G_2(x,g_1,g_2;0,t') = e^{-g_1(x)}G_1(x,g_2;t').$$

Un développement en puissance de $g_2 = \nu$ donne les conditions initiales pour les R_m et P_m et se répercute de la manière suivante sur les $\widetilde{R}_m$:

$$\widetilde{R}_m(0,g_1;t,t') = 0, \qquad \widetilde{R}_m(x,g_1;0,t') = e^{-g_1(x)}. \tag{5.7}$$

5.2.2 Description de la dynamique modifiée

La similitude entre (5.4) et (5.6) et les conditions initiales identiques entre les $\widetilde{R}_m$ et G_1 font que les fonctions $\widetilde{R}_m$ peuvent être interprétées comme les fonctions génératrices d'un processus de branchement-diffusion modifié. Une démarche plus rigoureuse [SD08] consiste à montrer que les fonctions génératrices du processus modifié défini ci-dessous satisfont les mêmes équations d'évolution que les $\widetilde{R}_m$ et possèdent les mêmes conditions initiales[2].

La présence de deux équations couplées (5.6a) et (5.6b) indique qu'il faut considérer non plus une mais deux espèces d'individus A_0 et A_1. La fonction $\widetilde{R}_0$ est la fonction génératrice de g_1 sachant que l'individu initial est de type A_0, alors que $\widetilde{R}_1$ correspond à un individu initial de type A_1. Les coefficients de diffusion et les dérives ainsi que les taux de division des particules de chaque type se lisent directement sur (5.6) et sont résumés dans le tableau 5.1. Par exemple, le terme $k\beta_k Q_e^{k-1}(\widetilde{R}_1\widetilde{R}_0^{k-1} - \widetilde{R}_1)$ s'interprètent comme la division d'une particule A_1 en une particule A_1 et $k-1$ particules A_0 avec un taux $k\beta_k Q_e^{k-1}$.

Le processus modifié est défini sur l'intervalle $[0, T]$ de la manière suivante :

- *la population est divisée en deux classes A_0 et A_1 ;*
- *chaque particule suit une dynamique de branchement-diffusion ;*
- *toutes les particules ont le même coefficient de diffusion D ;*
- *les particules A_1 ont une dérive :*

$$\vec{v}_1(\vec{x}, t, T) = \vec{v}(\vec{x}) + 2D\frac{\vec{\nabla}P_1}{P_1}(\vec{x}, T-t); \tag{5.8}$$

- *les particules A_0 ont une dérive :*

$$\vec{v}_0(\vec{x}, t, T) = \vec{v}(\vec{x}) + 2D\frac{\vec{\nabla}Q_e}{Q_e}(\vec{x}, T-t); \tag{5.9}$$

- *les particules A_1 se divisent selon :*

$$A_1 \xrightarrow{\beta_k^{(1)}(\vec{x},t,T)} A_1 + (k-1)A_0 \tag{5.10}$$

  pour $k \geq 2$ avec un taux

$$\beta_k^{(1)}(\vec{x}, t, T) = k\beta_k(\vec{x})Q_e(\vec{x}, t, T-t)^{k-1}; \tag{5.11}$$

- *les particules A_0 se divisent selon :*

$$A_0 \xrightarrow{\beta_k^{(0)}(\vec{x},t,T)} kA_0 \tag{5.12}$$

[2]Les fonctions $\widetilde{R}_m$ sont paramétrées par t et t' où t' est la durée (fixée) entre l'instant d'observation t et l'instant final $t+t'$ de conditionnement sur la taille et les dérivées temporelles (5.6) sont à t' constant. Dans le problème initial, au contraire, le conditionnement se fait à un instant final T fixé et on fait varier le temps d'observation de 0 à T et donc t' varie. Après passage de (t, t') à (t, T), on peut vérifier (cf. (5.3)) que la fonction génératrice de g_1 à un instant t conditionnée par un unique survivant à T est bien donnée par $\widetilde{R}_1(\vec{x}, g_1; t, T-t)$.

Processus	Type	Dérive	Division	Taux
originel	A	$\vec{v}(\vec{x})$	$A \to kA$	$\beta_k(\vec{x})$ $k \geq 0$
modifié	A_0	$\vec{v}(\vec{x}) + 2D\frac{\vec{\nabla}Q_e}{Q_e}(\vec{x}, T-t)$	$A_0 \to kA_0$	$\beta_k(\vec{x})Q_e(\vec{x}, T-t)^{k-1}$ $k \geq 0$
	A_1	$\vec{v}(\vec{x}) + 2D\frac{\vec{\nabla}P_1}{P_1}(\vec{x}, T-t)$	$A_1 \to A_1$ $+(k-1)A_0$	$k\beta_k(\vec{x})Q_e(\vec{x}, T-t)^{k-1}$ $k \geq 2$

TAB. 5.1: Dynamique modifiée des particules de type A_1 et A_0. La première ligne correspond au processus non conditionné, les deuxième et troisième lignes donnent la dérive $\vec{v}(\vec{x})$ et les taux de division $\beta_k(\vec{x})$ des particules dans le processus stochastique biaisé. Le coefficient de diffusion D reste le même pour tous les types de particules.

pour $k \geq 2$ avec un taux

$$\beta_k^{(0)}(\vec{x}, t, T) = \beta_k(\vec{x})Q_e(\vec{x}, t, T-t)^{k-1}; \tag{5.13}$$

– *les particules A_0 meurent ($A_0 \to \emptyset$) avec un taux $\beta_0^{(0)}(\vec{x}, t, T-t) = \beta_k(\vec{x})/Q_e(\vec{x}, t, T)$.*

Ce processus modifié exhibe toutes les propriétés attendues. Tout d'abord, le nombre de particules A_1 est conservé sur tout l'intervalle $[0, T]$: les particules de type A_1 ne peuvent pas mourir spontanément (il n'y a pas de $\beta_0^{(1)}(\vec{x})$), les réactions de type $A_1 \to A_1 + (k-1)A_0$ n'affectent pas leur nombre et enfin elles ne peuvent pas toucher le bord absorbant du domaine $\mathcal{D}$. En effet, nous avons $P_1(\vec{x}, T-t) = 0$ sur le bord de $\mathcal{D}$ et $P_1(\vec{x}, T-t) > 0$ à l'intérieur de $\mathcal{D}$: la dérive $v_1(\vec{x})$ repousse alors les particules A_1 vers l'intérieur du domaine[3].

Ainsi, si le processus commence avec une seule particule A_1, alors il n'y en a qu'une sur tout l'intervalle $[0, T]$. Nécessairement, l'unique survivant à T est alors la particule A_1 de départ elle-même. La particule A_1 qui apparaît quand on étudie $\widetilde{R}_1$ dans le cadre du processus modifié n'est autre que la particule qui aura survécu sur tout l'intervalle $[0, T]$ dans le processus initial conditionné (cf. figure 5.1). Elle correspond à *l'épine dorsale* introduite dans les travaux mathématiques [LPP95, HH07] à partir de laquelle toute la population est générée. La section 5.3.1 vise à décrire les propriétés de cette *épine dorsale* au cours du temps.

Au contraire, les particules A_0 disparaissent avec probabilité 1 quand $t = T$. En effet la probabilité $Q_e^{(0)}(\vec{x}, t; T)$ d'extinction à un temps t d'une particule A_0 initialement à $\vec{x}$ est donnée par la limite $\widetilde{R}_0(\vec{x}, \infty; t, T-t)$. Or, nous avons par construction :

$$\widetilde{R}_0(\vec{x}, \infty; t, T-t) = \frac{R_0(\vec{x}, \infty; t, T-t)}{P_0(\vec{x}, T)} = \frac{Q_e(\vec{x}, t)}{Q_e(\vec{x}, T)}$$

[3]Pour le montrer rigoureusement, il faut étudier la forme précise de P_1 au bord du domaine et montrer que la dérive diverge suffisamment sur le bord de $\mathcal{D}$. On supposera cependant que c'est le cas par construction pour toute dimension. Le vérifier en dimension 1 est trivial.

et donc $Q_e^{(0)}(\vec{x}, t; T) = 1$, ce qui correspond à une extinction certaine des particules A_0 avant l'instant final.

De plus, nous lisons directement sur les dérives et taux du tableau 5.1 que les particules A_0 sont *chassées* des zones où la probabilité d'extinction est faible : d'une part la dérive éloigne les particules A_0 des régions où Q_e est faible et, d'autre part, les taux de division $\beta_k^{(0)}$ pour $k \geq 2$ s'effondrent dans ces régions alors que le taux de mort spontanée $\beta_0^{(0)} = \beta_0(\vec{x})/Q_e(\vec{x}, t, T-t)$ augmente. Cela se comprend aisément en repensant à l'origine du processus modifié : le seul moyen d'avoir un unique survivant à l'instant final T est d'empêcher l'extinction totale (d'où la particule A_1 qui ne peut pas entrer en contact avec le mur) et, par ailleurs, d'interdire un trop fort développement de la population avant le temps final T : les particules ne doivent pas entrer dans les zones où elles ont une forte probabilité de croître exponentiellement *i.e.* où Q_e est faible.

5.2.3 Lien avec l'existence d'un état quasi-stationnaire

La limite $T \to \infty$ avec t fixé permet d'étudier les propriétés statistiques de la population à un temps t séparé du temps de conditionnement par $T - t \to \infty$: on s'attend, *a priori*, à ce que les propriétés du système au temps t ne dépendent pas des détails précis du conditionnement (extinction à T, taille 1 ou 2 à T, etc.), puisque la convergence vers l'état final peut se faire au dernier moment. En revanche, les propriétés du système à t fini dépendent des conditions initiales à $t = 0$: nous nous attendons alors à ce que, dans la limite $T \to \infty$ à t fixé, les taux de division et dérives du processus modifié convergent vers des valeurs limites.

À supposer que les dérives v_0 et v_1 et les taux de division $\beta_k^{(i)}$ aient une limite lorsque $T \to \infty$, cela n'implique pas pour autant l'existence d'un régime quasi-stationnaire, mais seulement l'existence d'une dynamique modifiée à tout temps t lorsque $T \to \infty$: indépendamment de l'existence du régime quasi-stationnaire, cette dynamique modifiée qui consiste à éviter à jamais l'état absorbant est appelée *Q-processus* par les mathématiciens [CCL$^+$07, Lam07]. Pour observer un régime quasi-stationnaire, il faut encore que cette dynamique modifiée conduise à un état stationnaire lorsque $t \to \infty$. L'ordre des limites $T \to \infty$ et $t \to \infty$ est ainsi important.

Pour qu'un régime quasi-stationnaire existe, il faut donc, dans le cadre du processus biaisé, que :

1. les dérives et taux de division aient une limite finie lorsque $T \to \infty$;
2. la dynamique ainsi modifiée conduise à un état stationnaire.

De plus, le régime stationnaire de la dynamique modifiée est alors identique (par construction) au régime quasi-stationnaire de la dynamique de départ.

Revenons à la marche aléatoire avec branchements et mur absorbant en dimension 1 que nous avons étudiée au chapitre 3. Le critère précédent appliqué à ce modèle permet d'expliquer les différences entre les résultats numériques des figures 4.1 et 4.3 [DS07] : pour $v < v_c$, la probabilté $Q_e(x, t)$ tend uniformément vers $Q_e^*(x)$ et la dynamique aux temps longs de P_1 ressemble à

$$\partial_t P_1 \simeq D\Delta P_1 + \vec{v}(\vec{x}) \cdot \vec{\nabla} P_1 + f'(Q_e^*) P_1. \tag{5.14}$$

Cela correspond à la forme linéarisée de l'équation KPP autour du point fixe Q_e^*. Dans le cas $v < v_c$, le système possède un temps de relaxation (3.40) le plus lent isolé $\tau_1 = -1/\lambda_1$ associé au vecteur propre ϕ_1 (3.41). Nous obtenons ainsi $P_1(x,t) \propto \phi_1(x)e^{-t/\tau_1}$ puis :

$$v_1^*(x) = \lim_{T\to\infty} v_1(x,t,T) = -v + 2\frac{\phi_1'(x)}{\phi_1(x)}. \tag{5.15}$$

La forme (3.41) permet de vérifier que $v_1^*(x) \propto 1/x$ pour x proche de 0 et que $v_1^*(x) < 0$ pour $x > L$ (région $Q_e^*(x) \ll 1$) : la particule A_1 est ainsi piégée dans la région $0 < x < L$ et on peut vérifier de la même manière que les particules A_0 sont aussi confinées dans cette région. Comme nous allons le voir en détail ci-dessous, ces expressions mènent au régime quasi-stationnaire décrit au chapitre précédent pour $v < v_c$.

Pour $v > v_c$ cependant, l'*ansatz* (3.27) aux temps longs pour Q_e et P_1 montre que $Q_e^*(x) = 1$ (aucun taux de division des particules A_0 n'est affecté) alors que v_1 devient :

$$v_1(x,t,T) \simeq -v + 2\left(\frac{(v/2)\sin(\pi x/L_{T-t}) + (\pi/L_{T-t})\cos(\pi x/L_{T-t})}{\sin(\pi x/L_{T-t})}\right) \simeq \frac{2}{x}. \tag{5.16}$$

pour tout $x > 0$. La dérive $v_1^*(x) = 2/x$ de la particule A_1 l'éloigne ainsi indéfiniment du mur au cours du temps alors que les taux de division et la dérive $v_0^*(x) = -v$ sont uniformes : la particule A_1 va ainsi s'éloigner en générant des particules A_0 toujours plus loin du mur et ainsi la population va croître indéfiniment dans cette dynamique. Puisque la taille ne se stabilise pas, le système ne peut donc pas atteindre de régime quasi-stationnaire, comme observé en figure 4.3. Néanmoins, cette dynamique modifiée permet de comprendre l'évolution conditionnée du système pour $t \ll T$: en particulier, lorsque T grandit, la partie de la courbe de la figure 4.3 à t petit et constant ne change pas.

5.3 Densités dans le régime quasi-stationnaire

Dans le régime quasi-stationnaire, les taux de division et les dérives du processus biaisé sont obtenus en prenant la limite $T \to \infty$ des taux résumés dans le tableau 5.1 :

$$\begin{aligned} \vec{v}_0^*(\vec{x}) &= \vec{v}(\vec{x}) + 2D\tfrac{\nabla Q_e^*}{Q_e^*}(\vec{x}), \quad \beta_k^{(0),*}(\vec{x}) = \beta_k(\vec{x})Q_e^*(\vec{x})^{k-1} \\ \vec{v}_1^*(\vec{x}) &= \vec{v}(\vec{x}) + 2D\tfrac{\nabla \phi_1}{\phi_1}(\vec{x}), \quad \beta_k^{(1),*}(\vec{x}) = k\beta_k(\vec{x})Q_e^*(\vec{x})^{k-1} \end{aligned} \tag{5.17}$$

où $Q_e^*(\vec{x})$ est le point fixe stable de l'équation (5.4) et $\phi_1(\vec{x})$ le vecteur propre correspondant au temps de relaxation le plus lent au voisinage de $Q_e^*(\vec{x})$ (extension directe de (3.15) à une géométrie générale).

5.3.1 Probabilité de présence du survivant

Comme expliqué précédemment, la particule A_1, la *colonne vertébrale* du processus de branchement-diffusion, survit éternellement et est la source de toutes les particules A_0. Elle retrace ainsi la position de l'individu dont la lignée aura survécu sur tout l'intervalle $[0,T]$. Il s'agit à présent de caractériser sa position au cours du temps par sa probabilité

$\rho_{1,\mathrm{st}}(x,t)$ d'être en x au temps t. Cette probabilité satisfait l'équation de Fokker-Planck suivante avec la dérive (5.8) :

$$\partial_t \rho_{1,\mathrm{st}} = D\vec{\nabla}^2 \rho_{1,\mathrm{st}}(\vec{x}) - \vec{\nabla}\cdot(\vec{v}_1(\vec{x},t,T)\rho_{1,\mathrm{st}})\,. \tag{5.18}$$

L'évolution de $\vec{v}_1(x,t,T)$ n'est pas simple *a priori* et la formule précédente n'est exploitable que numériquement. Cependant, dans le régime quasi-stationnaire, *i.e.* dans les limites successives $T\to\infty$ puis $t\to\infty$ lorsque $v<v_c$, la densité ρ_1 satisfait l'équation suivante indépendante de t :

$$D\vec{\nabla}^2\left(\rho_{1,\mathrm{st}}(\vec{x})\right) = \vec{\nabla}\cdot(\vec{v}_1^*(\vec{x})\rho_{1,\mathrm{st}})$$

où $\vec{v}_1^*(\vec{x})$ est donnée par (5.15). Une première intégration donne :

$$\frac{\vec{\nabla}(\rho_{1,\mathrm{st}})}{\rho_{1,\mathrm{st}}} = \frac{\vec{v}}{D} + 2\frac{\vec{\nabla}\phi_1}{\phi_1}$$

Nous nous limiterons ici aux modèles où le champ de dérive est le gradient d'un potentiel $U(\vec{x})$:

$$\vec{v} = -\vec{\nabla}U. \tag{5.19}$$

Dans ce cas, la densité quasi-stationnaire ρ_1 de l'*épine dorsale* prend la forme suivante :

$$\boxed{\rho_{1,\mathrm{st}}(\vec{x}) = \frac{1}{Z_1}\phi_1(\vec{x})^2 e^{-U(\vec{x})/D}} \tag{5.20}$$

où Z_1 est la constante de normalisation assurant $\int_{\mathcal{D}}\rho_{1,\mathrm{st}} = 1$.

Cette formule contient deux termes : le premier, en $e^{-U(\vec{x})/D}$, correspond à la distribution de probabilité d'une particule brownienne dans le potentiel U sans souci d'absorption aux bords ; le deuxième, en $\phi_1(\vec{x})^2$, correspond à la condition de non-absorption de la particule A_1 par les bords absorbants du domaine ($\phi_1(\vec{x})$ s'annule sur les bords). Les lieux où $\phi_1(\vec{x})^2$ est petit correspondent aux domaines interdits à la particule A_1 dans lesquels il pourrait y avoir prolifération ou extinction de la population et ainsi violation de la condition finale de taille unité.

5.3.2 Densité d'individus A_0

Dans le régime conditionné, les individus A_0 sont créés par la particule A_1 et éliminés soit spontanément (avec un taux $\beta_0^{(0)}$), soit sur les bords du domaine. L'équation donnant la densité $\rho_{0,\mathrm{st}}(\vec{x},t)$ d'individus A_0 à un instant t au point $\vec{x}$ est similaire à (5.18) avec deux termes de sources supplémentaires liés à la division des individus :

$$\partial_t\rho_{0,\mathrm{st}} = D\vec{\nabla}^2\rho_{0,\mathrm{st}} - \vec{\nabla}\cdot(\vec{v_0}\rho_{0,\mathrm{st}}) + \sum_k (k-1)\beta_k^{(0),*}\rho_{0,\mathrm{st}} + \sum_k (k-1)\beta_k^{(1),*}\rho_{1,\mathrm{st}} \tag{5.21}$$

Là encore, cette équation permet une étude numérique de la densité ρ_0 au cours du temps mais pour une étude analytique, cela reste trop complexe. Nous allons nous

placer dans le régime quasi-stationnaire en prenant successivement $T \to \infty$ puis $t \to \infty$ et en supposant qu'un régime quasi-stationnaire existe. Nous avons ainsi l'équation stationnaire suivante :

$$D\vec{\nabla}^2\rho_{0,\mathrm{st}} - \vec{\nabla}\left(\vec{v_0}\rho_{0,\mathrm{st}}\right) + \sum_k \beta_k^{(0)}(k-1)\rho_{0,\mathrm{st}} = -\sum_k \beta_k^{(1)}(k-1)\rho_{1,\mathrm{st}} \tag{5.22}$$

dont nous connaissons déjà le second terme (5.20).

Pour résoudre cette équation dans un milieu **unidimensionnel** avec un coefficient de diffusion D constant et une dérive $\vec{v} = -\partial_x U$, on peut effectuer le changement de variable suivant :

$$\rho_{0,\mathrm{st}}(\vec{x}) = e^{-U(\vec{x})/D} Q_e^*(\vec{x})\psi(\vec{x}) \tag{5.23}$$

qui ramène (5.22) à l'équation suivante :

$$\mathcal{L}[\psi] = -\rho_1 e^{U/D} f_e''(Q_e^*) \tag{5.24}$$

où $\mathcal{L}$ est l'opérateur linéarisé introduit en (3.38) qui nous avait permis d'étudier la relaxation d'une solution de l'équation F-KPP vers son point fixe Q_e^*. Dans le cas général où le domaine $\mathcal{D}$ est multi-dimensionnel, résoudre l'équation aux dérivées partielles (5.24) n'est pas simple.

Dans le cas de la demi-droite $[0, +\infty[$, (5.24) est une équation différentielle linéaire de degré 2 dont nous connaissons déjà une solution homogène. En effet, en dérivant (3.11), il apparaît immédiatement que $\partial_x Q_e^*(x)$ satisfait $\mathcal{L}[\partial_x Q_e^*] = 0$, et il devient possible d'exprimer ψ puis $\rho_{0,\mathrm{st}}$ uniquement en fonction de Q_e^* et ϕ_1. Un calcul détaillé [SD08] montre que la solution générale pour un problème **unidimensionnel** est donnée par :

$$\boxed{\begin{aligned}\rho_{0,\mathrm{st}}(x) =& e^{-U(x)/D} Q_e^*(x)\partial_x Q_e^*(x) \\ &\times \int_0^x \frac{dy}{[\partial_x Q_e^*(y)]e^{-U(y)/D}} \int_y^\infty \rho_{1,\mathrm{st}}(z)\partial_x Q_e^*(z) f_e''(Q_e^*(z))dz\end{aligned}} \tag{5.25}$$

où les constantes d'intégration ont été fixées de telle sorte à imposer $\rho_0(x) = 0$ pour $x = 0$ et $x \to \infty$ aux bords du domaine $\mathcal{D}$.

Dans le régime quasi-stationnaire conditionné par une taille finale unité, la construction d'un processus modifié nous a permis d'étudier tout le profil de densité quasi-stationnaire $\langle\rho\rangle_{\mathrm{qs}}$ dans un certain nombre de cas simples en le décomposant en deux contributions. Ainsi, pour $\langle\rho\rangle_{\mathrm{qs}}$, il ne reste plus qu'à rassembler les résultats exacts (5.20, 5.25) :

$$\boxed{\langle\rho(x)\rangle_{\mathrm{qs}} = \rho_{0,\mathrm{st}}(x) + \rho_{1,\mathrm{st}}(x)}\,. \tag{5.26}$$

Comme pour $\rho_{1,\mathrm{st}}(\vec{x})$, le profil $\rho_{0,\mathrm{st}}(\vec{x})$ est donné par la distribution $e^{-U(\vec{x})/D}$ modulée par une quantité ne dépendant que de la probabilité de survie $Q_e^*(\vec{x})$ et de $\phi_1(\vec{x})$. Cette formule permet d'estimer numériquement $\rho_0(x)$ dans un grand nombre de cas et une

étude analytique pourra être entreprise à chaque fois que l'on connaîtra les propriétés des solutions de l'équation F-KPP sur un domaine $\mathcal{D}$ avec les résultats (5.20,5.25) pour point de départ.

Dans notre travail, deux situations nous ont intéressés plus particulièrement et seront étudiées au prochain chapitre. Elles concernent toutes deux la marche aléatoire unidimensionnelle avec branchements près de la vitesse critique v_c sur les domaines $\mathcal{D}$ suivants :

1. $\mathcal{D} = [0, +\infty[$, $v(x) = -v$ et $v \to v_c^-$: cela correspond à un mur absorbant se déplaçant à vitesse v vers les x positifs (cf. chapitre 3) ; en particulier, un régime quasi-stationnaire existe pour $v < v_c$.
2. $\mathcal{D} = [0, l]$ et $v(x) = -v$: cela correspond à un domaine borné avec bords absorbants. Un régime quasi-stationnaire existe des deux côtés de la vitesse critique.

Conclusion

L'étude des fonctions génératrices à un et deux temps du processus de branchement-diffusion a permis de construire un processus stochastique biaisé équivalent au processus initial conditionné par une taille unité au temps final. La démarche, relativement générale, donne ainsi accès à une simulation directe de ce processus conditionné, le prix à payer étant la résolution (éventuellement numérique) de deux équations aux dérivées partielles.

Dans un second temps, nous avons vu que, lorsque la dérive des particules se met sous la forme $\vec{v} = -\vec{\nabla} U$, il devient possible d'écrire exactement la distribution quasi-stationnaire (5.20) de la position de la particule qui survit tout au long du processus ainsi que la densité quasi-stationnaire moyenne (5.20,5.26) de la population dans le cas uni-dimensionnel.

La construction de la dynamique modifiée est valide tant que les particules sont indépendantes les unes des autres : c'est en effet cette propriété qui a permis d'écrire la relation de récurrence (exacte) satisfaite par les fonctions génératrices. Lorsque les recombinaisons $A + A \to A$ sont autorisées, il n'est plus possible de suivre l'ancêtre du survivant final en remontant le temps de $t = T$ à $t = 0$ et la démarche précédente ne semble pas se généraliser directement. Il serait intéressant de comprendre de manière plus approfondie la construction du processus modifié afin d'envisager une extension aux modèles avec interactions.

Chapitre 6

Régime quasi-stationnaire au voisinage du point critique en dimension 1

Ce chapitre regroupe, dans une première moitié, les résultats principaux que nous avons publiés dans [SD08] sur le régime quasi-stationnaire d'une marche aléatoire unidimensionnelle avec branchements en présence d'un unique mur absorbant (ligne semi-infinie). Dans une seconde partie (contribution non publiée), nous présentons les équivalents de ces résultats en présence de deux murs absorbants éloignés d'une distance l fixée (boîte de longueur l). Les calculs sont relativement limités puisque tous les résultats découlent de ceux obtenus dans les chapitres précédents. Le chapitre 5 a montré qu'il suffisait de connaître la solution Q_e^* de l'équation F-KPP et le vecteur propre ϕ_1 pour décrire la densité quasi-stationnaire $\langle \rho(x) \rangle_{\text{qs}}$ d'individus et d'autre part les propriétés de Q_e^* et ϕ_1 ont déjà été détaillées dans les chapitres 2 et 3.

Nous revenons à l'étude, commencée au chapitre 3, d'une marche aléatoire avec branchements en présence d'un bord absorbant se déplaçant à une vitesse v, lorsque la vitesse v est proche de la vitesse critique v_c obtenue en (3.4). Dans le régime où elle est conditionnée à avoir une taille $N_T = 1$ à un temps final T, cette marche aléatoire avec branchements conduit à un régime quasi-stationnaire pour une vitesse du mur v inférieure à la vitesse critique (cf. chapitre 4) et elle peut être décrite par un processus modifié que nous avons construit au chapitre 5 : en dimension 1, cette démarche a conduit à une expression intégrale (5.25) de la densité quasi-stationnaire de la population pour $v < v_c$. Nous nous proposons, dans ce chapitre, de voir ce que deviennent ces expressions au voisinage du point critique. Étant donné qu'il est de plus en plus difficile de survivre lorsque la vitesse du mur s'approche de v_c, il est légitime de penser que la taille quasi-stationnaire de la population doit croître lorsque $v \to v_c^-$ afin d'assurer la survie jusqu'à l'instant final et nous nous attacherons à caractériser cette divergence.

6.1 En présence d'un unique mur absorbant : ligne semi-infinie

6.1.1 Densité quasi-stationnaire pour $v < v_c$

La marche aléatoire avec branchements se déplace sur $\mathcal{D} = [0, \infty[$ avec une dérive $-v$ vers le mur absorbant positionné en $x = 0$ (référentiel du mur). Nous avons vu au chapitre 3 que la probabilité d'extinction aux temps longs $Q_e^* = 1 - Q_s^*$ présentait, au voisinage de la vitesse critique v_c, deux régions correspondant à $Q_e^*(x) \simeq 1$ (région I de taille L donnée par (3.24)) et $Q_e^*(x) \ll 1$ (région II). D'après les formules (5.20, 5.25), il faut donc s'attendre à des expressions différentes pour les densités de particules A_1 et A_0 dans les régions I et II.

À cette dérive $-v$ vers le mur dans le référentiel de celui-ci est associé un potentiel $U(x) = vx$. La probabilité de présence de la particule A_1 est simplement donnée par la formule (5.20) et les expressions (3.41,3.42) de $\phi_1(x)$. Dans la région I, correspondant à $0 \leq x < L$, de taille $L \propto (v_c - v)^{-1/2}$, $\phi_1(x)$ a une forme universelle, alors que dans la seconde (région II), $\phi_1(x)$ ressemble à $\partial_x Q_e^*(x)$ à l'ordre dominant. En réunissant les différents résultats, nous obtenons :

$$\rho_{1,\mathrm{st}}(x) \simeq \frac{1}{Z_1} \begin{cases} \left(\frac{A_c v_c L e^{v_c(L+B_c/A_c)/2}}{\pi} \right)^2 \sin^2\left(\frac{\pi x}{L}\right) & \text{(région I, } x < L\text{)}, \\ [\partial_x Q_{v_c}(x - L - B_c/A_c)]^2 e^{-v_c x} & \text{(région II, } x > L\text{)}. \end{cases} \tag{6.1}$$

Dans la région I, la densité ρ_1 de la particule A_1 est simplement donnée par un terme $\sin^2(\pi x/L)$ qui ne dépend donc pas du détail des non-linéarités (3.12) de l'équation F-KPP. Dans la région II, la densité est décroissante et se raccorde à la région I dans la région $x - L = O(1)$ où la densité ρ_1 est déjà négligeable par rapport à l'intérieur de la région I (facteur $1/L^2$) : la particule A_1 n'entre donc pratiquement pas dans la région II. Cela se comprend bien si l'on se souvient que cette région correspond à une survie probable et donc une prolifération exponentielle probable, incompatible avec le conditionnement final.

La contribution de la région II est donc négligeable au premier ordre dans la détermination de la constante de normalisation Z_1. Nous obtenons alors pour Z_1 :

$$Z_1 \simeq \int_0^L \left(\frac{A_c v_c L e^{v_c(L+B_c/A_c)/2}}{\pi} \right)^2 \sin^2\left(\frac{\pi x}{L}\right) dx \simeq \left(\frac{A_c v_c L e^{v_c(L+B_c/A_c)/2}}{\pi} \right)^2 \frac{2}{L} \tag{6.2}$$

puis l'expression simplifiée suivante de ρ_1 dans la région I :

$$\boxed{\rho_{1,\mathrm{st}}(x) \underset{v \to v_c^-}{\simeq} \frac{2}{L} \sin^2\left(\frac{\pi x}{L}\right), \qquad (x < L)} \,. \tag{6.3}$$

La densité de particules A_0 est donnée par (5.25). En analysant les contributions dominantes des différentes intégrales et en utilisant les expressions approchées (3.23,

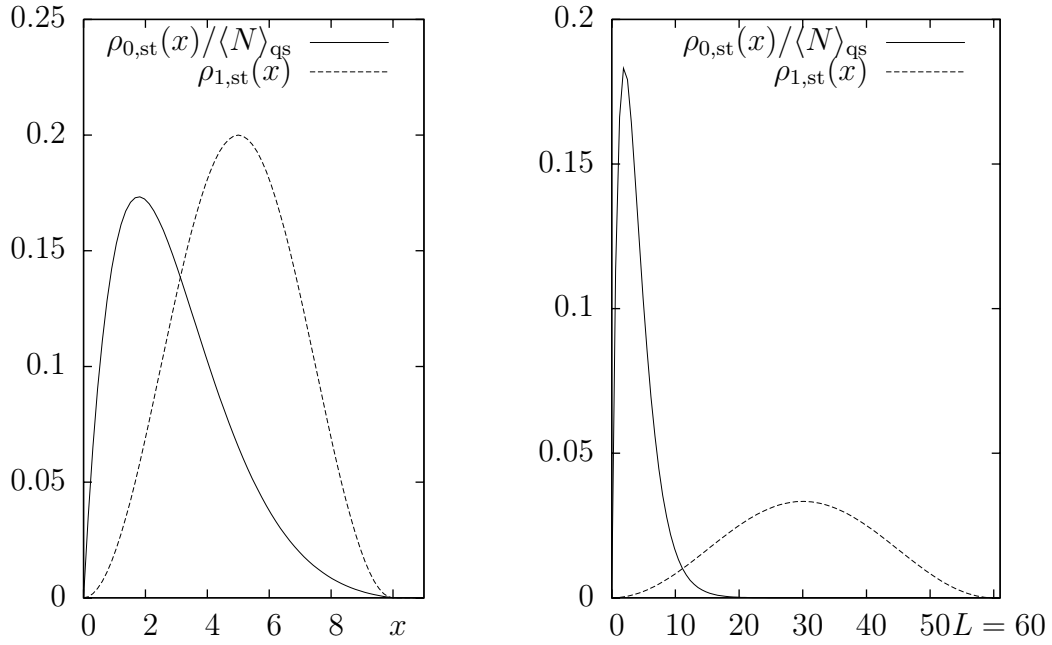


FIG. 6.1: Allure des densités $\rho_{1,\mathrm{st}}(x)$ et $\rho_{0,\mathrm{st}}(x)/\langle N\rangle_{\mathrm{qs}}$ pour v proche de v_c dans le régime quasi-stationnaire pour $L = 10$ (gauche) et $L = 60$ (droite) : lorsque l'on se rapproche de v_c, L augmente. L'individu A_0 est, en moyenne, au milieu du segment $[0, L]$ (distribution (6.3)) alors que les individus A_0 sont de plus en plus nombreux et de plus en plus près du mur en $x = 0$.

3.25) et la forme obtenue pour $\rho_{1,\mathrm{st}}$ ci-dessus, nous avons montré dans [SD08] que la densité $\rho_{0,\mathrm{st}}(x)$ prend la forme suivante dans la région I :

$$\boxed{\rho_{0,\mathrm{st}}(x) \underset{v\to v_c^-}{\simeq} K\frac{v_c^2}{4\pi L^2}\sin\left(\frac{\pi x}{L}\right)e^{-v_c(x-L)/2}, \qquad (x<L)} \tag{6.4}$$

et reste décroissante dans la région II, si bien que la contribution de cette dernière à la taille totale reste négligeable face à la contribution exponentiellement grande pour x petit. Les allures de $\rho_{1,\mathrm{st}}(x)$ et $\rho_{0,\mathrm{st}}(x)$ sont représentées en figure 6.1.1.

La constante K de la formule ci-dessous peut aussi être déterminée exactement (cf. [SD08]) et est donnée par :

$$K = \frac{32\pi^2 e^{\frac{1}{2}v_c B_c/A_c}}{A_c^3 v_c^4}\int_{-\infty}^{+\infty}[\partial_x Q_{v_c}(z)]^3 f_s''(Q_{v_c}(z))e^{-v_c z}dz. \tag{6.5}$$

La fonction $Q_{v_c}(x)$ est la solution de l'équation F-KPP sur la ligne infinie à la vitesse critique v_c à partir de laquelle tout un calcul perturbatif a été développé pour v proche de v_c (cf. fin du chapitre 2). C'est à partir des fronts Q_v qu'est obtenue la probabilité

d'extinction $Q_e^*(x)$. Les constantes A_c et B_c sont définies[1] par la forme asymptotique (2.11) de $Q_{v_c}(x)$ pour $x \to -\infty$. Ainsi, tous les termes entrant dans l'expression (6.5) de K peuvent être obtenus en étudiant uniquement *la forme d'un front stationnaire à la vitesse critique*. De même, la forme de $\rho_{0,\mathrm{qs}}$ dans la région II, bien que négligeable et non universelle, peut être exprimée à partir de Q_{v_c}.

À la constante multiplicative K près, l'équation (6.4) est universelle puisque les seules quantités qui apparaissent sont v_c et L, où la longueur L est définie par (3.24) et ne dépend que du taux de croissance $\beta = \sum_k \beta_k(k-1)$.

Malgré l'aspect négligeable de la région II en terme de contribution à la densité quasi-stationnaire, le calcul perturbatif à partir de la vitesse critique, en particulier la détermination des vecteurs et valeurs propres en sections 2.4.2 et 3.6, a utilisé de manière abondante les conditions de raccordement entre les deux régions I et II : les constantes multiplicatives de la densité (en particulier le facteur $1/L^2$ en (6.4)) sont intimement liées à la manière dont les raccordements entre les deux régions sont faits. De plus, les corrections aux ordres supérieurs dépendent explicitement de l'allure de Q_{v_c} dans la région II.

La densité quasi-stationnaire moyenne $\langle\rho(x)\rangle_{\mathrm{qs}}$ est donnée par (5.26) : les calculs précédents montrent que, dans la région I près du mur, elle est dominée par la contribution des particules A_0 qui sont exponentiellement plus nombreuses. Au voisinage de la vitesse critique, nous avons ainsi dans la région I la forme universelle suivante :

$$\boxed{\langle\rho(x)\rangle_{\mathrm{qs}} \simeq \rho_{0,\mathrm{st}}(x) \simeq K\frac{v_c^2}{4\pi L^2}\sin\left(\frac{\pi x}{L}\right)e^{-v_c(x-L)/2}}. \tag{6.6}$$

6.1.2 Taille quasi-stationnaire de la population

Singularité à la vitesse critique

La taille quasi-stationnaire moyenne $\langle N\rangle_{\mathrm{qs}}$ de la population est dominée par la contribution de la région I où le nombre de particules A_0 est exponentiellement grand. L'intégration de la densité $\rho_{1,\mathrm{st}}$ sur $[0,+\infty[$ donne 1 et est négligeable. On a ainsi :

$$\boxed{\langle N\rangle_{\mathrm{qs}} \simeq \frac{K}{L^3}e^{v_cL/2}} \tag{6.7}$$

où K est donnée par (6.5). Puisque $L \propto (v_c - v)^{-1/2}$ au voisinage du point critique, la divergence de N près de v_c a donc la forme singulière $(v_c - v)^{3/2}\exp(-C(v_c - v)^{-1/2})$.

Cette expression montre que la taille quasi-stationnaire se comporte de manière universelle puisqu'elle ne dépend que de L et est indépendante du détail des taux de division

[1] Les définitions exactes de Q_{v_c}, A_c et B_c faisaient intervenir une position arbitraire x_0 qui permettait de lever l'invariance par translation : on peut vérifier que l'expression (6.5) n'en dépend pas : une translation $x_0 \to x_0 + \delta$ entraîne, par définition de A_c et B_c, des variations $A_c \to A_c e^{-\delta v_c/2}$ et $B_c \to (B_c - A_c\delta)e^{-\delta v_c/2}$ et la translation de l'intégrale, via l'exponentielle, fait sortir un facteur $e^{-v_c\delta}$.

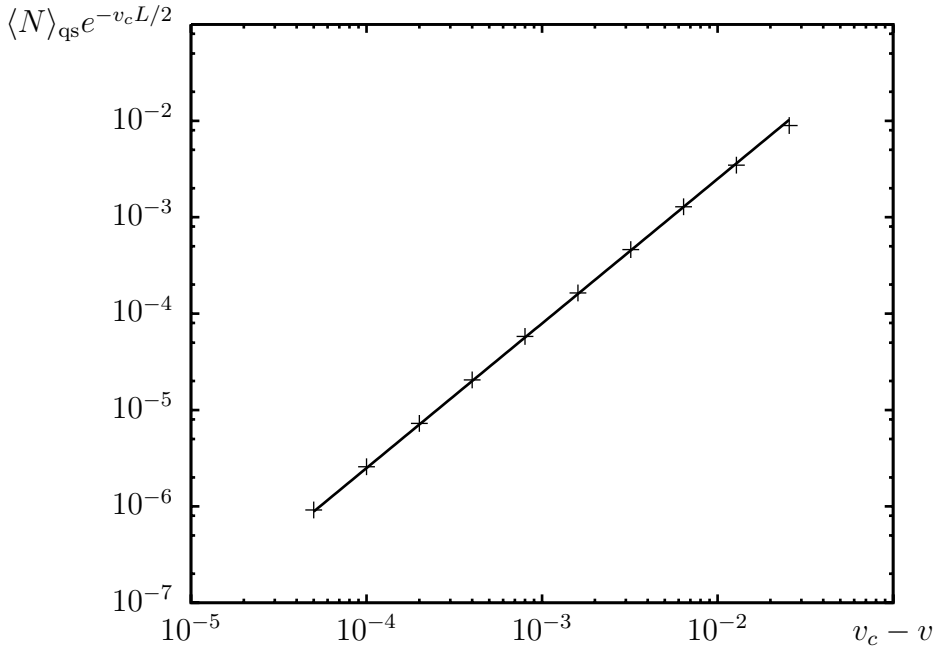


FIG. 6.2: Taille moyenne quasi-stationnaire corrigée par $e^{-v_c L/2}$: comparaison entre les données numériques (points) obtenues pour le modèle discret par la méthode de la section 4.4.1 et la formule (6.7) adaptée (ligne) au modèle discret (cf. section 3.3). La longueur L est calculée grâce à (3.24). L'accord entre les données numériques et la correction en $1/L^3$ prévue par (6.7) est excellent.

β_k. Seule l'amplitude K n'est pas universelle car elle est reliée à la forme du front $Q_{v_c}(x)$ à sa vitesse critique.

Nous ne nous sommes préoccupés jusqu'à présent que des valeurs moyennes quasi-stationnaires de la densité et de la taille, qui se révèlent être universelles au voisinage de la vitesse critique dans une région de taille L près du mur. Cette universalité se prolonge-t-elle à toute la mesure quasi-stationnaire ou bien seulement à certaines quantités bien particulières ? Les simulations numériques que nous avons présentées dans [DS07], facilement programmables pour un certain nombre de modèles discrets, semblent indiquer que le rapport $\langle N^2\rangle_{\rm qs}/\langle N\rangle^2_{\rm qs}$ tend vers 2 (cf. figure 6.3 pour le second moment) à la vitesse critique. Des simulations complémentaires (non représentées ici) semblent indiquer que le rapport $\langle N^3\rangle_{\rm qs}/\langle N\rangle^3_{\rm qs}$ se rapproche de 6. Bien que les simulations soient limitées à des valeurs de $v_c - v$ de l'ordre de 10^{-4} à cause de la divergence exponentielle (6.7), elles sont compatibles avec les résultats obtenus de manière exacte pour un modèle proche exactement soluble (cf. chapitre suivant) et tendraient ainsi à montrer que la distribution $P(x)$ de la taille de la population normalisée $x = N/\langle N\rangle_{\rm qs}$ serait exponentielle pour $v \to v_c$:

$$\mathrm{Prob}_{\rm qs}\left(\frac{N}{\langle N\rangle_{\rm qs}} = x\right) \stackrel{?}{\simeq} e^{-x}. \tag{6.8}$$

L'étude générale de cette distribution semble cependant pour l'instant hors de portée

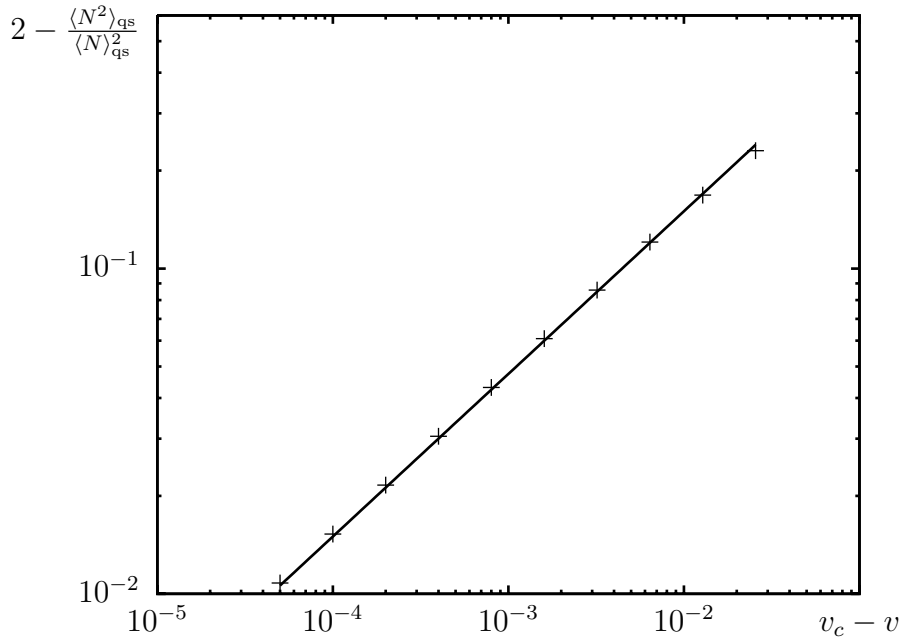


FIG. 6.3: Variance normalisée $\langle N^2\rangle_{\rm qs}/\langle N\rangle^2_{\rm qs}$ en fonction de $v_c - v$ pour $v < v_c$ dans le modèle discret mesurée par la méthode de la section 4.4.1. Les points représentent les données numériques et la droite correspond à l'ajustement $\langle N^2\rangle_{\rm qs}/\langle N\rangle^2_{\rm qs} \simeq 2 - C/L$ où C est une constante et L la longueur (3.24).

des outils développés ici mais il serait intéressant d'avoir une preuve mathématique de cette conjecture.

Comparaison avec la sélection à taille constante

Dans [BDMM06b], Brunet *et al.* étudient le problème d'une population qui diffuse et se divise avec la même loi que dans notre cas mais en présence de sélection interne, c'est-à-dire qu'à chaque pas de temps, seuls les N meilleurs individus sont sélectionnés (cf. chapitre 1). Dans ce cas, la population progresse vers les x positifs avec une vitesse *fluctuante* qui dépend de la taille *fixe* de la population N. Dans la limite de grande taille $N \to \infty$ et dans le régime stationnaire, la vitesse moyenne tend vers la même vitesse critique v_c que dans notre modèle. Néanmoins la convergence est lente et les effets de la taille finie N sur la vitesse moyenne de progression sont donnés par [BDMM06b] :

$$\langle v\rangle_N \simeq v_c - \frac{\pi^2 v''(\gamma_c)\gamma_c^2}{2\ln^2 N} + v''(\gamma_c)\pi^2\gamma_c^2\frac{3\ln\ln N}{\ln^3 N}. \tag{6.9}$$

Il est intéressant de remarquer que, dans la limite des grandes tailles, les expressions (6.7,6.9) qui relient, d'une part, $\langle N\rangle_{\rm qs}$ et v dans le régime quasi-stationnaire que nous avons étudié et, d'autre part, N et $\langle v\rangle_N$ dans ce modèle sont identiques aux ordres dominants ci-dessus. La raison de cette équivalence reste encore mystérieuse et il serait

intéressant de poursuivre cette étude afin de déterminer la généralité de ces relations entre taille et vitesse.

Les points communs dans la construction de ces deux modèles sont les suivants : les taux de division et la diffusion sont les mêmes et, dans les deux modèles, la taille est maintenue finie et non nulle aux temps longs. Au cours de l'analyse, les deux problèmes se ramènent à l'étude d'une équation de propagation de front, *avec bruit* dans le cas de la sélection à taille constante et *sans bruit mais avec condition aux bords nulle* dans le cas de la sélection avec mur. Dans les deux cas, c'est ce qui passe loin à l'avant du front (dans la phase instable) qui contrôle les propriétés du système (vitesse, temps de relaxation, etc.) mais, néanmoins, les similarités d'analyse sont difficiles à établir.

Brunet *et al.* ont développé dans [BDMM06b, BDMM07] une image phénoménologique qui prédit (6.9) en analysant le rôle des fluctuations à l'avant du front. Il serait intéressant d'essayer de voir ce qui peut être adapté à l'analyse du régime quasi-stationnaire que nous avons étudié. En particulier, les prédictions de [BDMM06b, BDMM07] concernent aussi le coefficient de diffusion D_N de la position $Y_N(t)$ du front[2] lorsque la taille est fixée à la valeur N et les moments d'ordre supérieur $\langle (Y_N(t) - Y_0)^p \rangle$ de la diffusion du front. Ces moments prennent, à l'ordre dominant, les valeurs universelles suivantes :

$$\begin{aligned} D_N &\simeq \gamma_c v''(\gamma_c) \frac{\pi^4}{3 \ln^3 N} + \dots \\ \frac{\langle (Y_N(t) - Y_0)^n \rangle}{t} &\simeq \gamma_c^{3-n} v''(\gamma_c) \frac{\pi^2 n! \zeta(n)}{\ln^3 N} + \dots \end{aligned}$$

où γ_c et $v''(\gamma_c)$ sont obtenus à partir de la relation de dispersion (2.9) en minimisant la vitesse par rapport à γ. D'autre part, dans le régime quasi-stationnaire de la marche aléatoire avec branchements en présence d'un mur absorbant, les résultats numériques semblent être universels pour les moments de la taille N de la population. Il serait intéressant d'approfondir l'analogie entre les deux modèles pour déterminer, par exemple, si la variance de la taille quasi-stationnaire est reliée au coefficient de diffusion ci-dessus.

6.1.3 Sélection adoucie

Dans les deux modes de sélection (mur ou taille constante), la sélection a lieu suffisamment loin à l'arrière de la population, où la densité de particules est suffisamment grande pour que l'on puisse espérer que les *détails* deviennent non pertinents. Un moyen de tester cette hypothèse, au moins numériquement, serait d'affaiblir le mode de sélection de la manière suivante :

- dans le cadre de la sélection interne, il serait possible de ne plus conserver exactement les N meilleurs mais d'utiliser, par exemple, une statistique de Fermi-Dirac à température non nulle ; nous nous attendrions alors à ce que la vitesse moyenne ne soit pas modifiée aux ordres dominants pour des températures suffisamment faibles.
- dans le cas du mur, on pourrait envisager de remplacer la condition abrupte d'élimination d'un individu par un taux de mort spontanée $\beta_0(x - vt)$ tel que $\beta_0(z) \to 0$

[2] Cette position peut être définie comme la position du centre de masse de la population de N individus ou la position médiane, ou encore la position du dernier individu : la différence entre ces quantités aux temps longs est négligeable aux ordres considérés.

lorsque $z \to +\infty$ et $\beta_0(z) \to +\infty$ pour $z \to -\infty$. Nous nous attendrions alors à ce que, si $\beta_0(z)$ a une décroissance suffisamment abrupte autour de 0, la taille quasi-stationnaire ne soit pas modifiée aux ordres dominants.

Cette question est encore ouverte pour nous et des simulations numériques sont envisagées pour tester ces conjectures.

6.2 Résultats numériques au-dessus de la vitesse critique pour un unique mur absorbant

Les résultats précédents ne sont valides que pour $v < v_c$. Au-delà de la vitesse critique, le régime conditionné est bien différent puisque les simulations numériques ne semblent indiquer aucun régime quasi-stationnaire (cf. section 4.4.1).

La difficulté analytique pour étudier le secteur $v > v_c$ provient du fait que $Q_e(x,t)$ ne tend pas vers $Q_e^*(x) = 1$ avec un temps de relaxation bien déterminé mais plutôt avec un comportement de type front. Toute l'analyse du chapitre 4, qui consistait à écrire qu'aux temps longs nous avions $Q_e = Q_e^* + \mathcal{A}\phi_1(x)e^{-t/\tau_1}$ puis à étudier l'amplitude $\mathcal{A}$, n'est plus valable. En se fondant sur les méthodes que nous avons présentées ici, nous pouvons songer à deux pistes principales pour aborder ce problème :

1. la première consisterait à reprendre l'*ansatz* de la section 3.5 où le rôle de l'amplitude $\mathcal{A}$ serait alors joué par la position du front $L_t - (v - v_c t)$, qui dépend elle aussi des conditions initiales ;
2. la seconde consisterait à considérer le processus modifié, qui reste toujours valable, puis à étudier la trajectoire de la particule A_1 entre $[0, T]$ (non stationnaire) pour remonter ensuite au profil des particules A_0.

6.3 Exploration numérique en dimension $d \geq 2$

En dimension $d \geq 2$, le calcul numérique expliqué en section 4.4.1 permet d'explorer la transition de phase vers l'état absorbant $Q_s^*(\vec{x}) = 0$. La figure 6.4 montre la probabilité de survie $Q_s^*(\vec{x})$ dans un modèle discret à deux dimensions au voisinage du point critique. Pour cela, nous avons considéré un modèle discret similaire à celui étudié en section 3.3 : les individus évoluent sur un réseau carré bidimensionnel ($\mathbb{N} \times \mathbb{N}$) limité au premier quadrant ($x \geq 0$ et $y \geq 0$). À chaque pas de temps, chaque individu se divise en deux individus puis, dans le même pas de temps, chaque enfant saute du site (x, y) à l'un des sites $(x+1, y)$, $(x+2, y)$, $(x-1, y)$, $(x-2, y)$ (mouvements horizontaux jusqu'au second voisin), $(x, y+1)$ et $(x, y-1)$ (mouvements verticaux) avec des probabilités respectives p_1, p_2, q_1, q_2, p_1', q_1' ou reste sur place avec une probabilité r.

Dès que l'une des coordonnées d'un individu devient négative ou nulle, l'individu meurt instantanément (bords absorbants en $x = 0$ et $y = 0$). Lorsque les probabilités de sauts varient, on observe également une transition de phase vers un état absorbant : si la dérive vers les bords absorbants est trop grande, l'extinction de la population est certaine. Nous nous sommes intéressés à l'étude numérique de la forme de la probabilité

de survie $Q_s^*(x,y)$ d'un individu situé initialement en (x,y) ainsi qu'à la taille quasi-stationnaire de la population .

Comme au chapitre 3, la forme de $Q_s^*(\vec{x})$ met en évidence l'existence de deux régions : la première (région I), proche de l'origine, correspond à une probabilité de survie $Q_s^*(\vec{x}) \ll 1$ et la seconde (région II) correspond à $Q_s^*(\vec{x}) \simeq 1$. Près du point critique, la taille de la première zone diverge. Néanmoins, l'étude du raccordement de $Q_s^*(x,y)$ n'est pas aussi simple qu'en dimension 1 puisque, par exemple, la courbe $Q_s^*(x) = 1/2$ n'est plus réduite à un point mais est une ligne dans le plan (x,y), comme le montre la figure 6.4 : il ne suffit plus de connaître les deux coefficients A_c et B_c introduits en (2.11) pour traiter le raccordement entre les deux régions.

En dimension 1, la taille quasi-stationnaire (6.7) diverge exponentiellement lorsque l'on se rapproche de la vitesse critique. Cette divergence s'observe aussi en dimension 2 lorsque la dérive $(-v_x, -v_y)$ s'approche de la courbe de dérive critique[3]. D'autre part, le rapport $\langle N^2\rangle_{\rm qs}/\langle N\rangle_{\rm qs}^2$ semble tendre vers 2 (non montré ici), comme dans le cas unidimensionnel.

Bien que l'étude soit plus difficile qu'en dimension 1, on peut s'attendre à ce que $Q_s^*(x,y)$ soit décrite par l'équation F-KPP linéarisée autour de 0 dans la région I près de l'origine : il faut alors s'attendre à ce que $Q_s^*(x)$ soit du type (3.23), à des coefficients prêts qui dépendent du raccordement à la partie non-linéaire. Bien que, pour $d = 2$, l'intégration des densités (5.20, 5.22) des particules A_0 et A_1 soit plus ardue, nous pouvons néanmoins nous attendre au même type de structure (6.4) en exponentielle-sinus dans la région I.

6.4 Régime quasi-stationnaire dans une boîte de taille constante à la vitesse critique

La situation présentée dans cette section est celle d'une marche aléatoire avec branchements qui évolue dans un domaine borné unidimensionnel de longueur l. La dérive et les taux de branchements sont les mêmes que précédemment.

Supposons que nous travaillions à dérive v et longueur l fixée. Le point critique a été étudié en section 3.7.2 du chapitre 3 et est caractérisé par (3.46). Dans cette section, nous avons également vu que les temps de relaxation et les vecteurs propres étaient semblables au-dessus et en-dessous du point critique. C'est pourquoi nous nous limiterons ici au cas $\beta < v^2/4 + \pi^2/l^2$ tel que nous soyons dans la phase inactive $Q_e^*(x) = 1$. D'après (3.51, 5.20), la particule A_1 a pour densité (formule exacte) :

$$\boxed{\rho_1(x) = \frac{2}{l}\sin^2\left(\frac{\pi x}{l}\right)}. \tag{6.10}$$

D'autre part, comme précédemment, le changement de variable $\rho_0(x) = e^{-vx}\psi(x)$ donne

[3] C'est-à-dire que, lorsque la vitesse tend vers la vitesse critique *à direction fixée*, nous obtenons à nouveau $\ln\langle N\rangle_{\rm qs} \propto (v_c - v)^{-1/2}$.

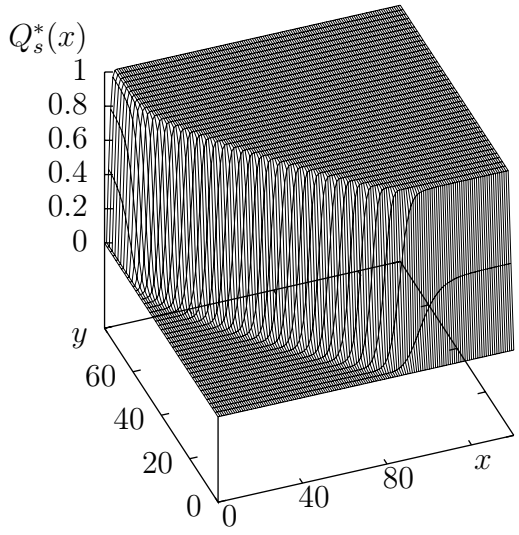


FIG. 6.4: Probabilité de survie dans un modèle discret à deux dimensions (cf. section 6.3). Les deux bords $x = 0$ et $y = 0$ sont absorbants et les valeurs des taux de sauts ($p_1 = 8,675.10^{-3}$, $p_1' = 12,15.10^{-3}$, $p_2 = 27,82.10^{-3}$, $r = 55,6.10^{-3}$, $q_1 = q_1' = 0.226$, $q_2 = 0.444$ sont choisies de manière à se placer près du point critique et correspondent à une dérive $(v_x, v_y) \simeq (0.25, -0.41)$. À deux dimensions, deux régions qui correspondent à $Q_s^*(\vec{x}) \ll 1$ et $Q_s^*(\vec{x}) \simeq 1$ se distinguent comme en dimension 1.

l'équation différentielle suivante sur ψ :

$$\partial_x^2\psi - v\partial_x\psi + \beta\psi = -f_e''(1)\frac{2}{l}\sin^2\left(\frac{\pi x}{l}\right)e^{vx} \tag{6.11}$$

avec les conditions aux bords :

$$\rho_0(0) = \rho_0(l) = \psi(0) = \psi(l) = 0. \tag{6.12}$$

La solution générale pour $\psi(x)$ est donnée par :

$$\begin{aligned}\psi(x) =& A_+e^{(v/2)x+i\pi x/L} + A_-e^{(v/2)x-i\pi x/L} \\ &+ \frac{f_e''(1)}{2l}\left[\frac{e^{vx+2i\pi x/l}}{\beta - 4\pi^2/l^2 + 2vi\pi/l} + \frac{e^{vx-2i\pi x/l}}{\beta - 4\pi^2/l^2 - 2vi\pi/l} - \frac{2e^{vx}}{\beta}\right].\end{aligned}$$

On notera, dans cette expression, l'intervention de deux longueurs : la longueur l du domaine considéré et la longueur minimale L dont a besoin la marche aléatoire avec branchements pour survivre (cf. définition (3.24) et équation (3.47)). Les conditions aux

bords fixent les valeurs de A_+ et A_- :

$$\begin{aligned} A_+ &= \frac{C}{2i\sin(\pi l/L)}\left(e^{vl/2} - e^{-i\pi l/L}\right), \\ A_- &= \frac{C}{2i\sin(\pi l/L)}\left(-e^{vl/2} + e^{i\pi l/L}\right), \\ C &\hat{=} \frac{f_e''(1)}{2l}\left[\frac{2}{\beta} - \frac{1}{\beta - 4\pi^2/l^2 + 2vi\pi/l} - \frac{1}{\beta - 4\pi^2/l^2 - 2vi\pi/l}\right] \end{aligned}$$

La densité $\rho_0(x)$ s'écrit donc de manière exacte :

$$\begin{aligned} \rho_0(x) = &\frac{C}{\sin(\pi l/L)}\left[e^{\frac{v}{2}(l-x)}\sin\left(\frac{\pi x}{L}\right) + e^{-\frac{v}{2}x}\sin\left(\frac{\pi(l-x)}{L}\right)\right] \\ &+\frac{f_e''(1)}{2l}\left[\frac{e^{2i\pi x/l}}{\beta - 4\pi^2/l^2 + 2vi\pi/l} + \frac{e^{-2i\pi x/l}}{\beta - 4\pi^2/l^2 - 2vi\pi/l} - \frac{2}{\beta}\right]. \end{aligned} \tag{6.13}$$

Près du point critique, C tend vers une valeur finie et le seul terme divergent dans cette équation est le dénominateur $\sin(\pi l/L)$. À l'intérieur du domaine $[0, l]$, la densité quasi-stationnaire de particules est alors dominée par :

$$\langle\rho(x)\rangle_{\rm qs} \simeq \frac{C}{\pi(L-l)}\left[e^{\frac{v}{2}(l-x)}\sin\left(\frac{\pi x}{L}\right) + e^{-\frac{v}{2}x}\sin\left(\frac{\pi(l-x)}{L}\right)\right] \tag{6.14}$$

$$\simeq \frac{C}{\pi(L-l)}e^{-vx/2}\sin\left(\frac{\pi x}{L}\right)\left[e^{vl/2} - 1\right]. \tag{6.15}$$

Il s'ensuit que la taille moyenne quasi-stationnaire est obtenue en intégrant $\langle\rho(x)\rangle_{\rm qs}$ sur $[0, l]$ et se comporte comme :

$$\boxed{\langle N\rangle_{\rm qs} \simeq \frac{2C\sinh(vl/2)}{\beta l}\frac{1}{L-l} \propto (L-l)^{-1}}. \tag{6.16}$$

La taille quasi-stationnaire diverge ainsi au point critique comme l'inverse de l'écart entre la longueur du domaine l et la longueur L minimale de survie.

Contrairement à (6.7) pour la ligne semi-infinie, la formule précédente n'exhibe pas de divergence exponentielle puisqu'ici le domaine où les particules sont confinées est de taille l fixée. De plus la constante C n'est pas de même nature que la constante K de l'équation (6.5) puisque cette dernière nécessite de connaître un front critique sur la ligne infinie alors que la constante C n'est liée qu'aux propriétés de f_e près de 1 (linéarisation de l'équation F-KPP).

Il serait intéressant d'étudier le *cross-over* entre la ligne semi-infinie et le segment de taille finie fixée : pour cela, il faudrait considérer le cas d'un segment de taille finie l et faire varier l en même temps que v pour v proche de v_c. En effet, on s'attend à ce que la partie non-linéaire du front se manifeste lorsque $l > L \propto (v_c - v)^{-1/2}$ et qu'il ne soit plus possible de linéariser l'équation F-KPP sur tout l'intervalle $[0, l]$ comme cela a été fait ici.

Chapitre 7

Un cas particulier exactement soluble : le modèle exponentiel

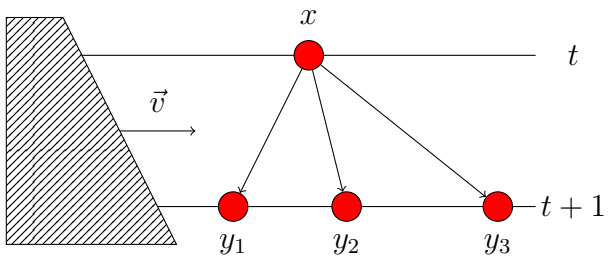
Les chapitres précédents ont montré que certaines propriétés du régime quasi-stationnaire sont universelles autour de la vitesse critique et des simulations numériques semblent prouver que cette universalité est plus large. Nous nous proposons dans ce chapitre d'étudier un modèle particulier, le modèle exponentiel, qui a l'avantage de pouvoir être résolu exactement et de partager un certain nombre de similarités avec les modèles étudiés précédemment. Ce chapitre reprend les résultats que nous avons développés dans la dernière section de [SD08].

7.1 Définition

7.1.1 Classe de modèles

Cette classe de modèle, introduite dans [BDMM06a, BDMM07], peut, par l'aspect non-local de la diffusion des individus et le grand nombre d'enfants potentiellement généré par un individu, être critiquée du point de vue de la modélisation biologique. Néanmoins, elle est reliée à d'autres modèles de physique statistique, comme les polymères dirigés [BD04] et donne lieu à un modèle exactement soluble.

Les individus sont à nouveau repérés par une unique coordonnée réelle x. Le temps est discret si bien qu'un individu à la génération t est remplacé par l'ensemble de ses enfants à la génération $t+1$. Ici diffusion et division sont mêlées puisque les enfants d'un individu sont distribués selon un *processus ponctuel de Poisson* : pour tout intervalle $[y, y+dy]$, un parent situé à la position x a un enfant dans cet intervalle avec une probabilité $\psi(y-x)dy$. Le nombre d'enfants d'un individu sur un intervalle donné I est ainsi une variable aléatoire poissonienne de paramètre $\int_I \psi(y-x)dy$. En particulier, la probabilité que, sur un intervalle $I = [a, b]$ donné, il y ait n enfants situés en $y_1 < \ldots < y_n$ est



Probabilité $= \psi(y_1+v-x)\psi(y_2+v-x)\psi(y_3+v-x)e^{-\int_0^\infty \psi(y+v-x)dy}$

FIG. 7.1: Reproduction en présence d'un mur se déplaçant à la vitesse v dans le modèle exponentiel $\psi(z) = e^{-z}$ (les coordonnées des individus sont définies par rapport au mur).

donnée par :

$$P_{[a,b]}(y_1 < \ldots < y_n | x) = \psi(y_1 - x) \ldots \psi(y_n - x) e^{-\int_a^b \psi(y-x)dy}. \tag{7.1}$$

La sélection élimine les individus avec les positions x les plus faibles, c'est-à-dire ceux qui ont une position inférieure à un seuil ξ donné. Selon le type de sélection (interne ou externe), ce seuil est la position de la N-ième particule ou celle du mur et met une borne inférieure aux positions possibles des enfants d'un individu. D'autre part, en sélection interne, afin que la position de la N-ième particule soit bien définie, il faut s'assurer que la reproduction produise une population de taille supérieure à N. Pour cela nous supposerons que ψ n'est pas intégrable au voisinage de $-\infty$. Dans le cas d'une sélection avec seuil, cette hypothèse peut être levée. De plus, afin que la taille ne diverge pas d'une génération à la suivante, nous supposerons donc que la fonction ψ est intégrable au voisinage de $+\infty$.

Dans le cas d'un mur absorbant qui se déplace à la vitesse v vers les x positifs, cette dynamique induit l'évolution suivante pour la probabilité d'extinction $Q_e(x,t)$ de la lignée d'un individu initial situé à une distance x du mur :

$$\begin{aligned} Q_e(x,t+1) &= \sum_{n=0}^{+\infty} \int_{0<y_1<\ldots<y_n} dy_1 \ldots dy_n \prod_{i=1}^{n} \big(Q_e(y_i,t)\psi(y_i+v-x)\big) e^{-\int_0^\infty \psi(y+v-x)dy} \\ &= \exp\left[-\int_0^{+\infty} \psi(y+v-x)(1-Q_e(y,t))dy\right]. \end{aligned} \tag{7.2}$$

Cette équation joue le rôle de l'équation F-KPP pour la dynamique de la probabilité de survie $Q_e(x,t)$.

7.1.2 Quelques propriétés remarquables du modèle exponentiel

Le modèle exponentiel correspond à une distribution $\psi(x) = e^{-x}$, qui est intégrable en $+\infty$ et diverge en $-\infty$. Son avantage majeur est la factorisation permise entre les positions y_i des enfants et la position x des parents. En particulier, l'évolution (7.2) de

la probabilité de survie $Q_e(x,t)$ se simplifie à $t > 1$ pour ne dépendre que d'un unique paramètre $c(t)$:

$$Q_e(x,t) = \exp\left[-c(t)e^x\right] \tag{7.3a}$$

$$c(t+1) = e^{-v}\int_0^\infty e^{-y}(1-e^{-c(t)e^y})dy \tag{7.3b}$$

Aux temps courts, nous avons $c(0) = \infty$ et $c(1) = e^{-v}$. Cette propriété sera largement utilisée par la suite pour étudier le régime quasi-stationnaire.

Une seconde propriété remarquable jouera un rôle dans la partie suivante : il s'agit du découplage entre les générations. En effet, la probabilité de passer d'une population $0 < x_1 < \ldots < x_n$ à une population d'enfants $0 < y_1 < \ldots < y_m$ est donnée par la somme suivante :

$$\text{Prob}\left(0 < y_1 < \ldots < y_m \middle| 0 < x_1 < \ldots < x_n\right) = \sum_{\wp:\{1,\ldots,m\}\to\{1,\ldots,n\}} \prod_{i=1}^m \psi(y_i - x_{\wp(i)}) \prod_{j=1}^m e^{-\int_0^\infty \psi(y-x_j)dy}$$

où $\wp(i)$ est le numéro du parent de l'individu i dans la génération précédente. La structure du modèle exponentiel $\psi(x) = e^{-x}$ permet ainsi de l'écrire sous la forme suivante :

$$\text{Prob}\left(0 < y_1 < \ldots < y_m \middle| 0 < x_1 < \ldots < x_n\right) = \left(\prod_{i=1}^m e^{-(y_i - X)}\right) e^{-\int_0^\infty e^{-(y-X)}dy} \tag{7.4}$$

où X est une position effective définie par :

$$\boxed{e^X = \sum_{i=1}^n e^{x_i}}. \tag{7.5}$$

La forme (7.4) montre que tirer m enfants en $y_1,\ldots,y_m$ de n parents à des positions $x_1,\ldots,x_n$ est identique en probabilité à tirer m enfants pour un unique parent effectif à la position X définie par (7.5). Ensuite, pour établir la parenté, il suffit pour chaque enfant à $t+1$ de choisir son parent i dans la génération t avec un poids $e^{x_i}/\sum_{j=1}^n e^{x_j}$.

7.2 Résultats en présence d'un mur absorbant

7.2.1 Probabilité de survie

La dynamique (7.3) montre que la probabilité d'extinction tend, pour $t \to +\infty$, vers

$$Q_e^*(x) = e^{-c_* e^x} \tag{7.6}$$

où le coefficient c_* est le point fixe de l'application h suivante :

$$h(c) = e^{-v}\int_0^\infty e^{-y}(1-e^{-ce^y})dy = e^{-v}\left(1 - \int_0^\infty e^{-y-ce^y}dy\right) \tag{7.7}$$

L'étude de la fonction h montre que, pour toute vitesse v finie, il existe une unique solution stable $c_* > 0$. Par conséquent, le modèle exponentiel n'exhibe pas de transition vers une phase absorbante à une vitesse finie. Néanmoins, pour $v \to \infty$, le point fixe c_* tend vers 0, de telle sorte que :

$$Q_e^*(x) \xrightarrow{v\to+\infty} 1. \tag{7.8}$$

Pour être précis, h n'est pas dérivable en 0 et s'annule comme $c \ln c$; le développement de h au voisinage de 0 aux ordres suivants

$$h(c) = e^{-v}\left(-c\ln c + (1-\gamma_E)c + \frac{c^2}{2} - \frac{c^3}{12} + O(c^4)\right) \tag{7.9}$$

montre que c_* converge vers 0 comme :

$$c_* \underset{v\to+\infty}{\simeq} e^{1-\gamma_E}\exp(-e^v), \tag{7.10}$$

où γ_E désigne la constante d'Euler[1]. D'autre part, la relaxation de $c(t)$ vers c_* est caractérisée par un temps de relaxation $e^{-1/\tau_1} = \Lambda = h'(c_*)$. Dans la limite $v \to \infty$, Λ se comporte comme :

$$\Lambda = h'(c_*) \simeq 1 - e^{-v}. \tag{7.11}$$

La forme (7.6) de la probabilité d'extinction montre l'existence de deux régions (cf. figure 7.2) comme dans le cas de la marche aléatoire avec branchements :

1. une région I de longueur $L = -\ln c_*$ dans laquelle $c_* e^x \simeq 0$ et $Q_e^*(x) \simeq 1$, qui correspond à une région où les individus s'éteignent avec une forte probabilité ;
2. une région II au-delà de L où $Q_e^*(x) \ll 1$ décroît comme une double exponentielle et dans laquelle un individu a une forte probabilité de survie aux temps longs.

La longueur L de la région I diverge pour $v \to \infty$:

$$\boxed{L = -\ln c_* \underset{v\to\infty}{\simeq} e^v}. \tag{7.12}$$

Il faut néanmoins noter que, dans la région I, nous avons $Q_e^*(x) \simeq 1 - c_* e^x$: il n'y a pas de correction sinusoïdale, contrairement à (3.23), et la probabilité d'extinction ne vaut pas rigoureusement 1 en $x = 0$. Cela s'explique par le caractère non-local de la reproduction (fonction ψ) : un individu loin devant le mur peut produire à la génération suivante un grand nombre d'enfants à proximité du mur.

[1] Elle est définie par la limite suivante :

$$\gamma_E = \lim_{n\to+\infty}\left(\sum_{k=1}^{n}\frac{1}{k} - \ln n\right) \simeq 0,577215665\ldots$$

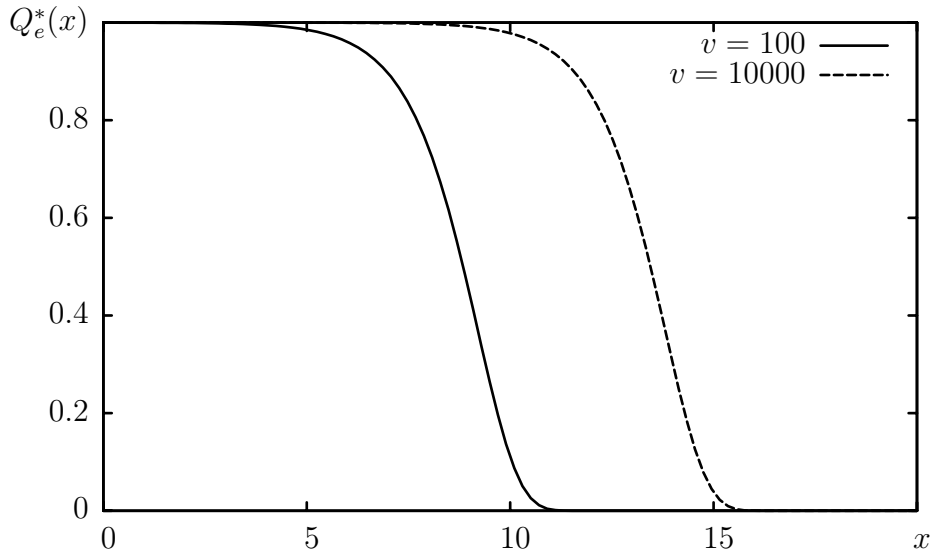


FIG. 7.2: Allure de la probabilité d'extinction aux temps longs $Q_e^*(x)$ d'un individu initialement situé à une distance x du mur.

7.2.2 Régime quasi-stationnaire

La réduction à un paramètre (7.3) de la dynamique (7.2) rend l'étude du régime quasi-stationnaire similaire à celle du processus de Galton-Watson faite au chapitre 4. En effet, la dynamique (7.3) implique pour $t \to \infty$:

$$\begin{aligned} Q_e(x,t) &= \exp\left[-e^x\left(c_* + \Lambda^{t-1}\mathcal{B}(c(1)) + \ldots\right)\right] \\ &= e^{-c_* e^x} + \Lambda^{t-1}\mathcal{B}\big(c(1)\big)\left(-e^{x-c_* e^x}\right) + o(\Lambda^t). \end{aligned} \tag{7.13}$$

Cette équation décrit également le comportement aux temps longs des fonctions génératrices $G_1(x, g_1; t)$ et $G_2(x, g_1, g_2; t, t')$ définies par (4.2), puisque celles-ci suivent la même équation d'évolution que $Q_e(x,t)$ (cf. chapitre 4), aux conditions initiales près. La différence se situe ainsi dans l'amplitude $\mathcal{B}(c(1))$ qui dépend à présent de la condition initiale $G_1(x, g_1; 0)$ via le coefficient $c(1)$. Ce coefficient $c(1)$ est obtenu en itérant une fois la fonctionnelle (7.2) sur la condition initiale $G_1(y, g_1, 0) = e^{-g_1(x)}$ et dépend de celle-ci comme :

$$c(1) \hat{=} C(e^{-g_1(x)}) = e^{-v}\int_0^{+\infty} e^{-y}\left(1 - e^{-g_1(x)}\right) dy. \tag{7.14}$$

L'équation (7.13) est similaire à (4.8), à condition de définir la fonctionnelle $\mathcal{A}$ par

$$\mathcal{A}(Q) = \mathcal{B}(C(Q))/\Lambda$$

et le vecteur propre $\phi_1(x)$ par

$$\phi_1(x) = -e^x e^{-\exp(x)}.$$

Toutes les propriétés du régime quasi-stationnaire sont ensuite données par l'équation (4.30). En particulier (cf. [SD08] pour le détail des calculs), la densité de particules au point x est donnée par :

$$\boxed{\langle \rho(x) \rangle_{\rm qs} = \frac{e^{-c_* e^x}}{\Lambda} + \big(- \mathcal{B}''(c_*) \big) e^{-v} e^{-x - c_* e^{-x}}} \,. \tag{7.15}$$

Ce profil quasi-stationnaire exhibe aussi deux régions I et II correspondant aux régions où $Q_e^*(x) \simeq 1$ et $Q_e^*(x) \ll 1$. Nous remarquons que, comme la marche aléatoire avec branchements, les individus sont confinés dans la région I de taille L.

7.2.3 Au voisinage du point critique

Pour étudier le régime quasi-stationnaire près du point critique, il reste à déterminer les propriétés de la fonction $\mathcal{B}$ pour $v \to \infty$. Par définition, $\mathcal{B}(c)$ satisfait $\Lambda \mathcal{B}(c) = \mathcal{B}(h(c))$ (cf. section 4.2, équation (4.10)). En dérivant cette équation en $c = c_*$, nous obtenons récursivement les dérivées de $\mathcal{B}$ en c_* en fonction de celles de la fonction d'itération h. Pour la dérivée seconde intervenant dans (7.15), nous obtenons asymptotiquement :

$$\mathcal{B}''(c_*) \simeq -\frac{1}{c_*}. \tag{7.16}$$

Au voisinage du point critique, la densité quasi-stationnaire est alors dominée par le terme $e^{-v} e^{-x - c_* e^{-x}} / c_*$ et la taille moyenne quasi-stationnaire diverge comme :

$$\boxed{\langle N \rangle_{\rm qs} \simeq \frac{e^{-v}}{c_*} \simeq \frac{e^L}{L}} \,. \tag{7.17}$$

Cette équation contient le même type de divergence exponentielle que la marche aléatoire avec branchements en présence d'un mur absorbant (cf. équations (6.4) et (6.7) du chapitre précédent). Néanmoins la correction en $1/L^3$ est ici remplacée par une correction en $1/L$.

Dans ce modèle particulier, il est possible d'aller plus loin et d'obtenir toute la distribution de la taille quasi-stationnaire. La fonction génératrice de la taille s'obtient directement en prenant $g_1 = \mu$ dans (4.30) et devient ici (cf. [SD08]) :

$$\langle e^{-\mu N} \rangle_{\rm qs} = e^{-\mu} \mathcal{B}' \left(c_* + (1 - e^{-\mu}) J_0 \right) \tag{7.18}$$

avec $J_0 = e^{-v} \int_0^\infty e^{-y - c_* e^y} dy \simeq e^{-v}$. Dans la limite $v \to \infty$, il faut considérer μ d'ordre $1/\langle N \rangle_{\rm qs}$, *i.e.* d'ordre c_*/e^{-v}. Ainsi, nous avons besoin de connaître $\mathcal{B}$ dans une région de taille c_* autour du point fixe c_*. Nous avons montré dans [SD08], à partir du développement de h (7.9), que $\mathcal{B}(c_* + c_* u) \to \ln(1+u)$ lorsque $v \to \infty$ et u reste d'ordre 1. La fonction génératrice (7.18) devient ainsi dans la limite $v \to v_c$:

$$\langle e^{-\mu (N / \langle N \rangle_{\rm qs})} \rangle_{\rm qs} \to \frac{1}{1+\mu}. \tag{7.19}$$

Cette fonction génératrice correspond à une distribution exponentielle de la taille quasi-stationnaire de la population. En corollaire immédiat, nous déduisons les moments suivants :

$$\begin{aligned} \frac{\langle N^2 \rangle_{\rm qs}}{\langle N \rangle_{\rm qs}^2} &\simeq 2, \\ \frac{\langle N^3 \rangle_{\rm qs}}{\langle N \rangle_{\rm qs}^3} &\simeq 6. \end{aligned}$$

Ces valeurs sont comparables à celles que nous avons obtenues numériquement pour une marche aléatoire avec branchements (cf. figure 4.1). Pour le processus de Galton-Watson (sans structure spatiale), la distribution de la taille quasi-stationnaire normalisée est universelle et est donnée au contraire par $4xe^{-2x}$ (cf. (4.21)) : l'annulation de cette distribution en $x = 0$ s'explique par le fait que le système ne doit pas approcher de l'état absorbant, *i.e.* de $x = 0$, dans le régime quasi-stationnaire. Dans le cas spatial qui nous occupe, au contraire, cette taille est plus libre de fluctuer autour des tailles petites, pourvu que les quelques individus ne s'approchent pas du mur absorbant.

Le modèle exponentiel est exactement soluble car l'étude de la fonctionnelle $\mathcal{A}$ se ramène à l'étude d'une fonction $\mathcal{B}$ pour laquelle les développements limités près du point critique sont bien contrôlés. Comme nous l'avons déjà dit au chapitre 4, il serait intéressant de contrôler proprement les développements de Taylor de la fonctionnelle $\mathcal{A}$ dans une démarche similaire à celle que nous avons suivie pour la fonction $\mathcal{B}$.

Nous reviendrons sur ce modèle dans la partie suivante où les généalogies seront traitées en détail et montrent aussi des similarités avec d'autres modèles de sélection.

Deuxième partie

Temps de coalescence et généalogies

Chapitre 8

Modèles de coalescence en champ moyen

Ce chapitre est une introduction aux processus de coalescence, auxquels nous allons nous intéresser dans toute la suite de cette partie. Après une courte description de nos motivations, une section est consacrée à la présentation de la théorie mathématique du Λ-coalescent, qui permet de décrire une large classe de modèles de coalescence en champ moyen. Les deux dernières sections décrivent deux cas particuliers de Λ-coalescent, le coalescent de Kingman et celui de Bolthausen-Sznitman, qui jouent un rôle particulier dans les chapitres ultérieurs. Bien qu'il ne contienne pas de contribution originale, ce chapitre introduit les notions et notations à la base de nos travaux présentés dans les trois chapitres ultérieurs.

8.1 Introduction

Le point de vue des modèles de coalescence est complémentaire du point de vue des marches aléatoires avec branchements étudiées dans la partie précédente. Ces dernières correspondent à un processus de croissance, de division et de mort d'individus. Cette évolution introduit des relations de parenté entre les individus qui survivent : les modèles de coalescence visent à étudier ces relations en se focalisant sur la structure généalogique.

Dans un processus de branchement tel que ceux décrits dans la partie précédente, cela consiste à prendre des individus à un temps arbitraire t et à regarder, *dans leur passé*, les évolutions qu'ils ont suivies depuis qu'ils se sont séparés de leur ancêtre commun le plus récent. Pour tout groupe de n individus, une généalogie (cf. figure 8.1) est caractérisée par :

1. l'*arbre* des parentés (topologie) construit de la manière suivante : tout individu est relié à son parent (lignée) jusqu'à atteindre un ancêtre commun à tout le groupe. Les *noeuds* de l'arbre correspondent ainsi aux moments où des lignées se sont séparées.

2. les longueurs des branches, ou *temps de coalescence*, qui sont les intervalles de temps entre deux séparations de lignées.

Le terme de *coalescence* provient du fait que nous ne regardons pas l'évolution future des lignées d'individus qui se reproduisent mais que nous nous tournons vers le passé : en remontant les parentés des individus, des lignées distinctes aboutissent généralement à des ancêtres communs et fusionnent en lignées communes à plusieurs individus. Le terme de *coalescence* permet un contraste avec le processus de *reproduction* suivant la direction temporelle dans laquelle nous regardons.

Nous appellerons T_n l'âge de l'*ancêtre commun le plus récent* (MRCA pour *most recent common ancestor* dans la littérature anglophone) d'un groupe de n individus, c'est-à-dire la durée entre l'instant présent (génération où vivent les n individus considérés) et la génération où leurs lignées se sont séparées pour la première fois. Remonter les lignées revient à remonter l'histoire des individus. Contrairement à la partie précédente, nous ne nous intéressons plus à l'évolution globale de la population mais à ses conséquences sur les lignées de petits groupes.

L'intérêt d'étudier les temps T_n est multiple. Considérons tout d'abord un exemple issu de la génétique. Les individus sont caractérisés par des séquences ADN que des mutations peuvent affecter. La diversité génétique dans une population est le fruit des mutations qui ont affecté les lignées depuis l'ancêtre commun le plus récent de la population. D'autre part, l'âge T de l'ancêtre commun le plus récent de toute la population correspond à l'échelle de temps de décorrélation de la diversité génétique : en effet, la population au temps t descend d'un unique individu de la population au temps $t-T$ est n'est donc pas sensible à la diversité génétique antérieure à cette date-ci. Cet âge T est aussi révélateur de la compétition entre individus et révèle certains aspects de la sélection. En effet, dans le cas où une mutation bénéfique affecte un individu, sa lignée se reproduit plus vite que les autres et régénère dans un délai plus court la population ; ainsi l'individu muté devient rapidement l'ancêtre commun le plus récent de la population. La relation entre sélection et généalogies est étudiée en section 8.4.

De manière assez générale, si l'âge de l'ancêtre commun le plus récent d'une population est petit, peu de mutations ont eu le temps de survenir dans les lignées et on s'attend à observer une population relativement homogène : l'objectif du chapitre 11 est d'explorer dans quelques exemples simples la corrélation à laquelle nous nous attendons entre généalogie et diversité génétique. De plus, dans un cadre biologique, hormis l'étude des fossiles, il n'est pas possible d'accéder expérimentalement aux données généalogiques directement et nous avons recours à des modèles probabilistes [Cha99, Kin82b, Ewe04] ; néanmoins, il est facile de prendre des échantillons de la population à l'instant présent et d'étudier leur diversité génétique (séquençage de génomes) : pour retracer les relations de parenté, l'étude des corrélations entre temps de coalescences et diversité est nécessaire [Wak07, TBGD97, GL05, FL97]. De nombreuses études et de nombreux modèles existent en absence de sélection [Don91, HSW05] mais relativement peu en présence de sélection [KDH88, BB03].

Au-delà de la biologie, un certain nombre de processus physiques peuvent être reformulés en termes de coalescences et de généalogies : c'est le cas, par exemple, des polymères

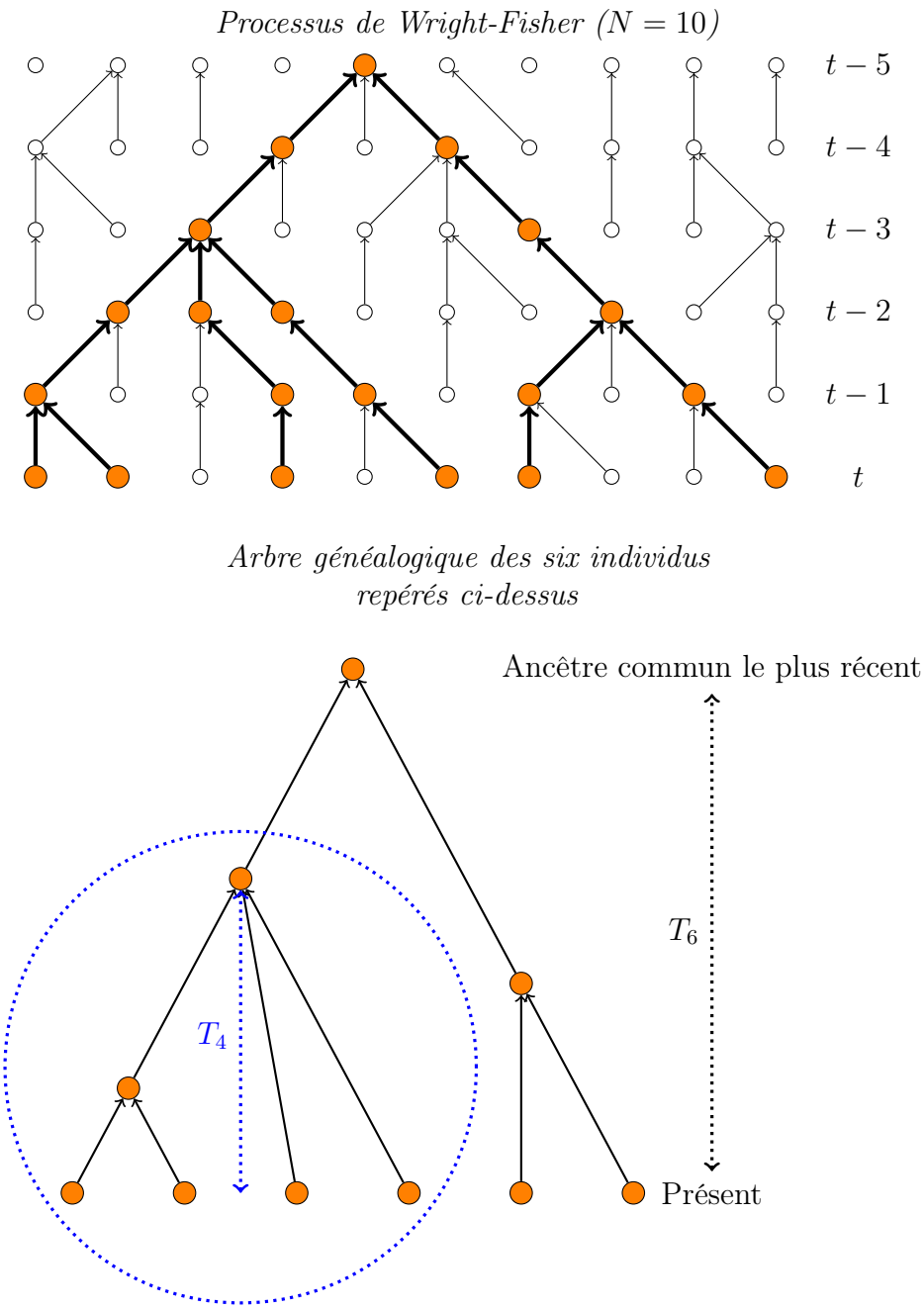


FIG. 8.1: Arbre généalogique d'un groupe de six individus dans une population de dix individus. En haut, processus de Wright-Fisher (cf. section 1.2.2) pour $N = 10$ individus : nous nous intéressons plus particulièrement à la généalogie des six individus marqués par des cercles pleins dans la génération présente. En bas, généalogie épurée de ces six individus. Une généalogie est caractérisée par la forme topologique de l'arbre et par les longueurs des branches (temps de coalescence). T_6 désigne l'âge de l'ancêtre commun le plus récent des 6 individus, tandis que T_4 désigne l'âge de l'ancêtre commun le plus récent des quatre premiers individus seulement ($T_6 = 5$ et $T_4 = 3$ dans l'exemple considéré).

dirigés en physique statistique (cf. chapitre 9) ou des chocs dans l'équation de Burgers [SAF92, Sin92, Gir01]. L'un des buts est alors de chercher des classes d'universalité dans les généalogies qui apparaissent dans les différents modèles.

8.2 Le champ moyen : le Λ-coalescent

8.2.1 Définition

Nous allons tout d'abord nous intéresser à des modèles de *champ moyen*, c'est-à-dire à des modèles sans structure spatiale dans lesquels tous les individus peuvent subir une coalescence avec tous les autres. Considérons un groupe de b individus dans une population de taille infinie à un instant t. Nous supposerons que, lorsque nous remontons d'un pas de temps dt dans le passé, chaque sous-groupe de k individus parmi les b a une probabilité $\lambda_{b,k}dt$ de subir une coalescence qui le réduit à un unique individu. Nous nous intéressons alors aux propriétés de l'arbre généalogique de ces b individus. Nous supposerons ici qu'à chaque pas de temps dt, une seule coalescence à k individus peut se produire et nous ne parlerons pas des généralisations où des coalescences multiples simultanées peuvent se produire [Möh06, Sch00].

Les coefficients $\lambda_{b,k}$ ne peuvent pas être pris de manière quelconque. En effet, tout groupe de b individus peut être considéré comme faisant partie d'un groupe de taille $b+1$ en ajoutant un individu pris dans le reste de la population. Ainsi, du point de vue du groupe de taille $b+1$, pour qu'un sous-groupe donné de taille k pris parmi les b premiers subisse une coalescence, il faut soit que ces mêmes k individus subissent une coalescence dans le groupe élargi de taille $b+1$, soit qu'ils subissent une coalescence commune (à $k+1$ individus donc) avec le $(b+1)$-ème individu du groupe élargi. Les taux $\lambda_{b,k}$ doivent ainsi satisfaire l'équation de récurrence suivante :

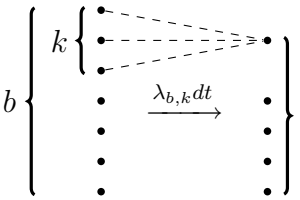


FIG. 8.2: Définition du taux de coalescence $\lambda_{b,k}$ avec lequel un sous-groupe de taille k subit une coalescence.

$$\lambda_{b,k} = \lambda_{b+1,k} + \lambda_{b+1,k+1}. \tag{8.1}$$

Pitman et Sagitov ont montré [Pit99, Sag99] que cette récurrence implique l'existence d'une mesure Λ sur $[0,1]$ telle que les taux $\lambda_{b,k}$ se mettent sous la forme :

$$\boxed{\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k}\Lambda(x)dx}. \tag{8.2}$$

La mesure $\Lambda(x)$ résume ainsi les degrés de liberté disponibles dans le choix des taux $\lambda_{b,k}$. Toutes les quantités pertinentes qui concernent les généalogies peuvent être exprimées en fonction de cette mesure. Par exemple, la probabilité $\gamma_b dt$ qu'une coalescence ait lieu

dans un groupe de b individus pendant un intervalle de temps dt est donnée par :

$$\gamma_b = \sum_{k=2}^{b} C_b^k \lambda_{b,k} = \int_0^1 \frac{1}{x^2} \left[1 - (1-x)^b - bx(1-x)^{b-1}\right] \Lambda(x) dx \tag{8.3}$$

où C_b^k désigne le coefficient binomial $b!/(k!(b-k)!)$.

Le temps moyen de coalescence T_2 de deux individus est alors donné par :

$$\langle T_2 \rangle = \frac{1}{\lambda_{2,2}} = \left(\int_0^1 \Lambda(x) dx \right)^{-1}. \tag{8.4}$$

Nous voyons ainsi que la normalisation $\int_0^1 \Lambda(x)dx$ donne l'échelle de temps des processus de coalescence. Le temps moyen de coalescence T_3 de trois individus contient deux contributions : soit les trois individus subissent une coalescence simultanée au même pas de temps avec une probabilité $\lambda_{3,3}dt$, soit deux d'entre eux seulement subissent d'abord une coalescence (trois possibilités avec chacune une probabilité $\lambda_{3,2}dt$) puis l'individu ainsi obtenu subit ensuite une coalescence avec le troisième. Ainsi, nous avons :

$$\langle T_3 \rangle = \frac{1 + 3\lambda_{3,2}\langle T_2 \rangle}{\lambda_{3,3} + 3\lambda_{3,2}}. \tag{8.5}$$

Le rapport $\langle T_3 \rangle / \langle T_2 \rangle$, sur lequel nous reviendrons plus loin, est alors indépendant de l'échelle de temps des processus de coalescence puisqu'il est donné par le rapport

$$\boxed{\frac{\langle T_3 \rangle}{\langle T_2 \rangle} = \frac{\lambda_{2,2} + 3\lambda_{3,2}}{\lambda_{3,3} + 3\lambda_{3,2}}}. \tag{8.6}$$

qui est indépendant de la normalisation de la mesure Λ. De la même manière, nous pourrions calculer les moyennes de tous les temps de coalescence $\langle T_n \rangle$.

La structure même de ce processus a deux corollaires immédiats :

1. les distributions des délais entre coalescences successives dans un même groupe sont exponentielles et ont pour valeurs moyennes les taux γ_b : ce modèle est donc incapable de décrire les processus où la distribution des événements de coalescence n'est pas exponentielle comme nous en rencontrerons par la suite en section 9.2 ;
2. le processus ne présente qu'une seule échelle de temps caractéristique contenue dans la normalisation de la mesure Λ puisque les rapports $\langle T_p \rangle / \langle T_2 \rangle$ ont une valeur finie indépendante de cette normalisation.

Le paragraphe suivant présente les temps de coalescence obtenus dans trois cas particuliers et nous verrons ensuite en section 8.2.3 une situation physique où ils émergent naturellement. Ils réapparaîtront quand nous traiterons le cas de processus de coalescence avec sélection et structure spatiale au chapitre suivant.

8.2.2 Temps de coalescence dans quelques cas particuliers

Nous nous concentrons ici sur les trois cas suivants :
– le coalescent de Kingman [Kin82a, Kin82b] :

$$\Lambda_K(x) = \delta(x); \tag{8.7}$$

– le Bêta-coalescent [BBC+05] de paramètre η :

$$\Lambda_\eta(x) = \frac{x^{1-\eta}(1-x)^{\eta-1}}{\Gamma(2-\eta)\Gamma(\eta)} \quad \text{avec } 1 \leq \eta < 2 \tag{8.8}$$

où $\Gamma(x)$ est la fonction Γ d'Euler[1] ;
– le coalescent de Bolthausen-Sznitman [BS98] :

$$\Lambda_{BS}(x) = 1. \tag{8.9}$$

Le premier et le dernier sont des cas limites de Bêta-coalescent pour les valeurs limites $\eta = 2$ et $\eta = 1$. Les calculs ci-dessous seront ainsi réalisés pour le Bêta-coalescent avec un η quelconque.

L'interprétation du coefficient η sera présentée dans la section suivante où nous montrerons comment il peut être déduit d'un processus de branchement équivalent. Dans toute la suite, nous nous concentrerons uniquement sur les rapports des temps de coalescence $\langle T_p\rangle/\langle T_2\rangle$ pour les premières valeurs de p.

Le calcul explicite des taux $\lambda_{b,k}$ à partir des formules (8.2, 8.3) donne :

$$\lambda_{b,k} = \frac{\Gamma(k-\eta)}{\Gamma(2-\eta)}\frac{\Gamma(b-k+\eta)}{\Gamma(\eta)}\frac{1}{\Gamma(b)} \tag{8.10}$$

puis un taux de coalescence γ_b total :

$$\gamma_b = \frac{\Gamma(b-1+\eta)}{\Gamma(\eta+1)\Gamma(b-1)}. \tag{8.11}$$

À partir de ces valeurs et des expressions de type (8.6) pour les temps $\langle T_p\rangle$, nous pouvons obtenir des expressions simples pour les rapports $\langle T_p\rangle/\langle T_2\rangle$. Afin de les comparer aux valeurs numériques que nous présenterons au chapitre suivant, nous présentons dans le tableau 8.1 les valeurs remarquables des rapports $\langle T_3\rangle/\langle T_2\rangle$ et $\langle T_4\rangle/\langle T_2\rangle$ pour les trois cas particuliers précédents. La forme des arbres pour des groupes de trois et quatre individus et leurs poids statistiques sont présentés dans le tableau 8.2.

8.2.3 Émergence des coalescents dans un modèle de branchement à taille constante

Les trois cas particuliers de Λ-coalescent précédents émergent naturellement dans un modèle de Wright-Fisher modifié [BBC+05]. Dans le modèle de Wright-Fisher originel, le

[1] Elle est définie par $\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}dt$ et satisfait, entre autres, les propriétés $\Gamma(x+1) = x\Gamma(x)$ et $\Gamma(x) = (x-1)!$ pour $x \in \mathbb{N}^*$.

Rapport	Kingman ($\eta \to 2$)	Bêta-coalescent ($1 \leq \eta < 2$)	Bolthausen-Sznitman ($\eta \to 1$)
$\dfrac{\langle T_3 \rangle}{\langle T_2 \rangle}$	$\dfrac{4}{3}$	$\dfrac{2+3\eta}{2+2\eta}$	$\dfrac{5}{4}$
$\dfrac{\langle T_4 \rangle}{\langle T_2 \rangle}$	$\dfrac{3}{2}$	$\dfrac{5\eta^2+14\eta+6}{3(1+\eta)(2+\eta)}$	$\dfrac{25}{18}$

TAB. 8.1: Rapports des premiers temps de coalescence pour différentes cas particuliers de Λ-coalescent (Kingman, Bêta, Bolthausen-Sznitman).

Arbre	Kingman	B.-S.
	1	$\dfrac{3}{4}$
	0	$\dfrac{1}{4}$

Arbre	Kingman	B.-S.
	$\dfrac{2}{3}$	$\dfrac{1}{3}$
	$\dfrac{1}{3}$	$\dfrac{1}{6}$
	0	$\dfrac{1}{6}$
	0	$\dfrac{2}{9}$
	0	$\dfrac{1}{9}$

TAB. 8.2: Poids des arbres dans les coalescents de Kingman et de Bolthausen-Sznitman (B.-S.) pour des groupes de trois (gauche) et quatre individus (droite).

parent d'un individu est choisi aléatoirement et uniformément parmi les individus de la génération précédente. Si nous nous intéressons au processus direct dans le temps, cela correspond (dans la limite de grande taille N de population) à procéder de la manière suivante :

- chaque individu i à la génération G a n_i enfants où n_i est distribué selon une loi p_n, conditionnée de sorte que $n_1 + n_2 + \ldots + n_N \geq N$;
- parmi les $N' = n_1 + n_2 + \ldots + n_N$ individus ainsi produits, nous n'en conservons que N choisis aléatoirement à la génération $G + 1$.

Le modèle de Wright-Fisher introduit au chapitre 1 est obtenu en prenant une distribution poissonnienne p_n du nombre d'enfants. De fait, pour N grand, on retrouve le modèle de Wright-Fisher dès lors que p_n est de *variance finie*. La structure des arbres généalogiques est alors donnée, à l'échelle de temps près, par le coalescent de Kingman introduit ci-dessus.

Lorsque la distribution p_n a une moyenne finie mais une variance infinie, la structure des arbres change (on passe du modèle du coalescent de Kingman au Bêta-coalescent puis au coalescent de Bolthausen-Sznitman), ainsi que l'échelle des temps de coalescence à k individus. Une telle distribution p_n se caractérise par une décroissance lente lorsque n est grand :

$$p_n \simeq \frac{A}{n^{1+\eta}} \tag{8.12}$$

avec un exposant η dans l'intervalle $]1, 2[$.

Que la variance soit finie ou non, la fonction génératrice $\widehat{p}(s)$ a le développement suivant pour s proche de 0 :

$$\widehat{p}(s) \hat{=} \langle e^{-sn} \rangle = 1 - s\langle n \rangle + Cs^\eta + o\left(s^\eta\right) \tag{8.13}$$

où la constante C est reliée à la constante A ou à la variance finie selon les cas. Lorsque la variance est finie, l'exposant η vaut[2] exactement 2.

Si nous notons q_k la probabilité que k individus aient le même parent à la génération précédente, nous allons voir, selon les cas, que tous les q_k sont du même ordre ou bien que q_2 seul domine. Pour qu'une coalescence à k individus se produise, les k individus doivent faire partie du même groupe de taille n_k engendré par leur parent. La probabilité q_k est ainsi donnée par la probabilité que k individus appartiennent à la même composante n_i :

$$q_k = \sum_{i=1}^{N} \left\langle \frac{n_i(n_i-1)\ldots(n_i-k+1)}{N'(N'-1)\ldots(N'-k+1)} \right\rangle \tag{8.14}$$

où $N' = n_1 + \ldots + n_N$.

Selon la décroissance de la distribution p_n, la fréquence des groupes de grande taille varie. Lorsque la variance est finie, les q_k se comportent comme $1/N$ pour $k = 2$ et comme au plus $1/N^2$ pour $k \geq 3$: chaque fraction $n_i/(n_1 + \ldots + n_N)$ est d'ordre $1/N$. Ainsi seules les coalescences *binaires* sont autorisées et, à chaque coalescence, la taille

[2] Plus généralement, si tous les moments de la distribution p_n sont finis, la fonction génératrice $\widehat{p}(s)$ se développe en série entière. Si seuls les m premiers moments sont finis, alors $\widehat{p}(s)$ n'admet un développement de Taylor que jusqu'à l'ordre m.

du groupe décroît d'une unité : cette simplification permet d'aller plus loin dans l'étude du coalescent de Kingman, comme nous allons le voir dans la section 8.3. Lorsque la variance est infinie, les groupes de grande taille sont plus fréquents (cf. expression ci-dessous). Puisque seules les grandes tailles $n_i \gg 1$ importent à l'ordre dominant, il est possible de remplacer asymptotiquement l'expression précédente de q_k par :

$$\begin{aligned} q_k &= \sum_{i=1}^{N} \Big\langle \left(\frac{n_i}{n_1 + n_2 + \ldots + n_N} \right)^k \Big\rangle = \sum_{i=1}^{N} \int_0^\infty s^{k-1} \Big\langle n_i^k e^{-s(n_1+\ldots+n_N)} \Big\rangle ds \\ &= N \int_0^\infty s^{k-1} (-1)^k \widehat{p}^{(k)}(s) \widehat{p}(s)^N ds. \end{aligned}$$

Puisque $k \geq 2$, la quantité $\widehat{p}^{(k)}(s)$ est dominée par une contribution du type $s^{\eta-k}$ lorsque s est petit. Le changement de variable $s = t/N$ donne ainsi asymptotiquement $q_k \propto 1/N^{\eta-1}$. L'exposant est cette fois-ci indépendant de k et cela prouve que des coalescences à $k \geq 2$ individus peuvent se produire.

Plus généralement, il est possible de calculer d'une manière similaire les taux $\lambda_{b,k}$ définis dans la section 8.2. À l'ordre dominant pour une variance infinie, nous avons ainsi :

$$\lambda_{b,k} \simeq N^{b-k+1} \Big\langle \frac{n_1^k n_2 \ldots n_{b-k+1}}{(n_1 + n_2 + \ldots + n_N)^b} \Big\rangle \tag{8.15}$$

où le coefficient N^{b-k+1} dénombre le choix des $b-k+1$ parents dans la génération précédente. Une démarche similaire à celle suivie pour q_k conduit à :

$$\lambda_{b,k} \simeq N^{b-k+1} \int_0^\infty s^{b-1} (-1)^k \widehat{p}^{(k)}(s) (-1)^{b-k} \widehat{p}^{(1)}(s)^{b-k} \widehat{p}(s)^{N-b} ds \tag{8.16}$$

Le changement de variable $s = t/N$ oblige à utiliser les développements $\widehat{p}^{(1)}(s) \simeq -\langle n \rangle$ et $\widehat{p}^{(k)}(s) \simeq (-1)^k C s^{\eta-k} \Gamma(k-\eta)/\Gamma(-\eta)$. Nous obtenons ainsi :

$$\lambda_{b,k} \simeq \left[\frac{C}{\langle n \rangle^\eta \Gamma(-\eta)} \cdot \frac{1}{N^{\eta-1}} \right] \cdot \Gamma(k-\eta)\Gamma(b-k+\eta) \tag{8.17}$$

qui est le résultat présenté dans [BBC+05] à une normalisation près. On montre par la même méthode que m coalescences ont une probabilité négligeable $1/N^{m(\eta-1)}$ d'intervenir simultanément. Ces résultats montrent ainsi que, à une mise à l'échelle du temps près, le processus de coalescence à k individus correspond au Bêta-coalescent (8.8) où le coefficient η est donné par l'exposant de décroissance (8.12) de la distribution p_n.

La conclusion à retenir de cet exemple est que les arbres généalogiques et l'échelle des temps de coalescence sont intimement liés aux *fractions* de la population générées par les différents individus [BBC+05, BL00]. Le passage de la variance d'une valeur finie à une valeur infinie provoque l'émergence de fractions macroscopiques qui facilitent les coalescences entre individus de deux manières : d'abord en raccourcissant l'échelle de temps (passage de N à $N^{\eta-1}$ avec $\eta-1 < 1$), puis en autorisant les coalescences multiples. Dans tous les modèles avec sélection que nous étudierons par la suite, c'est ce même effet qui modifiera les généalogies. D'autre part, cette notion de fractions macroscopiques de la population générées par quelques individus correspond à l'image phénoménologique introduite dans [BDMM06b, BDMM07] dans l'étude de fronts bruités.

8.3 Coalescent de Kingman et généalogies dans le modèle de Wright-Fisher

8.3.1 Temps de coalescence dans la limite de grande taille

Nous revenons ici sur le modèle de Wright-Fisher décrit au chapitre 1 et considérons une population de taille N : à chaque nouvelle génération, la population est remplacée par N nouveaux individus dont les parents sont tirés aléatoirement et *uniformément* (nous travaillons sans sélection) dans la génération précédente. Deux individus donnés ont donc le même parent avec une probabilité $1/N$. Plus généralement, p individus ont le même parent avec une probabilité $1/N^p$: ainsi, si nous ne nous intéressons qu'à un groupe de taille finie n dans une population de taille N très grande, les seules coalescences pertinentes, *i.e.* les événements où plusieurs individus du groupe ont le même parent, sont celles à deux individus. De plus, cela implique que les temps caractéristiques à considérer sont de l'ordre de la taille N de la population. Après mise à l'échelle, nous retrouvons ainsi le coalescent de Kingman [Kin82a, Kin82b]. Nous allons à présent donner quelques propriétés supplémentaires de ce coalescent.

Si nous considérons à un temps t un groupe de n individus dans une population de taille N grande, alors, à l'ordre dominant en $1/N$, il existe une série de dates $t_1 < t_2 < \ldots < t_{n-1} < t_n = t$ dans le passé telles que, à la génération t_p, la taille du groupe des ancêtres décroît d'une unité car deux individus ont le même parent dans la génération précédente. Un arbre généalogique est alors entièrement spécifié par les dates des coalescences et les individus qui participent à chacune de celles-ci.

Nous appellerons $\tau_p = t_p - t_{p-1} > 0$ le délai entre deux coalescences successives. En temps discret, la distribution des temps τ_p peut être obtenue en remarquant que, pendant $\tau_p - 1$ générations, les p individus doivent avoir des parents différents et, à la τ_p-ème génération, deux individus ont le même parent :

$$\begin{aligned}
\text{Prob}\left(\tau_p = \tau\right) &= C_p^2 \left[\frac{1}{N}\frac{N-1}{N}\ldots\frac{N-(p-2)}{N}\right] \\
&\quad \times \left[\frac{N-1}{N}\frac{N-2}{N}\ldots\frac{N-(p-1)}{N}\right]^{\tau_p-1} \\
&= C_p^2 \left(1-\frac{1}{N}\right)^{\tau_p} \ldots \left(1-\frac{p-2}{N}\right)^{\tau_p} \left(1-\frac{p-1}{N}\right)^{\tau_p-1}
\end{aligned}$$

où le coefficient binomial $C_p^2 = p(p-1)/2$ compte le nombre de paires possibles qui peuvent donner lieu à une coalescence. Cette expression montre que τ_p varie comme N lorsque N devient grand. Après changement d'échelle, nous avons la distribution suivante :

$$\boxed{\text{Prob}\left(\frac{\tau_p}{N} = \tau\right) \xrightarrow{N\to\infty} C_p^2 e^{-C_p^2 \tau}} \tag{8.18}$$

et nous retrouvons les distributions exponentielles attendues pour le coalescent de Kingman. Puisque seuls les coefficients binomiaux C_p^2 interviendront par la suite, nous allè-

gerons les notations en définissant, pour $p \geq 2$, les coefficients :

$$\boxed{c_p \hat{=} C_p^2 = \frac{p(p-1)}{2}} \,. \tag{8.19}$$

De plus, nous ne travaillerons désormais que dans la limite $N \to \infty$ et tous les temps considérés auront été mis à l'échelle $t \to t/N$. Le temps de coalescence T_n de n individus est la somme des temps τ_p pour $2 \leq p \leq n$ et a donc pour fonction génératrice :

$$\boxed{\langle e^{-sT_n} \rangle = \prod_{p=2}^{n} \frac{c_p}{s + c_p}} \,. \tag{8.20}$$

En particulier, la valeur moyenne de l'âge de l'ancêtre commun d'un groupe de n individus est donnée, après normalisation par la taille de la population, par :

$$\langle T_n \rangle = 2 \left(1 - \frac{1}{n} \right) . \tag{8.21}$$

Si nous nous intéressons à présent non plus à des sous-groupes de la population mais à la population entière, nous pouvons décomposer la dynamique de coalescence de la population globable en deux phases :

1. une brève phase de coalescences multiples où la taille du groupe d'ancêtres passe de N, la taille de la population, à un nombre fini n d'individus ; en effet, pour N grand, en une unique génération, le nombre d'ancêtres passe, en moyenne, de N à Ne^{-1} et décroît exponentiellement : au bout d'un temps typique $\propto \ln N$, le nombre d'ancêtres est d'ordre 1 ;
2. une phase longue où les n ancêtres laissés à la fin de la phase précédente subissent des coalescences rares (temps d'ordre N) et qui est décrite par les propriétés de coalescence, détaillées ci-dessus, d'un petit groupe d'individus dans une grande population.

Ainsi, l'âge de l'ancêtre commun le plus récent de toute la population, noté désormais T (sans indice) peut être obtenu en considérant des sous-groupes de tailles de plus en plus grandes et, après normalisation par la taille de la population, sa fonction génératrice est donnée par le produit infini :

$$\boxed{\langle e^{-sT} \rangle = \lim_{n \to \infty} \langle e^{-sT_n} \rangle = \prod_{p=2}^{\infty} \frac{c_p}{s + c_p}} \,. \tag{8.22}$$

La distribution de probabilité $\rho(T)$ correspondante est donnée par la série

$$\rho(T) = \sum_{p=2}^{\infty} (-1)^p (2p-1) c_p e^{-c_p T} \tag{8.23}$$

et nous avons les expressions suivantes des deux premiers moments :

$$\langle T \rangle = 2 \quad (8.24)$$

$$\langle T^2 \rangle - \langle T \rangle^2 = \frac{4\pi^2}{3} - 12 \simeq 1.16 \quad (8.25)$$

Il est intéressant de remarquer que l'âge T de l'ancêtre commun le plus récent de la population n'est pas auto-moyennant dans la limite $N \to \infty$: la largeur de sa distribution provient majoritairement du haut de l'arbre généalogique où le temps τ_2, qui est un temps de coalescence à deux individus seulement, contribue pour moitié à T.

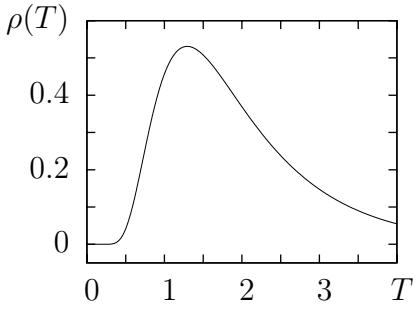


FIG. 8.3: Distribution stationnaire $\rho(T)$ de l'âge T du MRCA.

8.3.2 Quelques propriétés statistiques des arbres

Mesure sur les généalogies

Le modèle de Wright-Fisher définit une mesure stationnaire sur les généalogies de groupes d'individus. Dans ce modèle, un arbre généalogique est représenté par les temps τ_p entre deux coalescences successives et par la donnée, à chaque coalescence, de la paire d'individus qui ont le même parent[3].

Dans le coalescent de Kingman, tous les individus jouent le même rôle et toutes les coalescences de paires ont le même poids $1/c_n$ dans un groupe à n individus. Le nombre total d'arbres $\mathcal{T}$ possibles est donc donné par :

$$S(n) = \prod_{p=2}^{n} c_p = \frac{n!(n-1)!}{2^{n-1}}. \quad (8.26)$$

Il est ainsi possible de définir une mesure μ_n sur les généalogies $(\mathcal{T}, \tau_2, \ldots, \tau_n)$ d'un groupe de n individus dans le coalescent de Kingman. Celle-ci se factorise en une partie temporelle et une partie topologique et, dans la limite des grandes populations, devient :

$$\mu_n(\mathcal{T}, \tau_2, \ldots, \tau_n) = \frac{1}{S(n)} \prod_{p=2}^{n} \left(c_p e^{-c_p \tau_p} \right). \quad (8.27)$$

Nous verrons au chapitre 11 comment cette mesure est modifiée lorsque des informations supplémentaires sont connues à propos de la diversité génétique d'un groupe d'individus.

[3]Une façon de décrire les arbres à n individus est de considérer les partitions croissantes d'un ensemble à n éléments : une coalescence à k individus correspond à passer d'une partition à la suivante en faisant l'union disjointe de k ensembles. Les n individus de la génération présente sont donc repérés par la partition $\{\{1\}, \{2\}, \ldots, \{n\}\}$ alors que l'ancêtre commun le plus récent de tout le groupe l'est par l'ensemble $\{1, 2, \ldots, n\}$.

Partition d'un groupe suivant sa lignée ancestrale

Étant donné que les coalescences sont binaires, tout groupe de n individus peut être divisé en deux sous-groupes (cf. figure 8.1) selon la lignée issue de l'ancêtre commun le plus récent du groupe à laquelle ils appartiennent : juste avant d'atteindre l'ancêtre commun le plus récent du groupe, il ne reste plus que deux ancêtres dans une même génération et il est alors possible de classer les individus selon qu'ils descendent de l'un ou l'autre de ces deux ancêtres.

Soit $a_{p,n}$ la probabilité qu'un groupe de taille n soit divisé par cette méthode en deux sous-groupes de tailles p et $n-p$. Cette propriété ne fait intervenir que la structure de l'arbre et non les temps de coalescence. Cette probabilité est donnée par :

$$a_{p,n} = \frac{1}{2}\frac{S(p)S(n-p)}{S(n)} \cdot C_n^p \cdot C_{n-2}^{p-1}. \tag{8.28}$$

Dans cette formule, les facteurs $S(n-p)$ et $S(p)$ dénombrent les formes possibles des arbres de chacune des deux lignées, le facteur $1/2$ tient compte de la symétrie entre les deux sous-groupes et le coefficient binomial C_n^p compte le nombre de façons de partitionner le groupe de taille n en deux sous-groupes de taille p et $n-p$. Le dernier coefficient binomial dénombre les possibilités d'arrangement chronologique des coalescences dans les deux lignées : la première contient $p-1$ coalescences, la deuxième $n-p-1$, et il faut décider dans quel ordre elles s'intercalent.

L'équation précédente se simplifie pour donner :

$$a_{p,n} = \frac{1}{n-1} \tag{8.29}$$

et la dépendance en p disparaît. Ainsi, la taille des deux sous-groupes est uniformément distribuée sur $\{1, \ldots, n-1\}$.

Dans la limite $n \to \infty$, la distribution de probabilité $a(x)$ de la fraction $x = p/n$ devient uniforme sur $[0,1]$ et représente effectivement le partitionnement de la population selon les deux lignées issues de l'ancêtre commun le plus récent pour x et $1-x$ d'ordre 1. En effet, dans la limite $N \to \infty$, il devient possible [Ser05] d'écrire une équation de diffusion de Wright-Fisher pour la fraction x dont la solution stationnaire est bien la solution uniforme :

$$a(x) = 1. \tag{8.30}$$

Si l'on considère une population de taille finie N, il existe des corrections de taille finie pour x et $1-x$ d'ordre $1/N$ mais elles ne seront pas pertinentes pour la suite.

8.3.3 Nombre d'ancêtres à une hauteur donnée : fonctions z_m

Nous nous intéresserons enfin aux probabilités stationnaires $z_m(T)$ d'observer un nombre d'ancêtres égal à m à une hauteur T dans l'arbre généalogique d'une population, qui nous seront très utiles dans la section 11.2. Elles ont été introduites dans l'étude de modèles d'évolution de populations sans sélection dans [DB88]. Il est possible de les calculer par récurrence à partir des coefficients binomiaux $c_p = C_p^2 = p(p-1)/2$:

pour avoir m ancêtres au temps $t-T$, il faut[4], ou bien qu'il y ait m ancêtres au temps $t-T+dt$ et aucune coalescence dans l'intervalle dt, ou bien qu'il y ait $m+1$ ancêtres au temps $t-T+dt$ et une coalescence dans l'intervalle dt. Ainsi, les fonctions $z_m(T)$ satisfont pour $m \geq 1$:

$$\frac{dz_m}{dT}(T) = c_{m+1} z_{m+1}(T) - c_m z_m(T). \tag{8.31}$$

D'autre part, si $\frac{d}{dT} z_1(T)$ est la probabilité que le nombre d'ancêtres de la population passe de 2 à 1 au temps T, alors, par définition, elle est égale à la distribution $\rho(T)$ de l'âge de l'ancêtre commun le plus récent de la population, que nous avons obtenue en (8.23). De plus, nous avons $z_1(0) = 0$ puisque la taille de la population est plus grande que 1 et que T varie comme la taille de la population. Nous obtenons (cf. [DB88]) ainsi l'expression de $z_1(T)$ puis celles de toutes les fonctions $z_m(T)$ à partir de (8.31) :

$$z_m(T) = \sum_{p=m}^{\infty} (-1)^{p+m} (2p-1) \frac{(m+p-2)!}{m!(m-1)!(p-m)!} e^{-c_p t}. \tag{8.32}$$

Il est possible de vérifier sur cette expression la normalisation $\sum_{m=1}^{\infty} z_m(T) = 1$ en utilisant l'identité hypergéométrique[5] :

$$\sum_{m=1}^{p} (-1)^{m-1} \frac{(m+p-2)!}{m!(m-1)!(p-m)!} = \begin{cases} 1 & \text{si } p = 1, \\ 0 & \text{si } p \geq 2. \end{cases}$$

8.4 Coalescent de Bolthausen-Sznitman et généalogies dans les marches aléatoires avec branchements avec sélection

8.4.1 Généralités

Le coalescent de Bolthausen-Sznitman, introduit en (8.9), est défini par $\Lambda_{BS}(x) = 1$ et les taux de coalescence $\lambda_{b,k}$ correspondants sont donnés par :

$$\lambda_{b,k} = \frac{\Gamma(b-k+1)\Gamma(k-1)}{\Gamma(b)} = \frac{(b-k)!(k-2)!}{(b-1)!}. \tag{8.33}$$

[4] Nous travaillons directement dans l'échelle de temps normalisée $t' = t/N$ et la limite $N \to \infty$.

[5] Nous avons en effet [GR94] :

$$\sum_{m=1}^{p} (-1)^{m-1} z^m \frac{(m+p-2)!}{m!(m-1)!(p-m)!} \propto J_{p-1}^{(1,-1)}(1-2z)$$

où $J_{p-1}^{(1,-1)}(x)$ est un polynôme de Jacobi défini par

$$J_{l-1}^{(1,-1)}(x) = \frac{(-1)^{l-1}}{2^{l-1}(l-1)!} \frac{1+x}{1-x} \frac{d^{l-1}}{dx^{l-1}} \left((1-x)^l (1+x)^{l-2} \right)$$

Le temps d'attente de la première coalescence dans un groupe de b individus est donc distribué exponentiellement avec une moyenne donnée par $1/\gamma_b$ avec $\gamma_b = b - 1$. Deux différences majeures avec le coalescent Kingman font qu'un certain nombre de résultats de la section 8.3 ne se généralisent pas aisément au coalescent de Bolthausen-Sznitman : la première est l'apparition de coalescences multiples (cf. tableau 8.2) qui font que les événements de coalescence ne peuvent plus être repérés seulement par le nombre d'individus présents au moment de la coalescence (les temps τ_p ne sont plus tous présents) ; la seconde est la divergence de $\sum_b 1/\gamma_b$ qui implique[6] la divergence du temps moyen $\langle T_p \rangle$ de coalescence d'un groupe de taille p (divergence du type $\langle T_p \rangle \propto \langle T_2 \rangle \ln \ln p$ observée dans [FM07, BDMM07]) : dans la limite de grande taille de population, l'âge de l'ancêtre commun le plus récent de la population complète n'est pas du même ordre de grandeur que le temps de coalescence à 2 individus. Les fonctions $z_m(t)$ introduites en (8.32) pour m fini ne se généralisent pas directement dans le cas du coalescent de Bolthausen-Sznitman [FM07] et seuls des résultats partiels [BBL07, FM07] sont pour l'instant disponibles.

Le coalescent de Bolthausen-Sznitman a été introduit pour la première fois dans l'étude des verres de spin en champ moyen [BS98] : le calcul de l'énergie libre par la méthode des répliques fait apparaître des arbres aléatoires dont les statistiques découlent des taux (8.2) du coalescent de Bolthausen-Sznitman. Nous allons voir dans les sections qui suivent comment ce coalescent apparaît également dans des problèmes de sélection.

8.4.2 Résultats numériques pour des marches aléatoires avec branchements

Nous avons vu au chapitre 1 comment il est possible de décrire des populations en présence de compétition interne et de mutations par des marches aléatoires avec branchements. Chaque individu est représenté par une marche aléatoire sur l'axe du *fitness* (la diffusion joue le rôle des mutations au cours du temps) et la reproduction est modélisée par un branchement d'une marche aléatoire avec une probabilité β_k en k nouvelles marches aléatoires. La sélection est imposée en gardant constant le nombre de marches : dès que la reproduction fait passer de N marches à $N + k - 1$ marches, les $k - 1$ marches les plus à gauche sur l'axe des *fitnesses* sont immédiatement supprimées.

À un temps t, si nous considérons p marches parmi les N, il est possible de retracer leurs généalogies et de mesurer, au moins numériquement, les distributions des temps de coalescence T_p. Cette approche numérique a été réalisée dans [BDMM06a] et a conduit à des valeurs des rapports $\langle T_p \rangle / \langle T_2 \rangle$ analogues à celles du coalescent de Bolthausen-Sznitman pour $p = 3$ et $p = 4$ lorsque la taille N de la population est grande (cf. figure 8.4).

Les échelles de temps observées pour les modèles de la figure 8.4 dépendent des détails du modèle, bien qu'elles soient toutes de la forme $(\log N)^\alpha$ lorsque la taille N de la population est grande. L'exposant α mesuré numériquement par les auteurs pour le modèle qui est comparable au nôtre que nous avons décrit en section 3.3 n'est pas

[6] À cause des coalescences multiples, tous les termes $1/\gamma_b$ ne sont pas présents mais la somme des temps de coalescence reste divergente.

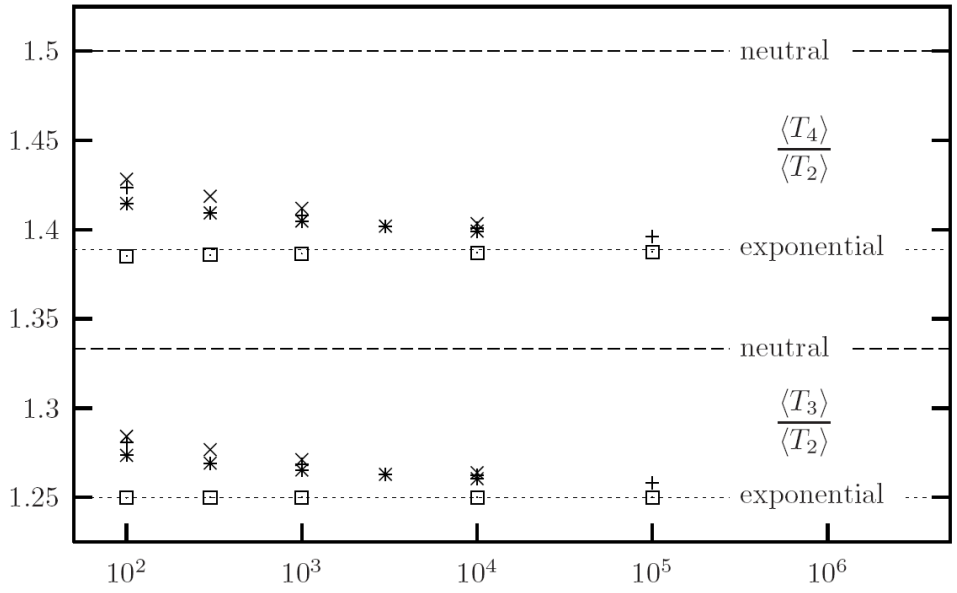


FIG. 8.4: Rapports des temps de coalescence $\langle T_p\rangle/\langle T_2\rangle$ pour $p = 3$ et $p = 4$ en fonction de la taille N dans différents modèles de marches aléatoires avec branchements maintenues à une taille constante N par la sélection (figure extraite de [BDMM06a]). Les pointillés correspondent aux valeurs théoriques des coalescents de Kingman (*neutral*) et de Bolthausen (*exponential*). Les carrés correspondent aux valeurs numériques mesurées pour le modèle exponentiel, les croix ($\times$) et les astérisques correspondent au mode de reproduction du chapitre 7 pour $\psi(x) = \theta(-x)$ et $\psi(x) = (-x)^3\theta(-x)$. Les signes + correspondent au modèle numérique que nous avons utilisé et qui est défini en section 3.3.

incompatible avec la valeur $\alpha_{\text{theo}} = 3$ qui est prédite dans [BDMM07]. Cette dernière valeur est d'ailleurs comparable à l'échelle de temps (3.31) qui apparaît dans le processus quasi-stationnaire des chapitres 5 et 6. De la même manière, la valeur numérique de α pour le modèle exponentiel dans [BDMM06a] est compatible avec la valeur théorique $\alpha_{\text{expo}} = 1$.

L'origine de ces temps de coalescence n'a pour l'instant qu'une explication phénoménologique [BDMM07]. Celle-ci est essentiellement basée sur l'étude de la fraction de la population générée par les individus dont le *fitness* devient supérieur aux autres individus de la population. Néanmoins, le modèle exponentiel du chapitre 7 permet des calculs analytiques complets qui aboutissent au coalescent de Bolthausen-Sznitman.

8.4.3 Généalogies dans le modèle exponentiel : un cas exactement soluble

Le modèle exponentiel à taille constante N est construit en prenant $a = y_N$ et $b = +\infty$ dans l'équation (7.1) afin que seuls les N meilleurs individus puissent subsister. Pour des parents aux positions $x_1,\ldots, x_N$, la distribution de probabilité des positions (classées

par ordre croissant) des N enfants est donnée par :

$$\text{Prob}\,(y_N < \ldots < y_2 < y_1 | x_N < \ldots < x_1) = e^{-(y_N - X)} e^{-(y_{N-1} - X)} \ldots e^{-(y_1 - X)} e^{-e^{-(y_N - X)}} \tag{8.34}$$

et ne dépend que de la variable réduite X définie par :

$$e^X = e^{x_1} + e^{x_2} + \ldots + e^{x_N}.$$

Une manière alternative de tirer les positions y_k des individus est la suivante : on tire tout d'abord la position $y_{N+1} = X + z_{N+1}$ du $(N+1)$-ème individu (qui ne survit pas) selon la loi $\exp(-(N+1)z_{N+1} - e^{-z_{N+1}})/N!$; puis, une fois dans le référentiel de cet $(N+1)$-ème individu, on tire N variables $z_k > 0$ avec une loi exponentielle e^{-z} pour former ensuite les positions $y_k = y_{N+1} + z_k$ des N individus, qu'il ne reste plus qu'à trier par ordre croissant de *fitness*. L'avantage de cette méthode est de fournir un *référentiel* naturel où tous les y_i (non classés) ont la même loi et sont décorrélés (cf. ci-dessous).

D'autre part, le parent de l'individu j (de position x_j) est choisi aléatoirement dans la génération précédente avec la probabilité :

$$\text{Prob}\left(j \xrightarrow{\text{parent}} i\right) = \frac{e^{x_i}}{e^{x_1} + \ldots + e^{x_N}} \tag{8.35}$$

et le choix du parent pour chaque individu est *indépendant* du choix des parents des autres individus. Cette propriété, qui n'est vraie que pour le modèle exponentiel, a l'avantage de le rendre exactement soluble. On remarquera que la probabilité (8.35) est invariante par translation de tous les individus et ne dépend que de leurs positions relatives.

D'après la propriété (8.34), nous voyons que dans le référentiel où $X = 0$ les y_i sont distribués comme les N individus les plus à droite d'un processus ponctuel de Poisson de densité exponentielle $e^{-y}dy$. Cela permet ainsi de calculer, de manière similaire à la section 8.2.3, la probabilité $\lambda_{b,k}$ d'observer une coalescence à k individus parmi b :

$$\lambda_{b,k} = \sum_{i_1, i_2, i_{b-k+1} \text{ différents}} \left\langle \frac{e^{k x_{i_1}} e^{x_{i_2}} \ldots e^{x_{i_{b-k+1}}}}{(e^{x_1} + \ldots + e^{x_N})^b} \right\rangle. \tag{8.36}$$

Le moyen le plus simple [BDMM07] d'évaluer cette expression pour N grand consiste à se placer dans le référentiel du $(N+1)$-ème individu (fictif) de telle sorte que les positions des N individus soient indépendantes et distribuées exponentiellement. Nous obtenons ainsi dans la limite de grande taille N :

$$\begin{aligned} \lambda_{b,k} &= N^{b-k+1} \int_0^\infty s^{b-1} \left\langle e^{k x_1} e^{x_2} \ldots e^{x_{b-k+1}} e^{-s(e^{x_1} + \ldots + e^{x_N})} \right\rangle ds \\ &= N^{b-k+1} \int_0^\infty (-1)^b s^{b-1} \mu^{(k)}(s) \mu'(s)^{b-k} \mu(s)^{N-b} ds \end{aligned}$$

où la fonction génératrice $\mu(s)$ est donnée par :

$$\mu(s) = \langle e^{-s e^x} \rangle = \int_0^\infty e^{-s e^x} e^{-x} dx \underset{u = e^x}{=} \int_1^\infty e^{-su} \frac{1}{u^2} du.$$

Nous nous retrouvons ainsi dans une situation tout à fait similaire à (8.16) à ceci près que le rôle de $\hat{p}(s)$ est joué ici par $\mu(s)$. Le nombre d'enfants de la section 8.2.3 est alors joué par $u = e^x$ et l'expression ci-dessus montre que u est distribué selon une loi de puissance d'exposant $\eta = 1$. *Mutatis mutandis*, la fin des calculs dans le modèle exponentiel est semblable à celui de la section 8.2.3 avec un exposant $\eta \to 1$: nous obtenons alors une échelle des temps de coalescence en

$$T_p \propto \log N \tag{8.37}$$

et les taux $\lambda_{b,k}$ du coalescent de Bolthausen-Sznitman.

Le modèle exponentiel ressemble beaucoup au cas sans *fitness* de la section 8.2.3 car, mise à part la progression de la coordonnée réduite e^X le long de l'axe de *fitness*, à chaque pas de temps les N *fitnesses* sont distribués indépendamment les uns des autres et sont directement reliés à la distribution du nombre d'enfants par individu (qui a une décroissance lente). Dans le cas des marches aléatoires avec branchements présentées en section 8.4.2 néanmoins, ce découplage ne se produit pas à chaque pas de temps. Nous avons vu aux chapitres 3 et 6 qu'il existe deux échelles de temps caractéristiques $\tau_1 \propto (v_c - v)^{-3/2} \propto \log^3 N$ et $\tau_2 \propto (v_c - v)^{-1} \propto \ln^2 N$ et la hauteur des arbres généalogiques est d'ordre $\tau_1 \propto \log^3 N$: il n'est donc pas impossible qu'à une échelle de temps plus grande que τ_2 et de l'ordre de τ_1, un découplage effectif semblable à celui qui se produit dans le modèle exponentiel se produise ici aussi et donne alors le coalescent de Bolthausen-Sznitman observé numériquement. Néanmoins, il n'a pas été encore possible de le montrer analytiquement, par exemple en renormalisant l'échelle de temps.

Chapitre 9

Influence de la structure spatiale sur les généalogies

Ce chapitre est consacré à l'étude de modèles de coalescence avec structure spatiale : les individus sont caractérisés à la fois par leur *fitness* et par leur position. Ce chapitre se divise en deux sous-parties, l'une consacrée aux modèles sans sélection, l'autre aux modèles en présence de sélection. En absence de sélection, nous rappelons les résultats existants au-delà de la dimension critique supérieure et dérivons la distribution des temps T_p en $d = 1$. En présence de sélection, nous mettons en évidence (cf. [BDS08]) le lien entre modèles d'évolution et modèles de polymères dirigés ; en particulier, nous présentons les résultats numériques obtenus pour les temps de coalescence à 2, 3 et 4 individus.

Le modèle du Λ-coalescent est un modèle dit de *champ moyen* puisqu'il suppose que tout individu peut subir une coalescence avec tout autre à tout moment. De même, dans les modèles précédents en présence de sélection interne, la compétition est présente entre tous les individus. En effet, lorsqu'un individu de haut *fitness* se divise, l'individu de plus bas *fitness* est éliminé instantanément, sans autre considération de distance entre ces deux individus.

Dans un certain nombre de situations cependant, la compétition n'est pas globale mais seulement locale. En particulier, si nous supposons que les individus sont distribués dans l'espace, alors la sélection n'agit que pour conserver localement les meilleurs. C'est alors la seule diffusion qui permet ensuite de faire interagir à grande échelle tous les individus. Il suffit de considérer les probabilités d'intersection de marches aléatoires en différentes dimensions (transience et récurrence) pour se rendre compte que la structure spatiale peut n'être pas négligeable.

Dans ce chapitre, nous présentons comment il est possible d'introduire une structure spatiale supplémentaire dans les modèles de reproduction avec ou sans sélection. La section 9.1 présente une version spatiale du modèle de Wright-Fisher et étudie la convergence des généalogies vers le coalescent de Kingman. La section 9.2 montre comment

il est possible de formuler une version spatiale des modèles de sélection de la première partie : le *fitness* est alors un nombre qui est porté par les individus qui diffusent sur un réseau et qui gouvernent leur compétition *locale*, telle l'énergie dans les polymères dirigés.

9.1 Marches aléatoires avec coalescence et généalogies sans sélection

9.1.1 Définition du modèle

Examinons le modèle de branchement suivant qui est localement un modèle de Wright-Fisher (cf. figure 9.1) :

- un individu est placé sur chaque site d'un réseau cubique de dimension d et de taille L avec conditions aux bords périodiques (tore de dimension d) ;
- d'une génération à la suivante, les individus sont placés sur les centres[1] des cubes du réseau de la génération précédente (cf. figure 9.1) ;
- à chaque génération, chaque nouvel individu choisit pour parent l'un des 2^d voisins du réseau précédent (sommets du cube).

La différence avec le modèle de Wright-Fisher est la restriction du choix du parent parmi les plus proches voisins.

Si nous remontons le temps, les lignées des individus effectuent des marches aléatoires en dimension d qui s'annihilent instantanément selon $A + A \to A$ dès qu'elles sont sur le même site (voir figure 9.2 en dimension $d = 1$). Dans ce langage, le modèle de Wright-Fisher en champ moyen, correspond à des marches aléatoires des lignées sur un graphe à N sommets totalement connecté (dimension « infinie »).

Il est à nouveau possible de définir les temps T_p comme les âges des ancêtres communs les plus récents de p individus. Ces âges correspondent aux durées au bout desquelles il ne reste plus qu'une seule marche aléatoire sachant qu'on a commencé avec p marches. *A priori*, ces temps dépendent des positions initiales respectives des différentes marches : si les positions initiales ne sont pas précisées, T_p désignera le temps de coalescence de p marches moyenné sur les $(L^d)^p$ positions initiales possibles des marches.

9.1.2 Généalogies en dimension $d \geq 2$

Les modèles de marches aléatoires avec coalescence ont été étudiés par les mathématiciens [Cox89, LS06] et peuvent être reformulés en termes de modèles de votants à d dimensions où chaque individu a tendance à aligner son opinion sur celle de ses voisins : les marches aléatoires avec coalescence sont les duaux de ces modèles et exhibent les mêmes propriétés. En particulier, le temps T au bout duquel il ne reste plus qu'une marche est l'analogue du temps de consensus au bout duquel tous les votants d'une population ont la même opinion.

[1] Le nouveau réseau est lui aussi cubique et est déplacé de $(\vec{e}_1 + \ldots + \vec{e}_d)/2$ par rapport au précédent, où les vecteurs $\vec{e}_i$ sont les vecteurs de la base canonique de $\mathbb{Z}^d$.

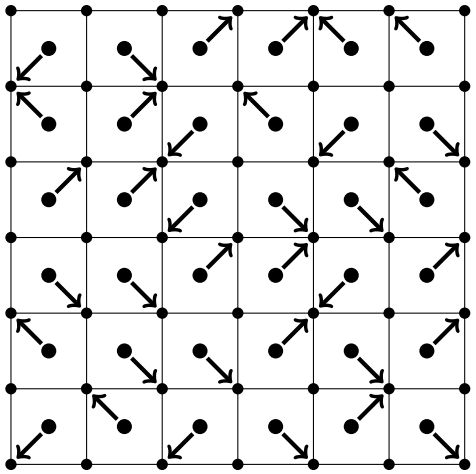


FIG. 9.1: Généralisation du modèle de Wright-Fisher avec une structure spatiale (ici $d = 2$) : les petits cercles désignent les individus de la génération t, les grands cercles ceux de la génération $t + 1$ et les flèches montrent l'attribution des parents : au lieu de choisir le parent parmi les N individus de la génération précédente, le choix se fait parmi les 2^d plus proches voisins.

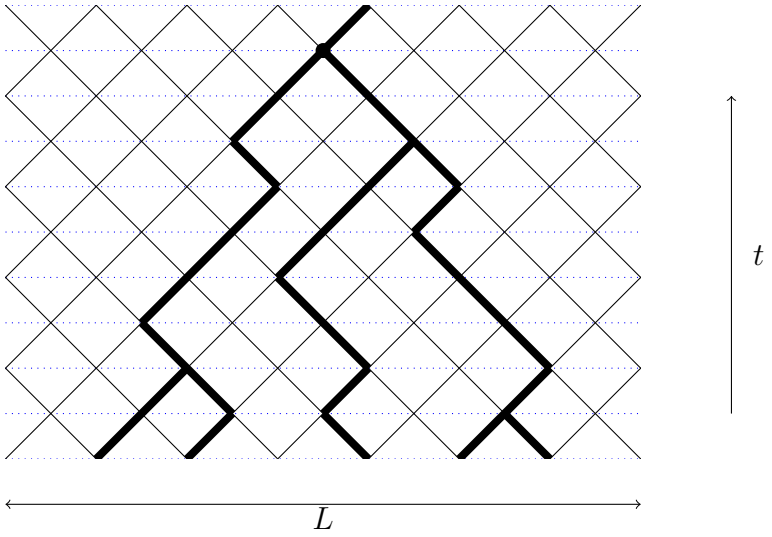


FIG. 9.2: Marches aléatoires avec coalescence sur réseau. Nous supposerons ici des conditions aux limites périodiques : chaque marche saute sur l'un des deux sites voisins avec probabilité 1/2 à chaque pas de temps. Deux marches sur le même site subissent une coalescence instantanée. Le point épais représente le moment où les cinq marches initiales n'en forment plus qu'une : c'est l'équivalent de l'ancêtre commun le plus récent dans les modèles de population.

Selon la dimension, les temps de coalescence T_p moyennés sur les positions initiales des marches aléatoires divergent comme :

$$T_p \propto \begin{cases} L^2 & \text{pour } d = 1, \\ L^2 \ln L & \text{pour } d = 2, \\ L^d & \text{pour } d > 2. \end{cases} \tag{9.1}$$

Parmi les résultats remarquables [Cox89, LS06], nous pouvons citer le fait qu'en dimension plus grande que deux, les temps de coalescence T_p moyennés sur les positions initiales des marches sont distribués, dans la limite $L \to \infty$ et après changement d'échelle, comme dans le coalescent de Kingman (cf. section 8.3). Nous notons en particulier que les valeurs suivantes des rapports des temps de coalescence sont universelles dès que $d \geq 2$:

$$\frac{\langle T_3 \rangle}{\langle T_2 \rangle} \xrightarrow{L\to\infty} \frac{4}{3}, \quad \frac{\langle T_4 \rangle}{\langle T_2 \rangle} \xrightarrow{L\to\infty} \frac{3}{2} \dots \tag{9.2}$$

et ne dépendent pas du détail du réseau.

Ces propriétés de *champ moyen* au-delà de la dimension critique $d_c = 2$ sont à mettre en relation avec les propriétés de transience des marches aléatoires pour $d > 2$. En effet, l'écart entre deux marches aléatoires est encore une marche aléatoire et la coalescence des deux marches se produit lorsque l'écart entre elles passe par l'originie $(0, \dots, 0)$. Lorsque L devient grand et que $d > 2$, le temps de passage à l'origine diverge comme L^d et est distribué exponentiellement. Pour $p > 2$ marches aléatoires, les coalescences multiples deviennent négligeables : si deux d'entre elles subissent une coalescence, les coalescences avec les autres prennent un temps du même ordre de grandeur même si ces autres ne sont pas éloignées du lieu de coalescence des deux premières. Ainsi, après changement d'échelle, la dimension spatiale n'est plus pertinente dès que $d > 2$.

Nous nous contenterons ici de montrer comment, en toute dimension, la distribution du temps T_2 est reliée à la probabilité de premier passage par l'origine d'une marche aléatoire et nous montrerons que, pour $d \geq 2$, la distribution du temps T_2 de coalescence de deux marches de positions initiales aléatoires est exponentielle, comme cela est attendu d'après [Cox89, LS06]. En dimension 1, où les généalogies ne sont plus données par le coalescent de Kingman mais où les calculs sont plus simples, nous déterminons la distribution de tous les temps de coalescence T_p de p marches aléatoires de positions initiales aléatoires et uniformément distribuées.

9.1.3 Distribution du temps T_2 en toute dimension

Nous nous intéressons dans cette section au temps de coalescence de deux marches aléatoires sur un réseau de taille L en dimension d. Leur temps de coalescence ne dépend que de la distance relative initiale des deux marches. Soit $g_2(\alpha, \vec{x}) = \langle e^{-\alpha T_2} \rangle_{\vec{x}}$ la fonction génératrice du temps de coalescence de deux marches séparées initialement d'un écart $\vec{x}$. Si les deux marches aléatoires sont au même point ($x_i = 0$ ou L pour tout i), alors nous avons coalescence instantanée et $T_2 = 0$, de telle sorte que :

$$g_2(\alpha, \vec{0}) = 1. \tag{9.3}$$

Une formulation équivalente consiste à dire que T_2 est le temps de premier passage par l'origine (modulo L dans chaque direction) de l'écart entre les deux marches aléatoires. Dans le réseau en *diamants* précédent (cf. figure 9.2), chaque marche aléatoire se déplace d'un vecteur $(\sigma_1/2, \ldots, \sigma_d/2)$ avec $\sigma_i = \pm 1$. L'écart entre les deux marches effectue donc une marche aléatoire sur $\mathbb{Z}^d$ (modulo L dans chaque direction) et se déplace à chaque pas de temps d'un vecteur $\vec{e} = (\epsilon_1, \ldots, \epsilon_n)$ avec les ϵ_i qui valent chacun -1, 0 ou 1 avec des probabilités $p(\vec{e})$ égales à 1/4, 1/2 et 1/4.

Pour cette marche sur $\mathbb{Z}^d$, la probabilité $P(\vec{0}, t|\vec{x}, 0)$ d'être en $\vec{0}$ à t en partant de $\vec{x}$ satisfait pour $t > 1$:

$$P(\vec{0}, t|\vec{x}, 0) = \sum_{\tau=1}^{t} P(\vec{0}, t-\tau|\vec{0}, 0) P_{\text{premier}}(\vec{0}, \tau|\vec{x}, 0) \tag{9.4}$$

où $P_{\text{premier}}(\vec{0}, \tau|\vec{x}, 0)$ est la probabilité d'être pour la *première fois* en $\vec{0}$ à t en partant de $\vec{x}$. Cette dernière probabilité est, par définition, la distribution de T_2. La transformée de Laplace de l'équation précédente donne pour $\vec{x} \neq \vec{0}$:

$$\begin{aligned} g_2(\alpha, \vec{x}) &= \frac{\widetilde{P}(\vec{x}, \alpha)}{\widetilde{P}(\vec{0}, \alpha)}, \\ \widetilde{P}(\vec{x}, \alpha) &= \sum_{t=0}^{\infty} e^{-\alpha t} P(\vec{0}, t|\vec{x}, 0). \end{aligned}$$

D'autre part, $P(\vec{0}, t|\vec{x}, 0)$ satisfait l'équation de diffusion suivante :

$$P(\vec{0}, t+1|\vec{x}, 0) = \sum_{\vec{e}} p(\vec{e}) P(\vec{0}, t|\vec{x} + \vec{e}, 0) \tag{9.5}$$

où $\vec{e}$ relie $\vec{x}$ à tous les voisins où la marche peut sauter avec une probabilité $p(\vec{e})$. On obtient ainsi :

$$\widetilde{P}(\vec{x}, \alpha) = \sum_{m_1, \ldots, m_d = 0, \ldots, L-1} \frac{1}{L^d} \frac{e^{2i\pi \vec{m} \cdot \vec{x}/L}}{e^{\alpha} - \sum_{\vec{e}} p(\vec{e}) e^{2\pi i \vec{m} \cdot \vec{e}/L}}. \tag{9.6}$$

D'autre part, la distribution du temps de coalescence moyenné sur les positions initiales est obtenue en intégrant sur $\vec{x}$:

$$\overline{g}_2(\alpha) = \frac{1}{L^d} \sum_{\vec{x} \in \{0, \ldots, L-1\}^d} g_2(\alpha, \vec{x}) = \frac{1}{\widetilde{P}(\vec{0}, \alpha)} \frac{1}{L^d} \sum_{\vec{x} \in \{0, \ldots, L-1\}^d} \widetilde{P}(\vec{x}, \alpha). \tag{9.7}$$

En utilisant (9.6), nous obtenons la couple d'équations :

$$\frac{1}{L^d} \sum_{\vec{x} \in \{0, \ldots, L-1\}^d} \widetilde{P}(\vec{x}, \alpha) = \frac{1}{L^d} \frac{1}{e^{\alpha} - 1}, \tag{9.8}$$

$$\widetilde{P}(\vec{0}, \alpha) = \sum_{m_1, \ldots, m_d = 0, \ldots, L-1} \frac{1}{L^d} \frac{1}{e^{\alpha} - \sum_{\vec{e}} p(\vec{e}) e^{2\pi i \vec{m} \cdot \vec{e}/L}}. \tag{9.9}$$

Tant que α est strictement positif, l'expression précédente est bien définie. Lorsque α tend vers 0, le terme $\vec{m} = \vec{0}$ fait diverger la fonction $\widetilde{P}(\vec{0}, \alpha)$. Lorsque L est grand et $\alpha = o(1/L^2)$, la somme sur $\vec{m} \neq \vec{0}$ dans (9.9) se comporte de la manière suivante lorsque $d > 2$:

$$\sum_{\vec{m}\in\{0,\dots,L\}^d, \vec{m}\neq\vec{0}} \frac{1}{e^{\alpha} - \sum_{\vec{e}} p(\vec{e}) e^{2\pi i \vec{m}\cdot\vec{e}/L}} \simeq \sum_{\vec{m}} \frac{L^d}{4\pi^2 \sum_{\vec{e}} p(\vec{e})(\vec{e}\cdot\vec{m})^2} \simeq a_d L^d \tag{9.10}$$

où a_d est une constante qui dépend de la dimension, du réseau et des taux de sauts $p(\vec{e})$.

Ainsi, en prenant $\alpha = \alpha'/(a_d L^d)$ pour $d > 2$ dans les équations (9.7, 9.8, 9.9), nous obtenons $\overline{g}_2(\alpha) \to 1/(1+\alpha')$ dans la limite $L \to \infty$ puis la distribution limite suivante pour l'âge T_2 :

$$\boxed{\text{Prob}\left(\frac{T_2}{a_d L^d} = \tau_2\right) \xrightarrow{d>2} e^{-\tau_2}} \tag{9.11}$$

et nous retrouvons l'échelle (9.1). Pour $d = 2$, une étude asymptotique plus précise montre qu'il faut substituer un coefficient $a_2 \log L$ au coefficient a_d ci-dessus. En prenant $\alpha = \alpha'/(a_2 L^2 \ln L)$, la formule (9.11) reste valable. La distribution de T_2 pour $d \geq 2$ est donc exponentielle, comme dans un modèle de Λ-coalescent en champ moyen.

En dimension 1, en revanche, la seconde somme sur $m \neq 0$ dans (9.10) reste convergente lorsque $L \to \infty$ et α varie comme α'/L^2 et sa limite dépend de α'. La somme (9.10) se comporte alors comme $L^2 \sum_{m\leq 1} 1/(\alpha' + m^2)$ et la dépendance en α' de la somme (9.10) ne disparaît plus. Ainsi, en dimension 1, le temps T_2 de coalescence de deux marches aléatoires de positions initiales aléatoires est distribué selon la loi :

$$\boxed{\text{Prob}\left(\frac{T_2}{a_1 L^2} = \tau_2\right) \xrightarrow{d=1} \sum_{m\in\mathbb{Z}} e^{-(m+1/2)^2\pi^2\tau_2}} \tag{9.12}$$

où a_1 est une constante qui dépend du réseau et des taux de sauts[2]. Une telle distribution correspond à la fonction génératrice du temps T_2 normalisé suivante :

$$\langle e^{-\alpha'(T_2/a_1 L^2)} \rangle \to \frac{\tanh(\sqrt{\alpha'})}{\sqrt{\alpha'}} = \sum_{m\in\mathbb{Z}} \frac{1}{\alpha' + (m+1/2)^2\pi^2} \tag{9.13}$$

que nous allons retrouver ci-dessous par une démarche légèrement différente. Les résultats ci-dessus sont compatibles avec les résultats de [LS06] qui démontrent que le coalescent de Kingman est valable dès la dimension 2 dans la limite $L \to \infty$, quitte à redéfinir l'échelle de temps pour tenir compte de la diffusion des marches aléatoires.

[2] Avec ces conventions, nous obtenons $\langle T_2 \rangle/(a_1 L^2) \to 3 \neq 1$. Pour se ramener à une moyenne égale à 1, il suffit de changer l'expression de a_1 d'un facteur 3 mais nous conserverons la normalisation présente afin de simplifier les calculs.

9.1.4 Distribution des temps de coalescence T_p en dimension 1

En dimension 1, nous pouvons obtenir une expression exacte [BDS08] pour les temps T_p dans la limite $T_p \to \infty$. Les L individus sont donc disposés sur un cercle de L sites. Considérons p individus situés aux positions $0 \leq x_1 \leq \ldots \leq x_p \leq L$. Définissons les distances d_i entre deux individus successifs par $d_i = x_{i+1} - x_i$ pour $i < p$ et $d_p = L - x_p + x_1$, de telle sorte que $d_1 + \ldots + d_p = L$. Les lignées diffusent sur le réseau et deux lignées successives subissent une coalescence dès qu'elles parviennent sur le même site, *i.e.* dès que $d_i = 0$.

Introduisons de manière similaire à la section précédente les fonctions génératrices :

$$g_p(x_1, \ldots, x_p; \alpha) = \langle e^{-\alpha T_p(x_1,\ldots,x_p)} \rangle \tag{9.14}$$

où T_p est le temps au bout duquel il ne reste plus qu'une seule lignée. L'invariance par translation implique que g_p ne dépend que des distances d_i entre individus successifs et est invariante sous le décalage $d_i \mapsto d_{i+1}$ et $d_p \mapsto d_1$. De plus, dès que k distances d_i sont égales à 0, T_p est égal au temps de coalescence T_{p-k} des $p-k$ marches restantes. Ainsi, $g_p(\{d_i\}; \alpha) = g_{p-k}(\{d_i | d_i \neq 0\}; \alpha)$.

La diffusion des p lignées (sur le réseau en diamants de la figure 9.2) induit l'équation aux différences finies suivante pour la fonction génératrice $g_p(x_1, \ldots, x_p; \alpha)$:

$$g_p(x_1, \ldots, x_p; \alpha) = \frac{1}{2^d} \sum_{\sigma_1 = \pm 1/2} \ldots \sum_{\sigma_p = \pm 1/2} g_p(x_1 + \sigma_1, \ldots, x_p + \sigma_p; \alpha) e^{-\alpha}. \tag{9.15}$$

On vérifie alors que la solution générale de l'équation précédente qui satisfait les conditions aux bords ci-dessus est donnée par :

$$g_p(x_1, \ldots, x_p; \alpha) = \sum_{i=1}^{p} \frac{\sinh(\gamma d_i)}{\sinh(\gamma L)} \tag{9.16}$$

où le coefficient γ satisfait l'équation

$$e^{\alpha} = \frac{1}{4} e^{\gamma} + \frac{1}{4} e^{-\gamma} + \frac{1}{2}. \tag{9.17}$$

Dans la limite $L \to \infty$ et pour des distances entre individus qui varient comme $d_i = \delta_i L$, l'expression (9.16) montre qu'il faut considérer γ de l'ordre de $1/L$ et, d'après (9.17), il faut α d'ordre $1/L^2$. Nous avons ainsi la limite suivante :

$$g_p\left(\xi_1 L, \ldots, \xi_p L; \frac{\alpha'}{L^2}\right) \xrightarrow{L \to \infty} \tilde{g}_p(\xi_1, \ldots, \xi_p; \alpha') = \sum_{i=1}^{p} \frac{\sinh(2\sqrt{\alpha'}\delta_i)}{\sinh(2\sqrt{\alpha'})} \tag{9.18}$$

avec la notation $\delta_i = d_i/L = \xi_{i+1} - \xi_i$.

Le temps de coalescence normalisé T_p/L^2 de p marches aléatoires de positions initiales aléatoires uniformément distribuées est ainsi donné, dans la limite $L \to \infty$ et en dimension 1, par la fonction génératrice :

$$\boxed{\overline{g}_p(\alpha') = \int_{[0,1]^L} \tilde{g}_p(\xi_1, \ldots, \xi_p; \alpha') d^p\xi = n(n-1) \int_0^1 (1-x)^{n-2} \frac{\sinh(2\sqrt{\alpha'}x)}{\sinh(2\sqrt{\alpha'})} dx}. \tag{9.19}$$

L'expression de $\overline{g}_2$ est conforme à (9.13) et montre que T_2 n'est pas distribué de manière exponentielle. Néanmoins, il est possible de calculer les rapports des moyennes des temps de coalescence T_p et nous obtenons :

$$\boxed{\frac{\langle T_p \rangle}{\langle T_2 \rangle} \xrightarrow{L\to\infty} \frac{2(n+4)(n-1)}{(n+2)(n+1)}} \tag{9.20}$$

et en particulier les valeurs $\langle T_3 \rangle / \langle T_2 \rangle \to 7/5$ et $\langle T_4 \rangle / \langle T_2 \rangle \to 8/5$, qui sont conformes aux valeurs numériques mesurées [BDS08].

9.2 Modèles spatiaux avec sélection

9.2.1 Quelques modèles reliés à une évolution avec sélection

Pour construire des modèles spatiaux avec sélection, nous reprenons le même schéma que dans la section précédente, en ajoutant une caractéristique à chaque individu sur chaque site, son *fitness*. Il faut alors spécifier les deux mécanismes supplémentaires suivants :

- comment le choix du parent dans les plus proches voisins de la génération précédente est biaisé par les *fitnesses* de ces parents potentiels (les hauts *fitnesses* se reproduisent plus en moyenne) ;
- comment, une fois le parent choisi, le *fitness* de l'individu est obtenu en fonction du *fitness* de son parent (mutations bénéfiques ou délétères).

Nous allons voir que cette situation correspond déjà à un certain nombre de modèles de physique statistique. Pour chacun d'eux, nous préciserons la manière dont est choisi le parent et le bruit lié aux mutations.

Les polymères dirigés

Le modèle des polymères dirigés a pour but de décrire les configurations d'un polymère en milieu désordonné. Nous considérons ici une géométrie spatiale identique à celle de la section précédente (cf. le réseau de la figure 9.2). L'espace où évolue le polymère est de dimension $1+d$ où la première dimension joue un rôle différent des autres : le polymère ne peut aller que dans une direction dans cette dimension (d'où l'adjectif « dirigé » dans le nom du modèle). Cette dimension peut ainsi servir à paramétrer la chaîne du polymère. Nous allons voir dans la suite qu'elle joue un rôle analogue au temps dans les modèles de croissance et les modèles d'évolution : pour cette raison nous noterons τ ou t la coordonnée du polymère dans cette direction et la chaîne du polymère peut être paramétrée par une fonction $x(\tau)$ où $x(\tau)$ est le vecteur d-dimensionnel qui donne la position du polymères dans les d dimensions supplémentaires.

Chaque lien du réseau porte une énergie ϵ indépendante des énergies des autres liens et cette énergie est distribuée aléatoirement selon une loi $\mu(\epsilon)$: le paysage énergétique dans lequel croît le polymère est donc désordonné.

Un polymère dirigé est décrit comme une ligne brisée qui part d'un site x_0 à $t = 0$ et arrive sur un site x_t à l'instant t, en passant par les liens (x_0, x_1), (x_1, x_2),..., (x_{t-1}, x_t). L'énergie totale de la chaîne est la somme des énergies des liens parcourus :

$$E\big((x_\tau)_{\tau=0,\dots,t}\big) = \sum_{\tau=0}^{t-1} \epsilon_{(x_\tau,\tau)\to(x_{\tau+1},\tau+1)} \tag{9.21}$$

où $\epsilon_{(x,\tau)\to(y,\tau+1)}$ est l'énergie du lien reliant le point x atteint à la génération τ au point y atteint à la génération suivante.

Pour connaître les propriétés thermodynamiques d'un polymère dirigé, il faudrait déterminer l'énergie libre de ce système en fonction de la température. Néanmoins, même à température nulle, la détermination de l'énergie de l'état fondamental d'un tel polymère n'est pas triviale puisqu'il existe de nombreux minima locaux d'énergie. C'est pourquoi nous ne nous intéresserons ici ni à la caractérisation de la rugosité du polymère, ni à la détermination de son énergie libre, mais seulement aux propriétés de recouvrement des configurations d'énergie minimale de plusieurs polymères d'extrêmités fixées (cf. [KZ87]), qui sont reliées aux corrélations des énergies entre sites. Pour un désordre donné, intéressons-nous aux énergies $E_t(x, y_1)$ et $E_t(x, y_2)$ des états fondamentaux de deux polymères qui commencent tous deux en x à l'instant $\tau = 0$ et aboutissent en y_1 pour le premier et y_2 pour le second à l'instant t. Il existe un temps T_2 tel que les polymères passent par les mêmes liens pour $\tau < t - T_2$ et passent par des liens différents pour $t - T_2 \leq \tau \leq t$. L'énergie des deux polymères se décompose alors en une partie commune pour $\tau < t - T_2$ et deux contributions différentes pour $\tau > t - T_2$. En appelant X le site où se trouvent les deux chaînes au temps $t - T_2$, nous avons ainsi :

$$\begin{cases} E_t(x, y_1) & = E_{t-T_2}(x, X) + E_{T_2}(X, y_1), \\ E_t(x, y_2) & = E_{t-T_2}(x, X) + E_{T_2}(X, y_2). \end{cases} \tag{9.22}$$

La durée T_2 est donc reliée au recouvrement entre les deux polymères qui est à l'origine de la contribution commune $E_{t-T_2}(x, X)$. Les corrélations entre énergies fondamentales de p polymères qui partent d'un même site x de départ et arrivent en p points différents au temps t font ainsi intervenir des temps T_p qui correspondent précisément aux temps où les polymères se rejoignent sur un même site.

Écrivons à présent l'algorithme de construction du chemin d'énergie minimale d'un polymère en termes d'évolution de population. Considérons un polymère qui part de l'origine $\vec{x} = \vec{0}$ à $\tau = 0$ et arrive au point $\vec{y}_t$ au temps $\tau = t$ tout en ayant l'énergie la plus petite possible. Ce polymère est donc passé à l'instant $t - 1$ par une position $\vec{y}'_{t-1}$ avec une énergie $E_{t-1}(\vec{x}, \vec{y}'_{t-1})$. Pour construire trouver le chemin optimal de $\vec{0}$ à $\vec{y}_t$, il faut donc trouver la position $\vec{y}'_{t-1}$ qui minimise $E_{t-1}(\vec{x}, \vec{y}'_{t-1}) + \epsilon(\vec{y}'_{t-1} \to \vec{y}_t)$. Une fois, cette position trouvée, le polymère optimal à t est obtenue en ajoutant le lien $\vec{y}'_{t-1} \to \vec{y}_t$ au polymère optimal d'énergie $E_{t-1}(\vec{x}, \vec{y}'_{t-1})$ à $t - 1$. Nous avons ainsi la construction récursive suivante pour $E_t(\vec{x}, \vec{y}_t)$:

$$\boxed{E_t(\vec{x}, \vec{y}_t) = \min_{\vec{y}'_{t-1}, \vec{y}_t \text{ voisins}} \Big(E_{t-1}(\vec{x}, \vec{y}'_{t-1}) + \epsilon(\vec{y}'_{t-1} \to \vec{y}_t) \Big)}. \tag{9.23}$$

La minimisation précédente fournit ainsi un unique[3] site parent $\vec{y}'_{t-1}$ à l'instant $t-1$ dont le site $\vec{y}_t$ hérite l'énergie à un bruit $\epsilon(\vec{y}'_{t-1} \to \vec{y}_t)$ près. Les polymères dirigés en dimension $1+d$ entrent ainsi dans le cadre de modèles de sélection spatiaux en dimension d et nous allons comparer les généalogies avec celles obtenues en champ moyen.

Modèles de croissance de surface

Un modèle du même type est celui de la croissance d'une surface d dimensionnelle dans un espace de dimension $d+1$ au cours du temps. Nous nous intéressons à la hauteur $h(\vec{x}, t)$ de la surface au point $\vec{x} = (x_1, \ldots, x_d)$ au cours du temps. Lors de la croissance d'une surface, les nouveaux atomes se fixent préférentiellement de façon à maximiser le nombre de voisins : ainsi, un nouvel atome se fixera préférentiellement dans un trou ou dans un coin plutôt qu'au milieu d'une surface plane. Cet attachement préférentiel a tendance à lisser la surface, alors que le dépôt aléatoire au milieu d'une surface lisse tend à la rendre rugueuse. Un mécanisme simple de croissance est le modèle de croissance polynucléaire (*PNG : polynuclear growth model* [PS02, Joh03]) qui est relié à l'équation de Kardar-Parisi-Zhang (KPZ) [KPZ86] et fait donc partie de la même classe d'universalité que les polymères dirigés. Sa dynamique est la suivante :

- à chaque pas de temps et en chaque point $\vec{x}$, des atomes s'ajoutent de manière déterministe de telle sorte que la surface au point $\vec{x}$ se mette au niveau de son voisin le plus haut (s'il existe un voisin plus haut que la hauteur en $\vec{x}$) ;
- des aspérités apparaissent aléatoirement en un point $\vec{x}$ et correspondent à une augmentation aléatoire de la hauteur en un point.

Nous avons ainsi l'évolution suivante (cf. figure 9.3) :

$$h(\vec{x}, t+1) = \max_{\vec{x} \text{ ou } \vec{x}' \text{ voisins de } \vec{x}} \Big(h(\vec{x}', t) \Big) + \omega(\vec{x}, t+1) \tag{9.24}$$

où les $\omega(x, t)$ sont des variables aléatoires positives indépendantes.

À chaque pas de temps, cela revient à considérer le modèle de sélection suivant : sur chaque site est présent un unique individu de *fitness* $h(x, t)$ et, à chaque pas de temps, cet individu choisit comme parent dans la génération précédente celui, parmi les sites voisins, qui a la hauteur la plus haute. Son nouveau *fitness* est alors donné par la hauteur $h(x, t)$ à laquelle s'ajoute une mutation aléatoire $\omega(x, t+1)$.

Les deux types de modèles précédents sont connus pour être dans la classe d'universalité de l'équation KPZ [HHZ95, KPZ86]. L'analogie avec les modèles de sélection incite à penser qu'en grande dimension, les généalogies sont décrites par le coalescent de Bolthausen-Sznitman. Dans la prochaine section, nous nous intéressons également à ces généalogies lorsque la dimension est basse.

9.2.2 Résultats numériques sur les temps de coalescence

Nous avons réalisé des simulations pour les modèles précédents en plusieurs dimensions ($d = 1$, 2, 3 et en champ moyen) et nous avons mesuré les quantités suivantes :

[3] Cela est vrai dès que le bruit ϵ n'est pas à valeurs discrètes.

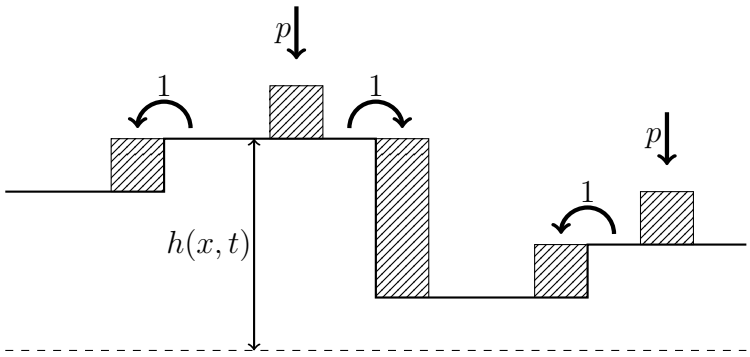


FIG. 9.3: Modèle de croissance de surface (PNG) unidimensionnel : à chaque discontinuité de hauteur, le plateau le plus haut envahit le plateau le plus bas avec une probabilité 1. Des atomes supplémentaires se déposent aléatoirement en chaque point avec une probabilité p.

– l'échelle du temps moyen $\langle T_2 \rangle$ de coalescence à 2 individus (figure 9.4),
– les rapports $\langle T_4 \rangle / \langle T_2 \rangle$ et $\langle T_3 \rangle / \langle T_2 \rangle$ qui donnent un premier aperçu (figure 9.5) d'une possible appartenance à l'une des classes d'universalité (cf. tableau 8.1, page 109).

Les modèles de polymères dirigés que nous avons considérés sont les suivants :

– en dimension finie, le réseau est un réseau cubique décalé de $(1/2, \ldots, 1/2)$ à chaque génération de telle sorte que le parent d'un individu est choisi parmi ces 2^d voisins de la génération précédente (cf. figure 9.2).
– en dimension infinie (champ moyen), le parent peut être tout individu de la génération précédente[4].
– l'énergie aléatoire ϵ sur chaque lien a une distribution[5] uniforme sur $[0, 1]$.

Dans les figures présentées ci-dessous et la suite du texte, L désigne l'extension spatiale du système et N, le nombre total de sites du système ; ces deux nombres sont reliés par la relation $N = L^d$ où d désigne la dimension. La figure 9.4 montre que l'échelle de temps caractéristique des généalogies correspond bien à l'échelle de temps attendue pour les polymères dirigés dans les différentes dimensions (les droites correspondent à des valeurs des exposants critiques obtenues dans d'autres travaux numériques antérieurs [HHZ95]).

Une fois les temps de coalescence normalisés par $\langle T_2 \rangle$, la figure 9.5 montre que, pour les grandes tailles N, les rapports $\langle T_4 \rangle / \langle T_2 \rangle$ et $\langle T_3 \rangle / \langle T_2 \rangle$ semblent converger vers des valeurs limites qui dépendent de la dimension. Quelle que soit la dimension, ces valeurs s'éloignent notablement des valeurs attendues dans le coalescent de Kingman, qui correspond à des modèles sans sélection. En champ moyen, elles semblent être en accord avec celles du coalescent de Bolthausen-Sznitman qui ont déjà été observées dans les modèles de marches aléatoires avec branchements des sections 8.4.2 et 8.4.3. L'idée d'une telle relation n'est pas nouvelle puisque les polymères dirigés sur des arbres désordonnés sont

[4]Nous avons aussi considéré dans les simulations numériques la variante suivante : pour chaque individu, deux (ou tout autre nombre fini) parents potentiels sont choisis aléatoirement dans la génération précédente et ensuite l'optimisation de l'énergie est faite sur ces deux parents. Les rapports des temps de coalescence sont inchangés et nous retrouvons les courbes des figures 9.4, 9.5.

[5]Là encore, considérer une distribution exponentielle à la place ne change rien.

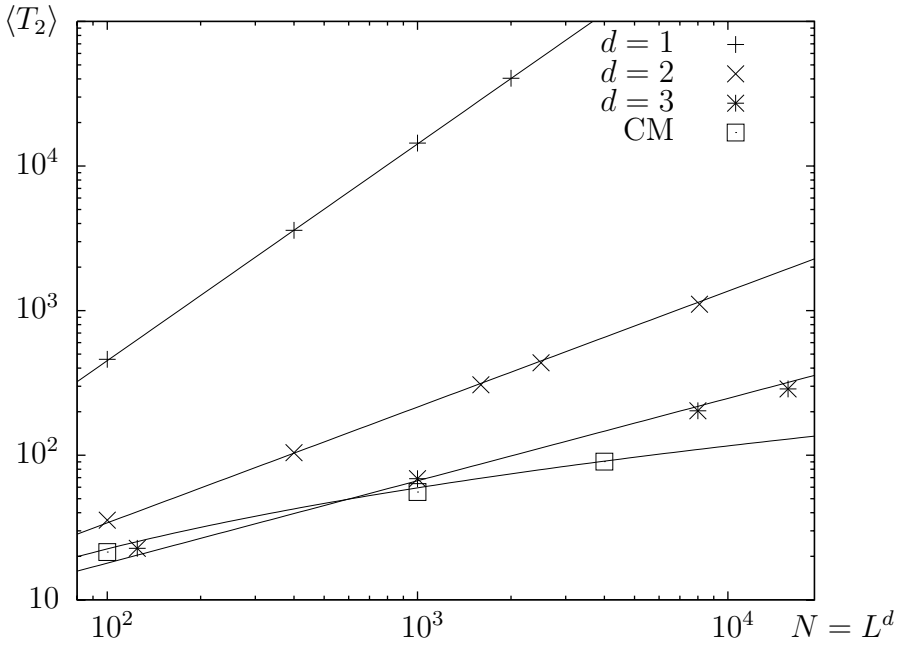


FIG. 9.4: Temps de coalescence moyen $\langle T_2 \rangle$ de deux individus en fonction de la taille du système pour différentes dimensions dans le modèle de polymères dirigés. La taille est donnée par $N = L^d$ où L est la longueur du système. Les droites représentent l'échelle $N^{z/d}$ attendue pour les temps caractéristiques des polymères dirigés dans les différents cas et correspondent aux exposants critiques $z = 3/2$ (exact), $z = 1.60$ pour $d = 2$ et $z = 1.709$ pour $d = 3$ [HHZ95]. En champ moyen, la ligne représente l'échelle $\ln N + 3 \ln \ln N$ attendue.

connus [DS88] pour être reliés à des propagations de fronts (au niveau du calcul de la fonction de partition), qui, comme nous l'avons vu dans la première partie, apparaissent aussi dans l'étude des marches aléatoires avec branchements.

En faible dimension ($d = 1$ et 2), les valeurs mesurées de $\langle T_3 \rangle / \langle T_2 \rangle$ et $\langle T_4 \rangle / \langle T_2 \rangle$ diffèrent sensiblement de celles obtenues en champ moyen et semblent indiquer d'autres classes d'universalité en basse dimension. Les simulations ont néanmoins une convergence trop lente pour espérer déceler dans les temps de coalescence une signature de la dimension critique supérieure : pour $d = 3$, la valeur du rapport $\langle T_4 \rangle / \langle T_2 \rangle$ semble en accord avec la valeur attendue en champ moyen alors que le rapport $\langle T_3 \rangle / \langle T_2 \rangle$ semble tendre vers une valeur proche mais différente.

Ces résultats numériques mettent en évidence l'influence de la dimension spatiale dans les problèmes de sélection. Dans le chapitre précédent, nous avons vu que les généalogies étaient reliées aux fractions de la population générées par chaque individu (sections 8.2.3 et 8.4.3) : si nous considérons les fractions $u_i(t)$ de la population au temps t qui descendent de l'individu initial i, au cours du temps certaines d'entre elles s'éteignent alors que d'autres grandissent, pour aboutir finalement à la domination de l'une d'entre elles. Il serait intéressant de comprendre comment la structure spatiale modifie l'évolu-

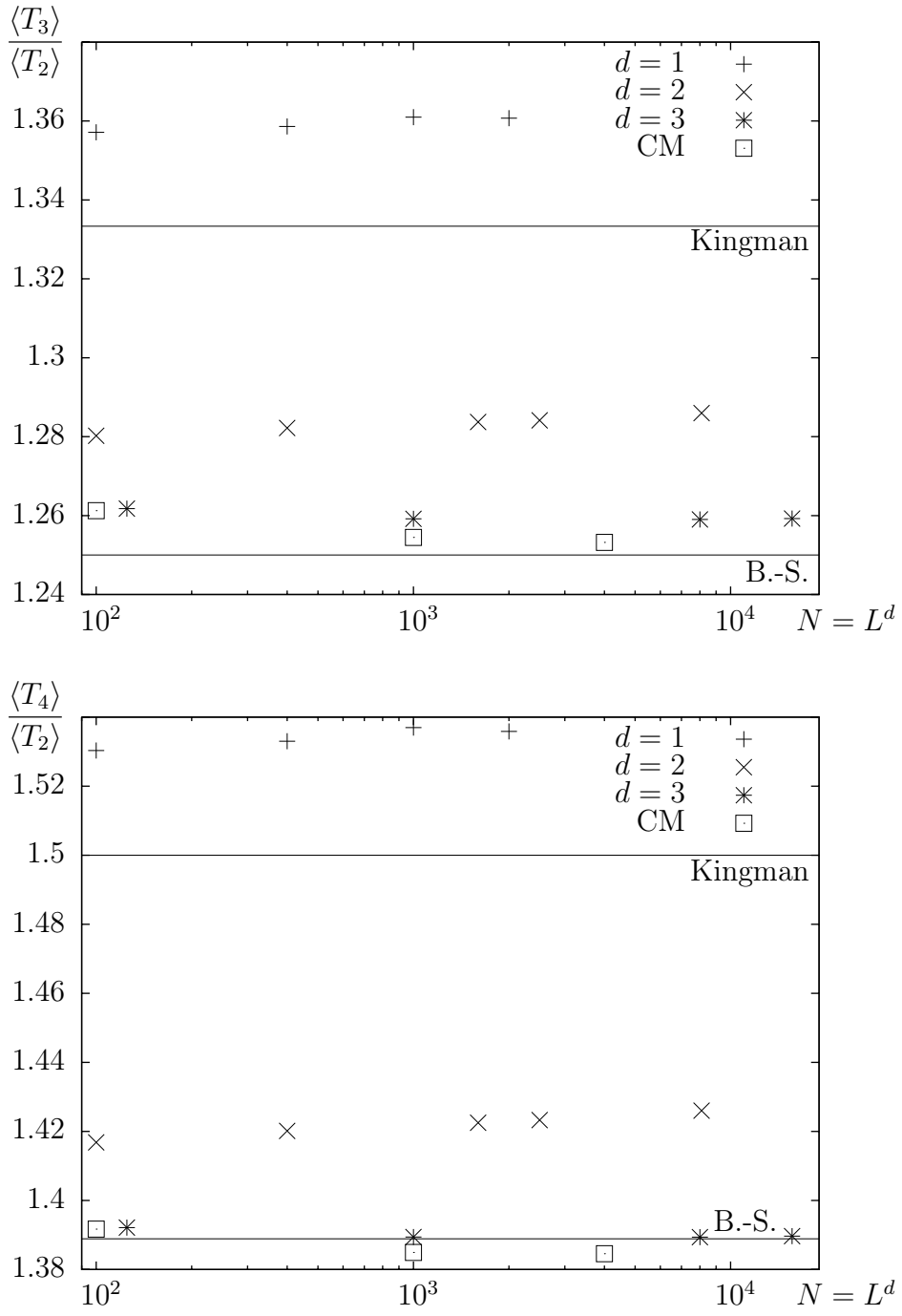


FIG. 9.5: Rapport des temps de coalescence $\langle T_3\rangle/\langle T_2\rangle$ et $\langle T_4\rangle/\langle T_2\rangle$ dans le modèle de polymères dirigés en dimensions $d = 1, 2, 3$ et en champ moyen. La distribution $\rho(\epsilon)$ de l'énergie aléatoire sur chaque lien est uniforme sur $[0, 1]$. L'attribution du parent dans le modèle de champ moyen se fait de la manière suivante : pour chaque individu à la génération $t+1$, deux individus sont tirés au hasard dans la génération précédente puis le parent est choisi parmi eux de manière à minimiser l'énergie totale au temps t d'après (9.23).

tion de ces fractions au cours du temps. Alors qu'en dimension $d > 1$, cela demande de considérer la géométrie des frontières entre les domaines de taille $u_i L^d$ de la population (potentiellement non connexes), en dimension 1 au contraire, les frontières sont réduites à des points qui diffusent (en interagissant) sur le cercle et s'annihilent : l'existence de résultats exacts [PS02, Joh03] pour les modèles de polymères dirigés et de croissance de surface entrouve une possibilité pour étudier les généalogies en dimension $d = 1$.

Chapitre 10

Dynamique des généalogies en absence de sélection

Ce chapitre présente les résultats relatifs à la dynamique des temps de coalescence que nous avons obtenus durant cette thèse et publiés dans [SD06]. Les propriétés stationnaires du modèle de Wright-Fisher ont été rappelées au chapitre 8 et nous étudions dans ce chapitre l'aspect dynamique des généalogies et de l'âge de l'ancêtre commun d'une population. La majorité des résultats sont obtenus dans la limite de grandes populations. La première section consiste en une approche numérique de la dynamique des généalogies. La seconde section construit un processus de Markov sur les généalogies qui permet de comprendre analytiquement cette dynamique et la troisième section utilise ce processus pour obtenir les fonctions d'autocorrélation de l'âge de l'ancêtre commun le plus récent d'une population.

10.1 Dynamique de l'âge de l'ancêtre commun le plus récent d'une population : premiers résultats

L'âge T de l'ancêtre commun le plus récent de toute la population a une valeur moyenne *finie* (8.24). En effet, aux temps longs, parmi les N individus initialement présents, toutes les lignées sauf une finissent par s'éteindre (diffusion de Wright-Fisher). Néanmoins, au cours du temps, T est une variable dynamique et nous allons, dans cette section, caractériser son évolution. La première section présente les résultats numériques et les interprétations basiques que l'on peut en tirer et les deux suivantes expliquent, par un calcul analytique, l'allure des résultats numériques.

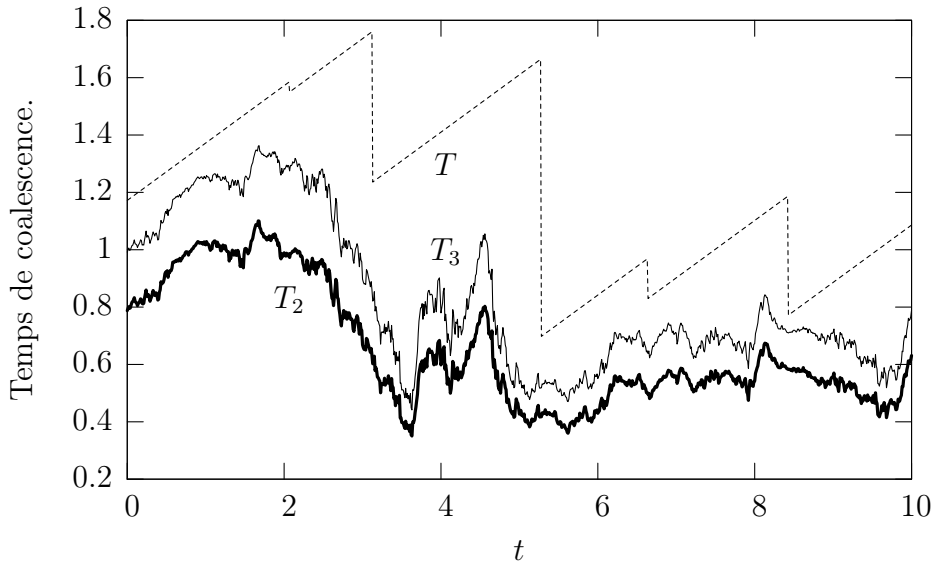


FIG. 10.1: Évolution de l'âge $T(t)$ de l'ancêtre commun le plus récent d'une population de $N = 500$ individus dans le modèle de Wright-Fisher, sur une durée normalisée $\Delta t = 10$ correspondant à 5000 générations. La ligne pointillée représente l'évolution de $T(t)$ alors que les lignes continues représentent les évolutions des temps de coalescence moyennées sur la population à l'instant t de 2 individus ($T_2(t)$, trait gras) et 3 individus ($T_3(t)$, trait fin). On remarquera que les sauts de $T(t)$ sont précédés d'une décroissance des quantités $T_2(t)$ et $T_3(t)$.

10.1.1 Simulations numériques

Nous avons utilisé la dynamique de Wright-Fisher décrite précédemment pour réaliser une simulation sur une population de 500 individus, dont les résultats sont présentés dans la figure 10.1. La dynamique observée pour l'âge T de l'ancêtre commun le plus récent de la population au cours du temps a une allure simple faite de périodes de durées variables d'augmentation linéaire de T (pente 1) séparées par des sauts abrupts décroissants sur une unique génération. Des observations similaires avaient été réalisées précédemment dans [Ser05].

Comme expliqué en section 8.3.2, la population peut être divisée en deux groupes selon les deux branches issues de l'ancêtre commun le plus récent. Tant qu'aucun des deux groupes ne disparaît lors du passage de la génération t à la génération $t + 1$, l'ancêtre commun le plus récent est toujours le même et son âge passe ainsi de T à $T+1$. Au contraire, à l'instant même où l'un des deux sous-groupes disparaît alors l'ancêtre commun le plus récent précédent n'est plus le plus récent et il est ainsi remplacé par un individu qui aura vécu dans une génération ultérieure (voir schéma de la figure 10.2). L'âge de l'ancêtre commun le plus récent passe ainsi de T à $T' < T+1$ et provoque ainsi l'une des discontinuités observées dans la figure 10.1.

Nous pouvons ainsi numéroter les discontinuités successives par un entier k et définir les quantités suivantes :

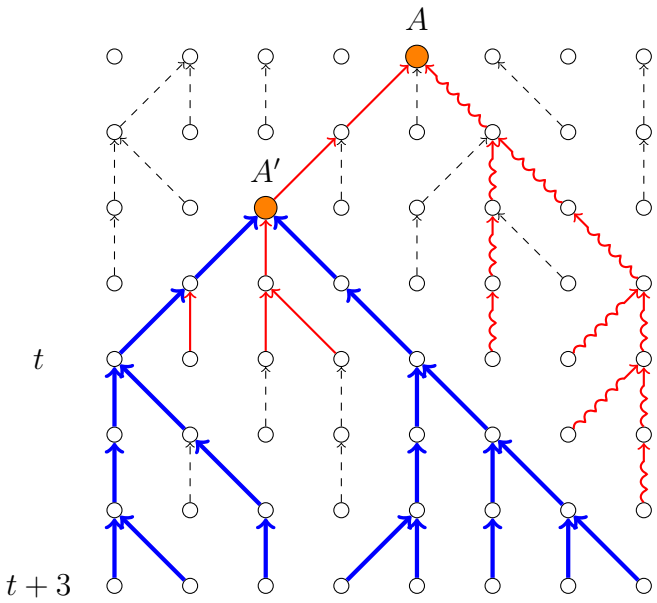


FIG. 10.2: Mécanisme d'une discontinuité de l'âge T de l'ancêtre commun le plus récent d'une population : explication pour une taille $N = 8$ (le temps s'écoule vers le bas). Les traits pointillés correspondent aux relations de parenté qui ne sont pas pertinentes. Les traits pleins correspondent aux généalogies de la population aux temps $t = 5$ et $t = 8$. Les traits *ondulés* représentent la lignée responsable du changement d'ancêtre commun le plus récent qui se produit à la dernière génération. Au temps t, les trois individus sur la droite descendent de l'ancêtre commun le plus récent A par une autre lignée que les cinq premiers. Aux générations $t+1$ et $t+2$, la taille de leur groupe diminue mais l'ancêtre commun le plus récent reste toujours A. À la génération $t+3$ cependant, ce groupe s'éteint : toute la partie ondulée de la généalogie disparaît de l'arbre généalogique de la population et l'ancêtre commun le plus récent de la population devient l'individu A'. Il se produit alors une discontinuité de T.

– D_k est le délai entre les $(k-1)$-ème et k-ème discontinuités ;
– H_k est la hauteur de la k-ème discontinuité.

La connaissance des propriétés statistiques de ces quantités suffit à caractériser entièrement l'évolution de l'âge T de l'ancêtre commun le plus récent. La figure 10.3 représente les histogrammes des délais D_k et des hauteurs H_k accumulés sur environ 10^4 discontinuités et met en évidence une distribution exponentielle pour les D_k et les H_k. Les corrélations temporelles entre ces différentes quantités ont été mesurées sur le même échantillon et nous obtenons, à une précision de 0,01 près, les valeurs suivantes (cf. [SD06]) :

$$\langle D_k D_{k-1}\rangle - \langle D_k\rangle\langle D_{k-1}\rangle \simeq -0,005 \tag{10.1a}$$
$$\langle H_k H_{k-1}\rangle - \langle H_k\rangle\langle H_{k-1}\rangle \simeq -0,006 \tag{10.1b}$$
$$\langle D_k H_{k-1}\rangle - \langle D_k\rangle\langle H_{k-1}\rangle \simeq -0,002 \tag{10.1c}$$
$$\langle H_k D_k\rangle - \langle H_k\rangle\langle D_k\rangle \simeq 0,84 \tag{10.1d}$$
$$\langle H_k D_{k-1}\rangle - \langle H_k\rangle\langle D_{k-1}\rangle \simeq 0,12. \tag{10.1e}$$

Il ne semble exister de corrélations qu'entre la hauteur H_k d'une discontinuité et les délais $D_{k'}$, $k' \leq k$, qui la précèdent. Le but des sections qui viennent est de comprendre les distributions exponentielles de la figure 10.3 ainsi que les corrélations (10.1).

10.1.2 Distribution des délais entre deux discontinuités

Nous allons tout d'abord calculer la probabilité $P_{\text{diff}}(t_0, t_1)$ que les ancêtres communs les plus récents de la population au temps t_0 et au temps $t_1 > t_0$ soient différents, c'est-à-dire que t_0 et t_1 soient séparés par au moins une discontinuité de hauteur H_k. Pour cela, décomposons la généalogie de la population prise à l'instant t_1 en deux parties : une partie inférieure qui correspond à la partie de l'arbre généalogique située entre les instants t_0 et t_1 et une partie supérieure qui correspond, quant à elle, à la partie de l'arbre antérieure à t_0 (cf. schéma de la figure 10.4).

Introduisons le nombre m d'ancêtres au temps t_0 de la population au temps t_1 : comme nous l'avons vu en section 8.3, ce nombre est distribué selon $z_m(t_1 - t_0)$ où les z_m sont les fonctions définies en (8.32). D'autre part, la population à l'instant t_0 peut être divisée en deux groupes de taille xN et $(1-x)N$ selon les deux lignées issues de l'ancêtre commun le plus récent de la population (cf. section 8.3.2 pour la manière de construire cette partition). Ainsi, la seule possibilité pour que les ancêtres communs les plus récents des populations aux instants t_0 et t_1 soient distincts correspond au cas où les m ancêtres au temps t_0 de la population au temps t_1 appartiennent tous à l'un des deux groupes de tailles xN et $(1-x)N$ (cf. figures 10.2 et 10.4). Ainsi, la probabilité $P_{\text{diff}}(t_0, t_1)$ est obtenue en effectuant les sommations sur les valeurs possibles de m et x puis en utilisant

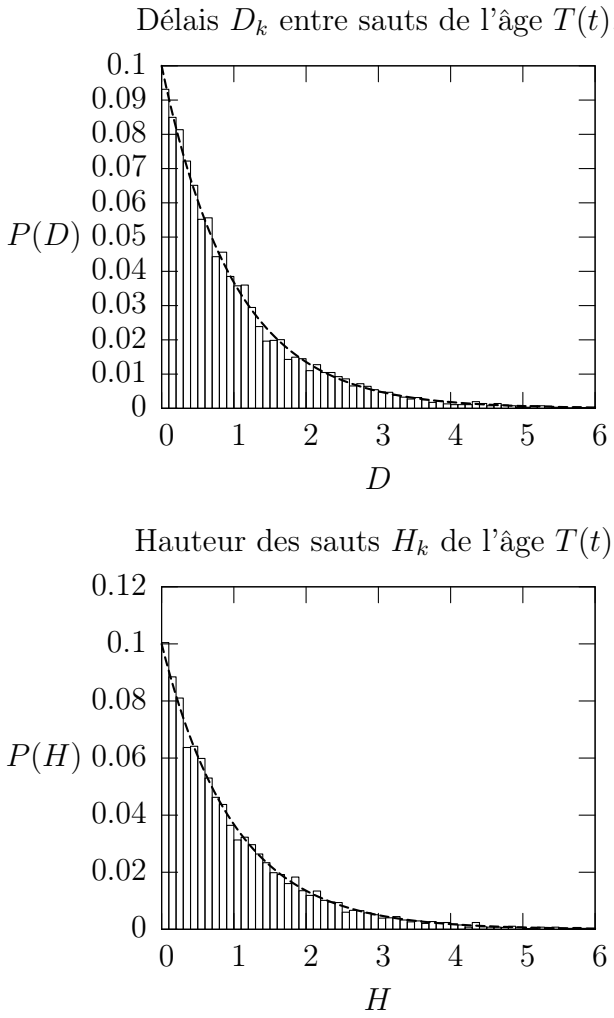


FIG. 10.3: Histogrammes des délais D_k entre sauts (en haut) et hauteurs H_k des sauts (en bas) de l'âge de l'ancêtre commun le plus récent de la population dans le modèle de Wright-Fisher. La taille de la population est $N = 500$ et les temps sont normalisés par la taille de la population. Les statistiques sont accumulées sur 9169 discontinuités. Les lignes pointillées correspondent aux prévisions théoriques (distributions exponentielles de moyenne un).

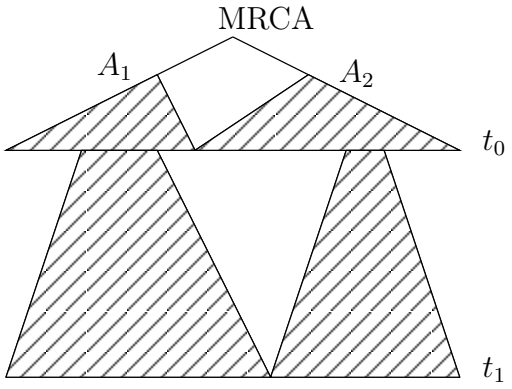


FIG. 10.4: Décomposition de l'arbre de la population à des temps différents : la population au temps t_1 au temps t_1 a m ancêtres dans la population au temps t_0. Cette dernière peut aussi être partitionnée selon la lignée issue de l'ancêtre commun le plus récent à laquelle les individus appartiennent.

les expressions (8.30, 8.32) :

$$\begin{aligned} P_{\text{diff}}(t_0,t_1) &= \int_0^1 dx a(x) \sum_{m=1} z_m(t_1-t_0)\left[x^m+(1-x)^m\right] \\ &= \sum_{m=1} z_m(t_1-t_0)\frac{2}{m+1} \\ &= 2\sum_{p=1}^{\infty}(-1)^p(2p-1)e^{-c_p(t_1-t_0)}\left[\sum_{m=1}^{p}(-1)^m\frac{(m+p-2)!}{(m+1)!(m-1)!(p-m)!}\right]. \end{aligned}$$

L'utilisation de l'identité hypergéométrique[1]

$$\sum_{m=1}^{p}(-1)^m\frac{(m+p-2)!}{(m+1)!(m-1)!(p-m)!} = \begin{cases} -\frac{1}{2} & \text{si } p=1 \\ -\frac{1}{6} & \text{si } p=2 \\ 0 & \text{si } p\geq 3 \end{cases}$$

ramène l'expression de $P_{\text{diff}}(t_0,t_1)$ à une somme de deux termes :

$$\boxed{P_{\text{diff}}(t_0,t_1) = 1-e^{-(t_1-t_0)}}. \tag{10.2}$$

[1]Nous avons en effet [GR94] :

$$\sum_{m=1}^{p}(-1)^{m-1}z^{m-1}\frac{(m+p-2)!}{(m+1)!(m-1)!(p-m)!} \propto J_{p-1}^{(2,-2)}(1-2z)$$

où $J_{p-1}^{(2,-2)}(z)$ est le polynôme de Jacobi défini par :

$$J_n^{(a,b)}(z) = \frac{(-1)^n}{2^n n!}(1-x)^{-a}(1+x)^{-b}\frac{d^n}{dx^n}\left[(1-x)^{a+n}(1+x)^{b+n}\right]$$

Puisque, par définition, les délais D_k séparent les phases où l'ancêtre commun le plus récent de la population reste le même, la distribution stationnaire des délais $p_{\text{délai}}(D)$ est donnée par :

$$\boxed{p_{\text{délai}}(D) = -\frac{d}{dt} P_{\text{diff}}(0,t)\Big|_{t=D} = e^{-D}} \tag{10.3}$$

et est en accord avec les résultats numériques de la figure 10.3.

10.1.3 Hauteur des discontinuités

Nous nous attachons maintenant à caractériser les hauteurs H_k des discontinuités de $T(t)$. Le schéma de la figure 10.2 donne une interprétation des changements d'ancêtre commun le plus récent. Regardons, dans cette image, ce qu'il advient des intervalles de temps entre coalescence en haut de l'arbre en nous plaçant dans l'arbre généalogique de la population à la hauteur où il ne reste plus que n ancêtres. Le haut de l'arbre généalogique de la population est l'arbre de ces n individus. On peut alors définir les temps τ_p comme les délais entre les coalescences faisant passer de p à $p-1$ individus dans l'arbre. Les propriétés stationnaires de ces temps ont été données en section 8.3.1.

Lors de l'avant dernière coalescence avant l'ancêtre commun le plus récent, nous passons de trois individus à deux individus. Appelons A_1 et A_2 ces deux derniers individus (cf. figure 10.2). Lorsque T subit une discontinuité, cela signifie nécessairement que l'une des lignées de A_1 et A_2 s'éteint. Supposons que ce soit celle de A_2. Le nouvel ancêtre commun devient alors A_1 : la discontinuité de T est donc donnée par le temps τ_2 juste avant la discontinuité (durée entre les deux dernières coalescences de l'arbre). Ainsi, la hauteur des discontinuités doit être distribuée comme τ_2 : d'après (8.18) c'est donc, après normalisation par la taille N de la population, une distribution exponentielle de moyenne 1

$$\boxed{p_{\text{hauteur}}(H) = e^{-H}}, \tag{10.4}$$

qui est compatible avec l'histogramme de la figure (10.3).

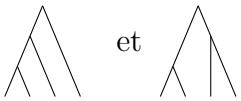
Ce résultat n'est valable que dans la limite $N \to \infty$: en effet, pour que le nouvel ancêtre commun le plus récent ne soit pas l'individu A_1, il faudrait que *simultanément* l'une des deux lignées issues de A_1 et celle de A_2 s'éteignent (ce qui ferait que le nouvel ancêtre commun le plus récent serait plus bas dans l'arbre), or de tels événements ont une probabilité négligeable.

10.2 Dynamique des généalogies : la cascade des longueurs de branches

10.2.1 Description du mécanisme

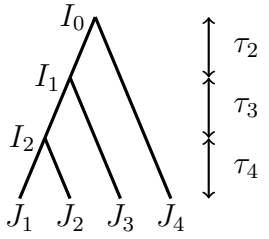
Nous allons à présent essayer de comprendre l'origine des corrélations (10.1). Pour cela, focalisons-nous sur le haut de l'arbre généalogique de la population jusqu'au niveau

$n = 4$. Nous obtenons les deux topologies suivantes :



Si l'une des quatre lignées issues des quatre individus du bas s'éteint dans l'arbre de droite, cela change la structure de l'arbre mais ne provoque pas de saut de l'ancêtre commun le plus récent ; dans l'arbre de gauche, au contraire, l'extinction de la lignée du quatrième individu, et elle seule, provoque un saut de l'ancêtre commun le plus récent de la population.

Dans l'arbre de gauche, nommons les individus qui nous intéressent de la manière suivante et rappelons les définitions des temps τ_p :



Nous avons ainsi les évolutions suivantes selon celle des quatre lignées qui s'éteint :

1. Si la lignée issue de J_4 s'éteint, nous avons une discontinuité de $T(t)$ et l'ancêtre commun le plus récent de la population qui était préalablement I_0 devient I_1 ; le rôle de I_1 est alors joué par I_2, etc. Les nouveaux délais τ'_p sont alors donnés par $\tau'_p = \tau_{p+1}$.
2. si c'est la lignée de J_3 qui s'éteint, alors l'ancêtre commun le plus récent reste I_0 mais le rôle de I_1 est alors joué par I_2. Le temps τ_2 devient alors $\tau'_2 = \tau_2 + \tau_3$ et $\tau'_3 = \tau_4$.
3. dans tous les cas néanmoins, quelle que soit la lignée qui s'éteint, le temps τ_4 est remplacé par la valeur de τ_5, qui n'est pas représentée sur le schéma. De plus, le rôle de I_2 est alors joué par l'un des J_k qui dépend de la structure de l'arbre jusqu'à $n = 5$.

Si nous tronquons l'arbre à la hauteur K, l'extinction de l'une des K lignées change la structure de l'arbre et redéfinit les temps τ_p car l'une des branches est effacée. Si cette branche est rattachée à la k-ième coalescence à partir du sommet de l'arbre, nous avons le processus suivant pour les τ_p :

$$\begin{cases} \tau'_p = \tau_p & \text{si } p < k, \\ \tau'_k = \tau_k + \tau_{k+1}, & \\ \tau'_p = \tau_{p+1} & \text{si } p > k \text{ et } p \neq K, \\ \tau'_K = \epsilon_K. & \end{cases} \tag{10.5}$$

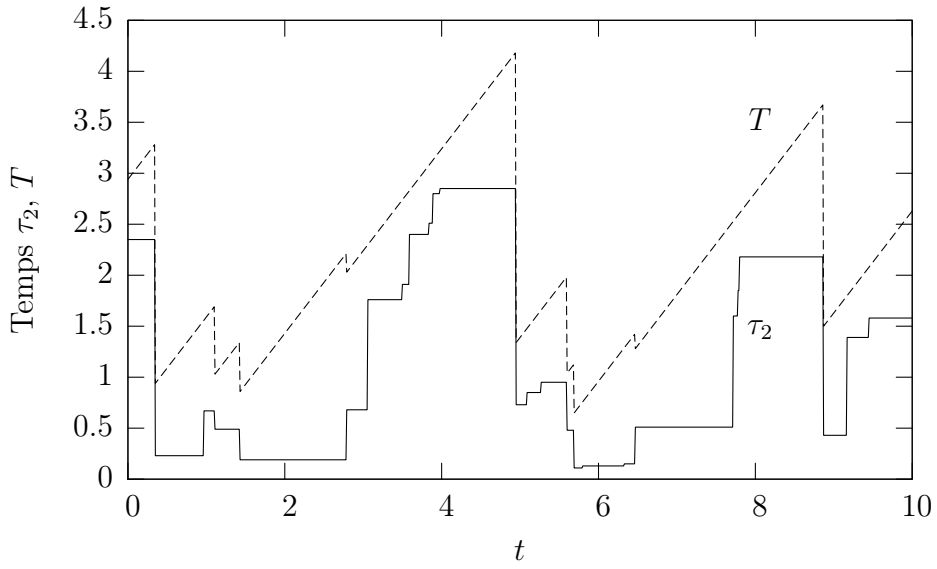


FIG. 10.5: Dynamique des temps de coalescence en haut de l'arbre généalogique : τ_2 est constant par morceaux. Les discontinuités de T ont pour hauteur la valeur de τ_2 au moment précédant le saut et la nouvelle valeur de τ_2 est l'ancienne valeur de τ_3.

Si aucune des K lignées ne s'éteint, alors les τ_p restent inchangés. Si $k = 1$, les deux premières lignes de (10.5) ne jouent aucun rôle : τ_2 disparaît et *tous* les temps sont décalés selon les troisième et quatrième lignes ci-dessus. Puisque T est la somme des τ_p, cela revient à lui ôter la valeur de τ_2 et cela correspond à une discontinuité de T. Au contraire, pour $k \geq 2$, la dynamique précédente ne fait qu'ajouter ϵ_K à la somme des τ_p : dans la limite $K \to \infty$ (cf. ci-dessous), cela correspond à une période d'augmentation linéaire de T. Ainsi, dans la limite $K \to \infty$, nous retrouvons bien la dynamique observée en figure (10.1).

La propriété remarquable de ces temps τ_i que nous avons montré dans [SD06] est que le processus précédent est *markovien* (du moins dans la limite $N \to \infty$) et les temps τ_p évoluent de la manière suivante :

- soit, pendant dt, rien ne change avec une probabilité $1 - (\sum_k q_k)dt$;
- soit l'événement (10.5) se produit avec une probabilité $q_k dt$, avec des taux q_k qui seront déterminés plus bas en section 10.2.2.

Cette dynamique est en accord avec les observations numériques de la figure 10.5.

Il reste à présent à préciser la valeur de ϵ_K. *A priori*, puisque le processus (10.5) ne fait que résumer les événements qui affectent le haut de l'arbre, ϵ_K est donné par τ_{K+1} au moment du saut et doit évoluer au cours du temps. Ainsi, pour tronquer à l'ordre K la dynamique (10.5), il faut prendre pour ϵ_K une variable aléatoire qui soit au moins cohérente avec la distribution stationnaire des τ_p. Il faut donc nécessairement que $\langle \epsilon_K \rangle = \langle \tau_K \rangle$. En fait, dans la limite $K \to \infty$, la condition précédente est aussi suffisante. En effet, la probabilité de renouveler ϵ_K pendant dt vaut $\sum_{k=1}^{K-1} q_k dt$ et diverge pour K

grand : la variable ϵ_K est donc renouvelée de plus en plus souvent si K est grand et sa contribution au reste des temps de coalescence τ_p se concentre autour de sa valeur moyenne. Il suffit ainsi de prendre une valeur constante $\epsilon_K = \langle \tau_K \rangle = 1/c_K$, pour que, dans la limite $K \to \infty$, la dynamique des τ_p pour p fini soit bien celle que nous observons.

Il ne reste alors plus qu'à déterminer la valeur des taux de cascade q_k ainsi que la valeur de ϵ_K dans l'équation (10.5), en prouvant par là-même que le processus de cascade est bien markovien.

10.2.2 Taux de cascade q_k

Deux contributions entrent dans les valeurs des q_k. Si nous considérons l'arbre tronqué à la hauteur K, une cascade (10.5) se produit à la hauteur k lorsque l'une des K lignées s'éteint *et* que, dans l'arbre tronqué, la lignée correspondante est directement reliée au niveau k de l'arbre.

La probabilité qu'une extinction se produise parmi les K lignées au temps t est tout simplement $c_K dt$. Cette valeur se lit en fait directement sur l'équation différentielle (8.31) satisfaite par la fonction $z_K(t)$. En effet, puisque $z_K(t)$ est la probabilité d'avoir K ancêtres lorsque l'on remonte d'une hauteur t dans l'arbre, elle ne varie que lorsque l'une des K lignées s'éteint. De plus, le fait que l'on ait une équation différentielle du type (8.31) pour les fonctions $z_m(t)$ indique que le processus est bien markovien et qu'il n'y a pas d'effet de mémoire.

D'autre part, la probabilité que la lignée qui subit l'extinction soit directement reliée au niveau k de l'arbre est le produit des probabilités que les $(k+1)$-ème, $(k+2)$-ème, ..., $(K-1)$-ème coalescences (en partant du sommet de l'arbre) n'affectent pas cette lignée et la k-ème l'affecte. En comptant les paires possibles pour chacune de ces coalescences, cette probabilité est donc donnée par le produit :

$$\frac{k}{c_{k+1}} \prod_{p=k+2}^{K} \frac{c_p - (p-1)}{c_p} = \frac{k}{c_K}. \tag{10.6}$$

Ainsi, q_k est donné par le produit des probabilités $c_K dt$ et k/c_K et nous avons ainsi :

$$\boxed{q_k = k}. \tag{10.7}$$

Pendant un intervalle dt, l'événement (10.5) se produit ainsi avec la probabilité $k dt$. Ces résultats sont en accord avec les fréquences moyennes de ces événements mesurées numériquement. Lors de l'une de ces cascades, la procédure d'actualisation de l'arbre tronqué est alors aisée : la branche affectée par l'extinction est supprimée et une branche est ajoutée en bas de l'arbre et se raccorde aléatoirement à l'une des $K-1$ restantes.

10.2.3 Fonctions de corrélation des délais τ_i

Le processus stochastique défini par (10.5) avec les taux (10.7) est suffisamment simple pour nous permettre de déterminer les fonctions de corrélation du type :

$$\mathbf{C}_{ij}(t) = \langle \tau_i(t) \tau_j(0) \rangle - \langle \tau_i(t) \rangle \langle \tau_j(0) \rangle \tag{10.8}$$

où les τ_i sont les temps d'attente entre coalescences successives. Un temps τ_i n'est affecté que par les événements (10.5) tels que $k \leq i$. Étant donné les taux de transition (10.7), nous avons la dynamique réduite suivante pour τ_i :

$$\tau_i(t+dt) = \begin{cases} \tau_i(t) & \text{avec probabilité } 1 - c_{i+1}dt \\ \tau_i(t) + \tau_{i+1}(t) & \text{avec probabilité } idt \\ \tau_{i+1}(t) & \text{avec probabilité } c_i dt \end{cases} \tag{10.9}$$

où $c_i = i(i-1)/2 = 1+2+\ldots+(i-1)$. Cela permet, premièrement, de voir que $\tau_i(t)$ n'est corrélé qu'aux temps $\tau_j(0)$ avec $j \geq i$ et, deuxièmement, d'écrire la récurrence suivante[2] sur les fonctions de corrélation $\mathbf{C}_{ij}(t)$:

$$\partial_t \mathbf{C}_{ij} = -c_i \mathbf{C}_{ij}(t) + c_{i+1}\mathbf{C}_{i+1,j}(t). \tag{10.10}$$

Les conditions initiales sont données par les distributions stationnaires (8.18) et la résolution de l'équation précédente par transformée de Laplace donne :

$$\boxed{\int_0^\infty \mathbf{C}_{ij}(t)e^{-st}dt = \begin{cases} 0 & \text{si } i > j \\ \left(\dfrac{1}{c_i c_j^2}\right) \displaystyle\prod_{l=i}^{j} \dfrac{c_l}{s+c_l} & \text{si } i \leq j \end{cases}}. \tag{10.11}$$

Après décomposition en éléments simples, la fonction de corrélation est donnée pour $i \leq j$ par :

$$\mathbf{C}_{ij}(t) = \frac{1}{2c_i c_j^2} \frac{j!(j-1)!}{(i-1)!(i-2)!} \sum_{l=i}^{j} e^{-c_l t} \frac{(i+l-2)!(-1)^{l-i}(2l-1)}{(l-i)!(j-l)!(j+l-1)!}. \tag{10.12}$$

Un corollaire de ce résultat est la décorrélation des hauteurs de discontinuité H_k : en effet, la hauteur de la discontinuité H_{k+1} est égale au temps τ_2 juste avant cette discontinuité. Ce temps ne dépend que des valeurs des temps τ_p avec $p \geq 3$ juste avant la discontinuité H_k et est donc indépendant de H_k, qui est égal à τ_2. On retrouve ainsi le coefficient (10.1b) et le fait que toutes les hauteurs de sauts H_k sont décorrélées.

10.2.4 Application : calcul de la corrélation $\langle D_k H_k \rangle$

La dynamique (10.5, 10.7) permet d'obtenir une expression analytique pour la corrélation (10.1d). Pour cela, plaçons-nous juste après une discontinuité. Il est possible de montrer (cf. notre article [SD06]) le résultat intuitif suivant : juste avant un saut, les τ_i sont distribués selon leurs mesures stationnaires ρ_i, alors que, juste après un saut, chaque τ_i est distribué selon la mesure stationnaire ρ_{i+1}, puisque la dynamique (10.5) avec $k = 1$ ne fait que décaler les τ_i.

[2] On notera la similarité avec l'équation (8.31).

Prenons comme origine des temps une discontinuité de $T(t)$ et introduisons la variable $\eta(t)$ définie de la manière suivante pour $t > 0$:

$$\eta(t) = \begin{cases} 1 & \text{si aucune nouvelle discontinuité ne s'est produite,} \\ 0 & \text{sinon.} \end{cases}$$

Sachant que H_k est donnée par la valeur de τ_2 au moment de la discontinuité, nous avons par définition de $\eta(t)$:

$$\langle D_k H_k \rangle = \int_0^\infty t \langle \eta(t)\tau_2(t) \rangle dt. \tag{10.13}$$

La forme ci-dessus suggère d'introduire la famille de fonctions :

$$\psi_i(s) = \int_0^\infty e^{-st} \langle \eta(t)\tau_i(t) \rangle dt. \tag{10.14}$$

La cascade (10.5) permet de montrer la récurrence suivante :

$$\partial_t \langle \eta(t)\tau_i(t) \rangle = (c_{i+1} - 1)\langle \eta(t)\tau_{i+1}(t) \rangle - c_i \langle \eta(t)\tau_i(t) \rangle. \tag{10.15}$$

Sachant que τ_i est distribué selon ρ_{i+1} juste après la discontinuité initiale, nous obtenons :

$$s\psi_i(s) - \frac{1}{c_{i+1}} = (c_{i+1} - 1)\psi_{i+1}(s) - c_i \psi_i(s). \tag{10.16}$$

Étant donné que nous avons $\langle D_k H_k \rangle = -\psi_2'(0)$, nous pouvons nous contenter de développer au premier ordre :

$$\psi_i(s) = \frac{2}{i(i+1)} \left(u_i + s v_i + o(s) \right)$$

dans la récurrence (10.16). De plus, pour un arbre tronqué à la hauteur K, nous avons $\psi_{K+1}(s) = \langle \epsilon_K \rangle / (1 + \lambda)$. On peut alors vérifier que nous avons [SD06] :

$$\begin{aligned} u_i &= \left(\sum_{j=i}^{K} \frac{2}{j(j-1)} + \frac{\langle \epsilon_K \rangle (K+1)(K+2)}{2} \right) \xrightarrow{K \to \infty} \frac{2}{i-1} + 1, \\ v_i &= \left(\sum_{j=i}^{K} \frac{u_j}{j(j-1)} + \frac{\langle \epsilon_K \rangle (K+1)(K+2)}{2} \right) \xrightarrow{K \to \infty} \frac{2\pi^2}{3} + 1 - \frac{2}{i-1}. \end{aligned}$$

Nous obtenons ainsi la valeur finale suivante :

$$\boxed{\langle D_k H_k \rangle - \langle D_k \rangle \langle H_k \rangle = \frac{2\pi^2}{9} - \frac{4}{3} \simeq 0.86}, \tag{10.17}$$

qui est en accord avec l'observation numérique (10.1d).

Des calculs similaires avec deux variables $\eta_1(t)$ et $\eta_2(t)$ pour compter la deuxième discontinuité après la discontinuité de référence à $t = 0$ permet de calculer sur le même schéma la quantité (10.1c).

10.3 Fonctions d'auto-corrélation des temps de coalescence

Nous avons à présent tous les outils en main pour déterminer les fonctions de corrélations des signaux stochastiques observés en figure (10.1). Ces résultats sont ceux obtenus dans [SD06] où figurent les calculs détaillés de la section 10.3.2.

10.3.1 Temps de coalescence de paires et de triplets d'individus

La figure (10.1) présente, aux côtés de celles de T, les variations de l'âge $\overline{T_2}$ (respectivement $\overline{T}_3$) de l'ancêtre commun de deux (respectivement trois) individus moyenné sur la population. Plus précisément, si $T_{i,j}(t)$ (resp. $T_{i,j,k}(t)$) désigne l'âge de l'ancêtre commun le plus récent des individus i et j (resp. i, j et k), alors $\overline{T_2}(t)$ et $\overline{T_3}(t)$ sont définis par :

$$\begin{aligned}\overline{T_2}(t) &= \frac{2}{N(N-1)}\sum_{i<j} T_{i,j}(t),\\ \overline{T_3}(t) &= \frac{6}{N(N-1)(N-2)}\sum_{i<j<k} T_{i,j,k}(t).\end{aligned}$$

Afin de déterminer la corrélation $\langle\overline{T_2}(t)\overline{T_2}(0)\rangle$, considérons une paire d'individus (i,j) à l'instant t et une autre (k,l) à l'instant 0. Deux situations sont envisageables :

1. soit la coalescence de la paire (i,j) a eu lieu dans l'intervalle de temps $[0,t]$ et les temps de coalescence des deux paires sont indépendants ;
2. soit la coalescence de la paire (i,j) est antérieure au temps 0, auquel cas les lignées de i et j peuvent interagir avec celles des individus (k,l) et des corrélations peuvent alors s'établir entre les temps de coalescence. Dans ce dernier cas, dans la limite $N\to\infty$, la probabilité que l'un des individus (k,l) soit un ancêtre de l'un des individus (i,j) tend vers 0. Sur l'intervalle de temps $]-\infty,0]$, il faut donc considérer les généalogies à *quatre* individus : k, l, l'ancêtre de i et l'ancêtre de j.

La probabilité du premier cas est $1-e^{-t}$. De plus, comme toutes les paires jouent le même rôle, il nous suffit, dans le deuxième cas, de considérer les généalogies de quatre individus $(1,2,3,4)$ arbitraires. Nous obtenons ainsi la décomposition suivante :

$$\langle\overline{T_2}(t)\overline{T_2}(0)\rangle \underset{N\to\infty}{=} \left(\int_0^t \rho_2(\tau_2)\tau_2 d\tau_2\right)\langle T_{1,2}(0)\rangle + e^{-t}\Big\langle (t+T_{3,4}(0))T_{1,2}(0)\Big\rangle. \tag{10.18}$$

Le seul terme qui n'est pas immédiat à calculer est la moyenne $\langle T_{3,4}(0)T_{1,2}(0)\rangle$. Pour ce dernier, il suffit de considérer les généalogies à quatre individus et, selon les formes possibles des arbres, de décomposer les temps $T_{3,4}(0)$ et $T_{1,2}(0)$ sur les temps de coalescence élémentaires τ_2, τ_3 et τ_4. Le tableau 10.1 énumère les types de contributions possibles avec leurs poids (cf. section 8.3.2). Nous obtenons ainsi l'équation suivante :

$$\begin{aligned}\langle T_{3,4}(0)T_{1,2}(0)\rangle =& \frac{2}{18}\langle\tau_4(\tau_4+\tau_3)\rangle + \frac{4}{18}\langle(\tau_4+\tau_3+\tau_2)^2\rangle\\ &+\frac{4}{18}\langle\tau_4(\tau_4+\tau_3+\tau_2)\rangle + \frac{8}{18}\langle(\tau_4+\tau_3)(\tau_4+\tau_3+\tau_2)\rangle = \frac{11}{9}.\end{aligned} \tag{10.19}$$

Arbre	$T_{1,2}$	$T_{3,4}$	Poids
	τ_4	$\tau_4 + \tau_3$	$\frac{2}{18}$
	$\tau_4 + \tau_3 + \tau_2$	$\tau_4 + \tau_3 + \tau_2$	$\frac{4}{18}$
	τ_4	$\tau_4 + \tau_3 + \tau_2$	$\frac{4}{18}$
	$\tau_4 + \tau_3$	$\tau_4 + \tau_3 + \tau_2$	$\frac{8}{18}$

TAB. 10.1: Arbres généalogiques possibles de quatre individus (aux symétries $1 \leftrightarrow 2$ et $3 \leftrightarrow 4$ près) qui interviennent dans l'évaluation de la corrélation $\langle T_{3,4}(0)T_{1,2}(0)\rangle$. Les poids des quatre décompositions possibles de $T_{3,4}(0)$ et $T_{1,2}(0)$ prennent en compte les configurations équivalentes par symétrie.

puis la fonction de corrélation suivante :

$$\boxed{\langle\overline{T_2}(t)\overline{T_2}(0)\rangle - \langle\overline{T_2}(t)\rangle\langle\overline{T_2}(0)\rangle = \frac{2}{9}e^{-t}} \quad . \tag{10.20}$$

La même méthode s'applique à des fonctions de corrélation plus générales $\langle\overline{T_m}(t)\overline{T_n}(0)\rangle$ et demande d'énumérer les arbres généalogiques de $n+2$ à $n+m$ individus à l'instant 0. Nous nous contenterons ici de donner la fonction de corrélation de $\overline{T_3}(t)$:

$$\boxed{\langle\overline{T_3}(t)\overline{T_3}(0)\rangle = \frac{16}{9} + \frac{29}{60}e^{-t} - \frac{13}{900}e^{-3t}} \quad . \tag{10.21}$$

10.3.2 Auto-corrélation de l'âge de l'ancêtre commun le plus récent la population

L'évolution de l'âge $T(t)$ de l'ancêtre commun le plus récent de *toute* la population a été présentée, avec celles de $\overline{T_2}(t)$ et $\overline{T_3}(t)$, en figure 10.1, page 136. Sa fonction de corrélation $\langle T(t)T(0)\rangle$ peut être déterminée de deux manières différentes :

1. soit en passant par le processus de cascade des τ_i (10.5, 10.7), et en remarquant que $T(t) = \sum_{p=2}^{K} \tau_p$ pour K grand ;

2. soit en utilisant la même idée que pour la fonction de corrélation $\langle\overline{T_2}(t)\overline{T_2}(0)\rangle$, *i.e.* en séparant les cas selon le nombre d'ancêtres au temps 0 de la population au temps t.

La première méthode permet ainsi d'écrire la décomposition suivante :

$$\langle T(t)T(0)\rangle - \langle T(t)\rangle\langle T(0)\rangle = \sum_{i,j}\langle\tau_i(t)\tau_j(0)\rangle - \langle\tau_i(t)\rangle\langle\tau_j(0)\rangle = \sum_{i=2}^{\infty}\sum_{j=2}^{\infty}\mathbf{C}_{ij}(t) \quad (10.22)$$

où les fonctions de corrélation $\mathbf{C}_{ij}(t)$ ont été calculées en (10.12). Cette formule donne ainsi la fonction d'auto-corrélation de $T(t)$ sous la forme d'une somme d'exponentielles de la forme $e^{-c_j t}$.

La seconde méthode consiste, quant à elle, à décomposer sur le nombre d'ancêtres qui subsistent au temps 0 (cf. figure 10.4, page 140). Il faut donc introduire les fonctions $z_m(\tau)$ que nous avons définies en (8.32) et qui donnent la probabilité qu'il ne reste que m ancêtres après une durée τ dans le passé. Cela donne la décomposition suivante :

$$\langle T(t)T(0)\rangle = \int_0^t \tau z_1'(\tau)d\tau \times \langle T(0)\rangle + \sum_{m=2}^{\infty} z_m(t)\langle(t+T_m(0))T(0)\rangle.$$

Comme précédemment, il s'agit d'énumérer les arbres et de décomposer les temps $T_m(0)$ et $T(0)$ en sommes de délais τ_i. Le détail du calcul par cette méthode est présenté dans [SD06] et nous nous contenterons ici d'obtenir le résultat par la première méthode. La comparaison permet, de plus, de vérifier que le processus en cascade (10.5, 10.7) est acceptable. Le report de l'équation (10.12) dans (10.22) donne ainsi :

$$\langle T(t)T(0)\rangle - \langle T(t)\rangle\langle T(0)\rangle = \\ \sum_{l=2}^{\infty} 4(-1)^l(2l+1)e^{-c_l t}\left(\sum_{i=2}^{l}(-1)^i\frac{(i+l-2)!}{i!(i-1)!(l-i)!}\right)\left(\sum_{j=l}^{\infty}\frac{1}{j}\frac{(j-2)!(j-2)!}{(j-l)!(j+l+1)!}\right).$$

D'autre part, nous avons la relation suivante[3] :

$$\sum_{i=1}^{l}(-1)^i\frac{(i+l-2)!}{i!(i-1)!(l-i)!} = 0$$

[3] C'est un cas particulier de la formule suivante [GR94] :

$$\sum_{i=1}^{l}\frac{(i+l-2)!}{i!(i-1)!(l-i)!}z^{i-1} = \frac{1}{l}J_{l-1}^{(1,-1)}(2z+1)$$

où $J_{l-1}^{(1,-1)}(x)$ est le polynôme de Jacobi qui s'annule en $x=-1$ et qui est défini par :

$$J_{l-1}^{(1,-1)}(x) = \frac{(-1)^{l-1}}{2^{l-1}(l-1)!}\frac{1+x}{1-x}\frac{d^{l-1}}{dx^{l-1}}\left((1-x)^l(1+x)^{l-2}\right).$$

Nous pouvons de la sorte simplifier l'expression de $\langle T(t)T(0)\rangle - \langle T(t)\rangle\langle T(0)\rangle$ en :

$$\boxed{\begin{array}{l}\langle T(t)T(0)\rangle - \langle T(t)\rangle\langle T(0)\rangle = \\ \displaystyle\sum_{l=2}^{\infty} 4(-1)^l(2l-1)\left(\sum_{j=l}^{\infty}\frac{1}{j}\frac{(j-2)!^2}{(j-l)!(j+l-1)!}\right)e^{-c_l t}\end{array}}. \tag{10.23}$$

Ainsi, nous avons obtenu la fonction d'auto-corrélation de l'âge de l'ancêtre commun d'une population et caractérisé plus précisément l'évolution observée en figure 10.1. Toutes les fonctions de corrélations peuvent être mises sous la forme de sommes d'exponentielles $e^{-c_p t}$ qui proviennent de la distribution des temps τ_i. Les facteurs combinatoires devant ces termes exponentiels, certes longs à écrire, peuvent être obtenus entièrement à partir de la combinatoire des arbres.

Il serait intéressant d'étudier ce que ces corrélations deviennent pour d'autres structures d'arbres et de temps de coalescence, en particulier si nous considérons le coalescent de Bolthausen-Snitzman introduit en section 8.4 : néanmois, des difficultés supplémentaires apparaissent dans ce cas-là puisque l'âge T de l'ancêtre commun de toute la population n'est pas du même ordre de grandeur que les âges T_2, T_3, etc. Il faudrait, dans ce cas, adapter les techniques mathématiques développées dans [FM07].

Chapitre 11

Âge de l'ancêtre commun et diversité génétique sans sélection

Ce chapitre présente les résultats que nous avons publiés dans [SD06] (à la suite de ceux du chapitre précédent) et qui étudient quelques liens entre génétique des populations et modèles de coalescence. La section 11.1 est une introduction aux problèmes de génétiques de populations : nous y présentons quelques résultats simples célèbres et introduisons les modèles que nous avons considérés. La section 11.2 présente notre contribution orginale : l'influence de la connaissance d'un nombre restreint de génomes sur l'âge de l'ancêtre commun de toute une population.

11.1 Généalogies et génomes : introduction

11.1.1 Intérêt

L'une des caractéristiques principales de l'évolution des espèces vivantes est l'apparition de mutations au cours du temps. Dans beaucoup de cas, ces mutations ont un impact direct sur le développement et le fonctionnement des individus qui les portent : elles peuvent être bénéfiques ou délétères. Nous entrons alors dans les modèles de sélection déjà étudiés au chapitre 1. Dans ce chapitre, nous allons nous intéresser au cas de mutations *neutres* du point de vue de la sélection. La redondance du code génétique, ainsi que la présence de parties non codantes sur l'ADN, font que de telles mutations peuvent survenir sans affecter le fonctionnement des individus et n'apportent donc aucun avantage évolutif direct. L'accumulation de ces mutations constitue un témoignage de l'évolution des individus. Plus l'ancêtre commun de deux individus est ancien, plus la probabilité que des mutations se soient produites dans leurs lignées est grande : l'intuition nous dit alors que, plus les génomes de deux individus sont proches, plus il faut s'attendre à ce que leur ancêtre commun le plus récent soit récent et, plus leurs génomes

sont différents, plus leur ancêtre commun le plus récent doit être ancien. Le but de ce chapitre est d'étudier mathématiquement ces corrélations que nous nous attendons à observer entre similarité génétique et généalogies pour des mutations neutres.

Dans un modèle d'évolution neutre comme le modèle de Wright-Fisher, la distribution de l'âge T de l'ancêtre commun le plus récent d'une population, qui est donnée par (8.23) et est représentée en figure 8.3, reste large lorsque la taille N de la population devient grande. Cependant, puisque le séquençage de gènes est à présent possible, il est possible d'avoir accès aux traces génétiques laissées par l'évolution. Il est donc tentant d'utiliser cette information génétique pour affiner l'estimation de l'âge de l'ancêtre commun le plus récent de différents individus (ou différentes espèces) et *inférer* la forme de leur arbre généalogique (ou phylogénétique) [GL05, DNRS], ou bien d'estimer l'époque où une mutation s'est produite dans une lignée [SR00, Taj83].

Il s'agit alors d'utiliser les probabilités bayésiennes : à supposer que nous connaissions les règles d'évolution, les conditions initiales, nous voulons, à partir des données *présentes*, caractériser le processus *passé* qui a produit ces données. Pour reprendre le langage des probabilités conditionnelles, la formulation du modèle en terme de processus de Markov donne la probabilité $P(B, t'|A, t)$ d'un événement B à la date $t' > t$ sachant que l'événement A s'est produit à l'instant t et nous nous intéressons à la probabilité $P(A, t|B', t')$, le lien entre les deux étant la formule des probabilités conditionnelles déjà écrite en (4.1).

Plus précisément, si nous travaillons à longueur de séquence ADN *fixée*[1], il est possible de définir une distance, appelée *distance de Hamming*, entre deux séquences $\sigma = a_0a_1a_2\dots a_n$ et $\sigma' = b_0b_1b_2\dots b_n$ où les lettres a_i et b_i correspondent à l'une des quatre bases A, G, C et T (voir figure 11.1). Elle est donnée par :

$$d_H(\sigma, \sigma') = \sum_{k=1}^{n} (1 - \delta_{a_k, b_k}) \,. \qquad (11.1)$$

Elle est minimale et vaut 0 si les deux séquences sont identiques ; elle est maximale et vaut n si les deux séquences n'ont aucune base en commun.

$$\begin{cases} \text{Séquence } \sigma_1 : & AT\mathbf{G}A\mathbf{T}CGA\mathbf{C}G \\ \text{Séquence } \sigma_2 : & AT\mathbf{C}A\mathbf{A}CGA\mathbf{T}G \end{cases}$$
$$\Longrightarrow d_H(\sigma_1, \sigma_2) = 3$$

FIG. 11.1: Distance de Hamming entre deux séquences de 10 bases.

D'autre part les généalogies fournissent aussi une distance entre individus. Celle-ci est définie comme étant l'âge $T_2(i, j)$ de l'ancêtre commun le plus récent de deux individus i et j :

$$d_G(i, j) = T_2(i, j). \qquad (11.2)$$

La structure en arbre généalogique implique la propriété d'ultra-métricité $d_G(i, j) \leq \max(d_G(i, k), d_G(j, k))$ entre trois individus i, j et k. D'autre part elle est nulle si deux

[1] Nous négligeons donc les mutations qui consistent à ajouter ou retirer des bases d'une séquence ADN, ainsi que les recombinaisons [Hud83].

individus sont identiques (l'ancêtre commun le plus récent d'un unique individu étant, par convention, cet individu lui-même). Elle est d'autant plus grande que l'ancêtre commun le plus récent des deux individus est plus ancien.

La problématique est alors de trouver des corrélations aussi fortes que possible entre les deux distances d_H et d_G. Le problème général est difficile puisqu'il existe divers types de mutations possibles (erreur sur une base, *crossing-over*, etc.) et que la reproduction sexuée fait que chaque individu porte deux versions de chaque gène. C'est pourquoi de nombreux modèles, dont le nôtre, se restreignent à une reproduction asexuée (modèle de Wright-Fisher par exemple). Il est intéressant de noter cependant que, même en reproduction sexuée, un certain nombre de caractères se transmettent de manière asexuée : le chromosome Y chez les hommes, l'ADN mitochondrial chez les femmes, voire le nom de famille si l'on s'intéresse à des quantités autres que génétiques.

11.1.2 La diversité génétique

Étant donné un groupe d'individus de séquences $\sigma_1, \ldots, \sigma_n$, la quantité idéale pour mesurer sa diversité génétique est un indicateur qui aurait les propriétés suivantes :

- sa valeur doit être minimale si tous les individus portent des séquences identiques et maximale si toutes les séquences sont complètement différentes deux à deux ;
- dans la perspective qui est la nôtre, il devrait être corrélé aussi fortement que possible à l'âge de l'ancêtre commun le plus récent du groupe d'individus ;
- il faudrait, pour extraire facilement l'information généalogique de l'information génétique, que cet indicateur donne lieu autant que possible à des calculs analytiques « simples ».

Avant de pouvoir accéder aux temps de coalescence et aux généalogies, il faut tout d'abord connaître le taux θ d'apparition effectif des mutations neutres que nous voulons décrire. Nous allons voir dans cette section un exemple particulier de mesure de θ à partir de plusieurs estimateurs de la diversité génétique.

Plusieurs candidats ont été proposés dans la littérature. Prenons l'exemple particulier du *modèle à nombre infini de sites* [Kim69] : les individus portent une séquence de paires de bases infinie et chaque nouvelle mutation affecte une nouvelle paire de bases. Formulé autrement, cela revient à dire qu'aucune mutation ne se corrige par la suite et que nous pouvons ainsi garder trace de toutes les mutations survenues. Ce modèle, bien que déjà simpliste, ne permet pas d'aller très loin dans les calculs analytiques. Il n'est valable que pour de très longues chaînes ADN avec un taux de mutation par paire de bases très faible. La justification biologique sous-jacente est le faible nombre de sites différents observé entre les individus par rapport au nombre total de sites. La reproduction est, quant à elle, décrite par le modèle de Wright-Fisher dans la limite de grande population (tous les temps seront normalisés par la taille de la population N et nous travaillerons ainsi dans la limite de temps continu, comme au chapitre 10).

Considérons un groupe de n individus de séquences $\sigma_1, \ldots, \sigma_n$. D'après l'hypothèse de nombre infini de sites, nous pouvons définir S_n comme le nombre de sites pour lesquels au moins deux individus ont des bases différentes. Soit $\theta = \theta'/2$ le taux d'apparition d'une

nouvelle mutation[2] en *temps continu*[3]. Le premier événement rencontré en remontant l'arbre généalogie est soit une coalescence (probabilité $c_n/(c_n+n\theta)$, soit une mutation $n\theta/(c_n+n\theta)$. Nous obtenons ainsi la récurrence suivante sur la distribution $\mathrm{Prob}(S_n=k)$ [Kim69, HSW05] :

$$\begin{aligned}\mathrm{Prob}\,(S_n=k) &= \frac{n\theta}{c_n+n\theta}\mathrm{Prob}\,(S_n=k-1)+\frac{c_n}{c_n+n\theta}\mathrm{Prob}\,(S_n=k)\\ &= \frac{\theta'}{n-1+\theta'}\mathrm{Prob}\,(S_n=k-1)+\frac{n-1}{c-1+\theta'}\mathrm{Prob}\,(S_n=k)\,.\end{aligned}$$

La résolution de cette récurrence montre que :

$$\mathrm{Prob}\,(S_n=k)=\frac{n-1}{\theta'}\sum_{p=1}^{n-1}(-1)^{p-1}\frac{(n-2)!}{(i-1)!(n-i-1)!}\left(\frac{\theta'}{p+\theta'}\right)^k/$$

En particulier, nous obtenons :

$$\langle S_n\rangle=\theta' h_n \tag{11.3}$$

où h_n est la série harmonique $h_n=\sum_{p=1}^{n-1}\frac{1}{p}$. Ainsi, le nombre S_n de sites discriminants dans un groupe de n individus fournit une première estimation du taux de mutation. Cet indicateur est appelé *estimateur de Watterson* [Wat75] et est défini par :

$$\boxed{\hat\theta_W=\frac{S_n}{h_n}}\,. \tag{11.4}$$

On s'attend à ce que $\langle\hat\theta_W\rangle=\theta'$. D'autre part, dans un groupe de n individus, nous pouvons compter le nombre de différences $d_H(\sigma,\sigma')$ entre deux individus moyenné sur toutes les paires d'individus et définir ainsi l'estimateur de Tajima [Taj89] :

$$\boxed{\hat\pi=\frac{1}{c_n}\sum_{i\neq j}d_H(\sigma_i,\sigma_j)}\,. \tag{11.5}$$

Dans le modèle de mutations décrit ci-dessus, le nombre moyen de sites différents entre deux individus est égal à θ' (distribution poissonnienne du nombre de mutations sur une branche de longueur t). On s'attend ainsi à obtenir là encore $\langle\hat\pi\rangle=\theta'$.

En moyennant sur différents groupes, les deux indicateurs précédents permettent d'estimer directement le taux de mutation $\theta=\theta'/2$ et, si le modèle est valable, doivent donner la même valeur. Si ce n'est pas le cas (cf. [Taj89, HSW05] pour une quantification plus précise des écarts-type attendus), cela signifie que le modèle est faux (populations structurées avec séparation géographique, influence de la sélection, etc.).

[2]Le facteur 2 dans la définition du taux réduit θ' n'est introduit que pour alléger l'expression des résultats. La probabilité que deux individus aient exactement les mêmes génomes sachant que leur ancêtre commun le plus récent a pour âge t est donnée par $e^{-2\theta t}=e^{-\theta' t}$ puisqu'il faut qu'aucune mutation n'ait affecté chacune des *deux* branches.

[3]Cf. la section suivante pour une discussion de l'échelle de temps choisie.

À supposer que nous ayons une estimation fiable (cf. [KAL99]) du taux de mutation θ, les valeurs des estimateurs ci-dessus pour un groupe particulier (en particulier leurs écarts par rapport aux moyennes attendues) peuvent ensuite être utilisées pour obtenir plus d'information sur la généalogie du groupe.

11.1.3 Le modèle à nombre infini d'allèles et la formule d'Ewens

Le modèle à nombre infini d'allèles permet de simplifier à l'extrême la description des génomes et des mutations. Un *allèle* est une version particulière d'un gène, *i.e.* l'une des variantes parmi les séquences possibles du gène. Le modèle à nombre d'allèles infini, introduit par Kimura et Crow [KC64] en 1964, consiste à ne conserver que l'information minimale lors de la comparaison de deux génomes : soit ils sont identiques, soit ils sont différents. Autrement dit, chaque mutation crée un nouvel allèle et aucune information n'est conservée sur le nombre de mutations qui séparent deux allèles. La distance entre séquences est donc ainsi résumée en deux valeurs : 0 et 1, selon qu'elles sont identiques ou différentes. Un tel modèle induit nécessairement une perte d'information par rapport à la connaissance de la distance de Hamming entre individus. Néanmoins, le traitement mathématique s'en trouve simplifié et la section 11.2 ci-dessous vise à estimer quelle information sur les généalogies on peut extraire à partir de l'information génétique que nous donne ce modèle simplifié.

Comme dans le modèle à nombre infini de sites présenté ci-dessus, nous désignerons par $\theta dt = (\theta'/2)dt$ la probabilité, *en temps continu*, qu'une mutation affecte un individu pendant l'intervalle dt. Le véritable taux de mutation par individu et par génération vaut donc $\alpha = \theta/N$. Le nombre total de mutations dans toute la population en une génération est donc $N\alpha = \theta$. Le choix de l'échelle $\alpha = \theta/N$ provient de deux considérations. D'une part, les taux de mutation observés sont faibles et il y a peu de mutants à chaque génération dans les populations biologiques (non seulement le mécanisme de réplication de l'ADN a un taux d'erreur très faible mais de plus la plupart des mutations sont délétères et le taux effectif des mutations neutres s'en trouve diminué) ; d'autre part, les comportements les plus riches s'obtiennent dans la limite $\alpha = \theta/N$. En effet, les deux facteurs qui entrent en compte dans la diversité génétique d'un groupe sont l'âge de l'ancêtre commun et le délai moyen entre mutations successives : les motifs les plus riches sont obtenus lorsque ces deux temps sont du même ordre[4]. Ainsi, nous travaillerons avec un taux $\alpha = \theta/N$, tel que, dans la limite $N \to \infty$, le nombre de mutants parmi la population en une génération soit θ.

La population, de taille N, peut être divisée en groupes homogènes d'individus qui possèdent le même génome. La question est alors de caractériser la distribution stationnaire des tailles de ces groupes [Kim55, Ewe72]. Considérons une sous-population de taille $n \ll N$. Elle peut être décomposée en k groupes homogènes de tailles $n_1, \ldots, n_k$. Pour la commodité des calculs, introduisons les nombres a_i qui comptent le nombre de groupes homogènes de taille i. Nous avons ainsi $a_1 + 2a_2 + \ldots + na_n = n$, et, d'autre part la somme $a_1 + a_2 + \ldots + a_n$ compte le nombre de génomes différents dans le groupe de

[4] Si la fréquence des mutations est trop faible ($\alpha = o(\theta/N)$), alors la population est globalement homogène ; si elle est trop forte, alors aucun individu ne porte le même génome.

taille n. Soit $\mathbb{P}_n(\mathbf{a})$ la probabilité d'observer une répartition $\mathbf{a} = (a_1, \ldots, a_n)$. Nous pouvons écrire une récurrence reliant $\mathbb{P}_n(\mathbf{a})$ à $\mathbb{P}_{n-1}(\mathbf{a}')$ en comptant le nombre de mutations qui se produisent avant la première coalescence qui fait passer de n à $n-1$ individus. Le résultat est connu depuis [Ewe72] et nous nous contentons ici d'en donner la solution, appelée *formule d'échantillonnage d'Ewens* :

$$\boxed{\mathbb{P}_n(\mathbf{a}) = \frac{n!\Gamma(\theta')}{\Gamma(n+\theta')} \prod_{i=1}^{n} \left(\frac{\theta'}{i}\right)^{a_i} \frac{1}{a_i!}} \,. \tag{11.6}$$

Cette distribution des tailles de groupes homogènes présente plusieurs particularités. Tout d'abord elle est invariante par extraction de sous-population : si nous considérons une sous-population de taille n dont la distribution des tailles de groupes homogènes est donnée par (11.6) puis une sous-population de taille $m < n$ de celle-ci, alors la distribution des tailles des groupes homogènes dans la sous-population de taille m extraite de la première est encore donnée par (11.6), comme on s'y attend intuitivement. De plus, un échantillon typique de taille $n \gg 1$ est composé d'un nombre d'ordre $O(1)$ de groupes de tailles macroscopiques et d'une grande quantité de petits groupes homogènes de taille de l'ordre de l'unité[5].

À notre connaissance, à part quelques résultats partiels [Möh06, BBS07], il n'existe pas de formule générale donnant la répartition des allèles dans un groupe de taille n pour des coalescents plus généraux (cf. section 8.2). La formule d'Ewens, et ses variantes, est déjà utilisée [GL05] en génétique des populations pour estimer l'âge d'allèles : il serait intéressant de l'étendre à d'autres coalescents afin d'élargir les résultats connus aux régimes où la sélection devient pertinente.

11.2 Étude de corrélations entre l'âge de l'ancêtre commun le plus récent et la diversité génétique dans le modèle à nombre infini d'allèles

11.2.1 Quantités étudiées et relations de récurrence

Nous nous plaçons dans le cadre du modèle à nombre infini d'allèles et dans la limite de grandes populations $N \gg 1$. Nous avons calculé en (8.20) la distribution de l'âge de l'ancêtre commun de n individus sans aucune information sur leurs génomes. Supposons à présent que nous ayons quelques informations sur les génomes de ces n individus : *a priori*, cela doit modifier la distribution du temps de coalescence du groupe de taille n.

[5] Il semblerait [KP05] que les noms de famille en Corée, qui se transmettent de père en fils, suivent la loi d'Ewens (11.6) pour un paramètre $\theta \simeq 10.07$ (cela correspond à dix nouveaux noms de famille dans toute la population à chaque génération). La population a subi très peu d'immigration et les noms de famille y existent depuis plus de huit siècles. Les noms les plus fréquents sont Kim, Lee et Park, qui sont portés, à eux trois, par 45% de la population. On notera néanmoins que l'hypothèse de taille constante semble peu réaliste.

Prenons l'exemple, déjà complexe au niveau des résultats, où nous savons que $m \leq n$ individus parmi les n portent les mêmes allèles. L'existence d'un sous-groupe homogène de taille m réduit l'âge de l'ancêtre commun le plus récent de ce sous-groupe puisqu'aucune mutation n'a pu se produire.

Définissons les quantités suivantes :

- $p_{m,n}(T_n)$ désigne la probabilité que l'âge de l'ancêtre commun le plus récent du groupe de taille n soit égal à T_n *et* que les m premiers individus du groupe portent le même allèle,
- $p_n(T_n|m)$ désigne la probabilité que l'âge de l'ancêtre commun le plus récent du groupe de taille n soit égal à T_n *sachant que* les m premiers individus du groupe portent le même allèle,
- Y_m compte la fraction de m-uplets d'individus qui portent le même génome :

$$Y_m = \frac{m!}{N(N-1)\dots(N-m+1)} \sum_{i_1<i_2<\dots<i_m} \delta_{g(i_1),g(i_2),\dots,g(i_m)}$$

où $\delta_{g_1,\dots,g_m}$ vaut 1 si les m génomes $g_1,\dots,\ g_m$ sont tous identiques et 0 sinon.

La quantité Y_2 joue un rôle similaire à l'estimateur de Tajima (11.5), mais il faut prendre garde au fait qu'elle contient moins d'information : notamment, Y_2 ne compte pas le nombre de différences entre deux génomes mais détecte uniquement si *au moins* une mutation les différencie. Les quantités Y_m pour $m \geq 3$ mesurent les corrélations de génomes à plusieurs individus et sont reliées aux moments de Y_2.

Les différentes quantités définies ci-dessus sont reliées entre elles. En effet, le passage de $p_{m,n}(T_n)$ à $p_n(T_n|m)$ se fait par la formule des probabilités conditionnelles. La probabilité que m individus aient le même génome est donnée par la formule d'Ewens (11.6) pour la partition $\vec{a} = (0,\dots,0,1)$ et nous avons ainsi :

$$p_{m,n}(T_n) = p_n(T_n|m)\frac{(m-1)!\Gamma(\theta'+1)}{\Gamma(\theta'+m)}. \tag{11.7}$$

D'autre part, la fonction génératrice de $p_{m,n}(T_n)$ n'est autre que la fonction de corrélation entre Y_m et T_n :

$$\widehat{p}_{m,n}(s) \hat{=}= \int_0^\infty e^{-sT} p_{m,n}(T)dT = \langle Y_m e^{-sT_n}\rangle. \tag{11.8}$$

Des dérivations successives en $s = 0$ permettent ainsi, en principe, de calculer toutes les corrélations $\langle Y_m T_n^k\rangle$.

Nous allons à présent écrire une équation de récurrence sur les $p_{m,n}(T)$. Pour cela, regardons le premier intervalle dt dans leur généalogie. Si une coalescence se produit à l'intérieur du sous-groupe homogène de taille m, alors nous sommes ramenés au calcul de $p_{m-1,n-1}(T)$; si, au contraire, une coalescence affecte n'importe quelle autre paire d'individus, cela nous oblige à calculer $p_{m,n-1}(T)$. D'autre part, l'hypothèse d'homogénéité du sous-groupe de taille m implique qu'aucune mutation ne doit se produire sur les lignées des m individus. Nous obtenons ainsi pour $m \geq 2$ et $n \geq 3$:

$$\partial_T p_{m,n}(T) = c_m p_{m-1,n-1}(T) + (c_n - c_m)p_{m,n-1}(T) - (c_n + m\theta)p_{m,n}(T). \tag{11.9}$$

Le passage aux fonctions génératrices donne alors :

$$\boxed{\widehat{p}_{m,n}(s) = \frac{1}{s + c_n + m\theta} \left(c_m \widehat{p}_{m-1,n-1}(s) + (c_n - c_m) \widehat{p}_{m,n-1}(s) \right)} . \tag{11.10}$$

Pour $m = 1$, les mutations sur le premier individu n'importent plus et $\widehat{p}_{1,n}(s)$ est directement donnée par (8.20). Pour $n = 2$ et $m = 2$, nous avons tout simplement $\widehat{p}_{2,2}(s) = 1/(s + 1 + 2\theta)$.

La résolution générale de (11.10) n'est pas immédiate et mène à des expressions peu utilisables en général : nous ne les reproduirons pas ici et le lecteur intéressé pourra consulter [SD06] pour l'expression exacte et son interprétation en termes de combinatoire sur les arbres généalogiques.

11.2.2 Influence de la connaissance des génomes de deux individus

Nous nous limiterons ici au cas particulier où $m = 2$ et $n \to \infty$, *i.e.* à l'âge de l'ancêtre commun le plus récent de toute la population sachant que deux individus partagent le même génome. Pour $m = 2$ et $n \geq 3$, l'équation (11.10) se réduit, en utilisant (8.20), à :

$$\widehat{p}_{2,n}(s) = \frac{1}{s + c_n + 2\theta} \left((c_n - 1) \widehat{p}_{2,n-1}(s) + \prod_{k=2}^{n-1} \frac{c_k}{s + c_k} \right) . \tag{11.11}$$

Afin d'établir une formule sommatoire, à chaque itération de la récurrence (11.11), il faut choisir l'un des deux termes de droite ; si l'on choisit le deuxième alors la récurrence s'arrête. Appelons q la valeur de n lorsque le deuxième terme est choisi. La solution générale est alors donnée pour $n \geq 3$ par :

$$\begin{aligned} \widehat{p}_{2,n}(s) = & \frac{c_2}{s + c_2 + 2\theta} \prod_{l=3}^{n} \frac{c_l - 1}{s + c_l + 2\theta} \\ & + \sum_{q=3}^{n} \left(\prod_{l=q}^{n} \frac{1}{s + c_l + 2\theta} \right) \left(\prod_{k=2}^{q-1} \frac{c_k}{s + c_k} \right) \left(\prod_{l=q+1}^{n} (c_l - 1) \right) \end{aligned}$$

où, par convention, le dernier produit est égal à 1 si $l = n + 1$. Considérons directement la limite $n \to \infty$. Nous obtenons ainsi :

$$\boxed{\widehat{p}_{2,\infty}(s) = \langle Y_2 e^{-sT} \rangle = \sum_{q=2}^{\infty} \frac{2}{(q+1)q} \left(\prod_{l=q}^{n} \frac{c_l}{s + c_l + \theta'} \right) \left(\prod_{k=2}^{q-1} \frac{c_k}{s + c_k} \right)} \tag{11.12}$$

où, par convention, le dernier produit sur k vaut 1 pour $q = 2$.

De l'expression précédente, il est possible d'extraire, par transformation de Laplace inverse, l'expression analytique de la distribution de T sachant que deux individus au

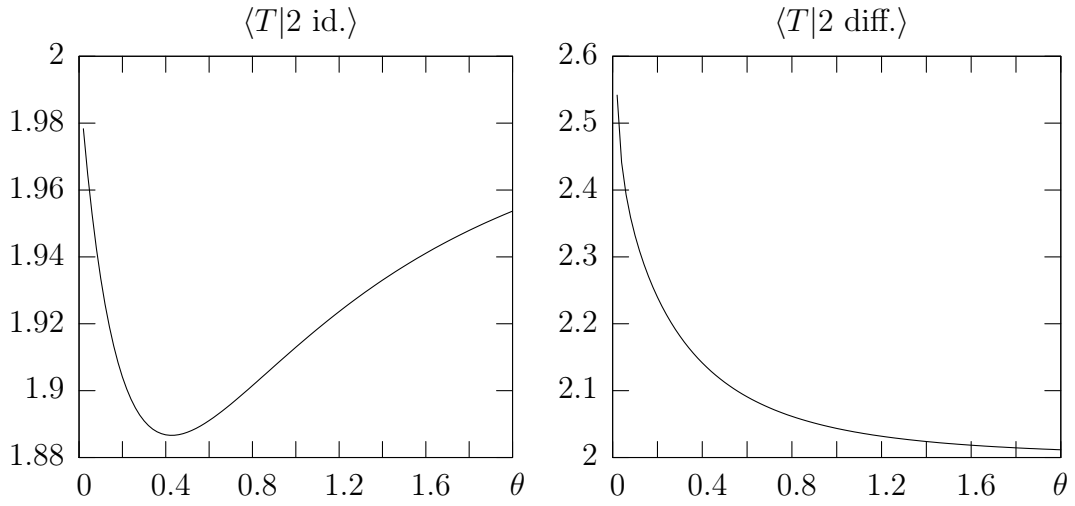


FIG. 11.2: Âge moyen de l'ancêtre commun le plus récent d'une population de taille $N \gg 1$ sachant que deux individus au moins ont le même génome (gauche) ou des génomes différents (droite). Sans information sur les génomes des individus, l'âge moyen attendu est $\langle T \rangle = 2$.

moins ont le même génome. Néanmoins, les expressions obtenues sont beaucoup moins simples que l'équation (8.23), qui était obtenue sans conditionnement. Nous nous contenterons ici de donner l'expression de l'âge moyen $\langle T|m = 2\rangle$ de l'ancêtre commun le plus récent de la population sachant qu'au moins deux individus ont le même génome. Cette expression s'obtient soit en dérivant l'équation de récurrence (11.11) et en résolvant la nouvelle équation de récurrence sur $\langle T_n|m = 2\rangle$ (cf. [SD06] pour le détail des calculs), soit en dérivant l'expression (11.12). Le résultat est le suivant :

$$\langle T|m = 2\rangle = 1 + \frac{1}{1+\theta'} - \theta' \sum_{p=3}^{\infty} \frac{(2p-1)(p+1)(p-2)}{2} \frac{(-1)^p}{c_p(c_p+\theta')} \tag{11.13}$$

où $\theta' = 2\theta$. Nous retrouvons l'expression $\langle T \rangle = 2$ en absence de mutations ($\theta = 0$).

La représentation graphique de ce résultat est donnée en figure 11.2. Elle met en évidence, pour $\theta \simeq 0{,}4$ une déviation maximale de l'ordre de 5% par rapport au cas sans conditionnement. Cette valeur, bien que faible, reste significative puisqu'elle signifie que la correction à l'âge de l'ancêtre commun le plus récent de *toute* la population lorsque nous avons des informations sur seulement un nombre fini d'individus (ici $m = 2$) ne tend pas vers 0 lorsque la taille totale N tend vers l'infini. Cette propriété n'est pas surprenante pour la raison suivante : en prenant une paire d'individus au hasard dans toute la population, la probabilité que leur ancêtre commun le plus récent coïncide avec celui de toute la population est d'ordre 1 (elle vaut 1/3, cf. [STW84]) et l'information sur deux individus se répercute de manière significative sur la hauteur de l'arbre généalogique de toute la population. Pour $m = 5$, les conséquences sur le temps de coalescence T sont représentées en figure (11.3) et confortent la tendance.

La résolution complète de (11.10) permet d'obtenir [SD06] une formule sommatoire

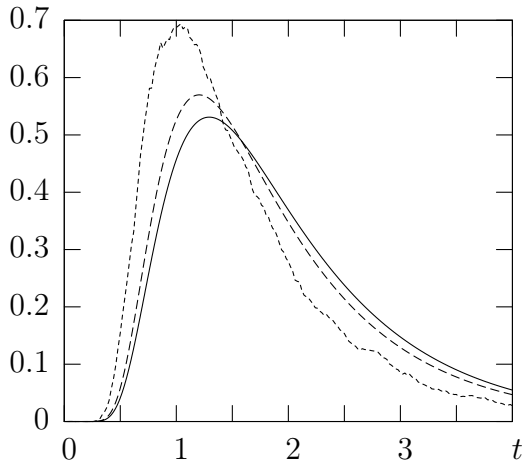


FIG. 11.3: Distribution de l'âge de l'ancêtre commun d'une population en présence d'information sur les génomes des individus. La valeur du taux de mutation est $\theta = 0.5$ pour toutes les courbes. La ligne représente la distribution stationnaire (8.23) de T sans information sur les génomes pour une population de taille $N \to \infty$. Les pointillés longs représentent la distribution du temps de coalescence T sachant que deux individus pris au hasard ont le même génome (obtenue à partir de (11.12)) ; les pointillés courts représentent la distribution du temps de coalescence T de la population sachant que 5 individus pris au hasard ont le même génome (résultat numérique pour une population de 50 individus). Les courbes mettent en évidence qu'il suffit de posséder des informations sur un nombre fini de génomes pour changer de manière significative l'estimation du temps de coalescence de la population complète.

pour $p_{m,n}(t)$ mais cela ne permet pas d'avoir une formule fermée simple, même pour la valeur moyenne de $\langle T_n \rangle$. Une généralisation à des distributions $p_{(m_1,m_2,\ldots,m_k),n}(T)$ sachant que k groupes de tailles m_l donnés sont homogènes permet seulement d'écrire des équations de récurrence de type (11.10) où l'on liste les coalescences autorisées, sans aboutir à une solution générale utilisable en pratique. Seuls quelques cas particuliers (deux groupes homogènes, plusieurs groupes homogènes de taille 2, etc.) peuvent être traités mais les corrections qu'ils apportent à l'âge T de l'ancêtre commun le plus récent restent du même ordre que celles observées en figures 11.2 et 11.3.

Conclusion et perspectives

Dans cette thèse, nous avons considéré différents modèles de dynamique de populations en présence ou non de sélection. Les modèles d'évolution constituent un domaine actif à l'interface entre biologie, mathématiques et physique statistique. La problématique générale de ces modèles est d'étudier la compétition entre, d'une part, la reproduction qui tend à augmenter la taille de la population et, d'autre part, les limitations extérieures du milieu qui tendent à réduire cette taille. Reproduction, déplacement et mutation des individus sont décrits par les outils probabilistes usuels que sont les marches aléatoires et les processus de branchement. Introduire de la sélection parmi les individus consiste à les classer selon leur adaptabilité à l'environnement et à biaiser la reproduction et la survie en faveur des individus les plus adaptés.

Les résultats présentés correspondent aux quatre travaux publiés durant ces années de thèse qui sont, par ordre chronologique : l'étude de la dynamique des généalogies dans le modèle de Wright-Fisher sans espace ni sélection, ainsi que la corrélation entre l'âge de l'ancêtre commun le plus récent d'une population et sa diversité génétique [SD06], l'étude de la probabilité de survie d'une marche aléatoire avec branchements en présence de conditions aux bords absorbantes et de la transition de phase vers l'extinction [DS07], l'étude de l'évolution de cette même marche aléatoire avec branchements conditionnée par une taille finale et la construction d'un processus biaisé qui nous a permis de simuler directement cette évolution conditionnée et de déterminer les propriétés du régime quasi-stationnaire près de la vitesse critique [SD08], et finalement l'étude des généalogies dans des modèles de sélection avec structure spatiale et le lien entre ces modèles et les polymères dirigés [BDS08].

Les marches aléatoires avec branchement entretiennent des liens étroits avec les équations de propagation de front : nous avons exploité les résultats connus sur les fronts qui se propagent sur une ligne infinie afin de les adapter au cas de marches aléatoires avec branchement en présence d'un mur absorbant (chapitre 3). Ce problème, qui est lié aux problèmes de survie de populations en présence de sélection, nous a obligés à considérer des solutions de fronts avec des conditions aux bords (annulation au bord). L'existence, ou non, et la forme de telles solutions, nous a permis d'étudier la transition de phase qui apparaît lorsque la pression de sélection est trop forte : au-delà d'une certaine vitesse critique v_c du mur absorbant, les marches aléatoires avec branchements sont rattrapées par le mur et la population s'éteint. Un traitement analytique nous a permis d'aboutir

à des formules universelles pour la probabilité de survie près de la vitesse critique.

Ce cadre est relativement différent des processus de réaction-diffusion avec états absorbants habituels [CT96, CT98, THVL05, CDC91] puisque la transition vers l'état absorbant n'est pas induite par des annihilations entre individus mais par les conditions aux bords du domaine. En particulier, il est plus difficile d'utiliser le formalisme de la théorie des champs. Les outils que nous avons utilisés sont spécifiques à l'étude des fronts en dimension $d = 1$ dans un milieu homogène. Des résultats mathématiques sur les fronts en dimension plus grande ou dans des milieux périodiques [BHN05] invitent à considérer l'effet de la structure spatiale sur des marches aléatoires avec branchements dans des domaines plus généraux que ceux considérés ici.

À la base de l'étude de l'évolution d'une marche aléatoire avec branchements conditionnée par sa taille finale (chapitres 4, 5, 6), réside le fait que la probabilité de survie de la marche et les fonctions génératrices satisfont des équations de fronts déterministes (de type F-KPP) avec des conditions aux bords idoines. Dans la limite des temps longs, le conditionnement a pour conséquence, lorsque $v < v_c$, de générer un régime quasi-stationnaire (chapitre 4). Ce dernier est relié au vecteur propre associé au premier temps de relaxation d'un front. La première étape fut ainsi d'étudier en détail l'effet des conditions aux bords sur les temps de relaxation de fronts (fin du chapitre 2).

Néanmoins, pour $v > v_c$, des calculs numériques tendent à montrer qu'il n'existe pas de régime quasi-stationnaire dans la limite des temps longs. De plus, la probabilité de survie, qui satisfait l'équation F-KPP, n'a plus de temps de relaxation discret et le formalisme du chapitre 4 ne s'applique plus. Il serait ainsi intéressant d'adapter ce formalisme aux cas où la relaxation de la probabilité de survie n'est pas exponentielle afin d'étudier le régime conditionné présenté en figure 4.3.

Pour $v < v_c$, nous avons déterminé dans le régime quasi-stationnaire le profil moyen et la taille de la population près du point critique. Les expressions obtenues (6.3, 6.4, 6.7) sont universelles et ne dépendent pas du mécanisme précis du branchement, ni d'une éventuelle discrétisation du *fitness*. Un aspect plus surprenant est la similarité des résultats avec des marches aléatoires avec branchements de taille N *fixée* (et de vitesse fluctuante) : la relation entre taille et vitesse semble être la même dans les deux, au moins aux deux premiers ordres. Il serait souhaitable de comprendre plus en profondeur cette universalité et d'en cerner aussi les limites.

Enfin, un dernier point intéressant de cette étude fut la construction d'un processus biaisé (chapitre 5) qui permet de prendre en compte le conditionnement d'une marche aléatoire avec branchement sur sa taille finale. Outre l'intérêt numérique d'un tel processus, il nous a permis de comprendre les propriétés du régime quasi-stationnaire. De plus, il a l'avantage d'être encore valable pour $v > v_c$ et peut constituer un point de départ pour l'étude du régime conditionné à $v > v_c$. Sa limitation actuelle réside dans le fait qu'il n'est valide que pour des marches aléatoires indépendantes et il serait intéressant de savoir s'il peut être généralisé au cas d'individus en interactions. Le processus dual, au sens de [DMS03], des marches aléatoires indépendantes est l'équation de front F-KPP déterministe : pour des marches aléatoires dont on autorise les recombinaisons $A + A \to A$, le processus dual est une équation F-KPP stochastique dont le bruit dépend du taux de recombinaison. Il serait alors intéressant de savoir si les termes de bruit sup-

plémentaires peuvent être interprétés en termes de particules aux propriétés modifiées (cf. chapitre 5).

Le deuxième thème de cette thèse fut l'étude des généalogies des individus dans les modèles d'évolution de populations. En absence de sélection, nous avons étudié les propriétés dynamiques de l'âge de l'ancêtre commun le plus récent d'une population : cela nous a amenés à construire un processus stochastique, non pas sur cet âge, mais sur toutes les longueurs de branches dans l'arbre généalogique (chapitre 10). Cette construction, qui donne au final un processus relativement simple, a permis de calculer diverses fonctions de corrélation liées à l'évolution de l'âge de l'ancêtre commun le plus récent. À notre connaissance, aucun résultat n'existe en présence de sélection sur l'aspect dynamique des généalogies : il serait intéressant de savoir ce qu'il advient lorsque la sélection entre en jeu. Un modèle-jouet simple pourrait être le régime quasi-stationnaire étudié dans la première partie, puisque, dans le processus biaisé, l'individu A_1 et les branches qui en partent correspondent au haut de l'arbre généalogique, ou bien le modèle exponentiel du chapitre 7 à taille constante pour lequel de nombreux résultats exacts peuvent être établis.

Sans sélection toujours, nous avons exploré les corrélations qui apparaissent entre ancêtre commun le plus récent et diversité génétique d'une population dans le modèle à nombre infini d'allèles. Dans le cas particulier où nous supposons avoir de l'information sur un nombre fini d'individus dans la population (n génomes identiques par exemple), nous avons déterminé la correction que cette information apporte dans l'estimation de l'âge de l'ancêtre commun le plus récent (chapitre 11). Les formules analytiques obtenues ne semblent pas être manipulables facilement dans le cas général, néanmoins il serait souhaitable de savoir s'il est possible de les utiliser au moins numériquement.

Si l'on introduit de la sélection tout en restant dans des modèles de type champ moyen, la structure des généalogies change de classe d'universalité (coalescent de Bolthausen-Sznitman au lieu du coalescent de Kingman). Ce changement est lié au fait que certains individus (de haut *fitness*) génèrent une plus grande partie de la population que d'autres (de bas *fitness*) qui ont plutôt tendance à s'éteindre. En dimension finie, le fait d'avoir une compétition locale entre les individus modifie les échelles de temps et les généalogies. Le chapitre 9 est une exploration numérique de ce qui se passe pour des modèles de polymères dirigés : si, en champ moyen, les résultats sont semblables aux marches aléatoires avec branchement (sur l'axe de *fitness*), en dimension finie au contraire, les effets géométriques sont prépondérants et les généalogies appartiennent à d'autres classes d'universalité.

La majorité de ces résultats sont numériques. Il serait souhaitable à présent d'en avoir une compréhension analytique. En champ moyen, il existe déjà très peu de modèles exactement solubles (modèle exponentiel du chapitre 7 et [BD04]). Dans un premier temps, il serait intéressant de comprendre rigoureusement pourquoi le coalescent de Bolthausen-Sznitman est universel en champ moyen. Dans un deuxième temps l'idéal serait d'étendre le résultat de [LS06] aux modèles avec sélection, c'est-à-dire de déterminer la dimension critique au-delà de laquelle les généalogies observées sont celles du champ moyen. Il faut néanmoins s'attendre à certaines difficultés, puisque, comme nous l'avons vu au chapitre 9, ce problème est relié à celui de la dimension critique des polymères dirigés.

Durant toute cette deuxième partie, nous avons comparé, au moins numériquement, les cas *sans* sélection et avec une sélection *forte* mais nous n'avons étudié aucun cas intermédiaire. Une sélection adoucie pourrait être obtenue par exemple non pas en sélectionnant les N meilleurs individus mais en sélectionnant les individus avec une statistique de type Fermi-Dirac. Les limites $\beta = 0$ et $\beta = \infty$ correspondent à des sélections nulle et maximale ; pour une « température » intermédiaire, on peut s'attendre à une transition de phase entre les régimes de Kingman et Bolthausen-Sznitman. Il serait intéressant de savoir si cette transition se produit à $\beta = 0$ ou à $\beta > 0$: l'étude de cette transition permettrait en effet de mieux comprendre l'origine de l'universalité des généalogies ainsi que les paramètres pertinents.

Le processus biaisé de la première partie a permis d'étudier la dynamique de l'individu A_1 qui, par construction, est l'ancêtre commun de toute la population à un temps ultérieur. En particulier, nous avons vu que par rapport aux autres individus de la population, il n'était pas à l'avant du front et son *fitness* moyen ne fait pas partie nécessairement des meilleurs *fitness* de la population. Cela met en évidence le rôle prépondérant des fluctuations, qui font que les meilleurs ne survivent pas nécessairement : il serait intéressant de transposer ces résultats par exemple au problème des polymères dirigés et d'étudier l'énergie des polymères qui ont des descendants, comparativement à l'énergie moyenne. Cela permettrait en outre de connaître la distribution du nombre de descendants d'un individu (respectivement d'un polymère) connaissant la place de son *fitness* (respectivement son énergie) par rapport à ceux des autres individus, qui est, comme nous l'avons vu au chapitre 8, la quantité-clef qui permet de déterminer analytiquement les généalogies.

Bibliographie

[ABTR07] T. Antal, K. B. Blagoev, S. A. Trugman et S. Redner, *Aging and immortality in a cell proliferation model*, Journal of Theoretical Biology, **248** (2007), 411–417.

[AW75] D. G. Aronson et H. F. Weinberger, *Nonlinear diffusion in population genetics, combustion, and nerve propagation*, Lecture notes in mathematics, **446** (1975), 5–49.

[AW78] ———, *Multidimensional nonlinear diffusion arising in population genetics*, Advances in Mathematics, **30** (1978), 33–78.

[BB03] M. Baake et E. Baake, *An exactly solved model for mutation, recombination and selection*, Canadian Journal of Mathematics, **55** (2003), 3.

[BBC+05] M. Birkner, J. Blath, M. Capaldo, A. M. Etheridge, M. Möhle, J. Schweinsberg et A. Wakolbinger, *Alpha-stable branching and Beta-coalescents*, Elect. Journ. Prob., **10** (2005), 303–325.

[BBL07] J. Berestycki, N. Berestycki et V. Limic, *The asymptotic number of blocks in a lambda coalescent*, To appear, (2007).

[BBS07] J. Berestycki, N. Berestycki et J. Schweinsberg, *Beta-coalescents and continuous stable random trees*, Ann. Probab., **35** (2007), 1835–1887.

[BD96a] R. D. Benguria et M. C. Depassier, *Speed of fronts of the reaction-diffusion equation*, Phys. Rev. Lett., (1996), 1171–1173.

[BD96b] ———, *Variational characterization of the speed of propagation of fronts for the nonlinear diffusion equation*, Comm. Math. Phys., **175** (1996), 221–227.

[BD97] É. Brunet et B. Derrida, *Shift in the velocity of a front due to a cutoff*, Phys. Rev. E, **57** (1997), 2597–2604.

[BD04] ———, *Exactly soluble noisy traveling-wave equation appearing in the problem of directed polymers in a random medium*, Phys. Rev. E, **70** (2004), 016106.

[BDMM06a] É. Brunet, B. Derrida, A. Mueller et S. Munier, *Noisy traveling waves : Effect of selection on genealogies*, Europhys. Lett., **76** (2006).

[BDMM06b] ———, *Phenomenological theory giving the full statistics of the position of fluctuating pulled fronts*, Phys. Rev. E, **73** (2006), 056126.

[BDMM07] ———, *Effect of selection on ancestry : an exactly soluble case and its phenomenological generalization*, Phys. Rev. E, **76** (2007), 041104.

[BDS08] É. Brunet, B. Derrida et D. Simon, *Coalescence times in presence and in absence of selection*, in preparation, (2008).

[BHN05] H. Berestycki, F. Hamel et N. Nadirashvili, *The speed of propagation for KPP type problems. I – Periodic framework*, J. European Math. Soc., **7** (2005), 173–213.

[BHR04] H. Berestycki, F. Hamel et L. Roques, *Reaction-diffusion equations and biological invasion models in periodic media*, Note C. R. Acad. Sc. Paris, **339** (2004), 549–554.

[BHR05a] ———, *Analysis of the periodically fragmented environment model : I – species persistence*, J. Math. Biol., **51** (2005), 75–113.

[BHR05b] ———, *Analysis of the periodically fragmented environment model : II – biological invasions and pulsating travelling front*, J. Math. Pures Appl., **84** (2005), 1101–1146.

[BJBD+85] E. Ben-Jacob, H. Brand, G. Dee, L. Kramer et J. S. Langer, *Pattern propagation in nonlinear dissipative systems*, Physica, **14** (1985), 348–364.

[BL00] J. Bertoin et J.-F. Le Gall, *The Bolthausen-Sznitman coalescent and the genealogy of continuous-state branching processes*, Probab. Th. Rel. Fields, **117** (2000), 249–266.

[BLQ+08] L. B. Barreiro, G. Laval, H. Quach, E. Patin et L. Quintana-Murci, *Natural selection has driven population differentiation in modern humans*, Nature genetics, (2008).

[Bra83] M. D. Bramson, *Convergence of solutions of the Kolmogorov equation to traveling waves*, Mem. Am. Math. Soc., **44** (1983), 1–190.

[BS98] E. Bolthausen et A.-S. Sznitman, *On Ruelle's probability cascades and an abstract cavity method*, Comm. Math. Phys., **197** (1998), 247–276.

[CBBV07] V. Colizza, M. Barthélémy, A. Barrat et A. Vespignani, *Epidemic modeling in complex realities*, C. R. Biologies, **330** (2007).

[CCL+07] P. Cattiaux, P. Collet, A. Lambert, S. Martinez, S. Méléard et J. San Martin, *Quasi-stationary distributions and diffusion models in population dynamics*, http ://arxiv.org/abs/math/0703781, (2007).

[CDC91] S. Cornell, M. Droz et B. Chopard, *Role of fluctuations for inhomogeneous reaction-diffusion phenomena*, Phys. Rev. A, **44** (1991), 4826–4832.

[Cha99] J. T. Chang, *Recent common ancestors of all present-day individuals*, Advances in Appl. Prob., **31** (1999), 1002–1026.

[CMS95] P. Collet, S. Martinez et J. San Martin, *Asymptotic laws for one-dimensional diffusions conditionned to non-absorption*, Ann. Prob., **23** (1995), 1300–1314.

[Cox89] J. T. Cox, *Coalescing random walks and voter model consensus times on the torus in* $\mathbb{Z}^d$, Ann. Probab., **17** (1989), 1333–1366.

[CR88] B. Chauvin et A. Rouault, *KPP equation and supercritical branching Brownian motion in the subcritical speed area. application to spatial trees.*, Probab. Th. Rel. Fields, **80** (1988), 299–314.

[CT96] J. L. Cardy et U. C. Täuber, *Theory of branching and annihilating random walks*, Phys. Rev. Lett., **77** (1996), 4780–4783.

[CT98] J. L. Cardy et U. C. Täuber, *Field theory of branching and annihilating random walks*, Journal of Statistical Physics, **90** (1998), 1–56.

[DB88] B. Derrida et D. Bessis, *Statistical properties of valleys in the annealed random map model*, J. Phys. A. : Math. Gen., **21** (1988), L509–L515.

[DF07] M. Desai et D. Fisher, *Beneficial mutation-selection balance and the effect of linkage on positive selection*, Genetics, **176** (2007), 1759–1798.

[Dic02] R. Dickman, *Nonequilibrium phase transitions in epidemics and sandpiles*, Physica A, **306** (2002), 90–97.

[DMS03] C. R. Doering, C. Mueller et P. Smereka, *Interacting particles, the stochastic Fisher–Kolmogorov–Petrovsky–Piscounov equation, and duality*, Physica A, **325** (2003), 243–259.

[DNRS] A. J. Drummond, G. K. Nichols, A. J. Rodrigo et W. Solomon, *Estimating mutation parameters, population history, and genealogies simultaneously using temporally spaced sequence data*, Genetics, **161**, 1307–1322.

[Don91] P. Donnelly, *Weak convergence to a markov chain with an entrance boundary : ancestral processes in population genetics*, The Annals of Probability, **19** (1991), 1102–1117.

[DP82] B. Derrida et L. Peliti, *Evolution in a flat fitness landscape*, Bulletin of mathematical biology, **53** (355–382), 1991.

[DS88] B. Derrida et H. Spohn, *Polymers on disordered trees, spin glasses and traveling waves*, J. Stat. Phys., **51** (1988), 817.

[DS07] B. Derrida et D. Simon, *The survival probability of a branching random walk in presence of an absorbing wall*, EPL, **78** (2007), 60006.

[DV03] R. Dickman et R. Vidigal, *Path integrals and perturbation theory for stochastic processes*, Brazilian Journal of physics, **33** (2003), 73–93.

[Ewe72] W. Ewens, *The sampling theory of selectively neutral alleles*, Theoretical Population Biology, **3** (1972), 87–112.

[Ewe04] ———, *Mathematical population genetics*, Springer-Verlag, seconde édition (2004).

[FH04] P. François et V. Hakim, *Design of genetic networks with specified functions by evolution* in silico, Proc. Natl. Acad. Sci. USA, **101** (2004), 580–585.

[Fis30] R. Fisher, *The Genetical Theory of Natural Selection*, Clarendon Press, Oxford (1930).

[Fis37] ———, Annals of Eugenics, **7** (1937), 355.

[FL97] Y.-X. Fu et W.-H. Li, *Estimating the age of the common ancestor of a sample of DNA*, Mol. Biol. Evol., **14** (1997), 195–199.

[FM07] F. Freund et M. Möhle, *On the time back to the most recent common ancestor and the external branch length of the Bolthausen-Sznitman coalescent*, submitted to Markov Process. Related Fields, (2007).

[FMM97] P. A. Ferrari, S. Martinez et J. S. Martin, *Phase transition for absorbed brownian motion with drift*, J. Stat. Phys., **86** (1997), 213–231.

[FMP91] P. A. Ferrari, S. Martinez et P. Picco, *Some properties on quasi-stationary distributions in the birth and death chains*, Instabilities and non-equilibrium structures III, Mathematics and its applications, Kluwer (1991) p. 177–187.

[FMP92] ———, *Existence of nontrivial quasi stationary distributions in the birth and death chain*, Adv. Appl. Probability, **24** (1992), 795–813.

[Gir01] C. Giraud, *Genealogies of shocks in Burgers turbulence with white noise initial velocity*, Comm. Math. Phys., **1** (2001), 67–86.

[GKP06] C. Giardinà, J. Kurchan et L. Peliti, *Direct evaluation of large-deviation functions*, Phys. Rev. Lett., **96** (2006), 120603.

[GL98] P. J. Gerrish et R. E. Lenski, *The fate of competing beneficial mutations in an asexual population*, Genetica, **102/103** (1998), 127–144.

[GL05] R. C. Griffiths et S. Lessard, *Ewens' sampling formula and related formulae : combinatorial proofs, extension to variable population size and applications to age of alleles*, Theor. Popul. Biol., **68** (2005), 167–177.

[Gou02] S. Gould, *The structure of evolutionary theory*, Harvard University Press, Cambridge (Mass.) ; London (2002).

[GR94] I. Gradshteyn et I. Ryzhik, *Table of integrals, series, and products*, Academic Press, cinquième édition (1994).

[GW74] F. Galton et H. Watson, *On the probability of extinction of families*, Journal of the Anthropological Institute, **3** (1874), 308–311.

[Har62] T. E. Harris, *The theory of branching processes*, Springer-Verlag (1962).

[HH04] R. Hardy et S. C. Harris, *A new formulation of the spine approach to branching diffusions*, Mathematics Preprint, **0404** (2004).

[HH07] J. W. Harris et S. C. Harris, *Survival probabilities for branching brownian motion with absorption*, Elect. Comm. in Probab., **12** (2007), 81–92.

[HHK06] J. W. Harris, S. C. Harris et A. E. Kyprianou, *Further probabilistic analysis of the Fisher-Kolmogorov-Petrovskii-Piscounov equation : one sided travelling-waves.*, Ann. Inst. H. Poincaré Probab. Statist., **42** (2006), 125–145.

[HHZ95] T. Halpin-Healy et Y. Zhang, *Kinetic roughening phenomena, stochastic growth, directed polymers and all that. Aspects of multidisciplinary statistical mechanics*, Phys. Rep., (1995), 215–414.

[Hin00] H. Hinrichsen, *Non-equilibrium critical phenomena and phase transitions into absorbing states*, Advances in physics, **49** (2000), 815–958.

[HR75] K. Hadeler et F. Rothe, *Travelling fronts in nonlinear diffusion equations*, J. Math. Biol., **2** (1975), 251–263.

[HSW05] J. Hein, M. Schierup et C. Wiuf, *Gene genealogies, variation and evolution : a primer in coalescent theory*, Oxford University Press (2005).

[Hud83] R. R. Hudson, *Properties of a neutral allele model with intragenic recombination*, Theoretical population biology, **23** (1983), 183–201.

[IMM05] E. Iancu, A. Mueller et S. Munier, *Universal behavior of qcd amplitudes at high energy from general tools of statistical physics*, Phys. Lett. B, **606** (2005), 342–350.

[Jai07] K. Jain, *Evolutionary dynamics of the most populated genotype on rugged fitness landscapes*, Phys. Rev. E, **76** (2007), 031922.

[JK06] K. Jain et J. Krug, *Deterministic and stochastic regimes of asexual evolution on rugged fitness landscapes*, Genetics, **175** (2006), 1275–1288.

[Joh03] K. Johansson, *Discrete polynuclear growth and determinantal processes*, Comm. Math. Phys., **242** (2003), 277–329.

[KAL99] E. K. Klein, F. Austerlitz et C. Larédo, *Some statistical improvements for estimating population size and mutation rate from segregating sites in DNA sequences*, Theor. Popul. Biol., **55** (1999), 235–247.

[KC64] M. Kimura et J. Crow, *The number of alleles that can be maintained in a finite population*, Genetics, **49** (1964), 725–738.

[KDH88] N. L. Kaplan, T. Darden et R. R. Hudson, *The coalescent process in models with selection*, Genetics, **120** (1988), 819–829.

[Kes78] H. Kesten, *Branching brownian motion with absorption*, Stoch. Process. Appl., **7** (1978), 9–47.

[Kha73] E. Khalili-Françon, *Processus de galton-watson*, Lecture notes in Math., **321** (1973), 122–135.

[Kim55] M. Kimura, *Stochastic processes and the distribution of gene frequencies under natural selection*, Cold Spring Harbor Symp. Quant. Biol., **20** (1955), 33–53.

[Kim69] ———, *The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations*, Genetics, **61** (1969), 893–903.

[Kin82a] J. Kingman, *The coalescent*, Stoch. Process., **13** (1982), 235–248.

[Kin82b] ———, *On the genealogy of large populations*, J. Appl. Probab., **19A** (1982), 27–43.

[KP05] B. J. Kim et S. M. Park, *Distribution of korean family names*, Physica A, **347** (2005), 683–694.

[KPP37] A. Kolmogorov, I. Petrovsky et N. Piscounov, Bull. Univ. Etat Moscou A, **1** (1937), 1.

[KPZ86] M. Kardar, G. Parisi et Y.-C. Zhang, *Dynamic scaling of growing interfaces*, Phys. Rev. Lett., **56** (1986), 889–892.

[KR85] K. Kang et S. Redner, *Fluctuation-dominated kinetics in diffusion-controlled reactions*, Phys. Rev. A, **32** (1985), 435–447.

[Kyp04] A. E. Kyprianou, *Travelling wave solutions to the KPP equation : alternatives to Simon Harris's probabilistic analysis.*, Ann. Inst. H. Poincaré Probab. Statist., **40** (2004), 53–72.

[KZ87] M. Kardar et Y.-C. Zhang, *Scaling of directed polymers in random media*, Phys. Rev. Lett., **58** (1987), 2087–2090.

[Lam07] A. Lambert, *Quasi-stationary distributions and the continuous-state branching process conditioned to be never extinct*, Elect. Journ. Prob., **12** (2007), 420–446.

[Les45] P. Leslie, *On the use of matrices in certain population mathematics*, Biometrika, **33** (1945), 184–212.

[LPP95] R. Lyons, R. Pemantle et Y. Peres, *Conceptual proofs of L Log L criteria for mean behavior of branching processes*, Ann. Prob., **23** (1995), 1125–1138.

[LS06] V. Limic et A. Sturm, *The spatial lambda-coalescent*, Elect. Journ. Prob., **11** (2006), 363–393.

[McK75] H. P. McKean, *Application of Brownian-motion to equation of Kolmogorov-Petrovskii-Piskunov*, Commun. Pure Appl. Math., **28** (1975), 323–331.

[MD86] J. Metz et O. Diekmann, *The dynamics of physiologically structured populations*, Lecture notes in biomathematics (68), Springer-Verlag (1986).

[MMQ08] C. Mueller, L. Mytnik et J. Quastel, *The asymptotic speed of a random traveling wave*, preprint, (2008).

[Möh06] M. Möhle, *On sampling distributions for coalescent processes with simultaneous multiple collisions*, Bernoulli, **12** (2006), 35–53.

[Mor58] P. Moran, *Random processes in genetics*, Proceedings of the Cambridge Philosophical Society, **54** (1958), 60–71.

[MPS05] C. Marquet, R. Peschanski et G. Soyez, *Traveling waves and geometric scaling at nonzero momentum transfer*, Nucl. Phys. A, **756** (2005), 399–418.

[MS95] C. Mueller et R. Sowers, *Random traveling waves for the KPP equation with noise*, J. Fun. Anal., **128** (1995), 439–498.

[Odo04] G. Odor, *Universality classes in nonequilibrium lattice systems*, Rev. Mod. Phys., **76** (2004), 663–724.

[Pel97] L. Peliti, *Lectures at the summer college on frustrated system*, cond-mat/9712027, (1997).

[Pin95] R. Pinsky, *K-P-P asymptotics for non-linear diffusion in a large ball with infinite boundary date and on $\mathbb{R}^d$ with infinite initial data outside a large ball*, Com. Part. Diff. Eq., **20** (1995), 1369–1393.

[Pit99] J. Pitman, *Coalescents with multiple collisions*, Ann. Probab., **27** (1999), 1870–1902.

[PK07] S.-C. Park et J. Krug, *Clonal interference in large populations*, submitted to PNAS USA, (2007).

[PS02] M. Prähofer et H. Spohn, *Scale invariance of the PNG droplet and the Airy process*, J. Stat. Phys., **108** (2002), 1076–1106.

[Pv02] D. Panja et W. van Saarloos, *Weakly pushed nature of "pulled" fronts with a cutoff*, Phys. Rev. E, **65** (2002), 057202.

[RBW07] I. Rouzine, É. Brunet et C. Wilke, *The traveling wave approach to asexual evolution : Muller's ratchet and speed of adaptation*, arXiv : 0707.3469, (2007).

[RWC03] I. Rouzine, J. Wakeley et J. M. Coffin, *The solitary wave of asexual evolution*, Proc. Natl. Acad. Sci. USA, **100** (2003), 587–592.

[SAF92] Z.-S. She, E. Aurell et U. Frisch, *The inviscid Burgers equation with initial data of brownian type*, Comm. Math. Phys., **148** (1992), 623–641.

[Sag99] S. Sagitov, *The general coalescent with asynchronous mergers of ancestral lines*, J. Appl. Prob., **36** (1999), 1116–1125.

[Sch00] J. Schweinsberg, *Coalescents with simultaneous multiple collisions*, Elect. Journ. Prob., **5** (2000), 1–50.

[SD06] D. Simon et B. Derrida, *Evolution of the most recent common ancestor of a population with no selection*, J. Stat. Mech., (2006), P05002.

[SD08] ———, *Quasi-stationary regime of a branching random walk in presence of an absorbing wall*, J. Stat. Phys., (2008).

[SE04] D. Steinsaltz et S. Evans, *Markov mortality models : Implications of quasistationarity and varying initial conditions*, Theoretical Population Biology, **65** (2004).

[SE05] D. Steinsaltz et S. N. Evans, *Quasistationary distributions for one-dimensional diffusions with killing*, Amer. Math. Soc., **359** (2005), 1285–1324.

[Ser05] M. Serva, *On the genealogy of populations : trees, branches and offsprings*, J. Stat. Mech., (2005).

[SH05] G. Sella et A. E. Hirsh, *The application of statistical physics to evolutionary biology*, PNAS, **120** (2005), 9541–9546.

[Sin92] Y. G. Sinai, *Statistics of shocks in solutions of inviscid Burgers equation*, Comm. Math. Phys., **148** (1992), 601–621.

[Sla68] R. S. Slack, *A branching process with mean one and possibly infinite variance*, Probab. Th. Rel. Fields, **9** (1968), 139–145.

[SR00] M. Slatkin et B. Rannala, *Estimating allele age*, Annu. Rev. Genomics Hum. Genet., **01** (2000), 225–249.

[Sto76] A. Stokes, *On two types of moving front in quasilinear diffusion*, Mathematical Bioscience, **31** (1976), 307–315.

[STW84] I. Saunders, S. Tavaré et G. Watterson, *On the genealogy of nested subsamples from a haploid population*, Adv. Appl. Prob., **16** (1984), 471–491.

[SV66] E. Seneta et D. Vere-Jones, *On quasi-stationary distributions in discrete-time markov chains with a denumerable infinity of states*, J. Appl. Prob., **3** (1966), 403–434.

[Taj83] F. Tajima, *Evolutionary relationship of DNA sequences in finite populations*, Genetics, **105** (1983), 437–460.

[Taj89] ———, *Statistical method for testing the neutral mutation hypothesis by dna polymorphism*, Gentics, **123** (1989), 585–595.

[TBGD97] S. Tavaré, D. J. Balding, R. C. Griffiths et P. Donnelly, *Inferring coalescence times from DNA sequence data*, Genetics, **145** (1997), 505–518.

[THVL05] U. Täuber, M. Howard et B. Vollmayr-Lee, *Applications of field-theoretic renormalization group methods to reaction-diffusion problems*, J. Phys. A, **38** (2005), R79.

[TLK96] L. S. Tsimring, H. Levine et D. A. Kessler, *RNA virus evolution via a fitness-space model*, Phys. Rev. Lett., **76** (1996), 4440–4443.

[van98] W. van Saarloos, *Three basics issues concerning interface dynamics in nonequilibrium pattern formation*, Phys. Rep., **301** (1998), 9–43.

[van03] ———, *Front propagation into unstable states*, Phys. Rep., **386** (2003), 29–222.

[WA01] J. Wakeley et N. Aliacar, *Gene genealogies in a metapopulation*, Genetics, **159** (2001), 893–905.

[Wak07] J. Wakeley, *Coalescent theory : an introduction*, Roberts & Co Publishers (2007).

[Wat75] E. Watterson, *On the number of segregating sites in genetical models without recombination*, Theor. Popul. Biol., **7** (1975), 256–276.

[Weg76] F. J. Wegner, *The Critical State, General Aspects*, *Phase Transitions and Critical Phenomena*, tome 6, Academic Press (1976) .

[WH98] H. Wilkinson-Herbots, *Genealogy and subpopulation differentiation under various models of population structure*, J. Math. Biol., **37** (1998), 535–585.

[Wri31] S. Wright, *Evolution in mendelian populations*, Genetics, **16** (1931), 97–159.

[Yag47] M. A. Yaglom, *Certain limit theorems of the theory of branching random processes*, Reports of the Academy of Sciences of USSR, **56** (1947), 795–798.

Résumé :

Cette thèse traite de la survie et des généalogies de populations en présence de sélection dans quelques modèles simples de physique statistique inspirés de la biologie.

La première partie étudie l'évolution de marches aléatoires avec branchements unidimensionnelles en présence d'un seuil de survie qui croît linéairement au cours du temps. En reliant les propriétés de ces marches aléatoires à une équation de propagation de fronts, nous étudions la transition vers l'extinction de ces marches lorsque la vitesse du seuil croît et obtenons les comportements critiques de la probabilité de survie. Nous construisons également un processus biaisé décrivant une population de telles marches conditionnée sur sa taille à un instant final. Cette construction permet d'étudier le régime quasi-stationnaire près de la vitesse critique. Enfin, nous présentons un modèle exactement soluble sur lequel plusieurs conjectures peuvent être vérifiées.

Dans une seconde partie, nous étudions des populations de taille constante du point de vue des généalogies et des temps de coalescence. Nous expliquons dans quelle mesure certains modèles d'évolution avec sélection se rapprochent des modèles de polymères dirigés et montrons plusieurs résultats numériques qui mettent en évidence l'existence de classes d'universalité dans les généalogies. En absence de sélection, nous étudions la dynamique des temps de coalescence et de l'âge de l'ancêtre commun d'une population, ainsi que les corrélations de ce dernier avec la diversité génétique dans un cas simple.

Mots-clefs : marche aléatoire avec branchements – transition vers un état absorbant – propagation de fronts – régime quasi-stationnaire – généalogies – dynamique de populations – effet de la sélection

Abstract :

This thesis presents a series of works dealing with the survival and the genealogies of populations in presence of selection in several simple models of statistical physics related to biology.

The first part focuses on the evolution of one-dimensional branching random walks in presence of an absorbing threshold which increases linearly in time. We relate the properties of these walks to travelling waves and we study the transition to extinction which occurs as the velocity of the threshold increases as well as the critical behaviour of the survival probability. We also develop a biased process which allows us to study a population of such walks conditioned on its size at a given final time. This process is used in order to study the quasi-stationary regime near the critical velocity. Finally, we present a exactly solvable model, for which several conjectures can be verified.

In the second part, we study populations with a constant size from the point of view of the genealogies and of the coalescence times. We explain how some evolutionary models with selection can be related to models of directed polymers and we present numerical results which tend to show the existence of universality classes in the genealogies. In absence of selection, we study the dynamics of the coalescence times and of the age of the most recent common ancestor of a population, as well as the correlations between this age and the genetic diversity in a simple case.

Keywords : branching random walks – transition to an absorbing state – travelling waves – quasi-stationary regime – genealogies – population dynamics – effect of selection