UNIVERSITÉ PARIS DESCARTES
UFR de Mathématiques et Informatique
École doctorale Mathématiques Paris Centre

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PARIS DESCARTES

Discipline : Mathématiques

Présentée par

Arno SIRI-JÉGOUSSE

Étude des généalogies dans des modèles de génétique des populations

Soutenue le 26 novembre 2009 devant le jury composé de :

Brigitte CHAUVIN	Présidente du jury
Jean-François DELMAS	Directeur de thèse
Jean-Stéphane DHERSIN	Directeur de thèse
Olivier FRANÇOIS	Rapporteur
Valentine GENON-CATALOT	
Amaury LAMBERT	
Martin MÖHLE	Rapporteur

À Wissem, Samir, Alexandre
et à tous ceux qui auraient pu
beaucoup apporter à la recherche
(mais qui ont préfére faire
du contrôle de gestion)

Remerciements

Le jour le plus important de la vie d'un thésard est celui où il apprend qu'il y aura des remerciements dans son manuscrit. À partir de ce moment, chaque anecdote, chaque bon mot est ponctué d'un « il faudra que je le mette dans mes remerciements ». Je me rends compte aujourd'hui de l'intérêt d'avoir un calepin dans sa poche, j'ai tout oublié. Je m'excuse donc auprès de tous ceux qui auraient dû apparaître dans ces bonnes feuilles.

Avant toute chose, et n'en déplaise à Arthur (que je remercierai plus tard), mes pensées vont à ceux qui ont lu ce travail (parfois plusieurs fois), qui l'ont commenté, m'ont soutenu et m'ont permis de l'améliorer. Mes directeurs, Jean-François Delmas et Jean-Stéphane Dhersin, mes rapporteurs, Olivier François et Martin Möhle, ainsi que Brigitte Chauvin, Valentine Genon-Catalot et Amaury Lambert qui m'ont honoré de leur présence, de leur intérêt et de leur pertinence lors de la soutenance. Cette thèse est l'aboutissement de dix années passées à l'université (moins une semaine en prépa), elles furent extraordinaires. Peut-être est-ce dû au hasard, mais avoir Amaury Lambert, Nathanaël Enriquez ou Arnak Dalalyan comme chargé de TD, suivre les cours de Jean-François Le Gall, de Zhan Shi, de Sacha Tsybakov ou d'Alison Etheridge, rencontrer Katia Mezaini, Mohamed Hebiri ou Vincent Bansaye à la sortie d'un amphi, présenter ses résultats devant Warren Ewens et Serik Sagitov, je n'en demandais pas tant ! Je me rends bien compte de la chance que j'ai d'être payé pour observer l'intelligence humaine, d'autres ne l'ont pas eue, je pense aux thésards d'autres domaines, devant monter des plans d'une ingéniosité sans pareille pour obtenir un financement.

Je tiens à remercier les membres de l'ANR MAEV pour l'organisation d'événements si intéressants, agréables et motivants, les membres du MAP5, Christine Graffigne, Annie Raoult et Marie-Hélène Gbaguidi pour leur disponibilité et ce qu'elles ont fait pour que je sois dans les meilleures conditions possibles. Et bien sûr, tous ceux qui ont partagé mon bureau, j'ai particulièrement apprécié les joutes verbales que Dali même n'aurait pas imaginées.

Si je devais remercier tous les gens qui m'ont accompagné ces quatre années... Oh et puis si je vais quand même le faire (ou presque). Avant tout, je ne remercie pas Matthias qui a préféré faire bronzette dans les Vosges que de venir à ma soutenance. Je remercie en revanche le reste de l' équipe de colocataires fous. Arthur, Théo, Koya qui ont transformé Stendhal en un club de vacances à l'année (activités : micro machines et chandelles). Olivier, qui a cru m'apprendre que dormir est une perte de temps et Riad qui, au moment

où j'écris ces lignes, vient de terminer le 743è dessin de son film d'animation (courage vieux, plus que 2000...). Je ne remercie pas du tout Felix pour tous ces vendredis gâchés à boire des coups. Si au moins on avait noté tout ce qu'on a dit, peut-être aurions nous révolutionné le surréalisme et été édités par Chalumeau. Je remercie très très fort Clarisse, Anne et Cécile, garantes de l'apéro, du festif et du culturel, et de beaucoup de bons moments. Merci à Xavier et Théo pour la musique et les bonnes soirées. À ce propos, je remercie Theo Parrish, Soil & Pimp Sessions et Jneiro Jarel. Merci à Alexis, mon plus vieux pote, pour ses cadeaux en gros caractères, son vase Ming, et ces grands moments en Avignon. Une bise à toute la compagnie Los Figaros, d'ailleurs. Paul, l'homme aux idées foisonnantes, Layek et Kebba, toujours prêts à le suivre dans des plans incroyables, les mecs nos idées sont bonnes. Just et Camille (ma biche), merci pour vos compiles. Si ces années ont été aussi bonnes, c'est sûrement grâce à toutes ces occasions que j'ai eues de varier les plaisirs, de participer à tant de projets et d'en voir toujours plus. Paul, encore une fois, Matthias, Félicie, Koya, merci pour ces voyages de par le monde. Une dédicace à mes basketteurs du dimanche (Cap 18, les filets en fer et Says le meneur d'hommes). Merci à Noy pour son canard au basilic, merci à RFI, merci à Jean-Paul II. Je tiens à dire que Ninnin est un champion.

Enfin, mes dernières fleurs vont à Matthias et Nora (pour beaucoup trop de raisons), Maman Siri et Papa Jégousse et à mon frère, Théo, qui n'a pas besoin de longues études pour cartonner, et je pense qu'il est bon de le rappeler au moment de rendre sa thèse.

Table des matières

Introduction et présentation des résultats **2**

A Étude des généalogies dans des modèles de génétique des populations 13

1 Les modèles classiques de la génétique de populations **15**
1.1 Le modèle de Cannings 16
1.2 Approximation par des diffusions 17
1.3 Le coalescent de Kingman 23
1.4 Dualité entre le coalescent de Kingman et la diffusion de Wright-Fisher . . 29
1.5 Ajouter des mutations 30

2 Les deux plus anciennes familles d'une population sans sélection **41**
2.1 Le modèle de Moran 43
2.2 Le processus de Fleming-Viot 45
2.3 Une construction dénombrable : le processus du look-down 46
2.4 L'étude des deux plus anciennes familles 50

3 Coalescent à collisions multiples et estimation du taux de mutation **59**
3.1 Le coalescent à collisions multiples 60
3.2 L'arbre de coalescence et les taux de mutation 64

B Articles 77

4 Les deux plus anciennes familles dans le processus... **79**
4.1 Introduction 80
4.2 Presentation of the main results on the conditional distribution 84
4.3 Stationary distribution of the relative size for the two oldest families 93
4.4 Proofs 96

5 Résultats asymptotiques sur la longueur d'arbres de coalescence 109
5.1 Introduction 110
5.2 Law of the first jump 115
5.3 Asymptotics for the number of jumps 122
5.4 First approximation of the length of the coalescent tree 128
5.5 Limit distribution of $\hat{L}_t^{(n)}$ 130
5.6 Proof of the main result 134

Bibliographie 139

Introduction et résultats principaux

Il n'aura échappé à personne que les individus d'une même espèce ne sont pas identiques. Les différences entre les êtres sont expliquées en partie par la variabilité génétique. Une variabilité si forte qu'à de rares exceptions près, chacun peut se considérer comme unique. La génétique des populations a pour but de comprendre d'où provient ce foisonnement de caractères, et dans quelle mesure la variabilité génétique est partagée entre des facteurs démographiques, le type de reproduction, la sélection naturelle ou des fluctuations dues au hasard. La complexité de cette étude n'est plus à démontrer et la modélisation de certains phénomènes s'avère être une tâche extrêmement délicate.

L'une des premières, et l'une des plus célèbres, représentations mathématiques de l'évolution des populations est le modèle de (Bienaymé-)Galton-Watson (Watson et Galton (1875) reprenant les travaux de Bienaymé développés au siècle précédent), introduit afin d'étudier la persistance des patronymes des familles nobles anglaises. Ce modèle considère une population dont chaque individu se reproduit suivant une même loi et indépendamment les uns des autres.

Au cours du XIX^{e} siècle, les découvertes de Mendel (Figure 1), publiées en 1866 (Mendel (1866)) mais redécouvertes par la communauté scientifique trois décennies plus tard, permettent d'établir les premières représentations des mécanismes de transmission héréditaire des caractères. L'idée de modéliser l'évolution des populations afin de prédire la persistance ou l'extinction d'un type (génétique bien qu'il n'en soit pas encore question) fait alors son chemin. La loi de Hardy-Weinberg (des noms d'un mathématicien anglais et d'un physicien allemand qui développent indépendamment ce résultat en 1908, voir le Chapitre 1 de Barton *et al.* (2007)) établit le postulat d'un équilibre de la fréquence des caractères au cours des générations. Par la suite, Fisher (1930) et Wright (1931) créent un schéma d'évolution de populations à taille fixe qui est encore de nos jours intensément utilisé. Les probabilités ne sont pas le seul outil mathématique appliqué à ce domaine. Le développement de méthodes statistiques modernes suit de près les découvertes en biologie de l'évolution puisque Pearson (l'inventeur du test du χ^2) a participé aux débats houleux sur l'évolution dans l'Angleterre de la fin du XIX^{e} siècle.

Dans les premiers modèles, les hypothèses sont simplifiées à l'extrême : on y considère une population

– haploïde : la reproduction est asexuée,

– homogène : le nombre d'enfants de chaque individu suit la même loi,

FIGURE 1 – Gregor Mendel (1822-1884)

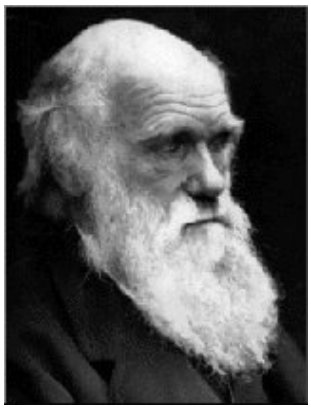

FIGURE 2 – Charles Darwin (1809-1882)

– sans chevauchement dans les générations : tous les individus se reproduisent au même moment.

En considérant que l'information génétique se transmet à l'identique de mère en fille, ces représentations permettent d'ores et déjà d'étudier les fluctuations aléatoires de la fréquence des types présents dans la population. La découverte de l'ADN et celle de nouveaux mécanismes de transmission, ainsi que les avancées dans le domaine des probabilités ont permis par la suite d'appréhender des modèle plus complexes.

Très rapidement, des mécanismes de reproduction sexuée sont intégrés (la loi de Hardy-Weinberg en est un fameux exemple). Dans ce cadre, chaque individu possède de l'information génétique provenant de ses deux parents (des paires de chromosomes), la population est dite diploïde. Il s'avère que des phénomènes de recombinaison peuvent se produire lors de la création des gamètes : des échanges entre le deux chromosomes d'une même paire qui intensifient le mélange de l'information.

Au moment de la reproduction, des modifications peuvent s'opérer dans l'ADN transmis d'une génération à l'autre. Les mutations sont l'ingrédient principal de la variabilité génétique. Sans elles la création de nouveaux caractères serait impossible. Une mutation peut être neutre, avantager ou désavantager l'individu qui la porte. La sélection naturelle compte parmi les grands principes de l'évolution. Elle est introduite par Darwin (Figure 2) en 1859 dans son ouvrage « On the origin of species » (Darwin (1859)) et, après des décennies de controverse, est apparue comme le mécanisme de base de l'évolution. Certains individus, dotés d'un type génétique mieux adapté aux conditions extérieures voient leur propension à se reproduire augmentée, ce qui facilite la probabilité de fixation de ce type dans la population - un type se fixant lorsque tout le monde le possède. Il est donc possible de considérer des modèles avec sélection, tout à fait justifiés par de multiples exemples d'adaptation au milieu. L'hégémonie de la sélection naturelle est remise en cause, grâce aux premières observations à un niveau moléculaire, par Kimura (1968) (Figure 3) lorsqu'il introduit la théorie neutraliste. Il affirme alors que la variation génétique

à l'intérieur des espèces est trop importante pour n'être expliquée que par des mutations soumises à la sélection naturelle. Il en vient à la conclusion qu'une très grande partie de la variabilité génétique provient de mutations neutres. La controverse est alors relancée. à l'heure actuelle nous ne savons toujours pas quelle fraction du génome est maintenue par la sélection naturelle. Toujours est-il que les modèles mathématiques neutres retrouvent de l'intérêt auprès des scientifiques dans les années 70 et que les recherches continuent sur les deux fronts, avec ou sans sélection.

FIGURE 3 – Motoo Kimura (1924-1994)

Enfin, l'hypothèse de générations sans chevauchement peut être relaxée en supposant que chacun se reproduit à des temps aléatoires indépendants ou non, identiquement distribués ou non. Nous pouvons facilement justifier la pertinence de cette relaxation pour de nombreuses espèces.

D'autres modèles permettent de considérer des populations dont la taille varie dans le temps. Le premier d'entre eux est celui de Galton-Watson. Il est aussi possible d'intégrer des phénomènes de migration. Au fil du temps, certains individus peuvent quitter la population tandis que d'autres peuvent arriver de l'extérieur. L'évolution peut être très fortement affectée par ces arrivées. L'exemple souvent utilisé est celui d'une île où débarquent sporadiquement des individus du continent.

Les objets modélisant l'évolution des populations sont assez différents suivant que l'on avance ou que l'on recule dans le temps. Si l'on fixe un instant et que l'on décide de tracer la généalogie des individus vivant à cet instant, il est à noter que les lignées ayant disparu avant n'apparaissent pas dans l'arbre obtenu. Les liens entre les processus construits en avançant et en remontant dans le temps ne sont d'ailleurs pas toujours simples à établir. Nous nous attacherons, dans cette thèse, à les mettre en évidence.

L'une des grandes avancées proposées par les mathématiques, consiste en l'établissement de limites, en faisant croître la taille de la population, pour les processus correc-

tement renormalisés (en temps et en taille). Kimura (1957), étudiant l'évolution de la fréquence d'un type génétique dans une population, établit la convergence en loi de ce processus vers une *diffusion* : la diffusion de Wright-Fisher. Les différentes versions, suivant que le modèle soit avec ou sans mutations et sélection sont résumées dans Ewens (2004). En remontant le temps à partir d'un instant fixé, les objets limites obtenus sont des processus de *coalescence*, ils s'avèrent très différents suivant que le modèle soit neutre ou que certains individus de la population soient avantagés. Le premier cas mène au coalescent *standard*, introduit par Kingman (1982a,b,c) (Figure 4), tandis que l'on parlera de coalescent *structuré* dans le second (Kaplan *et al.* (1988)).

FIGURE 4 – Sir John Kingman (1939-)

Les résultats proposés dans cette thèse sont obtenus pour des modèles neutres. Ils reposent, à la différence des modèles avec sélection, sur la propriété d'*échangeabilité*, indispensable à leur construction. Notre Chapitre 1 détaille les principaux modèles neutres : le modèle de Cannings, la diffusion de Wright-Fisher, le coalescent de Kingman, leur définition, leurs premières propriétés et leur intérêt en génétique des populations. Nous renvoyons le lecteur à Durrett (2008) pour une revue d'effectif des autres modèles appliqués à ce domaine.

Lorsque l'on trace les généalogies d'une population, l'arbre obtenu remonte jusqu'au plus récent ancêtre commun. Ainsi la totalité de la population étudiée peut être regroupée en quelques grandes familles. Pour savoir leur nombre il suffit de regarder combien de branches fusionnent lors de la dernière coalescence, celle qui mène au plus récent ancêtre commun (Figure 5). Supposons par exemple qu'il n'y ait que deux grandes lignées ancestrales. Il peut s'avérer intéressant d'étudier le comportement de ces deux plus anciennes familles lorsque le temps évolue, leur fréquence dans la population actuelle, le moment où l'une des deux disparaît et et comment, à ce moment, est regroupée la population en nouvelles familles ancestrales, issues d'un ancêtre commun qui aura changé... Il peut

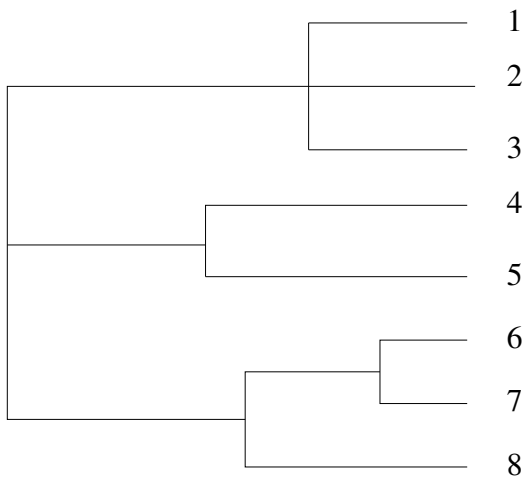

FIGURE 5 – Les généalogies de 8 individus. L'ensemble de la population peut être regroupé en 3 familles.

apparaître dès à présent au lecteur peu habitué au sujet que la première difficulté consiste ici à donner des explications claires pour décrire les problématiques et les résultats. Nous nous efforcerons de proposer une présentation complète, rigoureuse et illustrée de notre étude dans le Chapitre 2 de ce document.

Nous nous proposons, dans le Chapitre 3, de contribuer à l'étude des généalogies d'un groupe d'individus à un temps fixé. Nous aurons l'occasion de voir que, suivant le comportement des populations au cours du temps, les arbres généalogiques qui en découlent ont des formes limites qui peuvent varier (les lignées de populations marines sont par exemple différentes de celles de bactéries ou de mammifères terrestres, nous expliquerons pourquoi dans le Chapitre 3). En observant le génome d'un groupe d'individus d'une même espèce, il est possible de quantifier le nombre de différences entre chacun de leur génome. Nous nous intéresserons ainsi à la théorie de la coalescence et expliquerons en quoi une étude fine de la longueur des arbres permet d'estimer le taux de mutation apparues dans les lignées ancestrales d'une population.

Présentation des résultats principaux

Cette section propose au lecteur habitué aux concepts mathématiques de la génétique des populations les résultats établis dans ce document de thèse. Nous nous contentons ici d'énoncer les notations et les propositions. Nous proposons dans la Partie A des prélimi-

naires détaillés et le cheminement permettant d'accéder aux résultats principaux.

Évolution des deux plus anciennes familles d'une population sans sélection

Considérons une population dont les généalogies, correctement renormalisées, sont représentées à la limite par un coalescent de Kingman (Kingman (1982a)). Cette population peut donc être représentée, en avançant dans le temps, par un processus à valeurs mesures, le processus de Fleming-Viot (Fleming et Viot (1978, 1979)). Celui-ci peut être représenté de manière dénombrable par la construction look-down de Donnelly et Kurtz (1999).

Dans ce cadre, l'ensemble de la population peut être regroupé en deux lignées ancestrales. Nous nous intéressons au comportement, en régime stationnaire, de la fréquence de ces deux plus anciennes familles. Les résultats que nous obtenons sont dans la continuité des travaux de Pfaffelhuber et Wakolbinger (2006), ils sont détaillés dans l'article

On the two oldest families in a Wright-Fisher process

écrit avec Jean-François Delmas et Jean-Stéphane Dhersin et admis avec révisions pour parution dans *Electronic Journal of Probability* (Delmas *et al.* (2009)), reporté sans modification dans le Chapitre 4.

Les deux familles les plus anciennes sont issues des deux descendants du plus récent ancêtre commun dans l'arbre de coalescence. Les valeurs auxquelles nous nous intéressons sont les suivantes :

- A : le temps de naissance du plus récent ancêtre commun de la population actuelle.
- $\tau \geq 0$: le temps à attendre avant la disparition de l'une des deux familles. En d'autres termes, le temps à attendre avant le prochain changement de plus récent ancêtre commun dans la population.
- $L \in \mathbb{N}^*$: le nombre d'individus en vie actuellement et dont certains des descendants vivront au temps τ.
- $Z \in \{0, \ldots, L\}$: le nombre d'individus en vie actuellement et dont un descendant deviendra le plus récent ancêtre commun de la population dans le futur.
- $Y \in (0, 1)$: la fréquence de la famille qui restera lors du changement de plus récent ancêtre commun.
- $X \in (0, 1)$: la fréquence de l'une des deux plus anciennes familles, choisie au hasard.

La stationnarité nous permet d'omettre les indices de temps dans nos notations.

L'argument principal utilisé ici est la propriété PASTA (Poisson Arrivals See Time Average) qui peut être trouvée dans Brémaud *et al.* (1992) et qui nous permet de considérer la population au moment du changement de plus récent ancêtre commun pour déduire des résultats à un temps quelconque, étant entendu que ces temps aléatoires sont les temps de saut d'un processus de Poisson (rvoir Pfaffelhuber et Wakolbinger (2006)).

Notons $(E_k, k \in \mathbb{N}^*)$ une suite de variables aléatoires exponentielles indépendantes de

paramètre 1,

$$T_T = \sum_{k\geq 2} \frac{2}{k(k+1)} E_k$$

et

$$T_K = E_1 + T_T.$$

Théorème 1. *i) A est indépendante de (Y, X, τ, L, Z) et a la même loi que T_K à un temps fixé et que T_T au moment du changement de plus récent ancêtre commun de la population.*

à un temps fixé t ou au moment de changement de plus récent ancêtre commun, nous avons

ii) Conditionnellement à Y, X et (τ, L, Z) sont indépendantes.

iii) Conditionnellement à (Y, L), τ et Z sont indépendantes.

iv) Conditionnellement à Y, $X = \varepsilon Y + (1-\varepsilon)(1-Y)$ avec ε une variable aléatoire de Bernoulli de paramètre 1/2 indépendante de Y.

v) Conditionnellement à Y, L suit une loi géométrique de paramètre $1-Y$.

vi) Conditionnellement à (Y, L), $\tau = \sum_{k=L}^{\infty} \frac{2}{k(k+1)} E_k$ où les variables E_k sont i.i.d., exponentielles de paramètre 1, et indépendantes de (Y, L).

vii) Pour $u \in [0,1]$ et $a \geq 1$,

$$\mathbb{E}[u^Z | Y, L = a] = \begin{cases} 1 & si \quad a = 1 \\ \frac{u}{3}\frac{a+1}{a-1}\prod_{k=2}^{a-1}\left(\frac{2u}{(k-1)(k+2)}\right) & si \quad a \geq 2 \end{cases}$$

avec la convention que $\prod_{\emptyset} = 1$.

Nous verrons aussi que les individus qui deviendront le plus récent ancêtre commun naissent suivant un processus de Poisson de paramètre 1 (Pfaffelhuber et Wakolbinger (2006)). Sachant Y et L, la somme de τ (qui correspond au temps qu'il reste avant le prochain changement de plus récent ancêtre commun) et de Z variables exponentielles (qui représente intuitivement le temps qu'il s'est écoulé depuis le précédent changement de plus récent ancêtre commun) correspond à la taille d'un coalescent de Kingman. Lr résultat suivant traduit cette relation intuitive entre τ et Z :

Proposition 1. *Pour tout $\lambda \geq 0$*

$$\mathbb{E}[\mathrm{e}^{-\lambda\tau} | Y, L] = \mathbb{E}[\mathrm{e}^{-\lambda T_K}]\mathbb{E}[(1+\lambda)^Z | Y, L].$$

Nous nous attacherons par la suite à étudier la stationnarité des processus $(X_t, t \geq 0)$ et $(Y_t, t \geq 0)$. Ces processus peuvent être définis comme des diffusions de Wright-Fisher (voir la Section 1.2) ressuscitées suivant une mesure μ, représentant la taille de l'une des deux nouvelles familles, lorsqu'elles touchent les bornes de leur domaine de définition $[0, 1]$. Dans le second cas, puisque l'on représente la famille qui se fixe dans la population, la diffusion considérée est conditionnée à toucher 1 avant 0. Le résultat obtenu est le suivant :

Proposition 2. *i) Si la loi de resurrection de $(X_t, t \geq 0)$ est uniforme sur $(0,1)$, alors c'est aussi sa loi stationnaire.*
ii) Si la loi de resurrection de $(Y_t, t \geq 0)$ est la loi Beta(2,1), alors c'est aussi sa loi stationnaire.

Coalescent à collisions multiples et estimation du taux de mutation

Les Chapitres 3 et 5 décrivent le comportement asymptotique, lorsque la taille de la population initiale tend vers l'infini, des processus de coalescence représentant les généalogies limites d'un groupe d'individus considéré à un instant fixé. Ces travaux sont détaillés dans l'article

Asymptotic results on the length of coalescent trees

écrit avec Jean-François Delmas et Jean-Stéphane Dhersin et paru en 2008 dans *The Annals of Applied Probability* (Delmas *et al.* (2008)), reporté sans modification dans le Chapitre 5.

Considérons une population de taille n dont les généalogies sont représentées par un coalescent à collisions multiples (Pitman (1999), Sagitov (1999)). Les dynamiques de ce processus sont entièrement décrites par une mesure finie Λ sur $[0,1]$ (voir (3.5)).

Notons

$$\rho(t) = \int_t^1 x^{-2}\Lambda(dx)$$

et supposons que

$$\rho(t) = C_0 t^{-\alpha} + \mathcal{O}(t^{-\alpha+\zeta})$$

avec $\alpha \in (1,2)$, $C_0 > 0$ et $\zeta > 1 - 1/\alpha$. Cette condition est vérifiée par le $Beta$-coalescent, dont la mesure Λ est celle d'une loi $Beta(2-\alpha, \alpha)$, qui apparaît comme la limite des généalogies de certains modèles naturels en génétique des populations (Schweinsberg (2003)).

Dans le cadre du modèle à infinité de sites (*infinite sites model*, Kimura (1969)), le nombre total de mutations apparues dans les lignées d'un échantillon d'individus correspond au nombre de sites de ségrégation observés dans leur ADN. Notons $S^{(n)}$ cette valeur lorsque l'échantillon est de taille n..

Supposons que les mutations apparaissent à taux θ dans chaque lignée ancestrale. Conditionnellement à $L^{(n)}$, la longueur totale de l'arbre de coalescence de la population, le nombre total de mutations $S^{(n)}$ suit alors une loi de Poisson de paramètre $\theta L^{(n)}$.

L'estimation de θ est un problème tout à fait intéressant en génétique des populations. Elle permet d'évaluer le taux de mutation (nombre de mutations par paire de baes par génération) dans une population (les plus élevés sont de l'ordre de 10^{-4} dans l'ARN de certains virus et l'on observe des taux de l'ordre de 10^{-10} chez les humains). Le traitement de ce problème pour de nombreux modèles est résumé par Tavaré (2004).

Nous nous proposons donc de poursuivre les travaux de Berestycki *et al.* (2008) afin de décrire les fluctuations de $S^{(n)}$, lorsque n tend vers l'infini, dans le cadre du Beta-coalescent.

Pour ce faire, nous devons d'abord établir un résultat intermédiaire sur le comportement asymptotique de τ_n, le nombre total d'événements de coalescence jusqu'au plus récent ancêtre commun des n individus initiaux. Ce résultat rejoint les travaux parus simultanément de Gnedin et Yakubovich (2007) et Iksanov et Möhle (2008) obtenus avec des méthodes différentes. Soit $V = (V_t, t \geq 0)$ un processus de Lévy α-stable avec des sauts négatifs dont l'exposant de Laplace $\psi(u)$ est $u^\alpha/(\alpha - 1)$.

Proposition 3. *Supposons que* $\rho(t) = C_0 t^{-\alpha} + \mathcal{O}(t^{-\alpha+\zeta})$ *avec* $\alpha \in (1, 2)$, $C_0 > 0$ *et* $\zeta > 1 - 1/\alpha$. *La convergence en loi suivante s'opère lorsque* $n \to \infty$ *:*

$$n^{-1/\alpha}\left(n - \frac{\tau_n}{\alpha - 1}\right) \xrightarrow{d} V_{\alpha-1}.$$

Nous obtenons un résultat partiel de convergence sur la longueur de l'arbre. Notons $\lfloor x \rfloor$ la partie entière de x. Soit $L_t^{(n)}$ la longueur de l'arbre coupé au $\lfloor nt \rfloor^{\text{e}}$ événement de coalescence. Pour t dans $(0, \alpha - 1)$, soit

$$v(t) = \int_0^t \left(1 - \frac{r}{\alpha - 1}\right)^{1-\alpha} dr,$$

alors, lorsque $n \to \infty$,

$$L_t^{(n)} \sim n^{2-\alpha} \frac{v(t)}{C_0 \Gamma(2 - \alpha)}.$$

Lorsque n croît, τ_n se comporte comme $n(\alpha - 1)$. En approchant $L^{(n)} = L^{(n)}_{\frac{\tau_n}{n}}$ par $n^{2-\alpha} \frac{v(\alpha-1)}{C_0\Gamma(2-\alpha)}$, nous retrouvons le résultat de convergence de Berestycki *et al.* (2008).

Nous déterminons la distribution asymptotique du nombre total de mutations dans l'arbre, du moins jusqu'à la $\lfloor nt \rfloor^{\text{e}}$ coalescence. Notons $S_t^{(n)}$ cette quantité.

Théorème 2. *Supposons que* $\rho(t) = C_0 t^{-\alpha} + \mathcal{O}(t^{-\alpha+\zeta})$ *avec* $\alpha \in (1, 2)$, $C_0 > 0$ *et* $\zeta > 1 - 1/\alpha$. *Soient* t *dans* $(0, \alpha - 1)$ *et* G *une variable aléatoire gaussienne centrée réduite indépendante du processus* V.

1. *Soit* $\alpha \in (1, \sqrt{2})$. *Alors*

$$n^{-1+\alpha-1/\alpha}\left(S_t^{(n)} - \theta n^{2-\alpha} \frac{v(t)}{C_0\Gamma(2-\alpha)}\right) \xrightarrow{d} \theta \frac{\alpha - 1}{C_0\Gamma(2-\alpha)} \int_0^t dr \left(1 - \frac{r}{\alpha - 1}\right)^{-\alpha} V_r$$

lorsque $n \to \infty$.

2. *Soit* $\alpha \in (\sqrt{2}, 2)$. *Alors*

$$n^{-1+\alpha/2}\left(S_t^{(n)} - \theta n^{2-\alpha}\frac{v(t)}{C_0\Gamma(2-\alpha)}\right) \xrightarrow{d} \sqrt{\theta\frac{v(t)}{C_0\Gamma(2-\alpha)}}G$$

lorsque $n \to \infty$.

3. *Soit* $\alpha = \sqrt{2}$. *Alors* $-1 + \alpha - \frac{1}{\alpha} = -1 + \frac{\alpha}{2}$ *et*

$$n^{-1+\alpha-1/\alpha}\left(S_t^{(n)} - \theta n^{2-\alpha}\frac{v(t)}{C_0\Gamma(2-\alpha)}\right) \xrightarrow{d} \sqrt{\theta\frac{v(t)}{C_0\Gamma(2-\alpha)}}G + \theta\frac{\alpha-1}{C_0\Gamma(2-\alpha)}\int_0^t dr\left(1-\frac{r}{\alpha-1}\right)^{-\alpha}V_r$$

lorsque $n \to \infty$.

Nous observons une transition de phase. Lorsque α appartient à l'intervalle $(1, \sqrt{2})$, les fluctuations de la longueur de l'arbre l'emportent sur celles du processus de Poisson des mutations. Le rapport de force est inversé sur l'intervalle $(\sqrt{2}, 2)$ et lorsque $\alpha = 2$, les deux sont présents et indépendants.

Première partie

Étude des généalogies dans des modèles de génétique des populations

Chapitre 1

Les modèles classiques de la génétique de populations

1.1 Le modèle de Cannings

Introduisons tout d'abord un modèle d'évolution général (Cannings (1974, 1975)). Cette présentation rapide peut être complétée par la lecture de Durrett (2008) et de Lambert (2008). Considérons une population de taille fixe N, haploïde, évoluant en générations discrètes ne se chevauchant pas. Étiquetons les individus au hasard, de manière uniforme. Pour $r \in \mathbb{N}$ et $1 \leq i \leq N$, soit Υ_i^r la taille de la descendance à la génération $r+1$ de l'individu i de la génération r. La taille fixe de la population au cours du temps impose que

$$\Upsilon_1^r + \cdots + \Upsilon_N^r = N.$$

Soit $\Upsilon^r = (\Upsilon_1^r, \ldots, \Upsilon_N^r)$. On suppose que la taille des familles est indépendante et homogène en temps :

$$\text{les } \Upsilon^r, r \geq 0 \text{ sont des copies i.i.d. d'une même loi } \Upsilon$$

et la loi Υ est *échangeable* :

$$(\Upsilon_1, \ldots, \Upsilon_N) \stackrel{d}{=} (\Upsilon_{\pi(1)}, \ldots, \Upsilon_{\pi(N)}) \text{ pour toute permutation } \pi \text{ de } \{1, \ldots, N\}.$$

Ces conditions impliquent entre autre que $\mathbb{E}[\Upsilon_1] = 1$. Afin d'éviter le cas trivial où $\mathbb{P}(\Upsilon_1 = \cdots = \Upsilon_N = 1) = 1$, nous supposerons que $\mathrm{Var}(\Upsilon_1) > 0$.

Un cas particulier du modèle de Cannings est le classique *modèle de Wright-Fisher* (Fisher (1930), Wright (1931)) où la loi Υ est multinomiale de paramètres $(N, 1/N, \ldots, 1/N)$ (voir Figure 1.1). En d'autres termes,

$$\mathbb{P}(\Upsilon_1 = n_1, \ldots, \Upsilon_N = n_N) = \frac{N!}{n_1! \ldots n_N!} \frac{1}{N^N}.$$

Ceci revient à considérer que chaque individu de la population choisit son parent au hasard parmi les éléments de la génération précédente (voir l'exposé pédagogique de Birkner (2005)).

Considérons donc un modèle de Cannings en excluant le cas où $\Upsilon_1 = 1$ p.s.. Pour tout $r \in \mathbb{N}$, soit $M^r(k)$ le nombre de descendants à la génération r des k premiers individus de la génération 0. La suite

$$(M^r, r \in \mathbb{N}) := (M^r(k), r \in \mathbb{N})$$

est une chaîne de Markov à espace d'états fini. $M^0(k) = k$ et, conditionnellement à $M^r(k)$,

$$M^{r+1}(k) = \sum_{i=1}^{M^r(k)} \Upsilon_i^r.$$

Notons que dans le cas du modèle de Wright-Fisher, la loi conditionnelle de $M^{r+1}(k)$ est celle d'une variable aléatoire binomiale de paramètres N et $M^r(k)/N$:

$$\mathbb{P}(M^{r+1}(k) = i | M^r(k) = j) = \binom{N}{i} \left(\frac{j}{N}\right)^i \left(1 - \frac{j}{N}\right)^{N-i}. \tag{1.1}$$

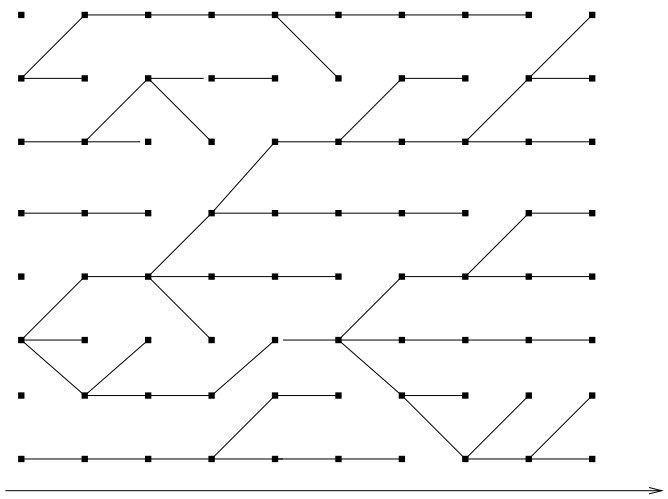

FIGURE 1.1 – Le modèle de Cannings.

Il y a deux états absorbants : 0 et N. Par conséquent,

$$M^\infty(k) = \lim_{r\to\infty} M^r(k) \text{ existe et appartient à } \{0, N\}.$$

La chaîne $(M^r, r \in \mathbb{N})$ est une martingale bornée. On peut en déduire que

$$k = \mathbb{E}[M^\infty(k)] = N\mathbb{P}(M^\infty(k) = N).$$

Définition 1.1. *Un groupe de k individus se fixe dans la population s'il existe une génération à laquelle toute la population descend de ce groupe. En d'autres termes, il y a fixation de ce groupe si l'événement $\{M^\infty(k) = N\}$ se réalise.*

Nous avons dans le cas du modèle de Cannings,

$$\mathbb{P}(M^\infty(k) = N) = \frac{k}{N}. \tag{1.2}$$

1.2 Approximation par des diffusions

Une question naturelle qui apparaît alors concerne le temps de fixation ou de disparition, d'une famille de taille k, c'est-à-dire la génération (aléatoire)

$$\tau_k^N := \inf\{r, M^r(k) \in \{0, N\}\}.$$

Ce résultat peut être calculé directement grâce à la théorie des chaînes de Markov. Une autre approche, du moins lorsque la population a une taille élevée, est de caractériser le processus limite, lorsque $N \to \infty$, de la fraction d'une famille initiée en

$$x = \frac{k}{N} \in [0, 1].$$

Pour ce faire, nous allons énoncer quelques outils de la théorie des *diffusions*. Nous proposons une introduction informelle, afin de donner une intuition des résultats. De nombreux ouvrages sont consacrés au sujet. Nous invitons le lecteur à se reporter à Dynkin (2006), Karlin et Taylor (1981), Knight (1981), Ethier et Kurtz (1986), Revuz et Yor (1999) ou Stroock et Varadhan (2006) pour un introduction exhaustive à cette théorie.

Définition 1.2. *Une diffusion $U := (U_t, t \geq 0)$ est un processus de Markov fort à trajectoires continues dans $\mathbb{R}$.*

Une diffusion n'est pas nécessairement à valeurs dans $\mathbb{R}$ tout entier. Dans ce document, nous nous intéresserons à des diffusions sur $(0, 1)$ mais pour cette introduction, nous les considérerons à valeurs dans un intervalle (a, b) (éventuellement en union avec $\{a, b\}$), avec $-\infty \leq a < b \leq +\infty$.

Soit h une fonction deux fois continûment dérivable sur (a, b). Éventuellement, suivant le comportement de U à proximité des bornes, cette fonction devra satisfaire des conditions en a et b (voir la Section 6 de Karlin et Taylor (1981) ou la la Section 2.3 de Etheridge (2009)). Pour $s, t > 0$, soit

$$\Delta_s h(U_t) = h(U_{t+s}) - h(U_t).$$

Dans les cas réguliers, le générateur infinitésimal de U est obtenu, pour tout x dans (a, b), par

$$\mathcal{L}h(x) := \lim_{s \to 0} \frac{1}{s} \mathbb{E}[\Delta_s h(U_t)] = f(x)h'(x) + \frac{1}{2} g^2(x)h''(x). \tag{1.3}$$

Pour évier tout problème, nous supposerons que les fonctions f et g^2 sont continues bornées et que g^2 est strictement positive sur (a, b). En revanche, g^2 peut s'annuler en a ou en b. Ces conditions peuvent être relaxées dans la théorie générale.

Soient $h_1 : x \mapsto x$ et $h_2 : x \mapsto x^2$. Sous de bonnes hypothèses sur f et g, la propriété de Markov et (1.3) impliquent que

$$\mathcal{L}h_1(U_t) = \lim_{s \to 0} \frac{1}{s} \mathbb{E}[\Delta_s U_t | U_t] = f(U_t).$$

Ainsi

$$\mathbb{E}[\Delta_s U_t | U_t] = sf(U_t) + o(s). \tag{1.4}$$

En remarquant que $(U_{t+s} - U_t)^2 = U_{t+s}^2 - U_t^2 - 2U_t(U_{t+s} - U_t)$,

$$\mathcal{L}h_2(U_t) - 2U_t \mathcal{L}h_1(U_t) = \lim_{s \to 0} \frac{1}{s} \mathbb{E}[(\Delta_s U_t)^2 | U_t] = g^2(U_t),$$

ce qui permet d'obtenir

$$\mathbb{E}[(\Delta_s U_t)^2 | U_t] = s g^2(U_t) + o(s). \tag{1.5}$$

Les équations (1.4) et (1.5) expliquent l'appellation de dérive et variance infinitésimales pour les termes f et g^2.

En fait, si un processus de Markov fort càdlàg $(U_t, t \geq 0)$, satisfait (1.4), (1.5) et, pour tout $p > 2$, la condition suivante :

$$\lim_{s \to 0} \frac{1}{s} \mathbb{E}[(\Delta_s U_t)^p | U_t = x] = 0,$$

où la convergence est uniforme en (x, t) sur les ensembles compacts de $(a, b) \times \mathbb{R}_+$, alors ce processus est nécessairement une diffusion (Karlin et Taylor (1981), Chapitre 15, Section 1, Lemme 1.1).

Introduisons à présent un outil fondamental pour établir des théorèmes de convergence de processus discrets vers des diffusions.

Définition 1.3. *Soient Ξ un espace métrique séparable complet, $\mathcal{D}$ une classe de fonctions réelles continues bornées et $\mathcal{L}$ un opérateur de $\mathcal{D}$ vers les fonctions mesurables bornées sur Ξ. Un processus $U := (U_t, t \geq 0)$ à valeurs dans Ξ est une solution au problème de martingale pour $\mathcal{L}$ et $\mathcal{D}$ si, pour toute fonction h de $\mathcal{D}$,*

$$h(U_t) - h(U_0) - \int_0^t \mathcal{L}h(U_s) ds \tag{1.6}$$

est une martingale (par rapport à la filtration naturelle engendrée par U).

On dit qu'un problème de martingale pour $\mathcal{L}$ et $\mathcal{D}$ est bien posé si U est l'unique solution au problème de martingale (1.6).

Une introduction aux problèmes de martingales se trouve dans Ethier et Kurtz (1986). Le cas des diffusions est traité exhaustivement dans Stroock et Varadhan (2006). Si f et g^2 sont telles que le problème de martingale associé à $\mathcal{L}$, défini par (1.3), et $\mathcal{D} = \mathcal{C}^2(a, b)$ est bien posé, alors la formule d'Itô permet d'exprimer une diffusion comme l'unique solution faible de l'équation différentielle stochastique

$$dU_t = f(U_t)dt + g(U_t)dW_t \tag{1.7}$$

où $(W_t, t \geq 0)$ est un mouvement brownien standard.

Rappelons que $\lfloor x \rfloor$ désigne la partie entière d'un réel x. En utilisant les notations de la section précédente, soit

$$U_t^N = \frac{1}{N} M^{\lfloor Nt \rfloor}(k),$$

la fréquence, dans le modèle de Wright-Fisher, de la famille issue d'une fraction x de la population initiale, où le temps a été accéléré d'un facteur N. Introduisons à présent la diffusion sur laquelle nous porterons notre étude par la suite.

Définition 1.4. *On appelle diffusion de Wright-Fisher le processus de Markov continu* $(U_t, t \geq 0)$ *de* $\mathbb{R}_+$ *dans* $[0,1]$ *(0 et 1 sont des états absorbants) dont le générateur infinitésimal est*

$$\mathcal{L}h(x) = \frac{1}{2}\, x(1-x)h''(x) \tag{1.8}$$

pour toute fonction h *deux fois continûment dérivable.*

Le problème de martingale associé à la diffusion de Wright-Fisher est bien posé. Nous renvoyons le lecteur au Chapitre 10 de Ethier et Kurtz (1986) ou à Feller (1951). Par conséquent, cette diffusion est l'unique solution de l'équation différentielle stochastique

$$dU_t = \sqrt{U_t(1-U_t)}dW_t. \tag{1.9}$$

Il est alors possible de prouver la convergence faible de la fréquence d'une famille dans un modèle de Wright-Fisher discret (Théorème 10.1 de Ethier et Kurtz (1986), voir aussi le Théorème 8.7.1 de Durrett (1996)).

Théorème 1.1. *Si la suite* $(U_0^N, N \geq 1)$ *converge en loi vers* U_0*, alors* $(U_t^N, t \geq 0)$ *converge étroitement vers la diffusion de Wright-Fisher* $(U_t, t \geq 0)$ *dans l'espace des processus càdlàg* $\mathcal{D}(\mathbb{R}_+, [0,1])$ *pour la topologie de Skorohod.*

Énonçons à présent un résultat dont les conséquences seront très utiles pour les applications à la génétique des populations (Karlin et Taylor (1981), Dynkin (2006)). Soient τ_a et τ_b les temps d'atteinte par une diffusion U de a et b. Notons

$$\tau := \tau_a \wedge \tau_b = \inf\{\tau_a, \tau_b\} = \inf\{t \geq 0, U_t \in \{a,b\}\}$$

le temps d'atteinte des bornes par une diffusion. Pour toutes fonctions c et d respectivement lipschitzienne et positive sur (a,b), les fonctionnelles additives positives de la forme

$$h(x) = \mathbb{E}_x\left[\int_0^\tau c(U_s)ds + d(U_\tau)\right], \tag{1.10}$$

sont les uniques solutions de

$$\left\{\begin{array}{rcll} -\mathcal{L}h(x) & = & c(x) & \text{si } x \in (a,b) \\ h(x) & = & d(x) & \text{si } x \in \{a,b\} \end{array}\right. \tag{1.11}$$

En choisissant $c \equiv 0$ et $d(x) = \mathbf{1}_{\{b\}}$, nous obtenons que $\mathbb{P}_x(\tau_b < \tau_a)$ est solution de

$$\left\{\begin{array}{rcll} -\mathcal{L}h(x) & = & 0 & \text{si } x \in (a,b) \\ h(a) & = & 0 & \\ h(b) & = & 1 & \end{array}\right. \tag{1.12}$$

D'autre part, en choisissant $c \equiv 1$ et $d \equiv 0$, nous obtenons que $\mathbb{E}_x[\tau]$ est solution de

$$\left\{ \begin{array}{rcll} -\mathcal{L}h(x) & = & 1 & \text{si } x \in (a,b) \\ h(a) & = & 0 & \\ h(b) & = & 0 & \end{array} \right. \tag{1.13}$$

Nous obtenons, d'après (1.12), les probabilités que U touche 1 avant 0, qui n'est autre que la probabilité pour la famille dont la fréquence évolue comme une diffusion de Wright-Fisher de se fixer dans la population :

$$v(x) := \mathbb{P}_x(\tau_1 < \tau_0) = x. \tag{1.14}$$

En utilisant à présent (1.13), il est aisé de donner une valeur de l'espérance du temps d'atteinte de 0 ou 1 :

$$\mathbb{E}_x[\tau] = -2x \log x - 2(1-x)\log(1-x). \tag{1.15}$$

Ces systèmes nous permettront de déterminer une approximation, en grande population, de la probabilité de fixation d'une famille et du temps de fixation dans un modèle de Wright-Fisher. Ainsi, posons $f \equiv 0$ et $g(x) = \sqrt{x(1-x)}$. On retrouve en (1.14) un résultat analogue à celui énoncé en (1.2) dans le cas d'une population finie. L'équation (1.15), prise en $x = k/N$, permet d'obtenir une valeur approchée de $N^{-1}\mathbb{E}[\tau_k^N]$. En effet,

$$\begin{array}{rcl} \mathbb{E}[\tau_k^N] & = & \displaystyle\sum_{\lfloor Nt \rfloor \geq 0} \mathbb{P}(\tau_k^N > \lfloor Nt \rfloor) \\ & = & \displaystyle\sum_{\lfloor Nt \rfloor \geq 0} \mathbb{P}(M^{\lfloor Nt \rfloor}(k) \notin \{0, N\}) \\ & \sim & \displaystyle N \int_0^\infty \mathbb{P}_{\frac{k}{N}}(U_t \notin \{0,1\}) \\ & = & N\mathbb{E}_{\frac{k}{N}}[\tau]. \end{array}$$

Plutôt que d'observer l'évolution de la fréquence d'une famille jusqu'à sa fixation ou son extinction, nous allons dorénavant nous intéresser à la famille qui va se fixer dans la population. En d'autres termes, nous allons conditionner la diffusion de Wright-Fisher à toucher 1. Notons $(U_t^{(1)}, t \geq 0)$ ce nouveau processus, $\mathcal{L}^{(1)}$ son générateur et $P_t^{(1)}(x, dy)$ son semi-groupe de transition. Nous allons utiliser $v(x)$ définie dans (1.14). En remarquant que $\mathcal{L}v \equiv 0$, nous pouvons déduire, dans un premier temps, que $(v(U_t), t \geq 0)$ est une martingale positive bornée. Soient $x > 0$ et h une fonction deux fois continûment dérivable

et bornée sur $[0,1]$. Nous avons

$$\begin{aligned} P_t^{(1)}h(x) &= \mathbb{E}_x[h(U_t^{(1)})] \\ &= \mathbb{E}_x[h(U_t^{(1)})|\tau_1 < \tau_0] \\ &= \frac{\mathbb{E}_x[h(U_t)\mathbf{1}\{\tau_1 < \tau_0\}]}{\mathbb{P}_x(\tau_1 < \tau_0)} \\ &= \frac{\mathbb{E}_x[h(U_t)v(U_t)]}{v(x)} \\ &= \frac{P_t(vh)(x)}{v(x)}, \end{aligned}$$

où P_t est le semi-groupe de transition de la diffusion standard de Wright-Fisher $(U_t, t \geq 0)$. On en déduit que

$$\mathcal{L}^{(1)}h(x) = \frac{\mathcal{L}(vh)(x)}{v(x)}. \tag{1.16}$$

Cette méthode a été développée par Doob sous le nom de *h-transforms* (voir Doob (2001)). Ainsi, par (1.16), nous pouvons calculer explicitement le générateur infinitésimal de $U^{(1)}$:

$$\mathcal{L}^{(1)}h(x) = \frac{1}{2}(1-x)(vh)''(x) = \frac{1}{2}(1-x)(v''h + 2v'h' + vh'')(x)$$

soit

$$\mathcal{L}^{(1)}h = (1-x)h' + \frac{1}{2}\,x(1-x)h'' \tag{1.17}$$

et en déduire l'équation différentielle stochastique associée

$$dU_t^{(1)} = (1-U_t^{(1)})dt + \sqrt{U_t^{(1)}(1-U_t^{(1)})}dW_t. \tag{1.18}$$

Remarque 1.1. *L'application principale de la diffusion de Wright-Fisher est l'étude de l'évolution de la fréquence d'un type génétique (allèle) dans une population à deux types : a et A. Tous les résultats énoncés plus haut permettent de quantifier l'influence de la dérive génétique (c'est-à-dire du hasard dû à la reproduction) dans l'évolution d'une population. Il est tout à fait possible d'obtenir des résultats de convergence de modèles discrets vers des diffusions en donnant la possibilité à chaque individu de muter (de passer du type a au type A et inversement) ou en avantageant un des deux types présents (pour représenter des phénomènes de sélection naturelle). Les processus obtenus sont des diffusions de Wright-Fisher avec mutations et/ou sélection. Leur variance infinitésimale est la même que dans le modèle neutre. En revanche, une dérive apparaît dans les deux cas (voir par exemple Ewens (2004) qui reprend les travaux de Kimura (1955a,b,c, 1957)).*

1.3 Le coalescent de Kingman

La théorie de la *coalescence* puise ses origines en physique et en génétique. En physique, des modèles sont développés afin de représenter des objets de masses différentes se déplaçant dans l'espace dans le cadre de l'équation de Smoluchowski. Quand deux objets de masse x et y se rapprochent, ils peuvent fusionner en un seul, de masse $x + y$. Ces coalescences ont lieu à taux $K(x, y)$ (Evans et Pitman (1997)). Aldous et Pitman (1998) étudient le cas additif, $K(x, y) = x + y$, tandis que Aldous (1997) (voir aussi Aldous et Limic (1998)) traite le cas multiplicatif, $K(x, y) = xy$. Les principaux résultats sont résumés dans Aldous (1999).

En génétique des populations, les premières idées à propos de coalescence apparaissent dans Griffiths (1980) mais la théorie est due à Kingman (1982a,b,c, 2000). Son utilisation est justifiée par le besoin de modéliser les généalogies de populations. Nous allons proposer un exemple et invitons le lecteur à se référer à Hudson (1991), Donnelly et Tavaré (1995), Li et Fu (1999), ou encore Nordborg (2001) pour plus d'informations.

Revenons au modèle discret de Cannings pour une population haploïde introduit dans la Section 1.1. Si l'on considère deux générations $r_1 < r_2$, tous les individus de la génération r_2 sont des descendants d'individus de la génération r_1. En revanche, tous les éléments de r_1 ne sont pas des ancêtres d'éléments de r_2. Le tracé des lignes généalogiques implique donc une réduction du nombre d'individus concernés au fur et à mesure que l'on remonte les générations. S'il n'en reste plus qu'un, il s'agit du *plus récent ancêtre commun* (voir Figure 1.2).

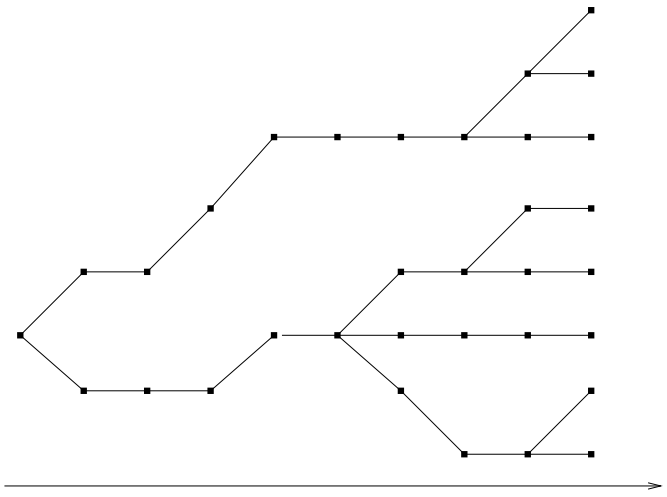

FIGURE 1.2 – Généalogies des individus de la dernière génération de la Figure 1.1.

Comme nous l'avons déjà vu, il est agréable de s'affranchir de calculs exacts, qui peuvent s'avérer fastidieux ou délicats, en considérant une population de grande taille et en décrivant l'arbre généalogique limite qui en découle. Sous certaines hypothèses, le processus limite qui apparaît ici est le *coalescent de Kingman* (Kingman (1982a,c)).

Considérons donc une population de taille N évoluant en régime stationnaire suivant un modèle de Cannings. Échantillonnons $n < N$ individus dans la population actuelle et étiquetons-les aléatoirement de 1 à n. Rappelons que Υ_i^r représente le nombre d'enfants d'un individu i à la génération r et que, pour tout $r \geq 0$, les variables aléatoires $\Upsilon^r = (\Upsilon_1^r, \ldots, \Upsilon_N^r)$ sont des copies indépendantes d'une même loi $\Upsilon = (\Upsilon_1, \ldots, \Upsilon_N)$. La probabilité que deux individus choisis au hasard à une génération donnée aient le même ancêtre à la génération précédente est

$$c_N = \frac{\sum_{i=1}^N \mathbb{E}\left[\binom{\Upsilon_i}{2}\right]}{\binom{N}{2}} = \frac{\sum_{i=1}^N \mathbb{E}[\Upsilon_i(\Upsilon_i - 1)]}{N(N-1)} = \frac{\mathbb{E}[\Upsilon_1(\Upsilon_1 - 1)]}{N-1} = \frac{\mathrm{Var}(\Upsilon_1)}{N-1} \tag{1.19}$$

où l'on a utilisé successivement dans ce calcul l'échangeabilité des $(\Upsilon_i, i \in \{1, \ldots, N\})$ et le fait que $\mathbb{E}[\Upsilon_1] = 1$.

Introduisons la relation d'équivalence suivante : si i et j sont des éléments de

$$[n] := \{1, \ldots, n\},$$

vivant à une génération donnée, alors $i \overset{r}{\sim} j$ lorsque i et j ont un ancêtre commun r générations auparavant. Nous pouvons alors définir la chaîne de Markov décrite par les classes d'équivalence correspondantes, à valeurs dans $\mathcal{P}^n$ l'ensemble des partitions de $[n]$. Nous la noterons $(\mathfrak{R}_r^{(N,n)}, r \in \mathbb{N})$. $\mathfrak{R}_0^{(N,n)}$ est la partition triviale ne contenant que des singletons.

Supposons que

$$c_N \to 0 \text{ lorsque } N \to \infty. \tag{1.20}$$

L'échangeabilité du modèle de Cannings permettant de considérer les individus 1 et 2 plutôt que i et j quelconques, nous avons

$$\mathbb{P}(1 \overset{1}{\sim} 2) = c_N$$

et

$$\mathbb{P}(1 \overset{r}{\sim} 2) = 1 - (1 - c_N)^r \sim 1 - \mathrm{e}^{-rc_N}\,. \tag{1.21}$$

Pour tout $t \geq 0$, ce calcul mène à une limite non-triviale lorsque $r = \lfloor t/c_N \rfloor$. Ceci nous donne une idée de la renormalisation en temps que nous allons devoir effectuer pour obtenir un processus limite, lequel sera une chaîne de Markov en temps continu à valeurs dans $\mathcal{P}^n$. D'après (1.21), il apparaît que dans ce processus deux lignées fusionnent après un temps exponentiel (de paramètre 1). Allons un peu plus loin dans la description (heuristique) de ce processus limite qui s'avérera être le *coalescent de Kingman* (Kingman (1982a), Möhle et Sagitov (2001)).

Est-il possible, dans le coalescent de Kingman, que plus de deux lignées fusionnent au même instant ? Pour répondre à cette question, introduisons d_N, la probabilité que trois individus choisis au hasard aient le même ancêtre à la génération précédente :

$$d_N = \mathbb{P}(1 \overset{1}{\sim} 2 \overset{1}{\sim} 3) = \frac{\sum_{i=1}^{N} \mathbb{E}\left[\binom{\Upsilon_i}{3}\right]}{\binom{N}{3}} = \frac{\mathbb{E}[\Upsilon_1(\Upsilon_1 - 1)(\Upsilon_1 - 2)]}{(N-1)(N-2)}.$$

Nous allons nous placer dans le cas où

$$\frac{d_N}{c_N} \to 0 \text{ lorsque } N \to \infty. \tag{1.22}$$

En utilisant l'échangeabilité, il apparaît que la probabilité que, à la génération où deux lignées fusionnent, une troisième se joigne à l'événement de coalescence est

$$\mathbb{P}(\{1 \overset{1}{\sim} 2\} \cap \{\exists\ 3 \leq k \leq n,\ k \overset{1}{\sim} 1\}) \leq (n-2)\mathbb{P}(1 \overset{1}{\sim} 2 \overset{1}{\sim} 3) = o(\mathbb{P}(1 \overset{1}{\sim} 2).$$

Sous l'hypothèse (1.22), il n'y aura donc pas de collisions multiples dans le processus limite.

Est-il possible, dans le coalescent de Kingman, que deux couples distincts de lignées fusionnent au même instant ? Cela revient à regarder la probabilité suivante

$$\mathbb{P}(1 \overset{1}{\sim} 2 \overset{1}{\nsim} 3 \overset{1}{\sim} 4) = \frac{\sum_{1 \leq i < j \leq N} \mathbb{E}\left[\binom{\Upsilon_i}{2}\binom{\Upsilon_j}{2}\right]}{\binom{N}{4}} = 3\frac{\mathbb{E}[\Upsilon_1(\Upsilon_1 - 1)\Upsilon_2(\Upsilon_2 - 1)]}{(N-2)(N-3)}.$$

Il y a ici quelques calculs à mener pour obtenir une réponse à notre question. Nous renvoyons le lecteur à la Section 1.2.2 de Birkner (2005) pour plus de précisions. L'hypothèse (1.22) implique que

$$\lim_{N \to \infty} \frac{\mathbb{P}(1 \overset{1}{\sim} 2 \overset{1}{\nsim} 3 \overset{1}{\sim} 4)}{\mathbb{P}(1 \overset{1}{\sim} 2)} = 0.$$

Il en découle que la probabilité que, à la génération où deux lignées fusionnent, un autre couple de lignées fusionne aussi est

$$\mathbb{P}(\{1 \overset{1}{\sim} 2\} \cap \{\exists 3 \leq k < l \leq n, k \overset{1}{\sim} l \overset{1}{\nsim} 1\}) \leq (n-2)\mathbb{P}(1 \overset{1}{\sim} 2 \overset{1}{\nsim} 3 \overset{1}{\sim} 4) = o(\mathbb{P}(1 \overset{1}{\sim} 2)).$$

Pour résumer, lorsque $N \to \infty$, en supposant que les hypothèses (1.20) et (1.22) soient vérifiées et en renormalisant le temps par c_N, les seuls événements qui restent visibles sont des coalescences de couples de lignées, apparaissant à taux 1.

Définition 1.5. *Soit $n \in \mathbb{N}^*$, le n-coalescent de Kingman est une chaîne de Markov en temps continu $(\mathcal{K}_t^{(n)}, t \geq 0)$, à valeurs dans $\mathcal{P}^n$ dont la matrice de transition est donnée, pour $\xi \neq \eta$, par*

$$Q(\xi, \eta) = \begin{cases} 1 & \text{si } \eta \text{ est obtenu en fusionnant exactement deux classes de } \xi \\ 0 & \text{sinon} \end{cases} \tag{1.23}$$

Le lecteur ne sera donc pas surpris par le résultat suivant, qui est un cas particulier du Théorème 2.1. de Möhle et Sagitov (2001).

Théorème 1.2. *Notons, pour tout* $t \geq 0$,

$$\mathcal{K}_t^{(N,n)} := \mathfrak{R}^{(N,n)}_{\lfloor c_N^{-1} t \rfloor}.$$

Si (1.20) *et* (1.22) *sont vérifiées, alors* $(\mathcal{K}_t^{(N,n)}, t \geq 0)$ *converge étroitement vers* $(\mathcal{K}_t^{(n)}, t \geq 0)$ *dans l'espace des processus càdlàg* $\mathcal{D}(\mathbb{R}_+, \mathcal{P}^n)$.

Remarque 1.2. *La réciproque est aussi vraie. La preuve de l'équivalence se trouve dans Möhle (2000) et Möhle et Sagitov (2001). Des taux de convergence sont proposés dans Möhle (2000).*

Dans le cas du modèle de Wright-Fisher,

$$\Upsilon_1 \sim \mathcal{B}\left(N, \frac{1}{N}\right)$$

où $\mathcal{B}(n,p)$ désigne une loi binomiale de paramètres n et p. Il est donc aisé de calculer c_N et d_N.

$$\begin{aligned} c_N &= \frac{1}{N-1}\,\frac{N-1}{N} = \frac{1}{N}, \\ d_N &= \frac{1}{(N-1)(N-2)}\,\frac{(N-1)(N-2)}{N^2} = \frac{1}{N^2}\,. \end{aligned}$$

Par conséquent, les généalogies d'un échantillon de n individus du modèle de Wright-Fisher, accélérées N fois, peuvent être approchées par un n-coalescent de Kingman.

Comme indiqué par sa matrice de transition, le nombre de lignées, ou de *blocs*, du n-coalescent de Kingman décroît de 1 en 1 à des temps exponentiels. Plus précisément, si $\mathcal{K}_t^{(n)}$ contient b blocs, alors T_b, le temps à attendre pour sauter de b à $b-1$ blocs, suit une loi exponentielle de paramètre $\binom{b}{2}$. La propriété de Markov assure que les variables $T_n, \ldots, T_2$ sont indépendantes et, comme le n-coalescent débute de la partition triviale $(\{1\}, \ldots, \{n\})$, le temps mis par les n lignées initiales pour fusionner en une seule, c'est-à-dire le temps mis pour atteindre le plus récent ancêtre commun, est

$$T^{(n)} = \sum_{b=2}^{n} T_b.$$

Remarquons que

$$\begin{aligned} \mathbb{E}\left[T^{(n)}\right] &= \sum_{b=2}^{n} \frac{2}{b(b-1)} \\ &= 2\left(1 - \frac{1}{n}\right). \end{aligned} \tag{1.24}$$

Soit $m < n$. Si $\mathcal{K}^{(n)}_{|[m]}$ désigne la restriction du n-coalescent aux partitions de $[m]$, nous pouvons remarquer que

$$\mathcal{K}^{(n)}_{|[m]} \overset{d}{=} \mathcal{K}^{(m)}.$$

Cette propriété de compatibilité permet de définir un processus sur $\mathcal{P}$, l'espace des partitions de $\mathbb{N}^*$, dont la restriction à $\mathcal{P}^n$ est de même loi que le n-coalescent. Nous nous reposons pour cela sur le théorème d'extension de Kolmogorov (Théorème 2.2 de Karatzas et Shreve (1991)).

Définition et Théorème 1.1. *(Kingman (1982a)) Le coalescent de Kingman, noté*

$$(\mathcal{K}_t, t \geq 0),$$

est la chaîne de Markov en temps continu sur $\mathcal{P}$ telle que $\mathcal{K}_{|[n]} \overset{d}{=} \mathcal{K}^{(n)}$, $\forall n \geq 1$.

Associons au coalescent de Kingman le processus $(R_t, t \geq 0)$ défini par

$$R_t = |\mathcal{K}_t|, \tag{1.25}$$

le nombre de lignées encore présentes au temps t. Il s'agit d'un processus de mort sur $\mathbb{N}^*$, inhomogène, dont le taux de transition entre n et $n-1$ est $\binom{n}{2}$. En notant q le générateur de R, et puisque ce processus ne peut perdre qu'un bloc à la fois,

$$qf(n) := \sum_p q(n,p) f(p) = \frac{n(n-1)}{2}(f(n-1) - f(n)). \tag{1.26}$$

Puisque la suite $(T^{(n)}, n \geq 2)$ est croissante et que, d'après (1.24), $\mathbb{E}[T^{(n)}] \leq 2$ pour tout $n \geq 2$, alors $T_K = \lim_{n\to\infty} T^{(n)}$ est fini presque sûrement. Le coalescent de Kingman *descend de l'infini*, c'est-à-dire que partant de $\mathcal{K}_0 = (\{1\}, \{2\}, \dots)$ (dont le nombre de blocs est infini), R_t est fini presque sûrement pour tout $t > 0$. Plus précisément, nous pouvons établir la vitesse à laquelle le coalescent de Kingman descend de l'infini. Heuristiquement (voir Berestycki (2009)), nous pouvons considérer que, lorsque t est proche de 0 (i.e. lorsque le nombre de blocs est grand)

$$\frac{dR_t}{dt} \sim -\frac{N_t^2}{2}$$

ce qui mène à

$$R_t \sim \frac{2}{t+c}$$

et comme $R_0 = +\infty$, le résultat suivant apparaît naturellement (c'est un cas particulier d'un résultat plus général de Berestycki *et al.* (2009)) :

Proposition 1.1. *Presque sûrement, $\frac{tR_t}{2} \to 1$ lorsque $t \to 0$.*

D'après (1.24), le temps total de coalescence vaut 2 en espérance. D'autre part, le temps à attendre pour que les deux dernières lignées coalescent est 1 en moyenne. Le coalescent de Kingman est donc très « broussailleux » dans le sens où les branches externes sont très courtes (voir Figure 1.3, utilisée avec l'aimable autorisation de Bob Griffiths). Des résultats sur la taille d'une branche externe choisie au hasard sont développés dans Blum et François (2005) et Caliebe *et al.* (2007). Ils recoupent les heuristiques de Rauch et Bar-Yam (2004).

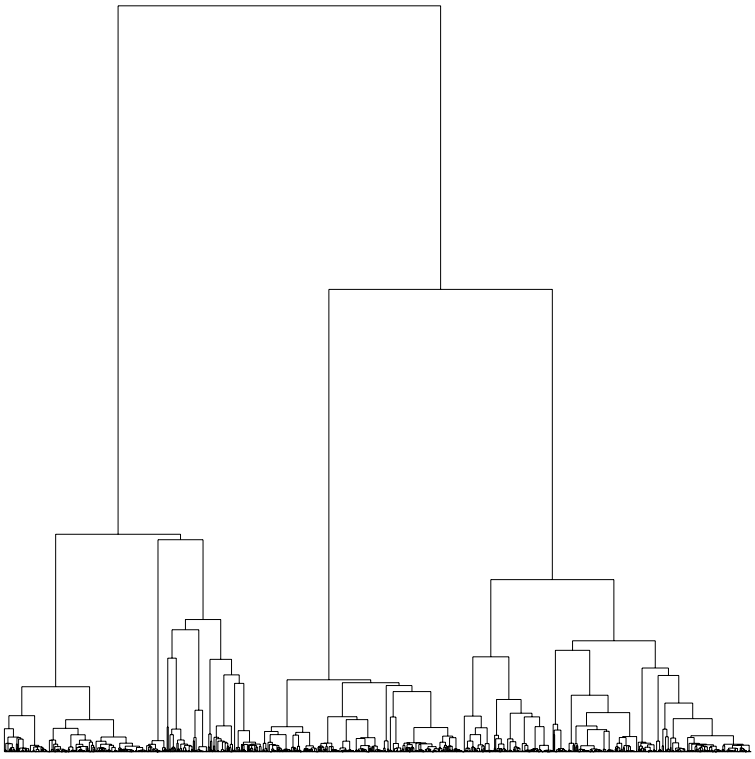

FIGURE 1.3 – Simulation d'un coalescent de Kingman.

Remarque 1.3. *1. Il est possible de considérer d'autres types de populations dont les généalogies limites sont représentées par un coalescent de Kingman. C'est le cas, par exemple, pour des individus diploïdes (qui portent l'information sur une paire de chromosomes), des critères de convergence sont proposés dans Möhle et Sagitov (2003). La valeur par laquelle il faut renormaliser le temps pour obtenir des généalogies limites non-triviales est appelée taille effective. Elle varie selon le modèle utilisé (Nordborg et Krone (2002), Sjödin* et al. *(2005), Barton* et al. *(2007)). Un exemple précis est proposé dans Nordborg et Donnelly (1997). D'autres populations, dont les généalogies sont représentées par des coalescents à collisions multiples, feront l'objet d'une étude approfondie dans le Chapitre 3.*

2. *Supposons que la population, de taille N, est diploïde et hermaphrodite (afin de ne pas avoir à classer les individus en deux groupes). Chaque individu choisit deux parents au hasard. Dans ce cas, plusieurs individus peuvent être ancêtres de toute la population. Chang (1999) montre que le nombre de générations à remonter pour atteindre le plus récent ancêtre commun est de l'ordre de* $\log_2 N$. *De plus, au bout de* $1,77 \log_2 N$ *générations, tous les individus qui sont ancêtres sont en fait ancêtres de toute la population actuelle.*

1.4 Dualité entre le coalescent de Kingman et la diffusion de Wright-Fisher

La diffusion de Wright-Fisher et le coalescent de Kingman permettent de décrire l'évolution d'une population en avançant et en remontant dans le temps. Nous proposons à présent un résultat de dualité intéressant qui permet de relier ces deux objets (ce résultat estutilisé dans la preuve du Théorème 1 de Birkner (2005)). Soit $x \in [0,1]$. Soit $(U_t, t \geq 0)$ une diffusion de Wright-Fisher, définie comme la solution de (1.9), débutant en $U_0 = x$. Si x vaut 0 ou 1, nous nous trouvons dans le cas dégénéré d'un processus constant. Soit $(R_t, t \geq 0)$ le processus du nombre de blocs du coalescent de Kingman, défini en (1.25). Définissons $(R_t^{(n)}, t \geq 0)$ comme le processus du nombre de blocs du n-coalescent :

$$(R_t^{(n)}, t \geq 0) \stackrel{d}{=} (R_{t+s}, t \geq s).$$

$R^{(n)}$ a les mêmes dynaliques que R mais $R_0^{(n)} = n$.

Proposition 1.2. *Quels que soient $x \in [0,1]$ et $n \geq 1$. Pour tout $t \geq 0$, la relation de dualité suivante est vérifiée :*

$$\mathbb{E}[(U_t)^n] = \mathbb{E}[x^{R_t^{(n)}}]. \tag{1.27}$$

Pour vérifier cette relation, notons, pour $n \geq 1$ fixé, $t \geq 0$ et $x \in [0,1]$,

$$u(t,x) = \mathbb{E}[x^{R_t^{(n)}}].$$

La définition (1.26) du générateur de R implique que pour toute fonction $f : \mathbb{N} \to \mathbb{R}_+$ bornée,

$$\mathcal{N}_t^f = f(R_t^{(n)}) - f(R_0^{(n)}) - \int_0^t \frac{R_s^{(n)}(R_s^{(n)} - 1)}{2}(f(R_s^{(n)} - 1) - f(R_s^{(n)}))ds$$

est une martingale d'espérance nulle. En choisissant $f(n) = x^n$, nous obtenons

$$\begin{aligned} x^{R_t^{(n)}} &= x^n + \int_0^t \frac{R_s^{(n)}(R_s^{(n)}-1)}{2}(x^{R_s^{(n)}-1} - x^{R_s^{(n)}})ds + \mathcal{N}_t^f \\ &= x^n + \frac{x(1-x)}{2}\int_0^t R_s^{(n)}(R_s^{(n)}-1)x^{R_s^{(n)}-2}ds + \mathcal{N}_t^f \end{aligned}$$

et nous pouvons en déduire que

$$u(t,x) = u(0,x) + \frac{x(1-x)}{2}\int_0^t \frac{\partial u}{\partial x^2}(s,x)ds.$$

En utilisant la formule d'Itô, on peut vérifier que

$$(\mathcal{M}_s^t, s \leq t) := (u(t-s, U_s), s \leq t)$$

est une martingale. La relation de dualité apparaît alors clairement puisque $\mathbb{E}[\mathcal{M}_0^t] = \mathbb{E}[\mathcal{M}_t^t]$.

1.5 Ajouter des mutations

L'information génétique est contenue dans l'ADN (ou l'ARN) des individus, qui n'est autre qu'une longue chaîne de *nucléotides*. Ces éléments codants sont représentés par quatre lettres A, T, G, C. Si l'information génétique est transmise sans changement, il ne peut y avoir d'évolution dans la population. La variation génétique s'explique par les *mutations* qui interviennent dans la population. Il y a plusieurs types de mutations mais, pour simplifier, nous ne considérerons que des mutations ponctuelles, affectant une base de nucléotide. Les taux de mutation varient suivant les espèces et l'emplacement des bases touchées. Ils sont assez faibles. Les plus élevés sont de l'ordre de 10^{-4} mutations par paire de bases par génération dans l'ARN de certains virus et l'on observe des taux de l'ordre de 10^{-10} chez les humains (nous invitons le lecteur à se reporter au Chapitre 12 de Barton *et al.* (2007) pour plus d'informations).

Pour ajouter des mutations dans un modèle de Cannings à N individus, nous supposons que la probabilité de muter pour chaque individu à chaque génération est constante égale à μ. En remontant une lignée ancestrale, le nombre de générations à attendre avant de voir une mutation est géométrique de paramètre μ. Rappelons que c_N, défini par (1.19), désigne la probabilité que deux individus aient un ancêtre commun à la génération précédente. En supposant que

$$\frac{\mu}{c_N} \to \theta \text{ lorsque } N \to \infty,$$

et en renormalisant le temps de manière à obtenir un coalescent de Kingman, le temps d'attente pour observer une mutation sur chaque lignée est exponentiel de paramètre θ.

Ainsi, en échantillonnant n individus et en passant à la limite, le premier événement (mutation ou coalescence) se produit au bout d'un temps exponentiel : le minimum entre n variables exponentielles de paramètre θ et une autre de paramètre $\binom{n}{2}$, toutes indépendantes.

En fait, l'objet limite peut aussi être obtenu en considérant un processus ponctuel de Poisson d'intensité θdl sur les branches de l'arbre de Kingman.

Remarque 1.4. *Puisque, dans le modèle discret, le nombre de générations à attendre avant que deux lignées choisies au hasard coalescent est, en moyenne, c_N^{-1}, le nombre de mutations qui s'est produit sur ces deux lignées est de l'ordre de $2\mu c_N^{-1}$. Avec les notations que nous avons introduites, la limite de cette dernière valeur est 2θ. Elle est souvent noté θ dans la littérature. Par convention, dans ce document, nous continuerons à noter θ le paramètre du processus de Poisson des mutations.*

Intéressons nous maintenant à la manière de modéliser l'apparition d'une mutation dans une chaîne de nucléotides. Deux modèles introduits dans les années 60 font encore référence aujourd'hui. Nous allons proposer une rapide description qui pourra être complétée par les sections 1.3 et 1.4 de Durrett (2008) (voir aussi la section 1.3.3 de Berestycki *et al.* (2007)).

Le premier est appelé *modèle à infinité d'allèles*, « infinite alleles model ». Il est introduit par Kimura et Crow (1964). Il est justifié par la remarque suivante (Kimura (1971)). Considérons un gène de 500 nucléotides qui forment une séquence d'ADN. Lors d'une mutation, le gène peut atteindre l'une des $3 \times 500 = 1500$ séquences « voisines ». La probabilité pour un allèle de retourner à l'identique après deux mutations est donc $1/1500$ et, comme la probabilité de revenir à l'identique en plus de deux mutations est négligeable devant cette dernière, nous pouvons supposer qu'il y a une infinité d'allèles et que chaque mutation crée un nouvel allèle. Dans ce modèle, si un individu est affecté par une mutation, toute sa descendance est affectée sauf ceux qui seront touchés par une mutation plus tardive, ceux-ci auront un allèle différent. Ce modèle a longtemps primé pour des raisons pratiques. S'il présente une certaine simplicité au niveau expérimental puisqu'il permet de cibler les observations (très coûteuses à l'époque) sur un endroit précis de la chaîne d'ADN, il a un défaut de taille : si deux mutations ont lieu sur une même branche d'un arbre généalogique, seule la seconde est observable (voir Figure 1.4). Le nombre d'allèles différents observés dans une population n'est pas égal au nombre total de mutations dans les lignées ancestrales. Nous pouvons alors séparer la population en groupes d'individus ayant le même allèle. Cette opération nous permet de définir la *partition allélique*. Notons $K^{(n)}$ le nombre de groupes dans un échantillon de taille n et $K^{(k,n)}$ le nombre de blocs dans la partition allélique contenant k éléments. Notons que

$$\sum_{k=1}^{n} kK^{(k,n)} = n.$$

Dans la Figure 1.4, $K^{(8)} = 5$, $K^{(1,8)} = 2$ et $K^{(2,8)} = 3$.

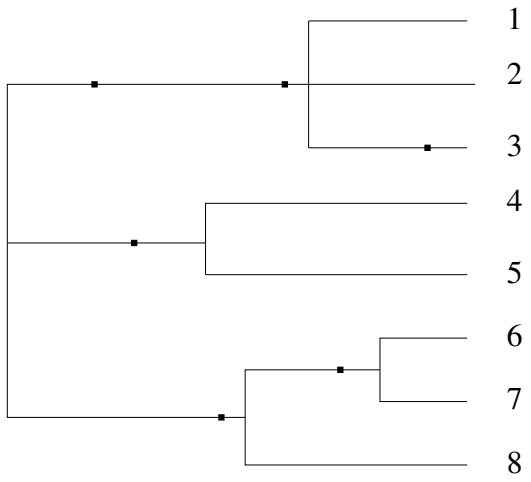

FIGURE 1.4 – Il y a 6 mutations. Dans le modèle à infinité d'allèles, $K^{(8)} = 5$ allèles différents sont portés par chaque groupe d'individus $\{1,2\}$, $\{3\}$, $\{4,5\}$, $\{6,7\}$ et $\{8\}$. Dans le modèle à infinité de sites, $S^{(8)} = 6$ mutations différentes sont portées par chaque groupe d'individus $\{1,2,3\}$, $\{1,2,3\}$, $\{3\}$, $\{4,5\}$, $\{6,7\}$ et $\{6,7,8\}$.

Le résultat le plus célèbre dans ce domaine est la formule d'échantillonnage d'Ewens (1972) (*Ewens' sampling formula*) qui permet d'exprimer la loi de la partition allélique pour une population dont les généalogies peuvent être représentées par le coalescent de Kingman (une preuve reposant sur cette représentation est fournie par Kingman (1982b)). En supposant que, le long des branches du coalescent, les mutations apparaissent à taux θ,

$$\mathbb{P}(K^{(1,n)} = a_1, \ldots, K^{(n,n)} = a_n) = \frac{n!}{(2\theta)_{(n)}} \prod_{i=1}^{n} \frac{\theta^{a_i}}{i^{a_i} a_i!}$$

où $(2\theta)_{(n)} = 2\theta(2\theta+1)\ldots(2\theta+n-1)$.

Ce résultat a un avantage considérable : sachant le nombre d'allèles $K^{(n)}$, la loi de la partition allélique ne dépend pas de θ (le résultat précis peut se trouver dans Ewens (2004)). Pour tester l'adéquation entre les généalogies et le coalescent de Kingman dans un modèle à infinité d'allèles, il n'est donc pas nécessaire d'estimer le taux de mutation θ.

Cette formule nous permet aussi d'obtenir un estimateur de θ. En effet, Watterson (1975) a décrit le comportement asymptotique de $K^{(n)}$ comme suit

$$\mathbb{E}[K^{(n)}] = 2\theta \log n + \mathcal{O}(1)$$

$$\mathrm{Var}(K^{(n)}) = 2\theta \log n + \mathcal{O}(1)$$

lorsque $n \to \infty$ et, puisque $\frac{K^{(n)} - \mathbb{E}[K^{(n)}]}{\sqrt{\mathrm{Var}(K^{(n)})}} \xrightarrow{d} G$ où G est une variable aléatoire gaussienne centrée réduite,

$$\frac{K^{(n)} - 2\theta \log n}{\sqrt{2\theta \log n}} \xrightarrow{d} G.$$

Ce résultat permet d'obtenir un estimateur asymptotiquement normal de θ dans le modèle à infinité d'allèles : l'estimateur de Watterson $K^{(n)}/2 \log n$. En revanche, sa vitesse de convergence logarithmique le rend très peu pratique puisque très lent.

Maintenant que nous sommes capables de séquencer rapidement, et à prix raisonnable, un génome entier (le premier séquençage du génome humain est reporté par Anderson *et al.* (1981)), le second modèle, dit *à infinité de sites*, « infinite sites model », peut s'avérer plus pertinent. Il est introduit par Kimura (1969). Il repose sur le très grand nombre de nucléotides dans la totalité de l'ADN de chaque individu et suppose que chaque mutation apparaît à un nouveau site. Le grand avantage de ce modèle est que le nombre de mutations apparues dans les généalogies d'une population actuelle de n individus correspond au nombre de *sites de ségrégations*, i.e. le nombre de sites où une différence apparaît. Nous noterons $S^{(n)}$ cette valeur. Remarquons que, conditionnellement à la longueur de l'arbre de coalescence $L^{(n)}$, $S^{(n)}$ suit une loi de Poisson de paramètre $\theta L^{(n)}$. Nous appelons aussi $S^{(k,n)}$ le nombre de mutations affectant exactement k individus dans l'échantillon. Dans la Figure 1.4, $S^{(8)}$ vaut 6, $S^{(1,8)} = 1$, $S^{(2,8)} = 2$ et $S^{(3,8)} = 3$. Le vecteur composé des $S^{(k,n)}$ est appelé *spectre des fréquences de sites*.

À l'instar de $K^{(n)}$, des résultats asymptotiques pour $S^{(n)}$ sont proposés par Watterson (1975). Notons

$$h_n = \sum_{k=1}^{n-1} \frac{1}{k},$$

alors

$$\mathbb{E}[S^{(n)}] = 2\theta h_n$$
$$\mathrm{Var}(S^{(n)}) = 2\theta h_n + \mathcal{O}(1)$$

lorsque $n \to \infty$ et

$$\frac{S^{(n)} - 2\theta h_n}{\sqrt{2\theta h_n}} \xrightarrow{d} G. \tag{1.28}$$

L'estimateur de Watterson de θ dans le modèle à infinité de sites est alors $S^{(n)}/2h_n$. Ces derniers résultats seront l'objet d'une explication plus précise dans la Section 3.2.

Remarque 1.5. *Le long de ce document, nous étudions des modèles neutres, c'est-à-dire qu'aucun des types présents dans la population n'est avantagé. Nous avons pourtant parlé dans l'Introduction de l'importance de la sélection naturelle dans l'évolution. Il se trouve que chacun des modèles que nous avons étudié jusqu'à présent ont été modifiés afin d'y*

ajouter de la sélection. Une description du modèle discret et de la diffusion de Wright-Fisher avec sélection se trouve par exemple dans Ewens (2004). Ils ont été établi avant les années 60.

Il faut, en revanche, attendre le XXI^{e} siècle pour trouver une description rigoureuse du coalescent avec sélection (Barton et Etheridge (2004), Barton et al. *(2004)) bien que les premiers pas aient été effectués quinze ans plus tôt (Kaplan* et al. *(1988), Darden* et al. *(1989)). La construction repose sur une structuration de la population en deux groupes, l'un avec le type sélectionné, l'autre sans. Les lignées doivent appartenir à un même groupe pour coalescer et les taux de coalescence dans chacun des groupes dépendent du temps. Il est en outre possible pour une lignée de passer d'un groupe à l'autre par le biais de mutations.*

Bibliographie

Aldous, D. J. (1997). Brownian excursions, critical random graphs and the multiplicative coalescent. *Ann. Probab.*, 25:812–854.

Aldous, D. J. (1999). Deterministic and stochastic models for coalescence (aggregation and coagulation) : a review of the mean-field theory for probabilists. *Bernoulli*, 5(1):3–48.

Aldous, D. J. et Limic, V. (1998). The entrance boundary of the multiplicative coalescent. *Electron. J. Probab.*, 3:1–59.

Aldous, D. J. et Pitman, J. (1998). The standard additive coalescent. *Ann. Probab.*, 26:1703–1726.

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H. L., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. H., Staden, R. et Young, I. G. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, 290:457–465.

Barton, N. H., Briggs, E. G., Eisen, J. A., Goldstein, D. B. et Patel, N. H. (2007). *Evolution*. Cold Spring Harbour Laboratory Press, Cold Spring Harbor, NY.

Barton, N. H. et Etheridge, A. M. (2004). The effect of selection on genealogies. *Genetics*, 166(2):1115–1131.

Barton, N. H., Etheridge, A. M. et Sturm, A. K. (2004). Coalescence in a random background. *Ann. Appl. Probab.*, 14(2):754–785.

Berestycki, J., Berestycki, N. et Limic, V. (2009). The Λ-coalescent speed of coming down from infinity. *http ://arxiv.org/abs/0807.4278. To appear.*

Berestycki, J., Berestycki, N. et Schweinsberg, J. (2007). Beta-coalescents and continuous stable random trees. *Ann. Probab.*, 35(5):1835–1887.

Berestycki, N. (2009). *Recent progress in coalescent theory. www.statslab.cam.ac.uk/ beresty/rp2.pdf. Work in progress.*

Birkner, M. (2005). Stochastic models from population biology. *http : //evol.bio.lmu.de/ birkner/lehre$_a$rchiv/smpb$_S$S07. To appear.*

Blum, M. G. B. et François, O. (2005). Minimal clade size and external branch length under the neutral coalescent. *Adv. in Appl. Probab.*, 37(3):647–662.

Caliebe, A., Neininger, R., Krawczak, M. et Rösler, U. (2007). On the length distribution of external branches in coalescence trees : genetic diversity within species. *Theoret. Population Biol.*, 72(2):245–252.

Cannings, C. (1974). The latent roots of certain Markov chains arising in genetics : a new approach. I. Haploid models. *Adv. in Appl. Probab.*, 6:260–290.

Cannings, C. (1975). The latent roots of certain Markov chains arising in genetics : a new approach. II. Further haploid models. *Adv. in Appl. Probab.*, 7:264–282.

Chang, J. T. (1999). Recent common ancestors of all present-day individuals. *Adv. in Appl. Probab.*, 31(4):1002–1038.

Darden, T., Kaplan, N. L. et Hudson, R. R. (1989). A numerical method for calculating moments of coalescent times in finite populations with selection. *J. Math. Biol.*, 27(3):355–368.

Donnelly, P. et Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. *Ann. Rev. Genet.*, 29:401–421.

Doob, J. L. (2001). *Classical potential theory and its probabilistic counterpart.* Classics in Mathematics. Springer-Verlag, Berlin. Reprint of the 1984 edition.

Durrett, R. (1996). *Stochastic calculus : a practical introduction.* Probability and Stochastics Series. CRC Press, Boca Raton,FL.

Durrett, R. (2008). *Probability models for DNA sequence evolution.* Probability and its Applications. Springer, New York, NY. second edition.

Dynkin, E. B. (2006). *Theory of Markov processes.* Dover Publications Inc., Mineola, NY. Reprint of the 1961 edition.

Etheridge, A. M. (2009). Some mathematical models from population genetics. *In École d'Été de Probabilités de Saint-Flour XXXIX—2009. To appear.*

Ethier, S. N. et Kurtz, T. G. (1986). *Markov processes : characterization and convergence.* Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, NY.

Evans, S. N. et Pitman, J. (1997). Construction of markovian coalescents. *Ann. Inst. H. Poincaré Probab. Stat.*, 34:339–383.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoret. Population Biol.*, 3(1):87–112.

Ewens, W. J. (2004). *Mathematical population genetics. I. Theoretical introduction*, volume 27 de *Interdisciplinary Applied Mathematics*. Springer-Verlag, New York, NY. second edition.

Feller, W. (1951). Two singular diffusion problems. *Ann. Math.*, 54(1):173–182.

Fisher, R. (1930). *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford.

Griffiths, R. C. (1980). Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theoret. Population Biol.*, 17(1):37–50.

Hudson, R. R. (1991). Gene genealogies and the coalescent process. *Oxford Survey Evol. Biol.*, 7:1–44.

Kaplan, N. L., Darden, T. et Hudson, R. R. (1988). The coalescent process in models with selection. *Genetics*, 120(3):819–829.

Karatzas, I. et Shreve, S. E. (1991). *Brownian motion and stochastic calculus*, volume 113 de *Graduate Texts in Mathematics*. Springer-Verlag, New York, NY. second edition.

Karlin, S. et Taylor, H. M. (1981). *A second course in stochastic processes*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, NY.

Kimura, M. (1955a). Random genetic drift in a multi-allelic locus. *Evolution*, 9:419–435.

Kimura, M. (1955b). Solution of a process of random genetic drift with a continuous model. *Proceedings. National Academy of Sciences (United States of America)*, 41:144–150.

Kimura, M. (1955c). Stochastic processes and distribution of gene frequencies under natural selection. *Cold Spring Harbor Symposia on Quantitative Biology*, 20:33–53.

Kimura, M. (1957). Some problems of stochastic processes in genetics. *Ann. Math. Statist.*, 28:882–901.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61:893–903.

Kimura, M. (1971). Theoretical foundations of population genetics at the molecular level. *Theoret. Population Biol.*, 2:174–208.

Kimura, M. et Crow, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics*, 49:725–738.

Kingman, J. F. C. (1982a). The coalescent. *Stochastic Process. Appl.*, 13(3):235–248.

Kingman, J. F. C. (1982b). Exchangeability and the evolution of large populations. *In Exchangeability in probability and statistics (Rome, 1981)*, pages 97–112. North-Holland, Amsterdam.

Kingman, J. F. C. (1982c). On the genealogy of large populations. *J. Appl. Probab.*, 19A:27–43.

Kingman, J. F. C. (2000). Origins of the coalescent 1974–1982. *Genetics*, 156:1461–1463.

Knight, F. B. (1981). *Essentials of Brownian motion and diffusion*, volume 18 de *Mathematical Surveys*. American Mathematical Society, Providence, R.I.

Lambert, A. (2008). Population dynamics and random genealogies. *Stoch. Models*, 24(suppl. 1):45–163.

Li, W. et Fu, Y. (1999). Coalescent theory and its applications in population genetics. *In Statistics in Genetics*, pages 45–79. Springer-Verlag, New York, NY.

Möhle, M. (2000). Total variation distances and rates of convergence for ancestral coalescent processes in exchangeable population models. *Adv. in Appl. Probab.*, 32(4):983–993.

Möhle, M. et Sagitov, S. (2001). A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.*, 29(4):1547–1562.

Möhle, M. et Sagitov, S. (2003). Coalescent patterns in diploid exchangeable population models. *J. Math. Biol.*, 47(4):337–352.

Nordborg, M. (2001). Coalescent theory. *In Handbook of Statistical Genetics*, pages 179–212. John Wiley & Sons Inc., Chichester.

Nordborg, M. et Donnelly, P. (1997). The coalescent process with selfing. *Genetics*, 146(3):1185–1195.

Nordborg, M. et Krone, S. (2002). Separation of time scales and convergence to the coalescent in structured populations. *In Modern Developments in Theoretical Population Genetics : The Legacy of Gustave Malécot*, pages 194–232. Oxford University Press, Oxford.

Rauch, E. et Bar-Yam, Y. (2004). Theory predicts the uneven distribution of genetic diversity within species. *Nature*, 431:449–452.

Revuz, D. et Yor, M. (1999). *Continuous martingales and Brownian motion*, volume 293 de *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin. Third.

Sjödin, P., Kaj, I., Krone, S., Lascoux, M. et Nordborg, M. (2005). On the meaning and existence of an effective population size. *Genetics*, 169:1061–1070.

Stroock, D. W. et Varadhan, S. R. S. (2006). *Multidimensional diffusion processes*. Classics in Mathematics. Springer-Verlag, Berlin. Reprint of the 1997 edition.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoret. Population Biol.*, 7:256–276.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16(2):97–159.

Chapitre 2

Les deux plus anciennes familles d'une population sans sélection

Nous allons dorénavant porter notre attention sur le plus récent ancêtre commun (que nous nommerons MRCA pour *most recent common ancestor*) de populations échangeables, haploïdes et de taille fixe N représentées par des modèles de Cannings (Section 1.1) ou de Moran (celui-ci sera introduit dans la Section 2.1). Ce plus récent ancêtre commun n'est autre que la profondeur de l'arbre généalogique des individus actuellement en vie. Rappelons que pour un modèle de Cannings, les généalogies limites obtenues sont représentées par un coalescent de Kingman (Section 1.3) lorsque les hypothèses (1.20) et (1.22) sont vérifiées et que l'on a renormalisé le temps en le multipliant par c_N^{-1} (défini en (1.19)). Dans ce cas, chaque couple de lignées fusionne à taux 1. Le MRCA est donc atteint au bout d'un temps

$$T_K = \sum_{k\geq 1} \frac{2}{k(k+1)} E_k \tag{2.1}$$

où $(E_k, k \geq 1)$ est une suite de variables aléatoires indépendantes exponentielles de paramètre 1. L'espérance de T_K vaut 2 (comme nous l'avons vu avec (1.24)). Nous en déduisons donc que, dans un modèle de Cannings vérifiant (1.20) et (1.22), le MRCA est atteint , en moyenne, au bout de $2c_N^{-1}$ générations. L'exemple le plus célèbre est le modèle de Wright-Fisher où le MRCA vit, en espérance, $2N$ générations auparavant.

Si l'on considère à présent des populations évoluant dans le temps, une autre valeur intéressante est D_t, le temps nécessaire pour atteindre le MRCA de tous les individus vivant au temps t. Il est possible, afin d'étudier le processus $(D_t, t \geq 0)$, d'introduire un processus à valeurs arbres comme l'ont fait Greven *et al.* (2009). De récents articles proposent une étude complète du processus de l'âge du MRCA $(A_t, t \geq 0)$ défini par $A_t = t - D_t$. Nous renvoyons le lecteur à Pfaffelhuber et Wakolbinger (2006) et Simon et Derrida (2006) lorsque les généalogies sous-jacentes sont des coalescents de Kingman ou à Evans et Ralph (2008) dans un cas particulier de grandes populations branchantes. Le processus A évolue linéairement en temps jusqu'à ce que le MRCA change. À ce moment s'opère un saut négatif (voir Figure 2.1). Pour $t_0 \geq 0$, nous noterons τ_{t_0} le premier temps de saut de A_t après le temps t_0 :

$$\tau_{t_0} = \inf\{t \geq t_0, A_t \neq A_{t_0} + (t - t_0)\}. \tag{2.2}$$

Si un modèle de Cannings vérifie les hypothèses (1.20) et (1.22), ses généalogies limites sont représentées par un coalescent de Kingman. C'est aussi le cas d'autres modèles, correctement renormalisés (voir la Remarque1.3.1). Toute la population ne descend alors que des deux dernières lignées coalescentes. Il est donc possible de regrouper tous les individus en *deux plus anciennes familles*. Soit $\mathcal{K}^{\downarrow t_0} := (\mathcal{K}_t^{\downarrow t_0}, t \geq 0)$ l'arbre de coalescence retraçant les généalogies de la population vivant au temps t_0. En reprenant les notations de la Section 1.3, ces deux familles sont les classes d'équivalence de

$$\mathcal{K}^{\downarrow t_0}_{D_{t_0}-} \stackrel{d}{=} \mathcal{K}_{T_K-}.$$

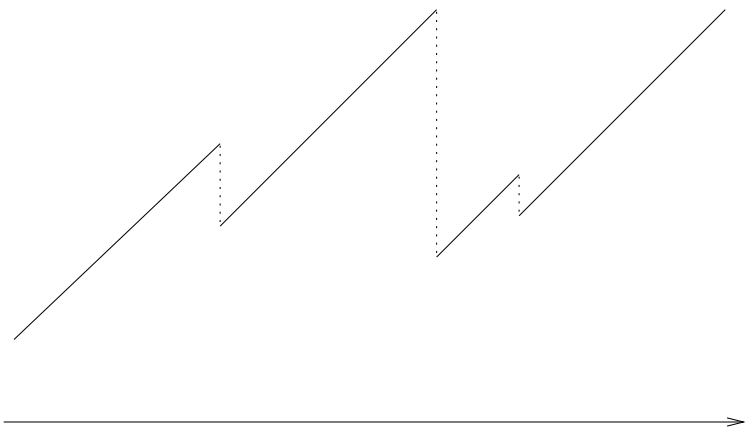

FIGURE 2.1 – Le processus de l'âge du MRCA. Les sauts négatifs correspondent à des changements de MRCA dans la population.

Faisons évoluer les généalogies dans le temps. Les temps de saut du processus d'âge sont les instants où l'une des deux familles ancestrales disparaît. En d'autres termes, si X_t et $1 - X_t$ représentent les proportions de ces deux lignées à un temps t quelconque, alors

$$\tau_{t_0} = \inf\{t \geq t_0, X_t \in \{0, 1\}\}. \tag{2.3}$$

Au temps τ_{t_0}, un nouveau MRCA est établi et la population se décompose en deux nouvelles plus anciennes familles, issues des deux lignées descendant du nouveau MRCA. À cet instant, et à l'instar de A, le processus $(X_t, t \geq 0)$ saute pour arriver à la proportion de l'une des deux nouvelles plus anciennes familles, choisie au hasard. L'étude préliminaire menée dans la Section 1.2 montre que, entre les sauts, le processus $(X_t, t \geq 0)$ est une diffusion de Wright-Fisher, vérifiant (1.9), dont nous avons étudié quelques caractéristiques telles que le temps de fixation (1.15) et la probabilité de toucher l'une ou l'autre des bornes 0 et 1 (1.14).

Avant d'aller plus loin dans notre étude des deux plus anciennes familles, introduisons plus précisément les modèles de Moran (1958) et de Fleming-Viot (1978, 1979) cités plus haut et l'outil qui nous permettra de formaliser notre problème : la construction look-down de Donnelly et Kurtz (1996, 1999).

2.1 Le modèle de Moran

Le modèle de Moran (1958) est un grand classique de la génétique des populations. Il est décrit dans de nombreux ouvrages, voir par exemple la Section 1.2.2 de Lambert (2008) ou la Section 1.5 de Durrett (2008). Comme nous allons le voir, il ressemble fort à

un cas particulier du modèle de Cannings puisqu'il conserve les propriétés d'échangeabilité et d'indépendance des lois de reproduction successives. En revanche, dans ce modèle, les événements de reproduction ne se déroulent plus à temps fixe mais à des temps aléatoires exponentiels. Il est défini comme suit :

Considérons à nouveau une population haploïde, de taille fixe N. À taux $\binom{N}{2}$ a lieu un événement de reproduction. À chacun de ces événements, parmi les N individus vivants,

- un individu est choisi uniformément et donne naissance à un enfant.
- un individu est choisi uniformément (ce peut être le même que celui qui se reproduit) et meurt.

Notons qu'il n'existe pas de convention pour le choix du taux de reproduction. Nous avons choisi $\binom{N}{2}$ afin que les généalogies d'une population à un temps fixé soient directement représentées par un coalescent de Kingman.

Si l'on considère qu'il y a deux groupes (deux types, deux familles) dans la population, et si U_t^N désigne la proportion de l'un des deux groupes au temps t, alors les transitions de U_t^N sont les suivantes :

$$\frac{i}{N} \to \frac{i+1}{N} \text{ à taux } \frac{(N-1)i(N-i)}{2N}$$

$$\frac{i}{N} \to \frac{i-1}{N} \text{ à taux } \frac{(N-1)i(N-i)}{2N}.$$

Notons

$$x = \frac{i}{N}.$$

Nous pouvons calculer, grâce aux taux de transition ci-dessus, la dérive et la variance infinitésimales de U_t^N :

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}_x[U_t^N - x] &= \frac{1}{N}\frac{(N-1)i(N-i)}{2N} - \frac{1}{N}\frac{(N-1)i(N-i)}{2N} = 0 \\
\frac{d}{dt}\mathbb{E}_x[(U_t^N - x)^2] &= \frac{1}{N^2}\frac{(N-1)i(N-i)}{2N} + \frac{1}{N^2}\frac{(N-1)i(N-i)}{2N} = \frac{2}{N^2}\frac{(N-1)i(N-i)}{2N} \\
&= x(1-x) + \mathcal{O}\left(\frac{1}{N}\right).
\end{aligned}$$

Il suffit alors d'utiliser les résultats développés dans la section 7.4 de Ethier et Kurtz (1986) ou dans le Théorème 8.7.1 de Durrett (1996) pour obtenir que

Théorème 2.1. *Si U_0^N converge en loi vers U_0, alors $(U_t^N, t \geq 0)$ converge étroitement vers $(U_t, t \geq 0)$, la diffusion de Wright-Fisher, dans l'espace des processus càdlàg $\mathcal{D}(\mathbb{R}_+, [0,1])$ muni de la topologie de Skorohod.*

2.2 Le processus de Fleming-Viot

Dans cette section, nous allons présenter le processus de Fleming-Viot (Fleming et Viot (1978, 1979)) qui apparaît comme le processus dual du coalescent de Kingman. Il appartient à une plus grande classe de processus : les *superprocessus* (ou processus à valeurs mesures, Dawson (1993)).

Nous donnons ici une définition rapide et les liens qui existent entre le processus de Fleming-Viot et celui de Kingman. Dans la Section 2.3 nous proposerons une construction plus récente, la *construction look-down* de Donnelly et Kurtz (1999), qui nous sera très utile par la suite. Pour une introduction plus exhaustive au processus de Fleming-Viot, nous renvoyons le lecteur à Etheridge (2000).

Soient E un espace métrique séparable complet et $\mathcal{M}^1(E)$ l'espace des probabilités sur E. Considérons des fonctionnelles de la forme

$$\begin{array}{rccl} G: & \mathcal{M}^1(E) & \to & \mathbb{R} \\ & \mu & \mapsto & G(\mu) = \int_{E^p} \mu(da_1)\dots\mu(da_p) f(a_1,\dots,a_p) \end{array} \tag{2.4}$$

où $p \in \mathbb{N}$ et $f : E^p \to \mathbb{R}$ est mesurable bornée. Pour tout vecteur $a = (a_1,\dots,a_p) \in E^p$ et $J \subseteq \{1,\dots,p\}$, notons

$$a_i^J = \begin{cases} a_{\min J} & \text{si } i \in J \\ a_i & \text{si } i \notin J \end{cases} \tag{2.5}$$

Définition 2.1. *On appelle processus de Fleming-Viot le processus $(\rho_t, t \geq 0)$ à valeurs mesures sur E dont le générateur infinitésimal est défini par*

$$\mathcal{T}^K G(\mu) = \sum_{J\subseteq\{1,\dots,p\},|J|=2} \int_{E^p} \mu(da_1)\dots\mu(da_p)(f(a_1^J,\dots,a_p^J) - f(a_1,\dots,a_p)). \tag{2.6}$$

L'exposant K est pour « Kingman ». Il existe en effet une relation de dualité entre le processus de Fleming-Viot et le coalescent de Kingman, Nous nous contentons d'énoncer ce résultat. Une preuve complète se trouve dans Dawson et Hochberg (1982) (voir aussi la Section 1.12 de Etheridge (2000)).

Théorème 2.2. *Pour tout $p \in \mathbb{N}$, toute fonction $f : E^p \to \mathbb{R}$ mesurable bornée et tout $t \geq 0$,*

$$\mathbb{E}\left[\int \rho_t(da_1)\dots\rho_t(da_p) f(a_1,\dots,a_p)\right] = \mathbb{E}\left[\int db_1\dots db_{|\mathcal{K}_t^p|} f_{\mathcal{K}_t^p}(b_1,\dots,b_{|\mathcal{K}_t^p|})\right] \tag{2.7}$$

où $\mathcal{K}^p$ est un p-coalescent et, pour toute partition $\pi = (C_1,\dots,C_q)$ de $\{1,\dots,p\}$

$$f_\pi(b_1,\dots,b_q) := f(a_1,\dots,a_p)$$

avec $a_i = b_k$ si $i \in C_k$.

En considérant a dans (2.6) comme un échantillon tiré suivant μ, le passage de a à a^J n'est autre que la coalescence des deux lignées $a_i, i \in J$, lorsque l'on remonte le temps. Ces coalescences ont lieu à taux 1 (d'après les propriétés du coalescent de Kingman).

Pour mieux se représenter le processus de Fleming-Viot, exprimons-le comme limite d'une suite de processus liés au modèle de Moran (Etheridge (2000)). Dans ce dernier, chaque individu de la population est supposé avoir un type (un type génétique par exemple) qui est un élément de E. Il est tout à fait possible pour nos individus de se déplacer dans E au cours du temps (ce qui reviendrait à autoriser des mutations dans notre modèle) mais nous ne nous intéressons ici qu'au modèle neutre. En représentant chaque individu par un *atome*, nous pouvons lui associer une masse. Dans notre cas, chaque individu a une masse 1/N, de manière à ce que le système ait une masse totale de 1. Nous pouvons alors nous représenter la population totale au temps t comme une mesure ponctuelle de probabilité ρ_t^N sur E. Pour tout $z \in E$, $\rho_t^N(dz)$ désigne la proportion de la population ayant le type z au temps t dans le modèle de Moran à N éléments. Le processus $(\rho_t, t \geq 0)$ apparaît comme la limite étroite, dans l'espace des mesures de probabilité sur E, de $(\rho_t^N, t \geq 0)$ lorsque N tend vers l'infini.

Nous voyons bien les généalogies de la population en filigrane dans cette description. Une construction fondée sur cette intuition est donc notre prochaine étape.

2.3 Une construction dénombrable : le processus du look-down

Afin de fournir une représentation dénombrable du processus de Fleming-Viot (et de bien d'autres dont le classique processus de Dawson-Watanabe (Watanabe (1968), Dawson (1993), Etheridge (2000), Perkins (2002)), ce qui revient à donner une solution au problème de martingale associé sous la forme d'un système dénombrable d'équations différentielles stochastiques, Donnelly et Kurtz (1996) ont introduit le processus du *look-down* (se reporter aussi au Chapitre 5 de Etheridge (2000)). Cette représentation a le grand avantage de fournir une description explicite de la généalogie de la population. Dans un second article, Donnelly et Kurtz (1999) proposent une construction du look-down *par persistance* qui prend en compte le temps de survie de la *ligne de descendance* de chaque individu et les classe en fonction. C'est ce processus du look-down *modifié* que nous utiliserons par la suite. Cependant, il est nécessaire de donner une description rapide de la première version (1996) afin de mieux comprendre la version modifiée (1999).

E désigne toujours un espace métrique séparable complet, que nous avons précédemment supposé être l'espace des types génétiques de la population. Soit $N \in \mathbb{N}$. Nous définissons un processus de Markov

$$(\xi_t^1, \ldots, \xi_t^N, t \geq 0),$$

à valeurs dans E^N, décrivant l'évolution des types d'une population de taille N de la façon

suivante :

- un déplacement suivant un processus markovien dans E (qui peut être vu comme des mutations neutres des individus dans le temps)
- un mécanisme de reproduction : les naissances arrivent à des temps exponentiels de paramètre $\binom{N}{2}$. Le parent est choisi au hasard dans la population, le nouveau-né a le même type que son parent et apparaît à un niveau choisi au hasard. L'individu qui se trouvait à ce niveau meurt.

Il s'agit d'un modèle de Moran avec déplacement spatial.

Précisons que le mécanisme de reproduction peut être compris de différentes manières. D'aucuns pourraient considérer qu'à un instant de naissance le parent, plutôt que de mettre au monde un individu et de continuer à vivre, pourrait disparaître et donner naissance à deux nouvelles entités (comme une bactérie qui se diviserait). Afin de lever l'ambiguïté, si lors d'un événement de reproduction, le parent se trouve au niveau i, nous confondrons les deux particules se trouvant au niveau i avant et après l'instant de reproduction. Ainsi, lorsque la mort d'un individu survient (remplacé par un autre qui naît à son niveau sans qu'il ait participé à l'événement de reproduction) nous parlerons plutôt de la disparition d'une *ligne de descendance*.

Il est très important de noter que nous sommes dans le cas d'un modèle neutre, aucun individu n'est avantagé. Toute la construction qui suit est en effet intimement liée au caractère échangeable du vecteur $(\xi_t^1, \ldots, \xi_t^N)$. Des constructions look-down adaptées à des modèles avec sélection ont été récemment introduites (voir Leocard (2009)). Dans notre cas, l'étiquetage est donc arbitraire et une permutation des indices ne change pas la distribution du vecteur. Introduisons la mesure

$$\zeta_t = \sum_{i=1}^{N} \delta_{\xi_t^i}.$$

À présent modifions la construction. L'idée est la suivante : regardons dans le futur et réordonnons les individus selon le temps de survie de leur ligne de descendance. Plus un individu est à un niveau élevé, plus sa ligne de descendance disparaîtra vite. À chaque événement de reproduction, le nouveau-né est inséré à un niveau choisi uniformément et cette naissance entraîne la relabélisation de tous les individus des niveaux supérieurs : si la naissance a lieu au niveau j, l'individu qui se trouvait au niveau j saute au niveau $j+1$, celui en $j+1$ saute en $j+2$ et ainsi de suite... La Figure 4.1 représente un événement de reproduction. Ceci permet d'introduire et de conserver un classement par persistance dans la population et, par conséquent, l'individu qui meurt est celui qui se trouvait au niveau le plus élevé à l'instant de la reproduction. Nous considérerons de plus que le parent est toujours à un niveau inférieur à celui de l'enfant. Ainsi si les deux niveaux impliqués dans la reproduction sont i et j, $i < j$, le nouveau-né sera donc placé en j et prendra le type de l'individu en i. Suivant la terminologie de Donnelly et Kurtz, nous dirons que j *regarde* i (voir Figure 4.1) .

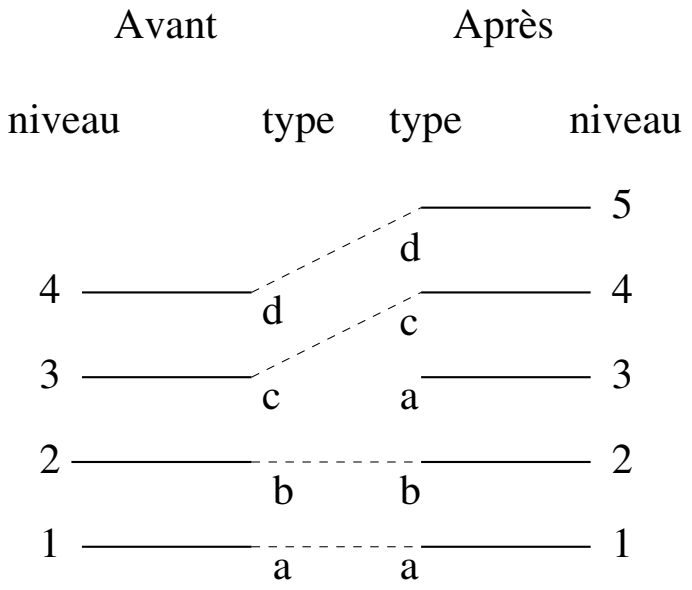

FIGURE 2.2 – Un événement de look-down entre les niveaux 1 et 3. Chaque individu vivant à un niveau au moins égal à 3 avant la naissance est décalé au niveau supérieur.

Ici la neutralité prend toute son importance : conditionnellement à toute l'information donnée par ζ jusqu'au temps t, chaque particule a la même chance d'avoir la plus longue ligne de descendance, la seconde plus longue ligne de descendance... Ce réordonnement n'est donc qu'une permutation aléatoire du système original et la mesure empirique est inchangée. Notons $(\tilde{\xi}_t^1, \dots, \tilde{\xi}_t^N)$ le processus réétiqueté. Le résultat suivant est tiré de Donnelly et Kurtz (1999).

Théorème 2.3. *Supposons que $(\tilde{\xi}_0^1, \dots, \tilde{\xi}_0^N)$ est échangeable. Définissons*

$$\tilde{\zeta}_t = \sum_{i=1}^{N} \delta_{\tilde{\xi}_t^i}.$$

Alors $\tilde{\zeta} \stackrel{d}{=} \zeta$ et le vecteur $(\tilde{\xi}_t^1, \dots, \tilde{\xi}_t^N)$ est échangeable.

Il est tout à fait possible, par des arguments de consistance, de construire ce processus sur $E^{\mathbb{N}}$. Nous obtenons alors une suite $(\tilde{\xi}_t^1, \tilde{\xi}_t^2, \dots)$ d'individus répartis sur une infinité de niveaux. Nous allons décrire le processus du look-down de manière plus rigoureuse puisque dans ce cas les événements de reproduction arrivent à des taux infinis. Par souci de simplicité, et puisque nous ne nous intéressons qu'aux généalogies des populations, on oublie le déplacement dans E.

Considérons donc une population infinie et soit $\mathcal{E} = \mathbb{R}_+ \times \mathbb{N}^*$. Chaque point (s, i) de $\mathcal{E}$ représente l'individu (unique) vivant au niveau i au temps s. Pour chaque couple de niveaux (i, j) avec $i < j$, notons $(B_{ij}(t), t \geq 0)$ des processus de Poisson indépendants de paramètre 1. Soit

$$\mathcal{B}_{ij} := \{t, B_{ij}(t) - B_{ij}(t-) = 1\}$$

l'ensemble des temps de saut de B_{ij}. Soit

$$s_0 \in \cup_{i<j} \mathcal{B}_{i,j}.$$

À l'instant s_0, j regarde à un niveau inférieur et un individu naît au niveau j, poussant tous les individus qui vivaient en s_0- à un niveau supérieur ou égal à j. Ce nouveau-né est représenté au cours du temps par sa ligne de descendance (Figure 2.3)

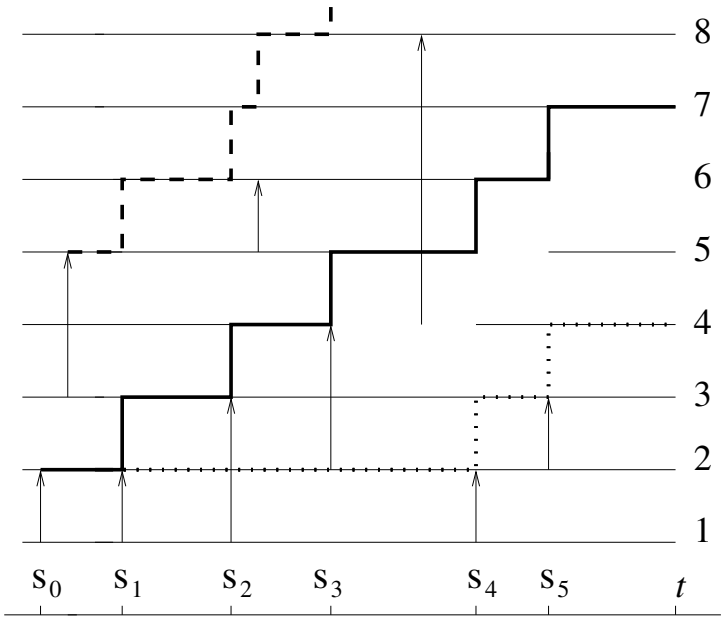

FIGURE 2.3 – Les trois ligne (grasse, hachurée et pointillée) représentent trois individus depuis leur naissance. La ligne grasse naît en $(s_0, 2)$. Les s_k, $k \geq 1$ sont les temps de saut de cette ligne.
Il y a deux courbes de fixation vivant au temps t. $Z_t = 1$, $L_0(t) = 7$ et $L_1(t) = 4$

$$G := ([s_0, s_1) \times \{j\}) \cup ([s_1, s_2) \times \{j+1\}) \cup ([s_2, s_3) \times \{j+2\}) \cup ...$$

Les $(s_k, k \geq 1)$ sont les temps où la ligne saute, en d'autres termes,

$$s_{k+1} := \inf\{s > s_k, s \in \cup_{i<j+k} \mathcal{B}ij\}.$$

Le processus du look-down $((\xi_t^k, \xi_t^k, \dots), t \geq 0)$ évolue donc comme suit : si $t \in \mathcal{B}_{ij}$,

$$\tilde{\xi}_t^k = \begin{cases} \tilde{\xi}_{t-}^k & \text{pour } k \leq j \\ \tilde{\xi}_{t-}^i & \text{pour } k = j \\ \tilde{\xi}_{t-}^{k-1} & \text{si } k > j \end{cases}$$

G décrit en fait les différents niveaux occupés par l'individu né en (s_0, j). Nous pouvons donc identifier un individu à cette ligne. Notons

$$b_G = s_0$$

le temps de naissance de l'individu G et

$$d_G = \lim_{k\to\infty} s_k$$

l'instant de disparition de sa ligne de descendance. Une particule au niveau j est poussée à taux $\binom{j}{2}$ (c'est le nombre de couples (i, i'), $1 \leq i < i' \leq j$). Ce taux est quadratique en j donc la disparition de la lignée d'un individu a lieu presque sûrement en un temps fini, à moins que cet individu ne vive dans la *ligne immortelle*

$$\iota = \mathbb{R}_+ \times \{1\}.$$

Si la construction look-down permet de représenter le processus de Fleming-Viot, alors le processus de coalescence induit est celui de Kingman. Afin de s'en convaincre, fixons t *geq*0 et supposons que le processus du look-down est défini sur $(-\infty, t)$. Pour $u \geq 0$ introduisons la relation d'équivalence suivante : $i \overset{u}{\sim} j$ si et seulement si les individus (t, i) et (t, j) ont un ancêtre commun au temps $t - u$. Notons les classes d'équivalence correspondantes $\mathcal{K}_u^{(t)}$ et remarquons que si $u \in \mathcal{B}_{ij}$, l'ancêtre au temps $u-$ de (u, j) et de tous les individus de la ligne contenant (u, j) est $(u-, i)$.

Pour prouver que $(\mathcal{K}_u^{(t)}, u \geq 0)$ est un coalescent de Kingman, il suffit de vérifier que la restriction de $\mathcal{K}^{(t)}$ à $[n] = \{1, \dots, n\}$ est un n-coalescent (grâce à la propriété de consistance du coalescent). Or les transitions de $\mathcal{K}^{(t)}$ ont lieu lorsque l'ancêtre d'une classe d'équivalence regarde le niveau de l'ancêtre d'une autre classe d'équivalence, ce qui arrive à taux 1 (voir Figure 4.2).

2.4 L'étude des deux plus anciennes familles

Dans la construction look-down, certaines lignes jouent un rôle particulier. Nous dirons que G est une *courbe de fixation* si $(b_G, 2) \in G$: l'individu correspondant est né au niveau 2, provenant d'un événement de look-down entre 2 et 1.

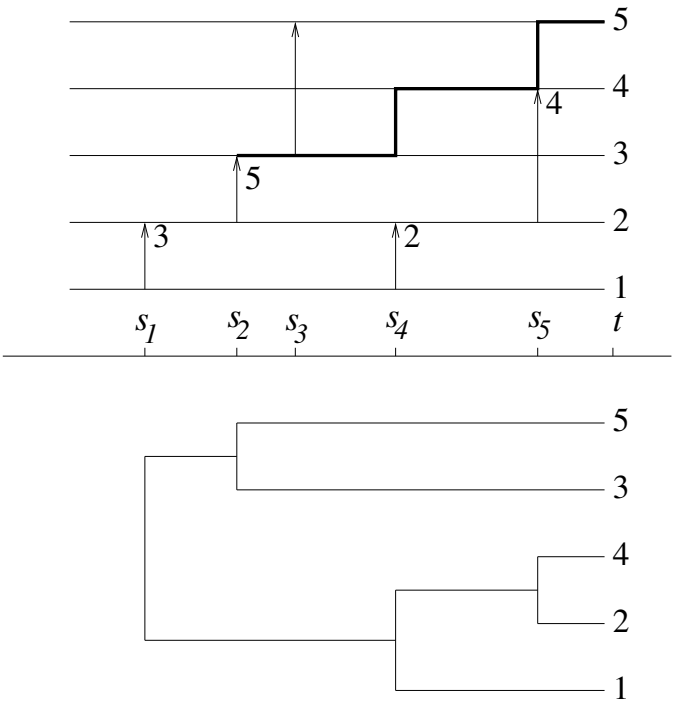

FIGURE 2.4 – Le processus du look-down et son arbre de coalescence associé, débuté au temps t pour les 5 premiers niveaux. À chaque événement de look-down se crée une nouvelle ligne de descendance. Nous indiquons à quel niveau se trouve cette ligne au temps t. En gras, la ligne contenant $(t, 5)$.

À un temps fixé t, notons $\mathcal{G}_t$ l'ensemble des lignes vivantes au temps t et soit G_t la lignée du MRCA de la population vivant au temps t. Notons que la naissance du MRCA est

$$A_t = \inf\{b_G, G \in \mathcal{G}_t\} = b_{G_t}.$$

Elle correspond au temps de naissance de la plus haute courbe de fixation vivant au temps t. La variable aléatoire $Z_t + 1$ désigne le nombre de courbes de fixation vivant au temps t. En d'autres termes, $Z_t \geq 0$ est le nombre de futurs MRCAs vivant déjà au temps t. Notons $L_0(t) > L_1(t) \cdots > L_{Z_t}(t)$ les niveaux décroissants des courbes de fixation vivant au temps t (voir Figure 2.3). La loi jointe de $(Z_t, L_0(t), L_1(t), \ldots, L_{Z_t}(t))$ est proposée par Pfaffelhuber et Wakolbinger (2006) (Théorème 2) et la loi de Z_t se trouve dans leur Théorème 3. Répartissons la population en deux plus anciennes familles grâce à

la relation d'équivalence $\mathcal{K}^{\downarrow t}_{D_t-}$. Ceci revient à classer les individus vivant au temps t selon que leur ancêtre est G_t ou la ligne immortelle. Nous appellerons Y_t la proportion de la sous-population dont l'ancêtre au temps A_t est la lignée immortelle. Cette sous-population est la plus ancienne famille contenant l'individu immortel. Soit X_t la proportion de l'une des deux plus anciennes familles choisie au hasard, c'est-à-dire que X_t vaut Y_t avec probabilité $1/2$ et $1 - Y_t$ avec probabilité $1/2$.

Nous déduisons de la stationnarité que la loi de

$$H_t := (X_t, Y_t, Z_t, L_0(t), L_1(t), \ldots, L_{Z_t}(t))$$

ne dépend pas de t. Entre deux morts de MRCA, le processus $(X_t, t \geq 0)$ est une diffusion de Wright-Fisher, solution de (1.9), le processus $(Y_t, t \geq 0)$ est une diffusion de Wright-Fisher conditionnée à toucher 1, solution de (1.18). Il est à noter que la loi conditionnelle de Z_t sachant X_t est tout à fait intéressante puisque la proportion des deux plus anciennes familles peut être estimée par l'analyse de l'ADN si les taux de mutation (neutre) sont assez forts.

Nous nous intéressons à la loi de H_t au moment (aléatoire) où le MRCA change, ainsi qu'à la loi des niveaux des individus d'une même plus ancienne famille. La distribution de H_t est la même suivant que l'on considère un temps fixé t ou ce temps aléatoire. L'argument est le même que celui utilisé dans la preuve du Théorème 2 de Pfaffelhuber et Wakolbinger (2006). Nous utilisons pour cela le fait que, comme les temps de naissance des MRCAs, les temps de changement de MRCA sont les temps de sauts d'un processus de Poisson sur $\mathbb{R}_+$. Il s'agit de la propriété PASTA (Poisson Arrivals See Time Average), nous invitons le lecteur à se reporter à Brémaud *et al.* (1992) pour plus d'informations sur le sujet. Pour cette raison, nous pouvons omettre l'indice t, écrire H pour H_t et proposer des preuves au moment de la mort d'un MRCA. Nos résultats donnent en particulier des preuves détaillées des arguments heuristiques des Remarques 3.2 et 7.3 de Pfaffelhuber et Wakolbinger (2006).

Considérons donc des populations (infinies) dont les généalogies sont représentées par un coalescent de Kingman et regardons-les au temps τ d'un changement de MRCA, c'est-à-dire au moment où une courbe de fixation atteint l'infini. Il y a alors deux nouvelles plus anciennes familles dont une qui se fixera dans la population. Si nous remontons à l'instant où il n'y a plus que N ancêtres, qui est aussi l'instant où la courbe de fixation ayant atteint l'infini en τ sautait au niveau $N + 1$, nous pouvons nous demander combien de ces N individus font partie de la famille qui se fixera plus tard. Cette quantité est notée V_N (Figure 2.5).

Nous prouvons dans le Théorème 4.2 que la loi de $(1 + V_{N+2}, N \in \mathbb{N})$ peut être vue comme le nombre de boules noires tirées dans une urne de Pólya initiée avec deux boules noires et une blanche. En particulier, le processus $(V_N, N \geq 2)$ est markovien et pour tout entier k entre 1 et $N - 1$,

$$\mathbb{P}(V_N = k) = \frac{2k}{N(N-1)}.$$

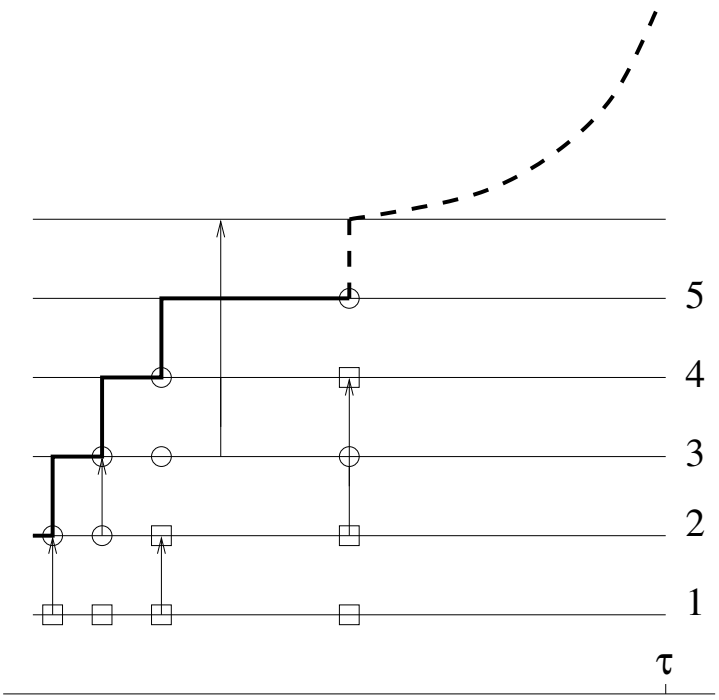

FIGURE 2.5 – Le temps aléatoire τ est un temps de changement de MRCA. Si une courbe de fixation se trouve au niveau $N+1$, V_N désigne le nombre de descendants de l'individu immortel sous la courbe. $V_2 = 1$, $V_3 = 1$, $V_4 = 2$ et $V_5 = 3$

La représentation look-down est adaptée à l'étude de ce processus. Elle nous permettra de prouver, toujours dans le Théorème 4.2, que conditionnellement à V_N les particules de chacune des deux plus anciennes familles sont uniformément réparties entre les niveaux 1 et N.

Une première application de ces résultats est l'obtention de la loi jointe de V_N et du niveau de la courbe de fixation L_1 (celle qui deviendra L_0 au moment du changement de MRCA). Ce résultat est proposé dans la Proposition 4.2. En faisant tendre N vers $+\infty$, nous retrouvons la loi de $L_0(\tau)$ énoncée dans Pfaffelhuber et Wakolbinger (2006).

Par la suite, nous énumérerons plusieurs résultats dont la loi du temps à attendre avant d'assister à un changement de MRCA sachant Y et $L = L_0$ ou la loi de Z sachant X (Théorème 4.1). Une relation intéressante entre τ et Z est ensuite énoncée dans la Proposition 4.1 : pour tout $\lambda \geq 0$,

$$\mathbb{E}\left[\mathrm{e}^{-\lambda\tau}\,|X\right] = \mathbb{E}\left[\mathrm{e}^{-\lambda T_K}\right]\mathbb{E}\left[(1+\lambda)^Z|X\right]. \tag{2.8}$$

Enfin, nous nous intéresserons à l'étude du processus $(X_t, t \geq 0)$. Nous avons vu qu'entre deux changements de MRCA, la fréquence d'une famille choisie au hasard se

comporte comme une diffusion de Wright-Fisher standard $(U_t, t \geq 0)$. Mais lorsque X_t touche 0 ou 1, il y a deux nouvelles lignées ancestrales, X est donc ressuscitée, c'est-à-dire que le processus saute suivant une loi μ. Deux questions naturelles se posent alors : μ est-elle une mesure stationnaire du processus ressuscité ? Et si c'est le cas, est-elle la seule ? Ferrari *et al.* (1995) ou Collet *et al.* (2000) ont mis en lumière que ces questions sont liées aux mesures quasi-stationnaires du processus tué lorsqu'il atteint ses bornes. Ce lien peut aussi être utilisé pour simuler des distributions quasi-stationnaires (Villemonais (2009)). Une mesure ν sur $[0, 1]$ est quasi-stationnaire pour le processus X si

$$\mathbb{P}_\nu(X_t \in dx | X_t \notin \{0, 1\}) = \nu(dx).$$

Ferrari *et al.* (1995) et Collet *et al.* (2000) montrent que la mesure de résurrection μ est une loi stationnaire de $(X_t, t \geq 0)$ si et seulement si μ est une loi quasi-stationnaire du processus tué.

En utilisant les propriétés des urnes de Pólya (voir Johnson et Kotz (1977)) et en regardant la fréquence limite d'une couleur, nous pouvons montrer que, à tout temps fixé, la diffusion suit une loi uniforme sur $(0, 1)$. Par ailleurs, l'uniformité asymptotique de la probabilité conditionnelle

$$\mathbb{P}(U_t \in dy | U_t \neq 0, 1) \sim dy \quad \text{lorsque } t \to \infty$$

est déjà montré dans Ewens (2004) et dans Huillet (2007). Une introduction aux notions de quasi-stationnarité est proposée dans le Chapitre 3 de Lambert (2008) ainsi que de nombreux exemples d'applications à la génétique des populations. La quasi-stationnarité d'autres diffusions appliquées aux dynamiques des populations est étudiée dans Cattiaux *et al.* (2009).

Dans la Section 4.3, nous adapterons ces résultats au processus $(Y_t, t \geq 0)$. Nous proposerons une condition nécessaire et suffisante sur la loi de résurrection μ afin qu'elle soit aussi la loi stationnaire du processus ressuscité construit à partir de la diffusion de Wright-Fisher conditionnée à toucher 1.

Remarque 2.1. *Un autre processus de coalescence, autorisant des collisions entre plus de deux lignées à la fois, sera l'objet d'une étude approfondie dans le Chapitre 3. Bertoin et Le Gall (2003) ont introduit une généralisation du processus de Fleming-Viot de manière à ce que ses généalogies soient représentées par des coalescents à collisions multiples. Ce superprocessus est appelé Fleming-Viot généralisé. Des relations de dualité, similaires à* (2.7)*, entre le processus de Fleming-Viot généralisé et le coalescent à collisions multiples sont prouvées dans Bertoin et Le Gall (2000) et Birkner* et al. *(2005). La construction look-down de Donnelly et Kurtz (1999) peut, elle aussi, être adaptée pour représenter de manière dénombrable le processus de Fleming-Viot généralisé. Une présentation de cette construction se trouve dans Birkner* et al. *(2005) ou dans Birkner et Blath (2007).*

Bibliographie

Bertoin, J. et Le Gall, J.-F. (2000). The Bolthausen-Sznitman coalescent and the genealogy of continuous-state branching processes. *Probab. Theory Related Fields*, 117(2):249–266.

Bertoin, J. et Le Gall, J.-F. (2003). Stochastic flows associated to coalescent processes. *Probab. Theory Related Fields*, 126(2):261–288.

Birkner, M. et Blath, J. (2007). Rescaled stable generalised Fleming-Viot processes : Flickering random measures. *http ://www.wias-berlin.de/main/publications/wias-publ/run.cgi ?template=abstract&type=Preprint&year=2007&number=1252. To appear.*

Birkner, M., Blath, J., Capaldo, M., Etheridge, A. M., Möhle, M., Schweinsberg, J. et Wakolbinger, A. (2005). Alpha-stable branching and Beta-coalescents. *Electron. J. Probab.*, 10(9):303–325.

Brémaud, P., Kannurpatti, R. et Mazumdar, R. (1992). Event and time averages : a review. *Adv. in Appl. Probab.*, 24(2):377–411.

Cattiaux, P., Collet, P., Lambert, A., Martinez, S., Méléard, S. et San Martin, J. (2009). Quasi-stationarity distributions and diffusion models in population dynamics. *http ://arxiv.org/abs/math/0703781. To appear.*

Collet, P., Martínez, S. et Maume-Deschamps, V. (2000). On the existence of conditionally invariant probability measures in dynamical systems. *Nonlinearity*, 13(4): 1263–1274.

Dawson, D. A. (1993). Measure-valued Markov processes. *In École d'Été de Probabilités de Saint-Flour XXI—1991*, volume 1541 de *Lecture Notes in Math.*, pages 1–260. Springer, Berlin.

Dawson, D. A. et Hochberg, K. J. (1982). Wandering random measures in the Fleming-Viot model. *Ann. Probab.*, 10(3):554–580.

Donnelly, P. et Kurtz, T. G. (1996). A countable representation of the Fleming-Viot measurable diffusion. *Ann. Probab.*, 24(2):698–742.

Donnelly, P. et Kurtz, T. G. (1999). Particle representations for measure-valued population models. *Ann. Probab.*, 27(1):166–205.

Durrett, R. (1996). *Stochastic calculus : a practical introduction*. Probability and Stochastics Series. CRC Press, Boca Raton,FL.

Durrett, R. (2008). *Probability models for DNA sequence evolution*. Probability and its Applications. Springer, New York, NY. second edition.

Etheridge, A. M. (2000). *An introduction to superprocesses*, volume 20 de *University Lecture Series*. American Mathematical Society, Providence, RI.

Ethier, S. N. et Kurtz, T. G. (1986). *Markov processes : characterization and convergence*. Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, NY.

Evans, S. N. et Ralph, P. L. (2008). Dynamics of the time to the most recent common ancestor in a large branching population. *http ://arxiv.org/abs/0812.1302. To appear.*

Ewens, W. J. (2004). *Mathematical population genetics. I. Theoretical introduction*, volume 27 de *Interdisciplinary Applied Mathematics*. Springer-Verlag, New York, NY. second edition.

Ferrari, P. A., Kesten, H., Martinez, S. et Picco, P. (1995). Existence of quasi-stationary distributions. A renewal dynamical approach. *Ann. Probab.*, 23(2):501–521.

Fleming, W. H. et Viot, M. (1978). Some measure-valued population processes. *In Stochastic analysis (Proc. Internat. Conf., Northwestern Univ., Evanston, Ill., 1978)*, pages 97–108. Academic Press, New York, NY.

Fleming, W. H. et Viot, M. (1979). Some measure-valued Markov processes in population genetics theory. *Indiana Univ. Math. J.*, 28(5):817–843.

Greven, A., Pfaffelhuber, P. et Winter, A. (2009). Tree-valued resampling dynamics : martingale problems and applications. *http ://arxiv.org/abs/0806.2224. To appear.*

Huillet, T. (2007). On Wright–Fisher diffusion and its relatives. *J. Stat. Mech.*, P1106.

Johnson, N. L. et Kotz, S. (1977). *Urn models and their application*. John Wiley & Sons Inc., New York, NY.

Lambert, A. (2008). Population dynamics and random genealogies. *Stoch. Models*, 24(suppl. 1):45–163.

Leocard, S. (2009). *Modèles probabilistes du balayage sélectif et auto-stop génétique*. Thèse de doctorat, Université Aix Marseille.

Moran, P. A. P. (1958). Random processes in genetics. *Proc. Cambridge Philos. Soc.*, 54:60–71.

Perkins, E. (2002). Dawson-Watanabe superprocesses and measure-valued diffusions. *In École d'Été de Probabilités de Saint-Flour XXIX—1999*, volume 1781 de *Lecture Notes in Math.*, pages 125–324. Springer, Berlin.

Pfaffelhuber, P. et Wakolbinger, A. (2006). The process of most recent common ancestors in an evolving coalescent. *Stochastic Process. Appl.*, 16(12):1836–1859.

Simon, D. et Derrida, B. (2006). Evolution of the most recent common ancestor of a population with no selection. *J. Stat. Mech.*, P05002.

Villemonais, D. (2009). Approximation of quasi-stationary distributions for 1-dimensional killed diffusions with unbounded drifts. *http ://arxiv.org/abs/0905.3636. To appear*.

Watanabe, S. (1968). A limit theorem of branching processes and continuous state branching processes. *J. Math. Kyoto Univ.*, 8:141–167.

Chapitre 3

Coalescent à collisions multiples et estimation du taux de mutation

Considérons une population de taille N évoluant suivant un modèle de Cannings introduit dans la Section 1.1. Rappelons que le nombre d'enfants de l'individu i de la génération r est donné par la variable aléatoire Υ_i^r et que les $(\Upsilon_1^r, \ldots, \Upsilon_N^r)$, $r \geq 0$, sont des copies i.I.d. d'une même variable $(\Upsilon_1, \ldots, \Upsilon_N)$. Rappelons de plus que la probabilité que deux individus, tirés au hasard, aient un ancêtre commun à la génération précédente est

$$c_N = \frac{\mathbb{E}[(\Upsilon_1 - 1)^2]}{N - 1}.$$

À une génération donnée, considérons un échantillon de taille n dans la population. Nous avons vu dans la Section 1.3 que si les conditions (1.20) et (1.22) sont respectées alors les généalogies d'un échantillon de n individus, accélérées c_N^{-1} fois, convergent en loi vers le n-coalescent de Kingman.

Supposons à présent que, lorsque $N \to \infty$, les trois conditions suivantes sont réalisées :

$$c_N \to 0, \tag{3.1}$$

$$c_N^{-1} \frac{\mathbb{E}[(\Upsilon_1 - 1)^k]}{N^{k-1}} \to \phi_k \geq 0 \text{ pour tout } k \geq 2 \tag{3.2}$$

et

$$c_N^{-1} \frac{\mathbb{E}[(\Upsilon_1 - 1)^2 \ldots (\Upsilon_a - 1)^2]}{N^{a-1}} \to 0 \text{ pour tout } a \geq 2. \tag{3.3}$$

La condition (3.1) est identique à (1.20). À la différence de (1.22), la condition (3.2) suggère qu'un événement de coalescence peut impliquer plus de deux lignées. En revanche, si un événement de coalescence a lieu, la condition (3.3) empêche qu'il y en ait d'autres au même instant.

En renormalisant le temps par c_N, le processus de coalescence qui apparaît est un *n-coalescent à collisions multiples* (Sagitov (1999)).

3.1 Le coalescent à collisions multiples

Le coalescent de Kingman, introduit dans la section 1.3, est un modèle utilisé pour des populations

- de taille constante
- sans sélection
- avec une loi de reproduction dont la variance est faible.

L'affaiblissement de ces hypothèses peut mener à rejeter le coalescent de Kingman pour approcher les généalogies (Tajima (1989), Fu et Li (1993)) et à introduire de nouveaux processus limites mieux adaptés. Imaginons par exemple une population où quelques individus ont de très larges portées et sont à l'origine d'une part non négligeable de la

population à la génération suivante. En remontant le temps, il sera plus pertinent de permettre à plusieurs lignées de coalescer en même temps. C'est le modèle retenu par Eldon et Wakeley (2006) pour des populations marines.

Le coalescent à collisions multiples, ou Λ-coalescent, est introduit de manière indépendante par Pitman (1999) et Sagitov (1999). Par collisions multiples, nous entendons que, lors d'un événement de coalescence, plusieurs lignées peuvent fusionner en une. En revanche, plusieurs événements de coalescence ne peuvent pas avoir lieu simultanément (dans ce cas on parlera de Ξ-coalescent, voir Schweinsberg (2000) ou Möhle et Sagitov (2001)).

Le coalescent à collisions multiples est une chaîne de Markov en temps continu,

$$\Pi := (\Pi_t, t \geq 0),$$

à valeurs dans $\mathcal{P}$, l'ensemble des partitions de $\mathbb{N}^*$. Sa valeur initiale est la partition ne contenant que des singletons et i et $j \in \mathbb{N}^*$ sont dans le même bloc de Π_t s'ils ont le même ancêtre au temps $-t$.

Comme le coalescent de Kingman, ce processus peut être construit, par des arguments de compatibilité, avec une famille de n-coalescents $(\Pi^{(n)} := (\Pi_t^{(n)}, t \geq 0), n \geq 1)$ à valeurs dans $\mathcal{P}^n$, l'espace des partitions de

$$[n] = \{1, \ldots, n\}.$$

Ces processus vérifient l'hypothèse suivante : pour tout couple (m, n) tel que $m \leq n$, si $\Pi^{(n)}_{|[m]}$ désigne la restriction de $\Pi^{(n)}$ aux partitions de $[m]$, alors

$$\Pi^{(n)}_{|[m]} \stackrel{d}{=} \Pi^{(m)}$$

(voir Bertoin (2006) et Pitman (2006) pour une étude plus détaillée). Comme ces n-coalescents sont des chaînes de Markov en temps continu (sur un espace fini), nous pouvons étudier leurs dynamiques.

Si $\Pi_t^{(n)}$ contient b blocs, soit $\lambda_{b,k}$ le taux auquel coalescent k blocs donnés, i.e. le taux auquel k lignées fusionnent en une seule. Consécutivement à la compatibilité de $\Pi^{(n)}$, ce taux ne dépend pas de n et il existe une relation de récurrence sur les taux (Pitman (1999)) :

$$\lambda_{b,k} = \lambda_{b+1,k} + \lambda_{b+1,k+1}. \tag{3.4}$$

Pour comprendre cette relation, il suffit d'imaginer comment peuvent fusionner les k blocs si l'on en rajoute un $b+1^{\text{e}}$. Soit ce dernier participe à l'événement de coalescence, soit il ne fusionne pas. Il y a donc compétition entre deux lois exponentielles, ce qui explique la somme de deux taux qui apparaît.

La condition (3.4) est en fait nécessaire et suffisante pour satisfaire la compatibilité. Il découle de la représentation de de Finetti de suites échangeables de 0 et de 1 (voir

Feller (1971), Section VII.4) que l'on peut établir une bijection entre les tableaux de $(\lambda_{b,k})$ vérifiant (3.4) et les mesures finies sur $[0,1]$.

Théorème 3.1. *Il existe une unique mesure finie Λ sur $[0,1]$ telle que, pour tout $b \geq k \geq 2$,*

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k}\Lambda(dx). \tag{3.5}$$

Ainsi peut être défini de manière unique le Λ-coalescent :

Définition et Théorème 3.1. *Soit $(\Pi^{(n)}, n \geq 1)$ une famille de n-coalescents dont les dynamiques satisfont (3.5). Il existe un unique (en loi) processus markovien $(\Pi_t, t \geq 0)$ à valeurs dans $\mathcal{P}$ tel que*

$$\forall n \geq 1, \Pi_{|[n]} \stackrel{d}{=} \Pi^{(n)}.$$

Ce processus est appelé coalescent à collisions multiples.

Citons deux mesures Λ menant à des coalescents remarquables.

Si $\Lambda = \delta_0$, pour tout $b \geq 2$, le coefficient $\lambda_{b,k}$ vaut 1 si $k = 2$ et 0 sinon. Nous retrouvons ainsi le coalescent de Kingman.

Si Λ est la mesure de Lebesgue sur $[0,1]$, nous obtenons le coalescent de Bolthausen-Szmitman (Bolthausen et Sznitman (1998)) utilisé dans des applications physiques telles que les verres de spin (voir le Chapitre 6 de Berestycki (2009)) Il peut être construit par le biais de généalogies de populations branchantes (Bertoin et Le Gall (2000)) et certains physiciens ont établi la conjecture que ce coalescent représente les généalogies des populations sous certaines conditions de sélection (Brunet *et al.* (2006, 2007)).

L'exemple des populations marines modélisées par Eldon et Wakeley (2006) est en fait un modèle de Cannings où la variance de la loi de reproduction est élevée. L'hypothèse (1.22) n'est alors pas vérifiée. Le Λ-coalescent apparaît comme processus limite d'après Sagitov (1999) (ou Möhle et Sagitov (2001)).

Une autre manière de modéliser les espèces marines part du constat que les individus se reproduisent en très grand nombre mais que seuls quelques-uns des enfants survivent : le rapport entre la population effective (qui donne la valeur par laquelle le temps doit être renormalisé afin d'obtenir un processus limite non trivial) et la population totale est très faible à chaque génération. Plusieurs observations ont été établies dans ce sens, par Hedgecock (1994) et Boom *et al.* (1994) pour des huîtres du Pacifique (*Crassostrea gigas*), ou par Árnason (2004) pour des morues de l'Atlantique (*Gadus morhua*), entre autres.

Plutôt que d'utiliser un modèle de Cannings, Schweinsberg (2003) a considéré un *modèle de Galton-Watson*, dont la taille n'est pas supposée fixe, surcritique et dont N éléments sont conservés à chaque génération.

Plus précisément, supposons que

$$\mathbb{P}(\Upsilon_1 > k) \sim Ck^{-\alpha},\ \alpha \geq 1$$

et

$$\mathbb{E}[\Upsilon_1] > 1.$$

Considérons le processus de Galton-Watson associé, partant de N individus, tel qu'à chaque génération seuls N individus, choisis au hasard, uniformément dans la population, sont conservés. Alors, si dans le cas où la variance de Υ_1 est finie ($\alpha \geq 2$), les généalogies, après rééchelonnement du temps, peuvent être approchées par un coalescent de Kingman, dans le cas où $\alpha \in (0, 2)$, le processus limite qui décrit les généalogies de cette population est un Λ-coalescent dont la mesure finie est une loi $Beta(2-\alpha, \alpha)$. Autrement dit,

$$\Lambda(dx) = \frac{1}{\Gamma(2-\alpha)\Gamma(\alpha)} x^{1-\alpha}(1-x)^{\alpha-1}dx. \tag{3.6}$$

Ces coalescents appartiennent à la classe des *Beta-coalescents* (le lecteur pourra se référer à Birkner *et al.* (2005), Bertoin et Le Gall (2006), Berestycki *et al.* (2007)). En fait, en toute généralité, le Beta-coalescent s'obtient lorsque Λ est la mesure d'une loi $Beta(a, b)$, avec a et b strictement positifs. Dans la suite de ce document, en l'absence de précisions, nous considérerons que la mesure Λ d'un Beta-coalescent est celle d'une loi $Beta(2-\alpha, \alpha)$, avec $\alpha \in (0, 2)$. Nous retrouvons le coalescent de Kingman pour $\alpha \to 2$ et le coalescent de Bolthausen et Szmitman pour $\alpha = 1$.

Ce dernier est tout à fait intéressant puisqu'il marque la frontière entre deux classes de coalescents dont les comportements sont très différents. En particulier, lorsque $\alpha \in (0, 1]$ les coalescents obtenus ne descendent pas de l'infini. C'est en revanche le cas lorsque $\alpha > 1$, Berestycki *et al.* (2009) en établissent la vitesse. Les résultats que nous énoncerons dans la section suivante concernent la cas $\alpha \in (1, 2)$. Intuitivement, plus α est proche de 2, plus les événements de coalescence impliquant deux (et seulement deux) lignées seront favorisés (voir Figure 3.1, utilisée avec l'aimable autorisation d'Emilia Huertas Sanchez).

La formule (3.5), associée à la représentation de de Finetti sous-jacente, jette les bases d'une construction poissonnienne du coalescent (Pitman (1999)). Soit $\bar{\Lambda}$ une mesure finie sur $[0, 1]$ telle que $\bar{\Lambda}(\{0\}) = 0$. Soit un processus ponctuel de Poisson sur $\mathbb{R} \times (0, 1]$ d'intensité

$$dt \otimes \bar{\nu}(dy)$$

où

$$\bar{\nu}(dy) = y^{-2}\bar{\Lambda}(dy).$$

À chaque atome (t_i, y_i) de ce processus correspond un événement de coalescence. Chaque bloc présent au temps t_i- tire de manière indépendante une pièce avec probabilité y_i de faire pile. Tous les blocs qui ont tiré pile fusionnent en t_i. Il est alors possible de reconstruire le Λ-coalescent avec

$$\Lambda(dx) = \rho\delta_0(dx) + \bar{\Lambda}(dx).$$

En d'autres termes, un Λ-coalescent peut être décomposé en deux parties : une partie « Kingman » où chaque paire de blocs coalesce à taux ρ et une partie « collisions multiples » où, à taux $\bar{\nu}(dx)$, une proportion x de la population coalesce.

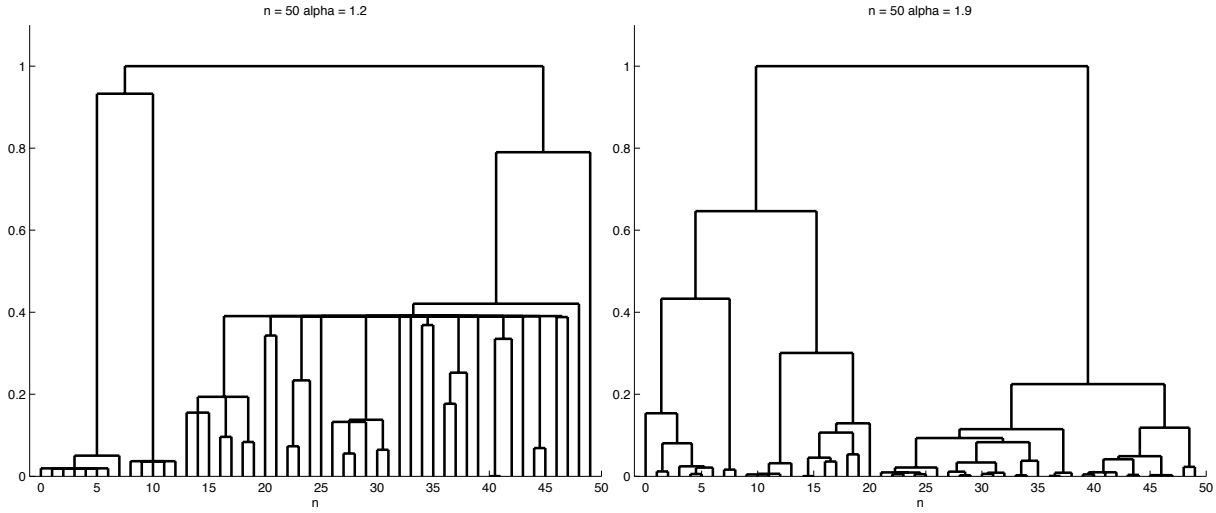

FIGURE 3.1 – Simulations de Beta-coalescents pour une population initiale de 50 individus.

Pourquoi est-il nécessaire que $\int_0^1 \Lambda(dx)$ soit finie ? Simplement parce que nous avons besoin, afin que le processus soit bien défini, que deux individus fixés coalescent à taux fini, sans quoi les événements de coalescence seraient invisibles. Or, soit les deux individus fusionnent suivant les dynamiques du coalescent de Kingman, soit suivant celles du Λ-coalescent. Si l'on se trouve dans le second cas, en utilisant le processus ponctuel de Poisson introduit plus haut, c'est que les deux individus ont tiré « pile » à une expérience de Bernoulli dont le paramètre est lui-même tiré suivant la mesure ν. Ce taux est donc $\rho + \int_0^1 x^2 \bar{\nu}(dx)$.

3.2 L'arbre de coalescence et les taux de mutation

Dans la Section 1.5, nous avons expliqué que le fait d'introduire des mutations dans les généalogies d'une population revient, asymptotiquement, à les répartir sur l'arbre suivant un processus de Poisson. Le paramètre de ce processus est noté θ.

La question de l'estimation de θ se pose alors. À ce sujet , et pour une introduction aux problèmes d'inférence statistique dans des modèles de génétique des populations, nous renvoyons le lecteur à Tavaré (2004). Dans le modèle à infinité de sites (voir Section 3.2), le nombre total de mutations $S^{(n)}$ est déterminé par le nombre de sites de ségrégation observés. Conditionnellement à la longueur de l'arbre $L^{(n)}$, $S^{(n)}$ suit une loi de Poisson de paramètre $\theta L^{(n)}$. Par conséquent, et puisque $L^{(n)} \xrightarrow{\mathbb{P}} \infty$,

$$\frac{S^{(n)} - \theta L^{(n)}}{\sqrt{\theta L^{(n)}}} \xrightarrow{d} G,$$

G étant une variable aléatoire gaussienne standard. Il en résulte que la connaissance du comportement asymptotique de $L^{(n)}$ permet d'établir le comportement asymptotique de $S^{(n)}$. En effet, s'il est possible de trouver deux suites $(a_n, n \geq 2)$ et $(b_n, n \geq 2)$ telles que

$$\frac{L^{(n)} - a_n}{b_n}$$

converge en loi vers une variable aléatoire non-dégénérée, il suffira d'utiliser la décomposition

$$\frac{S^{(n)} - \theta a_n}{\theta b_n} = \frac{S^{(n)} - \theta L^{(n)}}{\sqrt{\theta L^{(n)}}} \frac{\sqrt{\theta L^{(n)}}}{\theta b_n} + \frac{L^{(n)} - a_n}{b_n}$$

pour espérer trouver les fluctuations de $S^{(n)}$.

Dans le cas du Beta-coalescent avec $\alpha \in (1, 2)$, Berestycki *et al.* (2008) établissent le premier ordre du comportement asymptotique de $L^{(n)}$:

$$n^{\alpha-2} L^{(n)} \xrightarrow{\mathbb{P}} \frac{\Gamma(\alpha)\alpha(\alpha-1)}{2-\alpha}, \tag{3.7}$$

ce qui entraîne que

$$n^{\alpha-2} S^{(n)} \xrightarrow{\mathbb{P}} \theta \frac{\Gamma(\alpha)\alpha(\alpha-1)}{2-\alpha}. \tag{3.8}$$

Nous nous proposons de déterminer le second ordre du développement asymptotique de $L^{(n)}$ et $S^{(n)}$.

Considérons un Λ-coalescent $\Pi = (\Pi_t, t \geq 0)$ et son n-coalescent associé $\Pi^{(n)}$ et, comme dans (1.25) pour le coalescent de Kingman, notons

$$R_t^{(n)} = \left| \Pi_t^{(n)} \right| \tag{3.9}$$

le nombre de blocs de $\Pi^{(n)}$ au temps t. $R_0^{(n)} = n$ et $R_t^{(n)}$ peut être vu comme le nombre d'ancêtres vivant au temps $-t$. Le temps d'apparition du plus récent ancêtre commun est

$$\inf\{t \geq 0, R_t^{(n)} = 1\}.$$

Nous omettrons l'exposant n si cela n'entraîne pas de confusions. Intéressons-nous de nouveau aux dynamiques du processus $R = (R_t, t \geq 0)$ dans le cadre plus général des coalescents à collisions multiples. Le nombre de choix possibles de $l+1$ blocs parmi k est $\binom{k}{l+1}$ (pour tout $1 \leq l \leq k-1$) et chaque groupe de $l+1$ blocs fusionne à taux $\lambda_{k,l+1}$ défini par (3.5). Ainsi, lorsque R est en k, le temps d'attente avant premier saut est une variable aléatoire exponentielle de paramètre

$$g_k = \sum_{l=1}^{k-1} \binom{k}{l+1} \lambda_{k,l+1} = \int_0^1 (1 - (1-x)^k - kx(1-x)^{k-1}) \frac{\Lambda(dx)}{x^2} \tag{3.10}$$

et est donc distribué comme E/g_k, où E est une variable aléatoire exponentielle de paramètre 1. Nous retrouvons bien

$$g_k = \binom{k}{2}$$

pour le coalescent de Kingman. Pour le Beta-coalescent, nous sommes capables d'établir que, lorsque $k \to \infty$,

$$g_k \sim \frac{1}{\alpha\Gamma(\alpha)}\, k^\alpha. \tag{3.11}$$

Ce résultat est une conséquence du Lemme 4 de Bertoin et Le Gall (2006)). Un développement au second ordre est établi dans le Lemme 5.2.

Définissons la suite des temps de saut $(T_k, k \geq 0)$ du coalescent de manière récurrente par $T_0 = 0$ et, pour $k \geq 1$,

$$T_k = \inf\{t > T_{k-1}, R_t \neq R_{T_{k-1}}\}$$

Par convention, $\inf \emptyset = 0$. Définissons aussi τ_n comme le nombre de sauts du processus $R^{(n)}$ jusqu'à ce qu'il atteigne l'état absorbant 1, c'est-à-dire le nombre total de coalescences du processus $\Pi^{(n)}$:

$$\tau_n = \inf\{k, R_{T_k} = 1\}$$

Dans le cas du coalescent de Kingman, puisque seules deux lignées peuvent fusionner à la fois,

$$\tau_n = n - 1.$$

Dans le cas du Beta-coalescent, nous prouvons le résultat de convergence suivant :

Proposition 3.1. *Si $\Lambda \sim Beta(2-\alpha, \alpha)$ avec $\alpha \in (1, 2)$,*

$$n^{-1/\alpha}\left(n - \frac{\tau_n}{\alpha - 1}\right) \xrightarrow{d} V_{\alpha-1} \tag{3.12}$$

où $V = (V_t, t \geq 0)$ est un processus de Lévy α-stable avec des sauts négatifs dont l'exposant de Laplace $\psi(u)$ est $u^\alpha/(\alpha-1)$, en d'autres termes $\mathbb{E}[\mathrm{e}^{-uV_t}] = \mathrm{e}^{tu^\alpha/(\alpha-1)}$.

Un résultat plus général se trouve, avec sa preuve, dans le Chapitre suivant sous le nom de Proposition 5.1. Le lecteur peut se référer au Chapitre VII de Bertoin (1996) pour plus de détails sur les processus stables. Ce résultat est aussi obtenu par Gnedin et Yakubovich (2007) et Iksanov et Möhle (2008).

En fait, les cas $\alpha \in (0, 1)$ (Iksanov et Möhle (2008)) et $\alpha = 1$ (Panholzer (2004), Drmota *et al.* (2007), Iksanov et Möhle (2008) ou Drmota *et al.* (2009)) ont aussi été étudiés . Nous résumons les résultats dans le tableau suivant :

α	Comportement asymptotique de τ_n
$\alpha \to 2$	$\tau_n = n - 1$
$1 < \alpha < 2$	$n^{-1/\alpha}\left(n - \frac{\tau_n}{\alpha-1}\right) \xrightarrow{d} V_{\alpha-1}$, ($\alpha$-stable)
$\alpha = 1$	$\frac{(\log n)^2}{n}\tau_n - \log(n \log n) \xrightarrow{d} V$, (1-stable)
$0 < \alpha < 1$	$\frac{\tau_n}{\Gamma(2-\alpha)n^\alpha} \xrightarrow{d} \int_0^\infty \mathrm{e}^{-U_t}\, dt$, ($U$ est un subordinateur)

Ce cadre peut être généralisé à celui de Beta-coalescents dont la mesure Λ est celle d'une loi $Beta(a, b)$ avec a et b strictement positifs. Puisque seul le comportement de la mesure Λ à proximité de 0 importe, les résultats énoncés précédemment correspondent au cas $0 < a \leq 2$. Lorsque $a > 2$, le comportement asymptotique de τ_n est étudié dans Gnedin *et al.* (2008).

Remarque 3.1. *En écho à la Section 1.3, le comportement asymptotique de la taille d'une branche externe, choisie au hasard dans un Beta-coalescent, est établi par Gnedin* et al. *(2008) pour Λ suivant une loi $Beta(a, b)$ avec $a > 2$, et dans Freund et Möhle (2009) pour le coalescent de Bolthausen-Szmitman. Ces deux articles décrivent aussi le temps nécessaire pour remonter au plus récent ancêtre commun de la population actuelle (voir aussi Möhle (2004) pour des modèles discrets de population). Rappelons que le cas du coalescent de Kingman est traité dans Caliebe* et al. *(2007).*

Soit $Y = (Y_k, k \geq 1)$ la chaîne induite du processus R. Plus précisément, $Y_0 = R_0$ et, pour $k \geq 1$,

$$Y_k = R_{T_k}.$$

Par convention, $\inf \emptyset = +\infty$ et $Y_k = 1$ lorsque $k \geq \tau_n$. Nous serons amené à écrire $Y^{(n)}$ plutôt que Y lorsque nous souhaiterons appuyer le fait que Y débute en n. Les probabilités de transition de Y sont

$$P(k, k-l) = \frac{\binom{k}{l+1}\lambda_{k,l+1}}{g_k}, 1 \leq l \leq k-1. \tag{3.13}$$

Pour calculer la longueur totale de l'arbre de coalescence, il suffit de sommer toutes ses longueurs de branches jusqu'au plus récent ancêtre commun. D'après (3.10), ces longueurs sont les variables exponentielles de paramètre g_k. La loi de la longueur totale est donc définie par

$$L^{(n)} := \sum_{k=0}^{\tau_n - 1} \frac{Y_k^{(n)}}{g_{Y_k^{(n)}}} E_k$$

où les $(E_k, k \geq 0)$ sont des variables aléatoires exponentielles indépendantes, de paramètre 1 et indépendantes de $Y^{(n)}$.

Nous faisons, dans ce document, un premier pas vers la détermination du second ordre du développement de $L^{(n)}$ (et de $S^{(n)}$) en établissant un résultat partiel sur la loi asymptotique de la longueur de l'arbre jusqu'à la $\lfloor nt \rfloor^{\text{e}}$ coalescence :

$$L_t^{(n)} := \sum_{k=0}^{\lfloor nt \rfloor \wedge (\tau_n - 1)} \frac{Y_k^{(n)}}{g_{Y_k^{(n)}}} E_k \tag{3.14}$$

Rappelons la définition de la mesure ν :

$$\nu(dx) = x^{-2}\Lambda(dx) \tag{3.15}$$

et soit

$$\rho(t) = \nu((t, 1]).$$

Nous nous placerons dans le cas où

$$\rho(t) = C_0 t^{-\alpha} + \mathcal{O}(t^{-\alpha+\zeta}) \tag{3.16}$$

avec $\alpha \in (1, 2)$, $C_0 > 0$ et $\zeta > 1 - 1/\alpha$. Si la mesure Λ est celle d'une loi $Beta(2 - \alpha, \alpha)$, elle satisfait cette condition, avec

$$C_0 = \frac{1}{\alpha\Gamma(\alpha)\Gamma(2 - \alpha)}\,.$$

Nous conservons aussi le comportement asymptotique de τ_n énoncé en (3.12).

La première étape est d'approximer $L_t^{(n)}$ en remplaçant les E_k par leurs moyennes, 1, et les $g_{Y_k^{(n)}}$ par leurs équivalents donnés en (3.11). Définissons alors

$$\hat{L}_t^{(n)} := \sum_{k=0}^{\lfloor nt \rfloor \wedge (\tau_n - 1)} (Y_k^{(n)})^{1-\alpha}. \tag{3.17}$$

La proximité de $L_t^{(n)}$ et $\frac{\hat{L}_t^{(n)}}{C_0\Gamma(2-\alpha)}$ est démontrée dans les Lemmes 5.8 et 5.9.

Pour $t \in [0, \alpha - 1]$, posons

$$v(t) = \int_0^t \left(1 - \frac{r}{\alpha - 1}\right)^{1-\alpha} dr.$$

Alors, le Théorème 5.1 établit que, si $t_0 \in [0, \alpha - 1)$ et $\delta > 0$, alors lorsque $n \to \infty$

$$n^{(\alpha-1)/2-\delta} \sup_{0 \le t \le t_0} \left| n^{-2+\alpha} \hat{L}_t^{(n)} - v(t) \right| \xrightarrow{\mathbb{P}} 0 \tag{3.18}$$

et, pour tout t dans $[0, \alpha - 1)$,

$$n^{-1+\alpha-1/\alpha}(\hat{L}_t^{(n)} - n^{2-\alpha} v(t)) \xrightarrow{d} (\alpha - 1) \int_0^t dr \left(1 - \frac{r}{\alpha - 1}\right)^{-\alpha} V_r, \tag{3.19}$$

ce qui permet de déduire notre résultat principal de la Partie (ce résultat est prouvé dans le Chapitre 5 sous le nom de Théorème 5.2) :

Théorème 3.2. *Supposons que la condition* (3.16) *est vérifiée.*

1. *Soient $t_0 \in [0, \alpha - 1)$ et $\delta > 0$. Alors*

$$n^{-((5-3\alpha)_+/2)-\delta} \sup_{0 \le t \le t_0} \left| L_t^{(n)} - n^{2-\alpha} \frac{v(t)}{C_0 \Gamma(2-\alpha)} \right| \xrightarrow{\mathbb{P}} 0 \tag{3.20}$$

   lorsque $n \to \infty$.

2. *Soit $\alpha \in (1, (1+\sqrt{5})/2)$. Pour tout $t \in (0, \alpha - 1)$, nous avons*

$$n^{-1+\alpha-1/\alpha} \left(L_t^{(n)} - n^{2-\alpha} \frac{v(t)}{C_0 \Gamma(2-\alpha)} \right) \xrightarrow{d} \frac{\alpha - 1}{C_0 \Gamma(2-\alpha)} \int_0^t dr \left(1 - \frac{r}{\alpha - 1}\right)^{-\alpha} V_r \tag{3.21}$$

   lorsque $n \to \infty$.

Notons que

$$\alpha = \frac{1+\sqrt{5}}{2} \Rightarrow -1 + \alpha - \frac{1}{\alpha} = 0.$$

Intuitivement, la proximité de τ_n et de $n(\alpha - 1)$ laisse à penser que $L_{\alpha-1}^{(n)}$ et $L^{(n)}$ ne sont pas très éloignées. En particulier, nous nous attendons à ce que $n^{\alpha-2} L^{(n)}$ converge en probabilité vers $\frac{v(\alpha-1)}{C_0\Gamma(2-\alpha)}$. Or, dans le cas du Beta-coalescent du moins, cette valeur limite correspond à celle trouvée dans le Théorème 1.9 de Berestycki *et al.* (2008), et énoncée en (3.7). Nous proposons dans la Figure 3.2 des simulations de la longueur totale de Beta-coalescents permettant d'observer la vitesse à laquelle converge $n^{\alpha-2} L^{(n)}$. Il est à noter que la convergence semble plus rapide pour des valeurs de α proches de 1.7.

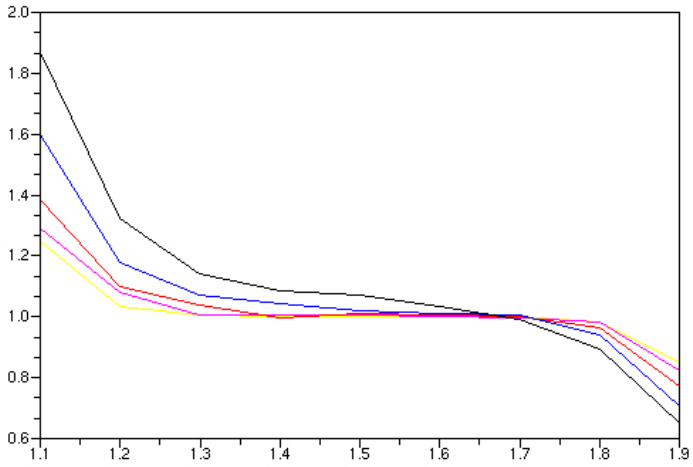

FIGURE 3.2 – Simulations de $\frac{2-\alpha}{\Gamma(\alpha)\alpha(\alpha-1)} n^{\alpha-2} L^{(n)}$ pour un Beta-coalescent avec $\alpha \in \{1.1, \ldots; 1.9\}$ (abscisses). Valeurs pour $n = 10^4$ (noir), $n = 10^5$ (bleu), $n = 10^6$ (rouge), $n = 10^7$ (rose), $n = 10^8$ (jaune).

La loi limite obtenue en (3.21) n'est pas définie pour $t = \alpha - 1$. Nous ne pouvons donc pas savoir si les fluctuations obtenues dans le Théorème 3.2 sont aussi les fluctuations de $L^{(n)}$, avec une autre loi limite, ou si elles sont différentes.

Nous ne connaissons pas les fluctuations de $L_t^{(n)}$ pour $\alpha \geq (1+\sqrt{5})/2 \approx 1.62$. Nous pourrons toutefois obtenir celes de $S_t^{(n)}$ pour α entre 1 et 2. En effet, lorsque $\alpha > \sqrt{2}$ (notons que $\sqrt{2} < (1+\sqrt{5})/2$), l'approximation de $S^{(n)}$ par la loi de Poisson l'emporte sur la loi limite de $L^{(n)}$. La distribution asymptotique du nombre total de mutations dans l'arbre, du moins jusqu'à la $\lfloor nt \rfloor^{\text{e}}$ coalescence, se déduit alors.

Corollaire 3.1. *Sous les conditions* (3.16)*, soient* t *dans* $(0, \alpha-1)$ *et* G *une loi gaussienne centrée réduite indépendante du processus* V*.*

1. *Soit* $\alpha \in (1, \sqrt{2})$*. Alors*

$$n^{-1+\alpha-1/\alpha} \left(S_t^{(n)} - \theta n^{2-\alpha} \frac{v(t)}{C_0 \Gamma(2-\alpha)} \right) \xrightarrow{d} \theta \frac{\alpha-1}{C_0 \Gamma(2-\alpha)} \int_0^t dr \left(1 - \frac{r}{\alpha-1}\right)^{-\alpha} V_r \tag{3.22}$$

   lorsque $n \to \infty$*.*

2. *Soit* $\alpha \in (\sqrt{2}, 2)$*. Alors*

$$n^{-1+\alpha/2} \left(S_t^{(n)} - \theta n^{2-\alpha} \frac{v(t)}{C_0 \Gamma(2-\alpha)} \right) \xrightarrow{d} \sqrt{\theta \frac{v(t)}{C_0 \Gamma(2-\alpha)}} G \tag{3.23}$$

lorsque $n \to \infty$.

3. *Soit* $\alpha = \sqrt{2}$. *Alors* $-1 + \alpha - \frac{1}{\alpha}$ $-1 + \frac{\alpha}{2}$ *et*

$$n^{-1+\alpha-1/\alpha}\left(S_t^{(n)} - \theta n^{2-\alpha}\frac{v(t)}{C_0\Gamma(2-\alpha)}\right) \xrightarrow{d} \theta\frac{\alpha-1}{C_0\Gamma(2-\alpha)}\int_0^t dr\left(1-\frac{r}{\alpha-1}\right)^{-\alpha} V_r + \sqrt{\theta\frac{v(t)}{C_0\Gamma(2-\alpha)}}G \tag{3.24}$$

lorsque $n \to \infty$.

Même si ce dernier résultat est incomplet, puisqu'il ne donne pas le comportement du nombre de mutations sur tout l'arbre, il nous permet tout de même de conjecturer la vitesse de convergence de l'estimateur

$$\hat{\theta}_n = \frac{C_0\Gamma(2-\alpha)S^{(n)}}{v(\alpha-1)n^{2-\alpha}}.$$

Le tableau suivant résume les comportements asymptotiques de $S^{(n)}$ pour le Beta-coalescent avec α compris entre 0 et 2. Nous y inscrivons notre conjecture et citons les résultats de Watterson (1975), Möhle (2006) et Drmota *et al.* (2007). Nous nous contenterons d'indiquer les ordres de grandeur des suites a_n et b_n telles que

$$\frac{S^{(n)} - \theta a_n}{\theta b_n}$$

converge en loi.

α	a_n	b_n
$\alpha \to 2$	$\sum_{k=1}^{n-1}\frac{1}{k}$	$\sum_{k=1}^{n-1}\frac{1}{k}$
$\sqrt{2} \le \alpha < 2$	$n^{2-\alpha}$	$n^{1-\alpha/2}$
$1 < \alpha < \sqrt{2}$	$n^{2-\alpha}$	$n^{1-\alpha+1/\alpha}$
$\alpha = 1$	$\frac{n}{\log n} + \frac{n\log(\log n)}{(\log n)^2}$	$\frac{n}{(\log n)^2}$
$0 < \alpha < 1$	0	n

Remarque 3.2. *Une autre valeur d'intérêt est le spectre des fréquences de sites (et celui des fréquences d'allèles dans le modèle à infinité d'allèles, voir la section 1.5). Un grand enjeu des mathématiques appliquées à la génétique des populations est en effet de généraliser la formule d'échantillonnage d'Ewens (1972). Pour* α *dans* $(1, 2)$*, Berestycki* et al. *(2007) montrent que ceux-ci se comportent comme le nombre de sites de ségrégations (ou le nombre d'allèles différents observés), à savoir :*

$$S^{(k,n)} = \mathcal{O}(n^{2-\alpha})$$

et

$$K^{(k,n)} = \mathcal{O}(n^{2-\alpha}).$$

Les fréquences alléliques sont aussi étudiées par Basdevant et Goldschmidt (2008) dans le cadre du coalescent de Bolthausen-Szmitman. Ces résultats offrent une autre approche pour estimer θ *puisque les spectres de fréquence sont aussi des valeurs observables.*

Bibliographie

Basdevant, A.-L. et Goldschmidt, C. (2008). Asymptotics of the allele frequency spectrum associated with the Bolthausen-Sznitman coalescent. *Electron. J. Probab.*, 13:486–512.

Berestycki, J., Berestycki, N. et Limic, V. (2009). The Λ-coalescent speed of coming down from infinity. *http ://arxiv.org/abs/0807.4278. To appear.*

Berestycki, J., Berestycki, N. et Schweinsberg, J. (2007). Beta-coalescents and continuous stable random trees. *Ann. Probab.*, 35(5):1835–1887.

Berestycki, J., Berestycki, N. et Schweinsberg, J. (2008). Small-time behavior of Beta-coalescents. *Ann. Inst. H. Poincaré Probab. Stat.*, 44(2):214–238.

Berestycki, N. (2009). *Recent progress in coalescent theory. www.statslab.cam.ac.uk/ beresty/rp2.pdf. Work in progress.*

Bertoin, J. (1996). *Lévy processes*. Cambridge University Press, Cambridge.

Bertoin, J. (2006). *Random fragmentation and coagulation processes*, volume 102 de *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge.

Bertoin, J. et Le Gall, J.-F. (2000). The Bolthausen-Sznitman coalescent and the genealogy of continuous-state branching processes. *Probab. Theory Related Fields*, 117(2):249–266.

Bertoin, J. et Le Gall, J.-F. (2006). Stochastic flows associated to coalescent processes. III. Limit theorems. *Illinois J. Math.*, 50(1-4):147–181.

Birkner, M., Blath, J., Capaldo, M., Etheridge, A. M., Möhle, M., Schweinsberg, J. et Wakolbinger, A. (2005). Alpha-stable branching and Beta-coalescents. *Electron. J. Probab.*, 10(9):303–325.

Bolthausen, E. et Sznitman, A.-S. (1998). On Ruelle's probability cascades and an abstract cavity method. *Comm. Math. Phys.*, 197(2):247–276.

Boom, J. D. G., Boulding, E. et Beckenbach, A. (1994). Mitochondrial DNA variation in introduced populations of pacific oyster, crassostrea gigas, in british columbia. *Can. J. Fish. Aquat. Sci.*, 51:1608–1614.

Brunet, E., Derrida, B., Mueller, A. H. et Munier, S. (2006). Noisy traveling waves : effect of selection on genealogies. *Europhys. Lett.*, 76(1):1–7.

Brunet, É., Derrida, B., Mueller, A. H. et Munier, S. (2007). Effect of selection on ancestry : an exactly soluble case and its phenomenological generalization. *Phys. Rev. E (3)*, 76(4):041104, 20.

Caliebe, A., Neininger, R., Krawczak, M. et Rösler, U. (2007). On the length distribution of external branches in coalescence trees : genetic diversity within species. *Theoret. Population Biol.*, 72(2):245–252.

Drmota, M., Iksanov, A., Möhle, M. et Rösler, U. (2007). Asymptotic results concerning the total branch length of the Bolthausen-Sznitman coalescent. *Stochastic Process. Appl.*, 117(10):1404–1421.

Drmota, M., Iksanov, A., Möhle, M. et Rösler, U. (2009). A limiting distribution for the number of cuts needed to isolate the root of a random recursive tree. *Random Structures Algorithms*, 34(3):319–336.

Eldon, B. et Wakeley, J. (2006). Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172:2621–2633.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoret. Population Biol.*, 3(1):87–112.

Feller, W. (1971). *An introduction to probability theory and its applications. Vol. II.* John Wiley & Sons Inc., New York, NY. second edition.

Freund, F. et Möhle, M. (2009). On the time back to the most recent common ancestor and the external branch length of the Bolthausen-Sznitman coalescent. *Markov Process. Related Fields*, 15. *À paraître.*

Fu, Y. X. et Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics*, 133:693–709.

Gnedin, A., Iksanov, A. et Möhle, M. (2008). On asymptotics of exchangeable coalescents with multiple collisions. *J. Appl. Probab.*, 45(4):1186–1195.

Gnedin, A. et Yakubovich, Y. (2007). On the number of collisions in Λ-coalescents. *Electron. J. Probab.*, 12(56):1547–1567.

Hedgecock, D. (1994). *Genetics and evolution of aquatic organisms*, chapitre Does variance in reproductive success limit effective population size of marine organisms ?, page 122–134. Chapman and Hall, London.

Iksanov, A. et Möhle, M. (2008). On the number of jumps of random walks with a barrier. *Adv. in Appl. Probab.*, 40(1):206–228.

Möhle, M. (2004). The time back to the most recent common ancestor in exchangeable population models. *Adv. in Appl. Probab.*, 36(1):78–97.

Möhle, M. (2006). On the number of segregating sites for populations with large family sizes. *Adv. in Appl. Probab.*, 38(3):750–767.

Möhle, M. et Sagitov, S. (2001). A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.*, 29(4):1547–1562.

Panholzer, A. (2004). Destruction of recursive trees. *In Mathematics and computer science. III*, Trends Math., pages 267–280. Birkhäuser, Basel.

Pitman, J. (1999). Coalescents with multiple collisions. *Ann. Probab.*, 27(4):1870–1902.

Pitman, J. (2006). Combinatorial stochastic processes. *In École d'Été de Probabilités de Saint-Flour XXXII—2002*, volume 1875 de *Lecture Notes in Math.*, pages 1–256. Springer-Verlag, Berlin.

Sagitov, S. (1999). The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.*, 36(4):1116–1125.

Schweinsberg, J. (2000). Coalescents with simultaneous multiple collisions. *Electron. J. Probab.*, 5:1–50.

Schweinsberg, J. (2003). Coalescent processes obtained from supercritical Galton-Watson processes. *Stochastic Process. Appl.*, 106(1):107–139.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123(3):585–595.

Tavaré, S. (2004). Ancestral inference in population genetics. *In Lectures on probability theory and statistics*, volume 1837 de *Lecture Notes in Math.*, pages 1–188. Springer, Berlin.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoret. Population Biol.*, 7:256–276.

Árnason, E. (2004). Mitochondrial cytochrome b dna variation in the high-fecundity atlantic cod : Trans-atlantic clines and shallow gene genealogy. *Genetics*, 166:1871–1885.

Deuxième partie

Articles

Chapitre 4

Les deux plus anciennes familles dans le processus de Wright-Fisher

Version non modifiée de l'article
On the two oldest families for the Wright-Fisher process
admis avec révisions pour publication dans Electronic Journal of Probability.

4.1 Introduction

Many models have been introduced to describe population dynamics in population genetics. Fisher (1930), Wright (1931) and Moran (1958) have introduced two models for exchangeable haploid populations of constant size. A generalization has been given by Cannings (1974). Looking backward in time at the genealogical tree leads to coalescent processes, see Griffiths (1980) for one of the first papers with coalescent ideas. For a large class of exchangeable haploid population models of constant size, when the size N tends to infinity and time is measured in units of "N generations", the associated coalescent process is Kingman's coalescent (Kingman (1982)) (see also Pitman (1999), Sagitov (1999), Möhle and Sagitov (2001), Schweinsberg (2000) for general coalescent processes associated with Cannings' model). One of the associated object of interest is the most recent common ancestor (MRCA) of the population currently alive, which is also the depth of their genealogical tree (see Ewens (2004), Durrett (2008)). In the case of Kingman's coalescent, each couple of particle merges at rate one, which gives an MRCA of expectation 2, or an expectation equivalent to $2N$ generations in the discrete case (see Ewens (2004) for more results on this approximation and Fu (2006) for exact coalescent for the Wright-Fisher Model). MRCAs have been studied for many other models (see e.g. Chang (1999) for a more relevant model for human population, where the MRCA of a population of N individuals is almost $\log_2 N$ generations ago). In the special case of Fleming-Viot processes, which genealogies at a fixed time are given by Kingman's coalescent, Greven et al. (2009) have introduced a tree-valued process as solution of a martingale problem that represents the evolving genealogies.

In Moran model (finite population size) and in Wright-Fisher model with infinite population size, only two lineages can merge at a time and the genealogy at a given time is given by Kingman's coalescent. At time t the population is divided in two "oldest" families each one born from one of the two children of the MRCA. Let X_t and $1 - X_t$ denote the relative proportion of those two oldest families. One of this two oldest families will disappear (in the future). Let Y_t be the relative size of the oldest family which will fixate: either $Y_t = X_t$ or $Y_t = 1 - X_t$. In a sense Y_t is the size of the oldest family to which belongs the immortal line of descent (or the immortal individual). Notice that X_t can be estimated (for example using DNA analysis of neutral mutations) at time t whereas Y_t is not observable at time t. When X_t hits 0 or 1, that is when Y_t hits 1, one of the two oldest families disappears and there is a change of MRCA. At this time two new oldest families appear. This corresponds to a jump of the processes $\mathbf{X} = (X_t, t \in \mathbb{R})$ and $\mathbf{Y} = (Y_t, t \in \mathbb{R})$.

For the Wright-Fisher model with infinite population size, in between two jumps the process $\mathbf{X}$ is a Wright-Fisher (WF) diffusion on $[0, 1]$: $dX_t = \sqrt{X_t(1 - X_t)}dB_t$, where B is a standard Brownian motion. The two absorbing states 0 and 1 are reached in finite time. In between two jumps the process $\mathbf{Y}$ is a WF diffusion on $[0, 1]$ conditioned not to hit 0: $dY_t = \sqrt{Y_t(1 - Y_t)}dB_t + (1 - Y_t)dt$. The WF diffusion and its conditioned version have been largely used to model allelic frequencies in a neutral two-types population,

see Ewens (2004), Huillet (2007),Durrett (2008). Using the look-down representation for the genealogy introduced by Donnelly and Kurtz (1996, 1999), we prove rigorously, see Corollary 4.1, that at a jump time the law of $\mathbf{X}$, μ_0, is the uniform distribution on $[0,1]$, and that the law of $\mathbf{Y}$, μ_1, is the size-biased distribution of μ_0, that is the beta $(2,1)$ distribution. The process $\mathbf{X}$ can be seen as a resurrected WF diffusion with resurrection distribution μ_0. It is then easy to check that μ_0 is also the invariant distribution of $\mathbf{X}$. Indeed, according to Lemma 2.1 of Collet et al. (2000) (see also the pioneer work of Ferrari et al. (1995) in a discrete setting), μ is a quasy-stationary distribution (QSD) of a process killed when it reaches a set Δ if and only if μ is the stationary distribution of the corresponding resurrected process which jumps with resurrection distribution μ when it reaches the set Δ. See Section 4.3.1 for a precise statement. Then the conclusion follows as μ_0 (resp. μ_1) is a QSD of the (resp. conditioned) WF diffusion, see Ewens (2004), Huillet (2007) and also Cattiaux et al. (2009). The only QSD distribution of the WF diffusion is the uniform distribution, see Ewens (2004), p. 161, or Huillet (2007) for an explicit computation. In Section 4.3, we check that the distribution μ_1 is the only QSD for the conditioned WF diffusion, see Proposition 4.5. A similar result is also true for the Moran model. In this case also, the QSD can be seen as the distribution of the size of one of the two oldest families. There is no such interpretation for the WF model for finite population, see Remark 4.2.

To establish Corollary 4.1, we use the look-down process, which gives a representation of the genealogy for the WF model of a population with infinite size. Following Pfaffelhuber and Wakolbinger (2006), we are also interested in the distribution of the following quantities:

- A: the birth time of the MRCA for the current population.
- $\tau \geq 0$: the time to wait before a change of MRCA happens (the hitting time of $\{0,1\}$ for $\mathbf{X}$).
- $L \in \mathbb{N}^*$: the number of living individuals which will have descendants at time τ.
- $Z \in \{0,\dots,L\}$: the number of living individuals which will become MRCA in the future.
- $Y \in (0,1)$ the relative size of the oldest family to which belongs the immortal individual.
- $X \in (0,1)$ the relative size of one of the two oldest families taken at random (with probability one half it has the immortal individual).

Recent papers give an exhaustive study of birth dates and death times of MRCA, see Pfaffelhuber and Wakolbinger (2006) and also Simon and Derrida (2006) (see also Evans and Ralph (2008) for genealogies of continuous state branching processes). In particular the birth dates of MRCA, as well as the death times of MRCA for the WF model, are distributed according to a Poisson process, see Donnelly and Kurtz (2006) and Pfaffelhuber and Wakolbinger (2006).

The distribution of (τ, L, Z) is given in Pfaffelhuber and Wakolbinger (2006). In particular, τ is an exponential random variable with mean 1. We give, see Theorem 4.1

below, the joint distribution of (A, τ, L, Z) at time t conditionally on Y_t or X_t, where t is either fixed or an MRCA death time. The study of this conditional distribution is motivated by the fact that the relative size of the current two oldest families, X_t, can be inferred from available DNA data at time t. By stationarity, for fixed t, this distribution does not depend on t. It is also the same, but for A, at the death time of an MRCA (the argument is the same as in the proof of Theorem 2 in Pfaffelhuber and Wakolbinger (2006)). This property is the analogue of the so-called PASTA (Poisson Arrivals See Time Average) property in queuing theory, see Brémaud et al. (1992) for a review on this subject.

We now state the main result of this paper. Let $(E_k, k \in \mathbb{N}^*)$ be independent exponential random variables with mean 1.We denote by $T_K = \sum_{k\geq 1} \frac{2}{k(k+1)} E_k$ and $T_T = \sum_{k\geq 2} \frac{2}{k(k+1)} E_k$. Notice that T_K has the law of the lifetime of Kingman's coalescent process.

Theorem 4.1. *At a fixed time t or at the death time of an MRCA, we have:*

i) A is independent of (Y, X, τ, L, Z), and is distributed as T_K at a fixed time and as T_T at the death time of an MRCA.

ii) Conditionally on Y, X and (τ, L, Z) are independent.

iii) Conditionally on (Y, L), τ and Z are independent.

iv) Conditionally on Y, we have $X = \varepsilon Y + (1-\varepsilon)(1-Y)$ where ε is an independent random variable such that $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = 0) = 1/2$.

v) Conditionally on Y, L is geometric with parameter $1 - Y$.

vi) Conditionally on (Y, L), $\tau = \sum_{k=L}^{\infty} \frac{2}{k(k+1)} E_k$, where $(E_k, k \in \mathbb{N}^*)$ are independent exponential random variables with mean 1 and independent of (Y, L).

vii) For $u \in [0, 1]$, and $a \geq 1$,

$$\mathbb{E}[u^Z | Y, L = a] = \begin{cases} 1 & \text{if} \quad a = 1 \\ \dfrac{u}{3}\dfrac{a+1}{a-1}\displaystyle\prod_{k=2}^{a-1}\left(1 + \frac{2u}{(k-1)(k+2)}\right) & \text{if} \quad a \geq 2 \end{cases}$$

with the convention that $\prod_{\emptyset} = 1$.

We also give the first two moments of Z in Section 4.2 conditionally on (Y, L), Y or X, see (4.18), (4.21) and (4.24), as well as its generating function conditionally on Y or on X, see Corollaries 4.3 and 4.4. Our results also give a detailed proof of the heuristic arguments of Remarks 3.2 and 7.3 in Pfaffelhuber and Wakolbinger (2006). From the conditional distribution of τ given in vi), we give its first two moments, see (4.9), and we recover the formula from Kimura and Ohta (1969a), Kimura and Ohta (1969b) of its conditional expectation and second moment, see (4.12) and (4.13). See also (4.14) and (4.15) for the first and second moment conditionally on X. Notice the Laplace transform of τ conditionally on $X = x$, given by (4.11), solves the ODE: $\mathcal{L}^X f = \lambda f$, $f(0) = f(1) = 1$,

where $\mathcal{L}$ is the generator of the WF diffusion: $\mathcal{L}^X h(x) = x(1-x)h''(x)$ in $(0,1)$. We also recover (Corollary 4.2) that τ is an exponential random variable with mean 1, see Pfaffelhuber and Wakolbinger (2006) or Donnelly and Kurtz (2006).

We shall end by a formula linking Z and τ.

Proposition 4.1. *We have for all $\lambda \geq 0$ and $a \in \mathbb{N}^*$:*

$$\mathbb{E}\left[(1+\lambda)^Z | Y, L = a\right] = \prod_{k=1}^{a} \frac{k(k+1)+2\lambda}{k(k+1)},$$

$$\mathbb{E}\left[\mathrm{e}^{-\lambda\tau} | Y, L\right] = \mathbb{E}\left[\mathrm{e}^{-\lambda T_K}\right] \mathbb{E}\left[(1+\lambda)^Z | Y, L\right].$$

In particular, we deduce that

$$\mathbb{E}[\mathrm{e}^{-\lambda\tau} | X] = \mathbb{E}[\mathrm{e}^{-\lambda T_K}]\mathbb{E}[(1+\lambda)^Z | X]. \tag{4.1}$$

Notice that we also immediately get the following relations for the first moments:

$$\mathbb{E}[\tau | Y, L] = 2 - \mathbb{E}[Z | Y, L], \tag{4.2}$$

$$\mathbb{E}[\tau^2 | Y, L] = \mathbb{E}[Z^2 | Y, L] - 5\mathbb{E}[Z | Y, L] + \frac{4\pi^2}{3} - 8, \tag{4.3}$$

using that $\mathbb{E}[T_K] = 2$ for the first equality and that $\mathbb{E}[T_K^2] = \frac{4\pi^2}{3} - 8$ for the last.

Remark 4.1. *At the MRCA death time, we have a new MRCA which is born at A in the past and will die at τ in the future. On one hand, by looking at the death time of this new MRCA, Kingman's coalescent theory implies that $A + \tau$ is distributed as T_K. As the coalescent times are independent of the structure of the coalescent tree, we get that A and τ are independent. On the other hand, there are Z living future MRCA and one new MRCA. We deduce that the new MRCA is the $Z+1$-th point in the past of the birth dates of MRCA process. This latter is a Poisson point process with intensity 1, which is a direct consequence of the look-down representation of the genealogy. Intuitively, we could think that the new MRCA date of birth, A, is distributed as the sum of $Z+1$ independent exponential random variables with mean 1. This result is false, as one can easily check by computing Laplace transform; this is because the Poisson point process of the MRCA births is not independent of Z. However, this result is partially true at least for the conditional expectation thanks to* (4.2)*. The link between the distribution of T_K and the joint distribution of τ and Z (which are independent conditionally on (Y, L)) is given by equation* (4.1)*.*

The rest of the paper is organized as follows. In Section 4.2, we state the results on the distribution of X, Y, L, τ using the look-down process and ideas of Pfaffelhuber and Wakolbinger (2006). In Section 4.3, we present the QSD of the WF diffusion and the WF diffusion conditioned not to reach 0 as the distribution of the relative size of one of the two oldest families at an MRCA change (Moran model and discrete WF are also considered). The proofs are postponed to Section 4.4.

4.2 Presentation of the main results on the conditional distribution

4.2.1 The look-down process and notations

The look-down process and the modified look-down process have been introduced by Donnelly and Kurtz (1996, 1999) to give the genealogical process associated to a diffusion model of population evolution (see also Etheridge (2000) for a detailed construction for the Fleming-Viot process). This powerful representation is now currently used

We briefly recall the definition of the modified look-down process, without taking into account any spatial motion for the individuals. Consider an infinite size population evolving forward in time. Let $E = \mathbb{R} \times \mathbb{N}^*$. Each (s, i) in E denotes the (unique) individual living at time s and level i. This level is affected according to the persistence of each individual: the higher the level is, the faster the particle will die. Let $(N_{i,j}, 0 \leq i < j)$ be independent Poisson processes with rate 1. At a jumping time t of N_{ij}, the individual $(t-, i)$ reproduces and its unique child appears at level j. At the same time every particle having level at least j is pushed one level up (see Figure 4.1). These reproduction events involving levels i and j are called look-down events (as j looks down at i).

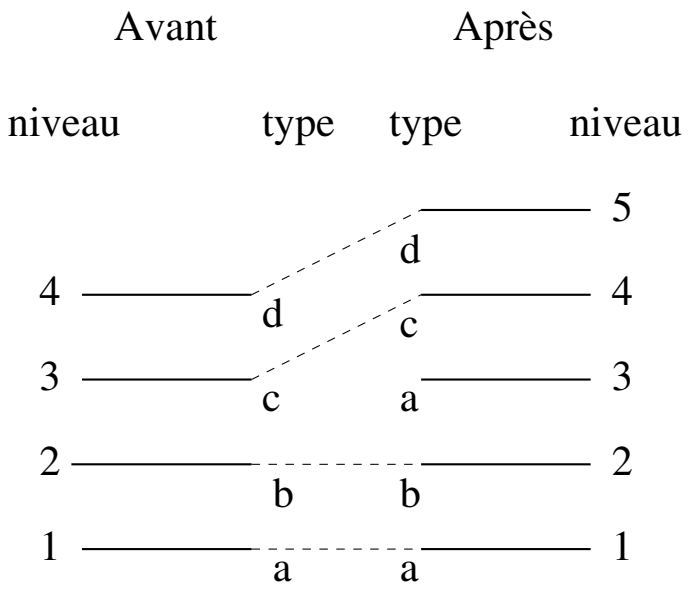

Figure 4.1: A look-down event between levels 1 and 3. Each individual living at level at least 3 before the look-down event is pushed one level up after it.

For any fixed time t_0, we can introduce the following family of equivalence relations $\mathcal{R}^{(t_0)} = (\mathcal{R}_s^{(t_0)}, s \geq 0)$: $i\mathcal{R}_s^{(t_0)}j$ if the two individuals i and j living at time t_0 have a common ancestor at time $t_0 - s$. It is then easy to show that the coalescent process on $\mathbb{N}^*$ defined by $\mathcal{R}^{(t_0)}$ is the Kingman's coalescent. See Figure 4.2 for a graphical representation.

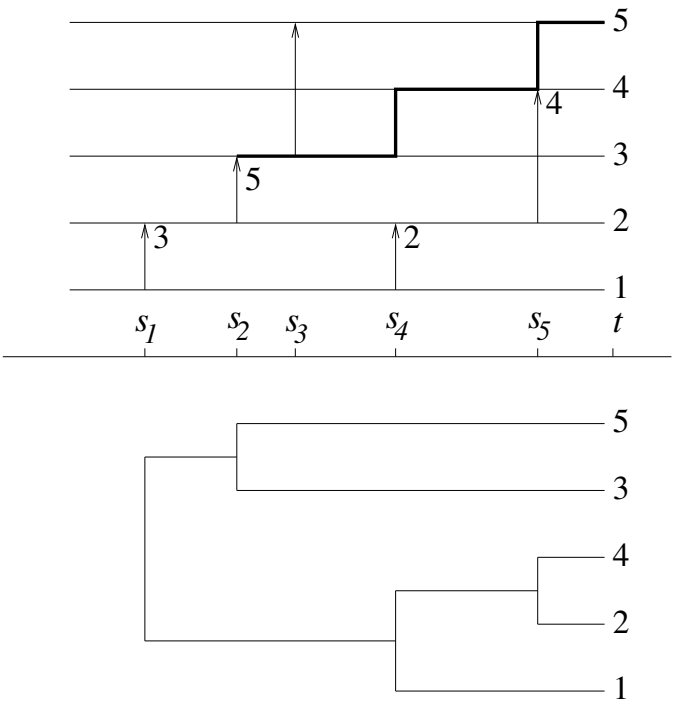

Figure 4.2: The look down process and its associated coalescent tree, started at time t for the 5 first levels. At each look-down event, a new curve is born. We indicate at which level this curve is at time t. The curve of the individual who is at level 5 at time t is bold.

We can describe the path of an individual born at level $j \geq 2$ at time s_0 as a curve in E

$$G = \bigcup_{k \in \mathbb{N}} [s_k, s_{k+1}) \times \{j + k\},$$

where for $k \in \mathbb{N}^*$, s_k is the first birth time after s_{k-1} of an individual with level less than $j + k + 1$. In particular G describes the different levels occupied by the individual born at time s_0. In fact, we shall identify an individual with its curve. We shall write $b_G = s_0$ for the birth time of individual G. We say that $d_G = \lim_{k \to \infty} s_k$ is the death time of

this individual. We say an individual or a curve is alive at time t if $b_G \leq t < d_G$, and k is the level of G at time t if $(t,k) \in G$. The set of all the curves, $\mathcal{G}$, is a partition of $E^* = \mathbb{R} \times \{2, \ldots\}$. We write $\mathcal{G}_t$ for the set of all curves alive at time t. Notice that the individual at level 1 is immortal and that by definition, its curve $\mathbb{R} \times \{1\}$ is not in $\mathcal{G}$. An individual at level j is pushed at rate $\binom{j}{2}$ to level $j+1$ (since there are $\binom{j}{2}$ possible independent look-down events which arrive at rate 1 and which push an individual living at level j). Since $\sum_{j\geq 2} 1/\binom{j}{2} < \infty$, we get that any individual but the one at level 1 dies in finite time.

In the study of MRCA, some curves will play a particular role. We say that a curve G is a fixation curve if $(b_G, 2) \in G$: the corresponding individual is born at level 2; the initial look-down event was from 2 to 1.

For a fixed time t, let G_t be the living MRCA of the whole population living at time t. Notice the birth time of the MRCA is $A_t = \inf\{b_G; G \in \mathcal{G}_t\} = b_{G_t}$. It corresponds to the birth time of the highest fixation curve living at time t. Let $Z_t + 1$ denote the number of fixation curves living at time t: $Z_t \geq 0$ is the number of future MRCA living at time t. We denote by $L_0(t) > L_1(t) > \cdots > L_{Z_t}(t)$ the decreasing levels of the fixation curves alive at time t. Notice $L(t) = L_0(t) - 1$ is the number of living individuals at time t which will have descendants at the next MRCA change. The joint distribution of $(Z_t, L_0(t), L_1(t), \ldots, L_{Z_t}(t))$ is given in Theorem 2 of Pfaffelhuber and Wakolbinger (2006), and the distribution of Z_t, the number of future MRCA alive at fixed time t, is given in Theorem 3 of Pfaffelhuber and Wakolbinger (2006). We consider the partition of the population into the two oldest families given by the equivalence relation $\mathcal{R}^{(t)}_{t-A_t}$. This corresponds to the partition of individuals alive at time t whose ancestor is either G_t or the immortal individual. We shall denote by Y_t the relative proportion of the sub-population (i.e. the oldest family) whose ancestor at time A_t is the immortal individual, that is the oldest family which contains the immortal individual. Let X_t be the relative proportion of an oldest family picked at random: with probability $1/2$ it is the one which contains the immortal individual and with probability $1/2$ the other one.

By stationarity, we have that the distribution of $H_t = (X_t, Y_t, Z_t, L(t), L_1(t), \ldots, L_{Z_t}(t))$ does not depend on t. In between two MRCA deaths, the process $(X_t, t \in \mathbb{R})$ is a Wright-Fisher diffusion with generator $\frac{1}{2}\, x(1-x)\partial_x^2$ and the process $(Y_t, t \in \mathbb{R})$ is a Wright-Fisher diffusion conditioned not to hit 0 with generator $\mathcal{L} = \frac{1}{2}\, x(1-x)\partial_x^2 + (1-x)\partial_x^1$, see Durrett (2008), Huillet (2007). Notice the distribution of Z_t conditionally on X_t is of interest, as the relative proportion of the two oldest families at time t, X_t, can be well estimated by DNA analysis if the (neutral) mutation rate is strong enough.

We are interested in the law of H_t at (random) times where the MRCA changes, as well as the distribution of the labels of the individuals of the same oldest family. The distributions of H_t is the same if we consider a fixed time t or this random time (the argument is the same as in the proof of Theorem 2 of Pfaffelhuber and Wakolbinger (2006)). This is the so-called PASTA (Poisson Arrivals See Time Average) property, see Brémaud et al. (1992) for a review on this subject, where the Poisson process considered

corresponds to the times where the MRCA changes. For this reason, we shall omit the subscript and write H, and carry out the proofs at the death time of an MRCA.

4.2.2 Size of the new two oldest families

We are interested in the description of the population, and more precisely in the relative size of the two oldest families at the time of death of an MRCA. More precisely, let G_* be a fixation curve and G be the next fixation curve: the individual G is the next MRCA after the MRCA G_*. Let $s_0 = b_{G_*}$ be the birth time of G_* and $(s_k, k \in \mathbb{N}^*)$ be the jumping times of G_*. Notice that $s_1 = b_G$ corresponds to the birth of the MRCA G. Let $N \geq 2$. Notice that at time s_{N-1}, only the individuals with level 1 to N will survive up to the death time d_G of G. They correspond to the ancestors at time s_{N-1} of the population living at time d_G. We consider the partition into 2 subsets given by $\mathcal{R}^{(s_{N-1})}_{s_{N-1}-s_0}$ which corresponds to the partition of individuals alive at time s_{N-1} with labels 1 to N whose ancestor is either G or the immortal individual. Consider the ancestor at time s_1 of the individual at level $k \in \{1, \ldots, N\}$ and time s_{N-1}, and let $\sigma_N(k) = 1$ if it is the immortal individual and $\sigma_N(k) = 0$ if it is G. Let $V_N = \sum_{k=1}^{N} \sigma_N(k)$ be the number of individuals at time s_{N-1} whose ancestor at time s_1 is the immortal individual, see Figure 4.3 for an example. Notice that $\lim_{N\to\infty} V_N/N$ will be the proportion of the oldest family which contains the immortal individual at the death time of the MRCA G_*. By construction the process $(\sigma_N, N \in \mathbb{N}^*)$ is Markov.

In order to give the law of (V_N, σ_N) we first recall some facts on Pólya's urns, see Johnson and Kotz (1977). Let $S_N^{(i,j)}$ be the number of green balls in an urn after N drawing, when initially there was i green balls and j of some other color in the urn, and where at each drawing, the chosen ball is returned together with one ball of the same color. The process $(S_N^{(i,j)}, N \in \mathbb{N})$ is a Markov chain, and for $\ell \in \{0, \ldots, N\}$

$$\mathbb{P}\left(S_N^{(i,j)} = i + \ell\right) = \binom{N}{\ell} \frac{(i+\ell-1)!(j+N-\ell-1)!(i+j-1)!}{(i-1)!(j-1)!(i+j+N-1)!}.$$

In particular, for $i = 2$, $j = 1$ and $k \in \{1, N+1\}$, we have

$$\mathbb{P}(S_N^{(2,1)} = k+1) = \frac{2k}{(N+2)(N+1)}. \tag{4.4}$$

Theorem 4.2. *Let $N \geq 2$.*

1. *The process $(1 + V_{N+2}, N \in \mathbb{N})$ is a Pólya's urn starting at $(2, 1)$. In particular, V_N has a size-biased uniform distribution on $\{1, \ldots, N-1\}$, i.e.*

$$\mathbb{P}(V_N = k) = \frac{2k}{N(N-1)}.$$

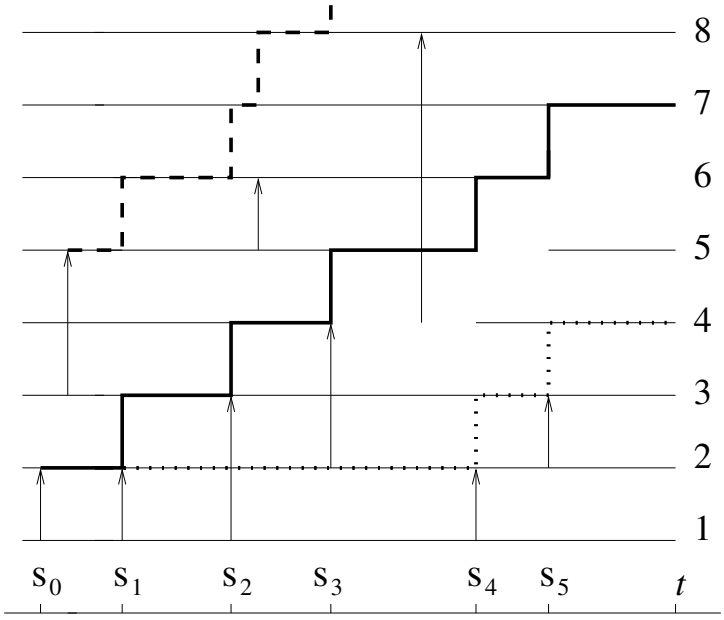

Figure 4.3: In this example, the fixation curve is in bold, and at time s_5, we have $\sigma_6 = (1,1,1,0,1,0)$ and $V_6 = 4$.

2. *Conditionally on* $(V_1, \ldots, V_N)$, σ_N *is uniformly distributed on the possible configurations:* $\{\sigma \in \{0,1\}^N; \sigma(1) = 1$ *and* $\sum_{k=1}^{N} \sigma(k) = V_N\}$.

Notice that, in general, if $N_0 \geq 3$, the process $(V_{N_0+N}, N \in \mathbb{N})$ conditionally on σ_{N_0} can not be described using Pólya's urns.

Results on Pólya's urns, see Section 6.3.3 of Johnson and Kotz (1977), give that $(V_N/N, N \in \mathbb{N}^*)$ converges a.s. to a random variable Y with a beta distribution with parameters $(2,1)$. This gives the following result.

Corollary 4.1. *When the MRCA changes, the relative proportion* Y *of the new oldest family which contains the immortal individual is distributed as a beta* $(2,1)$.

If one chooses a new oldest family at random (with probability 1/2 the one which contains the immortal individual and with probability 1/2 the other one), then its relative proportion X is uniform on $(0,1)$. This is coherent with the Remark 3.2 given in Pfaffelhuber and Wakolbinger (2006). Notice that Y has the size biased distribution of X, which corresponds to the fact that the immortal individual is taken at random from the two oldest families with probability proportional to their size.

4.2.3 Level of the next fixation curve

We keep notations from the previous section. Let $L^{(N)}+1$ be the level of the fixation curve G when the fixation curve G_* reaches level $N+1$, that is at time s_{N-1}. Notice that $L^{(N)}$ belongs to $\{1,\dots,V_N\}$. The law of $(L^{(N)},V_N)$ will be useful to give the joint distribution of (Z,Y), see Section 4.2.5. It also implies (4.6) which was already given by Lemma 7.1 of Pfaffelhuber and Wakolbinger (2006). The process $L^{(N)}$ is an inhomogeneous Markov chain, see Lemma 6.1 of Pfaffelhuber and Wakolbinger (2006). By construction, the sequence $(L^{(N)}, N\geq 2)$ is non-decreasing and converges a.s. to L defined in Section 4.2.1.

Proposition 4.2. *Let $N\geq 2$.*

i) For $1\leq i\leq k\leq N-1$, we have

$$\mathbb{P}(L^{(N)}=i,V_N=k)=2\frac{(N-i-1)!}{N!}\frac{k!}{(k-i)!}\frac{N-k}{N-1}, \tag{4.5}$$

and for all $i\in\{1,\dots,N-1\}$,

$$\mathbb{P}(L^{(N)}=i)=\frac{N+1}{N-1}\frac{2}{(i+1)(i+2)}. \tag{4.6}$$

ii) The sequence $((L^{(N)},V_N/N),N\in\mathbb{N}^*)$ converges a.s. to a random variable (L,Y), where Y has a beta $(2,1)$ distribution and conditionally on Y, L is geometric with parameter $1-Y$.

A straightforward computation gives that for $i\in\mathbb{N}^*$

$$\mathbb{P}(L=i)=\frac{2}{(i+1)(i+2)}.$$

This result was already in Proposition 3.1 of Pfaffelhuber and Wakolbinger (2006).

The level $L+1$ corresponds to the level of the MRCA, just after a change of MRCA. Recall $L_1(t)$ is the level at time t of the second fixation curve. We use the convention that $L_1(t)=1$ if there is only one fixation curve i.e. $Z(t)=0$. Just before the random time d_{G_*} of the death of the fixation curve G_*, we have $L_1(d_{G_*}-)=L_0(d_{G^*})=L+1$. At a fixed time t, by stationarity, the distribution of $L_1(t)$ does not depend on t, and equation (3.4) from Pfaffelhuber and Wakolbinger (2006) gives that $L_1(t)$ is distributed as L. In view of Remark 4.1 in Pfaffelhuber and Wakolbinger (2006), notice the result is also similar for $M/M/k$ queue where the invariant distribution for the queue process and the queue process just before arrivals time are the same. This is known as the PASTA property.

4.2.4 Next fixation time

We consider the time d_{G_*} of death of the MRCA. At this time, Y is the proportion of the oldest family which contains the immortal individuals. We denote by τ the time we have to wait for the next fixation time. It is the time needed by the highest fixation curve alive at time d_{G_*} to reach ∞. Hence, by the look-down construction, we get that

$$\tau = \sum_{k=L}^{\infty} \frac{2}{k(k+1)} E_k \tag{4.7}$$

where E_k are independent exponential random variables with parameter 1 and independent of Y and L. See also theorem 1 in Pfaffelhuber and Wakolbinger (2006).

Proposition 4.3. *Let $a \in \mathbb{N}^*$. The distribution of the waiting time for the next fixation time is given by: for $\lambda \in \mathbb{R}_+$,*

$$\mathbb{E}[\mathrm{e}^{-\lambda\tau} \,|Y, L = a] = \prod_{k=a}^{\infty} \left(\frac{k(k+1)}{k(k+1) + 2\lambda} \right). \tag{4.8}$$

Its first two moments are given by:

$$\mathbb{E}[\tau | Y, L = a] = \frac{2}{a} \quad \textit{and} \quad \mathbb{E}[\tau^2 | Y, L = a] = -\frac{8}{a} + 8 \sum_{k \geq a} \frac{1}{k^2}. \tag{4.9}$$

We also have: for $y, x \in (0,1)$ and $\lambda \in \mathbb{R}_+$,

$$\mathbb{E}[\mathrm{e}^{-\lambda\tau} \,|Y = y] = (1-y) \sum_{\ell=1}^{\infty} y^{\ell-1} \prod_{k=\ell}^{\infty} \left(\frac{k(k+1)}{k(k+1) + 2\lambda} \right). \tag{4.10}$$

$$\mathbb{E}[\mathrm{e}^{-\lambda\tau} \,|X = x] = x(1-x) \sum_{\ell=1}^{\infty} \left[x^{\ell-1} + (1-x)^{\ell-1} \right] \prod_{k=\ell}^{\infty} \left(\frac{k(k+1)}{k(k+1) + 2\lambda} \right). \tag{4.11}$$

We deduce from (4.9) that $\mathbb{E}[\tau | L = a] = \dfrac{2}{a}$, which was already in Theorem 1 in Pfaffelhuber and Wakolbinger (2006). Notice that using (4.10), we recover the following result.

Corollary 4.2. *The random variable τ is exponential with mean 1.*

Using (4.9) and the fact that L is geometric with parameter $1-Y$, we recover the well known results from Kimura and Ohta (1969a,b) (see also Ewens (2004)):

$$\mathbb{E}[\tau | Y = y] = -2 \frac{(1-y)\log(1-y)}{y}, \tag{4.12}$$

$$\mathbb{E}[\tau^2 | Y = y] = 8 \left(\frac{(1-y)\log(1-y)}{y} - \int_y^1 \frac{\log(1-z)}{z} \, dz \right). \tag{4.13}$$

The following Lemma is elementary.

Lemma 4.1. *Let Y be a beta (2,1) random variable and $X = \varepsilon Y + (1-\varepsilon)(1-Y)$ where ε is independent of Y and such that $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = 1/2$. Then X is uniform on $[0,1]$. Furthermore, if W is integrable and independent of ε, then we have $\mathbb{E}[W|X] = Xg(X) + (1-X)g(1-X)$ where $g(y) = \mathbb{E}[W|Y = y]$.*

We also get, thanks to the above Lemma that:

$$\mathbb{E}[\tau|X = x] = -2\left(x\log(x) + (1-x)\log(1-x)\right), \quad \text{and} \tag{4.14}$$

$$\mathbb{E}[\tau^2|X = x] = 8\left(x\log(x) + (1-x)\log(1-x) - x\int_x^1 \frac{\log(1-z)}{z}\,dz - (1-x)\int_{1-x}^1 \frac{\log(1-z)}{z}\,dz\right). \tag{4.15}$$

4.2.5 Number of future MRCA already living

We keep notations from Sections 4.2.1 and 4.2.3. We set $Z = Z_{d_{G_*}}$ the number of future MRCA living at time d_{G_*} of death of the MRCA G_*. Let $L_0 = L(d_{G^*}) + 1$ and $(L_0, L_1, \ldots, L_Z) = (L_0(d_{G_*}), \ldots, L_Z(d_{G_*}))$ be the levels of the fixation curves at the death time of G_*. Recall notations from Section 4.2.2. The following Lemma and Proposition 4.2 characterize the joint distribution of $(Y, Z, L, L_1, \ldots, L_Z)$.

Lemma 4.2. *Conditionally on (L, Y) the distribution of $(Z, L_1, \ldots, L_Z)$ does not depend on Y. Conditionally on $\{L = N\}$, $(Z, L_1, \ldots, L_Z)$ is distributed as follows:*

1. *$Z = 0$ if $N = 1$;*
2. *Conditionally on $\{Z \geq 1\}$, L_1 is distributed as $L^{(N)} + 1$.*
3. *For $N' \in \{1, \ldots, N-1\}$, conditionally on $\{Z \geq 1, L_1 = N'+1\}$, $(Z-1, L_2, \ldots, L_Z)$ is distributed as $(Z, L_1, \ldots, L_Z)$ conditionally on $\{L = N'\}$.*

We are now able to give the distribution of Z conditionally on Y or X.

Proposition 4.4. *Let $a \geq 1$. We have $\mathbb{P}(Z = 0|L = 1) = 1$ and for $k \geq 1$,*

$$\mathbb{P}(Z = k|Y, L = a) = \frac{2^{k-1}}{3}\frac{a+1}{a-1}\sum_{1<a_k<\cdots<a_2<a}\prod_{i=2}^{k}\frac{1}{(a_i-1)(a_i+2)}; \tag{4.16}$$

for all $u \in [0,1]$,

$$\mathbb{E}[u^Z|Y, L = a] = \begin{cases} 1 & \text{if} \quad a = 1, \\ \dfrac{u}{3}\dfrac{a+1}{a-1}\displaystyle\prod_{k=2}^{a-1}\left(1 + \frac{2u}{(k-1)(k+2)}\right) & \text{if} \quad a \geq 2, \end{cases} \tag{4.17}$$

with the convention that $\prod_{\emptyset} = 1$. *We also have*

$$\mathbb{E}[Z|Y, L = a] = 2 - \frac{2}{a} \quad \textit{and} \quad \mathbb{E}[Z^2|Y, L = a] = 18 - \frac{4\pi^2}{3} - \frac{18}{a} + 8\sum_{k\geq a} \frac{1}{k^2}\cdot \tag{4.18}$$

We deduce from ii) of Proposition 4.2 the next result.

Corollary 4.3. *Let* $y \in [0, 1]$. *We have* $\mathbb{P}(Z = 0|Y = y) = 1 - y$, *and, for all* $k \in \mathbb{N}^*$,

$$\mathbb{P}(Z = k|Y = y) = \frac{2^{k-1}}{3}(1 - y) \sum_{1<a_k<\cdots<a_1<\infty} (a_1 + 1)(a_1 + 2)y^{a_1-1} \prod_{i=1}^{k} \frac{1}{(a_i - 1)(a_i + 2)}; \tag{4.19}$$

for all $u \in [0, 1]$

$$\mathbb{E}[u^Z|Y = y] = (1 - y) + u\frac{1 - y}{3} \sum_{a=2}^{\infty} \frac{a + 1}{a - 1} y^{a-1} \prod_{\ell=2}^{a-1} \left(1 + \frac{2u}{(\ell - 1)(\ell + 2)}\right), \tag{4.20}$$

with the convention that $\prod_{\emptyset} = 1$. *We also have*

$$\mathbb{E}[Z|Y = y] = 2\left(1 + \frac{1 - y}{y} \log(1 - y)\right). \tag{4.21}$$

The next Corollary is a direct consequence of Lemma 4.1.

Corollary 4.4. *Let* $x \in [0, 1]$. *We have* $\mathbb{P}(Z = 0|X = x) = 2x(1-x)$, *and, for all* $k \in \mathbb{N}^*$,

$$\begin{aligned}&\mathbb{P}(Z = k|X = x)\\ &= \frac{2^{k-1}}{3}x(1 - x) \sum_{1<a_k<\cdots<a_1<\infty} (a_1 + 1)(a_1 + 2)\left(x^{a_1-2} + (1 - x)^{a_1-2}\right) \prod_{i=1}^{k} \frac{1}{(a_i - 1)(a_i + 2)};\end{aligned} \tag{4.22}$$

for all $u \in [0, 1]$,

$$\mathbb{E}[u^Z|X = x] = 2x(1-x)+u\frac{x(1 - x)}{3} \sum_{a=2}^{\infty} \frac{a + 1}{a - 1} \left(x^{a-1} + (1 - x)^{a-1}\right) \prod_{\ell=2}^{a-1} \left(1 + \frac{2u}{(\ell - 1)(\ell + 2)}\right), \tag{4.23}$$

with the convention that $\prod_{\emptyset} = 1$. *We also have*

$$\mathbb{E}[Z|X = x] = 2\left(1 + x \log(x) + (1 - x) \log(1 - x)\right). \tag{4.24}$$

The second moment of Z conditionally on Y (resp. X) can be deduced from (4.20) (resp. (4.23)) or from (4.3) and (4.13) (resp. (4.15)).

Some elementary computations give:

$$\begin{aligned}
\mathbb{P}(Z=0|X=x) &= 2x(1-x),\\
\mathbb{P}(Z=1|X=x) &= \frac{1}{3}\left[x^2+(1-x)^2-2x(1-x)\ln(x(1-x))\right],\\
\mathbb{P}(Z=2|X=x) &= \frac{2}{3}\left[\frac{11}{6}(x^2+(1-x)^2)-(1-x)\ln(1-x)-x\ln(x)\right]\\
&\qquad + \frac{2}{3}x(1-x)\left[2-\frac{\pi^2}{3}+2\ln(x)\ln(1-x)-\frac{1}{3}\ln(x(1-x))\right].
\end{aligned}$$

We recover by integration of the previous equations the following results from Pfaffelhuber and Wakolbinger (2006):

$$\mathbb{P}(Z=0)=\frac{1}{3},\quad \mathbb{P}(Z=1)=\frac{11}{27}\quad \text{and}\quad \mathbb{P}(Z=2)=\frac{107}{243}-\frac{2}{81}\pi^2.$$

4.3 Stationary distribution of the relative size for the two oldest families

4.3.1 Resurrected process and quasy-stationary distribution

Let E be a subset of $\mathbb{R}$. We recall that if $U=(U_t, t\geq 0)$ is an E-valued diffusion with absorbing states Δ, we say that a distribution ν is a quasy-stationary distribution (QSD) of U if for any Borel set $A\subset\mathbb{R}$,

$$\mathbb{P}_\nu(U_t\in A|U_t\notin\Delta)=\nu(A)\quad t\geq 0,$$

where we write $\mathbb{P}_\nu$ when the distribution of U_0 is ν. See also Steinsaltz and Evans (2007) for QSD for diffusions with killing.

Let μ and ν be two distributions on $E\backslash\Delta$. We define U^μ the resurrected process associated to U, with resurrection distribution μ, under $\mathbb{P}_\nu$ as follows:

1. U_0 is distributed according to ν and $U^\mu_t=U_t$ for $t\in[0,\tau_1)$, where $\tau_1=\inf\{s\geq 0; U_s\in\Delta\}$.
2. Conditionally on $(\tau_1,\{\tau_1<\infty\},(U^\mu_t, t\in[0,\tau_1)))$, $(U^\mu_{t+\tau_1}, t\geq 0)$ is distributed as U^μ under $\mathbb{P}_\mu$.

According to Lemma 2.1 of Collet et al. (2000), the distribution μ is a QSD of U if and only if μ is a stationary distribution of U^μ. See also the pioneer work of Ferrari et al. (1995) in a discrete setting.

The uniqueness of quasy-stationary distributions is an open question in general. We will give a genealogical representation of the QSD for the Wright-Fisher diffusion and the Wright-Fisher diffusion conditioned not to hit 0, as well as for the Moran model for the discrete case.

We also recall that the so-called Yaglom limit μ is defined by

$$\lim_{t\to\infty} \mathbb{P}_x(U_t \in A | U_t \notin \Delta) = \mu(A) \quad \forall A \in \mathcal{B}(\mathbb{R}),$$

provided the limit exists and is independent of $x \in E\backslash\Delta$.

4.3.2 The resurrected Wright-Fisher diffusion

From Corollary 4.1 and comments below it, we get that the relative proportion of one of the two oldest families at a change of MRCA is distributed according to the uniform distribution over $[0, 1]$. Then the relative proportion evolves according to a Wright-Fisher (WF) diffusion with generator $\frac{1}{2}\, x(1-x)\partial_x^2$. In particular it hits the absorbing state of the WF diffusion, $\{0, 1\}$, in finite time. At this time one of the two oldest families dies out and there is (again) a change of MRCA.

The QSD distribution of the WF diffusion exists and is the uniform distribution, see Ewens (2004), p. 161, or Huillet (2007) for an explicit computation. From Section 4.3.1, we get that in stationary regime, for fixed t (and of course at time when the MRCA changes) the relative size of one of the two oldest families taken at random, X_t, is uniform over $(0, 1)$.

Similar arguments as those developed in the proof of Proposition 4.5 yield that the uniform distribution is the only QSD of the WF diffusion. Lemma 2.1 in Collet et al. (2000) implies there is no other resurrection distribution which is also the stationary distribution of the resurrected process.

4.3.3 The oldest family with the immortal individual

Recall that $Y = (Y_t, t \in \mathbb{R})$ is the process of relative size for the oldest family containing the immortal individual. From Corollary 4.1, we get that Y at a change of MRCA is distributed according to the beta $(2, 1)$ distribution. Then Y evolves according to a WF diffusion conditioned not to hit 0; its generator is given by $\mathcal{L} = \frac{1}{2}\, x(1-x)\partial_x^2 + (1-x)\partial_x$, see Durrett (2008), Huillet (2007). Therefore Y is a resurrected Wright-Fisher diffusion conditioned not to hit 0, with beta $(2, 1)$ resurrection distribution.

The Yaglom distribution of the Wright-Fisher diffusion conditioned not to hit 0 exists and is the beta $(2, 1)$ distribution, see Huillet (2007) for an explicit computation. In fact the Yaglom distribution is the only QSD according to the next proposition.

Proposition 4.5. *The only quasy-stationary distribution of the Wright-Fisher diffusion conditioned not to hit* 0 *is the beta* $(2, 1)$ *distribution.*

Lemma 2.1 in Collet et al. (2000) implies that the beta $(2,1)$ distribution is therefore the stationary distribution of Y. Furthermore, the resurrected Wright-Fisher diffusion conditioned not to hit 0, with resurrection distribution μ has stationary distribution μ if and only if μ is the beta $(2,1)$ distribution.

4.3.4 Resurrected process in the Moran model

The Moran model has been introduced in Moran (1958). This mathematical model represents the neutral evolution of a haploid population of fixed size, say N. Each individual gives, at rate 1, birth to a child, which replaces an individual taken at random among the N individuals. Notice the population size is constant. Let ξ_t denote the size of the descendants at time t of a given initial group. The process $\xi = (\xi_t, t \geq 0)$ goes from state k to state $k+\varepsilon$, where $\varepsilon \in \{-1,1\}$, at rate $k(N-k)/N$. Notice that 0 and N are absorbing states. They correspond respectively to the extinction of the descendants of the initial group or its fixation. The Yaglom distribution of the process ξ is uniform over $\{1,\ldots,N-1\}$ (see Ewens (2004), p. 106). Since the state is finite, the Yaglom distribution is the only QSD.

Let μ be a distribution on $\{1,\ldots,N-1\}$. We consider the resurrected process $(\xi^\mu_t, t \geq 0)$ with resurrection distribution μ. The resurrected process has the same evolution as ξ until it reaches 0 or N, and it immediately jumps according to μ when it hits 0 or N. The process ξ^μ is a continuous time Markov process on $\{1,\ldots,N-1\}$ with transition rates matrix Λ^μ given by:

$$\begin{aligned}
\Lambda^\mu(1,k) &= \left(\mu(k) + \mathbf{1}_{\{k=2\}}\right)\frac{N-1}{N} \quad \text{for } k \in \{2,\ldots,N-1\},\\
\Lambda^\mu(k,k+\varepsilon) &= \frac{k(N-k)}{N} \quad \text{for } \varepsilon \in \{-1,1\} \text{ and } k \in \{2,\ldots,N-2\},\\
\Lambda^\mu(N-1,k) &= \left(\mu(k) + \mathbf{1}_{\{k=N-2\}}\right)\frac{N-1}{N} \quad \text{for } k \in \{1,\ldots,N-2\}.
\end{aligned}$$

We deduce from Ferrari et al. (1995), that μ is a stationary distribution for ξ^μ (i.e. $\mu\Lambda^\mu = 0$) if and only if μ is a QSD for ξ, hence if and only if μ is uniform over $\{1,\ldots,N-1\}$.

Using the genealogy of the Moran model, we can give a natural representation of the resurrected process ξ^μ when the resurrection distribution is the Yaglom distribution. Since the genealogy of the Moran model can be described by the restriction of the look-down process to $E^{(N)} = \mathbb{R}\times\{1,\ldots,N\}$, we get from Theorem 4.2 that the size of the oldest family which contains the immortal individual is distributed as the size-biased uniform distribution on $\{1,\ldots,N-1\}$ when there is a change of MRCA. The PASTA property also implies that this is the stationary distribution. If, when there is a change of MRCA, we consider at random one of the two oldest families (with probability $1/2$ the one with the immortal individual and with probability $1/2$ the other one), then the size process is distributed as $(\xi^\mu_t, t \in \mathbb{R})$ under its stationary distribution, with μ the uniform distribution.

Remark 4.2. *We can also consider the Wright-Fisher model (see e.g. Durrett (2008)) in discrete time with a population of fixed finite size N, $\zeta = (\zeta_k, k \in \mathbb{N})$. This is a Markov chain with state space $\{0, \dots, N\}$ and transition probabilities*

$$P(i,j) = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}.$$

There exists a unique quasy-stationary distribution, μ_N (which is not the uniform distribution), see Darroch and Seneta (1965). We deduce that the resurrected process ζ^μ has stationary distribution μ if and only if $\mu = \mu_N$. Notice, that in this example there is no biological interpretation of μ_N as the size of one of the oldest family at a change of MRCA.

4.4 Proofs

4.4.1 Proof of Theorem 4.2

We consider the set

$$A_N = \{(k_1, \dots, k_N); k_1 = 1, \text{ for } i \in \{1, \dots, N-1\}, k_{i+1} \in \{k_i, k_i + 1\}\}.$$

Notice that $\mathbb{P}(V_1 = k_1, \dots, V_N = k_N) > 0$ if and only if $(k_1, \dots, k_N) \in A_N$. To prove the first part of Theorem 4.2, it is enough to show that, for $N \geq 2$ and $(k_1, \dots, k_{N+1}) \in A_{N+1}$,

$$\mathbb{P}(V_{N+1} = k_{N+1} | V_N = k_N, \dots, V_1 = k_1) = \begin{cases} 1 - \frac{1+k_N}{N+1} & \text{if } k_{N+1} = k_N, \\ \frac{1+k_N}{N+1} & \text{if } k_{N+1} = 1 + k_N. \end{cases} \tag{4.25}$$

For p and q in $\mathbb{N}^*$ such that $q < p$, we introduce the set:

$$\Delta_{p,q} = \{a = (a_1, \dots, a_p) \in \{0,1\}^p, a_1 = 1, \sum_{i=1}^{p} a_i = q\}.$$

Notice that Card $(\Delta_{p,q}) = \binom{p-1}{q-1}$. Hence to prove the second part of Theorem 4.2, it is enough to show that: for all $(k_1, \dots, k_N) \in A_N$, and all $a \in \Delta_{N,k_N}$,

$$\mathbb{P}(\sigma_N = a | V_N = k_N, \dots, V_1 = k_1) = \frac{1}{\binom{N-1}{k_N - 1}}. \tag{4.26}$$

We proceed by induction on N for the proof of (4.25) and (4.26). The result is obvious for $N = 2$. We suppose (4.25) and (4.26) are true for a fixed N. We denote by I_N and J_N, $1 \leq I_N < J_N \leq N + 1$, the two levels involved in the look-down event at time s_N.

Notice that (I_N, J_N) and σ_N are independent. This pair is chosen uniformly so that, for $1 \leq i < j \leq N+1$,

$$\begin{aligned}
\mathbb{P}(I_N = i, J_N = j) &= \frac{2}{(N+1)N}, \\
\mathbb{P}(I_N = i) &= \frac{2(N-i+1)}{(N+1)N}, \\
\mathbb{P}(J_N = j) &= \frac{2(j-1)}{(N+1)N}.
\end{aligned}$$

For $a = \big(a(1), \dots, a(N+1)\big) \in \{0,1\}^{N+1}$ and $j \in \{1, \dots, N+1\}$, we set $a^j_\times = \big(a(1), \dots, a(j-1), a(j+1), \dots, a(N+1)\big) \in \{0,1\}^N$.

Let us fix $(k_1, \dots, k_{N+1}) \in A_{N+1}$, and $a = \big(a(1), \dots, a(N+1)\big) \in \Delta_{N+1,k_{N+1}}$. Notice that $\{\sigma_{N+1} = a\} \subset \{V_{N+1} = k_{N+1}\}$. We first compute

$$\mathbb{P}(\sigma_{N+1} = a | V_N = k_N, \dots, V_1 = k_1).$$

1st case: $k_{N+1} = k_N + 1$**.** We have:

$$\begin{aligned}
&\mathbb{P}(\sigma_{N+1} = a | V_N = k_N, \dots, V_1 = k_1) \\
&\quad = \sum_{1 \leq i < j \leq N+1} \mathbb{P}(I_N = i, J_N = j, \sigma_{N+1} = a | V_N = k_N, \dots, V_1 = k_1) \\
&\quad = \sum_{1 \leq i < j \leq N+1, a(i) = a(j) = 1} \mathbb{P}(I_N = i, J_N = j, \sigma_N = a^j_\times | V_N = k_N, \dots, V_1 = k_1) \\
&\quad = \sum_{1 \leq i < j \leq N+1, a(i) = a(j) = 1} \mathbb{P}(I_N = i, J_N = j) \mathbb{P}(\sigma_N = a^j_\times | V_N = k_N, \dots, V_1 = k_1) \\
&\quad = \sum_{1 \leq i < j \leq N+1, a(i) = a(j) = 1} \frac{2}{(N+1)N} \frac{1}{\binom{N-1}{k_N - 1}} \\
&\quad = \frac{2}{(N+1)N} \frac{1}{\binom{N-1}{k_N-1}} \frac{k_{N+1}(k_{N+1}-1)}{2} \\
&\quad = \frac{(k_N+1)!(N-k_N)!}{(N+1)!},
\end{aligned} \tag{4.27}$$

where we used the independence of (I_N, J_N) and σ_N for the third equality, the uniform distribution of σ_N conditionally on V_N for the fourth, and that $k_{N+1} = k_N + 1$ for the

sixth. Hence, we get

$$\begin{aligned}\mathbb{P}(V_{N+1}=k_N+1|V_N=k_N,\dots,V_1=k_1)=&\sum_{a\in\Delta_{N+1,k_{N+1}}}\mathbb{P}(\sigma_{N+1}=a|V_N=k_N,\dots,V_1=k_1)\\ =&\binom{N}{k_{N+1}-1}\frac{(k_N+1)!(N-k_N)!}{(N+1)!}\\ =&\frac{1+k_N}{N+1}.\end{aligned}\tag{4.28}$$

2nd case: $k_{N+1}=k_N$**.** Similarly, we have:

$$\begin{aligned}\mathbb{P}(\sigma_{N+1}=a|V_N=k_N,\dots,V_1=k_1)=&\sum_{1\le i<j\le N+1,a(i)=a(j)=0}\frac{2}{(N+1)N}\,\frac{1}{\binom{N-1}{k_N-1}}\\ =&\frac{2}{(N+1)N}\,\frac{1}{\binom{N-1}{k_N-1}}\,\frac{(N+1-k_N)(N-k_N)}{2}\\ =&\frac{(N-k_N)(k_N-1)!(N-k_N+1)!}{(N+1)!}.\end{aligned}\tag{4.29}$$

Hence, we get

$$\begin{aligned}\mathbb{P}(V_{N+1}=k_N|V_N=k_N,\dots,V_1=k_1)=&\sum_{a\in\Delta_{N+1,k_{N+1}}}\mathbb{P}(\sigma_{N+1}=a|V_N=k_N,\dots,V_1=k_1)\\ =&\binom{N}{k_{N+1}-1}\frac{(N-k_N)(k_N-1)!(N-k_N+1)!}{(N+1)!}\\ =&1-\frac{1+k_N}{N+1}.\end{aligned}\tag{4.30}$$

Equalities (4.28) and (4.30) imply (4.25). Moreover, we deduce from (4.27) and (4.29) that, for $k_{N+1}\in\{k_N,k_N+1\}$,

$$\begin{aligned}\mathbb{P}(\sigma_{N+1}=a|V_{N+1}=k_{N+1},\dots,V_1=k_1)=&\frac{\mathbb{P}(\sigma_{N+1}=a,V_{N+1}=k_{N+1}|V_N=k_N,\dots,V_1=k_1)}{\mathbb{P}(V_{N+1}=k_{N+1}|V_N=k_N,\dots,V_1=k_1)}\\ =&\frac{1}{\binom{N}{k_{N+1}-1}},\end{aligned}$$

which proves that (4.26) with N replaced by $N+1$ holds. This ends the proof.

4.4.2 Proof of Proposition 4.2

Theorem 4.2 shows that the distribution of σ_N conditionally on V_N is uniform. Then, if $V_N=k$, we can see $L^{(N)}$ as the number of draws (without replacement) we have to do

in a two-colored urn of size $N-1$ with $k-1$ black balls until we obtain a white ball. Hence, for $k \in \{1,\dots,N-1\}$ and $i \in \{1,\dots,k\}$,

$$\begin{aligned}\mathbb{P}(L^{(N)} = i | V_N = k) &= \frac{k-1}{N-1}\frac{k-2}{N-2}\cdots\frac{k-i+1}{N-i+1}\frac{N-k}{N-i} \\ &= \frac{(N-i-1)!}{(N-1)!}\frac{(k-1)!}{(k-i)!}(N-k).\end{aligned}$$

This and Theorem 4.2 give (4.5).

It is easy to prove by induction on j that for all $j \in \mathbb{N}$,

$$\sum_{k=i}^{i+j} \frac{k!}{(k-i)!} = \frac{(i+j+1)!}{j!(i+1)}. \tag{4.31}$$

Summing (4.5) over $k \in \{i,\dots,N-1\}$ gives:

$$\begin{aligned}\mathbb{P}(L^{(N)} = i) &= \frac{2(N-i-1)!}{N!(N-1)}\sum_{k=i}^{N-1}\frac{k!}{(k-i)!}(N-k) \\ &= \frac{2(N-i-1)!}{N!(N-1)}\left[(N+1)\sum_{k=i}^{N-1}\frac{k!}{(k-i)!} - \sum_{k=i}^{N-1}\frac{(k+1)!}{((k+1)-(i+1))!}\right] \\ &= \frac{2(N-i-1)!}{N!(N-1)}\left[\frac{(N+1)!}{(N-i-1)!(i+1)} - \frac{(N+1)!}{(N-i-1)!(i+2)}\right] \\ &= 2\frac{N+1}{N-1}\frac{1}{(i+1)(i+2)},\end{aligned}$$

where we used (4.31) twice in the third equality.

Since $(L^{(N)}, n \in \mathbb{N}^*)$ is non-decreasing, we deduce from Theorem 4.2 that the sequence $((L^{(N)}, V^{(N)}/N), N \in \mathbb{N}^*)$ converges a.s. to a limit (L, Y). Let $i \geq 1$ and $v \in [0,1)$. We have:

$$\begin{aligned}\mathbb{P}\left(L^{(N)} = i, \frac{V_N}{N} \leq v\right) &= \sum_{k=i}^{\lfloor Nv \rfloor}\mathbb{P}\left(L^{(N)} = i, V_N = k\right) \\ &= \sum_{k=i}^{\lfloor Nv \rfloor} 2\frac{(N-i-1)!}{N!}\frac{k!}{(k-i)!}\frac{N-k}{N-1} \\ &= \frac{2}{N}\sum_{k=i}^{\lfloor Nv \rfloor}\frac{k}{N-1}\frac{k-1}{N-2}\cdots\frac{k-i+1}{N-i}\left(1-\frac{k-1}{N-1}\right),\end{aligned}$$

which converges to $2\displaystyle\int_0^v y^i(1-y)dy$ as N goes to infinity. We deduce that $\mathbb{P}(L = i, Y \leq v) = 2\int_0^v y^i(1-y)dy$ for $i \in \mathbb{N}^*$ and $v \in [0,1)$. Thus Y has a beta $(2,1)$ distribution and conditionally on Y, L is geometric with parameter $1-Y$.

4.4.3 Proof of Proposition 4.3

The Laplace transform (4.8) comes from (4.7). To get the moments, we set $g(\lambda) = \mathbb{E}[\mathrm{e}^{-\lambda\tau}\,|Y, L = a] = \prod_{k\geq a} \frac{c_k}{c_k + 2\lambda}$, with $c_k = k(k+1)$. We get

$$g'(\lambda) = -g(\lambda) \sum_{k\geq a} \frac{2}{c_k + 2\lambda},$$

and thus

$$\mathbb{E}[\tau|Y, L = a] = -g'(0) = \sum_{k\geq a} \frac{2}{k(k+1)} = \frac{2}{a}\cdot$$

We also have

$$g''(\lambda) = g(\lambda) \sum_{k\geq a} \frac{4}{(c_k + 2\lambda)^2} + g(\lambda) \sum_{\ell,k\geq a} \frac{2}{c_k + 2\lambda} \frac{2}{c_\ell + 2\lambda}\cdot$$

Thus we get

$$\begin{aligned}
\mathbb{E}[\tau^2|Y, L = a] &= g''(0) \\
&= 4 \sum_{k\geq a} \frac{1}{k^2(k+1)^2} + 4 \sum_{\ell,k\geq a} \frac{1}{k(k+1)} \frac{1}{\ell(\ell+1)} \\
&= 8 \sum_{k\geq a} \frac{1}{k(k+1)} \sum_{\ell\geq k} \frac{1}{\ell(\ell+1)} \\
&= 8 \sum_{k\geq a} \frac{1}{k^2(k+1)} \\
&= 8 \sum_{k\geq a} \frac{1}{k^2} - 8 \sum_{k\geq a} \frac{1}{k(k+1)} \\
&= 8 \sum_{k\geq a} \frac{1}{k^2} - \frac{8}{a}\cdot
\end{aligned}$$

We get (4.10) from (4.8) and Proposition 4.2. We get (4.11) from (4.10) and Lemma 4.1.

4.4.4 Proof of Corollary 4.2

We give a direct proof. We set $c_k = k(k+1)$ and $b_k = c_k - 2 = (k-1)(k+2)$. Notice that $c_k + 2\lambda = b_k + 2(1+\lambda)$. We have from (4.10)

$$\begin{aligned}
\mathbb{E}[\mathrm{e}^{-\lambda\tau}] &= \int_0^1 2y\, dy \left(\sum_{a\geq 1} (1-y) y^{a-1} \prod_{k\geq a} \frac{c_k}{c_k + 2\lambda} \right) \\
&= 2 \sum_{a\geq 1} \frac{1}{(a+1)(a+2)} \prod_{k\geq a} \frac{c_k}{c_k + 2\lambda} \\
&= 2 \sum_{a\geq 1} \frac{1}{b_a + 2(\lambda+1)} \prod_{k\geq a+1} \frac{b_k}{b_k + 2(1+\lambda)} \\
&= \frac{1}{1+\lambda} \sum_{a\geq 1} \left(1 - \frac{b_a}{b_a + 2(1+\lambda)} \right) \prod_{k\geq a+1} \frac{b_k}{b_k + 2(1+\lambda)} \\
&= \frac{1}{1+\lambda},
\end{aligned}$$

where we used for the sixth equality that $\lim_{a\to\infty} \prod_{k\geq a+1} \frac{b_k}{b_k + 2(1+\lambda)} = 1$.

4.4.5 Proof of Lemma 4.2

Let us fix $N \geq 2$. We have introduced $L^{(N)} + 1$ as the level of the fixation curve G when the fixation curve G_* reaches level $N+1$, that is at time s_{N-1}. We denote by Z_N the number of other fixation curves alive at this time, and $L_1^{(N)} > L_2^{(N)} > \cdots > L_{Z_N}^{(N)} = 2$ their levels. By construction of the fixation curves, the result given by Lemma 4.2 is straightforward for $(V_N/N, Z_N, L^{(N)}, L_1^{(N)}, L_2^{(N)}, \ldots, L_{Z_N}^{(N)})$ instead of $(Y, Z, L, L_1, \ldots, L_Z)$. Now, using similar arguments as for the proof of the second part of Proposition 4.2, we get that $\big((V_N/N, Z_N, L^{(N)}, L_1^{(N)}, L_2^{(N)}, \ldots, L_{Z_N}^{(N)}), N \geq 2\big)$ converges a.s. to $(Y, Z, L, L_1, \ldots, L_Z)$ which ends the proof.

4.4.6 Proof of Propositions 4.4

Since conditionally on (Y, L), Z does not depend on Y thanks to Lemma 4.2, it is enough to compute the quantities $\mathbb{P}(Z = k | L = a)$. Those quantities are given in Pfaffelhuber and Wakolbinger (2006), but we recall their proofs. By definition of L and Z, $\mathbb{P}(Z = 0 | L = 1) = 1$ and $\mathbb{P}(Z = 0 | L = a) = 0$ for $a \geq 2$. We suppose that $a \geq 2$. We get

$\mathbb{P}(Z = k|L = a)$ by induction on k: for $k \geq 1$,

$$\begin{aligned}\mathbb{P}(Z = k|L = a) &= \sum_{1<a_2<a} \mathbb{P}(Z = k, L_1 = a_2 + 1|L = a) \\ &= \sum_{1<a_2<a} \mathbb{P}(Z = k|L_1 = a_2 + 1, L = a)\mathbb{P}(L_1 = a_2 + 1|L = a) \\ &= \sum_{1<a_2<a} \mathbb{P}(Z = k - 1|L = a_2)\mathbb{P}(L^{(a)} = a_2) \\ &= \sum_{1<a_k<\cdots a_2<a} \mathbb{P}(L^{(a_k)} = 1)\mathbb{P}(L^{(a_{k-1})} = a_k) \cdots \mathbb{P}(L^{(a)} = a_2),\end{aligned}$$

where we have used Lemma 4.2 for the third and last equalities. Using (4.6), equation (4.16) follows.

An expansion of $\prod_{k=2}^{a-1} \left(1 + \frac{2u}{(k-1)(k+2)}\right)$ and (4.16) immediately give (4.17). The result on the first two moments (4.18) follows from (4.9) and Proposition 4.1.

4.4.7 Proof of Theorem 4.1

The proof of i) is a direct consequence of Kingman's coalescent (for fixed t) or of Tajima (1999) (for the death time of MRCA) and the fact that the coalescent times (and thus the birth time of the MRCA A) does not depend on the coalescent tree shape. This last property can be deduced from Wiuf and Donnelly (1990), Section 3, see also Donnelly and Kurtz (2006). In particular, A does not depend on (X, Y, L, Z) neither on τ which conditionally on the past depends only on the coalescent tree shape (see Section 4.2.4). Properties ii) and iv) are straightforward by construction of X. We deduce iii) from (4.7), as the exponential random variables are independent of the past before d_{G_*}. Proposition 4.2 implies v). Proposition 4.3 implies vi) and Proposition 4.4 implies vii).

The properties ii)-vi) are proved at time d_{G_*}, but arguments as in the proof of Theorem 2 in Pfaffelhuber and Wakolbinger (2006) yields that the results also holds at fixed time.

4.4.8 Proof of Proposition 4.1

We set $c_k = k(k+1)$ and $b_k = c_k - 2 = (k-1)(k+2)$. Using (4.17), we have for $a \geq 3$

$$\begin{aligned}\mathbb{E}[(1+\lambda)^Z|Y, L = a] &= \frac{1+\lambda}{3}\frac{a+1}{a-1}\prod_{k=2}^{a-1}\frac{b_k + 2(1+\lambda)}{b_k} \\ &= \frac{1+\lambda}{3}\frac{a+1}{a-1}\prod_{k=2}^{a-1}\frac{c_k + 2\lambda}{b_k} \\ &= \prod_{k=1}^{a-1}\frac{c_k + 2\lambda}{c_k}.\end{aligned}$$

This equality is also true for $a = 2$. And for $a = 1$, we have $\mathbb{E}[(1+\lambda)^Z | Y, L = a] = 1$. The conclusion is then clear from vi) of Theorem 4.1 (see also (4.8)) as $\mathbb{E}[\mathrm{e}^{-\lambda T_K}] = \prod_{k=1}^{\infty} \frac{c_k}{c_k + 2\lambda}$.

4.4.9 Proof of Proposition 4.5

Let μ_1 be the beta $(2, 1)$ distribution. Using Collet et al. (2000), it is enough to prove that μ_1 is the only probability distribution μ on $[0, 1)$ such that μ is invariant for Y^μ. Since $x \mapsto \mathbb{E}_x[\tau]$ is bounded (see (4.14)), we get that $\mathbb{E}_\mu[\tau] < \infty$. For a measure μ and a function f, we set $\langle \mu, f \rangle = \int f \, d\mu$ when this is well defined. As $\mathbb{E}_\mu[\tau] < \infty$, it is straightforward to deduce from standard results on Markov chain having one atom with finite mean return time (see e.g. Meyn and Tweedie (1993) for discrete time Markov chains) that Y^μ has a unique invariant probability measure π which is defined by $\langle \pi, f \rangle = \mathbb{E}_\mu \left[\int_0^\tau f(Y_s) \, ds \right] / \mathbb{E}_\mu[\tau]$. Hence we have

$$\mathbb{E}_\mu \left[\int_0^\tau f(Y_s) ds \right] = \mathbb{E}_\mu[\tau] \langle \pi, f \rangle. \tag{4.32}$$

Let τ_n be the n-th resurrection time (i.e. n-th hitting time of 1) after 0 of the resurrected process Y^μ: $\tau_1 = \tau$ and for $n \in \mathbb{N}^*$, $\tau_{n+1} = \inf\{t > \tau_n; Y^\mu_{t-} = 1\}$. The strong law of large numbers implies that for any real measurable bounded function f on $[0, 1)$,

$$\mathbb{P}_\mu - a.s. \quad \frac{1}{\tau_n} \int_0^{\tau_n} f(Y_s) ds \xrightarrow[n \to \infty]{} \langle \pi, f \rangle.$$

Recall $\mathcal{L}$ is the infinitesimal generator of Y. For g any C^2 function defined on $[0, 1]$, the process $M_t = g(Y_t) - \int_0^t \mathcal{L}g(Y_s) \, ds$ is a martingale. Since $|M_t| \leq \|g\|_\infty + t(\|g'\|_\infty + \|g''\|_\infty)$ and $\mathbb{E}_\mu[\tau] < \infty$, we can apply the optional stopping theorem for $(M_t, t \geq 0)$ at time τ to get that

$$g(1) - \mathbb{E}_\mu \left[\int_0^\tau \mathcal{L}g(Y_s) \, ds \right] = \langle \mu, g \rangle \cdot$$

If a C^2 function g_λ is an eigenvector with eigenvalue $-\lambda$ (with $\lambda > 0$) such that $g_\lambda(1) = 0$, we deduce from (4.32) that $\langle \mu, g_\lambda \rangle = \lambda \mathbb{E}_\mu[\tau] \langle \pi, g_\lambda \rangle$. Therefore, if the resurrection measure is the invariant measure, we get:

$$\langle \mu, g_\lambda \rangle = \lambda \mathbb{E}_\mu[\tau] \langle \mu, g_\lambda \rangle. \tag{4.33}$$

Let $(a^\lambda_n, n \geq 0)$ be defined by $a^\lambda_0 = 1$ and, for $n \geq 0$,

$$a^\lambda_{n+1} = \frac{n(n+1) - 2\lambda}{(n+1)(n+2)} a^\lambda_n.$$

The function $\sum_{n=0}^{\infty} a_n^\lambda x^n$ solves $\mathcal{L}f = -\lambda f$ on $[0,1)$. For $N \in \mathbb{N}^*$ and $\lambda = \dfrac{N(N+1)}{2}$, notice that $P_N(x) = \sum_{n=0}^{\infty} a_n^\lambda x^n$ is a polynomial function of degree N. By continuity at 1, P_N is an eigenvector of $\mathcal{L}$ with eigenvalue $-N(N+1)/2$, and such that $P_N(1) = 0$ (as $\mathcal{L}f(1) = 0$ for any C^2 function f). Notice that $P_1(x) = 1-x$, which implies that $\langle \mu, P_1 \rangle > 0$. We deduce from (4.33) that $\mathbb{E}_\mu[\tau] = 1$ and $\langle \mu, P_N \rangle = 0$ for $N \geq 2$. As $P_N(1) = 0$ for all $N \geq 1$, we can write $P_N(x) = (1-x)Q_{N-1}(x)$, where Q_{N-1} is a polynomial function of degree $N-1$. For the probability distribution $\bar{\mu}(dx) = \dfrac{1-x}{\langle \mu, P_1 \rangle}\mu(dx)$, as $\frac{\langle \mu, P_{N+1} \rangle}{\langle \mu, P_1 \rangle} = 0$, we get that:

$$\langle \bar{\mu}, Q_N \rangle = 0, \quad \text{for all } N \geq 1. \tag{4.34}$$

Since $\bar{\mu}$ is a probability distribution on $[0,1]$, it is characterized by (4.34). To conclude, we just have to check that $\bar{\mu}_1$ satisfies (4.34). In fact, we shall check that $\langle \mu_1, g_\lambda \rangle = 0$ for any C^2 function g_λ eigenvector of L with eigenvalue $-\lambda$ such that $g_\lambda(1) = 0$ and $\lambda \neq 1$. Indeed, we have

$$\begin{aligned}
-\lambda \langle \mu_1, g_\lambda \rangle &= -\lambda \int_0^1 2x g_\lambda(x) dx \\
&= \int_0^1 x^2(1-x) g_\lambda''(x) dx + \int_0^1 2x(1-x) g_\lambda'(x) dx \\
&= \left[x^2(1-x) g_\lambda'(x)\right]_0^1 - \int_0^1 (2x(1-x) - x^2) g_\lambda'(x) dx + \int_0^1 2x(1-x) g_\lambda'(x) dx \\
&= \int_0^1 x^2 g_\lambda'(x) dx \\
&= \left[x^2 g_\lambda(x)\right]_0^1 - \int_0^1 2x g_\lambda(x) dx = -\langle \mu_1, g_\lambda \rangle,
\end{aligned}$$

which implies $\langle \mu_1, g_\lambda \rangle = 0$ unless $\lambda = 1$.

Bibliographie

Brémaud, P., Kannurpatti, R., and Mazumdar, R. (1992). Event and time averages : a review. *Adv. in Appl. Probab.*, 24(2) :377–411.

Cannings, C. (1974). The latent roots of certain Markov chains arising in genetics : a new approach. I. Haploid models. *Adv. in Appl. Probab.*, 6 :260–290.

Cattiaux, P., Collet, P., Lambert, A., Martinez, S., Méléard, S., and San Martin, J. (2009). Quasi-stationarity distributions and diffusion models in population dynamics. *http ://arxiv.org/abs/math/0703781. To appear.*

Chang, J. T. (1999). Recent common ancestors of all present-day individuals. *Adv. in Appl. Probab.*, 31(4) :1002–1038.

Collet, P., Martínez, S., and Maume-Deschamps, V. (2000). On the existence of conditionally invariant probability measures in dynamical systems. *Nonlinearity*, 13(4) :1263–1274.

Darroch, J. N. and Seneta, E. (1965). On quasi-stationary distributions in absorbing discrete-time finite Markov chains. *J. Appl. Probability*, 2 :88–100.

Donnelly, P. and Kurtz, T. G. (1996). A countable representation of the Fleming-Viot measurable diffusion. *Ann. Probab.*, 24(2) :698–742.

Donnelly, P. and Kurtz, T. G. (1999). Particle representations for measure-valued population models. *Ann. Probab.*, 27(1) :166–205.

Donnelly, P. and Kurtz, T. G. (2006). The Eve process. Manuscript, personal communication.

Durrett, R. (2008). *Probability models for DNA sequence evolution.* Probability and its Applications. Springer, New York, NY. second edition.

Etheridge, A. M. (2000). *An introduction to superprocesses*, volume 20 of *University Lecture Series.* American Mathematical Society, Providence, RI.

Evans, S. N. and Ralph, P. L. (2008). Dynamics of the time to the most recent common ancestor in a large branching population. *http ://arxiv.org/abs/0812.1302. To appear.*

Ewens, W. J. (2004). *Mathematical population genetics. I. Theoretical introduction*, volume 27 of *Interdisciplinary Applied Mathematics*. Springer-Verlag, New York, NY. second edition.

Ferrari, P. A., Kesten, H., Martinez, S., and Picco, P. (1995). Existence of quasi-stationary distributions. A renewal dynamical approach. *Ann. Probab.*, 23(2) :501–521.

Fisher, R. (1930). *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford.

Fu, Y. X. (2006). Exact coalescent for the Wright-Fisher model. *Theoret. Population Biol..*, 69(3) :1385–394.

Greven, A., Pfaffelhuber, P., and Winter, A. (2009). Tree-valued resampling dynamics : martingale problems and applications. *http ://arxiv.org/abs/0806.2224. To appear.*

Griffiths, R. C. (1980). Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theoret. Population Biol.*, 17(1) :37–50.

Huillet, T. (2007). On Wright–Fisher diffusion and its relatives. *J. Stat. Mech.*, P1106.

Johnson, N. L. and Kotz, S. (1977). *Urn models and their application*. John Wiley & Sons Inc., New York, NY.

Kimura, M. and Ohta, T. (1969a). The average number of generations until extinction of an individual mutant gene in a finite populationt. *Genetics*, 61 :763–771.

Kimura, M. and Ohta, T. (1969b). The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, 63(3) :701–709.

Kingman, J. F. C. (1982). Exchangeability and the evolution of large populations. In *Exchangeability in probability and statistics (Rome, 1981)*, pages 97–112. North-Holland, Amsterdam.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag, London.

Möhle, M. and Sagitov, S. (2001). A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.*, 29(4) :1547–1562.

Moran, P. A. P. (1958). Random processes in genetics. *Proc. Cambridge Philos. Soc.*, 54 :60–71.

Pfaffelhuber, P. and Wakolbinger, A. (2006). The process of most recent common ancestors in an evolving coalescent. *Stochastic Process. Appl.*, 16(12) :1836–1859.

Pitman, J. (1999). Coalescents with multiple collisions. *Ann. Probab.*, 27(4) :1870–1902.

Sagitov, S. (1999). The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.*, 36(4) :1116–1125.

Schweinsberg, J. (2000). Coalescents with simultaneous multiple collisions. *Electron. J. Probab.*, 5 :1–50.

Simon, D. and Derrida, B. (2006). Evolution of the most recent common ancestor of a population with no selection. *J. Stat. Mech.*, P05002.

Steinsaltz, D. and Evans, S. N. (2007). Quasistationary distributions for one-dimensional diffusions with killing. *Trans. Amer. Math. Soc.*, 359(3) :1285–1324.

Tajima, F. (1999). Relationship between DNA polymorphism and fixation time. *Genetics*, 56 :183–201.

Wiuf, C. and Donnelly, P. (1990). Conditional genealogies and the age of a neutral mutant. *Theoret. Population Biol.*, 125 :447–454.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16(2) :97–159.

Chapitre 5

Résultats asymptotiques sur la longueur d'arbres de coalescence

Version non modifiée de l'article *Asymptotic results on the length of coalescent trees.*
publié en 2008 dans
The Annals of Applied Probability. 18(3) :997-1025.

5.1 Introduction

5.1.1 Motivations

The Kingman coalescent, see Kingman (1982, 2000), allows to describe the genealogy of n individuals in a Wright-Fisher model, when the size of the whole population is very large and time is well rescaled. In what follows, we consider only neutral DNA mutations and the infinite sites model introduced by Kimura (1969), where each mutation occurs at a new site. In particular if an individual is affected by a mutation, all the descendants of this individual carry this mutation. Notice the total number of mutations observed among n individuals alive today, $S^{(n)}$, corresponds to the number of segregating sites. The Watterson estimator (Watterson (1975)) based on $S^{(n)}$ allows to estimate the rate of mutation for the DNA, θ. This estimator is consistent and converges at rate $1/\sqrt{\log(n)}$.

Other models of population where one individual can produce a large number of children give rise to more general coalescent processes than the Kingman coalescent, where multiple collisions appear, see Sagitov (1999) and Schweinsberg (2003) (such models may be relevant for oysters and some fish species (Boom et al. (1994), Eldon and Wakeley (2006))). In Birkner et al. (2005) and in Schweinsberg (2003) a natural family of one parameter coalescent processes arise to describe the genealogy of such populations: the Beta-$(2-\alpha, \alpha)$ coalescent with parameter $\alpha \in (1, 2)$. Results from Berestycki et al. (2008) give a consistent estimator, based on the observed total number, $S^{(n)}$, of mutations for the rate θ of mutation of DNA. This paper is a first step to study the convergence rate of this estimator or equivalently to the study of the asymptotic distribution of $S^{(n)}$. Results are also known for the asymptotic distribution of $S^{(n)}$ for other coalescent processes, see Drmota et al. (2007) and Möhle (2006).

For the Beta coalescent, the asymptotic distribution of $S^{(n)}$ depends on θ but also on the parameter α. In particular, if the mutation rate of the DNA is known, the asymptotic distribution of $S^{(n)}$ allows to deduce an estimation and a confidence interval for α, which in a sense characterize the size of a typical family according to Schweinsberg (2003).

5.1.2 The coalescent tree and mutation rate

We denote by $\mathbb{N}^*$ the set of positive integers. We consider at time $t = 0$ a number $n \in \mathbb{N}^*$ of individuals, and we look backward in time. Let $\mathcal{P}_n$ be the set of partitions of $\{1, \ldots, n\}$. For $t \geq 0$, let $\Pi_t^{(n)}$ be an element of $\mathcal{P}_n$ such that each block of $\Pi_t^{(n)}$ corresponds to the initial individuals which have a common ancestor at time $-t$. We assume that if we consider b blocks, k of them merge into 1 at rate $\lambda_{b,k}$, independently of the current number of blocks. Using this property and the compatibility relation implied when one consider a larger number of initial individuals, Pitman (1999), see also Sagitov (1999) for

a more biological approach, showed the transition rates are given by

$$\lambda_{b,k} = \int_{[0,1]} x^{k-2}(1-x)^{b-k}\Lambda(dx), \quad 2 \leq k \leq b,$$

for some finite measure Λ on $[0,1]$, and that $\Pi^{(n)}$ is the restriction of the so-called coalescent process defined on the set of partitions of $\mathbb{N}^*$. The Kingman coalescent corresponds to the case where Λ is the Dirac mass at 0, see Kingman (1982). In particular, in the Kingman coalescent, only two blocks merge at a time. The Bolthausen-Sznitman coalescent (Bolthausen and Sznitman (1998)) corresponds to the case where Λ is the Lebesgue measure on $[0,1]$. The Beta-coalescent introduced in Birkner et al. (2005) and in Schweinsberg (2003), see also Bertoin and Le Gall (2006) and Berestycki et al. (2007) , corresponds to $\Lambda(dx) = C_0 x^{\alpha-1}(1-x)^{1-\alpha}\mathbf{1}_{(0,1)}(x)\,dx$ for some constant $C_0 > 0$.

Notice $\Pi^{(n)} = (\Pi_t^{(n)}, t \geq 0)$ is a Markov process starting at the trivial partition of $\{1,\ldots,n\}$ into n singletons. We denote by $R_t^{(n)}$ the number of blocks of $\Pi_t^{(n)}$. We have, $R_0^{(n)} = n$, and $R_t^{(n)}$ can be seen as the number of ancestors alive at time $-t$. The apparition time of the most recent common ancestor (MRCA) is $\inf\{t > 0; R_t^{(n)} = 1\}$. We shall omit the superscript (n) when there is no confusion. The process $R = (R_t, t \geq 0)$ is a continuous time Markov process taking values in $\mathbb{N}^*$. The number of possible choices of $\ell+1$ blocks among k is $\binom{k}{\ell+1}$ (for $1 \leq \ell \leq k-1$) and each group of $\ell+1$ blocks merge at rate $\lambda_{k,\ell+1}$. So the waiting time of R in state k is an exponential random variable with parameter

$$g_k = \sum_{\ell=1}^{k-1} \binom{k}{\ell+1} \lambda_{k,\ell+1} = \int_{(0,1)} \Big(1-(1-x)^k - kx(1-x)^{k-1}\Big)\frac{\Lambda(dx)}{x^2} \tag{5.1}$$

and is distributed as E/g_k, where E is an exponential random variable with mean 1.

Let $Y = (Y_k, k \geq 1)$ be the different states of the process R. It is defined by $Y_0 = R_0$ and for $k \geq 1$, $Y_k = R_{T_k}$, where the sequence of jumping time $(T_k, k \geq 0)$ is defined inductively by $T_0 = 0$ and for $k \geq 1$, $T_k = \inf\{t > T_{k-1}; R_t \neq R_{T_{k-1}}\}$. We use the convention that $\inf\emptyset = +\infty$ and $Y_k = 1$ for $k \geq \tau_n$, where $\tau_n = \inf\{k; R_{T_k} = 1\}$ is the number of jumps of the process R until it reach the absorbing state 1. The number τ_n is the number of coalescences.

We shall write $Y^{(n)}$ instead of Y when it will be convenient to stress that Y starts at time 0 at point n. Notice Y is an $\mathbb{N}^*$-valued discrete time Markov chain, with probability transition

$$P(k, k-\ell) = \frac{\binom{k}{\ell+1}\lambda_{k,\ell+1}}{g_k}. \tag{5.2}$$

The sum of the lengths of all branches in the coalescent tree until the MRCA is distributed as

$$L^{(n)} = \sum_{k=0}^{\tau_n - 1} \frac{Y_k^{(n)}}{g_{Y_k^{(n)}}} E_k,$$

where $(E_k, k \geq 0)$ are independent exponential random variables with expectation 1.

In the infinite sites model, one assumes that (neutral) mutations appear in the genealogy at random with rate θ. In particular, conditionally on the length of the coalescent tree $L^{(n)}$, the total number $S^{(n)}$ of mutations is distributed according to a Poisson r.v. with parameter $\theta L^{(n)}$. Therefore, we have that $\dfrac{S^{(n)} - \theta L^{(n)}}{\sqrt{\theta L^{(n)}}}$ converges in distribution to a standard Gaussian r.v. (with mean 0 and variance 1). If the asymptotic distribution of $L^{(n)}$ is known, one can deduce the asymptotic distribution of $S^{(n)}$.

5.1.3 Known results

Kingman coalescence

For Kingman coalescence, a coalescence corresponds to the apparition of a common ancestor of only two individuals. In particular, we have for $0 \leq k \leq n-1$, $Y_k^{(n)} = n-k$. Thus we get $\tau_n = n-1$ as well as $g_{Y_k^{(n)}} = (n-k)(n-k-1)/2$. We also have $\dfrac{L^{(n)}}{2} = \displaystyle\sum_{k=0}^{n-2} \frac{1}{n-k-1} E_k = \sum_{k=1}^{n-1} \frac{1}{k} E_{n-k-1}$. The r.v. $L^{(n)}/2$ is distributed as the sum of independent exponential r.v. with parameter 1 to $n-1$, that is as the maximum on $n-1$ independent exponential r.v. with mean 1, see Feller (1971) section I.6. An easy computation gives that $L^{(n)}/(2\log(n))$ converges in probability to 1 and that $\dfrac{L^{(n)}}{2} - \log(n)$ converges in distribution to the Gumbel distribution (with density $\mathrm{e}^{-x-\exp -x}$) when n goes to infinity. It is then easy to deduce that $\dfrac{S^{(n)} - \theta\mathbb{E}[L^{(n)}]}{\sqrt{\theta\mathbb{E}[L^{(n)}]}}$ converges in distribution to the standard Gaussian distribution. This provides the weak convergence and the asymptotic normality of the Watterson estimator (Watterson (1975)) of θ: $\dfrac{S^{(n)}}{\mathbb{E}[L^{(n)}]} = \dfrac{S^{(n)}}{\sum_{k=1}^{n-1} \frac{1}{k}}$. See also the appendix in Drmota et al. (2007).

Bolthausen-Sznitman coalescence

In Drmota et al. (2007), the authors consider the Bolthausen-Sznitman coalescence: Λ is the Lebesgue measure on $[0,1]$. In this case they prove that $\dfrac{1}{n}\log(n)L^{(n)}$ converges in probability to 1 and that $\dfrac{L^{(n)} - a_n}{b_n}$ converges in distribution to a stable r.v. Z with Laplace transform $\mathbb{E}[\mathrm{e}^{-\lambda Z}] = \mathrm{e}^{\lambda\log(\lambda)}$ for $\lambda > 0$, where

$$a_n = \frac{n}{\log(n)} + \frac{n\log(\log(n))}{\log(n)^2} \quad \text{and} \quad b_n = \frac{n}{\log(n)^2}.$$

It is then easy to deduce that $\frac{S^{(n)} - \theta a_n}{\theta b_n}$ converges to Z.

The case $\int_{(0,1]} x^{-1}\Lambda(dx) < \infty$

In Möhle (2006), the author investigates the case where $x^{-1}\Lambda(dx)$ is a finite measure and consider directly the asymptotic distribution of $S^{(n)}$. In particular he gets that $S^{(n)}/n\theta$ converges in distribution to a non-negative r.v. Z uniquely determined by its moments: for $k \geq 1$,

$$\mathbb{E}[Z^k] = \frac{k!}{\prod_{i=1}^{k} \Phi(i)}, \quad \text{with} \quad \Phi(i) = \int_{[0,1]} (1-(1-x)^i) x^{-2} \Lambda(dx).$$

There is an equation in law for Z when Λ is a simple measure, that is when $\int_{(0,1]} x^{-2}\Lambda(dx) < \infty$.

Beta coalescent

The Beta-$(2-\alpha, \alpha)$ coalescent corresponds to the case where Λ is the Beta$(2-\alpha, \alpha)$ distribution, with $\alpha \in (1,2)$: $\Lambda(dx) = \frac{1}{\Gamma(2-\alpha)\Gamma(\alpha)} x^{1-\alpha}(1-x)^{\alpha-1}dx$. The Kingman coalescent can be viewed as the asymptotic case $\alpha = 2$ and the Bolthausen-Sznitman coalescence as the asymptotic case $\alpha = 1$.

The first order asymptotic behavior of $L^{(n)}$ is given in Berestycki et al. (2008), theorem 1.9: $n^{\alpha-2}L^{(n)}$ converges in probability to $\frac{\Gamma(\alpha)\alpha(\alpha-1)}{2-\alpha}$. We shall now investigate the asymptotic distribution of $L^{(n)}$.

5.1.4 Main result

In this paper we shall state a partial result concerning the asymptotic distribution of $L^{(n)}$. We shall only give the asymptotic distribution of the total length of the coalescent tree up to the $\lfloor nt \rfloor$-th coalescence:

$$L_t^{(n)} = \sum_{k=0}^{\lfloor nt \rfloor \wedge (\tau_n - 1)} \frac{Y_k^{(n)}}{g_{Y_k^{(n)}}} E_k, \tag{5.3}$$

where $\lfloor x \rfloor$ is the largest integer smaller or equal to x for $x \geq 0$.

We say $g = O(f)$, where f is a non-negative function and g a real valued function defined on a set E (mainly here $E = [0,1]$ or $E = \mathbb{N}^*$ or $E = \mathbb{N}^* \times [0,1]$), if there exists a finite constant $C > 0$ such that $|g(x)| \leq Cf(x)$ for all $x \in E$.

Let $\nu(dx) = x^{-2}\Lambda(dx)$ and $\rho(t) = \nu((t,1])$. We assume that $\rho(t) = C_0 t^{-\alpha} + O(t^{-\alpha+\zeta})$ for some $\alpha \in (1,2)$, $C_0 > 0$ and $\zeta > 1 - 1/\alpha$. This includes the Beta$(2-\alpha, \alpha)$ distribution for Λ. We have, see Lemma 5.2, that

$$g_n = C_0\Gamma(2-\alpha)n^\alpha + O(n^{\alpha-\min(\zeta,1)}).$$

Let $\gamma = \alpha - 1$. Let $V = (V_t, t \geq 0)$ be a α-stable Lévy process with no positive jumps (see chap. VII in Bertoin (1996)) with Laplace exponent $\psi(u) = u^\alpha/\gamma$: for all $u \geq 0$, $\mathbb{E}[\mathrm{e}^{-uV_t}] = \mathrm{e}^{tu^\alpha/\gamma}$.

We first give in Proposition 5.1 the asymptotic for the number of coalescences, τ_n:

$$n^{-\frac{1}{\alpha}}\left(n - \frac{\tau_n}{\gamma}\right) \xrightarrow[n\to\infty]{\text{(d)}} V_\gamma.$$

See also Gnedin and Yakubovich (2007) and Iksanov and Möhle (2008) for different proofs of this results under slightly different or stronger hypothesis. Then we give the asymptotics of $\hat{L}_t^{(n)}$ defined as $C_0\Gamma(2-\alpha)L_t^{(n)}$ but for the exponential r.v. E_k which are replaced by their mean that is 1 and for $g_{Y_k^{(n)}}$ which is replaced by its equivalent $C_0\Gamma(2-\alpha)\left(Y_k^{(n)}\right)^{2-\alpha}$:

$$\hat{L}_t^{(n)} = \sum_{k=0}^{\lfloor nt\rfloor \wedge (\tau_n - 1)} \left(Y_k^{(n)}\right)^{1-\alpha}. \tag{5.4}$$

For $t \in [0,\gamma]$, we set

$$v(t) = \int_0^t \left(1 - \frac{r}{\gamma}\right)^{-\gamma} dr.$$

Theorem 5.1 gives that the following convergence in distribution holds for all $t \in (0,\gamma)$

$$n^{-1+\alpha-1/\alpha}(\hat{L}_t^{(n)} - n^{2-\alpha}v(t)) \xrightarrow[n\to\infty]{\text{(d)}} (\alpha-1)\int_0^t dr\ (1-\frac{r}{\gamma})^{-\alpha}V_r. \tag{5.5}$$

Then we deduce our main result, Theorem 5.2. Let $\alpha \in (1, \frac{1+\sqrt{5}}{2})$. Then for all $t \in (0,\gamma)$, we have the following convergence in distribution

$$n^{-1+\alpha-1/\alpha}\left(L_t^{(n)} - n^{2-\alpha}\frac{v(t)}{C_0\Gamma(2-\alpha)}\right) \xrightarrow[n\to\infty]{\text{(d)}} \frac{\alpha-1}{C_0\Gamma(2-\alpha)}\int_0^t dr\ (1-\frac{r}{\gamma})^{-\alpha}V_r. \tag{5.6}$$

We also have that $n^{\alpha-2}L_t^{(n)}$ converges in probability to $\dfrac{v(t)}{C_0\Gamma(2-\alpha)}$ for $\alpha \in (1,2)$ uniformly on $[0,t_0]$ for any $t_0 \in [0,\gamma)$. See also analogous results in Basdevant and Goldschmidt (2008) for the Bolthausen-Sznitman coalescent. For $t = \gamma$, intuitively we have $L_\gamma^{(n)}$ close to $L^{(n)}$ as τ_n is close to n/γ. In particular, one expects that $n^{\alpha-2}L^{(n)}$ converges in probability

to $\dfrac{v(\gamma)}{C_0\Gamma(2-\alpha)}$. For the Beta-coalescent, $\Lambda(dx) = \dfrac{1}{\Gamma(2-\alpha)\Gamma(\alpha)}\, x^{1-\alpha}(1-x)^{\alpha-1}dx$, we have $C_0 = 1/\alpha\Gamma(2-\alpha)\Gamma(\alpha)$ and indeed, theorem 1.9 in Berestycki et al. (2008) gives that $n^{\alpha-2}L^{(n)}$ converges in probability to $\dfrac{\Gamma(\alpha)\alpha(\alpha-1)}{2-\alpha} = \dfrac{v(\gamma)}{C_0\Gamma(2-\alpha)}$. Notice theorem 1.9 in Berestycki et al. (2008) is stated for more general coalescents than the Beta-coalescent.

In Corollary 5.2, we give the asymptotic distribution of the total number $S_t^{(n)}$ of mutations on the coalescent tree up to the $\lfloor nt \rfloor$-th coalescent for $\alpha \in (1,2)$. In particular, for $\alpha > \sqrt{2}$, the approximations of the exponential r.v. by their mean are more important than the fluctuations of $\hat{L}^{(n)}$, and the asymptotic distribution is Gaussian.

5.1.5 Organization of the paper

In Section 5.2 we give estimates (distribution, Laplace transform) for the number of individuals involved in the first coalescence in a population of n individuals. We prove the asymptotic distribution of the number of collisions, τ_n, in Section 5.3, as well as an invariance principle for the coalescent process $Y^{(n)}$, see Corollary 5.1. In Section 5.4, we give error bounds on the approximation of $L_t^{(n)}$ by $\hat{L}_t^{(n)}/C_0\Gamma(2-\alpha)$. Section 5.5 is devoted to the asymptotic distribution of $\hat{L}_t^{(n)}$. Eventually, our main result, Theorem 5.2, on the asymptotic distribution of $L_t^{(n)}$, and Corollary 5.2, on the asymptotic distribution of the number of mutations $S_t^{(n)}$, and their proofs are given in Section 5.6.

In what follows, c is a non important constant which value may vary from line to line.

5.2 Law of the first jump

Let Y be a discrete time Markov chain on $\mathbb{N}^*$ with transition kernel P given by (5.2) and started at $Y_0 = n$. Let $\mathcal{Y} = (\mathcal{Y}_k, k \geq 0)$ be the filtration generated by Y. We set $X_k^{(n)} = Y_{k-1} - Y_k$ for $k \geq 1$. We give some estimates on the moment of $X_1^{(n)}$ and its Laplace transform.

For $n \geq 1$, $x \in (0,1)$, let $B_{n,x}$ be a binomial r.v. with parameter (n,x). Recall that for $1 \leq k \leq n$, we have

$$\mathbb{P}(B_{n,x} \geq k) = \frac{n!}{(k-1)!(n-k)!} \int_0^x t^{k-1}(1-t)^{n-k}\, dt. \tag{5.7}$$

Recall that $\nu(dx) = x^{-2}\Lambda(dx)$ and $\rho(t) = \nu((t,1])$. Use the first equality in (5.1) and (5.7)

to get

$$\begin{aligned} g_n &= \int_0^1 \sum_{k=2}^{n} \binom{n}{k} x^k (1-x)^{n-k} \nu(dx) \\ &= \int_0^1 \mathbb{P}(B_{n,x} \geq 2) \nu(dx) \\ &= n(n-1) \int_0^1 (1-t)^{n-2} t\rho(t)\, dt. \end{aligned} \tag{5.8}$$

Notice also that $\mathbb{P}(X_1^{(n)} = k) = P(n, n-k) = \dfrac{1}{g_n} \displaystyle\int_0^1 \mathbb{P}(B_{n,x} = k+1)\nu(dx)$ and thus

$$\mathbb{P}(X_1^{(n)} \geq k) = \frac{\int_0^1 \mathbb{P}(B_{n,x} \geq k+1)\nu(dx)}{g_n} = \frac{(n-2)!}{k!(n-k-1)!} \frac{\int_0^1 (1-t)^{n-k-1} t^k \rho(t)\, dt}{\int_0^1 (1-t)^{n-2} t\rho(t)\, dt}. \tag{5.9}$$

Let $\alpha \in (1,2)$ and $\gamma = \alpha - 1$. The following result on the asymptotic distribution of $(X_1^{(n)}, n \geq 2)$ is essentially in Bertoin and Le Gall (2006), lemma 4.

Lemma 5.1. *Assume that $\rho(t) = t^{-\alpha} L(t)$, where $L(t), t \in (0,1]$ is slowly varying at 0. Then $(X_1^{(n)}, n \geq 2)$ converges in distribution to the r.v. X taking values in $\mathbb{N}^*$ and such that for all $k \geq 1$,*

$$\mathbb{P}(X \geq k) = \frac{1}{\Gamma(2-\alpha)} \frac{\Gamma(k+1-\alpha)}{k!}.$$

We have $\mathbb{E}[X] = 1/\gamma$, $\mathbb{E}[X^2] = +\infty$ and its Laplace transform ϕ is given by: for $u \geq 0$,

$$\phi(u) = \mathbb{E}[\mathrm{e}^{-uX}] = 1 + \frac{\mathrm{e}^u - 1}{\alpha - 1} \left[(1 - \mathrm{e}^{-u})^{\alpha - 1} - 1 \right].$$

We shall use repeatedly the identity of the Beta distribution: for $a > 0$ and $b > 0$, we have

$$\int_0^1 t^{a-1}(1-t)^{b-1} dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{5.10}$$

Proof. Following lemma 4 from Bertoin and Le Gall (2006), it is easy to get that for fixed $k \geq 1$, as n goes to infinity, we have

$$\lim_{n\to\infty} n^{k+1-\alpha} L(1/n)^{-1} \int_0^1 (1-t)^{n-k-1} t^k \rho(t)\, dt = \Gamma(k+1-\alpha).$$

Therefore, we get that, for $k \in \mathbb{N}^*$,

$$\lim_{n\to\infty} \mathbb{P}(X_1^{(n)} \geq k) = \lim_{n\to\infty} \frac{(n-2)!}{k!(n-k-1)!} \frac{\int_0^1 (1-t)^{n-k-1} t^k \rho(t)\, dt}{\int_0^1 (1-t)^{n-2} t\rho(t)\, dt} = \frac{1}{\Gamma(2-\alpha)} \frac{\Gamma(k+1-\alpha)}{k!}.$$

This ends the first part of the Lemma. Since $\mathbb{P}(X \geq k) = \dfrac{1}{\Gamma(\alpha)\Gamma(2-\alpha)} \displaystyle\int_0^1 t^{k-\alpha}(1-t)^{\alpha-1}dt$, we deduce that

$$\mathbb{E}[X] = \sum_{k\geq 1} \mathbb{P}(X \geq k) = \frac{1}{\Gamma(\alpha)\Gamma(2-\alpha)} \int_0^1 \sum_{k\geq 1} t^{k-\alpha}(1-t)^{\alpha-1}dt = \frac{1}{\alpha - 1}\,.$$

Notice that $\mathbb{P}(X = k) = \mathbb{P}(X \geq k) - \mathbb{P}(X \geq k+1)$ and thus

$$\mathbb{P}(X = k) = \frac{1}{\Gamma(\alpha)\Gamma(2-\alpha)} \int_0^1 t^{k-\alpha}(1-t)^{\alpha}dt = \frac{\alpha}{\Gamma(2-\alpha)} \frac{\Gamma(k+1-\alpha)}{(k+1)!}. \tag{5.11}$$

The asymptotic expansion

$$\Gamma(z) = \sqrt{2\pi} z^{z-1/2}\, \mathrm{e}^{-z} \left(1 + \frac{1}{12z} + o\left(\frac{1}{z}\right)\right) \tag{5.12}$$

implies $\mathbb{P}(X = k) \sim_{+\infty} \dfrac{\alpha}{\Gamma(2-\alpha)} k^{-\alpha-1}$. Therefore we have $\mathbb{E}[X^2] = +\infty$. We compute the Laplace transform of X. Let $u \geq 0$, we have

$$\begin{aligned}
\phi(u) = \mathbb{E}[\mathrm{e}^{-uX}] &= \frac{\alpha}{\Gamma(2-\alpha)} \sum_{k\geq 1} \frac{1}{(k+1)!}\, \mathrm{e}^{-ku} \int_0^\infty x^{k-\alpha}\, \mathrm{e}^{-x}\ dx \\
&= \frac{\alpha\, \mathrm{e}^{u}}{\Gamma(2-\alpha)} \int_0^\infty \sum_{k\geq 2} \frac{1}{k!}\, \mathrm{e}^{-ku}\, x^{k-1-\alpha}\, \mathrm{e}^{-x}\ dx \\
&= \frac{\alpha\, \mathrm{e}^{u}}{\Gamma(2-\alpha)} \int_0^\infty x^{-1-\alpha}\, \mathrm{e}^{-x} (\mathrm{e}^{x\,\mathrm{e}^{-u}} - x\,\mathrm{e}^{-u} - 1)\ dx \\
&= 1 + \frac{\mathrm{e}^{u} - 1}{\alpha - 1} \left[(1 - \mathrm{e}^{-u})^{\alpha-1} - 1\right],
\end{aligned}$$

where we used (5.11) with $\Gamma(k+1-\alpha) = \displaystyle\int_0^\infty x^{k-\alpha}\, \mathrm{e}^{-x}\, dx$ for the first equality and two integrations by parts for the last. □

We give bounds on g_n.

Lemma 5.2. *Assume that $\rho(t) = C_0 t^{-\alpha} + O(t^{-\alpha+\zeta})$ for some $C_0 > 0$ and $\zeta > 0$. Then we have, for $n \geq 2$,*

$$g_n = C_0 \Gamma(2-\alpha) n^{\alpha} + O(n^{\alpha - \min(\zeta, 1)}). \tag{5.13}$$

Proof. Notice that

$$g_n = n(n-1) \int_0^1 (1-t)^{n-2} t \left(C_0 t^{-\alpha} + O(t^{-\alpha+\zeta})\right)\ dt = C_0 n(n-1) \frac{\Gamma(2-\alpha)\Gamma(n-1)}{\Gamma(n+1-\alpha)} + h_n,$$

where $h_n = n(n-1)\int_0^1 (1-t)^{n-2}t^{-\alpha+\zeta+1}O(1)\,dt$. In particular, using (5.12), we have for $n \geq 2$

$$|h_n| \leq cn(n-1)\int_0^1 (1-t)^{n-2}t^{-\alpha+\zeta+1} = cn(n-1)\frac{\Gamma(2-\alpha+\zeta)\Gamma(n-1)}{\Gamma(n+1-\alpha+\zeta)} \leq cn^{\alpha-\zeta}.$$

Using (5.12) again, we get that $\Gamma(n-1)/\Gamma(n+1-\alpha) = n^{\alpha-2} + O(n^{\alpha-3})$. This implies that

$$g_n = C_0\Gamma(2-\alpha)n^\alpha + O(n^{\max(\alpha-1,\alpha-\zeta)}).$$

□

We give an expansion of the first moment of $X_1^{(n)}$.

Lemma 5.3. *Assume that $\rho(t) = C_0 t^{-\alpha} + O(t^{-\alpha+\zeta})$ for some $C_0 > 0$ and $\zeta > 0$. Let $\varepsilon_0 > 0$. We set*

$$\varphi_n = \begin{cases} n^{-\zeta} & \text{if} \quad \zeta < \alpha - 1, \\ n^{1-\alpha+\varepsilon_0} & \text{if} \quad \zeta = \alpha - 1, \\ n^{1-\alpha} & \text{if} \quad \zeta > \alpha - 1. \end{cases} \tag{5.14}$$

There exists a constant $C_{5.15}$ s.t. for all $n \geq 2$, we have

$$\left|\mathbb{E}[X_1^{(n)}] - \frac{1}{\gamma}\right| \leq C_{5.15}\varphi_n. \tag{5.15}$$

Proof. We have

$$\begin{aligned}
\mathbb{E}[X_1^{(n)}] = \sum_{k\geq 1}\mathbb{P}(X_1^{(n)} \geq k) &= \frac{\int_0^1 \sum_{k\geq 1}\mathbb{P}(B_{n,x} \geq k+1)\nu(dx)}{g_n} \\
&= \frac{\int_0^1 (\mathbb{E}[B_{n,x}] - \mathbb{P}(B_{n,x} \geq 1))\nu(dx)}{g_n} \qquad (5.16) \\
&= \frac{\int_0^1 nx\nu(dx) - \int_0^1 (1-(1-x)^n)\nu(dx)}{g_n} \\
&= \frac{n\int_0^1 [1-(1-t)^{n-1}]\rho(t)\,dt}{g_n} \qquad (5.17) \\
&= \frac{\int_0^1 (1-t)^{n-2}\left(\int_t^1 \rho(r)\,dr\right)\,dt}{\int_0^1 (1-t)^{n-2}t\rho(t)\,dt},
\end{aligned}$$

using (5.9) for the first equality and (5.8) for the last. Notice that

$$\begin{aligned}\int_t^1 \rho(r)\,dr &= \frac{1}{\gamma}\,t\rho(t) + O(1) + \int_t^1 O(r^{-\alpha+\zeta})\,dr + O(t^{-\alpha+\zeta+1}) \\ &= \frac{1}{\gamma}\,t\rho(t) + O(t^{\min(-\alpha+\zeta+1,0)}) + O(|\log(t)|)\mathbf{1}_{\{\alpha-\zeta=1\}} \\ &= \frac{1}{\gamma}\,t\rho(t) + O(t^{\min(-\alpha+\zeta+1,0)}) + O(t^{-\varepsilon_0})\mathbf{1}_{\{\alpha-\zeta=1\}}.\end{aligned}$$

This implies that

$$\mathbb{E}[X_1^{(n)}] = \frac{1}{\gamma} + \frac{n(n-1)}{g_n}\int_0^1 (1-t)^{n-2}\left(O(t^{\min(-\alpha+\zeta+1,0)}) + O(t^{-\varepsilon_0})\mathbf{1}_{\{\alpha-\zeta=1\}}\right)\,dt.$$

Using (5.10), (5.12) and Lemma 5.2, we get

$$\begin{aligned}\left|\mathbb{E}[X_1^{(n)}] - \frac{1}{\gamma}\right| &\le c\frac{n(n-1)}{g_n}\int_0^1 (1-t)^{n-2}\left(t^{\min(-\alpha+\zeta+1,0)} + t^{-\varepsilon_0}\mathbf{1}_{\{\alpha-\zeta=1\}}\right)\quad dt \\ &\le cn^{2-\alpha}(n^{-1-\min(-\alpha+\zeta+1,0)} + n^{-1+\varepsilon_0}\mathbf{1}_{\{\alpha-\zeta=1\}}) \\ &\le c\varphi_n.\end{aligned}$$

□

We give an upper bound for the second moment of $X_1^{(n)}$.

Lemma 5.4. *Assume that $\rho(t) = O(t^{-\alpha})$. Then there exists a constant $C_{5.18}$ s.t. for all $n \ge 2$, we have*

$$\mathbb{E}\left[\left(X_1^{(n)}\right)^2\right] \le C_{5.18}\frac{n^2}{g_n}. \tag{5.18}$$

Proof. Using the identity $\mathbb{E}[Y^2] = \sum_{k\ge1}(2k-1)\mathbb{P}(Y \ge k)$ for $\mathbb{N}$-valued random variables,

we get

$$\begin{aligned}
\mathbb{E}\left[\left(X_1^{(n)}\right)^2\right] &= \frac{\int_0^1 \sum_{k\geq 1}(2k-1)\mathbb{P}(B_{n,x}\geq k+1)\nu(dx)}{g_n} \\
&= \frac{\int_0^1 \left(\sum_{k\geq 1}(2(k+1)-1)\mathbb{P}(B_{n,x}\geq k+1) - 2\sum_{k\geq 1}\mathbb{P}(B_{n,x}\geq k+1)\right)\nu(dx)}{g_n} \\
&= \frac{\int_0^1 \left(\mathbb{E}[B_{n,x}^2] - 2\mathbb{E}[B_{n,x}] + \mathbb{P}(B_{n,x}\geq 1)\right)\nu(dx)}{g_n} \\
&= \frac{\int_0^1 \left(\mathbb{E}[B_{n,x}^2] - \mathbb{E}[B_{n,x}]\right)\nu(dx)}{g_n} - \mathbb{E}[X_1^{(n)}] \\
&= \frac{\int_0^1 n(n-1)x^2\nu(dx)}{g_n} - \mathbb{E}[X_1^{(n)}] \\
&= 2n(n-1)\frac{\int_0^1 t\rho(t)\,dt}{g_n} - \mathbb{E}[X_1^{(n)}],
\end{aligned}$$

where we have used (5.16) for the fourth equality. Use $\int_0^1 t\rho(t)\,dt < \infty$ and $\mathbb{E}[X_1^{(n)}] \geq 0$ to conclude. □

We consider ϕ_n the Laplace transform of $X_1^{(n)}$: for $u \geq 0$, $\phi_n(u) = \mathbb{E}[\mathrm{e}^{-uX_1^{(n)}}]$.

Lemma 5.5. *Assume that $\rho(t) = C_0 t^{-\alpha} + O(t^{-\alpha+\zeta})$ for some $C_0 > 0$ and $\zeta > 0$. Let $\varepsilon_0 > 0$. Recall φ_n given by (5.14). Then we have, for $n \geq 2$,*

$$\phi_n(u) = 1 - \frac{u}{\gamma} + \frac{u^\alpha}{\gamma} + R(n,u), \tag{5.19}$$

where $R(n,u) = \left(u\varphi_n + u^2\right)h(n,u)$ with $\sup_{u\in[0,K],n\geq 2}|h(n,u)| < \infty$ for all $K > 0$.

Proof. We have

$$\begin{aligned}
&\phi_n(u)\\
&= \mathbb{E}\left[\mathrm{e}^{-uX_1^{(n)}}\right] = \sum_{k=1}^{n-1} \mathrm{e}^{-uk}\,\mathbb{P}(X_1^{(n)} = k)\\
&= \sum_{k=1}^{n-1} \mathrm{e}^{-uk}\,\mathbb{P}(X_1^{(n)} \geq k) - \sum_{k=2}^{n} \mathrm{e}^{-u(k-1)}\,\mathbb{P}(X_1^{(n)} \geq k)\\
&= \mathrm{e}^{-u} + \sum_{k=2}^{n-1} \mathrm{e}^{-uk}(1-\mathrm{e}^{u})\mathbb{P}(X_1^{(n)} \geq k)\\
&= \mathrm{e}^{-u} + (1-\mathrm{e}^{u}) \sum_{k=2}^{n-1} \frac{\mathrm{e}^{-uk}}{g_n} \int_0^1 \frac{n!}{k!(n-k-1)!} t^k (1-t)^{n-k-1} \rho(t)\, dt\\
&= \mathrm{e}^{-u} + (1-\mathrm{e}^{u}) \frac{n}{g_n} \int_0^1 \left[(1-t(1-\mathrm{e}^{-u}))^{n-1} - (1-t)^{n-1} - (n-1)\,\mathrm{e}^{-u}\, t(1-t)^{n-2}\right] \rho(t) dt\\
&= 1 + (1-\mathrm{e}^{u}) \frac{n}{g_n} \int_0^1 \left[(1-t(1-\mathrm{e}^{-u}))^{n-1} - (1-t)^{n-1}\right] \rho(t)\, dt,
\end{aligned}$$

where we used (5.8) for the last equality. Using (5.17), this implies

$$\phi_n(u) = 1 + (1-\mathrm{e}^{u}) \frac{n}{g_n} A + (1-\mathrm{e}^{u}) \mathbb{E}[X_1^{(n)}]. \tag{5.20}$$

with $A = \displaystyle\int_0^1 \left[(1-t(1-\mathrm{e}^{-u}))^{n-1} - 1\right] \rho(t)\, dt$.

Thanks to Lemma 5.3, we have that

$$(1-\mathrm{e}^{u}) \mathbb{E}[X_1^{(n)}] = -\frac{u}{\gamma} + \left(u^2 + u\varphi_n\right) h_1(n,u), \tag{5.21}$$

where, for all $K > 0$, $\sup_{u\in[0,K], n\geq 2} |h_1(n,u)| < \infty$.

To compute A, we set $a = (1-\mathrm{e}^{-u})$ and $f(t) = t^{-\max(\alpha-1-\zeta,0)} + t^{-\varepsilon_0} \mathbf{1}_{\{\alpha-\zeta=1\}}$. An integration by part gives

$$\begin{aligned}
A &= -a(n-1) \int_0^1 (1-at)^{n-2} \left(\int_t^1 \rho(r)\, dr \right) dt\\
&= -a(n-1) C_0 \int_0^1 (1-at)^{n-2} \left(\frac{t^{1-\alpha}}{\gamma} + O(f(t)) \right) dt\\
&= -A_1 + A_2,
\end{aligned}$$

with $A_1 = \dfrac{a(n-1)}{\gamma} C_0 \displaystyle\int_0^1 (1-at)^{n-2} t^{1-\alpha}\, dt$ and $A_2 = a(n-1) \displaystyle\int_0^1 (1-at)^{n-2} O(f(t))\, dt$. We have

$$\begin{aligned} A_1 &= \frac{a^{\alpha-1}(n-1)}{\gamma} C_0 \int_0^a (1-t)^{n-2} t^{1-\alpha}\, dt \\ &= \frac{a^{\alpha-1}(n-1)}{\gamma} C_0 \int_0^1 (1-t)^{n-2} t^{1-\alpha}\, dt - \frac{a^{\alpha-1}(n-1)}{\gamma} C_0 \int_a^1 (1-t)^{n-2} t^{1-\alpha}\, dt \\ &= \frac{a^{\alpha-1}(n-1)}{\gamma} C_0 \frac{\Gamma(n-1)\Gamma(2-\alpha)}{\Gamma(n+1-\alpha)} - \frac{a^{\alpha-1}(n-1)}{\gamma} C_0 \int_a^1 (1-t)^{n-2} t^{1-\alpha}\, dt \end{aligned}$$

Since $a \geq 0$, we have for $u \in [0, K]$ and $n \geq 2$

$$0 \leq \frac{a^{\alpha-1}(n-1)}{\gamma} \int_a^1 (1-t)^{n-2} t^{1-\alpha}\, dt \leq \frac{(n-1)}{\gamma} \int_a^1 (1-t)^{n-2}\, dt \leq \frac{1}{\gamma}\,.$$

Using (5.12) and Lemma 5.2, we get $|A_1 - \dfrac{a^{\alpha-1}}{\gamma}\dfrac{g_n}{n}| \leq c(1+n^{\alpha-1-\min(\zeta,1)}) \leq c n^{\max(\alpha-1-\zeta,0)}$, where c does not depend on n and $u \geq 0$. We also have, using (5.10) and (5.12)

$$|A_2| \leq ca(n-1) \int_0^1 (1-at)^{n-2} f(t)\, dt \leq c(n^{\max(\alpha-1-\zeta,0)} + n^{\varepsilon_0} \mathbf{1}_{\{\alpha-\zeta=1\}}).$$

We deduce, using Lemma 5.2 twice, that

$$|A + \frac{a^{\alpha-1}}{\gamma}\frac{g_n}{n}| \leq c(n^{\max(\alpha-1-\zeta,0)} + n^{\varepsilon_0} \mathbf{1}_{\{\alpha-\zeta=1\}}) \leq c\frac{g_n}{n}\varphi_n.$$

We deduce that

$$(1-\mathrm{e}^u)\frac{n}{g_n} A = (1-\mathrm{e}^u)\left(-\frac{(1-\mathrm{e}^{-u})^{\alpha-1}}{\gamma} + \varphi_n O(1)\right) = \frac{u^\alpha}{\gamma} + \left(u^{\alpha+1} + u\varphi_n\right) h_2(n,u), \quad (5.22)$$

where $\sup_{u\in[0,K], n\geq 2} |h_2(n,u)| < \infty$ for all $K > 0$. Then use the expression of ϕ_n given by (5.20) as well as (5.21) and (5.22) to end the proof. □

5.3 Asymptotics for the number of jumps

Let $\alpha \in (1,2)$. We assume that $\rho(t) = C_0 t^{-\alpha} + O(t^{-\alpha+\zeta})$ for some $C_0 > 0$ and $\zeta > 1 - 1/\alpha$.

Let $V = (V_t, t \geq 0)$ be a α-stable Lévy process with no positive jumps (see chap. VII in Bertoin (1996)) with Laplace exponent $\psi(u) = u^\alpha/\gamma$: for all $u \geq 0$, $\mathbb{E}[\mathrm{e}^{-uV_t}] = \mathrm{e}^{tu^\alpha/\gamma}$.

Lemma 5.1 implies that $(X_1^{(n)}, \ldots, X_k^{(n)})$ converges in distribution to $(X_1, \ldots, X_k)$ where $(X_k, k \geq 1)$ is a sequence of independent random variables distributed as X. Using

Lemma 5.1 and (5.12), we get that $\mathbb{P}(X \geq k) \sim_{+\infty} \frac{1}{\Gamma(2-\alpha)}\, k^{-\alpha}$. Hence Proposition 9.39 in Breiman (1992) implies that the law of X is in the domain of attraction of the α-stable distribution. We set $W_t^{(n)} = n^{-1/\alpha} \sum_{k=1}^{\lfloor nt \rfloor} (X_k - \frac{1}{\gamma})$ for $t \in [0, \gamma]$. An easy calculation using the Laplace transform of X shows that for fixed t the sequence $W_t^{(n)}$ converges in distribution to V_t. Then using Theorem 16.14 in Kallenberg (2002), we get that the process $(W_t^{(n)}, t \in [0, \gamma])$ converges in distribution to $V = (V_t, t \in [0, \gamma])$. We shall give in Corollary 5.1 a similar result with X_k replaced by $X_k^{(n)}$.

We first give a proof of the convergence of τ_n, see also Gnedin and Yakubovich (2007) and Iksanov and Möhle (2008) for a different proof. We will use that $\sum_{i=1}^{\tau_n} (X_i^{(n)} - \frac{1}{\gamma}) = n - 1 - \frac{\tau_n}{\gamma}$.

Proposition 5.1. *We assume that $\zeta > 1 - 1/\alpha$. We have the following convergence in distribution*

$$n^{-\frac{1}{\alpha}} \left(n - \frac{\tau_n}{\gamma} \right) \xrightarrow[n\to\infty]{(d)} V_\gamma.$$

Proof. Using Mukherjea et al. (2006), it is enough to prove that $\lim_{n\to\infty} \mathbb{E}[\mathrm{e}^{-un^{-\frac{1}{\alpha}}\left(n - \frac{\tau_n}{\gamma}\right)}] = \mathrm{e}^{u^\alpha}$ for all $u \geq 0$. Recall $\mathcal{Y} = (\mathcal{Y}_k, k \geq 0)$ is the filtration generated by Y. Notice τ_n is an $\mathcal{Y}$-stopping time. Recall that for $m \geq 1$, ϕ_m denotes the Laplace transform of $X_1^{(m)}$. For fixed n, and for any $v \geq 0$, the process $(M_{v,k}, k \geq 0)$ defined by

$$M_{v,k} = \prod_{i=1}^{k} \exp\left(-vX_i^{(n)} - \log \phi_{Y_{i-1}^{(n)}}(v) \right)$$

is a bounded martingale w.r.t. the filtration $\mathcal{Y}$. Notice that $\mathbb{E}[M_{v,k}] = 1$. As $X_i = 0$ for $i > \tau_n$, we also have

$$M_{v,k} = \prod_{i=1}^{k\wedge\tau_n} \exp\left(-vX_i^{(n)} - \log \phi_{Y_{i-1}^{(n)}}(v) \right). \tag{5.23}$$

Let $u \geq 0$ and consider a non-negative sequence $(a_n, n \geq 1)$ which converges to 0. Using (5.19), we get that :

$$M_{ua_n,k} = \exp\left(-ua_n \sum_{i=1}^{k\wedge\tau_n} X_i^{(n)} - \sum_{i=1}^{k\wedge\tau_n} \left(-\frac{ua_n}{\gamma} + \frac{u^\alpha a_n^\alpha}{\gamma} + R(Y_{i-1}^{(n)}, ua_n) \right) \right).$$

In particular, we have

$$M_{ua_n,\tau_n} = \exp\left(-ua_n(n-1-\frac{\tau_n}{\gamma}) - \frac{u^\alpha \tau_n a_n^\alpha}{\gamma} - \sum_{i=1}^{\tau_n} R(Y_{i-1}^{(n)}, ua_n)\right). \tag{5.24}$$

We first give an upper bound for $\sum_{i=1}^{\tau_n} R(Y_{i-1}^{(n)}, ua_n)$.

Lemma 5.6. *We assume that $\zeta > 1 - 1/\alpha$. Let $K > 0$. Let $\eta \geq \frac{1}{\alpha}$. There exist $\varepsilon_1 > 0$ and $C_{5.25}(K)$ a finite constant such that for all $n \geq 1$ and $u \in [0, K]$, a.s. with $a_n = n^{-\eta}$,*

$$\sum_{i=1}^{\tau_n} \left| R(Y_{i-1}^{(n)}, ua_n) \right| \leq C_{5.25}(K) n^{-\varepsilon_1}. \tag{5.25}$$

Proof. Notice that $\tau_n \leq n-1$. We have seen in Lemma 5.5 that $R(n,u) = \left(u\varphi_n + u^2\right) h(n,u)$ with $\bar{h}(K) = \sup_{u\in[0,K], n\geq 2} |h(n,u)| < \infty$ and φ_n given by (5.14). We have $2 - \alpha - \frac{1}{\alpha} = -\alpha(1-1/\alpha)^2 < 0$. As $\varepsilon_0 > 0$ is arbitrary in (5.14), we can take ε_0 small enough so that $1 - \alpha + \varepsilon_0 < 0$ and $2 - \alpha + \varepsilon_0 - 1/\alpha < 0$. We have

$$a_n \sum_{i=1}^{\tau_n} \varphi_{Y_{i-1}^{(n)}} \leq n^{-1/\alpha} \sum_{j=1}^{n} \varphi_j \leq c \begin{cases} n^{1-\zeta-\frac{1}{\alpha}} & \text{if} \quad \zeta < \alpha - 1, \\ n^{2-\alpha+\varepsilon_0-\frac{1}{\alpha}} & \text{if} \quad \zeta = \alpha - 1, \\ n^{2-\alpha-\frac{1}{\alpha}} & \text{if} \quad \zeta > \alpha - 1. \end{cases}$$

For $\varepsilon_1 > 0$ less than the two positive quantities $-1 + \zeta + \frac{1}{\alpha}$ and $-2 + \alpha - \varepsilon_0 + \frac{1}{\alpha}$, we have $a_n \sum_{i=1}^{\tau_n} \varphi_{Y_{i-1}^{(n)}} \leq c n^{-\varepsilon_1}$. We deduce that, for $u \in [0, K]$,

$$\begin{aligned} \sum_{i=1}^{\tau_n} \left| R(Y_{i-1}^{(n)}, ua_n) \right| &\leq \bar{h}(K) \sum_{i=1}^{\tau_n} \left(\varphi_{Y_{i-1}^{(n)}} ua_n + (ua_n)^2 \right) \\ &\leq \bar{h}(K) \sum_{j=1}^{n} \left(\varphi_j K a_n + (Ka_n)^2 \right) \\ &\leq c\bar{h}(K)(Kn^{-\varepsilon_1} + K^2 n^{1-\frac{2}{\alpha}}), \end{aligned}$$

for some constant c independent of n, u and K. Taking $\varepsilon_1 > 0$ small enough so that $\varepsilon_1 < \frac{2}{\alpha} - 1$, we then get (5.25). □

Next we prove the following Lemma.

Lemma 5.7. *We assume that $\zeta > 1 - 1/\alpha$. Let $\varepsilon > 0$. The sequence $(n^{-(1/\alpha)-\varepsilon}(n - 1 - \frac{\tau_n}{\gamma}), n \geq 1)$ converges in probability to 0.*

Proof. We set $a_n = n^{-\frac{1}{\alpha}-\varepsilon}$. Notice that

$$\mathrm{e}^{-ua_n(n-1-\frac{\tau_n}{\gamma})} = M_{ua_n,\tau_n}\,\mathrm{e}^{\frac{u^\alpha \tau_n a_n}{\gamma}+\sum_{i=1}^{\tau_n} R(Y_{i-1}^{(n)},ua_n)}.$$

As $\tau_n \le n-1$, we have $0 \le \tau_n a_n^\alpha \le n^{-\alpha\varepsilon}$. Using (5.25), we get for $u \ge 0$

$$\mathbb{E}[M_{ua_n,\tau_n}]\,\mathrm{e}^{-C_{5.25}(u)n^{-\varepsilon_1}} \le \mathbb{E}[\mathrm{e}^{-ua_n(n-1-\frac{\tau_n}{\gamma})}] \le \mathbb{E}[M_{ua_n,\tau_n}]\,\mathrm{e}^{C_{5.25}(u)n^{-\varepsilon_1}+\frac{u^\alpha n^{-\alpha\varepsilon}}{\gamma}}.$$

As τ_n is bounded, the stopping time theorem gives $\mathbb{E}[M_{ua_n,\tau_n}] = 1$. We deduce that, for all $u \ge 0$, $\lim_{n\to\infty} \mathbb{E}[\mathrm{e}^{-ua_n(n-1-\frac{\tau_n}{\gamma})}] = 1$. Using Mukherjea et al. (2006), we get the convergence in law of $a_n(n-1-\frac{\tau_n}{\gamma})$ to 0, and then in probability as the limit is constant. □

Let $a_n = n^{-\frac{1}{\alpha}}$ and $u \ge 0$. We have

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{e}^{-ua_n(n-1-\frac{\tau_n}{\gamma})}\right] \\
&\quad= \mathbb{E}\left[\mathrm{e}^{-ua_n(n-1-\frac{\tau_n}{\gamma})}\left(1-\mathrm{e}^{-u^\alpha a_n^\alpha(\frac{\tau_n}{\gamma}-n)}\right)\right] + \mathbb{E}\left[\mathrm{e}^{-ua_n(n-1-\frac{\tau_n}{\gamma})}\,\mathrm{e}^{-u^\alpha a_n^\alpha(\frac{\tau_n}{\gamma}-n)}\right] \\
&\quad= I_1 + I_2,
\end{aligned} \tag{5.26}$$

with $I_1 = \mathbb{E}\left[\mathrm{e}^{-ua_n(n-1-\frac{\tau_n}{\gamma})}\left(1-\mathrm{e}^{-u^\alpha a_n^\alpha(\frac{\tau_n}{\gamma}-n)}\right)\right]$ and $I_2 = \mathbb{E}\left[M_{ua_n,\tau_n}\,\mathrm{e}^{u^\alpha+\sum_{i=1}^{\tau_n} R(Y_{i-1}^{(n)},ua_n)}\right]$.

Using (5.25) and $\mathbb{E}[M_{ua_n,\tau_n}] = 1$, we get

$$\mathrm{e}^{u^\alpha - C_{5.25}(u)n^{-\varepsilon_1}} \le I_2 \le \mathrm{e}^{u^\alpha + C_{5.25}(u)n^{-\varepsilon_1}}.$$

This implies that $\lim_{n\to\infty} I_2 = \mathrm{e}^{u^\alpha}$.

We now prove that $\lim_{n\to\infty} I_1 = 0$. Recall that $\tau_n \le n-1$ so that $\tau_n a_n^\alpha \le 1$ and thanks to (5.25), we get

$$\mathbb{E}[\mathrm{e}^{-ua_n(n-1-\frac{\tau_n}{\gamma})}] = \mathbb{E}\left[M_{ua_n,\tau_n}\,\mathrm{e}^{\frac{u^\alpha \tau_n a_n^\alpha}{\gamma}+\sum_{i=1}^{\tau_n} R(Y_{i-1}^{(n)},ua_n)}\right] \le M(u)\mathbb{E}[M_{ua_n,\tau_n}] = M(u),$$

where $M(u)$ is a constant which does not depend on n. By Cauchy-Schwarz' inequality, we get that

$$\begin{aligned}
I_1^2 = \mathbb{E}\left[\mathrm{e}^{-ua_n(n-1-\frac{\tau_n}{\gamma})}\left(1-\mathrm{e}^{-u^\alpha a_n^\alpha(\frac{\tau_n}{\gamma}-n)}\right)\right]^2 &\le \mathbb{E}\left[\mathrm{e}^{-2ua_n(n-1-\frac{\tau_n}{\gamma})}\right]\mathbb{E}\left[\left(1-\mathrm{e}^{-u^\alpha a_n^\alpha(\frac{\tau_n}{\gamma}-n)}\right)^2\right] \\
&\le M(2u)\mathbb{E}\left[\left(1-\mathrm{e}^{-u^\alpha \frac{1}{n}(\frac{\tau_n}{\gamma}-n)}\right)^2\right].
\end{aligned}$$

Notice $(\frac{1}{n}(\frac{\tau_n}{\gamma}-n), n \ge 1)$ is bounded from below and above by finite constants, and thanks to Lemma 5.7 it converges to 0 in probability. Hence, we deduce that

$$\lim_{n\to\infty} \mathbb{E}\left[\left(1-\mathrm{e}^{-u^\alpha \frac{1}{n}(\frac{\tau_n}{\gamma}-n)}\right)^2\right] = 0.$$

This implies that $\lim_{n\to\infty} I_1 = 0$.

From the convergence of I_1 and I_2, we deduce from (5.26) that $\lim_{n\to\infty} \mathbb{E}\left[\mathrm{e}^{-ua_n(n-1-\frac{\tau_n}{\gamma})}\right] = \mathrm{e}^{u^\alpha}$. This ends the proof of the Proposition. □

We now give a general result.

Proposition 5.2. *We assume that $\zeta > 1 - 1/\alpha$. Let $f_n : \mathbb{R}_+ \to \mathbb{R}_+$ be uniformly bounded functions such that*

$$\kappa = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{\lfloor n\gamma \rfloor} f_n(k/n)^\alpha$$

exists. Then we have the following convergence in distribution

$$V^{(n)}(f_n) := n^{-\frac{1}{\alpha}} \sum_{k=1}^{\tau_n} f_n(k/n)(X_k^{(n)} - \frac{1}{\gamma}) \xrightarrow[n\to\infty]{(d)} \kappa^{1/\alpha} V_1. \tag{5.27}$$

In particular, if $f : \mathbb{R}_+ \to \mathbb{R}_+$ is a bounded locally Riemann integrable function, then

$$V^{(n)}(f) = n^{-\frac{1}{\alpha}} \sum_{k=1}^{\tau_n} f(k/n)(X_k^{(n)} - \frac{1}{\gamma}) \xrightarrow[n\to\infty]{(d)} \int_0^\gamma f(t) dV_t, \tag{5.28}$$

where the distribution of $\int_0^\gamma f(t)dV_t$ is characterized by its Laplace transform: for $u \geq 0$,

$$\mathbb{E}[\exp(-u \int_0^\gamma f(t)dV_t)] = \exp\left(\frac{u^\alpha}{\gamma} \int_0^\gamma f^\alpha(t)\, dt\right). \tag{5.29}$$

If we apply this Proposition with step functions, we deduce the following result.

Corollary 5.1. *We assume that $\zeta > 1-1/\alpha$. Let $V_t^{(n)} = V^{(n)}(\mathbf{1}_{[0,t]}) = n^{-1/\alpha} \sum_{k=1}^{\lfloor nt \rfloor \wedge \tau_n} (X_k^{(n)} - \frac{1}{\gamma})$ for $t \in [0,\gamma)$, and $V_\gamma^{(n)} = V^{(n)}(\mathbf{1}) = n^{-1/\alpha}\left(n - 1 - \frac{\tau_n}{\gamma}\right)$. The finite-dimensional marginals of the process $(V_t^{(n)}, t \in [0,\gamma])$ converges in law to those of the process $(V_t, t \in [0,\gamma])$.*

Proof. Thanks to Mukherjea et al. (2006), it is enough to prove that

$$\mathbb{E}[\exp(-uV^{(n)}(f_n))] \xrightarrow[n\to\infty]{} \mathrm{e}^{\kappa u^\alpha/\gamma}.$$

Taking uf_n as f_n, we shall only consider the case $u = 1$.

We set $a = \sup_{n\geq 1, x\geq 0} |f_n(x)|$ and for any bounded function g,

$$A_n(g) = \exp \sum_{k=1}^{\tau_n} \left(-n^{-1/\alpha} g(k/n) X_k^{(n)} - \log \phi_{Y_{k-1}^{(n)}}(n^{-\frac{1}{\alpha}} g(k/n))\right).$$

A martingale argument provides that $\mathbb{E}[A_n(g)] = 1$. Using (5.19), we get that :

$$A_n(g) = \exp\left(-n^{-1/\alpha}\sum_{k=1}^{\tau_n} g(k/n)(X_k^{(n)} - \frac{1}{\gamma}) - n^{-1}\sum_{k=1}^{\tau_n}\frac{g^\alpha(k/n)}{\gamma} - \sum_{k=1}^{\tau_n} R(Y_{k-1}^{(n)}, n^{-\frac{1}{\alpha}}g(k/n))\right)$$
$$= \exp\left(-V^{(n)}(g) - n^{-1}\sum_{k=1}^{\tau_n}\frac{g^\alpha(k/n)}{\gamma} - \sum_{k=1}^{\tau_n} R(Y_{k-1}^{(n)}, n^{-\frac{1}{\alpha}}g(k/n))\right).$$

Let $\Lambda_n = n^{-1}\sum_{k=1}^{\lfloor n\gamma\rfloor}\frac{f_n^\alpha(k/n)}{\gamma} - n^{-1}\sum_{k=1}^{\tau_n}\frac{f_n^\alpha(k/n)}{\gamma}$ and write

$$\mathbb{E}\left[\mathrm{e}^{-V^{(n)}(f_n)}\right] = I_1 + I_2$$

with $I_1 = \mathbb{E}\left[\mathrm{e}^{-V^{(n)}(f_n)}\left(1-\mathrm{e}^{\Lambda_n}\right)\right]$ and $I_2 = \mathbb{E}\left[\mathrm{e}^{-V^{(n)}(f_n)}\,\mathrm{e}^{\Lambda_n}\right]$.

First of all, let us prove that I_1 converges to 0 when n tends to ∞. Recall that the functions f_n are uniformly bounded by a. Thanks to (5.25), we have

$$\mathbb{E}[\mathrm{e}^{-2V^{(n)}(f_n)}] = \mathbb{E}[\mathrm{e}^{-V^{(n)}(2f_n)}] = \mathbb{E}\left[A_n(2f_n)\,\mathrm{e}^{n^{-1}\sum_{k=1}^{\tau_n}\frac{2^\alpha f_n^\alpha(k/n)}{\gamma} + \sum_{k=1}^{\tau_n} R(Y_{k-1}^{(n)}, n^{-\frac{1}{\alpha}}2f_n(k))}\right] \le M,$$

where M is a finite constant which does not depend on n. By Cauchy-Schwarz' inequality, we get that

$$(I_1)^2 \le \left(\mathbb{E}\left[\mathrm{e}^{-V^{(n)}(f_n)}\left|1-\mathrm{e}^{\Lambda_n}\right|\right]\right)^2 \le \mathbb{E}\left[\mathrm{e}^{-V^{(n)}(2f_n)}\right]\mathbb{E}\left[\left(1-\mathrm{e}^{\Lambda_n}\right)^2\right] \le M\mathbb{E}\left[\left(1-\mathrm{e}^{\Lambda_n}\right)^2\right].$$

Moreover as $|1-\mathrm{e}^x| \le \mathrm{e}^{|x|}-1$ and $\Lambda_n \le \frac{a^\alpha}{n\gamma}|\lfloor n\gamma\rfloor - \tau_n|$, we get

$$\mathbb{E}\left[\left(1-\mathrm{e}^{\Lambda_n}\right)^2\right] \le \mathbb{E}\left[\left(1-\mathrm{e}^{\frac{|\lfloor n\gamma\rfloor-\tau_n|a^\alpha}{n\gamma}}\right)^2\right]. \tag{5.30}$$

The quantity $\frac{|\lfloor n\gamma\rfloor - \tau_n|a^\alpha}{n\gamma}$ is bounded and goes to 0 in probability when n goes to infinity. Therefore, the right-hand side of (5.30) converges to 0. This implies that $\lim_{n\to\infty} I_1 = 0$.

Let us now consider the convergence of I_2. Remark that

$$I_2 = \mathbb{E}\left[A_n(f_n)\exp\left(n^{-1}\sum_{k=1}^{\lfloor n\gamma\rfloor}\frac{f_n^\alpha(k/n)}{\gamma} + \sum_{k=1}^{\tau_n} R(Y_{k-1}^{(n)}, n^{-\frac{1}{\alpha}}f_n(k))\right)\right].$$

Recall that f_n is bounded by a and that $\mathbb{E}[A_n(f_n)] = 1$. Using Lemma 5.6, we get for some $\varepsilon > 0$

$$\exp\left(-C_{5.25}(a)n^{-\varepsilon_1} - n^{-1}\sum_{k=1}^{\lfloor n\gamma\rfloor}\frac{f_n^\alpha(k/n)}{\gamma}\right)$$
$$\leq \mathbb{E}\left[A_n(f_n)\exp\left(n^{-1}\sum_{k=1}^{\lfloor n\gamma\rfloor}\frac{f_n^\alpha(k/n)}{\gamma} + \sum_{k=1}^{\tau_n} R(Y_{k-1}^{(n)}, n^{-\frac{1}{\alpha}}f_n(k))\right)\right]$$
$$\leq \exp\left(C_{5.25}(a)n^{-\varepsilon_1} + n^{-1}\sum_{k=1}^{\lfloor n\gamma\rfloor}\frac{f_n^\alpha(k/n)}{\gamma}\right). \quad (5.31)$$

As $\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{\lfloor n\gamma\rfloor} f_n^\alpha(k/n) = \kappa$, we get that $\lim_{n\to\infty} I_2 = \mathrm{e}^{\kappa/\gamma}$, which achieves the proof of (5.27). To get (5.28), notice that $\kappa = \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{\lfloor n\gamma\rfloor} f(k/n)^\alpha = \int_0^\gamma f(t)^\alpha\, dt$. □

5.4 First approximation of the length of the coalescent tree

Let $\alpha \in (1,2)$. We assume that $\rho(t) = C_0 t^{-\alpha} + O(t^{-\alpha+\zeta})$ for some $C_0 > 0$ and $\zeta > 1 - 1/\alpha$.

Recall that the length of the coalescent tree up to the $\lfloor nt\rfloor$-th coalescence is, for $t \geq 0$, given by (5.3). The next Lemma gives an upper bound on the error when one replaces the exponential random variables by their mean.

Lemma 5.8. *For $t \geq 0$, let*

$$\tilde{L}_t^{(n)} = \sum_{k=0}^{\lfloor nt\rfloor\wedge(\tau_n-1)} \frac{Y_k^{(n)}}{g_{Y_k^{(n)}}}.$$

There exists a finite constant $C_{5.32}$ such that, we have

$$\mathbb{E}\left[\sup_{t\geq 0}(L_t^{(n)} - \tilde{L}_t^{(n)})^2\right] \leq C_{5.32}\begin{cases} n^{3-2\alpha} & \text{if } \alpha < 3/2, \\ \log(n) & \text{if } \alpha = 3/2, \\ 1 & \text{if } \alpha > 3/2. \end{cases} \quad (5.32)$$

Proof. Recall that $\mathcal{Y} = (\mathcal{Y}_k, k \geq 0)$ denotes the filtration generated by Y. Conditionally on $\mathcal{Y}$, the random variables $\frac{Y_k^{(n)}}{g_{Y_k^{(n)}}}(E_k - 1)$ are independent with zero mean. We deduce

that

$$\mathbb{E}\left[\sup_{t\geq 0}(L_t^{(n)}-\tilde{L}_t^{(n)})^2|\mathcal{Y}\right]=\mathbb{E}\left[\sup_{t\geq 0}\left(\sum_{k=0}^{\lfloor nt\rfloor\wedge(\tau_n-1)}\frac{Y_k^{(n)}}{g_{Y_k^{(n)}}}(E_k-1)\right)^2|\mathcal{Y}\right]$$
$$\leq 4\sum_{k=0}^{\tau_n-1}\left(\frac{Y_k^{(n)}}{g_{Y_k^{(n)}}}\right)^2$$
$$\leq 4\sum_{\ell=1}^{n}\left(\frac{\ell}{g_\ell}\right)^2,$$

where we used Doob inequality for martingale in the first inequality. Thanks to (5.13), we get

$$\mathbb{E}\left[\sup_{t\geq 0}(L_t^n-\tilde{L}_t^{(n)})^2|\mathcal{Y}\right]\leq c\sum_{\ell=1}^{n}\ell^{2-2\alpha}\leq c\begin{cases}n^{3-2\alpha} & \text{if } \alpha<3/2,\\ \log(n) & \text{if } \alpha=3/2,\\ 1 & \text{if } \alpha>3/2,\end{cases}$$

where c is non random. This implies the result. □

Lemma 5.9. *For $t\geq 0$, let*

$$\hat{L}_t^{(n)}=\sum_{k=0}^{\lfloor nt\rfloor\wedge(\tau_n-1)}\left(Y_k^{(n)}\right)^{-\gamma}.$$

There exists a finite constant $C_{5.33}$ such that for all $t\geq 0$, we have

$$|\tilde{L}_t^{(n)}-\frac{\hat{L}_t^{(n)}}{C_0\Gamma(2-\alpha)}|\leq C_{5.33}\begin{cases}n^{2-\alpha-\zeta} & \textit{if } \zeta<2-\alpha,\\ \log(n) & \textit{if } \zeta=2-\alpha,\\ 1 & \textit{if } \zeta>2-\alpha.\end{cases}\tag{5.33}$$

Proof. Use (5.13) to get that

$$\tilde{L}_t^{(n)}-\frac{\hat{L}_t^{(n)}}{C_0\Gamma(2-\alpha)}=\sum_{k=0}^{\lfloor nt\rfloor\wedge(\tau_n-1)}\left(Y_k^{(n)}\right)^{-\gamma}O\left(\left(Y_k^{(n)}\right)^{-\min(\zeta,1)}\right).$$

We deduce that

$$|\tilde{L}_t^{(n)}-\frac{\hat{L}_t^{(n)}}{C_0\Gamma(2-\alpha)}|\leq c\sum_{\ell=1}^{n}\ell^{-\alpha+1-\min(\zeta,1)}\leq c\begin{cases}n^{2-\alpha-\zeta} & \text{if } \zeta<2-\alpha,\\ \log(n) & \text{if } \zeta=2-\alpha,\\ 1 & \text{if } \zeta>2-\alpha.\end{cases}$$

□

5.5 Limit distribution of $\hat{L}_t^{(n)}$

Let $\alpha \in (1,2)$ and $\gamma = \alpha - 1$. For $t \in [0,\gamma]$, we set

$$v(t) = \int_0^t \left(1 - \frac{r}{\gamma}\right)^{-\gamma} dr.$$

Theorem 5.1. *We assume that $\rho(t) = C_0 t^{-\alpha} + O(t^{-\alpha+\zeta})$ for some $C_0 > 0$ and $\zeta > 1 - 1/\alpha$.*
(i) Let $t_0 \in [0,\gamma)$ and $\delta > 0$. The following convergence in probability holds:

$$n^{\frac{\alpha-1}{2}-\delta} \sup_{0\le t\le t_0} |n^{-2+\alpha}\hat{L}_t^{(n)} - v(t)| \xrightarrow[n\to\infty]{\mathbb{P}} 0. \tag{5.34}$$

(ii) Let $t \in [0,\gamma)$. The following convergence in distribution holds:

$$n^{-1+\alpha-1/\alpha}(\hat{L}_t^{(n)} - n^{2-\alpha}v(t)) \xrightarrow[n\to\infty]{(d)} (\alpha-1)\int_0^t dr\,(1-\frac{r}{\gamma})^{-\alpha}V_r. \tag{5.35}$$

Proof. Let $\varepsilon_2 \in (0,\gamma)$, $t_0 = \gamma - \varepsilon_2$ and $t \in [0,t_0]$. We use a Taylor expansion to get

$$\begin{aligned}
\hat{L}_t^{(n)} &= \sum_{k=0}^{\lfloor nt\rfloor\wedge(\tau_n-1)} \left(n - \sum_{i=1}^k X_i^{(n)}\right)^{-\gamma} \\
&= \sum_{k=0}^{\lfloor nt\rfloor\wedge(\tau_n-1)} \left(n - \frac{k}{\gamma} - \sum_{i=1}^k (X_i^{(n)} - \frac{1}{\gamma})\right)^{-\gamma} \\
&= \sum_{k=0}^{\lfloor nt\rfloor\wedge(\tau_n-1)} \left(n - \frac{k}{\gamma}\right)^{-\gamma} (1-\Delta_{n,k})^{-\gamma} \\
&= I_n(t) + \gamma J_n(t) + \gamma(\gamma+1)R_n(t)
\end{aligned} \tag{5.36}$$

with $\Delta_{n,k} = \dfrac{\sum_{i=1}^k (X_i^{(n)} - \frac{1}{\gamma})}{n - k/\gamma}$ and

$$\begin{aligned}
I_n(t) &= \sum_{k=0}^{\lfloor nt\rfloor\wedge(\tau_n-1)} \left(n - \frac{k}{\gamma}\right)^{-\gamma}, \\
J_n(t) &= \sum_{k=1}^{\lfloor nt\rfloor\wedge(\tau_n-1)} \left(n - \frac{k}{\gamma}\right)^{-\gamma-1} \sum_{i=1}^k (X_i^{(n)} - \frac{1}{\gamma}), \\
R_n(t) &= \sum_{k=1}^{\lfloor nt\rfloor\wedge(\tau_n-1)} \left(n - \frac{k}{\gamma}\right)^{-\gamma} \int_0^{\Delta_{n,k}} (\Delta_{n,k} - s)\,(1-s)^{-\gamma-2}\,ds.
\end{aligned}$$

Notice that a.s. $\Delta_{n,k} < 1$, so that $R_n(t)$ is well defined.
Convergence of $I_n(t)$. We write $I_n(t) = n^{2-\alpha} I_{n,1}(t)\mathbf{1}_{\{nt<\tau_n\}} + I_n(t)\mathbf{1}_{\{nt\geq\tau_n\}}$ with $I_{n,1}(t) = \frac{1}{n}\sum_{k=0}^{\lfloor nt\rfloor}\left(1-\frac{k}{n\gamma}\right)^{-\gamma}$. Standard computation yields

$$I_{n,1}(t) = v(t) + \frac{1}{n}\, h_3(n,t),$$

where $\sup_{t\in[0,t_0],n\geq 1} |h_3(n,t)| < \infty$. Hence, we have for $\varepsilon > 0$

$$\begin{aligned}
&\mathbb{P}\left(n^{-1+\alpha-1/\alpha}\sup_{0\leq t\leq t_0}\left|I_n(t) - n^{2-\alpha}v(t)\right| \geq \varepsilon\right)\\
&\qquad = \mathbb{P}\left(n^{1-1/\alpha}\sup_{0\leq t\leq t_0}|I_{n,1}(t) - v(t)| \geq \varepsilon, nt_0 < \tau_n\right)\\
&\qquad\quad + \mathbb{P}\left(n^{-1+\alpha-1/\alpha}\sup_{0\leq t\leq t_0}\left|I_n(t) - n^{2-\alpha}v(t)\right| \geq \varepsilon, nt_0 \geq \tau_n\right)\\
&\qquad \leq \mathbb{P}(n^{-1/\alpha}\sup_{0\leq t\leq t_0}|h_3(n,t)| \geq \varepsilon) + \mathbb{P}(nt_0 \geq \tau_n).
\end{aligned}$$

According to Lemma 5.7, τ_n/n converges in probability to γ. This implies that for all $t \in [0,\gamma)$

$$\lim_{n\to\infty}\mathbb{P}(nt \geq \tau_n) = 0. \tag{5.37}$$

As $n^{-1/\alpha}\sup_{0\leq t\leq t_0}|h_3(n,t)| < \varepsilon$ for n large enough, we deduce the following convergence in probability:

$$n^{-1+\alpha-1/\alpha}\sup_{0\leq t\leq t_0}\left|I_n(t) - n^{2-\alpha}v(t)\right| \xrightarrow[n\to\infty]{\mathbb{P}} 0. \tag{5.38}$$

Convergence of $J_n(t)$. Let $t \in [0,t_0]$. To get the convergence of $J_n(t)$, notice that

$$J_n(t) = \sum_{i=1}^{\lfloor nt\rfloor\wedge(\tau_n-1)}(X_i^{(n)} - \frac{1}{\gamma})\sum_{k=i}^{\lfloor nt\rfloor\wedge(\tau_n-1)}\left(n-\frac{k}{\gamma}\right)^{-\alpha} = n^{1-\alpha}J_{n,1}\mathbf{1}_{\{nt<\tau_n\}} + J_n(t)\mathbf{1}_{\{nt\geq\tau_n\}}, \tag{5.39}$$

with $J_{n,1} = \sum_{i=1}^{\lfloor nt\rfloor\wedge(\tau_n-1)} f_n(i/n)(X_i^{(n)} - \frac{1}{\gamma})$ and $f_n(r) = \frac{1}{n}\sum_{j=\lfloor nr\rfloor}^{\lfloor nt\rfloor}\left(1-\frac{j}{n\gamma}\right)^{-\alpha}$. The functions f_n are finite and uniformly bounded as for $n \geq 2/\varepsilon_2$,

$$0 \leq f_n(r) \leq f_n(0) = \frac{1}{n}\sum_{k=0}^{\lfloor nt\rfloor}\left(1-\frac{k}{n\gamma}\right)^{-\alpha} \leq \int_0^{\gamma-\varepsilon_2/2}\left(1-\frac{s}{\gamma}\right)^{-\alpha} ds < \infty.$$

Notice that

$$\kappa = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{\lfloor n\gamma \rfloor} f_n(k/n)^\alpha = \int_0^t dr \left(\int_r^t (1 - \frac{s}{\gamma})^{-\alpha}\, ds \right)^\alpha .$$

We deduce from Proposition 5.2 that $(n^{-\frac{1}{\alpha}} J_{n,1}, n \geq 2)$ converges in distribution to $\kappa^{1/\alpha} V_1$. For $\varepsilon' > 0$, we have $\mathbb{P}(\mathbf{1}_{\{nt \geq \tau_n\}} |J_n(t)| \geq \varepsilon') \leq \mathbb{P}(nt \geq \tau_n)$. Then we use (5.39) and (5.37) to conclude that the following convergence in distribution holds:

$$n^{-1+\alpha-1/\alpha} J_n(t) \xrightarrow[n\to\infty]{\text{(d)}} \kappa^{1/\alpha} V_1. \tag{5.40}$$

Convergence of $R_n(t)$. Let $t \in [0, t_0]$. We shall now prove that $n^{-1+\alpha-1/\alpha} R_n(t)$ converges to 0 in probability. Let $\varepsilon \in (0, \gamma)$. We have $R_n(t) = R_{n,1} + R_{n,2}$, with

$$\begin{aligned}
R_{n,1} &= \sum_{k=1}^{\lfloor nt \rfloor} \left(n - \frac{k}{\gamma} \right)^{-\gamma} \mathbf{1}_{\{k<\tau_n\}} R_{n,1,k}, \\
R_{n,1,k} &= \mathbf{1}_{\{\Delta_{n,k} < 1-\varepsilon\}} \int_0^{\Delta_{n,k}} (\Delta_{n,k} - s)\, (1-s)^{-\gamma-2}\, ds, \\
R_{n,2} &= \sum_{k=1}^{\lfloor nt \rfloor} \left(n - \frac{k}{\gamma} \right)^{-\gamma} \mathbf{1}_{\{k<\tau_n\}} \mathbf{1}_{\{\Delta_{n,k} \geq 1-\varepsilon\}} \int_0^{\Delta_{n,k}} (\Delta_{n,k} - s)\, (1-s)^{-\gamma-2}\, ds.
\end{aligned}$$

We have

$$\mathbb{E}[|R_{n,1,k}|] \leq \frac{\varepsilon^{-\gamma-2}}{2}\, \mathbb{E}[(\Delta_{n,k})^2] \leq \frac{c}{n^2}\, \mathbb{E}\left[\left(\sum_{i=1}^{k} (X_i^{(n)} - \frac{1}{\gamma}) \right)^2 \right],$$

where we used that $k \leq n(\gamma - \varepsilon_2)$ for the last inequality and c depends on ε and ε_2.

Recall $\mathcal{Y} = (\mathcal{Y}_k, k \geq 0)$ is the filtration generated by Y. We consider the $\mathcal{Y}$-martingale $N_r = \sum_{j=1}^{r} \Delta N_r$, with $\Delta N_r = X_r^{(n)} - \mathbb{E}[X_r^{(n)} | \mathcal{Y}_{r-1}]$. We have

$$\mathbb{E}\left[\left(\sum_{i=1}^{k} (X_i^{(n)} - \frac{1}{\gamma}) \right)^2 \right] \leq 2\mathbb{E}\left[N_k^2\right] + 2\mathbb{E}\left[\left(\sum_{i=1}^{k} (\mathbb{E}[X_i^{(n)} | \mathcal{Y}_{i-1}] - \frac{1}{\gamma}) \right)^2 \right].$$

Notice that

$$\mathbb{E}\left[N_k^2\right] = \mathbb{E}\left[\sum_{i=1}^{k} (\Delta N_i)^2 \right] \leq \mathbb{E}\left[\sum_{i=1}^{k} \mathbb{E}[(X_i^{(n)})^2 | \mathcal{Y}_{i-1}] \right] \leq \mathbb{E}\left[\sum_{i=1}^{k} (X_i^{(n)})^2 \right].$$

Using that, conditionally on $\mathcal{Y}_{i-1}$, $X_i^{(n)}$ and $X_1^{(Y_{i-1})}$ have the same distribution, we get that

$$\mathbb{E}\left[N_k^2\right] \leq \sum_{j=1}^{n} \mathbb{E}[(X_1^{(j)})^2].$$

Thanks to (5.18) and (5.13), we deduce that

$$\mathbb{E}\left[N_k^2\right] \le C_{5.18} \sum_{j=1}^{n} \frac{j^2}{g_j} \le c \sum_{j=1}^{n} j^{2-\alpha} \le c\, n^{3-\alpha}.$$

Using (5.15) and (5.13), we get

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{i=1}^{k}(\mathbb{E}[X_i^{(n)}|\mathcal{Y}_{i-1}] - \frac{1}{\gamma})\right)^2\right] &\le \mathbb{E}\left[\left(\sum_{i=1}^{k}|\mathbb{E}[X_i^{(n)}|\mathcal{Y}_{i-1}] - \frac{1}{\gamma}|\right)^2\right] \\
&\le \mathbb{E}\left[\left(\sum_{i=1}^{k} C_{5.15}\varphi_{Y_{i-1}}\right)^2\right] \\
&\le c\,\left(\sum_{j=1}^{n}\varphi_j\right)^2 \le c\, n^{3-\alpha},
\end{aligned}$$

where for the last inequality we used (5.14) with $\varepsilon_0 > 0$ small enough (such that $1+2\varepsilon_0 < \alpha$) and the fact that $\zeta > 1 - 1/\alpha$ implies $2 - 2\zeta \le 3 - \alpha$ as $\alpha \in (1,2)$. This implies that for $k \le n(\gamma - \varepsilon_2) = nt_0$

$$\mathbb{E}\left[\left(\sum_{i=1}^{k}(X_i^{(n)} - \frac{1}{\gamma})\right)^2\right] \le cn^{3-\alpha} \tag{5.41}$$

therefore $\mathbb{E}[|R_{n,1,k}|] \le c\, n^{1-\alpha}$ and $\mathbb{E}[|R_{n,1}|] \le c\, n^{3-2\alpha}$. Thus, we get that $(n^{-1+\alpha-1/\alpha}R_{n,1}, n \ge 1)$ converges in probability to 0 since $-1+\alpha-1/\alpha+3-2\alpha = -(\alpha-1)^2/\alpha < 0$ for $\alpha > 1$.

We now consider $R_{n,2}$. Suppose that $k \le \lfloor nt \rfloor - 1$ satisfies $\Delta_{n,k} \ge 1 - \varepsilon$ on $\{nt < \tau_n\}$. Then on $\{nt < \tau_n\}$, we have

$$\Delta_{n,k+1} = \Delta_{n,k} + \frac{X_{k+1}^{(n)} - \frac{1}{\gamma} + \frac{\Delta_{n,k}}{\gamma}}{n - (k+1)/\gamma} \ge \Delta_{n,k} + \frac{X_{k+1}^{(n)} - \frac{\varepsilon}{\gamma}}{n - (k+1)/\gamma} \ge \Delta_{n,k},$$

where we used that $\gamma > \varepsilon$ for the first inequality and $X_{k+1}^{(n)} \ge 1$ for the last. In particular, on $\{nt < \tau_n\}$, if $\Delta_{n,k} \ge 1 - \varepsilon$ for some $k \le \lfloor nt \rfloor$, then we have $\Delta_{n,\lfloor nt \rfloor} \ge 1 - \varepsilon$. This implies that $\mathbf{1}_{\{nt<\tau_n\}}R_{n,2} = \mathbf{1}_{\{\Delta_{n,\lfloor nt\rfloor} \ge 1-\varepsilon\}}\mathbf{1}_{\{nt<\tau_n\}}R_{n,2}$. With the notations of Corollary 5.1, we have

$$\{nt < \tau_n\} \cap \{\Delta_{n,\lfloor nt \rfloor} \ge 1 - \varepsilon\} \subset \{V_t^{(n)} \ge (1-\varepsilon)(n - \frac{\lfloor nt \rfloor}{\gamma})n^{-1/\alpha}\} \subset \{n^{-1+1/\alpha}V_t^{(n)} \ge c\},$$

and then for any $\varepsilon' > 0$

$$\begin{aligned}
\mathbb{P}(n^{-1+\alpha-1/\alpha}|R_{n,2}| \ge \varepsilon',\ nt < \tau_n) &= \mathbb{P}(\mathbf{1}_{\{\Delta_{n,\lfloor nt\rfloor} \ge 1-\varepsilon\}} n^{-1+\alpha-1/\alpha}|R_{n,2}| \ge \varepsilon',\ nt < \tau_n) \\
&\le \mathbb{P}(\Delta_{n,\lfloor nt \rfloor} \ge 1 - \varepsilon, ,\ nt < \tau_n) \\
&\le \mathbb{P}(n^{-1+1/\alpha}V_t^{(n)} \ge c).
\end{aligned}$$

Use the convergence of $V_t^{(n)}$, see Corollary 5.1, to get that the right-hand side of the last inequality converges to 0 as n goes to infinity. Then notice that $\mathbb{P}(n^{-1+\alpha-1/\alpha}|R_{n,2}| \geq \varepsilon',\ nt \geq \tau_n) \leq \mathbb{P}(nt \geq \tau_n)$ which converges to 0 thanks to (5.37).

Thus $n^{-1+\alpha-1/\alpha}R_n(t)$ converges in probability to 0. As $t \mapsto R_n(t)$ is non-negative and non-decreasing, we conclude that

$$n^{-1+\alpha-1/\alpha} \sup_{0\leq t\leq t_0} R_n(t) \xrightarrow[n\to\infty]{\mathbb{P}} 0. \tag{5.42}$$

We deduce from (5.36), (5.38), (5.40) and (5.42) that

$$n^{-1+\alpha-1/\alpha}\left(\hat{L}_t^{(n)} - n^{2-\alpha}v(t)\right) \xrightarrow[n\to\infty]{\text{(d)}} \gamma\left[\int_0^t dr\ \left(\int_r^t (1-\frac{s}{\gamma})^{-\alpha}ds\right)^{\alpha}\right]^{1/\alpha} V_1. \tag{5.43}$$

To obtain (5.35), use (5.29) to get that $\gamma\left[\int_0^t dr\left(\int_r^t (1-\frac{s}{\gamma})^{-\alpha}ds\right)^{\alpha}\right]^{1/\alpha} V_1$ is distributed as $\gamma\int_0^t dV_r \int_r^t (1-\frac{s}{\gamma})^{-\alpha}ds$ which in turn is equal to $\int_0^t dr\ (1-\frac{r}{\gamma})^{-\alpha}V_r$.

To get (5.34), thanks to (5.36), (5.38) and (5.42), we have to check that $n^{\frac{\alpha-1}{2}-\delta-2+\alpha} \sup_{0\leq t\leq t_0} |J_n(t)|$ converges to 0 in probability for any $\delta > 0$. Notice that for $t \in [0, t_0]$,

$$\begin{aligned}
|J_n(t)| &\leq \sum_{k=1}^{\lfloor nt\rfloor\wedge(\tau_n-1)} \left(n-\frac{k}{\gamma}\right)^{-\gamma-1}\left|\sum_{i=1}^{k}(X_i^{(n)}-\frac{1}{\gamma})\right| \\
&\leq \sum_{k=1}^{\lfloor nt_0\rfloor\wedge(\tau_n-1)} \left(n-\frac{k}{\gamma}\right)^{-\gamma-1}\left|\sum_{i=1}^{k}(X_i^{(n)}-\frac{1}{\gamma})\right|.
\end{aligned}$$

Use (5.41) to deduce that

$$\mathbb{E}\left[\sup_{0\leq t\leq t_0}|J_n(t)|^2\right]^{1/2} \leq cn^{(5-3\alpha)/2}.$$

This implies that $n^{\frac{\alpha-1}{2}-\delta-2+\alpha} \sup_{0\leq t\leq t_0} |J_n(t)|$ converges to 0 in L^2 and thus in probability. This ends the proof of (5.34).

□

5.6 Proof of the main result

Let $\alpha_0 = \dfrac{1+\sqrt{5}}{2}$. Notice that for $\alpha \in (1, \alpha_0)$, we have $-1+\alpha-1/\alpha < 0$, whereas for $\alpha \geq \alpha_0$, $-1+\alpha-1/\alpha \geq 0$. Recall $\gamma = \alpha - 1$. We define $a(t)$ for $t \in [0, \gamma]$ by

$$a(t) = \frac{v(t)}{C_0\Gamma(2-\alpha)}, \quad \text{where} \quad v(t) = \int_0^t \left(1-\frac{r}{\gamma}\right)^{-\gamma} dr.$$

We also set $V_t^* = \dfrac{\alpha-1}{C_0\Gamma(2-\alpha)}\displaystyle\int_0^t (1-\frac{r}{\gamma})^{-\alpha} V_r\, dr$ for $t \in (0,\gamma)$. Let $x_+ = \max(x,0)$ denote the positive part of x.

Theorem 5.2. *Let $\alpha \in (1,2)$. We assume that $\rho(t) = C_0 t^{-\alpha} + O(t^{-\alpha+\zeta})$ for some $C_0 > 0$ and $\zeta > 1 - 1/\alpha$.*

(i) Let $t_0 \in [0,\gamma)$ and $\delta > 0$. We have the following convergence in probability:

$$n^{-\frac{(5-3\alpha)_+}{2}-\delta} \sup_{0\le t\le t_0} |L_t^{(n)} - n^{2-\alpha}a(t)| \xrightarrow[n\to\infty]{\mathbb{P}} 0. \tag{5.44}$$

In particular, we have $n^{-2+\alpha}L_t^{(n)} \xrightarrow[n\to\infty]{\mathbb{P}} a(t)$ for all $t \in [0,\gamma)$.

(ii) If $\alpha \in (1,\alpha_0)$, for $t \in (0,\gamma)$, the following convergence in distribution holds:

$$n^{-1+\alpha-1/\alpha}\left(L_t^{(n)} - a(t)n^{2-\alpha}\right) \xrightarrow[n\to\infty]{(d)} V_t^*. \tag{5.45}$$

Proof. First of all, let us consider the case $\alpha \in (1,\alpha_0)$. Lemma 5.8 and Tchebychev inequality imply that for $\alpha \in (1,\alpha_0)$, we have the following convergence in probability

$$\lim_{n\to\infty} n^{-1+\alpha-1/\alpha} \sup_{t\ge 0} |L_t^{(n)} - \tilde{L}_t^{(n)}| = 0.$$

This and Lemma 5.9 imply that for $\alpha \in (1,\alpha_0)$, we have the following convergence in probability

$$\lim_{n\to\infty} n^{-1+\alpha-1/\alpha} \sup_{t\ge 0} |L_t^{(n)} - \frac{\hat{L}_t^{(n)}}{C_0\Gamma(2-\alpha)}| = 0.$$

Then (5.45) and (5.44) for $\alpha \in (1,\alpha_0)$ are a direct consequence of Theorem 5.1.

For $\alpha \in [\alpha_0, 2)$, note that $\alpha > 3/2$ and $-1+\alpha-1/\alpha \ge 0$. As $\zeta > 1-1/\alpha$ and $\alpha \ge \alpha_0$ i.e. $1-1/\alpha \ge 2-\alpha$, we get $\zeta > 2-\alpha$. We then use Lemma 5.8, Lemma 5.9 (only with $\zeta > 2-\alpha$) and Theorem 5.1 to get (5.44) for $\alpha \in [\alpha_0, 2)$. □

Let $S_t^{(n)}$ be the total number of mutations up to the $\lfloor nt \rfloor$-th coalescence, for $t \in (0,\gamma)$. conditionally on $L_t^{(n)}$, $S_t^{(n)}$ is a Poisson r.v. with parameter $\theta L_t^{(n)}$. The next Corollary is a consequence of Theorem 5.2.

Corollary 5.2. *We assume that $\rho(t) = C_0 t^{-\alpha} + O(t^{-\alpha+\zeta})$ for some $C_0 > 0$ and $\zeta > 1 - 1/\alpha$. Let $t \in (0,\gamma)$ and G be a standard Gaussian r.v., independent of V.*

(i) For $\alpha \in (1,\sqrt{2})$, we have

$$n^{-1+\alpha-1/\alpha}(S_t^{(n)} - \theta a(t)n^{2-\alpha}) \xrightarrow[n\to\infty]{(d)} \theta V_t^*.$$

(*ii*) *For* $\alpha \in (\sqrt{2}, 2)$, *we have*

$$n^{-1+\alpha/2}(S_t^{(n)} - \theta a(t) n^{2-\alpha}) \xrightarrow[n\to\infty]{(d)} \sqrt{\theta a(t)} G.$$

(*iii*) *For* $\alpha = \sqrt{2}$, *we have* $-1 + \alpha - \frac{1}{\alpha} = 1 - \frac{\alpha}{2}$ *and*

$$n^{-1+\alpha-1/\alpha}(S_t^{(n)} - \theta a(t) n^{2-\alpha}) \xrightarrow[n\to\infty]{(d)} \theta V_t^* + \sqrt{\theta a(t)} G.$$

Proof. Let us compute the characteristic function $\psi_n(u, v)$ of the 2-dimensional r.v. (G_n, H_n) with

$$G_n = \frac{S_t^{(n)} - \theta L_t^{(n)}}{\sqrt{\theta a(t) n^{2-\alpha}}} \quad \text{and} \quad H_n = n^{-1+\alpha-1/\alpha} \left(L_t^{(n)} - a(t) n^{2-\alpha} \right).$$

Using that, conditionally on $L_t^{(n)}$, the law of $S_t^{(n)}$ is a Poisson distribution with parameter $\theta L_t^{(n)}$, we have

$$\psi_n(u, v) = \mathbb{E}\left[\mathrm{e}^{iuG_n} \, \mathrm{e}^{ivH_n} \right] = \mathbb{E}\left[\mathrm{e}^{-\theta L_t^{(n)} \left(1 - \mathrm{e}^{iu/\sqrt{\theta a(t) n^{2-\alpha}}} + iu/\sqrt{\theta a(t) n^{2-\alpha}} \right)} \, \mathrm{e}^{ivH_n} \right].$$

Using (i) of Theorem 5.2, we get that

$$-\theta L_t^{(n)} \left(1 - \mathrm{e}^{iu/\sqrt{\theta a(t) n^{2-\alpha}}} + iu/\sqrt{\theta a(t) n^{2-\alpha}} \right)$$

tends to $-u^2/2$ in probability and has a non-negative real part.

We first consider the case $\alpha \in (1, \sqrt{2}]$. We have $\sqrt{2} < \alpha_0$. Hence, applying (ii) of Theorem 5.2, we get that (G_n, H_n) converges in distribution to (G, V_t^*), where G is a standard Gaussian r.v. independent of V. Notice that

$$S_t^{(n)} = \theta a(t) n^{2-\alpha} + \theta n^{1-\alpha+1/\alpha} H_n + \sqrt{\theta a(t)} n^{1-\alpha/2} G_n$$

and then

$$n^{-1+\alpha-1/\alpha}(S_t^{(n)} - \theta a(t) n^{2-\alpha}) = \theta H_n + \sqrt{\theta a(t)} n^{\alpha/2-1/\alpha} G_n.$$

When $\alpha \leq \sqrt{2}$, notice that $\alpha/2 - 1/\alpha < 0$ (resp. $= 0$) if $\alpha < \sqrt{2}$ (resp. $\alpha = \sqrt{2}$). This gives (i) and (iii) of the Corollary.

Now we consider the case $\alpha \in (\sqrt{2}, 2)$. We write

$$n^{-1+\alpha/2}(S_t^{(n)} - \theta a(t) n^{2-\alpha}) = \sqrt{\theta a(t)} G_n + n^{-1+\alpha/2}(L_t^{(n)} - a(t) n^{2-\alpha}).$$

We still have that G_n converges in law to G. Moreover, the convergences (5.45) and (5.44) imply that $n^{-1+\alpha/2}(L_t^{(n)} - a(t) n^{2-\alpha})$ converges to 0 in probability. This gives (ii). □

Bibliographie

Basdevant, A.-L. and Goldschmidt, C. (2008). Asymptotics of the allele frequency spectrum associated with the Bolthausen-Sznitman coalescent. *Electron. J. Probab.*, 13 :486–512.

Berestycki, J., Berestycki, N., and Schweinsberg, J. (2007). Beta-coalescents and continuous stable random trees. *Ann. Probab.*, 35(5) :1835–1887.

Berestycki, J., Berestycki, N., and Schweinsberg, J. (2008). Small-time behavior of Beta-coalescents. *Ann. Inst. H. Poincaré Probab. Stat.*, 44(2) :214–238.

Bertoin, J. (1996). *Lévy processes*. Cambridge University Press, Cambridge.

Bertoin, J. and Le Gall, J.-F. (2006). Stochastic flows associated to coalescent processes. III. Limit theorems. *Illinois J. Math.*, 50(1-4) :147–181.

Birkner, M., Blath, J., Capaldo, M., Etheridge, A. M., Möhle, M., Schweinsberg, J., and Wakolbinger, A. (2005). Alpha-stable branching and Beta-coalescents. *Electron. J. Probab.*, 10(9) :303–325.

Bolthausen, E. and Sznitman, A.-S. (1998). On Ruelle's probability cascades and an abstract cavity method. *Comm. Math. Phys.*, 197(2) :247–276.

Boom, J. D. G., Boulding, E., and Beckenbach, A. (1994). Mitochondrial DNA variation in introduced populations of pacific oyster, crassostrea gigas, in british columbia. *Can. J. Fish. Aquat. Sci.*, 51 :1608–1614.

Breiman, L. (1992). *Probability*, volume 7 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. Reprint of the 1968 edition.

Drmota, M., Iksanov, A., Möhle, M., and Rösler, U. (2007). Asymptotic results concerning the total branch length of the Bolthausen-Sznitman coalescent. *Stochastic Process. Appl.*, 117(10) :1404–1421.

Eldon, B. and Wakeley, J. (2006). Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172 :2621–2633.

Feller, W. (1971). *An introduction to probability theory and its applications. Vol. II.* John Wiley & Sons Inc., New York, NY. second edition.

Gnedin, A. and Yakubovich, Y. (2007). On the number of collisions in Λ-coalescents. *Electron. J. Probab.*, 12(56) :1547–1567.

Iksanov, A. and Möhle, M. (2008). On the number of jumps of random walks with a barrier. *Adv. in Appl. Probab.*, 40(1) :206–228.

Kallenberg, O. (2002). *Foundations of modern probability.* Probability and its Applications. Springer-Verlag, New York, NY. second edition.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61 :893–903.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Process. Appl.*, 13(3) :235–248.

Kingman, J. F. C. (2000). Origins of the coalescent 1974–1982. *Genetics*, 156 :1461–1463.

Möhle, M. (2006). On the number of segregating sites for populations with large family sizes. *Adv. in Appl. Probab.*, 38(3) :750–767.

Mukherjea, A., Rao, M., and Suen, S. (2006). A note on moment generating functions. *Statist. Probab. Lett.*, 76(11) :1185–1189.

Pitman, J. (1999). Coalescents with multiple collisions. *Ann. Probab.*, 27(4) :1870–1902.

Sagitov, S. (1999). The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.*, 36(4) :1116–1125.

Schweinsberg, J. (2003). Coalescent processes obtained from supercritical Galton-Watson processes. *Stochastic Process. Appl.*, 106(1) :107–139.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoret. Population Biol.*, 7 :256–276.

Bibliographie

Aldous, D. J. (1997). Brownian excursions, critical random graphs and the multiplicative coalescent. *Ann. Probab.*, 25:812–854.

Aldous, D. J. (1999). Deterministic and stochastic models for coalescence (aggregation and coagulation) : a review of the mean-field theory for probabilists. *Bernoulli*, 5(1):3–48.

Aldous, D. J. et Limic, V. (1998). The entrance boundary of the multiplicative coalescent. *Electron. J. Probab.*, 3:1–59.

Aldous, D. J. et Pitman, J. (1998). The standard additive coalescent. *Ann. Probab.*, 26:1703–1726.

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H. L., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. H., Staden, R. et Young, I. G. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, 290:457–465.

Barton, N. H., Briggs, E. G., Eisen, J. A., Goldstein, D. B. et Patel, N. H. (2007). *Evolution*. Cold Spring Harbour Laboratory Press, Cold Spring Harbor, NY.

Barton, N. H. et Etheridge, A. M. (2004). The effect of selection on genealogies. *Genetics*, 166(2):1115–1131.

Barton, N. H., Etheridge, A. M. et Sturm, A. K. (2004). Coalescence in a random background. *Ann. Appl. Probab.*, 14(2):754–785.

Basdevant, A.-L. et Goldschmidt, C. (2008). Asymptotics of the allele frequency spectrum associated with the Bolthausen-Sznitman coalescent. *Electron. J. Probab.*, 13:486–512.

Berestycki, J., Berestycki, N. et Limic, V. (2009). The Λ-coalescent speed of coming down from infinity. *http ://arxiv.org/abs/0807.4278. To appear.*

Berestycki, J., Berestycki, N. et Schweinsberg, J. (2007). Beta-coalescents and continuous stable random trees. *Ann. Probab.*, 35(5):1835–1887.

Berestycki, J., Berestycki, N. et Schweinsberg, J. (2008). Small-time behavior of Beta-coalescents. *Ann. Inst. H. Poincaré Probab. Stat.*, 44(2):214–238.

Berestycki, N. (2009). *Recent progress in coalescent theory. www.statslab.cam.ac.uk/ beresty/rp2.pdf. Work in progress.*

Bertoin, J. (1996). *Lévy processes*. Cambridge University Press, Cambridge.

Bertoin, J. (2006). *Random fragmentation and coagulation processes*, volume 102 de *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge.

Bertoin, J. et Le Gall, J.-F. (2000). The Bolthausen-Sznitman coalescent and the genealogy of continuous-state branching processes. *Probab. Theory Related Fields*, 117(2):249–266.

Bertoin, J. et Le Gall, J.-F. (2003). Stochastic flows associated to coalescent processes. *Probab. Theory Related Fields*, 126(2):261–288.

Bertoin, J. et Le Gall, J.-F. (2006). Stochastic flows associated to coalescent processes. III. Limit theorems. *Illinois J. Math.*, 50(1-4):147–181.

Birkner, M. (2005). Stochastic models from population biology. *http : //evol.bio.lmu.de/ birkner/lehre$_a$rchiv/smpb$_S$S07. To appear.*

Birkner, M. et Blath, J. (2007). Rescaled stable generalised Fleming-Viot processes : Flickering random measures. *http ://www.wias-berlin.de/main/publications/wias-publ/ run.cgi ?template=abstract&type=Preprint&year=2007&number=1252. To appear.*

Birkner, M., Blath, J., Capaldo, M., Etheridge, A. M., Möhle, M., Schweinsberg, J. et Wakolbinger, A. (2005). Alpha-stable branching and Beta-coalescents. *Electron. J. Probab.*, 10(9):303–325.

Blum, M. G. B. et François, O. (2005). Minimal clade size and external branch length under the neutral coalescent. *Adv. in Appl. Probab.*, 37(3):647–662.

Bolthausen, E. et Sznitman, A.-S. (1998). On Ruelle's probability cascades and an abstract cavity method. *Comm. Math. Phys.*, 197(2):247–276.

Boom, J. D. G., Boulding, E. et Beckenbach, A. (1994). Mitochondrial DNA variation in introduced populations of pacific oyster, crassostrea gigas, in british columbia. *Can. J. Fish. Aquat. Sci.*, 51:1608–1614.

Breiman, L. (1992). *Probability*, volume 7 de *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. Reprint of the 1968 edition.

Brémaud, P., Kannurpatti, R. et Mazumdar, R. (1992). Event and time averages : a review. *Adv. in Appl. Probab.*, 24(2):377–411.

Brunet, E., Derrida, B., Mueller, A. H. et Munier, S. (2006). Noisy traveling waves : effect of selection on genealogies. *Europhys. Lett.*, 76(1):1–7.

Brunet, É., Derrida, B., Mueller, A. H. et Munier, S. (2007). Effect of selection on ancestry : an exactly soluble case and its phenomenological generalization. *Phys. Rev. E (3)*, 76(4):041104, 20.

Caliebe, A., Neininger, R., Krawczak, M. et Rösler, U. (2007). On the length distribution of external branches in coalescence trees : genetic diversity within species. *Theoret. Population Biol.*, 72(2):245–252.

Cannings, C. (1974). The latent roots of certain Markov chains arising in genetics : a new approach. I. Haploid models. *Adv. in Appl. Probab.*, 6:260–290.

Cannings, C. (1975). The latent roots of certain Markov chains arising in genetics : a new approach. II. Further haploid models. *Adv. in Appl. Probab.*, 7:264–282.

Cattiaux, P., Collet, P., Lambert, A., Martinez, S., Méléard, S. et San Martin, J. (2009). Quasi-stationarity distributions and diffusion models in population dynamics. *http ://arxiv.org/abs/math/0703781. To appear.*

Chang, J. T. (1999). Recent common ancestors of all present-day individuals. *Adv. in Appl. Probab.*, 31(4):1002–1038.

Collet, P., Martínez, S. et Maume-Deschamps, V. (2000). On the existence of conditionally invariant probability measures in dynamical systems. *Nonlinearity*, 13(4): 1263–1274.

Darden, T., Kaplan, N. L. et Hudson, R. R. (1989). A numerical method for calculating moments of coalescent times in finite populations with selection. *J. Math. Biol.*, 27(3):355–368.

Darroch, J. N. et Seneta, E. (1965). On quasi-stationary distributions in absorbing discrete-time finite Markov chains. *J. Appl. Probability*, 2:88–100.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection.* John Murray, London.

Dawson, D. A. (1993). Measure-valued Markov processes. *In École d'Été de Probabilités de Saint-Flour XXI—1991*, volume 1541 de *Lecture Notes in Math.*, pages 1–260. Springer, Berlin.

Dawson, D. A. et Hochberg, K. J. (1982). Wandering random measures in the Fleming-Viot model. *Ann. Probab.*, 10(3):554–580.

Delmas, J.-F., Dhersin, J.-S. et Siri-Jegousse, A. (2008). Asymptotic results on the length of coalescent trees. *Ann. Appl. Probab.*, 18(3):997–1025.

Delmas, J.-F., Dhersin, J.-S. et Siri-Jegousse, A. (2009). On the two oldest families in a Wright-Fisher process. *To appear in Electron. J. Probab.*

Donnelly, P. et Kurtz, T. G. (1996). A countable representation of the Fleming-Viot measurable diffusion. *Ann. Probab.*, 24(2):698–742.

Donnelly, P. et Kurtz, T. G. (1999). Particle representations for measure-valued population models. *Ann. Probab.*, 27(1):166–205.

Donnelly, P. et Kurtz, T. G. (2006). The Eve process. Manuscript, personal communication.

Donnelly, P. et Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. *Ann. Rev. Genet.*, 29:401–421.

Doob, J. L. (2001). *Classical potential theory and its probabilistic counterpart*. Classics in Mathematics. Springer-Verlag, Berlin. Reprint of the 1984 edition.

Drmota, M., Iksanov, A., Möhle, M. et Rösler, U. (2007). Asymptotic results concerning the total branch length of the Bolthausen-Sznitman coalescent. *Stochastic Process. Appl.*, 117(10):1404–1421.

Drmota, M., Iksanov, A., Möhle, M. et Rösler, U. (2009). A limiting distribution for the number of cuts needed to isolate the root of a random recursive tree. *Random Structures Algorithms*, 34(3):319–336.

Durrett, R. (1996). *Stochastic calculus : a practical introduction*. Probability and Stochastics Series. CRC Press, Boca Raton,FL.

Durrett, R. (2008). *Probability models for DNA sequence evolution*. Probability and its Applications. Springer, New York, NY. second edition.

Dynkin, E. B. (2006). *Theory of Markov processes*. Dover Publications Inc., Mineola, NY. Reprint of the 1961 edition.

Eldon, B. et Wakeley, J. (2006). Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172:2621–2633.

Etheridge, A. M. (2000). *An introduction to superprocesses*, volume 20 de *University Lecture Series*. American Mathematical Society, Providence, RI.

Etheridge, A. M. (2009). Some mathematical models from population genetics. *In École d'Été de Probabilités de Saint-Flour XXXIX—2009. To appear.*

Ethier, S. N. et Kurtz, T. G. (1986). *Markov processes : characterization and convergence*. Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, NY.

Evans, S. N. et Pitman, J. (1997). Construction of markovian coalescents. *Ann. Inst. H. Poincaré Probab. Stat.*, 34:339–383.

Evans, S. N. et Ralph, P. L. (2008). Dynamics of the time to the most recent common ancestor in a large branching population. *http ://arxiv.org/abs/0812.1302. To appear.*

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoret. Population Biol.*, 3(1):87–112.

Ewens, W. J. (2004). *Mathematical population genetics. I. Theoretical introduction*, volume 27 de *Interdisciplinary Applied Mathematics*. Springer-Verlag, New York, NY. second edition.

Feller, W. (1951). Two singular diffusion problems. *Ann. Math.*, 54(1):173–182.

Feller, W. (1971). *An introduction to probability theory and its applications. Vol. II.* John Wiley & Sons Inc., New York, NY. second edition.

Ferrari, P. A., Kesten, H., Martinez, S. et Picco, P. (1995). Existence of quasi-stationary distributions. A renewal dynamical approach. *Ann. Probab.*, 23(2):501–521.

Fisher, R. (1930). *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford.

Fleming, W. H. et Viot, M. (1978). Some measure-valued population processes. *In Stochastic analysis (Proc. Internat. Conf., Northwestern Univ., Evanston, Ill., 1978)*, pages 97–108. Academic Press, New York, NY.

Fleming, W. H. et Viot, M. (1979). Some measure-valued Markov processes in population genetics theory. *Indiana Univ. Math. J.*, 28(5):817–843.

Freund, F. et Möhle, M. (2009). On the time back to the most recent common ancestor and the external branch length of the Bolthausen-Sznitman coalescent. *Markov Process. Related Fields*, 15. *À paraître.*

Fu, Y. X. (2006). Exact coalescent for the Wright-Fisher model. *Theoret. Population Biol.*., 69(3):1385–394.

Fu, Y. X. et Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics*, 133:693–709.

Gnedin, A., Iksanov, A. et Möhle, M. (2008). On asymptotics of exchangeable coalescents with multiple collisions. *J. Appl. Probab.*, 45(4):1186–1195.

Gnedin, A. et Yakubovich, Y. (2007). On the number of collisions in Λ-coalescents. *Electron. J. Probab.*, 12(56):1547–1567.

Greven, A., Pfaffelhuber, P. et Winter, A. (2009). Tree-valued resampling dynamics : martingale problems and applications. *http ://arxiv.org/abs/0806.2224. To appear.*

Griffiths, R. C. (1980). Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theoret. Population Biol.*, 17(1):37–50.

Hedgecock, D. (1994). *Genetics and evolution of aquatic organisms*, chapitre Does variance in reproductive success limit effective population size of marine organisms ?, page 122–134. Chapman and Hall, London.

Hudson, R. R. (1991). Gene genealogies and the coalescent process. *Oxford Survey Evol. Biol.*, 7:1–44.

Huillet, T. (2007). On Wright–Fisher diffusion and its relatives. *J. Stat. Mech.*, P1106.

Iksanov, A. et Möhle, M. (2008). On the number of jumps of random walks with a barrier. *Adv. in Appl. Probab.*, 40(1):206–228.

Johnson, N. L. et Kotz, S. (1977). *Urn models and their application*. John Wiley & Sons Inc., New York, NY.

Kallenberg, O. (2002). *Foundations of modern probability*. Probability and its Applications. Springer-Verlag, New York, NY. second edition.

Kaplan, N. L., Darden, T. et Hudson, R. R. (1988). The coalescent process in models with selection. *Genetics*, 120(3):819–829.

Karatzas, I. et Shreve, S. E. (1991). *Brownian motion and stochastic calculus*, volume 113 de *Graduate Texts in Mathematics*. Springer-Verlag, New York, NY. second edition.

Karlin, S. et Taylor, H. M. (1981). *A second course in stochastic processes*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, NY.

Kimura, M. (1955a). Random genetic drift in a multi-allelic locus. *Evolution*, 9:419–435.

Kimura, M. (1955b). Solution of a process of random genetic drift with a continuous model. *Proceedings. National Academy of Sciences (United States of America)*, 41:144–150.

Kimura, M. (1955c). Stochastic processes and distribution of gene frequencies under natural selection. *Cold Spring Harbor Symposia on Quantitative Biology*, 20:33–53.

Kimura, M. (1957). Some problems of stochastic processes in genetics. *Ann. Math. Statist.*, 28:882–901.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61:893–903.

Kimura, M. (1971). Theoretical foundations of population genetics at the molecular level. *Theoret. Population Biol.*, 2:174–208.

Kimura, M. et Crow, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics*, 49:725–738.

Kimura, M. et Ohta, T. (1969a). The average number of generations until extinction of an individual mutant gene in a finite populationt. *Genetics*, 61:763–771.

Kimura, M. et Ohta, T. (1969b). The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, 63(3):701–709.

Kingman, J. F. C. (1982a). The coalescent. *Stochastic Process. Appl.*, 13(3):235–248.

Kingman, J. F. C. (1982b). Exchangeability and the evolution of large populations. *In Exchangeability in probability and statistics (Rome, 1981)*, pages 97–112. North-Holland, Amsterdam.

Kingman, J. F. C. (1982c). On the genealogy of large populations. *J. Appl. Probab.*, 19A:27–43.

Kingman, J. F. C. (2000). Origins of the coalescent 1974–1982. *Genetics*, 156:1461–1463.

Knight, F. B. (1981). *Essentials of Brownian motion and diffusion*, volume 18 de *Mathematical Surveys*. American Mathematical Society, Providence, R.I.

Lambert, A. (2008). Population dynamics and random genealogies. *Stoch. Models*, 24(suppl. 1):45–163.

Leocard, S. (2009). *Modèles probabilistes du balayage sélectif et auto-stop génétique*. Thèse de doctorat, Université Aix Marseille.

Li, W. et Fu, Y. (1999). Coalescent theory and its applications in population genetics. *In Statistics in Genetics*, pages 45–79. Springer-Verlag, New York, NY.

Mendel, G. (1866). Versuche über pflanzen-hybriden. *Verh. Naturforsch. Ver. Brünn*, 4:3–47.

Meyn, S. P. et Tweedie, R. L. (1993). *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag, London.

Möhle, M. (2000). Total variation distances and rates of convergence for ancestral coalescent processes in exchangeable population models. *Adv. in Appl. Probab.*, 32(4):983–993.

Möhle, M. (2004). The time back to the most recent common ancestor in exchangeable population models. *Adv. in Appl. Probab.*, 36(1):78–97.

Möhle, M. (2006). On the number of segregating sites for populations with large family sizes. *Adv. in Appl. Probab.*, 38(3):750–767.

Möhle, M. et Sagitov, S. (2001). A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.*, 29(4):1547–1562.

Möhle, M. et Sagitov, S. (2003). Coalescent patterns in diploid exchangeable population models. *J. Math. Biol.*, 47(4):337–352.

Moran, P. A. P. (1958). Random processes in genetics. *Proc. Cambridge Philos. Soc.*, 54:60–71.

Mukherjea, A., Rao, M. et Suen, S. (2006). A note on moment generating functions. *Statist. Probab. Lett.*, 76(11):1185–1189.

Nordborg, M. (2001). Coalescent theory. *In Handbook of Statistical Genetics*, pages 179–212. John Wiley & Sons Inc., Chichester.

Nordborg, M. et Donnelly, P. (1997). The coalescent process with selfing. *Genetics*, 146(3):1185–1195.

Nordborg, M. et Krone, S. (2002). Separation of time scales and convergence to the coalescent in structured populations. *In Modern Developments in Theoretical Population Genetics : The Legacy of Gustave Malécot*, pages 194–232. Oxford University Press, Oxford.

Panholzer, A. (2004). Destruction of recursive trees. *In Mathematics and computer science. III*, Trends Math., pages 267–280. Birkhäuser, Basel.

Perkins, E. (2002). Dawson-Watanabe superprocesses and measure-valued diffusions. *In École d'Été de Probabilités de Saint-Flour XXIX—1999*, volume 1781 de *Lecture Notes in Math.*, pages 125–324. Springer, Berlin.

Pfaffelhuber, P. et Wakolbinger, A. (2006). The process of most recent common ancestors in an evolving coalescent. *Stochastic Process. Appl.*, 16(12):1836–1859.

Pitman, J. (1999). Coalescents with multiple collisions. *Ann. Probab.*, 27(4):1870–1902.

Pitman, J. (2006). Combinatorial stochastic processes. *In École d'Été de Probabilités de Saint-Flour XXXII—2002*, volume 1875 de *Lecture Notes in Math.*, pages 1–256. Springer-Verlag, Berlin.

Rauch, E. et Bar-Yam, Y. (2004). Theory predicts the uneven distribution of genetic diversity within species. *Nature*, 431:449–452.

Revuz, D. et Yor, M. (1999). *Continuous martingales and Brownian motion*, volume 293 de *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin. Third.

Sagitov, S. (1999). The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.*, 36(4):1116–1125.

Schweinsberg, J. (2000). Coalescents with simultaneous multiple collisions. *Electron. J. Probab.*, 5:1–50.

Schweinsberg, J. (2003). Coalescent processes obtained from supercritical Galton-Watson processes. *Stochastic Process. Appl.*, 106(1):107–139.

Simon, D. et Derrida, B. (2006). Evolution of the most recent common ancestor of a population with no selection. *J. Stat. Mech.*, P05002.

Sjödin, P., Kaj, I., Krone, S., Lascoux, M. et Nordborg, M. (2005). On the meaning and existence of an effective population size. *Genetics*, 169:1061–1070.

Steinsaltz, D. et Evans, S. N. (2007). Quasistationary distributions for one-dimensional diffusions with killing. *Trans. Amer. Math. Soc.*, 359(3):1285–1324.

Stroock, D. W. et Varadhan, S. R. S. (2006). *Multidimensional diffusion processes*. Classics in Mathematics. Springer-Verlag, Berlin. Reprint of the 1997 edition.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123(3):585–595.

Tajima, F. (1999). Relationship between DNA polymorphism and fixation time. *Genetics*, 56:183–201.

Tavaré, S. (2004). Ancestral inference in population genetics. *In Lectures on probability theory and statistics*, volume 1837 de *Lecture Notes in Math.*, pages 1–188. Springer, Berlin.

Villemonais, D. (2009). Approximation of quasi-stationary distributions for 1-dimensional killed diffusions with unbounded drifts. *http ://arxiv.org/abs/0905.3636. To appear.*

Watanabe, S. (1968). A limit theorem of branching processes and continuous state branching processes. *J. Math. Kyoto Univ.*, 8:141–167.

Watson, H. W. et Galton, F. (1875). On the probability of the extinction of families. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 4:138–144.

Watson, J. D. et Crick, F. H. C. (1953). A structure for deoxyribose nucleic acid. *Nature*, 171:737–738.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoret. Population Biol.*, 7:256–276.

Wiuf, C. et Donnelly, P. (1990). Conditional genealogies and the age of a neutral mutant. *Theoret. Population Biol.*, 125:447–454.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16(2):97–159.

Árnason, E. (2004). Mitochondrial cytochrome b dna variation in the high-fecundity atlantic cod : Trans-atlantic clines and shallow gene genealogy. *Genetics*, 166:1871–1885.