Modèles probabilistes de populations : branchement avec catastrophes et signature génétique de la sélection

Charline Smadi

UNIVERSITÉ PARIS-EST

THÈSE

présentée pour obtenir

LE GRADE DE DOCTEUR EN SCIENCES DE L'UNIVERSITÉ PARIS-EST

Spécialité : Mathématiques

par

Charline SMADI

Modèles probabilistes de populations : branchement avec catastrophes et signature génétique de la sélection

Soutenue le 5 mars 2015 devant un jury composé de :

M. Jean BERTOIN	rapporteur
M. Sylvain BILLIARD	examinateur
M. Loïc CHAUMONT	examinateur
M. Jean-François DELMAS	directeur de thèse
Mme. Alison ETHERIDGE	rapportrice
Mme. Sylvie MÉLÉARD	directrice de thèse

Remerciements

En premier lieu, ma gratitude va à mes directeurs, Jean-François Delmas et Sylvie Méléard, qui y sont pour beaucoup dans mon désir de faire des probabilités. C'est en suivant leurs enseignements à l'Ecole Polytechnique que poursuivre par une thèse en probabilités m'est apparu comme une évidence. Ils ont su me guider et me soutenir durant ces trois années et ont eu le courage de se plonger dans ma prose enchevêtrée et peu compréhensible. Un grand merci également à Vincent pour le temps qu'il m'a consacré, pour son désir incessant de transmettre et pour ses réponses patientes à mes questions.

Je suis extrêmement reconnaissante envers Jean Bertoin et Alison Etheridge, qui m'ont fait l'honneur de rapporter cette thèse, malgré leurs emplois du temps surchargés, et envers Sylvain Billiard et Loic Chaumont qui ont accepté de faire partie de mon jury.

Merci à ceux qui m'ont accompagnée et supportée au quotidien durant ces trois années, en particulier ceux avec qui j'ai partagé mon bureau, Aline, Camille, Guillerme, Hélène, Manon, Romain, Zhenjie, et à ceux qui ont animé les pauses café ; j'y croisais en particulier (régulièrement) Antoine, Aymeric, Etienne(s), Gwenael, Laurent, Lucas, Maxime, Pascal et Xavier, ainsi que tous ceux qui ne sont pas nommés ici mais grâce auxquels j'ai passé de très bons moments.

J'ai eu la chance durant ma thèse d'enseigner à l'ENSTA et à l'Ecole des Ponts et Chaussées. Je remercie Jean-François Delmas, Jean-Stéphane Dhersin et Benjamin Jourdain de m'avoir accordé leur confiance.

Merci à Laurence, Chi et Sylvain d'avoir rendu mes séjours à Lille si sympathiques.

Julien (également expert en import/export), après toutes ces années passées à étudier ensemble, tu as fait le choix discutable de te tourner vers la physique. Si tu retrouves la raison et souhaites t'intéresser à la biologie, je serai très heureuse de te guider dans cette voie.

Mon séjour à Gottingen a été très enrichissant, et j'en suis redevable à Anja Sturm et Rebekka Brink-Spalink.

Mes remerciements vont aussi bien sûr à Juan Carlos, qui sait rester mexicain en toutes circonstances et a une connaissance encyclopédique des processus stables.

Merci à Vladimir, pour son accueil en Russie, sa gentillesse, sa grande patience lorsque je parle russe, et ses encouragements.

Merci également à tous les membres du CMAP et du CERMICS, et en particulier au secrétaires pour leur gentillesse et leur efficacité, à Sylvain pour ses salutaires debuggages, à Sylvie Cach pour sa bienveillance malgré mes flirts constants avec les deadlines, et au groupe PEIPS pour la disponibilité de ses membres et la qualité scientifique de nos rencontres qui ont

constitué pour moi un cadre de recherche particulièrement stimulant.

Merci à ma famille pour son affection et son support indéfectible. Enfin et surtout merci à Pierre pour tout ce que nous avons vécu et pour tout ce qu'il nous reste à construire.

Résumé

Cette thèse porte sur l'étude probabiliste des réponses démographique et génétique de populations à certains événements ponctuels. Elle a donné lieu à trois travaux. Les deux premiers ont été publiés :

- [BPS13] Vincent Bansaye, Juan Carlos Pardo Millan, and Charline Smadi, *On the extinction of continuous state branching processes with catastrophes*, Electron. J. Probab. 18 (2013), no. 106, 1–31
- [Sma14] Charline Smadi, *An eco-evolutionary approach of adaptation and recombination in a large population of varying size*, Stochastic Processes and their Applications, DOI : 10.1016/j.spa.2014.12.007
- [BSS15a] Rebekka Brink-Spalink and Charline Smadi, *Genealogies of two neutral loci after a selective sweep in a large population of varying size*, in preparation

Ces trois travaux constituent, après une introduction au sujet, les trois chapitres de cette thèse.

Processus de branchement à états continus avec catastrophes

Les processus de branchement ont été introduits en temps et espace discrets au 19ème siècle pour modéliser la dynamique des noms de famille illustres en Grande Bretagne. Les individus meurent et se reproduisent de manière indépendante et identiquement distribuée ; il n'y a donc pas d'interactions entre eux, et en particulier pas de limitation de ressources. Ces processus ont été particulièrement utilisés en biologie, car l'indépendance des comportements individuels permet des calculs explicites. Afin de modéliser de grandes populations d'individus de petite taille, des processus de branchement à états continus ont été construits, comme limites d'échelle des processus discrets initiaux. Ils peuvent être réalisés en particulier comme l'unique solution forte de l'équation différentielle stochastique :

$$Z_t = Z_0 + \int_0^t g Z_s ds + \int_0^t \sqrt{2\sigma^2 Z_s} dB_s + \int_0^t \int_0^\infty \int_0^{Z_{s-}} z \widetilde{N}_0(ds, dz, du), \tag{1}$$

où $Z_0 > 0$ p.s. est la taille initiale de la population, B est un mouvement Brownien standard, $N_0(ds, dz, du)$ une mesure aléatoire de Poisson d'intensité $ds\mu(dz)du$ (où μ est une mesure

à support dans $(0,\infty)$ qui vérifie $\int_{(0,\infty)}(z\wedge z^2)\mu(dz))$ indépendante de B, et $\widetilde{N}_0$ la mesure compensée de N_0. Le dernier terme est un terme de sauts dont l'intensité est proportionnelle à la taille de la population. Il représente les gros événements de naissance, lorsqu'un individu "infinitésimal" donne naissance à un nombre d'individus "infinitésimaux" si grand que cet événement augmente la taille de la population d'une quantité macroscopique.

Dans le Chapitre 2, nous avons cherché avec Vincent Bansaye et Juan Carlos Pardo à comprendre l'impact de catastrophes tuant une fraction de la population, ou d'événements d'immigration proportionnels à la taille de la population survenant de manière répétée. Nous avons dans un premier temps construit une nouvelle classe de processus, les processus de branchement à états continus avec catastrophes, en les réalisant comme l'unique solution forte (sous certaines conditions) de l'équation différentielle stochastique :

$$\begin{aligned} Y_t = Y_0 + \int_0^t gY_s ds + \int_0^t \sqrt{2\sigma^2 Y_s} dB_s \quad &+ \quad \int_0^t \int_{[0,\infty)} \int_0^{Y_{s-}} z\widetilde{N}_0(ds,dz,du) \\ &+ \quad \int_0^t \int_{[0,\infty)} \Big(m-1\Big) Y_{s-} N_1(ds,dm), \end{aligned} \tag{2}$$

où $Y_0 > 0$ p.s., et N_1 est une mesure aléatoire de Poisson sur $[0,\infty)^2$ d'intensité $dt\nu(dm)$.

Deux termes de sauts de natures différentes apparaissent dans cette équation. Le premier, dont nous avons déjà parlé, correspond aux grands événements de naissances et est lié à la démographie intrinsèque de la population. Le second correspond aux événements extérieurs. Lorsqu'une catastrophe survient, la taille de la population est multlipliée par un facteur $m \in (0,1)$, et lorsque qu'un événement d'immigration se produit elle est multlipliée par un facteur $m > 1$.

Nous avons déterminé les conditions sous lesquelles la taille de la population tend vers 0. Enfin, dans les cas d'absorption presque sûre nous avons déterminé la vitesse d'absorption asymptotique du processus.

Ce dernier résultat a une application directe à la détermination du nombre de cellules infectées dans un modèle d'infection de cellules par des parasites. En effet, la quantité de parasites dans une lignée cellulaire suit dans ce modèle un processus de branchement à états continus, et les "catastrophes" surviennent lorsque, lors des divisions cellulaires, la quantité de parasites est partagée entre les deux cellules filles.

Signature génétique de la sélection

Un individu est caractérisé par son matériel génétique. Celui-ci détermine (pour une grande partie) son phénotype et en particulier certains traits quantitatifs comme le taux de naissance, le taux de mort intrinsèque, ou sa capacité d'interaction avec les autres individus (compétition, coopération,...). Mais son génotype seul ne détermine pas son histoire de vie, le fait qu'il soit plus ou moins bien “adapté” que les autres individus : l'espérance de vie d'un humain par exemple, n'est pas fonction seulement de son génotype, mais aussi de l'environnement dans lequel il vit (accès à l'eau potable, à des infrastructures médicales, qualité et quantité de l'alimentation, ...). L'approche éco-évolutive cherche à prendre en compte l'importance de l'environnement en modélisant les interactions entre les individus. Ainsi, en fonction de la

composition de la population (nombre d'individus de chaque trait) et du milieu (quantité et qualité des ressources), un individu vivra plus ou moins longtemps et laissera plus ou moins de descendants.
Lorsqu'une mutation ou une modification de l'environnement survient (catastrophe climatique, déplacement de la population, ...), des allèles peuvent envahir la population et s'y fixer au détriment des autres allèles initialement présents : c'est le phénomène de balayage sélectif. Ces événements évolutifs laissent des traces dans la diversité neutre au voisinage du locus auquel l'allèle s'est fixé. En effet ce dernier "emmène" avec lui des allèles qui se trouvent sur les loci physiquement liés au locus sous sélection : c'est le phénomène d'auto-stop. La seule possibilité pour un locus de ne pas être "emmené" avec l'allèle sous sélection est l'occurence d'une recombination génétique, qui l'associe à un autre haplotype dans la population.
Dans le Chapitre 3, je quantifie la signature laissée par un tel sweep sélectif sur la diversité neutre. Je me concentre dans un premier temps sur la variation des proportions neutres dans les loci voisins du locus sous sélection sous différents scénarios de balayages : balayages doux, balayages durs dans un régime de recombinaisons rares ou fréquentes. Je montre que ces différents scenari évolutifs laissent des traces bien distinctes sur la diversité neutre, qui peuvent permettre de les discriminer. Dans le Chapitre 4, nous nous intéressons avec Rebekka Brink-Spalink aux généalogies jointes de deux loci neutres au voisinage du locus sous sélection dans le cas du balayage dur. Cela nous permet en particulier de quantifier des statistiques attendues sous certains scenari de sélection, qui sont utilisées par les biologistes pour détecter des événements de sélection dans l'histoire évolutive de populations à partir de données génétiques actuelles. Dans les Chapitres 3 et 4, la population évolue suivant un processus de naissance et mort avec compétition. Si un tel modèle est plus réaliste que les processus de branchement, la non-linéarité introduite par les compétitions entre individus en rend l'étude plus complexe.

Table of contents

CHAPITRE 1

Introduction

La première partie de cette thèse est consacrée au comportement en temps long de processus de branchement soumis à des catastrophes répétées qui suppriment une fraction de la population (catastrophes climatiques, divisions en plusieurs sous-populations, grands événements de prédation, ...). Un intérêt particulier est porté aux conditions de survie de la population et à la vitesse d'absorption dans le cas de l'absorption presque sûre.

La seconde partie est dédiée aux réponses démographique et génétique de populations à une modification des conditions de sélection (arrivée d'un mutant favorable ou déplacement de la population dans un nouvel environnement par exemple). En particulier on détermine la signature génétique laissée sur la diversité neutre par la fixation d'un allèle au détriment des autres, ce qui peut permettre de détecter une sélection récente grâce aux données génétiques des populations actuelles.

1.1 Extinction des processus de branchement à états continus avec catastrophes

La première partie de cette thèse, qui a fait l'objet d'une publication en collaboration avec Vincent Bansaye et Juan-Carlos Pardo [BPS13], traite des processus de branchement continus avec catastrophes. Après une description de la problématique et des travaux antérieurs nous présentons les résultats du Chapitre 2 et les perspectives de ce travail.

Processus de branchement discrets en environnement aléatoire

Processus de Galton-Watson

Ces processus trouvent leur origine dans l'étude des probabilités d'extinction des noms de famille illustres en Grande-Bretagne au dix-neuvième siècle. En 1873, Galton soumit le problème suivant à l'*Educational Time* [Gal73] :

> PROBLEM 4001 : A large nation, of whom we will only concern ourselves with adult males, N in number, and who each bear separate surnames colonise a district. Their law of population is such that, in each generation, a0 per cent of

> the adult males have no male children who reach adult life ; a1 have one such male child ; a2 have two ; and so on up to a5 who have five. Find (1) what proportion of their surnames will have become extinct after r generations ; and (2) how many instances there will be of the surname being held by m persons.

Galton ne reçut pas de solution satisfaisante, et demanda de l'aide à Watson, qui utilisa les fonctions génératrices pour étudier le problème [WG75]. Mais il fallut attendre les années 1930 pour que le problème soit complètement résolu, grâce à la contribution de Fisher, Haldane, Erlang et Steffenson [Erl29, Fis30, Ken66, GASK95]. Ce processus de branchement à temps discret est le processus connu sous le nom de processus de Galton-Watson, ou processus de Bienaymé-Galton-Watson, du nom d'un probabiliste et statisticien français qui avait étudié indépendamment ces processus dès 1845 [Bie45], et dont le travail était passé inaperçu.

La présentation qui suit est inspirée de l'ouvrage de Athreya et Ney [AN72]. Soit ξ une variable aléatoire à valeurs entières de loi $\mathbb{P}(\xi = k) = p_k$, $k \geq 0$, et Z_n la taille de population au temps n. La génération $n+1$ est composée des descendants des individus de la génération n, et conditionnellement à Z_n, l'individu i $(1 \leq i \leq Z_n)$ de la génération n engendre $\xi_{(i,n)}$ descendants où les $\xi_{(i,n)}$ sont indépendants et de même loi que ξ, ce qu'on peut écrire :

$$Z_{n+1} = \sum_{i=1}^{Z_n} \xi_{(i,n)}.$$

Ce procédé est ensuite itéré, les $\xi_{(i,n)}$ étant indépendants des $\xi_{(j,p)}$ pour $p \neq n$.

La chaîne de Markov $(Z_n, n \geq 0)$ est appelée processus de Galton-Watson. Si $Z(x)$ est un processus de Galton-Watson avec un nombre d'individus initial $Z_0 = x$, on a la propriété de branchement qui découle directement de la définition du processus :

$$Z(x+y) \overset{\mathscr{L}}{=} Z^{(1)}(x) + Z^{(2)}(y)$$

où $Z^{(1)}$ et $Z^{(2)}$ sont des copies indépendantes de Z. En terme de modélisation, cette propriété implique l'absence de compétition entre individus. Elle est donc bien adaptée lorsque l'on considère de petites populations ou des populations ayant accès à une grande quantité de ressources.

On exclura les cas peu intéressants où $p_0 = 1$ ou $p_1 = 1$. En l'absence de précision, on supposera que $Z_0 = 1$. Pour $s \in [0,1]$, la fonction génératrice f de la variable aléatoire ξ est définie par

$$f(s) = \mathbb{E}[s^\xi] = \sum_{i=0}^{\infty} p_i s^i, \quad s \in [0,1].$$

On remarque alors que pour tout $(s,n) \in [0,1] \times \mathbb{N}$,

$$\mathbb{E}[s^{Z_{n+1}}] = \mathbb{E}[\mathbb{E}[s^{Z_{n+1}} | Z_n]] = \mathbb{E}[\mathbb{E}[s^{\xi_{(1,n)}+\xi_{(2,n)}+\ldots+\xi_{(Z_n,n)}} | Z_n]] = \mathbb{E}[f(s)^{Z_n}].$$

En itérant on obtient

$$\mathbb{E}[s^{Z_n}] = f^{(n)}(s) = \underbrace{f \circ \ldots \circ f}_{n \text{ fois}}(s), \quad (s,n) \in [0,1] \times \mathbb{N}.$$

Les moments du processus, lorsqu'ils existent, peuvent s'exprimer à l'aide des dérivées de $f(s)$:

$$\mathbb{E}(Z_1) = f'(1) \quad \text{et} \quad \mathbb{E}(Z_n) = (f^{(n)})'(1) = (f'(1))^n,$$

ce qui conduit à la définition suivante qui sera justifiée par la suite :

Définition 1. *Si* $\log f'(1)$ *est strictement inférieur à* 0*, égal à* 0*, ou strictement supérieur à* 0*, le processus est appelé sous-critique, critique ou surcritique respectivement.*

Cette classification revient à distinguer les processus qui en moyenne décroissent, restent stables ou croissent.

On se restreint au cas $p_0 + p_1 < 1$. f est alors strictement convexe et croissante sur $[0,1]$. Soit q la plus petite racine de f sur $[0,1]$. Elle vaut 1 dans les cas sous-critique et critique, et est strictement inférieure à 1 dans le cas surcritique (cf figure 1.1). C'est cette racine qui va nous indiquer le comportement en temps long du processus :

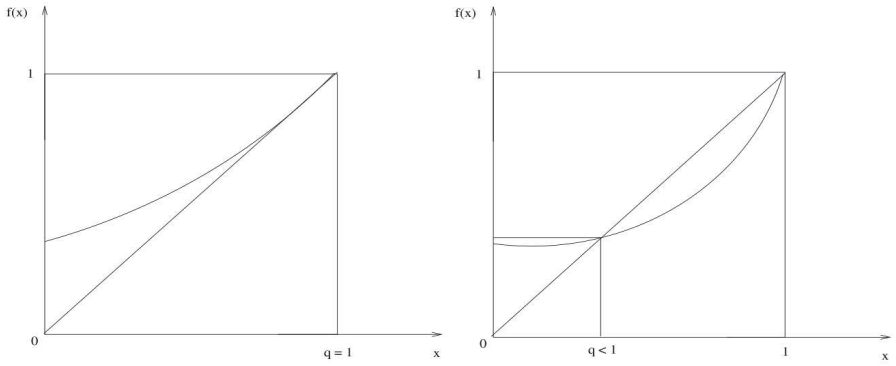

Figure 1.1: Position de la plus petite racine de f sur $[0,1]$: quand $m \leq 1$, pas de racine sur $[0,1[$, et quand $m > 1$, une unique racine sur $[0,1[$

Proposition 1. *Les processus sous-critique et critique s'éteignent presque sûrement. Le processus surcritique s'éteint avec probabilité* $q < 1$*, où* q *est la plus petite racine de* $x \mapsto f(x) - x$ *sur* $[0,1]$.

Connaissant la probabilité d'extinction des différents processus, une question naturelle est de s'intéresser aux vitesses d'extinction des processus qui s'éteignent presque sûrement, et au comportement en temps long des processus qui survivent. La chaîne de Markov :

$$W_n := Z_n (f'(1))^{-n}, \quad n \in \mathbb{N}$$

est une martingale positive, et donc converge presque sûrement vers une variable aléatoire positive W. La loi de W peut alors nous donner des informations sur le comportement en temps long d'un processus surcritique lorsqu'il ne s'éteint pas.

Théorème 1. • *Processus sous-critique : Si* $\log f'(1) < 0$ *et* $\mathbb{E}[Z_1 \log^+ Z_1] < \infty$*, il existe une constante finie* c *telle que*

$$\mathbb{P}(Z_n > 0) \sim c(f'(1))^n, \quad (n \to \infty).$$

- *Processus critique : Si* $\log f'(1) = 0$,

$$\mathbb{P}(Z_n > 0) \sim \frac{2}{nVar(Z_1)}, \quad (n \to \infty).$$

- *Processus surcritique : Si* $\log f'(1) > 0$ *et* $\mathbb{E}[Z_1 \log Z_1] < \infty$, *alors* $\mathbb{E}[W] = 1$, *et*

$$\mathbb{P}(W = 0) = \mathbb{P}(\exists n \in \mathbb{N}, Z_n = 0 | Z_0 = 1) = q.$$

Processus de Galton-Watson en environnement aléatoire

Dans le processus de Galton-Watson, la loi des naissances est la même à chaque génération, ce qui signifie que la population vit dans un environnement constant. Au début des années 70 Smith et Wilkinson [SW69], puis Athreya et Karlin [AK+71b, AK71a] ont cherché à comprendre l'effet de l'environnement sur la dynamique d'une population. Ils ont alors introduit une variabilité des lois de naissance traduisant une variabilité de l'environnement au cours des générations. Si on pense à une population de plantes annuelles, cette variabilité peut traduire la variabilité de l'ensoleillement, de la pluviométrie ou du taux de pollinisation par les abeilles d'une année à l'autre, variables qui ont une grande influence sur la quantité de descendants produits par les plantes. Ces processus ont par la suite été étudiés par de nombreux auteurs. Nous renvoyons à [BGK05, DVS11] pour les résultats connus sur les processus critique et sous-critique (cf Définition 2) et pour une bibliographie exhaustive à ce sujet. Nous renvoyons également à [BB11, BB13, Böi14] pour des résultats récents concernant le cas surcritique. Nous nous limiterons dans cette présentation au cas monotype et aux environnements indépendants et identiquement distribués.

Un processus de Galton-Watson $(Z_n, S_n, n \in \mathbb{N})$ en environnement aléatoire (GWEA) peut être décrit par la suite de ses fonctions génératrices. $\mathscr{E}_n$ est l'état de l'environnement au temps $n-1$. A cet état est associée une suite de réels $(p_k^{(n)}, k \in \mathbb{N})$ où $p_k^{(n)}$ est la probabilité qu'un individu de la génération $n-1$ produise k descendants dans la génération n. On obtient ainsi la suite de fonctions génératrices suivante :

$$f_n(s) := \mathbb{E}[s^{Z_n} | \mathscr{E}_n, Z_{n-1} = 1] = \sum_{k=0}^{\infty} p_k^{(n)} s^k, \quad (s, n) \in [0,1] \times \mathbb{N}.$$

La fonction génératrice de la chaîne de Markov inhomogène $(Z_n, n \in \mathbb{N})$ correspondant à la suite des tailles de population dans l'environnement $(\mathscr{E}_1, ..., \mathscr{E}_n, ...)$ est alors obtenue en itérant les fonctions génératrices successives :

$$\mathbb{E}[s^{Z_n} | \mathscr{E}_1, ..., \mathscr{E}_n] = f_1 \circ f_2 \circ ... \circ f_n(s), \quad (s, n) \in [0,1] \times \mathbb{N}.$$

En conséquence, la moyenne de la taille de population Z_n dans l'environnement $(\mathscr{E}_1, ..., \mathscr{E}_n)$ est :

$$\mathbb{E}[Z_n | \mathscr{E}_1, ..., \mathscr{E}_n] = f_1'(1) ... f_n'(1) = e^{S_n}, \quad n \in \mathbb{N}, \tag{1}$$

où la marche aléatoire S_n est définie par :

$$S_0 = 0, \quad S_n = \log f_1'(1) + ... + \log f_n'(1), \quad n \in \mathbb{N}. \tag{2}$$

L'exemple le plus simple de GWEA est le cas linéaire fractionnaire, caractérisé par les fonctions génératrices

$$f_n(s) = r_n + (1-r_n)\frac{t_n s}{1-(1-t_n)s},$$

où les paramètres (r_n, t_n) décrivant l'état de l'environnement à la génération $n-1$ sont dans $[0,1)\times(0,1]$ pour tout entier n. On peut alors exprimer la probabilité de survie jusqu'au temps n de la chaîne de Markov $(Z_n, n\in\mathbb{N})$:

$$\mathbb{P}(Z_n>0|\mathscr{E}_1,...,\mathscr{E}_n) = \left(e^{-S_n} + \sum_{k=1}^{n}\frac{1-t_k}{1-r_k}e^{-S_{k-1}}\right)^{-1}. \tag{3}$$

L'égalité (3) implique que sous certaines conditions de moments, la probabilité de survie du processus, $\mathbb{P}(Z_n>0|\mathscr{E}_1,...,\mathscr{E}_n)$, est gouvernée par le minimum de la marche aléatoire $(S_n, n\in\mathbb{N})$ définie dans (2). En effet le membre de droite de (3) est de l'ordre de $\exp(\min(S_0, S_1, ...S_n))$.

De manière générale on verra que la marche aléatoire $(S_n, n\in\mathbb{N})$ est intimement liée au comportement en temps long du processus $(Z_n, n\in\mathbb{N})$. Mais avant de préciser davantage cette relation, remarquons que :

$$\begin{aligned}\mathbb{P}(Z_n>0) = \mathbb{P}(\mathbb{P}(Z_n>0|\mathscr{E}_1,...,\mathscr{E}_n)) &= \mathbb{P}\Big(\min_{0\le i\le n}\mathbb{P}(Z_i>0|\mathscr{E}_1,...,\mathscr{E}_n)\Big)\\ &\le \mathbb{P}\Big(\min_{0\le i\le n}\mathbb{E}(Z_i|\mathscr{E}_1,...,\mathscr{E}_n)\Big)\\ &= \mathbb{E}\Big[\exp(\min(S_0,S_1,...S_n))\Big],\end{aligned} \tag{4}$$

où on a appliqué l'inégalité de Markov, l'égalité (1) et la définition (2). Mais si la marche aléatoire $(S_n, n\in\mathbb{N})$ n'est pas dégénérée (c'est-à-dire qu'on exclut le cas $S\equiv 0$), elle a seulement trois comportements possibles à l'infini [Fel71] : $\lim_{n\to\infty} S_n = +\infty$, $\lim_{n\to\infty} S_n = -\infty$, et $\limsup_{n\to\infty} S_n = -\liminf_{n\to\infty} S_n = +\infty$, qui correspondent respectivement aux cas $\mathbb{E}[\log f_1'(1)] < 0$, >0, et $=0$ lorsque $\mathbb{E}[|\log f_1'(1)|] < \infty$. Dans la suite de cette présentation on se restreindra au cas où $\log f_1'(1)$ est intégrable. Cette condition peut être relaxée dans la plupart des résultats mentionnés, mais cela permet d'en alléger la présentation. On peut définir une classification des GWEA analogue à la Définition 1 :

Définition 2. *Si $\mathbb{E}[\log f_1'(1)]$ est strictement inférieur à 0, égal à 0, ou strictement supérieur à 0, le processus de Galton-Watson en environnement aléatoire est appelé sous-critique, critique ou surcritique.*

L'inégalité (4) suffit à montrer que le processus $(Z_n, n\in\mathbb{N})$ s'éteint presque sûrement dans les cas sous-critique et critique. On a en fait l'équivalent du Théorème 1 :

Proposition 2 (Théorème 3.1 [SW69]). *Les processus sous-critique et critique s'éteignent presque sûrement. Le processus surcritique survit avec une probabilité strictement positive si $\mathbb{E}[|\log(1-f_1(0))|] < \infty$*

La condition $\mathbb{E}[|\log(1-f_1(0))|] < \infty$ exclut la possibilité qu'une suite d'environnements "catastrophiques" conduise à la mort de la population en quelques générations. Enfin, on a l'équivalent suivant du Théorème 1 concernant la vitesse d'extinction du cas critique et le comportement en temps long du processus surcritique dans le cas de la survie :

Théorème 2 ([AK71a, K$^+$74, Koz76, GK00]). • *Processus critique : Si* $\mathbb{E}[\log f_1'(1)] = 0$, $0 < \mathrm{Var}[\log f_1'(1)] < \infty$ *et* $f_1'(1)$ *a une distribution non lattice*[1]*, il existe des constantes positives finies* c_1 *et* c_2 *telles que*

$$\mathbb{P}(Z_n > 0) \sim c_1 \mathbb{P}(\min(S_0, ..., S_n) \geq 0) \sim \frac{c_2}{\sqrt{n}}, \quad (n \to \infty).$$

• *Processus surcritique : Si* $\mathbb{E}[Z_1 \log Z_1 / f_1'(1)] < \infty$ *et* $\mathbb{E}[\log f_1'(1)] > 0$*, alors la martingale* $(Z_n \exp(-S_n), n \in \mathbb{N})$ *a une limite finie non nulle sur l'événement de non-extinction du processus :*

$$\lim_{n\to\infty} Z_n e^{-S_n} = W \quad a.s., \quad \mathbb{P}(W > 0) = \mathbb{P}(\forall n \in \mathbb{N}, Z_n > 0 | Z_0 = 1) > 0.$$

Le cas sous-critique est plus complexe dans le cas des GWEA que dans le cas des processus de Galton-Watson classiques. En particulier, savoir que $\mathbb{E}[\log f_1'(1)] < 0$ n'est pas suffisant pour connaître le comportement du processus en temps long. La vitesse d'extinction dans le cas sous-critique a été étudiée, dans des cas de plus en plus généraux par Dekking [Dek87], D'Souza et Hambly [DH97], Guivarch et Liu [GL01], et enfin par Geiger, Kersting et Vatutin [GKV03]. L'heuristique que l'on va présenter pour bien comprendre l'origine des différents sous-cas est inspirée de [BGK05]. On rappelle la définition (2) de la marche aléatoire $(S_n, n \in \mathbb{N})$ et on suppose qu'il existe $\varepsilon > 0$ tel que $\mathbb{E}[e^{(1+\varepsilon)S_1}] < \infty$. On introduit, pour $\beta \in [0, 1]$, la probabilité $\mathbb{P}^{(\beta)}$:

$$\begin{aligned} \mathbb{E}^{(\beta)}[\phi(\mathscr{E}_1, ..., \mathscr{E}_n, Z_1, ..., Z_n)] &:= \frac{\mathbb{E}\left[\phi(\mathscr{E}_1, ..., \mathscr{E}_n, Z_1, ..., Z_n) e^{\beta S_n}\right]}{\mathbb{E}\left[e^{\beta S_n}\right]} \\ &= \left(\mathbb{E}\left[e^{\beta S_1}\right]\right)^{-n} \mathbb{E}\left[\phi(\mathscr{E}_1, ..., \mathscr{E}_n, Z_1, ..., Z_n) e^{\beta S_n}\right], \end{aligned}$$

pour ϕ fonction positive. La probabilité de survie au temps n peut donc s'exprimer en fonction de cette nouvelle probabilité : pour tout $\beta \in \mathbb{R}^+$ tel que $\mathbb{E}[e^{\beta S_1}] < \infty$,

$$\mathbb{P}(Z_n > 0) = \left(\mathbb{E}\left[e^{\beta S_1}\right]\right)^n \mathbb{E}^{(\beta)}\left[\mathbb{P}(Z_n > 0 | \mathscr{E}_1, ..., \mathscr{E}_n) e^{-\beta S_n}\right].$$

On poursuit maintenant l'heuristique en admettant que $\mathbb{P}(Z_n > 0 | \mathscr{E}_1, ..., \mathscr{E}_n)$ est de l'ordre de $e^{\min(S_0, S_1, ..., S_n)}$, et on s'intéresse au comportement limite de l'espérance :

$$\mathbb{E}\left[e^{\min(S_0, S_1, ..., S_n)}\right] = \left(\mathbb{E}\left[e^{\beta S_1}\right]\right)^n \mathbb{E}^{(\beta)}\left[e^{\min(S_0, S_1, ..., S_n) - \beta S_n}\right].$$

On va alors chercher un réel β tel que le dernier terme n'ait pas une croissance ou une décroissance exponentielle. Le terme $\mathbb{E}[\exp(\beta S_1)]$ nous donnera alors la décroissance exponentielle de la probabilité de survie, et le terme $\mathbb{E}^{(\beta)}[e^{\min(S_0, S_1, ..., S_n) - \beta S_n}]$ sa décroissance polynomiale. On peut déjà exclure les $\beta > 1$. En effet dans le cas sous-critique la marche aléatoire $(S_n, n \in \mathbb{N})$ dérive vers moins l'infini et le terme $\min(S_0, S_1, ..., S_n) - \beta S_n$ dérive vers l'infini pour tout

[1] On dit qu'une variable aléatoire X a une distribution lattice s'il existe $(a, b) \in \mathbb{R}$ tel que $\mathbb{P}(X \in a + b\mathbb{Z}) = 1$.

$\beta > 1$. Soit $\beta \in [0,1]$. On va voir qu'on peut distinguer trois cas suivant le comportement de la fonction $\beta \mapsto \mathbb{E}[\exp(\beta S_1)]$ sur l'intervalle $[0,1]$ (cf figure 1.2). Cette fonction est bien définie sous nos hypothèses et admet même une dérivée et une dérivée seconde sur $[0,1]$ qui ont pour expressions :

$$\partial_\beta \mathbb{E}[\exp(\beta S_1)] = \mathbb{E}[S_1 \exp(\beta S_1)], \quad \partial_{\beta\beta} \mathbb{E}[\exp(\beta S_1)] = \mathbb{E}[S_1^2 \exp(\beta S_1)] > 0.$$

C'est donc une fonction convexe qui atteint son minimum en $\tau \in (0,1)$ si $\mathbb{E}[S_1 \exp(S_1)] > 0$ et en 1 si $\mathbb{E}[S_1 \exp(S_1)] \leq 0$. On va donc considérer successivement les différents signes que peut prendre $\mathbb{E}[S_1 \exp(S_1)]$.

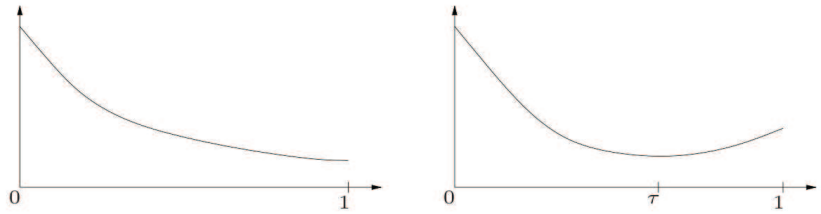

Figure 1.2: Comportement de $\beta \mapsto \mathbb{E}[\exp(\beta S_1)]$: lorsque $\mathbb{E}(S_1 \exp(S_1)) \leq 0$ le minimum est atteint en 1, et lorsque $\mathbb{E}(S_1 \exp(S_1)) > 0$ le minimum est atteint en $\tau \in (0,1)$

- $\mathbb{E}[S_1 \exp(S_1)] > 0$: Alors $\beta \mapsto \mathbb{E}[\exp(\beta S_1)]$ atteint son minimum sur $[0,1]$ en $\tau \in (0,1)$, et

  $$0 = \mathbb{E}[S_1 e^{\tau S_1}] = \mathbb{E}[e^{\tau S_1}] \mathbb{E}^{(\tau)}[S_1].$$

  La marche aléatoire $(S_n, n \in \mathbb{N})$ est de moyenne nulle et de variance finie sous $\mathbb{P}^{(\tau)}$. La seule possibilité pour que $\min(S_0, S_1, ..., S_n) - \tau S_n$ soit proche de 0 est que $\min(S_0, S_1, ..., S_n)$ et τS_n soient proches de 0. L'événement $\{S_0, ..., S_{n-1} \geq 0, S_n \leq 0\}$ a une probabilité de l'ordre de $n^{-3/2}$ sous $\mathbb{P}^{(\tau)}$ (voir [Épp79] Equation C du Théorème pour le cas non lattice, et [AD99] pour le cas lattice). On s'attend donc dans ce cas à

  $$\mathbb{E}^{(\tau)}\left[e^{\min(S_0,S_1,...,S_n)-\tau S_n}\right] \asymp n^{-3/2}.$$

- $\mathbb{E}[S_1 \exp(S_1)] = 0$: Alors de la même manière, $\mathbb{E}^{(1)}[S_1] = 0$ et $\min(S_0, S_1, ..., S_n) - S_n$ est proche de 0 si $(S_n, n \in \mathbb{N})$ atteint son minimum sur $\{0, ..., n\}$ près de n. Comme la probabilité de l'événement $\{S_0, ..., S_{n-1} \geq S_n\}$ est de l'ordre de $n^{-1/2}$ sous $\mathbb{P}^{(1)}$ (Equation (7.12) Chapitre XII de [Fel71]) , on s'attend à avoir :

  $$\mathbb{E}^{(1)}\left[e^{\min(S_0,S_1,...,S_n)-S_n}\right] \asymp n^{-1/2}.$$

- $\mathbb{E}[S_1 \exp(S_1)] > 0$: Alors $\mathbb{E}^{(1)}[S_1] < 0$ et $\min(S_0, S_1, ..., S_n) - S_n$ reste borné puisque $(S_n, n \in \mathbb{N})$ dérive vers moins l'infini sous la probabilité $\mathbb{P}^{(1)}$. On aura donc :

  $$\mathbb{E}^{(1)}\left[e^{\min(S_0,S_1,...,S_n)-S_n}\right] \asymp const.$$

Le traitement précis de cette heuristique mène aux asymptotiques suivantes pour la probabilité de survie dans le cas sous-critique :

Théorème 3 ([GKV03]). *On considère le cas sous-critique,* $\mathbb{E}[\log f_1'(1)] < 0$*, et on introduit*

$$\tau : \underset{\beta\in(0,1)}{argmin}\ \mathbb{E}[f_1'(1)^\beta].$$

Sous des hypothèses additionnelles de moments on a les asymptotiques suivantes, où c *désigne une constante finie dont la valeur peut changer d'une ligne à l'autre :*

- *Cas sous-critique fort : Si* $\mathbb{E}[f_1'(1)\log f_1'(1)] < 0$*, alors* $\tau = 1$ *et*

$$\mathbb{P}(Z_n > 0) \sim c(\mathbb{E}[f_1'(1)])^n, \quad (n \to \infty).$$

- *Cas sous-critique intermédiaire : Si* $\mathbb{E}[f_1'(1)\log f_1'(1)] = 0$*, alors* $\tau = 1$ *et*

$$\mathbb{P}(Z_n > 0) \sim c(\mathbb{E}[f_1'(1)])^n n^{-1/2}, \quad (n \to \infty).$$

- *Cas sous-critique faible : Si* $0 < \mathbb{E}[f_1'(1)\log f_1'(1)] < \infty$*, alors* $\tau \in (0,1)$ *et*

$$\mathbb{P}(Z_n > 0) \sim c(\mathbb{E}[f_1'(1)^\tau])^n n^{-3/2}, \quad (n \to \infty).$$

Ainsi le cas sous-critique fort, dans lequel l'environnement est peu variable est similaire au cas sous-critique du processus de Galton-Watson. Pour un taux de croissance moyen $\mathbb{E}[Z_1|Z_0 = 1] = \mathbb{E}[\log f_1'(1)] < 0$, la vitesse d'extinction exponentielle ne dépend de la loi de l'environnement que si ce dernier est suffisamment variable, plus précisément à partir du moment où $\mathbb{E}[f_1'(1)\log f_1'(1)] > 0$. Pour voir que la variabilité de l'environnement est liée à cette inégalité on peut remarquer que pour que cette inégalité soit vérifiée il faut que $f'(1)$ puisse prendre de grandes valeurs. De plus, puisque $\mathbb{E}[\log f_1'(1)]$ est négatif dans le cas sous-critique, $f'(1)$ doit également pouvoir prendre de petites valeurs.

Processus de branchement à états continus en environnement aléatoire

Processus de branchement à états continus

Les processus de branchement à états continus (CSBP pour Continuous State Branching Processes) sont l'analogue en temps et espace continus des processus de Galton-Watson. Ils ont été introduits par Jirina [Jir58] et étudiés par de nombreux auteurs. On mentionnera en particulier [Lam67a, Lam67b, Gre74, Gri74, Bin76] pour les premiers travaux sur le sujet, et [Kyp06, Li10] pour des reviews récentes. Un CSBP $Z = (Z_t, t \in \mathbb{R}^+)$ est un processus de Markov fort, càdlàg, à valeurs dans $\mathbb{R}^+$ et qui satisfait la propriété de branchement. Z admet 0 et ∞ comme points absorbants, et $\mathbb{P}_x$ désigne la loi du processus partant de x. Lamperti [Lam67b] a montré que les CSBP sont les seules limites d'échelles possibles des processus de Galton-Watson et que, réciproquement, tout CSBP peut être obtenu comme une telle limite d'échelle.

Ici encore dans un souci de simplicité nous supposerons que $(Z_t, t \in \mathbb{R}^+)$ admet un premier moment. La propriété de branchement du processus implique que son exposant de Laplace a la forme :

$$\mathbb{E}_x[e^{-\theta Z_t}] = e^{-x u_t(\theta)}, \quad (x,t,\theta) \in \mathbb{R}_+^3, \tag{5}$$

où ([Sil68]) la fonction $(u_t(\theta), (t,\theta) \in \mathbb{R}_+^2)$ est l'unique solution de l'équation différentielle :

$$\frac{\partial}{\partial t} u_t(\theta) + \psi(u_t(\theta)) = 0, \quad u_0(\theta) = \theta, \tag{6}$$

où pour $\lambda \geq 0$,

$$\psi(\lambda) = \sigma^2\lambda^2 - g\lambda + \int_{(0,\infty)} (e^{-\lambda z} - 1 + \lambda z)\mu(dz),$$

pour $\sigma, g \in \mathbb{R}$, et μ une mesure à support dans $(0,\infty)$ qui vérifie $\int_{(0,\infty)}(z \wedge z^2)\mu(dz)$.

Les processus de branchement continus sont spectralement positifs, les sauts représentant des événements de naissance macroscopiques (un individu infinitésimal donne naissance à un nombre suffisamment grand d'individus infinitésimaux pour que cela soit visible macroscopiquement). Cette propriété n'est pas directement apparente dans la définition du processus, elle le devient dans l'énoncé de la transformée de Lamperti. Ce dernier a montré [Lam67a] qu'un CSBP pouvait être réalisé comme un processus de Lévy spectralement positif changé de temps.

Un CSBP peut également être défini via une équation différentielle stochastique, représentation qui nous servira par la suite :

Théorème 4 ([FL10]). *Le processus $(Z_t, t \in \mathbb{R}^+)$ est l'unique solution forte de l'équation*

$$Z_t = Z_0 + \int_0^t g Z_s ds + \int_0^t \sqrt{2\sigma^2 Z_s} dB_s + \int_0^t \int_0^\infty \int_0^{Z_{s-}} z \widetilde{N}_0(ds, dz, du), \tag{7}$$

où B est un mouvement Brownien standard, $N_0(ds,dz,du)$ est une mesure aléatoire de Poisson d'intensité $ds\mu(dz)du$ indépendante de B, μ vérifie $\int_{(0,\infty)}(z \wedge z^2)\mu(dz)$, et $\widetilde{N}_0$ est la mesure compensée de N_0.

Remarque 1. *Nous nous sommes placés dans le cas où Z admet un premier moment. Cela entraîne en particulier $\int_{(0,\infty)}(z \wedge z^2)\mu(dz)$, ce qui permet de compenser également les grands sauts et de faire sortir le terme de drift g qui gouverne le comportement en temps long du processus. Dans le cas général on a seulement $\int_{(0,\infty)}(1 \wedge z^2)\mu(dz)$ et on ne compense que les saut de taille inférieure à 1.*

Le mécanisme de branchement ψ détermine de manière unique la loi du CSBP Z. C'est une fonction convexe nulle en 0. Sa forme dépend du signe de $\psi'(0)$ (cf figure 1.3). La dérivation de (5) en $\theta > 0$ donne :

$$\mathbb{E}_x[Z_t e^{-\theta Z_t}] = x \frac{\partial}{\partial \theta} u_t(\theta) e^{-x u_t(\theta)},$$

et en passant à la limite lorsque θ tend vers 0 on obtient :

$$\mathbb{E}_x[Z_t] = x \frac{\partial}{\partial \theta} u_t(0).$$

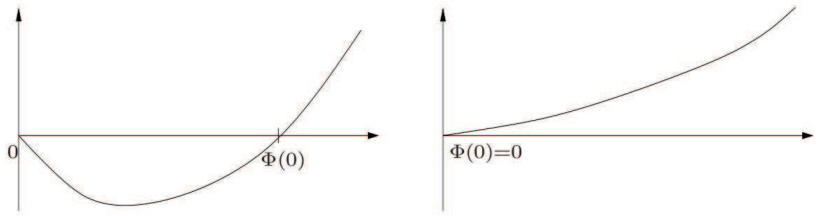

Figure 1.3: Forme de ψ : lorsque $g > 0$, $\psi(\eta) = 0$ pour un $\eta > 0$, qu'on note $\Phi(0)$, et lorsque $g \leq 0$, $\psi(\eta) = 0$ seulement en $\eta = 0$

D'autre part la dérivation en θ de (6) pour $(t,\theta) \in (\mathbb{R}_+^*)^2$ conduit à l'égalité :

$$\frac{\partial}{\partial t}\frac{\partial}{\partial \theta}u_t(\theta) + \psi'(u_t(\theta))\frac{\partial}{\partial \theta}u_t(\theta) = 0.$$

On en déduit qu'il existe une contante finie c telle que :

$$\frac{\partial}{\partial \theta}u_t(\theta) = ce^{-\int_0^t \psi'(u_s(\theta))ds}.$$

En prenant la limite lorsque t tend vers 0, on voit que $c = 1$, et la limite quand θ tend vers 0 nous permet de conclure :

$$\mathbb{E}_x[Z_t] = xe^{gt}.$$

On a alors l'analogue des Définitions 1 et 2 :

Définition 3. *Si g est strictement inférieur à 0, égal à 0, ou strictement supérieur à 0, le CSBP Z est appelé sous-critique, critique ou surcritique respectivement.*

L'événement d'extinction de la population est défini par :

$$\text{Ext} := \{\lim_{t\to\infty} Z_t = 0\}, \tag{8}$$

ce qui conduit à l'analogue des Propositions 1 et 2 :

Proposition 3. *Les processus sous-critique et critique s'éteignent presque sûrement, le processus surcritique survit avec probabilité positive. Plus précisément,*

$$\mathbb{P}_x(Ext) = e^{-x\Phi(0)}, \quad \forall x > 0,$$

où $\Phi(0)$ est la racine strictement positive de ψ.

Mais dans le cas des CSBP il faut faire la distinction entre deux événements, l'extinction et l'absorption du processus. Le premier a été défini en (8), le second correspond au cas où le processus touche 0 en temps fini :

$$\text{Abs} := \{\exists t < \infty, Z_t = 0\}. \tag{9}$$

Ces deux événements ne coincident pas nécessairement. Par exemple le CSBP $Z = (Z_0 e^{-t}, t \geq 0)$ vérifie pour tout $x \in \mathbb{R}_+^*$, $\mathbb{P}_x(\text{Ext}) = 1$ et $\mathbb{P}_x(\text{Abs}) = 0$. On peut donc avoir extinction et pas absorption. En revanche, si la probabilité que le processus soit absorbé est positive, les événements d'absorption et d'extinction sont presque sûrement confondus :

$$\begin{aligned} \mathbb{P}_x(\text{Abs}) > 0, \quad \forall x > 0 &\iff \text{Abs} = \text{Ext} \quad \text{p.s.} \\ &\iff \psi(\infty) = \infty \quad \text{et} \quad \int^{\infty} \frac{ds}{\psi(s)} < \infty. \end{aligned}$$

Les CSBP sont donc des objets plus complexes que les processus de Galton-Watson. En particulier, la détermination de leur vitesse d'extinction pose problème puisqu'ils peuvent tendre vers 0 sans s'annuler. Une classe de CSBP cependant est bien connue. Il s'agit des CSBP α-stables, pour α dans $(1,2]$. Leur mesure de sauts μ (voir Théorème 4) vérifie :

$$\mu(dx) = \frac{c\alpha}{\Gamma(2-\alpha)x^{\alpha+1}}, \quad x > 0.$$

L'équation différentielle (6) se résout alors explicitement et on obtient une expression de la vitesse d'extinction (voir par exemple Proposition 4 de [BPS13]) :

$$\mathbb{P}_{x_0}(Z_t > 0) = 1 - \exp\Big(- x_0 \Big(c(\alpha-1) \int_0^t e^{-(\alpha-1)gs} ds \Big)^{-1/(\alpha-1)} \Big).$$

Cela permet de dériver l'analogue suivant du Théorème 1 (on peut trouver le résultat pour le cas critique dans [KP08] p. 12) :

Théorème 5. *Soit $Z = (Z_t, t \geq 0)$ un CSBP α-stable avec α dans $(1,2]$:*

- *Si $g = 0$ (cas critique)*

$$\mathbb{P}_{x_0}(Z_t > 0) \sim \frac{x_0}{(c(\alpha-1)t)^{1/(\alpha-1)}}, \quad (t \to \infty).$$

- *Si $g < 0$ (cas sous-critique)*

$$\mathbb{P}_{x_0}(Z_t > 0) \sim x_0 \Big(\frac{|g|}{c} \Big)^{1/(\alpha-1)} e^{gt} = x_0 \Big(\frac{|g|}{c} \Big)^{1/(\alpha-1)} (\mathbb{E}_1[Z_1])^t, \quad (t \to \infty).$$

En particulier dans le cas $\alpha = 2$ (diffusion de Feller) on retrouve une vitesse d'extinction analogue à celle des processus de Galton-Watson dans le cas critique :

$$\mathbb{P}_{x_0}(Z_t > 0) \sim \frac{x_0}{\sigma^2 t}, \quad (t \to \infty).$$

Cela est une conséquence du fait que les diffusions de Feller sont des limites d'échelles de suites de processus de Galton-Watson dont la loi de reproduction a une variance finie. Si :

$$Z_t^{(n)} := \frac{1}{n} Z_{\lfloor nt \rfloor}^{(n)}, \quad t \geq 0$$

où $(Z^{(n)}, n \in \mathbb{N})$ est une suite de processus de GW de loi de naissance ξ et vérifiant $Z_0^{(n)} = n$ et $\mathbb{E}[(Z_1^{(n)})^2] < \infty$ pour tout $n \in \mathbb{N}$, alors $(Z_t^{(n)}, t \in [0,T])$ converge en loi lorsque n tend vers l'infini (ici T est un réel positif) vers une diffusion de Feller sur l'intervalle $[0,T]$.

Diffusions branchantes en environnement brownien

La construction des CSBP comme limites d'échelle de processus de Galton-Watson soulève alors une question naturelle. Peut-on, à l'instar des processus de Galton-Watson, plonger un processus de branchement continu dans un environnement aléatoire, pour prendre en compte l'évolution des conditions de vie et de reproduction des individus (infinitésimaux) constituant les populations modélisées ? Mais ici la question est plus délicate. En effet, faire varier l'environnement dans un cadre discret revenait à considérer de nouvelles lois de naissances à chaque pas de temps discret. Ici les variations d'environnement peuvent prendre des formes multiples. L'environnement peut varier continuement, brutalement, plus ou moins fréquemment et avec une amplitude variable. A notre connaissance, deux cas seulement ont été étudiés jusqu'à présent : des diffusions branchantes en environnement brownien et des CSBP soumis à des catastrophes qui surviennent à des temps poissonniens et multiplient la taille de la population par un réel positif. Le premier cas a fait l'objet d'un article de Böinghoff et Hutzenthaler [BH12] (voir aussi [Hut11] pour le cas surcritique) dont je vais maintenant présenter les principaux résultats, et le second cas est étudié dans le Chapitre 2 de cette thèse.

L'existence des diffusions branchantes en environnement brownien (DBEB), construites comme des limites de processus de Galton-Watson en environnement aléatoire a été conjecturée par Keiding [Kei75] et rigoureusement établie par Kurtz [Kur78]. On considère une suite $((Z^{(p)}, S^{(p)}), p \in \mathbb{N})$ de processus de Galton-Watson en environnement aléatoire. $\mathscr{E}^{(p)} := (\mathscr{Q}_1^{(p)}, \mathscr{Q}_2^{(p)}, ...)$ est une suite de probabilités sur $\mathbb{N}$ représentant les environnements successifs. Conditionnellement à la suite d'environnements $\mathscr{E}^{(p)}$ le processus de Galton-Watson en environnement aléatoire $(Z_n^{(p)}, n \in \mathbb{N})$ est défini récursivement par :

$$Z_{n+1}^{(p)} = \sum_{i=1}^{Z_n^{(p)}} \xi_{(i,n)}^{(p)}, \quad n \in \mathbb{N},$$

où $Z_0^{(p)}$ est indépendant de $\mathscr{E}^{(p)}$, conditionnellement à $\mathscr{E}^{(p)}$ les $(\xi_{(i,n)}^{(p)})_{(i,n)\in\mathbb{N}^2}$ sont des variables aléatoires indépendantes et les $(\xi_{(i,n)}^{(p)})_{i\in\mathbb{N}}$ sont des variables aléatoires identiquement distribuées avec pour fonction génératrice :

$$f_n^{(p)}(s) = \mathbb{E}(s^{\xi_{(1,n)}^{(p)}} | \mathscr{Q}_n^{(p)}) = \sum_{k=0}^{\infty} \mathscr{Q}_n^{(p)}(k) s^k, \quad s \in [0,1].$$

En conséquence, la moyenne de la taille de population $Z_n^{(p)}$ dans l'environnement $\mathscr{E}^{(p)}$ vérifie :

$$\mathbb{E}[Z_n^{(p)} | \mathscr{E}^{(p)}] = f_1^{(p)'}(1)...f_n^{(p)'}(1) = e^{S_n^{(p)}}, \quad (n,p) \in \mathbb{N}^2, \tag{10}$$

où la marche aléatoire $(S_n^{(p)}, n \in \mathbb{N})$ est définie par :

$$S_0^{(p)} = 0, \quad S_n^{(p)} = \log f_1^{(p)'}(1) + ... + \log f_n^{(p)'}(1), \quad (n,p) \in \mathbb{N}^2.$$

Les hypothèses suivantes assurent la convergence de la marche aléatoire $(S_n^{(p)}, n \in \mathbb{N})$ changée de temps :

$$\lim_{p\to\infty} p.\mathbb{E}[f_1^{(p)'}(1) - 1] = \alpha \in \mathbb{R}, \tag{11}$$

$$\lim_{p\to\infty} p.\mathbb{E}[(f_1^{(p)'}(1)-1)^2] = \sigma_e^2 \in [0,\infty), \tag{12}$$

$$\sup_{p\in\mathbb{N}} \mathbb{E}\Big[\sum_{k=0}^{\infty}\Big|\frac{k}{f_1^{(p)'}(1)}-1\Big|^3 \mathcal{Q}_1^{(p)}(k)\Big] < \infty. \tag{13}$$

Le processus de branchement est donc proche d'un processus critique, et σ_e^2 rend compte de la stochasticité environnementale. On fait également une hypothèse de convergence de la stochasticité démographique,

$$\lim_{p\to\infty} \mathbb{E}\Big[\sum_{k=0}^{\infty}\Big(\frac{k}{f_1^{(p)'}(1)}-1\Big)^2 \mathcal{Q}_1^{(p)}(k)\Big] = \sigma_b^2 \in [0,\infty). \tag{14}$$

Kurtz démontre le résultat suivant :

Théorème 6 ([Kur78]). *On suppose que $Z_0^{(p)}/p \to Z_0$ en loi lorsque $p\to\infty$. Sous les hypothèses (11) à (14),*

$$\Big(\frac{Z^{(p)}_{\lfloor pt\rfloor}}{p}, S^{(p)}_{\lfloor pt\rfloor}\Big)_{t\geq 0} \underset{p\to\infty}{\to} (Z_t, S_t)_{t\geq 0}$$

dans la topologie de Skorohod [2] *où la diffusion limite $(Z_t, S_t, t\geq 0)$ est la solution forte de*

$$\begin{cases} dZ_t = \frac{1}{2}\sigma_e^2 Z_t dt + Z_t dS_t + \sqrt{\sigma_b^2 Z_t}\, dW_t^{(b)} \\ dS_t = \alpha dt + \sqrt{\sigma_e^2}\, dW_t^{(e)}. \end{cases}, \tag{15}$$

avec $W^{(b)}$ et $W^{(e)}$ deux mouvements browniens indépendants.

En particulier, en combinant le dernier couple d'équations, on obtient :

$$dZ_t = (\alpha + \frac{1}{2}\sigma_e^2) Z_t dt + Z_t\sqrt{\sigma_e^2}\, dW_t^{(e)} + \sqrt{\sigma_b^2 Z_t}\, dW_t^{(b)}.$$

Cette représentation met en lumière la perte de la propriété de branchement lorsqu'on ne conditionne pas à l'environnement. En effet vivre dans le même environnement introduit une corrélation dans le comportement des individus. On peut également montrer à l'aide de cette repésentation l'égalité presque sure suivante :

$$\mathbb{E}[Z_t|S_t] = \mathbb{E}[Z_0]e^{S_t}, \tag{16}$$

à rapprocher de (1). En compensant par l'environnement aléatoire et en effectuant un changement de temps, on peut retrouver un processus branchant de loi connue :

[2] Soient D l'ensemble des fonctions càdlàg sur $[0,1]$,

$$\Lambda := \{\lambda : [0,1]\to[0,1], \mathscr{C}^0, \text{ strictement croissante, } \lambda(0)=0, \lambda(1)=1\}.$$

et la distance d pour $x, y \in D$:

$$d(x,y) := \inf\{\varepsilon>0, \exists\lambda\in\Lambda, \sup_{t\in[0,1]} |\lambda(t)-t| < \varepsilon, \sup_{t\in[0,1]} |x(t)-y(\lambda(t))| < \varepsilon\}.$$

Alors (D,d) est un espace métrique séparable mais pas complet.

Proposition 1 ([Kur78]). *Soient $\alpha \in \mathbb{R}$, $\sigma_b \in (0,\infty)$, $\sigma_e \in [0,\infty)$, et $(W_t^{(b)})_{t\geq 0}$ et $(W_t^{(e)})_{t\geq 0}$ deux mouvements browniens standards indépendants, $(F_t)_{t\geq 0}$ la solution forte de*

$$dF_t = \sqrt{F_t}\, dW_t^{(b)}, \quad t \geq 0,$$

et $S_t := \alpha t + \sigma_e W_t^{(e)}$, $t \geq 0$. De plus on définit le temps aléatoire

$$\tau(t) := \int_0^t e^{-S_s}\sigma_b^2 ds, \quad t \geq 0.$$

Alors

$$(F_{\tau(t)} e^{S_t}, S_t)_{t\geq 0}$$

est solution forte de (15).

En particulier, $(Z_t e^{-S_t})_{t\geq 0}$ est une diffusion de Feller changée de temps, ce qui permet d'obtenir une expression explicite de la transformée de Laplace du processus Z conditionnellement à l'environnement.

Corollaire 1 ([BH12]). *Soient $\alpha \in \mathbb{R}$, $\sigma_b \in (0,\infty)$, $\sigma_e \in [0,\infty)$, et $(Z_t, S_t)_{t\geq 0}$ la solution forte de (15). Alors*

$$\mathbb{E}_x[\exp(-\lambda Z_t)|(S_s)_{s\leq t}] = \exp\Big(-\frac{x}{\int_0^t \sigma_b^2 \exp(-S_s)ds/2 + \exp(-S_t)/\lambda}\Big),$$

pout tous $(t,x,\lambda) \in [0,\infty)^3$ presque surement. En particulier,

$$\mathbb{P}_x(Z_t > 0|(S_s)_{s\leq t}) = 1 - \exp\Big(-\frac{x}{\int_0^t \sigma_b^2 \exp(-S_s)ds/2}\Big),$$

pour tous $t \in (0,\infty)$ et $x \in [0,\infty)$ presque surement.

Ce lien avec les diffusions de Feller permet d'obtenir des formules explicites pour de nombreuses expressions. En particulier, cela permet de distinguer les contributions respectives des stochasticités démographique et environnementale dans les différentes vitesses d'extinction des DBEB. L'expression de la moyenne conditionnelle dans (16) indique que là encore le comportement du processus $(S_t)_{t\geq 0}$ va avoir une grande influence. Ce dernier tend vers $+\infty$, oscille ou tend vers $-\infty$ si $\alpha > 0$, $\alpha = 0$ ou $\alpha > 0$ respectivement, et en conséquence la DBEB sera appelée surcritique, critique ou sous-critique :

Théorème 7 ([BH12]). *Soient $\alpha \in \mathbb{R}$, $\sigma_b, \sigma_e \in (0,\infty)$, et $(Z_t, S_t)_{t\geq 0}$ la solution forte de (15). Alors :*

- *Cas surcritique : si $\alpha > 0$*

$$\mathbb{P}_x(Z_t > 0) \sim f_1(x,\alpha,\sigma_e,\sigma_b), \quad (t \to \infty).$$

- *Cas critique : si $\alpha = 0$*

$$\mathbb{P}_x(Z_t > 0) \sim \frac{1}{\sqrt{t}} f_2(x,\sigma_e,\sigma_b), \quad (t \to \infty).$$

- *Cas sous-critique :*

$$si\ \frac{\alpha}{\sigma_e^2} \in (-1,0), \quad \mathbb{P}_x(Z_t>0) \sim t^{-3/2} e^{-\frac{\alpha^2}{2\sigma_e^2}t} f_3(x,\alpha,\sigma_e,\sigma_b), \quad (t\to\infty)$$

$$si\ \frac{\alpha}{\sigma_e^2} = -1, \quad \mathbb{P}_x(Z_t>0) \sim t^{-1/2} e^{-\frac{\sigma_e^2}{2}t} x f_4(\sigma_e,\sigma_b), \quad (t\to\infty)$$

$$si\ \frac{\alpha}{\sigma_e^2} < -1, \quad \mathbb{P}_x(Z_t>0) \sim e^{(\alpha+\frac{\sigma_e^2}{2})t} x f_5(\alpha,\sigma_e,\sigma_b), \quad (t\to\infty)$$

pour tout $x\in(0,\infty)$ *où les* f_i *sont des fonctions à valeurs dans* $\mathbb{R}_+^*$.

Là encore on voit apparaître une transition de phase dans le cas sous-critique, qui dépend du rapport entre le drift et le bruit de l'environnement. Lorsque l'environnement est peu variable (σ_e petit), le comportement du processus est proche de celui des CSBP sous-critiques en environnement constant, dans le sens où $\mathbb{P}_x(Z_t>0) \sim const.\mathbb{E}_x(Z_t)$. Mais lorsque le bruit de l'environnement dépasse une certaine valeur la décroissance exponentielle de la vitesse d'extinction n'est plus égale au paramètre malthusien $\alpha+\sigma_e^2/2$ (qui peut être positif). Nous verrons dans le Chapitre 2 que nous obtenons des régimes de vitesses d'extinction similaires dans le cas des processus de branchement avec catastrophes, alors que la construction et les techniques de preuves utilisés sont très différents.

Résultats du Chapitre 2

Le Chapitre 2, résultat d'une collaboration avec V. Bansaye et J. C. Pardo a été motivé par des expériences du Laboratoire Tamara montrant une asymétrie entre les deux cellules filles issues d'une même mère chez E. Coli. Bansaye et Tran [BT11] ont alors développé un modèle d'infection de parasites dans une lignée cellulaire dans lequel les parasites étaient distribués de manière asymétrique entre les deux cellules filles : chaque cellule peut contenir des parasites, dont la dynamique suit un processus de Feller, et se divise indépendamment des autres cellules.

Nous nous sommes intéressés de manière plus générale au comportement en temps long de processus de branchement soumis de façon répétée à des catastrophes, comme l'est la quantité de parasites dans une lignée cellulaire à chaque division, et comme peut l'être également une population soumise à des catastrophes environnementales ou à l'arrivée de nouveaux prédateurs. Lorsqu'une catastrophe survient, elle tue chaque sous-population avec la même probabilité, et résulte donc en la multiplication de la taille de la population par une fraction aléatoire. Nous avons également considéré le cas des sauts positifs, qui peuvent représenter des événements d'immigration proportionnels à la taille de la population. On renvoie au Chapitre 12 de [DGC08] pour des exemples et explications de tels comportements d'agrégation, ou à [RLF+13], qui montre qu'ils peuvent résulter de manipulations par des parasites, dans le but d'accroître leur transmission.

Reprenant les notations du Théorème 4, nous introduisons une mesure aléatoire de Poisson N_1 sur $[0,\infty)^2$ d'intensité $dtv(dm)$. Cela nous permet de définir le CSBP (g,σ,μ) avec

catastrophes ν comme la solution de l'équation différentielle stochastique :

$$Y_t = Y_0 + \int_0^t g Y_s ds + \int_0^t \sqrt{2\sigma^2 Y_s} dB_s \quad + \quad \int_0^t \int_{[0,\infty)} \int_0^{Y_{s-}} z \widetilde{N}_0(ds, dz, du) + \quad \int_0^t \int_{[0,\infty)} \Big(m-1\Big) Y_{s-} N_1(ds, dm), \tag{17}$$

où $Y_0 > 0$ p.s. et N_1 est indépendante de B et N_0. Ainsi le processus Y se comporte comme un CSBP en dehors des catastrophes, et est multiplié par une variable aléatoire de loi ν à chaque nouvelle catastrophe. Si $m < 1$ la taille de la population décroît, et si $m > 1$ elle croît, ce qu'on peut interpréter comme un événement d'immigration. On introduit alors le processus de Lévy $\Delta = (\Delta_t, t \geq 0)$ représentant les catastrophes :

$$\Delta_t = \int_0^t \int_{(0,\infty)} \log(m) N_1(ds, dm) = \sum_{s \leq t} \log(m_s).$$

Le choix de cette représentation, qui est équivalent à la donnée de la mesure N_1, découle du constat suivant : en l'absence de catastrophes, le processus se comporte en moyenne comme e^{gt} et g est le paramètre Malthusien de la population. Lorsqu'une catastrophe d'intensité m survient, la taille de la population est multlipliée par un facteur $m = e^{\log m}$. En définitive, le processus va se comporter, en un certain sens qui sera précisé par la suite, comme le processus $e^{gt+\Delta_t}$ et l'introduction du processus $(\Delta_t, t \geq 0)$ permettra d'exprimer de manière explicite les conditions sous lesquelles la population peut survivre avec probabilité positive. Notre premier résultat en ce sens est le suivant :

Théorème 8. *L'équation différentielle stochastique* (17) *admet une unique solution forte positive Y pour tous $g \in \mathbb{R}, \sigma \geq 0$, μ et ν vérifiant*

$$\int_0^\infty (z \wedge z^2) \mu(dz) < \infty, \tag{18}$$

$$\nu(\{0\}) = 0 \qquad \textit{et} \qquad 0 < \int_{(0,\infty)} (1 \wedge |m-1|) \nu(dm) < \infty. \tag{19}$$

Alors le processus $Y = (Y_t, t \geq 0)$ est un processus de Markov càdlàg satisfaisant la propriété de branchement conditionnellement à Δ. De plus, pour tout $t \geq 0$,

$$\mathbb{E}_y \left[\exp\Big\{ -\lambda \exp\{-gt - \Delta_t\} Y_t \Big\} \Big| \ \Delta \right] = \exp\Big\{ -y v_t(0, \lambda, \Delta) \Big\} \qquad \textit{p.s.}, \tag{20}$$

où pour tous $\lambda \in \mathbb{R}_+$ et δ fonction à variations bornées de $\mathbb{R}_+$ dans $\mathbb{R}$, $v_t : s \in [0,t] \mapsto v_t(s, \lambda, \delta)$ est l'unique solution de l'équation différentielle rétrograde :

$$\frac{\partial}{\partial s} v_t(s, \lambda, \delta) = e^{gs + \delta_s} \psi_0\big(e^{-gs-\delta_s} v_t(s, \lambda, \delta)\big), \qquad v_t(t, \lambda, \delta) = \lambda, \tag{21}$$

et

$$\psi_0(\lambda) = \sigma^2 \lambda^2 + \int_0^\infty (e^{-\lambda z} - 1 + \lambda z) \mu(dz). \tag{22}$$

Nous restreignons donc notre étude à des CSBP qui sans les catastrophes seraient d'espérance finie (condition (18)) et à des processus de catastrophes Δ qui sont des processus de Lévy à variations bornées (condition (19)). L'existence et l'unicité des solutions de l'équation différentielle stochastique (17) n'est pas un résultat classique car les fonctions ne sont pas lipschitziennes. Elles découlent de la Proposition 2.2 et des Théorèmes 3.2 et 5.1 de [FL10]. Pour obtenir (20), l'idée est de compenser les sauts du processus Y afin d'obtenir une martingale locale. Soit $\widetilde{Z}_t = Y_t \exp\{-gt - \Delta_t\}$. La formule d'Itô conduit à

$$\widetilde{Z}_t = Y_0 + \int_0^t e^{-gs-\Delta_s}\sqrt{2\sigma^2 Y_s}dB_s + \int_0^t\int_0^\infty\int_0^{Y_{s-}} e^{-gs-\Delta_{s-}} z\widetilde{N}_0(ds,dz,du).$$

Le processus $\widetilde{Z}$ est donc une martingale locale conditionnellement à Δ. Enfin, en introduisant $F(s,x) := \exp\{-xv_t(s,\lambda,\Delta)\}$, avec $v_t(s,\lambda,\Delta)$ différentiable par rapport à s, positive et telle que $v_t(t,\lambda,\Delta) = \lambda$ pour $\lambda \geq 0$, on montre que $(F(s,\tilde{Z}_s), 0 \leq s \leq t)$ est une martingale si et seulement si pour tout $s \in [0,t]$

$$\frac{\partial}{\partial s} v_t(s,\lambda,\Delta) = e^{gs+\Delta_s}\psi_0\left(e^{-gs-\Delta_s}v_t(s,\lambda,\Delta)\right), \quad \text{p.s.},$$

où ψ_0 est définie en (22).

L'espérance de Y vérifie pour $t \geq 0$,

$$\mathbb{E}[Y_t|\Delta] = \mathbb{E}[Y_0]e^{gt+\Delta_t}.$$

Le processus de Lévy $t \mapsto gt + \Delta_t$ va donc jouer le même rôle que la marche aléatoire ou la diffusion S dans le cas des processus de Galton-Watson et des diffusions en environnement aléatoire respectivement. En particulier, le taux de croissance malthusien et les catastrophes interagissent de la manière suivante :

Corollaire 2. *On a les trois régimes suivants :*

i) Sous-critique : Si $(\Delta_t + gt)_{t\geq 0}$ *tend vers* $-\infty$, *alors* $\mathbb{P}(Y_t \to 0 \mid \Delta) = 1$ *p.s.*

ii) Critique : Si $(\Delta_t + gt)_{t\geq 0}$ *oscille, alors* $\mathbb{P}(\liminf_{t\to\infty} Y_t = 0 \mid \Delta) = 1$ *p.s.*

iii) Surcritique : Si $(\Delta_t + gt)_{t\geq 0}$ *tend vers* $+\infty$ *et qu'il existe* $\varepsilon > 0$ *tel que*

$$\int_0^\infty z\log^{1+\varepsilon}(1+z)\mu(dz) < \infty, \tag{23}$$

alors $\mathbb{P}(\liminf_{t\to\infty} Y_t > 0 \mid \Delta) > 0$ *p.s. et il existe une variable aléatoire positive* W *telle que*

$$e^{-gt-\Delta_t}Y_t \xrightarrow[t\to\infty]{W} \quad p.s., \qquad \{W = 0\} = \left\{\lim_{t\to\infty} Y_t = 0\right\}.$$

Le processus $\tilde{Z} = (Y_t \exp(-gt - \Delta_t) : t \geq 0)$ est une martingale locale positive. En particulier c'est une surmartingale positive et elle converge p.s. vers une variable aléatoire finie W. Cela prouve i)-ii). Pour prouver iii) on étudie finement les propriétés de la fonction $v_t(., \lambda, \Delta)$ solution de (21) afin de minorer sa valeur en 0 par une quantité strictement positive. Cela permet de montrer que

$$\mathbb{E}_y[\exp\{-W\} \mid \Delta] = \exp\left\{-y \lim_{t\to\infty} v_t(0,1,\Delta)\right\} < 1,$$

ce qui entraîne $\mathbb{P}(W > 0 \mid \Delta) > 0$ p.s.

Dans le cas ii) on n'a pas de résultat général concernant le comportement de la limite supérieure : on verra dans la Proposition 3 que dans le cas d'un CSBP stable avec catastrophes, $\limsup Y_t = \liminf Y_t = 0$; on a également des cas où $\limsup Y_t = \infty$ presque surement, le processus $Y_t = \exp(gt + \Delta_t)$ qui vérifie $\mu = \sigma \equiv 0$ entre dans cette catégorie.

La condition (23) est réminiscente des conditions des Théorèmes 1 et 2 pour le cas surcritique. Notre conviction profonde est que le ε n'est pas nécessaire mais nous ne sommes pas parvenus à montrer la validité de ce résultat pour $\varepsilon = 0$.

Comme dans le cas des CSBP on peut avoir extinction sans absorption dans les cas sous-critique et critique. Dans le cas critique, on a l'apparition d'un nouveau phénomène qui n'a pas lieu en environnement constant. Non seulement le CSBP avec catastrophes peut ne pas être absorbé en temps fini, mais on peut même avoir des phénomènes d'oscillations, avec $\limsup_{t\to\infty} Y_t = \infty$. Par exemple cela arrive si $\mu = 0$, $\sigma = 0$, et $Y_t = \exp(gt + \Delta_t)$.

Nous nous sommes ensuite penchés sur la vitesse d'extinction de ces processus (nous présenterons une application de ces résultats au modèle d'infection parasitaire d'une population de cellules à la fin de cette section). Nous avons en particulier étudié cette vitesse dans le cas des CSBP stables et ce pour deux raisons. D'une part les événements d'absorption et d'extinction coïncident pour cette classe de processus et la probabilité $\mathbb{P}(Y_t > 0)$ tend bien vers 0. D'autre part, comme on le verra dans le Corollaire 3, l'étude de cette classe de processus nous permet d'obtenir des bornes pour la vitesse d'extinction d'une large classe de CSBP avec catastrophes.

Le mécanisme de branchement d'un CSBP stable avec un taux de croissance g a la forme suivante :

$$\psi(\lambda) = -g\lambda + c_+\lambda^{\beta+1}, \tag{24}$$

avec $\beta \in (0,1]$, $c_+ > 0$ et g dans $\mathbb{R}$. Dans ce cas particulier, l'équation différentielle rétrograde (21) peut être résolue explicitement, ce qui fournit une expression de la probabilité d'absorption du processus :

Proposition 4. *On suppose (24) vérifiée. Alors pour tous $x_0 > 0$ et $t \geq 0$:*

$$\mathbb{P}_{x_0}(Y_t > 0 \mid \Delta) = 1 - \exp\left\{-x_0\left(c_+\beta \int_0^t e^{-\beta(gs+\Delta_s)}\, ds\right)^{-1/\beta}\right\} \qquad p.s. \tag{25}$$

De plus,

$$\mathbb{P}_{x_0}(\exists t > 0,\ Y_t = 0 \mid \Delta) = 1 \qquad p.s.,$$

si et seulement si le processus $(gt+\Delta_t, t\geq 0)$ *ne tend pas vers* $+\infty$.

Cette expression va nous permettre de déterminer l'asymptotique des vitesses d'extinction en temps long. Nous sommes donc ramenés à déterminer le comportement asymptotique d'une fonctionnelle exponentielle de processus de Lévy qui tend vers $+\infty$. L'approche naïve consistant à utiliser l'équivalent $1-\exp(-x)\sim x$ en 0 et à penser que l'équivalent suivant est vérifié

$$\mathbb{P}_{x_0}(Y_t>0)\sim_{t\to\infty} x_0\left(c_+\beta\int_0^t e^{-\beta(gs+\Delta_s)}ds\right)^{-1/\beta}$$

est pertinente dans les cas sous-critiques fort et intermédiaire mais s'avère fausse dans le cas sous-critique faible. La méthode que nous avons retenue consiste à discrétiser le processus Y avec des pas de temps $1/q$ pour q dans $\mathbb{N}$. Nous avons ensuite obtenu un équivalent de $\mathbb{P}(Y_{p/q}>0)$ pour p tendant vers $+\infty$ et en avons déduit un équivalent pour t tendant vers $+\infty$ dans $\mathbb{R}_+$ en écrivant $t=\lfloor qt\rfloor/q$ et en prenant t et q tendant vers l'infini. De ce fait nous avons eu recours à des résultats asymptotiques sur les fonctionnelles de marches aléatoires. Plus précisément, nous introduisons $K_t:=gt+\Delta_t$ et les variables aléatoires

$$A_{p,q}:=\sum_{i=0}^{p}\exp\{-\beta K_{i/q}\},\quad ((p,q)\in\mathbb{N}\times\mathbb{N}^*). \tag{26}$$

Cela nous permet d'encadrer la fonctionnelle intégrale par des fonctionnelles des variables $(A_{p,q},(p,q)\in\mathbb{N}\times\mathbb{N}^*)$ sous la forme

$$\frac{1}{q}f(-\sigma^{(+)}_{1/q})A^{(1)}_{\lfloor qt\rfloor-1,q}\leq\int_0^t e^{-\beta K_s}ds\leq\frac{1}{q}f(\sigma^{(-)}_{1/q})A^{(2)}_{\lfloor qt\rfloor,q},$$

où $f(x)\sim 1, x\to 0^+$ et pour tous $(p,q)\in\mathbb{N}\times\mathbb{N}^*$, $\sigma^{(+)}_{1/q}$ (resp $\sigma^{(-)}_{1/q}$) est indépendant de $A^{(1)}_{p,q}$ (resp $A^{(2)}_{p,q}$) et

$$A_{p,q}\overset{(d)}{=}A^{(1)}_{p,q}\overset{(d)}{=}A^{(2)}_{p,q}.$$

Les processus $\sigma^{(+)}$ et $\sigma^{(-)}$ qui apparaissent dans cette décomposition sont des subordinateurs de sauts purs indépendants et à accroissements bornés. Nous cherchons donc un équivalent de

$$\mathbb{E}\Big[1-\exp\Big(-x_0(c_+\beta y)^{-1/\beta}\Big)\Big],$$

lorsque y prend la valeur des variables aléatoires encadrant $\int_0^t e^{-\beta K_s}ds$ et voulons montrer qu'on obtient le même équivalent pour les deux encadrants. Comme la fonction $x\mapsto 1-\exp(-x)$ est croissante, cela permettra de conclure.

Sous les hypothèses additionnelles :

$$\exists\,\theta_{max}>0,\quad \phi(\lambda):=\log\mathbb{E}[e^{\lambda\Delta_1}]=\int_0^\infty(m^\lambda-1)\nu(dm)<\infty\quad\text{pour }\lambda\in[0,\theta_{max}), \tag{27}$$

et

$$\int_{(0,e^{-1}]\cup[e,\infty)}(\log m)^2\nu(dm)<\infty, \tag{28}$$

nous prouvons alors les asymptotiques suivantes pour les probabilités de survie des processus stables :

Théorème 9. *On suppose que ν vérifie (19) et (28), et que ψ et ϕ vérifient (24) et (27) respectivement.*

a/ Si $\phi'(0)+g<0$ (cas sous-critique) *et $\theta_{max}>1$, alors*

(i) Si $\phi'(1)+g<0$ (régime sous-critique fort), *alors il existe $c_1>0$ tel que pour tout $x_0>0$,*

$$\mathbb{P}_{x_0}(Y_t>0)\sim c_1 x_0 e^{t(\phi(1)+g)}, \qquad \textit{quand} \quad t\to\infty.$$

(ii) Si $\phi'(1)+g=0$ (régime sous-critique intermédiaire), *alors il existe $c_2>0$ tel que pour tout $x_0>0$,*

$$\mathbb{P}_{x_0}(Y_t>0)\sim c_2 x_0 t^{-1/2} e^{t(\phi(1)+g)}, \qquad \textit{quand} \quad t\to\infty.$$

(iii) Si $\phi'(1)+g>0$ (régime sous-critique faible) *et $\theta_{max}>\beta+1$, alors pour tout $x_0>0$, il existe $c_3(x_0)>0$ tel que*

$$\mathbb{P}_{x_0}(Y_t>0)\sim c_3(x_0) t^{-3/2} e^{t(\phi(\tau)+g\tau)}, \qquad \textit{quand} \quad t\to\infty,$$

où τ est la racine de $\phi'+g$ sur $]0,1[$: $\phi(\tau)+g\tau=\min_{0<s<1}\{\phi(s)+gs\}$.

b/ Si $\phi'(0)+g=0$ (cas critique) *et $\theta_{max}>\beta$, alors pour tout $x_0>0$, il existe $c_4(x_0)>0$ tel que*

$$\mathbb{P}_{x_0}(Y_t>0)\sim c_4(x_0) t^{-1/2}, \qquad \textit{quand} \quad t\to\infty.$$

Nous verrons une illustration de ce résultat dans le cas de l'infection parasitaire d'une population de cellules à la fin de cette section.

On retrouve donc ici encore des régimes similaires au cas des processus de branchement discrets en environnement aléatoire et à celui des diffusions branchantes en environnement brownien, mais les preuves sont très différentes du cas des diffusions branchantes. En particulier, on suppose que le processus de Lévy décrivant les catastrophes est à variations bornées, alors qu'il s'agit d'un mouvement Brownien dans [BH12]. La discrétisation du processus pose problème pour pouvoir étendre notre résultat au cas de l'environnement brownien. Cependant, cette similarité des régimes obtenus laisse penser qu'ils pourraient être généralisés à des processus de branchement et des environnements plus complexes. Il est possible, à partir du Théorème 9, de faire un pas dans cette direction et d'étudier les vitesses d'extinction de processus de branchement plus généraux que les processus stables, mais cette approche donne des encadrements et pas des équivalents de ces dernières, et on suppose toujours que le processus de Lévy des catastrophes est à variations bornées.

Corollaire 3. *On suppose (27) vérifiée et*

$$\int_{(0,\infty)} z^2\mu(dz)<\infty, \qquad \sigma^2>0, \qquad \int_{(0,\infty)} (\log m)^2\nu(dm)<\infty.$$

a/ Si $\phi'(0)+g<0$ *et* $\theta_{max}>1$, *alors*

(i) Si $\phi'(1)+g<0$, *il existe* $0<c_1\le c_1'<\infty$ *tels que pour tout* x_0,

$$c_1 x_0 e^{t(\phi(1)+g)} \le \mathbb{P}_{x_0}(Y_t>0) \le c_1' x_0 e^{t(\phi(1)+g)} \qquad \textit{pour } t \textit{ assez grand.}$$

(ii) Si $\phi'(1)+g=0$, *il existe* $0<c_2\le c_2'<\infty$ *tels que pour tout* x_0,

$$c_2 x_0 t^{-1/2} e^{t(\phi(1)+g)} \le \mathbb{P}_{x_0}(Y_t>0) \le c_2' x_0 t^{-1/2} e^{t(\phi(1)+g)} \quad \textit{pour } t \textit{ assez grand.}$$

(iii) Si $\phi'(1)+g>0$ *et* $\theta_{max}>\beta+1$, *pour tout* x_0, *il existe* $0<c_3(x_0)\le c_3'(x_0)<\infty$ *tels que*

$$c_3(x_0) t^{-3/2} e^{t(\phi(\tau)+g\tau)} \le \mathbb{P}_{x_0}(Y_t>0) \le c_3'(x_0) t^{-3/2} e^{t(\phi(\tau)+g\tau)} \quad (t>0),$$

où τ *est la racine de* $\phi'+g$ *sur* $]0,1[$.

b/ Si $\phi'(0)+g=0$ *et* $\theta_{max}>\beta$, *alors pour tout* x_0, *il existe* $0<c_4(x_0)<c_4'(x_0)<\infty$ *tels que*

$$c_4(x_0) t^{-1/2} \le \mathbb{P}_{x_0}(Y_t>0) \le c_4'(x_0) t^{-1/2} \quad (t>0).$$

La preuve consiste à encadrer le mécanisme de branchement du CSBP avec catastrophes par les mécanismes de branchement de diffusions de Feller avec catastrophes dont on connaît l'asymptotique (cas $\beta=1$). On rappelle (21) et (22). Alors la dérivée seconde de ψ_0 vérifie

$$2\sigma^2 \le \psi_0''(\lambda) = 2\sigma^2 + \int_{(0,\infty)} z^2 e^{-\lambda z}\mu(dz) \le 2\sigma^2 + \int_{(0,\infty)} z^2\mu(dz).$$

La formule de Taylor-Lagrange conduit alors, pour tout λ positif, à

$$\sigma^2\lambda^2 \le \psi_0(\lambda) \le \frac{1}{2}\Big(2\sigma^2 + \int_{(0,\infty)} z^2\mu(dz)\Big)\lambda^2.$$

Cela permet de comparer la fonctionnelle $v_t(0,\lambda,\Delta)$ avec la même fonctionnelle pour des diffusions de Feller, de coefficients de diffusion respectifs σ^2 et $\sigma^2+\int_{(0,\infty)} z^2\mu(dz)/2$, soumises aux mêmes catastrophes Δ, et on conclut grâce à l'égalité :

$$\mathbb{P}_y(Y_t=0) = \mathbb{E}\Big[\exp\Big\{-y\,v_t(0,\infty,\Delta)\Big\}\Big].$$

Pour finir, nous revenons au modèle d'infection cellulaire [BT11] qui a constitué la motivation initiale de ce travail. La quantité de parasites dans chaque cellule évolue selon une diffusion de Feller et chaque cellule, après une durée de vie exponentielle de paramètre r indépendante des autres cellules, donne naissance à deux cellules filles : une fraction Θ des parasites est transmise à une cellule fille, et une fraction $1-\Theta$ à l'autre cellule fille, où Θ est une variable aléatoire dans $(0,1)$. Les auteurs montrent que le nombre moyen de cellules infectées au temps t, N_t^*, est égal à $\mathbb{E}[N_t^*] = e^{rt}\mathbb{P}(Y_t>0)$, où le processus Y suit l'équation différentielle stochastique

$$Y_t = Y_0 + \int_0^t g Y_s ds + \int_0^t \sqrt{2\sigma^2 Y_s} dB_s + \int_0^t\int_0^1 (\theta-1) Y_{s_-}\rho(ds,d\theta). \tag{29}$$

$Y_0 >$ p.s., B est un mouvement Brownien et $\rho(ds, d\theta)$ une mesure aléatoire de Poisson d'intensité $2rds\mathbb{P}(\Theta \in d\theta)$.

Le processus Y se comporte donc comme un processus de Feller entre les temps de saut et est multliplié par une variable aléatoire Θ à chaque temps de saut. L'interprétation est la suivante. On veut connaître la probabilité qu'une cellule 'typique' soit infectée au temps t. On choisit donc une cellule uniformément au temps t et on suit l'évolution de la quantité de parasites le long de la lignée ancestrale de cette cellule, en particulier on veut savoir si cette quantité est strictement positive au temps t. Cependant la loi de Y n'est pas la loi de la quantité de parasites dans une lignée quelconque, puisque le taux de saut est $2r$ et non pas r. Ce biais est dû au fait qu'en choisissant une cellule uniformément au temps t, on a davantage de chances de choisir une lignée qui a connu de nombreuses divisions. Ainsi la lignée ancestrale d'un individu 'typique' au temps t a un taux de division $2r$.

Le Théorème 9 nous permet alors de déterminer le nombre asymptotique moyen de cellules infectées selon les paramètres de l'infection. Nous renvoyons au Chapitre 2 pour un énoncé précis. Nous pouvons cependant mentionner ici un cas particulier dans lequel on peut décrire simplement la dépendance du comportement en temps long de l'infection en fonction des paramètres du modèle : si on considère les variables aléatoires Θ qui vérifient $\mathbb{P}(\Theta = \theta) = \mathbb{P}(\Theta = 1-\theta) = 1/2$ pour un $\theta \in]0, 1/2[$, alors les differents regimes peuvent être décrits graphiquement (voir Figure 1.4).

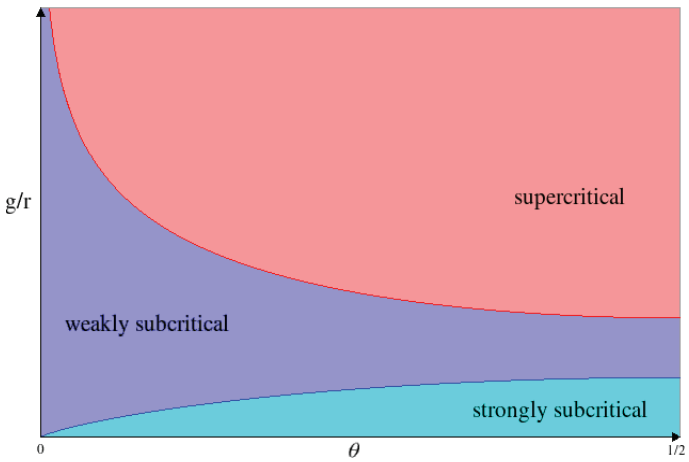

Figure 1.4: Régimes d'extinction dans le cas $\mathbb{P}(\Theta = \theta) = \mathbb{P}(\Theta = 1-\theta) = 1/2$. Les frontières entre les différents régimes sont données par $g/r = -\log(\theta(1-\theta))$ (surcritique et sous-critique) et $g/r = -\theta\log\theta - (1-\theta)\log(1-\theta)$ (fortement et faiblement sous-critiques).

1.2 Signature génétique d'un balayage sélectif dans une population sexuée

La seconde partie de cette thèse est consacrée à l'étude de la signature génétique laissée par un événement de sélection dans des modèles de populations de taille variable prenant en compte la compétition entre les individus. En particulier, nous allons voir comment les processus de recombinaison et d'adaptation interagissent au cours de l'évolution des populations sexuées. Je vais introduire dans un premier temps les notions biologiques nécessaires à la compréhension de ces processus, avant de présenter les travaux antérieurs consacrés à cette problématique et les résultats obtenus.

Quelques notions biologiques

Haploïdie, hermaphroditisme, reproduction sexuée et panmictique

Nous nous intéresserons à des populations haploïdes et hermaphrodites avec une reproduction sexuée et panmictique. Cette section est consacrée à la définition de ces termes.

L'ADN (Acide Désoxyribo-Nucléique) est un ensemble de molécules, présentes dans toutes les cellules vivantes, qui contiennent l'information génétique, et sont transmises lors de la reproduction. Chaque molécule d'ADN, associée à des protéines, constitue un chromosome, lui-même divisé, pour simplifier, en plusieurs unités d'information génétique, les gènes. Un gène détermine donc une propriété héréditaire d'un être vivant. Chaque gène peut exister dans la population sous différentes formes, ou allèles. La plupart des caractères sont polygéniques, c'est-à-dire dépendent de l'expression de plusieurs gènes, mais certains caractères sont monogéniques : par exemple une mutation du seul gène CFTR est responsable de la mucoviscidose.

Une cellule est dite haploïde lorsque les chromosomes qu'elle contient sont chacun en un seul exemplaire (n chromosomes), et diploïde lorsqu'ils sont en double exemplaire (2n chromosomes). On parle de reproduction sexuée lorsque deux individus d'une même espèce sont nécessaires pour donner naissance à un nouvel individu. Cette dernière implique donc un échange génétique et met en jeu des mécanismes de réduction (méiose) et d'augmentation (fécondation) de la répétition des chromosomes (ploïdie). Le cycle de vie d'un organisme eucaryote comprend donc nécessairement une alternance de stades avec des niveaux de répétition chromosomique différents : on parle d'alternance de phases. Chez les humains la phase haploïde (n) est très réduite. Elle correspond à la formation des gamètes : spermatozoïde ou ovocyte. Nous sommes donc des individus dits diploïdes. Chez les mousses (cf figure 1.5), chez certaines algues et champignons, la phase diploïde (2n) est au contraire beaucoup plus limitée (restreinte au zygote). Ces organismes sont alors appelés haploïdes.

Un individu est dit hermaphrodite lorsqu'il possède les fonctions sexuelles mâle et femelle. Par exemple, de nombreuses espèces de plantes présentent des organes de reproduction femelles et mâles sur un même organisme. C'est également le cas des escargots, des sangsues ou des lombrics.

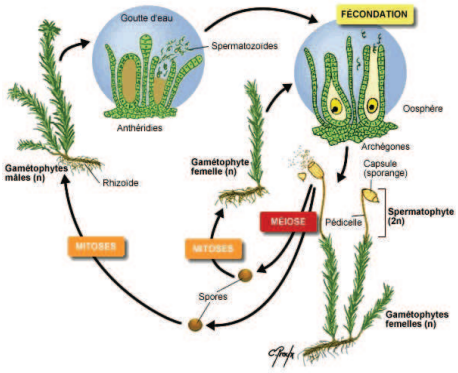

Figure 1.5: Cycle de vie des mousses [fig]

Enfin on parle de reproduction panmictique lorsque l'on considère que les individus sont répartis de manière homogène au sein de la population et choisissent leur partenaire uniformément parmi tous les individus de la population.

Processus d'adaptation et recombinaison génétique

L'adaptation est un processus de transformation d'organismes conduisant à une plus grande adéquation à leur environnement. Elle opère par la modification des propriétés héritables des organismes (mutations) et par l'augmentation progressive des fréquences des génotypes qui favorisent la survie et/ou la reproduction (sélection naturelle).

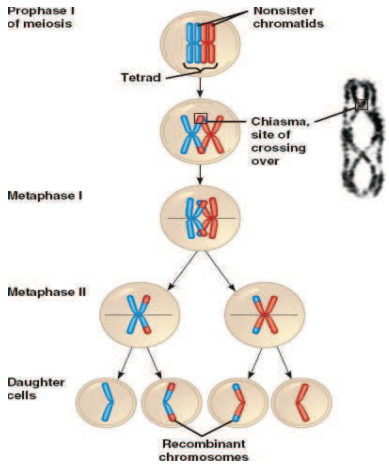

Figure 1.6: Méiose et recombinaison [mei]

La recombinaison génétique est "le phénomène conduisant à l'apparition, dans une cellule ou dans un individu, de gènes ou de caractères héréditaires dans une association différente de celle observée chez les cellules ou individus parentaux" [JO00]. Elle se produit au cours de la

méiose (cf figure 1.6), qui est la division cellulaire conduisant à la formation des gamètes ou des spores. La méiose donne quatre cellules haploïdes (n) à partir d'une cellule diploïde (2n) contenant un exemplaire 'maternel' et 'paternel' de chaque chromosome (on a donc également un doublement du matériel génétique durant ce processus). Au cours de l'appariement des chromosomes homologues, des enjambements et des échanges de fragments entre chromosomes homologues peuvent se produire (cf figure 1.6). On dit qu'il y a une recombinaison. Cela crée de nouvelles combinaisons d'allèles qui n'étaient présentes chez aucun des deux parents.

Nous venons de décrire deux phénomènes, l'adaptation et la recombinaison, qui ont des effets opposés sur la diversité génétique. D'une part la sélection naturelle entraîne l'augmentation de la fréquence de certains allèles, et en conséquence d'allèles voisins présents sur le même chromosome que l'allèle sélectionné. En effet, en l'absence de recombinaison, les chromosomes sont transmis tels quels aux descendants et la sélection positive d'un allèle entraîne l'augmentation en fréquence de la combinaison d'allèles du chromosome sur lequel il se trouve. D'autre part, la présence de recombinaisons chez les espèces qui se reproduisent de manière sexuée est à l'origine d'un brassage génétique et de la formation incessante de nouveaux chromosomes, qui diffèrent de ceux des générations précédentes par leurs combinaisons alléliques.

Nous nous sommes donc intéressés à l'interaction entre ces deux processus, et en particulier à ce que cette interaction pouvait nous apprendre sur l'histoire évolutive des populations. Quelles informations les variations de diversité génétique engendrées par ces deux processus nous donnent sur les régions chromosomiques qui ont connu une sélection récente ?

Actuellement, de nombreuses méthodes sont développées pour détecter les gènes récemment sélectionnés ou encore sous sélection à partir de données moléculaires [KS02, FW05, SSF+06]. Mais pour que ces méthodes de détection soient efficaces, il est nécessaire de s'appuyer sur des modèles les plus réalistes possibles. C'est l'objectif de la seconde partie de cette thèse.

Balayages sélectifs (selective sweeps)

Lorsqu'un allèle avantageux parvient à se fixer dans une population, les loci voisins du locus où l'allèle s'est fixé se retrouvent également dépourvus d'une part de leur variation neutre comme simple conséquence de leur liaison génétique, les allèles du mutant initial étant "auto-stoppés" par la fixation de l'allèle avantageux (voif figure 1.7). Ce phénomène s'appelle le "balayage sélectif" (selective sweep). A plus longue distance, l'effet de balayage s'estompe progressivement en raison des recombinaisons, jusqu'à ne plus être repérable.

La sélection laisse donc un "trou" de variation neutre dans le génome. C'est la signature génétique de la sélection naturelle. La longueur de la région affectée fournit de l'information. L'amplitude de l'effet d'auto-stop augmente avec la force de la sélection et dépend du scenario ayant conduit à la fixation. Elle décroît avec celle du taux de recombinaison, qui tend à découpler l'association entre régions voisines.
On distingue le balayage sélectif dur (hard sweep) du balayage sélectif doux (soft sweep). Le

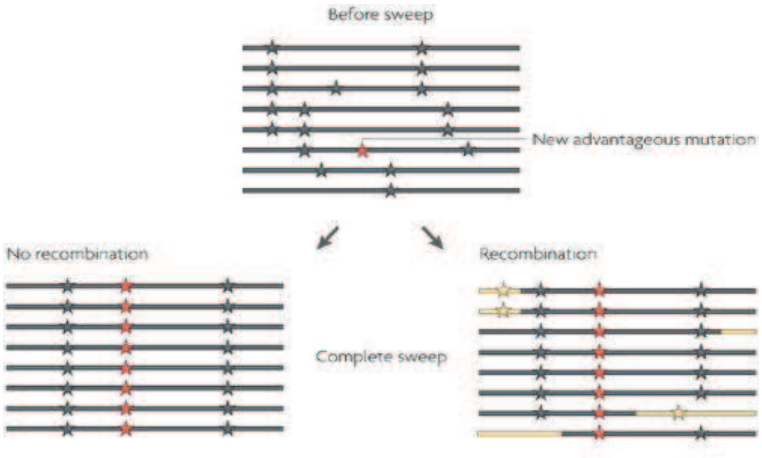

Figure 1.7: Schéma d'un balayage sélectif [NHH+07]

premier découle de l'apparition dans une population d'une nouvelle mutation positivement sélectionnée qui se fixe dans la population. Dans le second cas, un allèle déjà présent en plusieurs exemplaires dans la population voit ses conditions de sélection changer (déplacement de la population, apparition d'un nouveau prédateur, catastrophe climatique,...) et devient avantageux. Les nouvelles mutations sont des sources de diversité, et les balayages durs ont longtemps été les seuls vecteurs d'adaptation considérés. Les balayages doux permettent une adaptation plus rapide aux nouveaux environnements, et leur importance est croissante, tant dans les études théoriques qu'empiriques (Hermisson et Pennings [HP05, PH06a, PH06b], Prezeworski, Coop et Wall [PCW05], Barrett et Schluter [BS08]). En particulier Messer et Petrov [MP13] passent en revue un grand nombre d'exemples de soft sweep mis au jour récemment qui montrent que ces derniers sont courants dans de nombreux organismes, des virus aux insectes en passant par les mammifères.

Auto-stop génétique et invasion d'un mutant

Auto-stop génétique, une littérature foisonnante

Nous allons maintenant présenter les principaux travaux traitant de l'effet d'auto-stop génétique dans le cas des balayages durs sur lesquels nous nous sommes davantage penchés. Pour un résumé des travaux antérieurs à 2000, on pourra pour plus de précisions se référer à [Bar00].

Les premiers biologistes à avoir étudié cette question sont Maynard Smith et Haigh en 1974 [SH74]. Ils considèrent une population de taille infinie et à générations discrètes. Ils distinguent un locus sous sélection (allèle b ou B) et un locus neutre (allèle a ou A). L'allèle B est favorable. Sa fitness est caractérisée par un paramètre $s > 0$ et un individu portant l'allèle B laisse $1 + s$ fois plus de descendants dans la génération suivante qu'un individu portant l'allèle b. Les

fréquences des différents génotypes à la génération n sont notées :

Génotype	AB	aB	Ab	ab
Fréquence	$p_n Q_n$	$p_n(1-Q_n)$	$(1-p_n)R_n$	$(1-p_n)(1-R_n)$
Fitness	$1+s$	$1+s$	1	1

Q_n et R_n représentent donc les proportions d'allèles A dans les populations de types B et b respectivement, à la génération n. En examinant successivement les dix génotypes possibles pour les couples de parents à la génération n, les auteurs dérivent des expressions explicites des fréquences génotypiques à la génération $n+1$. En particulier ils obtiennent la relation suivante :

$$Q_{n+1} - R_{n+1} = (1-r)(Q_n - R_n) = \ldots = -R_0(1-r)^{n+1},$$

si on suppose que le premier mutant a le génotype aB. Cette égalité traduit le fait que les recombinaisons cassent les liaisons préférentielles entre allèles : tant que les proportions d'allèles A ne sont pas les mêmes dans les populations de types b et B, la recombinaison réduit l'écart entre ces deux proportions. Cette réduction est rapide, géometrique en le nombre de générations. Cela permet à Maynard-Smith et Haigh de déduire la proportion finale d'allèles A après fixation du mutant B :

$$Q_\infty = rR_0(1-p_0)\sum_{n=0}^{\infty} \frac{(1-r)^n}{1-p_0+p_0(1+s)^{n+1}}.$$

C'est le premier résultat quantitatif sur l'effet de la recombinaison sur les proportions neutres.

Ils dérivent ensuite des équations analogues dans le cas diploïde. Comme les expressions obtenues sont plus complexes ils les approchent par des équations différentielles partielles lorsque le coefficient de sélection s et la probabilité de recombinaison r sont suffisamment petits. Ils obtiennent ainsi un équivalent de dQ/dp. Pour en déduire la proportion finale Q_∞ d'allèles A il suffit donc d'intégrer cet équivalent en dp entre l'arrivée du premier mutant et la fixation de l'allèle B. Malheureusement dans le cas de la dominance[3] cette intégrale est divergente en $p=1$. Ils résolvent ce problème de manière arbitraire en intégrant entre $p=\varepsilon$ et $p=1-\varepsilon$ pour $\varepsilon>0$ petit.

Comme les auteurs de cet article ne manquent pas de le faire remarquer, ce modèle déterministe comporte deux faiblesses majeures. D'une part la croissance initiale du nombre de mutants est aléatoire et la population mutante, même favorable peut s'éteindre. Ainsi, la croissance initiale des populations mutantes qui arrivent effectivement à se fixer devrait être supérieure à celle du modèle déterministe. D'autre part dans le modèle déterministe la proportion finale d'allèles B tend vers 1 sans jamais l'atteindre, alors que les fluctuations réelles de fréquences dans la population permettent d'atteindre 1 en temps fini.

Dans un article de 1992 [SWL92], Stephan, Wiene et Lenz se penchent à leur tour sur la question de l'autostop génétique. Là encore ils introduisent un $\varepsilon>0$ petit et se concentrent sur l'intervalle de temps $[t_\varepsilon, 1-t_\varepsilon]$ pendant lequel la fréquence de l'allèle positivement sélectionné,

[3] On dit que l'allèle B est dominant si les individus de génotype (b,B) et (B,B) ont la même fitness $1+s$ alors que les individus de génotype (b,b) ont une fitness 1.

B, appartient à l'intervalle $[\varepsilon, 1-\varepsilon]$. Ils considèrent qu'il n'y a pas de recombinaison avant le temps t_ε. La fréquence de l'allèle sous sélection B suit une équation logistique déterministe,

$$\frac{dp(t)}{dt} = sp(t)(1-p(t)),$$

qui se résout explicitement

$$p(t) = \frac{\varepsilon}{\varepsilon + (1-\varepsilon)e^{-s(t-t_\varepsilon)}}, \quad t \geq t_\varepsilon.$$

Suivant l'approche de Ohta et Kimura [OK75] qui s'étaient intéressés à une question légèrement différente, ils modélisent la fréquence des allèles neutres par des diffusions avec des dérives résultant des échanges entre les populations B et b via les recombinaisons. En particulier ils obtiennent ainsi des équations différentielles ordinaires vérifiées par les moments d'ordre 1 et 2 de $Q(t)$ et $R(t)$, les fréquences respectives d'allèles A dans les populations B et b, toujours dans l'intervalle de temps $[t_\varepsilon, 1-t_\varepsilon]$, et en déduisent la variation d'hétérozygotie au locus neutre due au balayage sélectif. Dans leur modèle, la différence des proportions neutres satisfait :

$$\frac{d\mathbb{E}[Q(t)-R(t)]}{dt} = -r\mathbb{E}[Q(t)-R(t)].$$

On peut remarquer que les équations déterministes obtenues dans ce modèle (taille de la population de type B et moment d'ordre 1 pour la différence des proportions neutres) correspondent à un développement limité du premier ordre en $s \ll 1$ des équations continues (obtenues comme limites des équations pour les générations discrètes) de [SH74] ; en effet Maynard Smith et Haigh avaient dérivé les équations :

$$\frac{dp(t)}{dt} = sp(t)(1-p(t))/(1+sp(t)),$$

et

$$\frac{d\mathbb{E}[Q(t)-R(t)]}{dt} = -r\frac{1-sp(t)}{1+sp(t)}\mathbb{E}[Q(t)-R(t)].$$

Jusqu'à présent, les nombres d'individus portant les allèles B et b étaient modélisés par des processus déterministes. Barton reprend l'étude de l'auto-stop génétique en 1998 [Bar98] avec pour but la prise en compte de la stochasticité initiale de la taille de la population mutante, négligée dans les travaux précédents. Cependant il suppose comme ses prédécesseurs que la taille de la population totale est constante égale à $2N$ et que le coefficient de sélection est petit, mais grand devant $1/2N$ ($1/N \ll s \ll 1$). Il distingue quatre phases dans le processus de fixation de l'allèle B :

1. Durant une première phase stochastique le nombre de mutants k varie selon un processus de branchement de paramètre malthusien s. Seule une fraction $2s$ des trajectoires va conduire à la fixation.

2. Lorsque k a atteint une valeur suffisante ($ks > 2$) les mutants sont suffisamment nombreux pour considérer leur croissance comme déterministe mais constituent encore une fraction négligeable de la population ($k \ll 2N$)

3. Durant la troisième phase, qui a une durée de l'ordre de $1/s$, la fréquence du mutant dans la population augmente de manière déterministe jusqu'à atteindre une valeur proche de 1.

4. Enfin dans la quatrième et dernière phase l'allèle résidant b s'éteint.

Pour déterminer l'effet du balayage sélectif sur la distribution d'un allèle neutre voisin de l'allèle sous sélection, il s'intéresse à la généalogie de cet allèle neutre. Plus précisément, il échantillonne des individus à la fin du balayage et il remonte la généalogie de leurs allèles neutres pour savoir s'ils descendent du premier mutant ou s'ils étaient associés à un allèle b au moment de la mutation. Son approche cherche à corriger les estimations précédentes en prenant en compte la phase stochastique. En particulier Barton intègre le fait que le nombre de mutants augmente plus vite peu après la mutation si la trajectoire considérée conduit à la fixation que si elle conduit à la disparition du mutant, et que les probabilités de coalescer ou de recombiner pour deux lignées neutres dépendent du nombre d'allèles b et B à l'instant considéré.

Cette approche généalogique, introduite par Barton, a été reprise par la suite dans de nombreuses études du hitchhiking, en particulier par Schweinsberg et Durrett [SD05] et par Etheridge et ses coauteurs [EPW06], mais le papier reste heuristique et ne développe pas complètement les preuves mathématiques.

Durrett et Schweinsberg [DS04, SD05] reprennent cette idée de suivre les généalogies neutres et font une analyse rigoureuse des échanges d'allèles neutres durant le balayage sélectif. Là encore tous les individus portent l'allèle b jusqu'à l'arrivée d'un mutant favorable B. Ils modélisent la dynamique de la population par un processus de Moran avec sélection prenant en compte deux loci :

- La taille de population $2N$ est constante
- Chaque individu a une durée de vie indépendante des autres individus et qui suit une loi exponentielle de paramètre 1
- Lorsqu'un individu meurt il est remplacé par un autre individu :
    - avec probabilité $1 - r_N$ il n'y a pas de recombinaison et le nouveau né est choisi uniformément parmi tous les invididus (y compris celui qui va mourir)
    - avec probabilité r_N il y a une recombinaison et on choisit uniformément deux individus dans la population : le premier transmet son allèle sous sélection, b ou B, et le deuxième son allèle neutre, a ou A.
- Pour prendre en compte la meilleure fitness de la nouvelle mutation on ajoute la règle suivante : si l'individu qui meurt porte l'allèle mutant B et que l'individu choisi uniformément pour transmettre son allèle sous sélection porte l'allèle résident b, le remplacement est rejeté avec probabilité $0 < s < 1$ et l'individu portant l'allèle B reste en vie. Ainsi un individu portant l'allèle B a en moyenne davantage de descendants qu'un individu portant l'allèle b.

- le procédé est réitéré lorsqu'une nouvelle horloge exponentielle sonne.

Schweinsberg et Durrett cherchent à comprendre l'influence du balayage sélectif sur la diversité neutre en fonction de la probabilité de recombinaison. Ils échantillonnent donc n individus (n entier donné) à la fin du sweep sélectif et retracent leur généalogie pour savoir s'ils descendent du mutant initial ou d'un individu portant l'allèle b, et combien d'allèles neutres ont le même ancêtre. Ils montrent que les généalogies des allèles neutres échantillonnés sont asymptotiquement indépendantes, avec une erreur de l'ordre de $1/\log N$. L'allèle neutre de chaque individu échantilloné descend du mutant initial avec probabilité $p_N := e^{-r_N \log(2N)/s}$, et descend d'un individu de type b avec probabilité $1-p_N$. D'autre part, deux allèles neutres échappant au sweep descendent de deux individus b distincts.

Pour prouver ce résultat, l'idée consiste à suivre la généalogie des allèles neutres échantillonnés à la fin du balayage sélectif. Deux types d'événements ont une influence sur la généalogie :

1. les événements de coalescence : on dit que deux allèles neutres coalescent à la génération $n \in \mathbb{N}$ si l'un des allèles est porté par un individu qui vient de naître et a été transmis par l'autre individu. En d'autres terme les généalogies de ces deux allèles neutres sont confondues avant la génération n.

2. les événements de recombinaison : on dit qu'un allèle neutre recombine à la génération n s'il est porté par un individu qui vient de naître et si cet individu a hérité son allèle neutre et son allèle sous sélection de deux parents distincts.

La probabilité pour deux allèles neutres donnés de coalescer à la génération n ne dépend que du nombre d'individus de type b et B aux générations $n-1$ et n. Ainsi la première étape de la preuve consiste à compter le nombre de sauts de l à $l+1$, de l à $l-1$ et de l à l pour le processus représentant le nombre d'individus portant l'allèle B. Ensuite, il faut dénombrer le nombre d'événements de coalescence et de recombinaisons le long des généalogies, conditionnellement au processus de taille de la population B.

Ce résultat permet d'avoir une approximation des moments d'ordre d de fonctionnelles liées à la partition (sous réserve que N soit suffisamment grand), et quantifie l'erreur qui est commise en utilisant cette approximation. Durrett et Schweinsberg vont même plus loin en construisant une partition qui approche la distribution du modèle avec une erreur $1/\log^2 N$ ([SD05] Théorème 1.2).

Etheridge, Pfaffelhuber et Wakolbinger [EPW06] s'intéressent à leur tour à cette question, mais dans un cadre un peu différent. Ils considèrent une population de taille constante $2N$ constituée de deux types d'individus. Les individus portant l'allèle B ont un avantage sélectif par rapport aux individus portant l'allèle b. Ils s'intéressent à la proportion d'individus P porteurs de l'allèle B et approchent cette dernière, lorsque N est grand, par une diffusion de Wright-Fisher avec sélection :

$$dP = \sqrt{2P(1-P)}dW + \alpha P(1-P)dt,$$

où W est un mouvement brownien standard et $s = \alpha/2N$ est l'avantage sélectif de l'allèle B par individu. La probabilité de recombinaison entre l'allèle sous sélection et un allèle neutre voisin est r, ce qui conduit à un taux de recombinaison $\rho = 2Nr$. Ils considèrent la limite en grande sélection, $\alpha \gg 1$. La durée du sweep est de l'ordre de $\log\alpha/\alpha$. En conséquence, si l'on veut observer un nombre non trivial de recombinaisons, ρ doit être de l'ordre de $\alpha/\log\alpha$. Ils supposent donc :

$$\rho = \gamma\frac{\alpha}{\log\alpha}, \quad 0 < \gamma < \infty. \tag{30}$$

Ils introduisent la définition suivante pour la partition ancestrale au locus neutre :

Définition 4. *On échantillonne un allèle neutre chez n individus à la fin du balayage sélectif. On peut alors distinguer différentes familles en fonction de la généalogie :*

1. *Une famille est appelée non-recombinante si les lignées neutres de tous les individus de la famille sont restées dans la population de type B.*
2. *Une famille est appelée recombinante-primitive si aucune lignée n'a quitté la population B avant la première coalescence (en remontant le temps) mais si tous les individus de la famille descendent d'un individu de type b au début du sweep.*
3. *Une famille est appelée recombinante-tardive si au moins une lignée a quitté la population B avant la première coalescence (en remontant le temps) et si tous les individus de la famille descendent d'un individu de type b au début du sweep.*
4. *Dans les autres cas, la famille est appelée exceptionnelle.*

Ils dérivent alors une approximation de la partition ancestrale d'un échantillon de n individus à la fin du sweep et donnent la loi de répartition des quatre types de familles décrites, avec une erreur de l'ordre de $1/\log^2\alpha$.

La preuve de ce résultat suit essentiellement quatre étapes. De manière analogue à [SD05], ils raisonnent conditionnellement à la taille (ici la proportion) de la population B. Il s'agit donc de déterminer la loi de cette dernière lorsque l'allèle B parvient à se fixer. Conditionnellement à la fixation de l'allèle B, la proportion d'allèle B suit l'équation différentielle stochastique :

$$dX = \sqrt{2X(1-X)}dW + \alpha X(1-X)\coth\left(\frac{\alpha}{2}X\right)dt, \quad t > 0$$

et est issue de 0 au temps $t = 0$. La seconde étape consiste à prouver que les événements qui ont lieu au sein de la population de type b (coalescences et retour d'un allèle neutre dans la population de type B (en remontant le temps)) sont des événements négligeables. Ainsi, il suffit de prendre en compte les coalescences au sein de la population de type B et les recombinaisons de B vers b (toujours considérées en remontant le temps). Seule une fraction $(1-X)$ des recombinaisons est une recombinaison de B vers b. Ainsi, la troisième étape consiste à ajouter au coalescent décrivant la généalogie des allèles sélectionnés B un processus de Poisson d'intensité $\rho(1-X_t)dt$ pour représenter ces recombinaisons. Enfin, la dernière étape consiste à comparer l'évolution du nombre d'individus porteurs de l'allèle B au début du sweep à un processus de Yule de taux de branchement α.

Les auteurs font remarquer que leurs résultats sont en accord avec ceux obtenus dans [SD05] dans le sens suivant : si les paramètres α, ρ et γ vérifient

$$\alpha = 2Ns, \quad \rho = 2Nr, \quad \text{et} \quad \gamma = \frac{r}{s}\log\alpha,$$

alors la différence entre la deuxième approximation de [SD05] et la leur est d'ordre $1/\log^2 N$.

Nous concluons ici notre exposé bibliographique sur le phénomène du hitchhiking car il donne une idée assez précise de l'évolution des modèles étudiés. Dans un premier temps, les tailles des populations de types B et b ont été modélisées de manière déterministe, avec dans certains cas de la stochasticité dans l'évolution du nombre d'allèles neutres. Dans un deuxième temps les modèles ont pris en compte la stochasticité de la taille de la population mutante, mais les auteurs se sont restreint à une taille totale de population constante. Une des finalités de l'approche que nous allons présenter dans la section suivante est de relâcher cette hypothèse de taille de population fixe, difficile à justifier dans de nombreuses situations réelles. Mentionnons cependant avant de conclure cette introduction bibliographique quelques travaux ultérieurs qui étendent certains des résultats présentés. Pfaffelhuber et Studeny [PS07], puis Léocard [Leo09] ont étendu les résultats de [EPW06] à deux et à un nombre quelconque de loci neutres respectivement, et Barton et ses coauteurs [BEKV13] se sont intéressés à l'influence d'une structure spatiale sur le phénomène d'auto-stop génétique.

L'approche éco-évolutive

Les modèles que nous avons présentés ont deux points communs : d'une part, la population est de taille constante (finie ou infinie) ; d'autre part, chaque individu de la population a une fitness donnée, 1 pour les individus résidants et $1+s>1$ pour les individus mutants, qui ne dépend pas de l'état de la population. C'est une bonne approximation au début du balayage sélectif. En effet à ce moment là, les mutants sont en faible nombre et n'influencent pas la dynamique du résident largement majoritaire. Les résidents sont donc proches de l'équilibre et ont un paramètre malthusien nul alors que les mutants sont en compétition avec des individus moins bien adaptés à l'état actuel de la population et ont un paramètre malthusien strictement positif. En revanche lorsque le nombre de mutants augmente ces derniers sont en compétition avec des individus semblables de plus en plus nombreux et ils deviennent moins bien adaptés relativement à l'ensemble de la population.

Il est important de comprendre que ce qui fait qu'un individu est bien adapté ou non n'est pas seulement une caractéristique intrinsèque (son génotype ou son phénotype suivant l'échelle de modélisation) mais dépend du milieu dans lequel il évolue. Ainsi, les "traits" des individus vont avoir une influence sur ses interactions avec les autres individus (compétition, coopération, descendance plus ou moins grande dans la population), et ces interactions à leur tour vont générer de la sélection sur ces traits. C'est l'idée fondamentale de Darwin : la variabilité apparaît à l'échelle de l'individu qui subit les influences de toute la population ; ainsi la composition génétique de la population, sa taille, sa structure (spatiale, en âges, en types sexuels,...) évoluent de manière interdépendante. La formalisation mathématique de ce

type d'approche éco-évolutive a été initiée dans les années 90 dans le cadre déterministe par Hofbauer et Sigmund [HS90], et largement développée et étudiée par Metz et ses coauteurs [MGM+96, GMKM97, DMM08]. Bolker et Pacala [BP99] et Law et Dieckmann [LD00] se sont intéressés à des modèles aléatoires et ont dérivé des approximations déterministes des moments du processus au cours du temps. Fournier et Méléard [FM04] ont introduit et étudié de manière précise un processus stochastique microscopique exhibant des comportements macroscopiques analogues à ceux du modèle introduit par Bolker et Pacala. Cette approche probabiliste initiée par Fournier et Méléard a ensuite fait l'objet de nombreux développements, que ce soit dans le cas haploide asexué avec une ou plusieurs ressources par Champagnat, Ferrière, Méléard et leurs coauteurs (voir [Cha06, CFM06, CM11, BFMT13] et leurs références), ou dans le cas diploide sexué par Collet, Méléard et Metz [CMM11], puis Coron et ses coauteurs [CMPR13, Cor13, Cor14].

Les modèles de l'auto-stop génétique que nous avons présentés sont des modèles de génétique des populations. Cette dernière, qui a pris son essor à partir des travaux de Mendel, se concentre sur la distribution et les changements de fréquence des allèles dans les populations sous l'influence des pressions sélectives : sélection naturelle, dérive génétique, mutations, recombinaisons,... mais elle ne s'intéresse pas à l'effet de ces modifications de fréquences sur la taille de la population et sur l'évolution des interactions entre individus. La dynamique des populations quant à elle étudie la répartition et le développement quantitatif de populations d'individus. Elle s'intéresse aux mécanismes d'auto-régulation (compétition pour les ressources ou l'espace, contrôle des naissances, ...) des populations, au problème de l'extinction d'une population ou à l'existence d'un éventuel état stationnaire ou quasi-stationnaire.

Notre but dans l'étude de la signature de la sélection est de faire le lien entre génétique des populations et dynamique des populations, avec un point de vue unifié prenant en compte à la fois la composition génétique et la généalogie des populations étudiées et la dynamique quantitative du nombre d'individus et des caractéristiques des interactions entre les individus de ces populations.

Invasion d'un mutant positivement sélectionné

Champagnat décrit précisément dans [Cha06] la dynamique de l'invasion d'un mutant dans une population initialement monomorphe et à l'équilibre. Nous nous appuierons sur cette description pour étudier, dans les Chapitres 3 et 4, la signature laissée par un balayage sélectif dans une population haploide sexuée. C'est pourquoi nous présentons maintenant ce travail.

Chaque individu est caractérisé par un trait, a ou A et on note N_A et N_a les tailles des populations respectives. Les interactions entre les individus sont décrites par des paramètres écologiques qui dépendent des traits de chaque individu. Un individu de trait α

- donne naissance à taux f_α
- meurt de mort naturelle à taux D_α
- subit une compétition C_{α,α_1} de chaque individu de la population de trait α_1.

On introduit de plus un paramètre K qui permet de modifier la force de la compétition. Ce paramètre quantifie les ressources et l'espace disponibles. L'idée est de modéliser le fait que la compétition diminue lorsque la quantité de ressource augmente, et que la taille d'équilibre que peut atteindre une population dans un environnement donné dépend des ressources dont elle dispose. On notera N^K la population pour indiquer la capacité de charge, puisque celle-ci a une influence sur les taux de morts. Si l'on considère une population dans laquelle deux types d'individus A et a coexistent, le taux de naissance agrégé des individus portant l'allèle $\alpha \in \mathscr{A} := \{A, a\}$ est :

$$b_\alpha(N^K) = f_\alpha N^K_\alpha, \tag{31}$$

et le taux de mort agrégé :

$$d_\alpha(N^K) = \Big(D_\alpha + \frac{C_{\alpha,A}}{K} N^K_A + \frac{C_{\alpha,a}}{K} N^K_a\Big) N^K_\alpha, \tag{32}$$

où $N^K = (N^K_A, N^K_a)$ est l'état courant de taille de la population. On peut quantifier l'adaptation d'un mutant de type α dans une population de type $\bar{\alpha} \neq \alpha$ à l'équilibre à l'aide de la fitness d'invasion

$$S_{\alpha\bar{\alpha}} := f_\alpha - D_\alpha - C_{\alpha,\bar{\alpha}} \bar{n}_{\bar{\alpha}},$$

où la densité d'équilibre $\bar{n}_\alpha$ est définie par

$$\bar{n}_\alpha = \frac{f_\alpha - D_\alpha}{C_{\alpha,\alpha}}.$$

La fitness d'invasion $S_{\alpha\bar{\alpha}}$ est donc la croissance initiale d'un mutant de type α qui apparaît dans une population de type $\bar{\alpha}$ à l'équilibre. Le rôle de la fitness d'invasion $S_{\alpha\bar{\alpha}}$ et la définition de la densité d'équilibre $\bar{n}_\alpha$ découlent des propriétés du système compétitif de Lotka-Volterra :

$$\dot{n}^{(z)}_\alpha = (f_\alpha - D_\alpha - C_{\alpha,A} n^{(z)}_A - C_{\alpha,a} n^{(z)}_a) n^{(z)}_\alpha, \quad n^{(z)}_\alpha(0) = z_\alpha, \quad \alpha \in \mathscr{A}. \tag{33}$$

Si on suppose que $S_{Aa} < 0 < S_{aA}$, alors le système (33) a un unique équilibre stable $(0, \bar{n}_a)$ et deux états stationnaires instables $(0,0)$ et $(\bar{n}_A, 0)$. A l'aide du Théorème 2.1 p. 456 de Ethier and Kurtz [EK86] on peut montrer que si $N^K_A(0)$ et $N^K_a(0)$ sont de l'ordre de K, alors dans la limite en grande population ($K \to \infty$), le processus $(N^K_A(0)/K, N^K_a(0)/K)$ converge en probabilité vers la solution de (33) sur tout intervalle de temps fini (voir Proposition 2 de [Cha06] pour un énoncé précis de ce résultat). De plus, quand le mutant a apparaît dans une population monomorphe d'individus A à l'équilibre (taille $\bar{n}_A K$), la fitness d'invasion S_{aA} correspond à la croissance individuelle initiale de la population mutante. Ainsi, la dynamique de la population de type a est très dépendante des propriétés du système (33) et Champagnat a prouvé que sous l'hypothèse

Hypothèse 1.

$$\bar{n}_A > 0, \quad \bar{n}_a > 0, \quad \textit{et} \quad S_{Aa} < 0 < S_{aA},$$

l'allèle a a une probabilité positive de se fixer dans la population et de remplacer l'allèle résident A. Il donne même un équivalent de cette probabilité en grande population (Equation (39) de [Cha06]) :

$$\lim_{K\to\infty} \mathbb{P}(\text{fixation de } a | (N_A^K(0), N_a^K(0)) = (\bar{n}_A K, 1)) = \frac{S_{aA}}{f_a}. \tag{34}$$

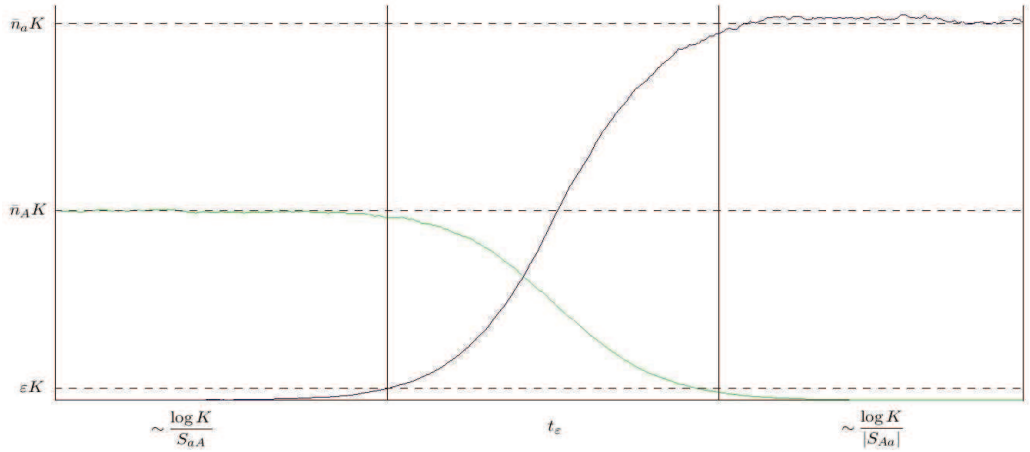

Figure 1.8: Les trois phases de l'invasion d'un mutant ; dans cette simulation, $K = 10\,000$ et $S_{aA} = |S_{Aa}| = 4$.

On peut diviser les trajectoires conduisant à la fixation du mutant en trois phases successives (cf figure 1.8). Durant la première phase, qui a une durée de l'ordre de $\log K$, le mutant constitue une fraction de la population inférieure à ε (choisi positif et petit) et a peu d'influence sur la dynamique du résident. Ce dernier reste donc proche de son état d'équilibre $\bar{n}_A K$ alors que la taille de la population mutante, comparable à un processus de naissance et mort surcritique de paramètres f_a (naissance) et $D_a + C_{a,A}\bar{n}_A = f_a - S_{aA}$ (mort) évolue de 1 (état initial) à $\lfloor \varepsilon K \rfloor$ (fin de la première phase). Une fois que les deux populations ont une taille de l'ordre de K, on peut comparer leur évolution à celle des solutions du système (33). Durant la phase 2, la solution du système dynamique atteint un état proche de l'équilibre stable. Plus précisément, on s'arrête lorsque $n_A(t) \leq \varepsilon^2$ et $|n_a(t) - \bar{n}_a| \leq \varepsilon$. Durant la troisième phase du balayage sélectif, les individus de type A sont rares. Il faut donc de nouveau étudier la population de manière stochastique. En effet, si les fluctuations étaient négligeables durant la phase 2 par rapport à l'effet de moyenne décrit par le système déterministe (33), elles deviennent de nouveau prépondérantes lorsque l'une des populations est en effectif restreint (en particulier la solution du système dynamique n_A tend vers 0 sans jamais l'atteindre). Durant cette dernière phase, qui a une durée de l'ordre de $\log K$, la population a reste proche de sa taille d'équilibre et la dynamique de la population A peut être approchée par un processus de naissances et morts sous-critique de paramètres f_A (naissance) et $f_A + |S_{Aa}|$ (mort) qui s'éteint presque sûrement.

Résultats du Chapitre 3

Dans le Chapitre 3, j'étudie la signature sur la diversité neutre laissée par un balayage sélectif dans une population haploïde sexuée. Comme je l'ai mentionné précédemment, je m'intéresse à des populations de taille variable et modélise de manière précise les interactions entre les individus. Les populations évoluent selon des processus de naissance et mort multitypes et non-linéaires. Nous considérons deux loci voisins sur un même chromosome : le premier porte l'allèle A ou l'allèle a, qui donnent lieu à des paramètres écologiques différents ; le second locus porte l'allèle b_1 ou l'allèle b_2 qui n'ont aucune influence sur les paramètres écologiques que l'on considère. On dira qu'ils sont neutres. Chaque individu a donc un génotype (α, β), avec $\alpha \in \mathscr{A} := \{A, a\}$ et $\beta \in \mathscr{B} := \{b_1, b_2\}$. Les paramètres écologiques ne dépendent que de l'allèle sous sélection, A ou a, et sont définis comme dans [Cha06]. On obtient donc des taux de mort analogues à (32) :

$$d^K_{\alpha\beta}(n) = \left[D_\alpha + C_{\alpha,A} n_A/K + C_{\alpha,a} n_a/K\right] n_{\alpha\beta}, \quad (\alpha,\beta) \in \mathscr{E} := \mathscr{A} \times \mathscr{B}. \tag{35}$$

Les taux de naissance en revanche sont différents, puisque l'on considère une population sexuée qui se reproduit de manière Mendélienne. Il faut maintenant deux invididus pour engendrer un descendant. Le paramètre f_α représente à la fois le taux auquel un individu donne naissance ("fonction femelle") et le taux de production de gamètes mâles ("fonction mâle"). Ainsi, un individu de type (α, β) engendre un descendant au taux f_α, et la proportion de gamètes mâles de type (α, β) est

$$p_{\alpha\beta}(n) = \frac{f_\alpha n_{\alpha\beta}}{f_A n_A + f_a n_a},$$

où $n = (n_{Ab_1}, n_{Ab_2}, n_{ab_1}, n_{ab_2})$ désigne le nombre courant d'individus des différents génotypes. Lorsque un individu de type α donne naissance, il choisit un partenaire uniformément parmi les gamètes disponibles. A chaque événement de naissance, une recombinaison peut avoir lieu avec une probabilité r_K. S'il n'y a pas de recombinaison, le nouveau-né est le clône d'un de ses parents (chacun avec probabilité 1/2), sinon il hérite son allèle sous sélection d'un de ses parents et son allèle neutre de son autre parent (là aussi avec la même probabilité), ce qu'on peut résumer ainsi : deux parents de génotypes respectifs $\alpha\beta$ et $\alpha'\beta'$ peuvent avoir un descendant parmi les suivants,

génotype possible	événement	probabilité
$\alpha\beta, \alpha'\beta'$	pas de recombinaison	$1 - r_K$
$\alpha\beta', \alpha'\beta$	recombinaison	r_K

Pour calculer le taux de naissance des individus de type (α, β) il faut donc prendre en compte tous les couples de parents qui peuvent engendrer un descendant avec ce génome. On détaille le calcul de $b^K_{Ab_1}(n)$:

$$\begin{aligned} b^K_{Ab_1}(n) &= f_A n_{Ab_1}[p_{Ab_1} + p_{Ab_2}/2 + p_{ab_1}/2 + (1-r_K)p_{ab_2}/2] + f_A n_{Ab_2}[p_{Ab_1}/2 + r_K p_{ab_1}/2] \\ &\quad + f_a n_{ab_1}[p_{Ab_1}/2 + r_K p_{Ab_2}/2] + f_a n_{ab_2}(1-r_K)p_{Ab_1}/2 \\ &= f_A n_{Ab_1} + r_K f_A f_a (n_{ab_1} n_{Ab_2} - n_{Ab_1} n_{ab_2})/(f_A n_A + f_a n_a). \end{aligned}$$

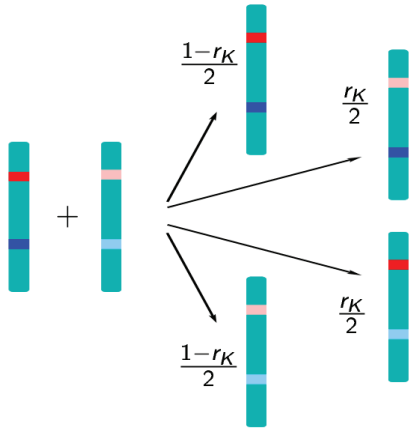

Figure 1.9: Les différents descendants possibles pour un couple de parents donné

Si on note $\bar{\alpha}$ (resp. $\bar{\beta}$) le complémentaire de α dans $\mathscr{A}$ (resp. β dans $\mathscr{B}$), on obtient de manière analogue :

$$b^K_{\alpha\beta}(n) \quad = \quad f_\alpha n_{\alpha\beta} + r_K f_a f_A \frac{n_{\bar{\alpha}\beta} n_{\alpha\bar{\beta}} - n_{\alpha\beta} n_{\bar{\alpha}\bar{\beta}}}{f_A n_A + f_a n_a}, \quad (\alpha,\beta) \in \mathscr{E}. \tag{36}$$

Si l'on se restreint aux populations de types A et a sans tenir compte des allèles neutres, on obtient les taux de naissance et de mort :

$$d^K_\alpha(n) = \sum_{\beta\in\mathscr{B}} d^K_{\alpha\beta}(n) = \left[D_\alpha + C_{\alpha,A}\frac{n_A}{K} + C_{\alpha,a}\frac{n_a}{K}\right] n_\alpha, \quad b^K_\alpha(n) = \sum_{\beta\in\mathscr{B}} b^K_{\alpha\beta}(n) = f_\alpha n_\alpha, \tag{37}$$

qui sont précisément les taux (31) et (32) décrits par Champagnat dans [Cha06]. On peut donc appliquer ses résultats sur la dynamique d'un balayage sélectif.

Comme nous allons le montrer, les balayages doux et durs laissent des signatures très différentes sur la diversité génétique. L'étude de cette signature peut donc permettre de discriminer entre différents scenari d'évolution.

Penchons nous tout d'abord sur le cas du balayage dur dans lequel un mutant apparaît et se fixe dans la population. La signature dépend nécessairement de la probabilité de recombinaison. En effet, si on suppose que cette dernière est nulle la fixation de l'allèle a va entraîner avec lui la fixation de l'allèle neutre porté par le mutant initial. *A contrario*, si la probabilité de recombinaison est élevée, cela va séparer les deux allèles du mutant initial chez de nombreux individus, et la signature laissée par le sweep sera plus faible. Reste à déterminer ce que l'on entend par "probabilité élevée" et à quantifier la variation de la signature en fonction de la probabilité de recombinaison. En fait la réponse à la première question est assez intuitive. On a vu que la durée du balayage sélectif était de l'ordre de $\log K$. On a alors deux possibilités : soit la probabilité de recombinaison est de l'ordre de $1/\log K$. Dans ce cas le nombre de

recombinaisons durant le balayage n'est pas trop élevé et ce dernier laissera une signature détectable ; soit la probabilité de recombinaison est grande devant $1/\log K$. Dans ce cas le nombre de recombinaisons sera suffisamment élevé pour répartir également les allèles neutres dans les populations A et a durant le balayage, et la répartition des allèles neutres dans la population a après la fixation sera la même que la répartition dans la population A au moment de l'arrivée de la mutation. On distingue donc deux régimes :

Hypothèse 2. *Régime de forte recombinaison*

$$\lim_{K\to\infty} r_K \log K = \infty.$$

Hypothèse 3. *Régime de faible recombinaison*

$$\limsup_{K\to\infty} r_K \log K < \infty.$$

On note Fix^K l'événement de fixation et T^K_{Fix} le temps de fixation qui correspond à la mort du dernier individu de type A. On définit également

$$P^K_{\alpha,\beta}(t) = \frac{N^K_{\alpha\beta}(t)}{N^K_\alpha(t)}, \quad (\alpha,\beta)\in\mathscr{E}, K\in\mathbb{N}, \tag{38}$$

la proportion d'allèles β dans la population de type α au temps t, avec la convention $0/0=0$, et on rappelle que la probabilité de l'événement de fixation Fix^K converge vers $S_{a,A}/f_a > 0$ lorsque K tend vers l'infini (34). Alors les proportions neutres vérifient les asymptotiques suivantes :

Théorème 10. *On suppose que l'hypothèse 1 est vérifiée et que* $(N^K_A(0), N^K_{ab_1}(0), N^K_{ab_2}(0)) = (\bar{n}_A K, 1, 0)$. *Alors sur l'événement de fixation* Fix^K *et sous l'hypothèse 2 ou 3, la proportion d'allèles* b_1 *au temps de fixation* T^K_{Fix} *converge en probabilité. Plus précisément, sous l'hypothèse 2,*

$$\lim_{K\to\infty} \mathbb{P}\Big(\mathbf{1}_{Fix^K}\Big|P^K_{a,b_1}(T^K_{Fix}) - P^K_{A,b_1}(0)\Big| > \varepsilon\Big) = 0, \quad \forall\varepsilon>0,$$

et sous l'hypothèse 3 ,

$$\lim_{K\to\infty} \mathbb{P}\Big(\mathbf{1}_{Fix^K}\Big|P^K_{a,b_1}(T^K_{Fix}) - \Big[P^K_{A,b_1}(0) + P^K_{A,b_2}(0)\exp\Big(-\frac{f_a r_K \log K}{S_{aA}}\Big)\Big]\Big| > \varepsilon\Big) = 0, \quad \forall\varepsilon>0.$$

Avant de présenter les idées de preuve de ces résultats nous allons introduire une représentation semi-martingale des proportions neutres dans les populations a et A, qui est à la base de notre approche :

Proposition 2. *Soit* (α, K) *dans* $\mathscr{A}\times\mathbb{N}$. *Le processus* $(P^K_{\alpha,b_1}(t), t\geq 0)$ *défini en* (38) *est une semi-martingale et on a la décomposition suivante :*

$$\begin{aligned} P^K_{\alpha,b_1}(t) &= P^K_{\alpha,b_1}(0) + M^K_\alpha(t) + r_K f_A f_a \int_0^t \frac{N^K_{\tilde{\alpha}b_1}(s)N^K_{\alpha b_2}(s) - N^K_{\alpha b_1}(s)N^K_{\tilde{\alpha}b_2}(s)}{(N^K_\alpha(s)+1)(f_A N^K_A(s) + f_a N^K_a(s))}ds \\ &= P^K_{\alpha,b_1}(0) + M^K_\alpha(t) + r_K f_A f_a \int_0^t \frac{N^K_A(s)N^K_a(s)(P^K_{\alpha,b_1}(s) - P^K_{\tilde{\alpha},b_1}(s))}{(N^K_\alpha(s)+1)(f_A N^K_A(s) + f_a N^K_a(s))}ds, \end{aligned} \tag{39}$$

où le processus $(M^K_\alpha(t), t\geq 0)$ *est une martingale bornée sur tout intervalle* $[0,t]$.

En particulier, on remarque dans la deuxième formulation de (39) que lorsque les proportions neutres dans les populations de types a et A sont proches, le terme de dérive est presque nul et P^K_{α,b_1} est proche d'une martingale. C'est l'idée que l'on va développer pour montrer le comportement asymptotique dans le régime fort de recombinaisons ($r_K \log K \to \infty$). Nous introduisons le processus stochastique

$$Y^K(t) = \mathbb{1}_{\{N^K_A(t)\geq 1, N^K_a(t)\geq 1\}}\left(P^K_{\alpha,b_1}(t) - P^K_{\bar{\alpha},b_1}(t)\right)^2 e^{r_K(f_A\wedge f_a)t}, \quad \forall t \geq 0.$$

Ce dernier vérifie alors

$$\mathbb{E}[Y^K(t\wedge T^K_\varepsilon \wedge \tilde{T}^K_\varepsilon)] \leq c e^{r_K(f_A\wedge f_a)t}\left(e^{-r_K(f_A\wedge f_a)t} + \frac{1}{Kr_K} + e^{-\frac{S_{aA}}{2(k_0+1)})t}\right), \quad \forall t \geq 0$$

pour un c fini qui peut être choisi indépendamment de ε si ce dernier est suffisamment petit. Cela permet de montrer qu'en un temps T satisfaisant $1/r_K \ll T \ll \log K$, les proportions neutres s'égalisent dans les populations de types a et A. L'équation (39) permet de montrer que la proportion neutre dans la population A varie peu (comme l'intégrale devient rapidement petite et la variation quadratique de la martingale M^K_a est de l'ordre de $1/K$), ce qui permet de conclure qu'à la fin de la première phase les proportions d'allèles b_1 dans les deux populations sont sensiblement égales à $P^K_{A,b_1}(0)$, c'est-à-dire la proportion initiale dans la population A. Durant la seconde phase du sweep, on peut comparer la dynamique de la population renormalisée par K à celle des solutions du système dynamique :

$$\dot{n}^{(z,K)}_{\alpha\beta} = \left(f_\alpha - (D_\alpha + C_{\alpha,A}n^{(z,K)}_A + C_{\alpha,a}n^{(z,K)}_a)\right)n^{(z,K)}_{\alpha\beta} + r_K f_A f_a \frac{n^{(z,K)}_{\bar{\alpha}\beta}n^{(z,K)}_{\alpha\bar{\beta}} - n^{(z,K)}_{\alpha\beta}n^{(z,K)}_{\bar{\alpha}\bar{\beta}}}{f_A n^{(z,K)}_A + f_a n^{(z,K)}_a}, \quad (\alpha,\beta)\in\mathscr{E}, \quad (40)$$

avec pour condition initiale $n^{(z,K)}_{\alpha\beta}(0) = z$. Mais l'introduction des fonctions

$$p^{(z,K)}_{\alpha,b_1} = n^{(z,K)}_{\alpha b_1}/n^{(z,K)}_\alpha, \quad \alpha\in\mathscr{A}, \quad \text{et} \quad g^{(z,K)} = p^{(z,K)}_{A,b_1} - p^{(z,K)}_{a,b_1},$$

permet d'écrire le système (40) sous la forme équivalente :

$$\begin{cases} \dot{n}^{(z,K)}_\alpha = (f_\alpha - (D_\alpha + C_{\alpha,A}n^{(z,K)}_A + C_{\alpha,a}n^{(z,K)}_a))n^{(z,K)}_\alpha, \quad \alpha\in\mathscr{A} \\ \dot{g}^{(z,K)} = -g^{(z,K)}\left(r_K f_A f_a (n^{(z,K)}_A + n^{(z,K)}_a)/(f_A n^{(z,K)}_A + f_a n^{(z,K)}_a)\right) \\ \dot{p}^{(z,K)}_{a,b_1} = g^{(z,K)}\left(r_K f_A f_a n^{(z,K)}_A/(f_A n^{(z,K)}_A + f_a n^{(z,K)}_a)\right). \end{cases} \quad (41)$$

En particulier, si la condition initiale $g^{(z,K)}(0)$ de (41) est proche de 0 elle le reste sur tout intervalle de temps (comme $\dot{(g^{(z,K)})^2} \leq 0$). Ainsi, les proportions neutres évoluent peu durant la seconde phase.

Enfin, pour montrer que la proportion neutre reste stable dans la population a durant la troisième phase on utilise de nouveau la décomposition martingale établie dans la Proposition 2. Elle entraîne que la proportion neutre dans la population a est proche d'une

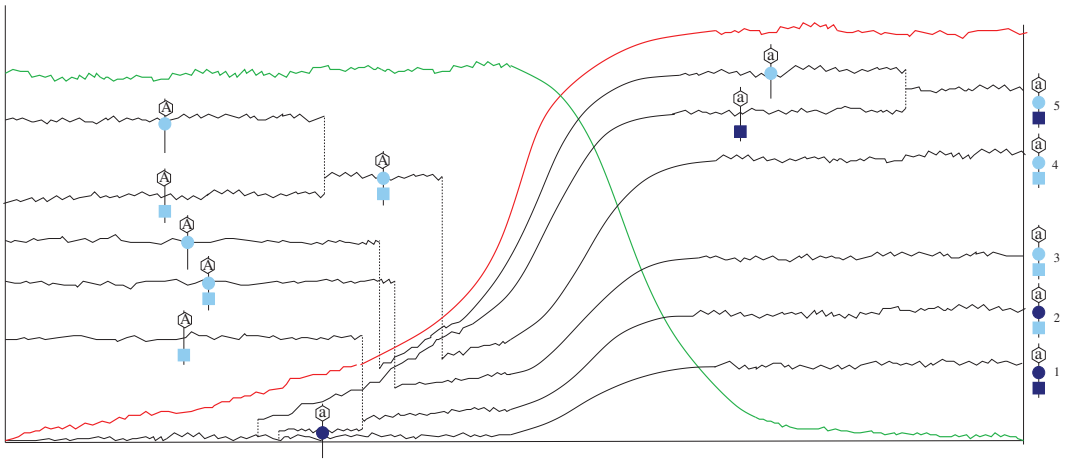

Figure 1.10: Exemple de généalogies de deux loci neutres d'un échantillon de taille 5 : les allèles neutres bleu foncé descendent du mutant et les bleu clair d'un individu de type A. On a indiqué l'allèle sous sélection, A ou a, associé aux allèles neutres durant le balayage sélectif. Il peut changer lors d'une recombinaison. Les courbes en gras représentent les tailles des populations de type A (vert) et a (rouge). Dans cet exemple, les deux allèles neutres de l'individu 1, le premier allèle neutre de l'individu 2 et le second allèle neutre de l'individu 5 descendent du mutant ; les deux allèles neurtes de l'individu 3 descendent du même individu de type A, et les deux allèles neutres de l'individu 4 descendent de deux individus A différents.

martingale dont la variation quadratique est de l'ordre de $1/K$. Comme la durée de la troisième phase est de l'ordre de $\log K$ qui est négligeable devant K, la proportion neutre varie peu.

La preuve du Théorème 10 dans le régime faible de recombinaisons requiert des outils très différents de celle du régime fort. Nous nous sommes inspirés des travaux de Schweinsberg et Durrett [SD05] étudiant une population de taille fixée. L'idée, introduite par Barton [Bar98], consiste à considérer un échantillon d'individus à la fin du sweep sélectif et de suivre leur généalogie neutre en remontant le temps pour déterminer si des recombinaisons ont eu lieu et si l'ancêtre des allèles neutres au moment de l'apparition de la mutation portait l'allèle a (auquel cas l'allèle neutre est nécessairement un b_1) ou l'allèle A (dans ce cas il peut s'agir d'un b_1 ou d'un b_2).

Dans la Figure 1.10 on a représenté un exemple de généalogie dans lequel on suit deux allèles neutres dans le voisinage de l'allèle sous sélection (j'ai fait ce choix de représentation car l'étude de la généalogie de deux allèles neutres liés à l'allèle sous sélection constitue l'étape suivante de mon travail de thèse).

Les recombinaisons ont lieu au moment des naissances. La première étape consiste donc à approcher la loi du nombre de naissances. Plus précisément, on approche la loi du nombre de naissances faisant passer la taille de la population mutante de k à $k+1$ ou de k à k (naissance d'un individu de type A) pour $k < \lfloor \varepsilon K \rfloor$. Pour de tels événements de naissance on peut quantifier précisément la probabilité d'avoir une recombinaison en fonction de k,

K et r_K. La dernière étape de la preuve pour la première phase consiste à approcher les proportions neutres dans la population de type A pour connaître la probabilité de choisir un allèle b_2 lorsqu'il y a une recombinaison avec la population de type A. La décomposition semi-martingale (39) nous permet de montrer que ces proportions restent inchangées durant la première phase. Comme la durée de la seconde phase du sweep est de l'ordre de 1, trop peu de recombinaisons se produisent durant cette phase pour que cela ait un effet sur les proportions neutres dans les populations de types A et a. Enfin, la troisième phase est traitée de la même manière que dans le régime de recombinaison forte.

Comme nous allons le voir maintenant, les balayages doux laissent une signature très différente de celle des balayages forts pour un même taux de recombinaison. En effet les loci neutres dont la distribution allélique était impactée dans le cas d'un balayage fort étaient ceux qui avaient un taux de recombinaison de l'ordre de $1/\log K$ et ce en raison de l'échelle de temps du balayage fort qui a une durée de l'ordre de $\log K$. Dans le cas d'un balayage doux, les allèles positivement sélectionnés sont déjà en grand nombre au début du sweep. Leur fixation va être plus rapide : en un temps qui ne tend pas vers l'infini avec la taille de la population. Ainsi la zone neutre impactée par le balayage sera plus grande.

Dans le cas des balayages doux, comme l'allèle positivement sélectionné constitue une fraction positive de la population au début du sweep, on peut comparer la dynamique de la population aux solutions du système dynamique (33) dès le début et il y a fixation avec grande probabilité :

$$\lim_{K\to\infty} \mathbb{P}(\mathrm{Fix}^K | (N_A^K(0), N_a^K(0)) = (n_A(0)K, n_a(0)K)) = 1, \quad \forall (n_A(0), n_a(0)) \in (\mathbb{R}_+^*)^2.$$

Afin de présenter le résultat on introduit la fonction F, définie pour tout $(z, r, t) \in (\mathbb{R}_+^{\mathscr{E}})^* \times [0,1] \times \mathbb{R}_+$ par

$$F(z,r,t) = \int_0^t \frac{r f_A f_a n_A^{(z)}(s)}{f_A n_A^{(z)}(s) + f_a n_a^{(z)}(s)} \exp\Big(- r f_A f_a \int_0^s \frac{n_A^{(z)}(u) + n_a^{(z)}(u)}{f_A n_A^{(z)}(u) + f_a n_a^{(z)}(u)} du \Big) ds,$$

où $(n_A^{(z)}, n_a^{(z)})$ est la solution du système dynamique (33). On peut montrer que $F(z,r,t)$ converge (uniformément en r) lorsque t tend vers l'infini vers une limite dans $[0,1]$. On peut donc définir :

$$F(z,r) := \lim_{t\to\infty} F(z,r,t) \in [0,1].$$

Comme nous l'avons dit précédemment, le balayage doux a une durée qui ne tend pas vers l'infini avec la taille de la population, et qui dépend du système dynamique (41). Les proportions finales dans le cas du balayage sélectif doux vérifient alors :

Théorème 11. *Soit z dans $\mathbb{R}_+^{A\times\mathscr{B}} \times (\mathbb{R}_+^{a\times\mathscr{B}})^*$. On suppose que $N^K(0) = z$ et que l'hypothèse 1 est vérifiée. Alors sur l'événement de fixation Fix^K, la proportion d'allèles b_1 au temps de fixation T_{Fix}^K converge en probabilité :*

$$\lim_{K\to\infty} \mathbb{P}\Big(\mathbf{1}_{Fix^K} \Big| P_{a,b_1}^K(T_{Fix}^K) - \Big[P_{A,b_1}^K(0) F(z, r_K) + P_{a,b_1}^K(0)(1 - F(z, r_K)) \Big] \Big| > \varepsilon \Big) = 0, \quad \forall \varepsilon > 0.$$

La limite est moins explicite que dans le cas du balayage fort, mais comme elle dépend seulement des solutions d'un système de Lotka-Volterra de dimension 2 elle est très facile à calculer numériquement.

La preuve de ce Théorème repose essentiellement sur une comparaison de la population aux solutions du système dynamique (40). En particulier, c'est le système équivalent (41), plus maniable, qui nous permet d'obtenir cette expression.

Résultats du Chapitre 4

Dans le Chapitre 3 nous avons déterminé la variation de fréquence des allèles neutres dans le voisinage d'un allèle positivement sélectionné après sa fixation. Mais les proportions neutres sur deux loci distincts au voisinage de l'allèle sélectionné ne nous donnent pas d'information sur la corrélation de ces proportions neutres, et de nombreux tests statistiques utilisés pour détecter la sélection à partir de données génomiques s'appuient sur ces corrélations. On peut citer notamment le déséquilibre de liaison [SSL06, McV07] ou la statistique D de Tajima [BHK+95].

Dans le Chapitre 4, qui est un travail commun avec Rebekka Brink-Spalink, nous étudions la généalogie de deux allèles neutres dans le voisinage de l'allèle sélectionné dans le cas d'un balayage fort lorsque la recombinaison est faible, c'est-à-dire de l'ordre de $1/\log K$ (on a vu qu'en cas de forte recombinaison le balayage fort ne laissait pas de signature sur la diversité neutre). Nous supposons comme précédemment que toute la population est de type A jusqu'à l'arrivée d'un mutant a positivement sélectionné. Nous considérons un locus sous sélection SL et deux loci neutres, $N1$ et $N2$. Nous prenons les mêmes notations que précédemment pour SL et $N1$, c'est-à-dire qu'ils portent des allèles dans $\mathscr{A}$ et $\mathscr{B}$. Les individus portent l'allèle $\gamma \in \mathscr{C} := \{c_1, c_2\}$ au locus neutre $N2$. On considère les deux géométries possibles : une géometrie dans laquelle les allèles neutres sont adjacents, $SL-N1-N2$, et une géométrie dans laquelle ils sont séparés par le locus sous sélection, $N1-SL-N2$. On note $r_{1,K}$ la probabilité de recombinaison entre SL et $N1$, et $r_{2,K}$ la probabilité de recombinaison entre $N1$ et $N2$ ou entre SL et $N2$ à chaque naissance et on suppose que les recombinaisons entre les deux paires de loci se produisent indépendamment. Ainsi, des parents de génotypes respectifs $\alpha_1\beta_1\gamma_1$ et $\alpha_2\beta_2\gamma_2$ peuvent donner naissance à un descendant d'un des types suivants dans la géométrie $SL-N1-N2$:

génotype possible	événement	probabilité
$\alpha\beta\gamma, \alpha'\beta'\gamma'$	pas de recombinaison	$(1-r_{1,K})\cdot(1-r_{2,K})$
$\alpha\beta'\gamma', \alpha'\beta\gamma$	une recombinaison entre SL et N1	$r_{1,K}\cdot(1-r_{2,K})$
$\alpha\beta\gamma', \alpha'\beta'\gamma$	une recombinaison entre N1 et N2	$(1-r_{1,K})\cdot r_{2,K}$
$\alpha\beta'\gamma, \alpha'\beta\gamma'$	deux recombinaisons	$r_{1,K}\cdot r_{2,K}$

Enfin, on se place en régime de recombinaisons faibles :

Hypothèse 4. *Les probabilités de recombinaison vérifient :*

$$\limsup_{K\to\infty} r_{j,K}\log K<\infty,\ j=1,2.$$

Ce travail constitue donc une généralisation du Théorème 10 dans le cas de la recombinaison faible.

Soit un entier d. On cherche à quantifier l'effet d'un balayage sélectif sur la diversité neutre. Notre méthode consiste à remonter les généalogies neutres de d individus échantillonnés à la fin du balayage, jusqu'au temps d'apparition de la mutation Deux types d'événements peuvent modifier les relations ancestrales des allèles échantillonnés : les coalescences correspondent à une fusion des généalogies neutres de deux individus à un ou plusieurs loci, et les recombinaisons redistribuent les allèles d'un individu en deux groupes portés par ses deux parents. Nous allons représenter les généalogies neutres à l'aide d'une partition Θ_d^K appartenant à l'ensemble $\mathscr{P}_d^*$ des partitions de $\{(i,k), i\in\{1,...,d\}, k\in\{1,2\}\}$ avec au plus un block distingué par la marque $*$. Dans cette notation $(i,1)$ (resp $(i,2)$) est l'allèle neutre au locus $N1$ (resp $N2$) du i-ème individu échantillonné. Θ_d^K est rigoureusement définie de la manière suivante :

Définition 5. *Soit d un entier. On échantillonne d individus uniformément et sans remplacement à la fin du balayage (temps T_{ext}^K). On suit les généalogies des premier et second allèles neutres du i-ème individu échantillonné, $(i,1)$ et $(i,2)$ pour $i\in\{1,...,d\}$. Il y a au plus $2d$ ancêtres pour ces $2d$ allèles. Alors la partition $\Theta_d^K\in\mathscr{P}_d^*$ est définie de la manière suivante : un bloc de la partition Θ_d^K est constitué des allèles neutres qui descendent d'un individu donné vivant au début du sweep ; le bloc qui contient les descendants du mutant a, s'il y en a un, est distingué par la marque $*$.*

Nous allons montrer dans les Théorèmes 1 et 2 qu'à la limite lorsque K est grand, la partition Θ_d^K appartient à un sous-ensemble Δ_d de $\mathscr{P}_d^*$,

Définition 6. *Soit d un entier. Δ_d est le sous-ensemble de $\mathscr{P}_d^*$ dont les blocs non marqués, s'il y en a, sont soit des singletons, soit des paires de la forme $\{(i,1),(i,2)\}$ pour un $i\in\{1,...,d\}$.*

Un échantillon de taille d dont la loi est représentée par une partition dans Δ_d vérifie la propriété suivante : deux allèles neutres de deux individus distincts ne peuvent pas descendre du même individu de type A.

Exemple 1. *Dans l'exemple de la Figure 1.10, la partition marquée $\pi^{(ex)}$ appartient à Δ_d :*

$$\pi^{(ex)}=\Big\{\{(1,1),(1,2),(2,1),(5,2)\}^*,\{(2,2)\},\{(3,1),(3,2)\},\{(4,1)\},\{(4,2)\},\{(5,1)\}\Big\}.$$

On définit, pour une partition $\pi\in\mathscr{P}_d^*$, le nombre d'individus avec certaines généalogies :

Définition 7. *Soient d un entier et π une partition marquée dans $\mathscr{P}_d^*$. Alors :*

$\pi_1=\#\{1\le i\le d$ *tel que* $(i,1)$ *et* $(i,2)$ *appartient au bloc marqué*$\}$

$\pi_2=\#\{1\le i\le d$ *tel que* $(i,1)$ *appartient au bloc marqué et* $(i,2)$ *est un bloc non marqué*$\}$

$\pi_3 = \#\{1 \le i \le d$ *tel que* $(i,2)$ *appartient au bloc marqué et* $(i,1)$ *est un bloc non marqué*$\}$

$\pi_4 = \#\{1 \le i \le d$ *tel que* $\{(i,1),(i,2)\}$ *est un bloc non marqué*$\}$

$\pi_5 = \#\{1 \le i \le d$ *tel que* $(i,1)$ *et* $(i,2)$ *sont deux blocs non marqués distincts*$\}$

Pour exprimer la distribution limite de la partition Δ_d on introduit :

$$q_1 := \exp\Big(-\frac{f_a r_1 \log K}{S_{aA}}\Big), \quad q_2 := \exp\Big(-\frac{f_a r_2 \log K}{S_{aA}}\Big),$$
$$\bar{q}_2 := \exp\Big(-\frac{f_a r_2 \log K}{|S_{Aa}|}\Big), \quad q_3 := \frac{f_a r_1}{f_a(r_1+r_2) - f_A r_2}\Big(q_2^{f_A/f_a} - q_1 q_2\Big). \tag{42}$$

Nous n'avons fait aucune hypothèse sur le signe de $f_a(r_1+r_2) - f_A r_2$, mais q_3 peut s'écrire sous la forme $\delta(e^{-\mu} - e^{-\nu})/(\nu - \mu)$ pour $\delta, \mu, \nu \in \mathbb{R}_+$ et est donc bien défini et positif. On vérifie facilement que $q_3 < 1$. Ces quatre réels nous permette d'introduire l'ensemble $(p_k, 1 \le k \le 5)$ qui va permettre de décrire la loi de la partition limite dans la Théorème 1 :

$$p_1 := q_1 q_2 [1 - (1-q_1)(1-\bar{q}_2)], \quad p_2 := q_1[(1-q_1 q_2) - q_2 \bar{q}_2 (1-q_1)],$$
$$p_3 := q_1 q_2 (1-\bar{q}_2)(1-q_1), \quad p_4 := \bar{q}_2 q_3, \quad p_5 := (1-q_1)(1-q_1 q_2(1-\bar{q}_2)) - \bar{q}_2 q_3. \tag{43}$$

On remarque que $\sum_{1 \le k \le 5} p_k = 1$. Finalement, on introduit une hypothèse qui résume des différentes hypothèses faites dans la Chapitre 4 :

Hypothèse 5. $(N_A^K(0), N_a^K(0)) = (\bar{n}_A K, 1)$ *et les Hypothèses 1 et 4 sont vérifiées.*

On rappelle les Définitions 1, 2 et 7. On peut maintenant énoncer notre principal résultat sur les relations ancestrales des allèles neutres échantillonnés à la fin du balayage :

Théorème 12 (Généalogie d'un échantillon, géométrie $SL-N1-N2$). *Soit un entier d. Alors sous l'Hypothèse 1, on a pour tout $\pi \in \mathscr{P}_d^*$*

$$\lim_{K\to\infty} \Big| \mathbb{P}(\Theta_d^K = \pi | Fix^K) - \mathbf{1}_{\{\pi \in \Delta_d\}} {p_1}^{\pi_1} {p_2}^{\pi_2} {p_3}^{\pi_3} {p_4}^{\pi_4} {p_5}^{\pi_5} \Big| = 0.$$

On remarque que lorsque K est grand, Θ_d^K appartient à Δ_d avec une probabilité proche de 1, et que $({p_1}^{|\pi|_1} {p_2}^{|\pi|_2} {p_3}^{|\pi|_3} {p_4}^{|\pi|_4} {p_5}^{|\pi|_5}, \pi \in \Delta_d)$ est une probabilité sur Δ_d. La première implication de ce résultat est l'indépendance asymptotique des généalogies neutres des individus échantillonnés. Les allèles neutres d'un individu échantillonné i soit descendent du mutant initial et sont dans le bloc marqué, soit échappent au balayage sélectif et descendent d'un individu A. Dans ce cas ils sont dans un bloc non marqué de la forme $\{(i,1)\}$, $\{(i,2)\}$ ou $\{(i,1),(i,2)\}$ d'après la Définition 7. En conséquence, le Théorème 12 implique que si des allèles neutres de deux individus distincts échappent au sweep, ils descendent de deux individus de type A distincts. Cependant, les généalogies des deux allèles neutres d'un individu échantillonné ne sont pas indépendantes. Par exemple la probabilité que $(i,1)$ et $(i,2)$ échappent au balayage est $p_4 + p_5$; la probabilité que $(i,1)$ (resp. $(i,2)$) échappe au balayage est $p_3 + p_4 + p_5$ (resp. $p_2 + p_4 + p_5$), et pour tout $K \in \mathbb{N}$ tel que $r_1 \neq 0$

$$(p_3+p_4+p_5)(p_2+p_4+p_5) = (1-q_1)(1-q_1 q_2) < (1-q_1)(1-q_1 q_2 + q_1 q_2 \bar{q}_2) = p_4 + p_5.$$

Cela vient du fait que si une recombinaison a d'abord lieu entre SL et $N1$, l'allèle neutre en $N2$, lié à $N1$, échappe aussi au balayage. Comme le terme $q_1 q_2 \bar{q}_2$ ne tend pas vers 0 quand K tend vers l'infini sous l'Hypothèse (4), la seule possibilité d'avoir une égalité à la limite est le cas $r_1 \ll 1/\log K$, ou, en d'autres termes, lorsque la probabilité de voir une recombinaison entre SL et $N1$ est négligeable.

On considère maintenant la seconde géometrie, $N1 - SL - N2$:

Théorème 13 (Généalogie d'un échantillon, géométrie $N1 - SL - N2$). *Soit d un entier. Alors sous l'Hypothèse 1, on a pour tout $\pi \in \mathscr{P}_d^*$*

$$\lim_{K\to\infty} \left| \mathbb{P}(\Theta_d^K = \pi | \mathit{Fix}^K) - \mathbf{1}_{\{\pi\in\Delta_d\}} [q_1 q_2]^{\pi_1} [q_1(1-q_2)]^{\pi_2} [(1-q_1)q_2]^{\pi_3} [(1-q_1)(1-q_2)]^{\pi_4} \right| = 0.$$

De nouveau, cela implique que les généalogies neutres des d individus échantillonnés sont asymptotiquement indépendantes. On a de plus indépendance entre les loci neutres dans la seconde géometrie. En effet, le résultat du Théorème 13 établit que chaque allèle neutre situé au locus Nk échappe au sweep indépendamment des autres allèles neutres (dont celui situé sur l'autre locus neutre du même individu) avec probabilité $1 - q_k$ et dans ce cas est le seul descendant d'un individu de type A vivant au début du sweep. Cela vient du fait que dans la deuxième géométrie, une recombinaison entre SL et un locus neutre n'affecte pas l'environnement génétique de l'allèle situé sur l'autre locus neutre. On remarque qu'il n'y a pas de bloc de la forme $\{(i,1),(i,2)\}$ dans la partiton limite, comme les deux allèles neutres d'un individu ont une très faible probabilité de recombiner lors du même événement de naissance.

Les preuves des Théorèmes 12 et 13 suivent les mêmes idées que la preuve du régime faible de recombinaisons dans le cas du sweep fort (Théorème 10). Comme on s'intéresse maintenant à deux loci neutres on a davantage de scenari possibles pour les généalogies neutres et on a besoin de quantifier plus précisément la distribution du nombre de naissances. Par exemple, les deux allèles neutres peuvent se séparer au sein de la population de type A après une première recombinaison (toujours en regardant le temps dans le sens rétrograde) entre SL et $N1$ qui les a liés simultanément au même allèle A. Pour approcher la probabilité de ce type d'événement, il faut quantifier précisément la loi du nombre de naissances dans la population A, ce qui n'était pas nécessaire lorsqu'on ne s'intéressait qu'à un locus neutre. Le traitement de la troisième phase est également plus précis que dans le Chapitre 3, dans lequel on s'intéresse seulement aux proportions neutres et on ne considère pas la généalogie.

Le déséquilibre de liaison est l'écart à l'association uniforme des allèles. Dans notre cas, c'est la valeur du déséquilibre de liaison entre les deux loci neutres à la fin du sweep qui va apporter de l'information. Il peut s'exprimer de la manière suivante :

$$D(\alpha,t) := \left| \frac{N^K_{\alpha b_i c_j}(t)}{N^K_\alpha(t)} - \frac{(N^K_{\alpha b_i c_1} + N^K_{\alpha b_i c_2})(t)}{N^K_\alpha(t)} \frac{(N^K_{\alpha b_1 c_j} + N^K_{\alpha b_2 c_j})(t)}{N^K_\alpha(t)} \right|, \quad (i,j) \in \{1,2\}^2,$$

où $N^K_{\alpha\beta\gamma}$ est le nombre d'individus portant les allèles (α,β,γ). En d'autres termes, c'est la valeur absolue de la différence entre la fréquence des individus portant les allèles b_i et c_j et

le produit des fréquences des individus portant l'allèle b_i et des individus portant l'allèle c_j. On peut vérifier aisément que la valeur du déséquilibre de liaison ne dépend pas du choix de $(i,j) \in \{1,2\}^2$.

Des études théoriques et expérimentales ont montré que le déséquilibre de liaison constitue une signature importante de la sélection [Asm86, PMH01, KS02, Prz02, SRH+02, WFF+02]. Au début, les biologistes se concentraient sur le déséquilibre de liaison entre le locus sous sélection et un locus voisin. L'effet attendu du balayage sélectif était alors d'accroître le niveau de déséquilibre de liaison, puisque la variation génétique était réduite. Mais comme l'a noté Gillepsie [Gil97], puis [KN04, SSL06, McV07], "linked selection can reduce variation without building up high levels of linkage disequilibrium, contrary to our intuition". Nous montrons également ce phénomène dans notre modèle. Plus précisément, il s'agit de distinguer les loci neutres séparés par le locus sous sélection, et les loci neutres qui se trouvent du même côté du locus sous sélection. Supposons que le déséquilibre de liaison entre les loci neutres était nul au début du sweep ($D(A,0) = 0$, c'est une hypothèse classique, faite par exemple dans [TK87]), et rappelons (42). On note $D^{(ga)}$ (et $D^{(gs)}$) les déséquilibres de liaison pour la géometrie $SL-N1-N2$ (et $N1-SL-N2$). Alors nous montrons que le déséquilibre de liaison prend les formes suivantes dans les deux géométries :

Proposition 5. *On suppose que $D(A,0)^{(ga)} = D(A,0)^{(gs)} = 0$ et que le premier mutant est de type (a,β,γ). On note u_β et u_γ les proportions initiales d'allèles β et γ dans la population A. Alors sous l'Hypothèse 1*

$$\lim_{K\to\infty} \left| \mathbb{E}[D^{(ga)}(a, T^K_{ext}) | \mathrm{Fix}^K] - (1-u_\beta)(1-u_\gamma)(1-q_1)q_1 q_2 \bar{q}_2 \right| = 0,$$

$$\lim_{K\to\infty} \mathbb{E}[D^{(gs)}(a, T^K_{ext}) | \mathrm{Fix}^K] = 0.$$

Ainsi, le déséquilibre de liaison entre deux loci neutres est très dépendant de la position relative du locus sous sélection (entre les deux loci neutres ou adjacent à ces derniers, plus ou moins proche de l'un d'entre eux,...). En utilisant cette propriété, on peut construire des tests statistiques pour distinguer les loci qui ont connu un événement récent de sélection et la force de sélection durant cet événement. On peut également donner une expression du déséquilibre de liaison lorsque sa valeur initiale est différente de zéro, mais le résultat est bien plus complexe dans ce cas.

1.3 Perspectives

Une question pertinente dans la suite du Chapitre 2 est l'étude de distributions quasi-stationnaires pour les CSBP avec catastrophes. La motivation biologique de l'étude de tels objets est de se demander si, sur une échelle de temps plus courte que l'extinction on observe une forme de stationnarité parmi les individus encore vivants. Cette question vient du constat que toutes les populations sont vouées à l'extinction et que ce que l'on observe est une population qui va s'éteindre mais qui n'est pas encore éteinte. On pourra consulter [MV12] pour une review récente sur les distributions quasi-stationnaires.

Un autre point particulièrement intéressant à étudier est la nature des environnements sur l'événement de non-extinction : si l'on conditionne à la non extinction du processus au temps t, que peut-on dire sur l'environnement dans lequel la population a évolué avant le temps t ? Dans le cas discret [Ban09], deux comportements distincts ont été exhibés : dans les régimes sous-critiques fort et intermédiaire, conditionner à la non-extinction ne change pas le caractère sous-critique de l'environnement, et la survie est expliquée par une reproduction exceptionnelle des individus malgré un environnement défavorable. Dans le cas sous-critique faible en revanche, conditionnellement à la non-extinction au temps n, la suite d'environnements est surcritique. La survie est donc expliquée par une reproduction normale dans un environnement particulièrement bon. Nous souhaiterions également étudier ces questions dans le cas continu.

Dans leur article [BS11], Bansaye et Simatos établissent une condition suffisante pour qu'une suite de processus de Galton-Watson en environnement fluctuant admette une limite. Cette condition est assez générale. En particulier elle inclut des cas où la loi de naissance a une variance infinie, des processus limites faisant apparaître des goulots d'étranglement, ou des cas où le processus peut exploser en temps fini. Ce travail montre l'existence d'une grande classe de processus en environnement fluctuant et motive l'étude de CSBP en environnement aléatoire (survie, taux d'extinction, distributions quasi-stationnaires, environnements permettant la survie, ...) plus généraux que les CSBP avec catastrophes auxquels nous nous sommes intéressés.

Dans la suite des travaux des Chapitres 3 et 4, une première idée consiste à généraliser l'étude de la signature laissée par un balayage fort à un nombre quelconque $d \in \mathbb{N}$ de loci neutres. Cependant il paraît malaisé d'étendre les techniques de preuves developpées, qui consistent à évaluer la probabilité de chaque scenario de généalogie des allèleles neutres. Un challenge serait de développer de nouvelles techniques de preuves généralisables à un nombre quelconque de loci neutres.

Dans les Chapitres 3 et 4 nous avons fait implicitement une hypothèse de mutations rares, ce qui permet d'étudier le balayage sélectif en supposant qu'il y a seulement des individus de type A et a dans la population entre la mutation $A \to a$ et la fixation de l'allèle a (ou la disparition de ce dernier s'il ne se fixe pas). Dans [LBD11, LRH+13] les auteurs suivent l'évolution de quarante populations de levures sur 1000 générations. Ils séquencent le génome de ces populations à des intervalles de temps successifs, ce qui leur permet de suivre l'apparition et l'évolution des mutations positivement sélectionnées. En particulier ils remarquent que plusieurs mutations surviennent et évoluent simultanément dans la population, et que plusieurs cohortes sont régulièrement présentes en même temps et sont en compétition les unes avec les autres. Une question particulièrement intéressante serait donc d'étudier l'auto-stop génétique dans le cas où plusieurs mutations favorables sont en compétition. A ma connaissance cette question a été peu étudiée. La seule référence que j'aie trouvée est [CB+08], mais dans ce modèle comme dans les modèles précédents de hitchhiking la population est de taille constante, les fitness s_1 et s_2 des nouveaux mutants ne dépendent pas de l'état de la population et sont supposées petites ($s_1, s_2 \ll 1$), ce qui permet aux auteurs de négliger les termes contenant le

produit des fitness. Comme la dynamique d'un tel processus est particulièrement complexe (au moins trois loci à considérer dont un seul neutre, quatre ensembles de paramètres écologiques suivant les allèles portés par un individu : pas de mutation, une des deux mutations ou les deux mutations, ...), un travail préliminaire consisterait à étudier le phénomène d'interférence clonale dans une population asexuée, et de s'appuyer sur cette dynamique comme nous nous étions appuyés sur la dynamique de l'invasion d'un mutant dans une population haploïde développée par Champagnat [Cha06] dans les Chapitres 3 et 4.

J'ai développé et étudié dans cette thèse des modèles probabilistes de populations soumises à des événements ponctuels, extérieurs dans le cas des processus de branchements continus soumis à des catastrophes et des balayages sélectifs doux, ou internes à la population dans le cas de l'apparition des mutations conduisant à des balayages forts. Je souhaite poursuivre le développement de ce type de modèles pour aider à comprendre l'influence de certains paramètres dans le comportement des populations. Un aspect, absent de cette thèse, que je souhaiterais également développer est l'influence de la dimension spatiale dans ces comportements : évolution en une ou plusieurs dimensions, espace homogène ou inhomogène, espace variant de manière déterministe ou aléatoire, ...

CHAPITRE 2

On the extinction of Continuous State Branching Processes with catastrophes

Introduction

Continuous state branching processes (CSBP) are the analogues of Galton-Watson (GW) processes in continuous time and continuous state space. They have been introduced by Jirina [Jir58] and studied by many authors including Bingham [Bin76], Grey [Gre74], Grimvall [Gri74], Lamperti [Lam67a, Lam67b], to name but a few.

A CSBP $Z = (Z_t, t \geq 0)$ is a strong Markov process taking values in $[0, \infty]$, where 0 and ∞ are absorbing states, and satisfying the branching property. We denote by $(\mathbb{P}_x, x > 0)$ the law of Z starting from x. Lamperti [Lam67b] proved that there is a bijection between CSBP and scaling limits of GW processes. Thus they may model the evolution of renormalized large populations on a large time scale.

The branching property implies that the Laplace transform of Z_t is of the form

$$\mathbb{E}_x\Big[\exp(-\lambda Z_t)\Big] = \exp\{-x u_t(\lambda)\}, \qquad \text{for } \lambda \geq 0,$$

for some non-negative function u_t. According to Silverstein [Sil68], this function is determined by the integral equation

$$\int_{u_t(\lambda)}^{\lambda} \frac{1}{\psi(u)} du = t,$$

where ψ is known as the branching mechanism associated to Z. We assume here that Z has finite mean, so that we have the following classical representation

$$\psi(\lambda) = -g\lambda + \sigma^2\lambda^2 + \int_0^\infty \Big(e^{-\lambda z} - 1 + \lambda z\Big)\mu(\mathrm{d}z), \tag{2.0.1}$$

where $g \in \mathbb{R}$, $\sigma \geq 0$ and μ is a σ-finite measure on $(0, \infty)$ such that $\int_{(0,\infty)} \big(z \wedge z^2\big)\mu(\mathrm{d}z)$ is finite. The CSBP is then characterized by the triplet (g, σ, μ) and can also be defined as the unique non-negative strong solution of a stochastic differential equation. More precisely, from Fu and Li [FL10] we have

$$Z_t = Z_0 + \int_0^t g Z_s \mathrm{d}s + \int_0^t \sqrt{2\sigma^2 Z_s}\mathrm{d}B_s + \int_0^t \int_0^\infty \int_0^{Z_{s-}} z\tilde{N}_0(\mathrm{d}s, \mathrm{d}z, \mathrm{d}u), \tag{2.0.2}$$

where B is a standard Brownian motion, $N_0(\mathrm{d}s, \mathrm{d}z, \mathrm{d}u)$ is a Poisson random measure with intensity $\mathrm{d}s\mu(\mathrm{d}z)\mathrm{d}u$ independent of B, and $\widetilde{N}_0$ is the compensated measure of N_0.

The stable case with drift, i.e. $\psi(\lambda) = -g\lambda + c\lambda^{1+\beta}$, with β in $(0,1]$, corresponds to the CSBP that one can obtain by scaling limits of GW processes with a fixed reproduction law. It is of special interest in this paper since the Laplace exponent can be computed explicitly and it can also be used to derive asymptotic results for more general cases.

In this work, we are interested in modeling catastrophes which occur at random and kill each individual with some probability (depending on the catastrophe). In terms of the CSBP representing the scaling limit of the size of a large population, this amounts to letting the process make a negative jump, i.e. multiplying its current value by a random fraction. The process that we obtain is still Markovian whenever the catastrophes follow a time homogeneous Poisson Point Process. Moreover, we show that conditionally on the times and the effects of the catastrophes, the process satisfies the branching property. Thus, it yields a particular class of CSBP in random environment, which can also be obtained as the scaling limit of GW processes in random environment (see [BS11]). Such processes are motivated in particular by a cell division model; see for instance [BT11] and Section 2.4.

We also consider positive jumps that may represent immigration events proportional to the size of the current population. Our motivation comes from the aggregation behavior of some species. We refer to Chapter 12 in [DGC08] for adaptive explanations of these aggregation behaviors, or [RLF+13] which shows that aggregation behaviors may result from manipulation by parasites to increase their transmission. For convenience, we still call these dramatic events catastrophes.

The process Y that we consider in this paper is then called *a CSBP with catastrophes*. Roughly speaking, it can be defined as follows : The process Y follows the SDE (2.0.2) between catastrophes, which are then given in terms of the jumps of a Lévy process with bounded variation paths. Thus the set of times at which catastrophes occur may have accumulation points, but the mean effect of the catastrophes has a finite first moment. When a catastrophe with effect m_t occurs at time t, we have

$$Y_t = m_t Y_{t-}.$$

We defer the formal definitions to Section 2.1. We also note that Brockwell has considered birth and death branching processes with another kind of catastrophes, see e.g. [Bro85].

First we verify that CSBP with catastrophes are well defined as solutions of a certain stochastic differential equation, which we give as (2.1.3). We characterize their Laplace exponents via an ordinary differential equation (see Theorem 1), which allows us to describe their long time behavior. In particular, we prove an extinction criterion for the CSBP with catastrophes which is given in terms of the sign of $\mathbb{E}[g + \sum_{s\leq 1} \log m_s]$. We also establish a central limit theorem conditionally on survival and under some moment assumptions (Corollary 2.2).

We then focus on the case when the branching mechanism associated to the CSBP with catastrophes Y has the form $\psi(\lambda) = -g\lambda + c\lambda^{1+\beta}$, for $\beta \in (0,1]$, i.e. the stable case. In this scenario, the extinction and absorption events coincide, which means that $\{\lim_{t\to\infty} Y_t = 0\} = \{\exists t \geq 0, Y_t = 0\}$. We prove that the speed of extinction is directly related to the asymptotic

behavior of exponential functionals of Lévy processes (see proposition 3). More precisely, we show that the extinction probability of a stable CSBP with catastrophes can be expressed as follows :

$$\mathbb{P}(Y_t > 0) = \mathbb{E}\left[F\left(\int_0^t e^{-\beta K_s} \mathrm{d}s\right)\right],$$

where F is a function with a particular asymptotic behavior and $K_t := gt + \sum_{s \leq t} \log m_s$ is a Lévy process of bounded variation that does not drift to $+\infty$ and satisfies an exponential positive moment condition. We establish the asymptotic behavior of the survival probability (see Theorem 2) and find four different regimes when this probability is equal to zero. Actually, such asymptotic behaviors have previously been found for branching processes in random environments in discrete time and space (see e.g. [GL01, GKV03, AGKV05]). Here, the regimes depend on the shape of the Laplace exponent of K, i.e. on the drift g of the CSBP and the law of the catastrophes. The asymptotic behavior of exponential functionals of Lévy processes drifting to $+\infty$ has been deeply studied by many authors, see for instance Bertoin and Yor [BY05] and references therein. To our knowledge, the remaining cases have been studied only by Carmona et al. (see Lemma 4.7 in [CPY97]) but their result focuses only on one regime. Our result is closely related to the discrete framework via the asymptotic behaviors of functionals of random walks. More precisely, we use in our arguments local limit theorems for semi direct products [LPP97, GL01] and some analytical results on random walks [Koz76, Hir98], see Section 2.3.

From the speed of extinction in the stable case, we can deduce the speed of extinction of a larger class of CSBP with catastrophes satisfying the condition that extinction and absorption coincide (see Corollary 2.3). General results for the case of Lévy processes of unbounded variation do not seem easy to obtain since the existence of the process Y and our approximation methods are not so easy to deduce. The particular case when $\mu = 0$ and the environment K is given by a Brownian motion has been studied in [BH12]. The authors in [BH12] also obtained similar asymptotics regimes using the explicit law of $\int_0^t \exp(-\beta K_s) ds$.

Finally, we apply our results to a cell infection model introduced in [BT11] (see Section 2.4). In this model, the infection in a cell line is given by a Feller diffusion with catastrophes. We derive here the different possible speeds of the infection propagation. More generally, these results can be related to some ecological problems concerning the role of environmental and demographical stochasticities. Such topics are fundamental in conservation biology, as discussed for instance in Chapter 1 in [LES03]. Indeed, the survival of the population may be either due to the randomness of the individual reproduction, which is specified in our model by the parameters σ and μ of the CSBP, or to the randomness (rate, size) of the catastrophes due to the environment. For a study of relative effects of environmental and demographical stochasticities, the reader is referred to [Lan93] and references therein.

The remainder of the paper is structured as follows. In Section 2, we define and study the CSBP with catastrophes. Section 2.2 is devoted to the study of the extinction probabilities where special attention is given to the stable case. In Section 2.3, we analyse the asymptotic behavior of exponential functionals of Lévy processes of bounded variation. This result is

the key to deducing the different extinction regimes. In Section 2.4, we apply our results to a cell infection model. Finally, Section A contains some technical results used in the proofs and deferred for the convenience of the reader.

2.1 CSBP with catastrophes

We consider a CSBP $Z = (Z_t, t \geq 0)$ defined by (2.0.2) and characterized by the triplet (g, σ, μ), where we recall that μ satisfies

$$\int_0^\infty (z \wedge z^2)\mu(\mathrm{d}z) < \infty. \tag{2.1.1}$$

The catastrophes are independent of the process Z and are given by a Poisson random measure $N_1 = \sum_{i\in I} \delta_{t_i, m_{t_i}}$ on $[0,\infty) \times [0,\infty)$ with intensity $\mathrm{d}t\nu(\mathrm{d}m)$ such that

$$\nu(\{0\}) = 0 \qquad \text{and} \qquad 0 < \int_{(0,\infty)} (1 \wedge |m-1|)\nu(\mathrm{d}m) < \infty. \tag{2.1.2}$$

The jump process

$$\Delta_t = \int_0^t \int_{(0,\infty)} \log(m) N_1(\mathrm{d}s, \mathrm{d}m) = \sum_{s\leq t} \log(m_s),$$

is thus a Lévy process with paths of bounded variation, which is non identically zero.

The CSBP (g, σ, μ) with catastrophes ν is defined as the solution of the following stochastic differential equation :

$$\begin{aligned} Y_t = Y_0 + \int_0^t g Y_s \mathrm{d}s + \int_0^t \sqrt{2\sigma^2 Y_s}\mathrm{d}B_s \quad &+ \quad \int_0^t \int_{[0,\infty)} \int_0^{Y_{s-}} z\tilde{N}_0(\mathrm{d}s, \mathrm{d}z, \mathrm{d}u) \\ &+ \quad \int_0^t \int_{[0,\infty)} \Big(m-1\Big) Y_{s-} N_1(\mathrm{d}s, \mathrm{d}m), \end{aligned} \tag{2.1.3}$$

where $Y_0 > 0$ a.s.

Let $\mathscr{BV}(\mathbb{R}_+)$ be the set of càdlàg functions on $\mathbb{R}_+ := [0,\infty)$ of bounded variation and C_b^2 the set of all functions that are twice differentiable and are bounded together with their derivatives, then the following result of existence and unicity holds :

Theorem 1. *The stochastic differential equation (2.1.3) has a unique non-negative strong solution Y for any $g \in \mathbb{R}, \sigma \geq 0$, μ and ν satisfying conditions (2.1.1) and (2.1.2), respectively. Then, the process $Y = (Y_t, t \geq 0)$ is a càdlàg Markov process satisfying the branching property conditionally on $\Delta = (\Delta_t, t \geq 0)$ and its infinitesimal generator $\mathscr{A}$ satisfies for every $f \in C_b^2$*

$$\begin{aligned} \mathscr{A}f(x) = gxf'(x) + \sigma^2 x f''(x) + \int_0^\infty \Big(f(mx) - f(x)\Big)\nu(dm) \\ + \int_0^\infty \Big(f(x+z) - f(x) - zf'(x)\Big)x\mu(dz). \end{aligned} \tag{2.1.4}$$

Moreover, for every $t \geq 0$,

$$\mathbb{E}_y\left[\exp\left\{-\lambda \exp\{-gt-\Delta_t\}Y_t\right\}\middle|\ \Delta\right] = \exp\left\{-y v_t(0,\lambda,\Delta)\right\} \qquad a.s.,$$

where for every $(\lambda,\delta) \in (\mathbb{R}_+, \mathscr{BV}(\mathbb{R}_+))$, $v_t : s \in [0,t] \mapsto v_t(s,\lambda,\delta)$ *is the unique solution of the following backward differential equation :*

$$\frac{\partial}{\partial s} v_t(s,\lambda,\delta) = e^{gs+\delta_s}\psi_0\big(e^{-gs-\delta_s} v_t(s,\lambda,\delta)\big), \qquad v_t(t,\lambda,\delta) = \lambda, \tag{2.1.5}$$

and

$$\psi_0(\lambda) = \psi(\lambda) - \lambda\psi'(0) = \sigma^2\lambda^2 + \int_0^\infty (e^{-\lambda z} - 1 + \lambda z)\mu(dz). \tag{2.1.6}$$

Démonstration. Under Lipschitz conditions, the existence and uniqueness of strong solutions for stochastic differential equations are classical results (see [IW89]). In our case, the result follows from proposition 2.2 and Theorems 3.2 and 5.1 in [FL10]. By Itô's formula (see for instance [IW89] Th.5.1), the solution of the SDE (2.1.3), $(Y_t, t \geq 0)$ solves the following martingale problem. For every $f \in C_b^2$,

$$\begin{aligned} f(Y_t) = f(Y_0) + \text{ loc. mart. } &+ g\int_0^t f'(Y_s)Y_s \mathrm{d}s \\ &+ \sigma^2 \int_0^t f''(Y_s)Y_s \mathrm{d}s + \int_0^t \int_0^\infty Y_s\Big(f(Y_s+z) - f(Y_s) - f'(Y_s)z\Big)\mu(\mathrm{d}z)\mathrm{d}s \\ &+ \int_0^t \int_0^\infty \Big(f(mY_s) - f(Y_s)\Big)\nu(\mathrm{d}m)\mathrm{d}s, \end{aligned}$$

where the local martingale is given by

$$\begin{aligned} &\int_0^t f'(Y_s)\sqrt{2\sigma^2 Y_s}\mathrm{d}B_s + \int_0^t \int_0^\infty \Big(f(mY_{s^-}) - f(Y_{s^-})\Big)\widetilde{N}_1(\mathrm{d}s,\mathrm{d}m) \\ &\qquad + \int_0^t \int_0^\infty \int_0^{Y_{s-}} \Big(f(Y_{s-}+z) - f(Y_{s-})\Big)\widetilde{N}_0(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u), \end{aligned} \tag{2.1.7}$$

and $\widetilde{N}_1$ is the compensated measure of N_1. Even though the process in (2.1.7) is a local martingale, we can define a localized version of the corresponding martingale problem as in Chapter 4.6 of Ethier and Kurtz [EK86]. We leave the details to the reader. From pathwise uniqueness, we deduce that the solution of (2.1.3) is a strong Markov process whose generator is given by (2.1.4).

The branching property of Y, conditionally on Δ, is inherited from the branching property of the CSBP and the fact that the additional jumps are multiplicative.

To prove the second part of the theorem, let us now work conditionally on Δ. Applying Itô's formula to the process $\widetilde{Z}_t = Y_t \exp\{-gt - \Delta_t\}$, we obtain

$$\widetilde{Z}_t = Y_0 + \int_0^t e^{-gs-\Delta_s}\sqrt{2\sigma^2 Y_s}\mathrm{d}B_s + \int_0^t \int_0^\infty \int_0^{Y_{s-}} e^{-gs-\Delta_{s-}} z\widetilde{N}_0(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u),$$

and then $\widetilde{Z}$ is a local martingale conditionally on Δ. A new application of Itô's formula ensures that for every $F \in C_b^{1,2}$, $F(t, \widetilde{Z}_t)$ is also a local martingale if and only if for every $t \geq 0$,

$$\int_0^t \frac{\partial^2}{\partial x^2} F(s, \widetilde{Z}_s)\sigma^2 \widetilde{Z}_s e^{-gs-\Delta_s} \mathrm{d}s + \int_0^t \frac{\partial}{\partial s} F(s, \widetilde{Z}_s)\mathrm{d}s \tag{2.1.8}$$
$$+ \int_0^t \int_0^\infty \widetilde{Z}_s \Big(\Big[F(s, \widetilde{Z}_s + z e^{-gs-\Delta_s}) - F(s, \widetilde{Z}_s) \Big] e^{gs+\Delta_s} - \frac{\partial}{\partial x} F(s, \widetilde{Z}_s) z \Big) \mu(\mathrm{d}z)\mathrm{d}s = 0.$$

In the vein of [IW89, BT11], we choose $F(s,x) := \exp\{-x v_t(s,\lambda,\Delta)\}$, where $v_t(s,\lambda,\Delta)$ is differentiable with respect to the variable s, non-negative and such that $v_t(t,\lambda,\Delta) = \lambda$, for $\lambda \geq 0$. The function F is bounded, so that $(F(s,\tilde{Z}_s), 0 \leq s \leq t)$ will be a martingale if and only if for every $s \in [0,t]$

$$\frac{\partial}{\partial s} v_t(s,\lambda,\Delta) = e^{gs+\Delta_s} \psi_0 \left(e^{-gs-\Delta_s} v_t(s,\lambda,\Delta) \right), \quad \text{a.s.},$$

where ψ_0 is defined in (2.1.6).

Proposition 5 in Section 6 ensures that a.s. the solution of this backward differential equation exists and is unique, which essentially comes from the Lipschitz property of ψ_0 (Lemma 8) and the fact that Δ possesses bounded variation paths. Then the process $(\exp\{-\tilde{Z}_s v_t(s,\lambda,\Delta)\}, 0 \leq s \leq t)$ is a martingale conditionally on Δ and

$$\mathbb{E}_y \left[\exp\left\{ -\tilde{Z}_t v_t(t,\lambda,\Delta) \right\} \middle| \Delta \right] = \mathbb{E}_y \left[\exp\left\{ -\tilde{Z}_0 v_t(0,\lambda,\Delta) \right\} \middle| \Delta \right] \quad \text{a.s.},$$

which yields

$$\mathbb{E}_y \left[\exp\left\{ -\lambda \tilde{Z}_t \right\} \middle| \Delta \right] = \exp\left\{ -y v_t(0,\lambda,\Delta) \right\} \quad \text{a.s.} \tag{2.1.9}$$

This implies our result. □

Referring to Theorem 7.2 in [Kyp06], we recall that a Lévy process has three possible asymptotic behaviors : either it drifts to ∞, $-\infty$, or oscillates a.s. In particular, if the Lévy process has a finite first moment, the sign of its expectation yields the regimes of above. We extend this classification to CSBP with catastrophes.

Corollary 2.1. *We have the following three regimes.*

i) If $(\Delta_t + gt)_{t\geq 0}$ *drifts to* $-\infty$, *then* $\mathbb{P}(Y_t \to 0 \mid \Delta) = 1$ *a.s.*

ii) If $(\Delta_t + gt)_{t\geq 0}$ *oscillates, then* $\mathbb{P}(\liminf_{t\to\infty} Y_t = 0 \mid \Delta) = 1$ *a.s.*

iii) If $(\Delta_t + gt)_{t\geq 0}$ *drifts to* $+\infty$ *and there exists* $\varepsilon > 0$, *such that*

$$\int_0^\infty z \log^{1+\varepsilon}(1+z)\mu(dz) < \infty, \tag{2.1.10}$$

then $\mathbb{P}(\liminf_{t\to\infty} Y_t > 0 \mid \Delta) > 0$ *a.s. and there exists a non-negative finite r.v.* W *such that*

$$e^{-gt-\Delta_t} Y_t \xrightarrow[t\to\infty]{} W \quad a.s., \qquad \{W = 0\} = \Big\{ \lim_{t\to\infty} Y_t = 0 \Big\}.$$

Remark 1. *In the regime (ii), Y may be absorbed in finite time a.s. (see the next section). But Y_t may also a.s. do not tend to zero. For example, if $\mu = 0$ and $\sigma = 0$, then $Y_t = \exp(gt + \Delta_t)$ and* $\limsup_{t\to\infty} Y_t = \infty$.

Assumption (iii) of the corollary does not imply that $\{\lim_{t\to\infty} Y_t = 0\} = \{\exists t : Y_t = 0\}$. *Indeed, the case $\mu(dx) = x^{-2}\mathbf{1}_{[0,1]}(x)dx$ inspired by Neveu's CSBP yields $\psi(u) \sim u\log u$ as $u \to \infty$. Then, according to Remark 2.2 in [Lam08],* $\mathbb{P}(\exists t : Y_t = 0) = 0$ *and* $0 < \mathbb{P}(\lim_{t\to\infty} Y_t = 0) < 1$.

Démonstration. We use (2.1.8) with $F(s,x) = x$ to get that $\tilde{Z} = (Y_t \exp(-gt - \Delta_t) : t \geq 0)$ is a non-negative local martingale. Thus it is a non-negative supermartingale and it converges a.s. to a non-negative finite random variable W. This implies the proofs of (i-ii).

In the case when $(gt+\Delta_t, t \geq 0)$ goes to $+\infty$, we prove that $\mathbb{P}(W > 0 \mid \Delta) > 0$ a.s. According to Lemma 9 in Section 6, the assumptions of (iii) ensure the existence of a non-negative increasing function k on $\mathbb{R}^+$ such that for all $\lambda > 0$,

$$\psi_0(\lambda) \leq \lambda k(\lambda) \quad \text{and} \quad c(\Delta) := \int_0^\infty k\Big(e^{-(gt+\Delta_t)}\Big)dt < \infty \quad \text{a.s.}$$

For every $(t,\lambda) \in (\mathbb{R}_+^*)^2$, the solution v_t of (2.1.5) is non-decreasing on $[0,t]$. Thus for all $s \in [0,t]$, $v_t(s,1,\Delta) \leq 1$, and

$$\begin{aligned}\psi_0(e^{-gs-\Delta_s} v_t(s,1,\Delta)) &\leq e^{-gs-\Delta_s} v_t(s,1,\Delta) k(e^{-gs-\Delta_s} v_t(s,1,\Delta)) \\ &\leq e^{-gs-\Delta_s} v_t(s,1,\Delta) k(e^{-gs-\Delta_s}) \qquad \text{a.s.}\end{aligned}$$

Then (2.1.5) gives

$$\frac{\partial}{\partial s} v_t(s,1,\Delta) \leq v_t(s,1,\Delta) k(e^{-gs-\Delta_s}),$$

implying

$$-\ln(v_t(0,1,\Delta)) \leq \int_0^t k(e^{-gs-\Delta_s})ds \leq c(\Delta) < \infty \quad \text{a.s.}$$

Hence, for every $t \geq 0$, $v_t(0,1,\Delta) \geq \exp(-c(\Delta)) > 0$ and conditionally on Δ there exists a positive lower bound for $v_t(0,1,\Delta)$. Finally from (2.1.9),

$$\mathbb{E}_y[\exp\{-W\} \mid \Delta] = \exp\Big\{-y \lim_{t\to\infty} v_t(0,1,\Delta)\Big\} < 1$$

and $\mathbb{P}(W > 0 \mid \Delta) > 0$ a.s.
Moreover, since Y satisfies the branching property conditionally on Δ, we can show (see Lemma 10 in Section 6) that

$$\{W = 0\} = \Big\{\lim_{t\to\infty} Y_t = 0\Big\} \qquad \text{a.s.},$$

which completes the proof. □

We now derive a central limit theorem in the supercritical regime :

Corollary 2.2. *Assume that* $(gt+\Delta_t, t\geq 0)$ *drifts to* $+\infty$ *and (2.1.10) is satisfied. Then, under the additional assumption*

$$\int_{(0,e^{-1}]\cup[e,\infty)} (\log m)^2 \nu(\mathrm{d}m) < \infty, \tag{2.1.11}$$

conditionally on $\{W>0\}$,

$$\frac{\log(Y_t) - \boldsymbol{m}t}{\rho\sqrt{t}} \xrightarrow[t\to\infty]{d} \mathcal{N}(0,1),$$

where $\xrightarrow{d}$ *means convergence in distribution,*

$$m := g + \int_{\{|\log x|\geq 1\}} \log m\, \nu(\mathrm{d}m) < \infty, \qquad \rho^2 := \int_0^\infty (\log m)^2 \nu(\mathrm{d}m) < \infty,$$

and $\mathcal{N}(0,1)$ *denotes a centered Gaussian random variable with variance equals 1.*

Démonstration. We use the central limit theorem for the Lévy process $(gt+\Delta_t, t\geq 0)$ under assumption (2.1.11) of Doney and Maller [DM02], see Theorem 3.5. For simplicity, the details are deferred to Section A. We then get

$$\frac{gt+\Delta_t - \mathbf{m}t}{\rho\sqrt{t}} \xrightarrow[t\to\infty]{d} \mathcal{N}(0,1). \tag{2.1.12}$$

From Corollary 2.1 part iii), under the event $\{W>0\}$, we get

$$\log Y_t - (gt+\Delta_t) \xrightarrow[t\to\infty]{a.s.} \log W \in (-\infty,\infty),$$

and we conclude using (2.1.12). □

2.2 Speed of extinction of CSBP with catastrophes

In this section, we first study the particular case of the stable CSBP with growth $g\in\mathbb{R}$. Then, we derive a similar result for another class of CSBP's.

The stable case

We assume in this section that

$$\psi(\lambda) = -g\lambda + c_+\lambda^{\beta+1}, \tag{2.2.1}$$

for some $\beta\in(0,1]$, $c_+>0$ and g in $\mathbb{R}$.

If $\beta=1$ (i.e. the Feller diffusion), we necessarily have $\mu=0$ and the CSBP Z follows the continuous diffusion

$$Z_t = Z_0 + \int_0^t g Z_s \mathrm{d}s + \int_0^t \sqrt{2\sigma^2 Z_s}\mathrm{d}B_s, \quad t\geq 0.$$

In the case when $\beta \in (0,1)$, we necessarily have $\sigma = 0$ and the measure μ takes the form $\mu(\mathrm{d}x) = c_+(\beta+1)x^{-(2+\beta)}\mathrm{d}x/\Gamma(1-\beta)$. In other words, the process possesses positive jumps with infinite intensity [Lam07]. Moreover,

$$Z_t = Z_0 + \int_0^t gZ_s\mathrm{d}s + \int_0^t Z_{s^-}^{1/(\beta+1)}\mathrm{d}X_s, \quad t \geq 0,$$

where X is a $(\beta+1)$-stable spectrally positive Lévy process.

For the stable CSBP with catastrophes, the backward differential equation (2.1.5) can be solved and in particular, we get

Proposition 3. *For all $x_0 > 0$ and $t \geq 0$:*

$$\mathbb{P}_{x_0}(Y_t > 0 \mid \Delta) = 1 - \exp\left\{-x_0\left(c_+\beta\int_0^t e^{-\beta(gs+\Delta_s)}\mathrm{d}s\right)^{-1/\beta}\right\} \qquad a.s. \tag{2.2.2}$$

Moreover,

$$\mathbb{P}_{x_0}(\textit{there exists } t > 0,\ Y_t = 0 \mid \Delta) = 1 \qquad a.s.,$$

if and only if the process $(gt + \Delta_t, t \geq 0)$ does not drift to $+\infty$.

Démonstration. Since $\psi_0(\lambda) = c_+\lambda^{\beta+1}$, a direct integration gives us

$$v_t(u,\lambda,\Delta) = \left[c_+\beta\int_u^t e^{-\beta(gs+\Delta_s)}\mathrm{d}s + \lambda^{-\beta}\right]^{-1/\beta},$$

which implies

$$\mathbb{E}_{x_0}\left[e^{-\lambda\tilde{Z}_t}\middle|\ \Delta\right] = \exp\left\{-x_0\left(c_+\beta\int_0^t e^{-\beta(gs+\Delta_s)}\mathrm{d}s + \lambda^{-\beta}\right)^{-1/\beta}\right\} \quad \text{a.s.} \tag{2.2.3}$$

Hence, the absorption probability follows by letting λ tend to ∞ in (2.2.3). In other words,

$$\mathbb{P}_{x_0}(Y_t = 0 \mid \Delta) = \exp\left\{-x_0\left(c_+\beta\int_0^t e^{-\beta(gs+\Delta_s)}\mathrm{d}s\right)^{-1/\beta}\right\} \quad \text{a.s.}$$

Since $\mathbb{P}_{x_0}(\text{there exists } t \geq 0 : Y_t = 0 \mid \Delta) = \lim_{t\to\infty}\mathbb{P}_{x_0}(Y_t = 0 \mid \Delta)$ a.s., we deduce

$$\mathbb{P}_{x_0}(\text{there exists } t \geq 0 : Y_t = 0 \mid \Delta) = \exp\left\{-x_0\left(c_+\beta\int_0^\infty e^{-\beta(gs+\Delta_s)}\mathrm{d}s\right)^{-1/\beta}\right\} \quad \text{a.s.}$$

Finally, according to Theorem 1 in [BY05], $\int_0^\infty \exp\{-\beta(gs+\Delta_s)\}\mathrm{d}s = \infty$ a.s. if and only if the process $(gt+\Delta_t, t \geq 0)$ does not drift to $+\infty$. This completes the proof. □

In what follows, we assume that the Lévy process Δ admits some positive exponential moments, i.e. there exists $\lambda > 0$ such that $\phi(\lambda) < \infty$. We can then define $\theta_{max} = \sup\{\lambda > 0, \phi(\lambda) < \infty\} \in (0, \infty]$ and we have

$$\phi(\lambda) := \log \mathbb{E}[e^{\lambda \Delta_1}] = \int_0^\infty (m^\lambda - 1)\nu(dm) < \infty \qquad \text{for } \lambda \in [0, \theta_{max}). \tag{2.2.4}$$

We note that ϕ can be differentiated on the right in 0 and also in 1 if $\theta_{max} > 1$:

$$\phi'(0) := \phi'(0+) = \int_0^\infty \log(m)\nu(dm) \in (-\infty, \infty), \qquad \phi'(1) = \int_0^\infty \log(m) m \nu(dm).$$

Recall that Δ_t / t converges to $\phi'(0)$ a.s. and that $g + \phi'(0)$ is negative in the subcritical case. proposition 3 then yields the asymptotic behavior of the quenched survival probability :

$$e^{-gt-\Delta_t} \mathbb{P}_{x_0}(Y_t > 0 |\ \Delta) \sim x_0 \Big(c_+ \beta \int_0^t e^{\beta(gt+\Delta_t - gs - \Delta_s)} ds \Big)^{-1/\beta} \qquad (t \to \infty),$$

which converges in distribution to a positive finite limit proportional to x_0. Then,

$$\frac{1}{t} \log \mathbb{P}_{x_0}(Y_t > 0 |\ \Delta) \to g + \phi'(0) \qquad (t \to \infty)$$

in probability.

Additional work is required to get the asymptotic behavior of the annealed survival probability, for which four different regimes appear when the process a.s. goes to zero :

Proposition 4. *We assume that ν satisfies (2.1.2) and (2.1.11), and that ψ and ϕ satisfy (2.2.1) and (2.2.4) respectively.*

a/ If $\phi'(0) + g < 0$ (subcritical case) *and $\theta_{max} > 1$, then*

(i) If $\phi'(1) + g < 0$ (strongly subcritical regime), *then there exists $c_1 > 0$ such that for every $x_0 > 0$,*

$$\mathbb{P}_{x_0}(Y_t > 0) \sim c_1 x_0 e^{t(\phi(1)+g)}, \qquad \text{as} \quad t \to \infty.$$

(ii) If $\phi'(1) + g = 0$ (intermediate subcritical regime), *then there exists $c_2 > 0$ such that for every $x_0 > 0$,*

$$\mathbb{P}_{x_0}(Y_t > 0) \sim c_2 x_0 t^{-1/2} e^{t(\phi(1)+g)}, \qquad \text{as} \quad t \to \infty.$$

(iii) If $\phi'(1) + g > 0$ (weakly subcritical regime) *and $\theta_{max} > \beta + 1$, then for every $x_0 > 0$, there exists $c_3(x_0) > 0$ such that*

$$\mathbb{P}_{x_0}(Y_t > 0) \sim c_3(x_0) t^{-3/2} e^{t(\phi(\tau)+g\tau)}, \qquad \text{as} \quad t \to \infty,$$

where τ is the root of $\phi' + g$ on $]0,1[$: $\phi(\tau) + g\tau = \min_{0<s<1} \{\phi(s) + gs\}$.

b/ If $\phi'(0)+g=0$ (critical case) *and* $\theta_{max}>\beta$, *then for every* $x_0>0$, *there exists* $c_4(x_0)>0$ *such that*

$$\mathbb{P}_{x_0}(Y_t>0)\sim c_4(x_0)\,t^{-1/2}, \qquad as \quad t\to\infty.$$

Démonstration. From Proposition 3 we know that

$$\mathbb{P}_{x_0}(Y_t>0)=1-\mathbb{E}\left[\exp\left\{-x_0\left(c_+\beta\int_0^t e^{-\beta(gs+\Delta_s)}\mathrm{d}s\right)^{-1/\beta}\right\}\right]=\mathbb{E}\left[F\left(\int_0^t e^{-\beta K_s}\mathrm{d}s\right)\right],$$

where $F(x)=1-\exp\{-x_0(c_+\beta x)^{-1/\beta}\}$ and $K_s=\Delta_s+gs$. The function F satisfies assumption (3.1.12) which is required in Theorem 2 (which is stated and proved in the next section). Hence Proposition 4 follows from a direct application of this Theorem. □

In the case of CSBP's without catastrophes ($\nu=0$), the subcritical regime is reduced to (i), and the critical case differs from b/, since the asymptotic behavior is given by $1/t$.
In the strongly and intermediate subcritical cases (ii) and (iii), $\mathbb{E}[Y_t]$ provides the exponential decay factor of the survival probability which is given by $\phi(1)+g$. Moreover the probability of non-extinction is proportional to the initial state x_0 of the population. We refer to the proof of Lemma 3 and Section 2.3 for more details.
In the weakly subcritical case (iii), the survival probability decays exponentially with rate $\phi(\tau)+g\tau$, which is strictly smaller than $\phi(1)+g$. In fact, as it appears in the proof of Theorem 2, the quantity which determines the asymptotic behavior in all cases is $\mathbb{E}[\exp\{\inf_{s\in[0,t]}(\Delta_s+gs)\}]$. We also note that c_3 and c_4 may not be proportional to x_0. We refer to [Ban09] for a result in this vein for discrete branching processes in random environment.

More generally, the results stated above can be compared to the results which appear in the literature of discrete (time and space) branching processes in random environment (BPRE), see e.g. [GL01, GKV03, AGKV05]. A BPRE $(X_n, n\in\mathbb{N})$ is an integer valued branching process, specified by a sequence of generating functions $(f_n, n\in\mathbb{N})$. Conditionally on the environment, individuals reproduce independently of each other and the offsprings of an individual at generation n has generating function f_n. We present briefly the results of Theorem 1.1 in [GK00] and Theorems 1.1, 1.2 and 1.3 in [GKV03]. To lighten the presentation, we do not specify here the moment conditions.
In the subcritical case, i.e. when $\mathbb{E}[\log f_0'(1)]<0$, we have the following three asymptotic regimes as n increases,

$$\mathbb{P}(X_n>0)\sim c\,a_n, \qquad \text{as} \quad n\to\infty,$$

where c is a positive constant and a_n is given by

$$a_n=\mathbb{E}\left[f_0'(1)\right]^n, \quad a_n=n^{-1/2}\mathbb{E}\left[f_0'(1)\right]^n \quad \text{or} \quad a_n=n^{-3/2}\left(\min_{0<s<1}\mathbb{E}\left[(f_0'(1))^s\right]\right)^n,$$

when $\mathbb{E}[f_0'(1)\log f_0'(1)]$ is negative, zero or positive, respectively.
In the critical case, i.e. $\mathbb{E}[\log f_0'(1)]=0$, we have

$$\mathbb{P}(X_n>0)\sim c\,n^{-1/2}, \qquad \text{as} \quad n\to\infty,$$

for some positive constant c. In the particular case when $\beta = 1$, these results on BPRE and the approximation techniques implemented in Section 2.3 can be used to get Proposition 4. We refer to Remarks 2 and 3 for more details.

Finally, in the continuous framework, such results have been established for the Feller diffusion case, i.e. $\beta = 1$, whose drift varies following a Brownian motion (see [BH12]). In other words the process K is given by a Brownian motion plus a drift. The techniques used by the authors rely on an explicit formula for the Laplace transform of exponential functionals of Brownian motion which we cannot find in the literature for the case of Lévy processes. These results have been completed in the surpercritical regime in [Hut11].

Beyond the stable case.

In this section, we prove a similar result to Proposition 4 for CSBP's with catastrophes in the case when the branching mechanism ψ_0 is not stable. For technical reasons, we assume that the Brownian coefficient is positive and the associated Lévy measure μ satisfies a second moment condition.

Corollary 2.3. *Assume that (2.2.4) holds and*

$$\int_{(0,\infty)} z^2 \mu(dz) < \infty, \qquad \sigma^2 > 0, \qquad \int_{(0,\infty)} (\log m)^2 \nu(dm) < \infty.$$

a/ If $\phi'(0) + g < 0$ and $\theta_{max} > 1$, then

(i) If $\phi'(1) + g < 0$, there exist $0 < c_1 \le c_1' < \infty$ such that for every x_0,

$$c_1 x_0 e^{t(\phi(1)+g)} \le \mathbb{P}_{x_0}(Y_t > 0) \le c_1' x_0 e^{t(\phi(1)+g)} \qquad \text{for sufficiently large } t.$$

(ii) If $\phi'(1) + g = 0$, there exist $0 < c_2 \le c_2' < \infty$ such that for every x_0,

$$c_2 x_0 t^{-1/2} e^{t(\phi(1)+g)} \le \mathbb{P}_{x_0}(Y_t > 0) \le c_2' x_0 t^{-1/2} e^{t(\phi(1)+g)} \quad \text{for sufficiently large } t.$$

(iii) If $\phi'(1) + g > 0$ and $\theta_{max} > \beta + 1$, for every x_0, there exist $0 < c_3(x_0) \le c_3'(x_0) < \infty$ such that

$$c_3(x_0) t^{-3/2} e^{t(\phi(\tau)+g\tau)} \le \mathbb{P}_{x_0}(Y_t > 0) \le c_3'(x_0) t^{-3/2} e^{t(\phi(\tau)+g\tau)} \quad (t > 0),$$

where τ is the root of $\phi' + g$ on $]0,1[$.

b/ If $\phi'(0) + g = 0$ and $\theta_{max} > \beta$, then for every x_0, there exist $0 < c_4(x_0) < c_4'(x_0) < \infty$ such that

$$c_4(x_0) t^{-1/2} \le \mathbb{P}_{x_0}(Y_t > 0) \le c_4'(x_0) t^{-1/2} \quad (t > 0).$$

Note that the assumption $\sigma^2 > 0$ is only required for the upper bounds.

Démonstration. We recall that the branching mechanism associated with the CSBP Z satisfies (2.0.1) for every $\lambda \geq 0$. So for every $\lambda \geq 0$,

$$2\sigma^2 \leq \psi''(\lambda) = 2\sigma^2 + \int_{(0,\infty)} z^2 e^{-\lambda z} \mu(dz).$$

Since $c := \int_0^\infty z^2 \mu(dz) < \infty$, ψ'' is continuous on $[0,\infty)$. By Taylor-Lagrange's Theorem, we get for every $\lambda \geq 0$, $\psi_-(\lambda) \leq \psi(\lambda) \leq \psi_+(\lambda)$, where

$$\psi_-(\lambda) = \lambda\psi'(0) + \sigma^2\lambda^2 \quad \text{and} \quad \psi_+(\lambda) = \lambda\psi'(0) + (\sigma^2 + c/2)\lambda^2.$$

We first consider the case $\nu(0,\infty) < \infty$, so that Δ has a finite number of jumps on each compact interval a.s., and we also introduce the CSBP's with catastrophes Y^- and Y^+ which have the same catastrophes Δ as Y, but with the characteristics $(g,\sigma^2,0)$ and $(g,\sigma^2+c/2,0)$, respectively. We denote $u_{-,t}$ and $u_{+,t}$ for their respective Laplace exponent, in other words for all $(\lambda,t) \in \mathbb{R}_+^2$,

$$\mathbb{E}\Big[\exp\{-\lambda Y_t^-\}\Big] = \exp\{-u_{-,t}(\lambda)\}, \qquad \mathbb{E}\Big[\exp\{-\lambda Y_t^+\}\Big] = \exp\{-u_{+,t}(\lambda)\}.$$

Thus conditionally on Δ, for every time t such that $\Delta_t = \Delta_{t-}$, we deduce, thanks to Theorem 1, the following identities

$$u'_{-,t}(\lambda) = -\psi_-(u_{-,t}), \qquad u'_{+,t}(\lambda) = -\psi_+(u_{+,t}), \qquad u'_t(\lambda) = -\psi(u_t).$$

Moreover for every t such that $\theta_t = \exp\{\Delta_t - \Delta_{t-}\} \neq 1$,

$$\frac{u_{-,t}(\lambda)}{u_{-,t-}(\lambda)} = \frac{u_t(\lambda)}{u_{t-}(\lambda)} = \frac{u_{+,t}(\lambda)}{u_{+,t-}(\lambda)} = \theta_t,$$

and $u_{-,0}(\lambda) = u_0(\lambda) = u_{+,0}(\lambda) = \lambda$. So for all t,λ, we have

$$u_{+,t}(\lambda) \leq u(t,\lambda) \leq u_{-,t}(\lambda).$$

The extension of the above inequality to the case $\nu(0,\infty) \in [0,\infty]$ can be achieved by successive approximations. We defer the technical details to Section A.
Having into account that the above inequality holds in general, we deduce, taking $\lambda \to \infty$, that

$$\mathbb{P}(Y_t^+ > 0) \leq \mathbb{P}(Y_t > 0) \leq \mathbb{P}(Y_t^- > 0).$$

The result then follows from the asymptotic behavior of $\mathbb{P}(Y_t^- > 0)$ and $\mathbb{P}(Y_t^+ > 0)$, which are inherited from Proposition 4. □

2.3 Local limit theorem for some functionals of Lévy processes

We proved in Proposition 3 that the probability that a stable CSBP with catastrophes becomes extinct at time t equals the expectation of a functional of a Lévy process. We now prove the key result of the paper. It deals with the asymptotic behavior of the mean of some Lévy

functionals.
More precisely, we are interested in the asymptotic behavior at infinity of

$$a_F(t) := \mathbb{E}\left[F\left(\int_0^t \exp\{-\beta K_s\}\mathrm{d}s\right)\right],$$

where K is a Lévy process with bounded variation paths and F belongs to a particular class of functions on $\mathbb{R}_+$. We will focus on functions which decrease polynomially at infinity (with exponent $-1/\beta$). The motivations come from the previous section. In particular, the Proposition 4 is a direct application of Theorem 2.

Thus, we consider a Lévy process $K = (K_t, t \geq 0)$ of the form

$$K_t = \gamma t + \sigma_t^{(+)} - \sigma_t^{(-)}, \qquad t \geq 0, \tag{2.3.1}$$

where γ is a real constant, $\sigma^{(+)}$ and $\sigma^{(-)}$ are two independent pure jump subordinators. We denote by Π, $\Pi^{(+)}$ and $\Pi^{(-)}$ the associated Lévy measures of K, $\sigma^{(+)}$ and $\sigma^{(-)}$, respectively. We also define the Laplace exponents of K, $\sigma^{(+)}$ and $\sigma^{(-)}$ by

$$\phi_K(\lambda) = \log\mathbb{E}\left[e^{\lambda K_1}\right], \quad \phi_K^+(\lambda) = \log\mathbb{E}\left[e^{\lambda\sigma_1^{(+)}}\right] \quad \text{and} \quad \phi_K^-(\lambda) = \log\mathbb{E}\left[e^{-\lambda\sigma_1^{(-)}}\right], \tag{2.3.2}$$

and assume that

$$\theta_{max} = \sup\left\{\lambda \in \mathbb{R}^+, \int_{[1,\infty)} e^{\lambda x}\Pi^{(+)}(\mathrm{d}x) < \infty\right\} > 0. \tag{2.3.3}$$

From the Lévy-Khintchine formula, we deduce

$$\phi_K(\lambda) = \gamma\lambda + \int_{(0,\infty)}\left(e^{\lambda x} - 1\right)\Pi^{(+)}(\mathrm{d}x) + \int_{(0,\infty)}\left(e^{-\lambda x} - 1\right)\Pi^{(-)}(\mathrm{d}x).$$

Finally, we assume that $\mathbb{E}[K_1^2] < \infty$, which is equivalent to

$$\int_{(-\infty,\infty)} x^2\Pi(dx) < \infty. \tag{2.3.4}$$

Theorem 2. *Assume that (2.3.1), (2.3.3) and (2.3.4) hold. Let $\beta \in (0,1]$ and F be a positive non increasing function such that for $x \geq 0$*

$$F(x) = C_F(x+1)^{-1/\beta}\left[1 + (1+x)^{-\varsigma}h(x)\right], \tag{2.3.5}$$

where $\varsigma \geq 1$, C_F is a positive constant, and h is a Lipschitz function which is bounded.

a/ If $\phi_K'(0) < 0$

(i) If $\theta_{max} > 1$ and $\phi_K'(1) < 0$, there exists a positive constant c_1 such that

$$a_F(t) \sim c_1 e^{t\phi_K(1)}, \quad \text{as} \quad t \to \infty.$$

(ii) If $\theta_{max} > 1$ and $\phi_K'(1) = 0$, there exists a positive constant c_2 such that

$$a_F(t) \sim c_2 t^{-1/2} e^{t\phi_K(1)}, \quad \text{as} \quad t \to \infty.$$

(iii) If $\theta_{max} > \beta + 1$ *and* $\phi'_K(1) > 0$, *there exists a positive constant* c_3 *such that*

$$a_F(t) \sim c_3 t^{-3/2} e^{t\phi_K(\tau)}, \quad as \quad t \to \infty,$$

where τ *is the root of* ϕ'_K *on* $]0,1[$.

b/ If $\theta_{max} > \beta$ *and* $\phi'_K(0) = 0$, *there exists a positive constant* c_4 *such that*

$$a_F(t) \sim c_4 t^{-1/2}, \quad as \quad t \to \infty.$$

This result generalizes Lemma 4.7 in Carmona et al. [CPY97] in the case when the process K has bounded variation paths. More precisely, the authors in [CPY97] only provide a precise asymptotic behavior in the case when $\phi'_K(1) < 0$.

The assumption on the behavior of F as $x \to \infty$ is finely used to get the asymptotic behavior of $a_F(t)$. Lemma 2 gives the properties of F which are required in the proof.

The strongly subcritical case (case (i)) is proved using a continuous time change of measure (see Section 2.3). For the remaining cases, we divide the proof in three steps. The first one (see Lemma 1) consists in discretizing the exponential functional $\int_0^t \exp(-\beta K_s)\mathrm{d}s$ using the random variables

$$A_{p,q} = \sum_{i=0}^{p} \exp\{-\beta K_{i/q}\} = \sum_{i=0}^{p} \prod_{j=0}^{i-1} \exp\Big\{-\beta\big(K_{(j+1)/q} - K_{j/q}\big)\Big\} \quad ((p,q) \in \mathbb{N} \times \mathbb{N}^*). \tag{2.3.6}$$

Secondly (see Lemmas 3, 4 and 5), we study the asymptotic behavior of the discretized expectation

$$F_{p,q} := \mathbb{E}\Big[F\Big(A_{p,q}/q\Big)\Big] \quad (q \in \mathbb{N}^*), \tag{2.3.7}$$

when p goes to infinity. This step relies on Theorem 2.1 in [GL01], which is a limit theorem for random walks on an affine group and generalizes theorems A and B in [LPP97].
Finally (see Sections 2.3 and 2.3), we prove that the limit of $F_{\lfloor qt \rfloor, q}$, when $q \to \infty$, and $a_F(t)$ both have the same asymptotic behavior when t goes to infinity.

Discretization of the Lévy process

The following result, which follows from the property of independent and stationary increments of the process K, allows us to concentrate on $A_{p,q}$, which has been defined in (2.3.6).

Lemma 1. *Let* $t \geq 1$ *and* $q \in \mathbb{N}^*$. *Then*

$$\frac{1}{q} e^{-\beta(|\gamma|/q + \sigma^{(+)}_{1/q})} A^{(1)}_{\lfloor qt \rfloor - 1, q} \leq \int_0^t e^{-\beta K_s} \mathrm{d}s \leq \frac{1}{q} e^{\beta(|\gamma|/q + \sigma^{(-)}_{1/q})} A^{(2)}_{\lfloor qt \rfloor, q},$$

where for every $(p,q) \in \mathbb{N} \times \mathbb{N}^*$, $\sigma^{(+)}_{1/q}$ *(resp* $\sigma^{(-)}_{1/q}$*) is independent of* $A^{(1)}_{p,q}$ *(resp* $A^{(2)}_{p,q}$*) and*

$$A_{p,q} \overset{(d)}{=} A^{(1)}_{p,q} \overset{(d)}{=} A^{(2)}_{p,q}.$$

Démonstration. Let (p,q) be in $\mathbb{N}\times\mathbb{N}^*$ and $s\in[p/q,(p+1)/q]$. Then

$$K_s \le K_{p/q} + |\gamma|/q + [\sigma^{(+)}_{(p+1)/q} - \sigma^{(+)}_{p/q}] \quad \text{and} \quad K_s \ge K_{p/q} - |\gamma|/q - [\sigma^{(-)}_{(p+1)/q} - \sigma^{(-)}_{p/q}]. \tag{2.3.8}$$

Now introduce

$$K^{(1)}_{p/q} = K_{p/q} + [\sigma^{(+)}_{(p+1)/q} - \sigma^{(+)}_{p/q}] - \sigma^{(+)}_{1/q} = \gamma p/q + [\sigma^{(+)}_{(p+1)/q} - \sigma^{(+)}_{1/q}] - \sigma^{(-)}_{p/q},$$

and

$$K^{(2)}_{p/q} = K_{p/q} - [\sigma^{(-)}_{(p+1)/q} - \sigma^{(-)}_{p/q}] + \sigma^{(-)}_{1/q} = \gamma p/q + \sigma^{(+)}_{p/q} - [\sigma^{(-)}_{(p+1)/q} - \sigma^{(-)}_{1/q}].$$

Then, we have for all $(p,q)\in\mathbb{N}\times\mathbb{N}^*$

$$(K_0, K_{1/q}, ..., K_{p/q}) \overset{(d)}{=} (K^{(1)}_0, K^{(1)}_{1/q}, ..., K^{(1)}_{p/q}) \overset{(d)}{=} (K^{(2)}_0, K^{(2)}_{1/q}, ..., K^{(2)}_{p/q}).$$

Moreover, the random vector $(K^{(1)}_0, K^{(1)}_{1/q}, ..., K^{(1)}_{p/q})$ is independent of $\sigma^{(+)}_{1/q}$ and $(K^{(2)}_0, K^{(2)}_{1/q}, ..., K^{(2)}_{p/q})$ is independent of $\sigma^{(-)}_{1/q}$. Finally, the definition of

$$A^{(i)}_{p,q} = \sum_{i=0}^{p} \exp\{-\beta K^{(i)}_{i/q}\}$$

for $i\in\{1,2\}$ and the inequalities in (2.3.8) complete the proof. □

Asymptotical behavior of the discretized process

First, we recall Theorem 2.1 of [GL01] in the case where the test functions do not vanish. This is the key result to obtain the asymptotic behavior of the discretized process.

Theorem 3 (Giuvarc'h, Liu 01). *Let $(a_n,b_n)_{n\ge0}$ be a $(\mathbb{R}_+^*)^2$-valued sequence of iid random variables such that $\mathbb{E}[\log(a_0)]=0$. Assume that $b_0/(1-a_0)$ is not constant a.s. and define $A_0=1$, $A_n=\prod_{k=0}^{n-1}a_k$ and $B_n=\sum_{k=0}^{n-1}A_kb_k$, for $n\ge1$. Let η,κ,ξ be three positive numbers such that $\kappa<\xi$, and $\tilde\phi$ and $\tilde\psi$ be two positive continuous functions on $\mathbb{R}_+$ such that they do not vanish and for a constant $C>0$ and for every $a>0$, $b\ge0$, $b'\ge0$, we have*

$$\tilde\phi(a)\le Ca^{\kappa}, \quad \tilde\psi(b)\le\frac{C}{(1+b)^{\xi}}, \quad \text{and} \quad |\tilde\psi(b)-\tilde\psi(b')|\le C|b-b'|^{\eta}.$$

Moreover, assume that

$$\mathbb{E}[a_0^{\kappa}]<\infty, \quad \mathbb{E}[a_0^{-\eta}]<\infty, \quad \mathbb{E}[b_0^{\eta}]<\infty \quad \text{and} \quad \mathbb{E}[a_0^{-\eta}b_0^{-\eta}]<\infty.$$

Then there exist two positive constants $c(\tilde\phi,\tilde\psi)$ and $c(\tilde\psi)$ such that

$$\lim_{n\to\infty} n^{3/2}\mathbb{E}\left[\tilde\phi(A_n)\tilde\psi(B_n)\right]=c(\tilde\phi,\tilde\psi) \qquad \text{and} \qquad \lim_{n\to\infty} n^{1/2}\mathbb{E}\left[\tilde\psi(B_n)\right]=c(\tilde\psi).$$

Let us now state a technical lemma on the tail of function F, useful to get the asymptotical behaviour of the disretized process. Its proof is deferred to Section A for the convenience of the reader.

Lemma 2. *Assume that F satisfies (3.1.12). Then there exist two positive finite constants η and M such that for all (x,y) in $\mathbb{R}_+^2$ and ε in $[0,\eta]$,*

$$\left|F(x)-C_F x^{-1/\beta}\right| \leq Mx^{-(1+\varepsilon)/\beta}, \tag{2.3.9}$$

$$\left|F(x)-F(y)\right| \leq M\left|x^{-1/\beta}-y^{-1/\beta}\right|. \tag{2.3.10}$$

Recall the definitions of $A_{p,q}$ and $F_{p,q}$ in (2.3.6) and (2.3.7), respectively. The three following lemmas study the asymptotic behavior of $F_{p,q}$ and the mean value of $(A_{p,q}/q)^{-1/\beta}$ in the regimes of (ii), (iii) and b/.

Lemma 3. *Assume that $|\phi_K'(0+)|<\infty$, $\theta_{max}>1$ and $\phi_K'(1)=0$. Then there exists a positive and finite constant $c_2(q)$ such that,*

$$F_{p,q}\sim C_F c_2(q)(p/q)^{-1/2}e^{(p/q)\phi_K(1)}, \quad as \quad p\to\infty, \tag{2.3.11}$$

and

$$\mathbb{E}\left[\left(A_{p,q}/q\right)^{-1/\beta}\right]\sim c_2(q)(p/q)^{-1/2}e^{(p/q)\phi_K(1)}, \quad as \quad p\to\infty. \tag{2.3.12}$$

Démonstration. Let us introduce the exponential change of measure known as the Escheer transform

$$\left.\frac{\mathrm{d}\mathbb{P}^{(\lambda)}}{\mathrm{d}\mathbb{P}}\right|_{\mathscr{F}_t} = e^{\lambda K_t-\phi_K(\lambda)t} \qquad \text{for } \lambda\in[0,\theta_{max}), \tag{2.3.13}$$

where $(\mathscr{F}_t)_{t\geq 0}$ is the natural filtration generated by K which is naturally completed.

The following equality in law

$$A_{p,q}=e^{-\beta K_{p/q}}\Big(\sum_{i=0}^{p}e^{\beta(K_{p/q}-K_{i/q})}\Big)\overset{(d)}{=}e^{-\beta K_{p/q}}\Big(\sum_{i=0}^{p}e^{\beta K_{i/q}}\Big),$$

leads to $e^{-(p/q)\phi_K(1)}\mathbb{E}\left[A_{p,q}^{-1/\beta}\right]=\mathbb{E}^{(1)}\left[\tilde{A}_{p,q}^{-1/\beta}\right]$, where $\tilde{A}_{p,q}=\sum_{i=0}^{p}e^{\beta K_{i/q}}$. Let $\varepsilon>0$ be such that (2.3.9) holds and observe that $\tilde{A}_{p,q}\geq 1$ a.s. for every (p,q) in $\mathbb{N}\times\mathbb{N}^*$. Thus,

$$\mathbb{E}^{(1)}\left[\tilde{A}_{p,q}^{-(1+\varepsilon)/\beta}\right]\leq\mathbb{E}^{(1)}\left[\tilde{A}_{p,q}^{-1/\beta}\right]\leq\mathbb{E}^{(1)}\left[\inf_{i\in[0,p]\cap\mathbb{N}}e^{-K_{i/q}}\right].$$

Since $\phi_K'(1)=0$ and $\mathbb{E}[K_{1/q}^2]<\infty$, Theorem A in [Koz76] implies

$$\mathbb{E}^{(1)}\left[\inf_{i\in[0,p]\cap\mathbb{N}}e^{-K_{i/q}}\right]\sim\hat{C}_q(p/q)^{-1/2}, \quad \text{as} \quad p\to\infty,$$

where $\hat{C}_q$ is a finite positive constant. We define for $z\geq 1$,

$$D_q(z,p)=(p/q)^{1/2}\mathbb{E}^{(1)}\left[\tilde{A}_{p,q}^{-z/\beta}\right].$$

Moreover, we note that there exists $p_0\in\mathbb{N}$ such that for $p\geq p_0$, $D_q(1,p)\leq 2\hat{C}_q$.

Our aim is to prove that $D_q(1,p)$ converges, as p increases, to a finite positive constant $d_2(q)$. Then, we introduce an arbitrary $x \in (0,(C_F/M)^{1/\varepsilon} q^{-1/\beta})$ and apply Theorem 3 with

$$\tilde{\psi}(z) = F(z), \quad \tilde{\phi}(z) = z^{1/(2\beta)}, \qquad (\eta,\kappa,\xi) = (1,1/(2\beta),1/\beta).$$

Observe that F is a Lipschitz function and that under the probability measure $\mathbb{P}^{(1)}$, $(a_n,b_n)_{n\geq 0} = (\exp(\beta(K_{(n+1)/q}-K_{n/q})), x^{-\beta}q^{-1})_{n\geq 0}$ is an i.i.d. sequence of random variables with $\mathbb{E}^{(1)}[\log(a_0)] = 0$, since $\phi'_K(1) = 0$. Moreover, a simple computation gives

$$\mathbb{E}^{(1)}[a_0^{-1}] = e^{(\phi_K(1-\beta)-\phi_K(1))/q} < \infty,$$

so that the moment conditions of Theorem 3 are satisfied. We apply the result with

$$B_n = q^{-1}x^{-\beta}\sum_{i=0}^{n-1} e^{\beta K_{i/q}}, \quad n \in \mathbb{N}^*$$

and we get the existence of a positive finite real number $b(q,x)$ such that

$$(p/q)^{1/2}\mathbb{E}^{(1)}\left[F\left(x^{-\beta}\tilde{A}_{p,q}/q\right)\right] \to b(q,x), \quad \text{as} \quad p\to\infty.$$

Taking expectation in (2.3.9) yields

$$\left|(p/q)^{1/2}\mathbb{E}^{(1)}\left[F\left(x^{-\beta}\tilde{A}_{p,q}/q\right)\right] - C_F x q^{1/\beta} D_q(1,p)\right| \leq M x^{1+\varepsilon} q^{(1+\varepsilon)/\beta} D_q(1+\varepsilon,p). \tag{2.3.14}$$

Defining $\underline{D}_q := \liminf_{p\to\infty} D_q(1,p)$ and $\overline{D}_q := \limsup_{p\to\infty} D_q(1,p)$, we combine the two last dispalys to get

$$C_F x q^{1/\beta}\overline{D}_q \leq b(q,x) + M x^{1+\varepsilon} q^{(1+\varepsilon)/\beta} \limsup_{p\to\infty} D_q(1+\varepsilon,p),$$

and

$$C_F x q^{1/\beta}\underline{D}_q \geq b(q,x) - M x^{1+\varepsilon} q^{(1+\varepsilon)/\beta} \limsup_{p\to\infty} D_q(1+\varepsilon,p).$$

Adding that $D_q(z,p)$ is non-increasing with respect to z, $D_q(1+\varepsilon,p) \leq D_q(1,p) \leq 2\hat{C}_q$ for every $p \geq p_0$ and

$$\overline{D}_q - \underline{D}_q \leq \frac{4M\hat{C}_q x^{\varepsilon} q^{\varepsilon/\beta}}{C_F}.$$

Finally, letting $x \to 0$, we get that $D_q(1,p)$ converges to a finite constant $d_2(q)$. Moreover, from (2.3.14), we get for every integer p :

$$(C_F x q^{1/\beta} + M x^{1+\varepsilon} q^{(1+\varepsilon)/\beta}) D_q(1,p) \geq (p/q)^{1/2}\mathbb{E}^{(1)}\left[F\left(x^{-\beta}\tilde{A}_{p,q}/q\right)\right].$$

Letting $p\to\infty$, we get that $d_2(q)$ is positive, which gives (2.3.12).

Now, using (2.3.9), we get

$$\mathbb{E}\left|F_{p,q} - C_F\left(A_{p,q}/q\right)^{-1/\beta}\right| \leq \mathbb{E}\left[\left(A_{p,q}/q\right)^{-(1+\varepsilon)/\beta}\right],$$

so the asymptotic behavior in (2.3.11) will be proved as soon as we show that

$$\mathbb{E}\Big[A_{p,q}^{-(1+\varepsilon)/\beta}\Big] = o\Big(\mathbb{E}\Big[A_{p,q}^{-1/\beta}\Big]\Big), \quad \text{as} \quad p\to\infty.$$

From the Escheer transform (2.3.13), with $\lambda = 1+\varepsilon$, and the independence of the increments of K, we have

$$\begin{aligned}
\mathbb{E}\Big[A_{p,q}^{-(1+\varepsilon)/\beta}\Big] &= e^{(p/q)\phi_K(1)}\mathbb{E}^{(1)}\Big[\Big(\sum_{i=0}^{p} e^{-\beta K_{i/q}}\Big)^{-\varepsilon/\beta}\Big(\sum_{i=0}^{p} e^{\beta(K_{p/q}-K_{i/q})}\Big)^{-1/\beta}\Big] \\
&\le e^{(p/q)\phi_K(1)}\mathbb{E}^{(1)}\Big[\inf_{0\le i\le\lfloor p/3\rfloor} e^{\varepsilon K_{i/q}} \inf_{\lfloor 2p/3\rfloor\le j\le p} e^{-(K_{p/q}-K_{j/q})}\Big] \\
&= e^{(p/q)\phi_K(1)}\mathbb{E}^{(1)}\Big[\inf_{0\le i\le\lfloor p/3\rfloor} e^{\varepsilon K_{i/q}}\Big]\mathbb{E}^{(1)}\Big[\inf_{0\le j\le\lfloor p/3\rfloor} e^{-K_{j/q}}\Big].
\end{aligned}$$

Using (2.3.4), we observe that $\mathbb{E}^{(1)}[K_{1/q}] = 0$ and $\mathbb{E}^{(1)}[K_{1/q}^2] < \infty$. We can then apply Theorem A in [Koz76] to the random walks $(-K_{i/q})_{i\ge1}$ and $(\varepsilon K_{i/q})_{i\ge1}$. Therefore, there exists $C(q) > 0$ such that

$$\mathbb{E}\Big[A_{p,q}^{-(1+\varepsilon)/\beta}\Big] \le (C(q)/p)e^{(p/q)\phi_K(1)} = o\Big(\mathbb{E}\Big[A_{p,q}^{-1/\beta}\Big]\Big), \quad \text{as} \quad p\to\infty.$$

Taking $c_2(q) = d_2(q)q^{1/\beta}$ leads to the result. □

Remark 2. *In the particular case when $\beta = 1$, it is enough to apply Theorem 1.2 in [GKV03] to a geometric BPRE $(X_n, n\ge0)$ whose p.g.f's satisfy*

$$f_n(s) = \sum_{k=0}^{\infty} p_n q_n^k s^k = \frac{p_n}{1-q_n s},$$

with $1/p_n = 1+\exp\{\beta(K_{(n+1)/q} - K_{n/q})\}$, and $q_n = 1-p_n$. Using $\mathbb{E}[A_{p,q}^{-1}] = \mathbb{P}(X_p > 0)$ and $\log f_0'(1) = K_{1/q}$, allows to get the asymptotic behavior of $\mathbb{E}[A_{p,q}^{-1}]$ from the speed of extinction of BPRE in the case of geometric reproduction law (with the extra assumption $\phi_K(2) < \infty$).

Recall that τ is the root of ϕ_K' on $]0,1[$, i.e. $\phi_K(\tau) = \min_{0<s<1}\phi_K(s)$.

Lemma 4. *Assume that $\phi_K'(0) < 0$, $\phi_K'(1) > 0$ and $\theta_{max} > \beta+1$. Then there exist two positive constants $d(q)$ and $c_3(q)$ such that*

$$F_{p,q} \sim c_3(q)(p/q)^{-3/2}e^{(p/q)\phi_K(\tau)}, \quad \text{as} \quad p\to\infty, \tag{2.3.15}$$

and

$$\mathbb{E}\Big[(A_{p,q}/q)^{-1/\beta}\Big] \sim d(q)(p/q)^{-3/2}e^{(p/q)\phi_K(\tau)}, \quad \text{as} \quad p\to\infty. \tag{2.3.16}$$

Démonstration. First we apply Theorem 3 where, for $z\ge0$,

$$\tilde\psi(z) = F(z), \qquad \tilde\phi(z) = z^{\tau/\beta}, \qquad (\eta,\kappa,\xi) = (1,\tau/\beta,1/\beta).$$

Again F is a Lipschitz function, and under the probability measure $\mathbb{P}^{(\tau)}$, $(a_n, b_n)_{n\geq 0} = (\exp(-\beta(K_{(n+1)/q} - K_{n/q})), q^{-1})_{n\geq 0}$, is an i.i.d. sequence of random variables such that $\mathbb{E}^{(\tau)}[\log(a_0)] = 0$, since $\phi'_K(\tau) = 0$. The moment conditions

$$\mathbb{E}^{(\tau)}\left[a_0^{\tau/\beta}\right] = e^{-\phi_K(\tau)/q} < \infty \qquad \text{and} \qquad \mathbb{E}^{(\tau)}\left[a_0^{-1}\right] = e^{(\phi_K(\beta+\tau)-\phi_K(\tau))/q} < \infty,$$

enable us to apply Theorem 3. In this case,

$$B_n = q^{-1}\sum_{i=0}^{n-1} e^{-\beta K_{i/q}}, \quad n \in \mathbb{N}^*.$$

Then there exists $c_3(q) > 0$ such that

$$\mathbb{E}\left[F(A_{p,q}/q)\right] e^{-(p/q)\phi_K(\tau)} = \mathbb{E}^{(\tau)}\left[F(A_{p,q}/q)e^{-\tau K_{p/q}}\right] \sim \quad c_3(q)(p/q)^{-3/2},$$

as $p \to \infty$. This gives (2.3.15).

To prove

$$\mathbb{E}\left[(A_{p,q}/q)^{-1/\beta}\right] \sim d(q)(p/q)^{-3/2} e^{\frac{p}{q}\phi_K(\tau)}, \quad \text{as} \quad p \to \infty$$

for $d(q) > 0$, we follow the same arguments as those used in the proof of Lemma 3. In other words, we define for $z \geq 1$,

$$D_q(z,p) = (p/q)^{3/2} e^{-(p/q)\phi_K(\tau)} \mathbb{E}\left[A_{p,q}^{-z/\beta}\right],$$

which is non-increasing with respect to z. We obtain the same type of inequalities as in Lemma 3, for the random variable A instead of $\tilde{A}$.

Again we take $\varepsilon > 0$ such that (2.3.9) holds. Then Lemma 7 in [Hir98] yields the existence of $C_q > 0$ such that for p large enough,

$$\mathbb{E}\left[A_{p,q}^{-(1+\varepsilon)/\beta}\right] \leq \mathbb{E}\left[A_{p,q}^{-1/\beta}\right] \leq \mathbb{E}\left[\inf_{i\in[0,p]\cap\mathbb{N}} e^{-K_{i/q}}\right] \sim C_q(p/q)^{-3/2} e^{(p/q)\phi_K(\tau)}.$$

Finally, we use Theorem 3 to get $0 < \liminf_{n\to\infty} D_q(1,n) = \limsup_{n\to\infty} D_q(1,n) < \infty$, which completes the proof. □

Lemma 5. *Assume that $\phi'_K(0) = 0$ and $\theta_{max} > \beta$. Then there exist two positive constants $b(q)$ and $c_4(q)$ such that*

$$F_{p,q} \sim c_4(q)(p/q)^{-1/2}, \quad \text{as} \quad p \to \infty, \tag{2.3.17}$$

and

$$\mathbb{E}\left[(A_{p,q}/q)^{-1/\beta}\right] \sim b(q)(p/q)^{-1/2}, \quad \text{as} \quad p \to \infty. \tag{2.3.18}$$

Démonstration. The proof is almost the same as the proof of Lemma 4. We first apply Theorem 3 to the same function $\tilde{\psi}$ and sequence $(a_n, b_n)_{n\geq 0}$ defined in Lemma 4 but under the probability measure $\mathbb{P}$ instead of $\mathbb{P}^{(\tau)}$. Then, we get

$$\mathbb{E}\left[F(A_{p,q}/q)\right] \sim c_4(q)(p/q)^{-1/2}, \quad \text{as} \quad p \to \infty.$$

Now, we define for $z \geq 1$,

$$D_q(z,p) = (p/q)^{1/2}\mathbb{E}\left[A_{p,q}^{-z/\beta}\right],$$

and from Theorem A in [Koz76] and Theorem 3, we obtain that $D_q(1,p)$ has a positive finite limit when p goes to infinity. □

From the discretized process to the continuous process

Up to now, the asymptotic behavior of the processes was depending on the step size $1/q$. By letting q tend to infinity, we obtain our results in continuous time. To do this we shall use several times a technical Lemma on limits of sequences.

Lemma 6. *Assume that the non-negative sequences* $(a_{n,q})_{(n,q)\in\mathbb{N}^2}$, $(a'_{n,q})_{(n,q)\in\mathbb{N}^2}$ *and* $(b_n)_{n\in\mathbb{N}}$ *satisfy for every* $(n,q)\in\mathbb{N}^2$ *:*

$$a_{n,q} \le b_n \le a'_{n,q},$$

and that there exist three sequences $(a(q))_{q\in\mathbb{N}}$, $(c^-(q))_{q\in\mathbb{N}}$ *and* $(c^+(q)_{q\in\mathbb{N}}$ *such that*

$$\lim_{n\to\infty} a_{n,q} = c^-(q)a(q), \quad \lim_{n\to\infty} a'_{n,q} = c^+(q)a(q), \quad \textit{and} \quad \lim_{q\to\infty} c^-(q) = \lim_{q\to\infty} c^+(q) = 1.$$

Then there exists a non-negative constant a *such that*

$$\lim_{q\to\infty} a(q) = \lim_{n\to\infty} b_n = a.$$

Démonstration. From our assumptions, it is clear that for every $q\in\mathbb{N}$

$$\limsup_{n\to\infty} b_n \le c^+(q)a(q) \quad \text{and} \quad c^-(q)a(q) \le \liminf_{n\to\infty} b_n.$$

Then letting q go to infinity, we obtain

$$\limsup_{n\to\infty} b_n \le \liminf_{q\to\infty} a(q) \quad \text{and} \quad \limsup_{q\to\infty} a(q) \le \liminf_{n\to\infty} b_n,$$

which ends the proof. □

Recalling the notations (2.3.11) to (2.3.18), we prove the following limits :

Lemma 7. *There exist five finite positive constants* b, d, c_2, c_3 *and* c_4 *such that*

$$(b(q), d(q), c_2(q), c_3(q), c_4(q)) \longrightarrow (b, d, c_2, c_3, c_4), \quad \textit{as} \quad q\to\infty. \tag{2.3.19}$$

Démonstration. First we prove the convergence of $d(q)$. From Lemma 1, we know that for every $n\in\mathbb{N}^*$

$$e^{\frac{\phi_K^-(1)-|\gamma|}{q}}\mathbb{E}\Big[\Big(A_{nq,q}/q\Big)^{-1/\beta}\Big] \le \mathbb{E}\Big[\Big(\int_0^n e^{-\beta K_u}\mathrm{d}u\Big)^{-1/\beta}\Big] \le e^{\frac{\phi_K^+(1)+|\gamma|}{q}}\mathbb{E}\Big[\Big(A_{nq-1,q}/q\Big)^{-1/\beta}\Big]. \tag{2.3.20}$$

A direct application of Lemma 6 with

$$a(q) = d(q), \quad c^-(q) = e^{(\phi_K^-(1)-|\gamma|)/q}, \quad \text{and} \quad c^+(q) = e^{(\phi_K^+(1)+|\gamma|)/q},$$

yields that $d(q)$ converges as $q\to\infty$ to a finite non-negative constant d. Let us now prove that d is positive. Let (q_1,q_2) be in $\mathbb{N}^2$. According to (2.3.16) and (2.3.20) there exists $n\in\mathbb{N}$ such that

$$0 < e^{\frac{\phi_K^-(1)-|\gamma|}{q_1}} d(q_1)/2 \le n^{3/2} e^{-n\phi_K(\tau)}\mathbb{E}\Big[\Big(\int_0^n e^{-\beta K_u}\mathrm{d}u\Big)^{-1/\beta}\Big] \le 2e^{\frac{\phi_K^+(1)+|\gamma|-\phi_K(\tau)}{q_2}} d(q_2).$$

Letting q_2 go to infinity, we conclude that $\liminf_{q\to\infty} d(q) > 0$. Similar arguments imply the convergence of $b(q)$ to a positive constant.

Now, we prove the convergence of $c_2(q)$, $c_3(q)$ and $c_4(q)$. Again the proofs of the three cases are very similar, so we only prove the second one. From Lemmas 1 and 4, we know that for every $(n,q) \in \mathbb{N}^2$,

$$\mathbb{E}\left[F\left(e^{\beta(|\gamma|/q+\sigma^{(-)}_{1/q})} A_{nq,q}/q\right)\right] \le a_F(n) \le \mathbb{E}\left[F\left(e^{-\beta(|\gamma|/q+\sigma^{(+)}_{1/q})} A_{nq-1,q}/q\right)\right].$$

Using (2.3.10), we obtain

$$F_{nq,q} + M\mathbb{E}\left[e^{-|\gamma|/q-\sigma^{(-)}_{1/q}} - 1\right]\mathbb{E}\left[\left(\frac{A_{nq,q}}{q}\right)^{-\frac{1}{\beta}}\right] \le a_F(n) \le F_{nq-1,q} + M\mathbb{E}\left[e^{|\gamma|/q+\sigma^{(+)}_{1/q}} - 1\right]\mathbb{E}\left[\left(\frac{A_{nq-1,q}}{q}\right)^{-\frac{1}{\beta}}\right].$$

Thus, dividing by $n^{-3/2}\exp(n\phi_K(\tau))$ in the above inequality, we get the convergence using Lemmas 4, 6 and Equation (2.3.10) with

$$a(q) = c_3(q), \quad c^-(q) = 1 - \frac{Md(q)(e^{(\phi_K^-(1)-|\gamma|)/q} - 1)}{c_3(q)} \quad c^+(q) = 1 + \frac{Md(q)(e^{(\phi_K^+(1)+|\gamma|)/q} - 1)}{c_3(q)}.$$

We then prove that $\lim_{q\to\infty} c_3(q)$ is positive using similar arguments as previously. □

Proof of Theorem 2

Proof of Theorem 2 a/ (i). Recall from Lemma II.2 in [BLG00] that the process $(K_t - K_{(t-s)^-}, 0 \le s \le t)$ has the same law as $(K_s, 0 \le s \le t)$. Then

$$\int_0^t e^{-\beta K_s} ds = \int_0^t e^{-\beta K_{(t-s)}} ds = e^{-\beta K_t}\int_0^t e^{\beta K_t - \beta K_{(t-s)}} ds \overset{(d)}{=} e^{-\beta K_t}\int_0^t e^{\beta K_s} ds.$$

We first note that for every $q \in \mathbb{N}^*$ and $t \ge 2/q$, Lemma 1 leads to

$$\begin{aligned}\mathbb{E}\left[\left(\int_0^t e^{-\beta K_s} ds\right)^{-1/\beta}\right] &\le \mathbb{E}\left[\left(\int_0^{2/q} e^{-\beta K_s} ds\right)^{-1/\beta}\right] \\ &\le q^{1/\beta} e^{|\gamma|/q}\mathbb{E}\left(e^{\sigma^{(+)}_{1/q}}(A^{(1)}_{1,q})^{-1/\beta}\right) \\ &\le q^{1/\beta}\exp\left(\frac{\phi_K(1) + |\gamma| + \phi_K^+(1)}{q}\right) < \infty,\end{aligned}$$

where ϕ_K^+ was defined in (2.3.2). Hence using (2.3.13), with $\lambda = 1$, we have

$$\mathbb{E}\left[\left(\int_0^t e^{-\beta K_s} ds\right)^{-1/\beta}\right] = \mathbb{E}\left[e^{K_t}\left(\int_0^t e^{\beta K_s} ds\right)^{-1/\beta}\right] = e^{t\phi_K(1)}\mathbb{E}^{(1)}\left[\left(\int_0^t e^{\beta K_s} ds\right)^{-1/\beta}\right].$$

The above identity implies that the decreasing function $t \mapsto \mathbb{E}^{(1)}[(\int_0^t e^{\beta K_s} ds)^{-1/\beta}]$ is finite for all $t > 0$. So it converges to a non-negative and finite limit c_1, as t increases. This limit is positive, since under the probability $\mathbb{P}^{(1)}$, K is still a Lévy process with negative mean $\mathbb{E}^{(1)}(K_1) = \phi'_K(1)$ and according to Theorem 1 in [BY05], we have

$$\int_0^\infty e^{\beta K_s} ds < \infty, \qquad \mathbb{P}^{(1)}\text{-a.s.}$$

Hence, we only need to prove

$$a_F(t) \quad \sim \quad C_F \mathbb{E}\Big[\Big(\int_0^t e^{-\beta K_s} ds\Big)^{-1/\beta}\Big], \quad \text{as} \quad t \to \infty. \tag{2.3.21}$$

Recall that $\theta_{max} > 1$ and $\phi'_K(1) < 0$. So we can choose $\varepsilon > 0$ such that (2.3.9) holds, $1 + \varepsilon < \theta_{max}$, $\phi_K(1+\varepsilon) < \phi_K(1)$ and $\phi'_K(1+\varepsilon) < 0$. Therefore

$$\left| F\left(\int_0^t e^{-\beta K_s} ds\right) - C_F \left(\int_0^t e^{-\beta K_s} ds\right)^{-1/\beta} \right| \le M \left(\int_0^t e^{-\beta K_s} ds\right)^{-(1+\varepsilon)/\beta}.$$

In other words, it is enough to show

$$\mathbb{E}\left[\left(\int_0^t e^{-\beta K_s} ds\right)^{-(1+\varepsilon)/\beta}\right] = o(e^{t\phi_K(1)}), \quad \text{as} \quad t \to \infty.$$

From the Escheer transform (2.3.13), with $\lambda = 1 + \varepsilon$, we deduce

$$\begin{aligned} \mathbb{E}\left[\left(\int_0^t e^{-\beta K_s} ds\right)^{-(1+\varepsilon)/\beta}\right] &= \mathbb{E}\left[e^{(1+\varepsilon)K_t}\left(\int_0^t e^{\beta K_s} ds\right)^{-(1+\varepsilon)/\beta}\right] \\ &= e^{t\phi_K(1+\varepsilon)} \mathbb{E}^{(1+\varepsilon)}\left[\left(\int_0^t e^{\beta K_s} ds\right)^{-(1+\varepsilon)/\beta}\right]. \end{aligned}$$

Again from Lemma 1, we obtain for $t \ge q/2$,

$$\mathbb{E}\left[\left(\int_0^t e^{-\beta K_s} ds\right)^{-\frac{1+\varepsilon}{\beta}}\right] \le q^{(1+\varepsilon)/\beta} \exp\left(\frac{\phi_K(1+\varepsilon) + |\gamma|(1+\varepsilon) + \phi_K^+(1+\varepsilon)}{q}\right) < \infty,$$

implying that the decreasing function $t \mapsto \mathbb{E}^{(1+\varepsilon)}[(\int_0^t \exp(\beta K_s) ds)^{-(1+\varepsilon)/\beta}]$ is finite for all $t > 0$. This completes the proof. □

Remark 3. *In the particular case when $\beta = 1$, it is enough to apply Theorem 1.1 in [GKV03] to the geometric BPRE $(X_n, n \ge 0)$ defined in Remark 2 to get the result.*

Proof of Theorem 2 a/ (ii), (iii), and b/. The proofs are very similar for the three regimes, for this reason we only focus on the proof of the regime in a/(iii).

Let $\varepsilon > 0$. Thanks to Lemma 7, we can choose $q \in \mathbb{N}^*$ such that $q \ge 1/\varepsilon$ and $(1-\varepsilon)c_3 \le c_3(q) \le (1+\varepsilon)c_3$. Then for every $t \ge 1$, the monotonicty of F yields

$$\mathbb{E}\Big[F(C_{\lfloor qt \rfloor, q} e^{\beta|\gamma|/q}/q)\Big] \le a_F(t) \le \mathbb{E}\Big[F(D_{\lfloor qt \rfloor - 1, q} e^{-\beta|\gamma|/q}/q)\Big].$$

Applying (2.3.10), we obtain :

$$\left|\mathbb{E}\left[F(C_{\lfloor qt\rfloor,q}e^{\beta|\gamma|/q}/q)\right]-F_{\lfloor qt\rfloor,q}\right|\leq(1-e^{-\varepsilon(|\gamma|-\phi_K^-(1))})M\mathbb{E}\left[(A_{\lfloor qt\rfloor,q}/q)^{-1/\beta}\right],$$
$$\left|\mathbb{E}\left[F(D_{\lfloor qt\rfloor-1,q}e^{-\beta|\gamma|/q}/q)\right]-F_{\lfloor qt-1\rfloor,q}\right|\leq(e^{\varepsilon(|\gamma|+\phi_K^+(1))}-1)M\mathbb{E}\left[(A_{\lfloor qt\rfloor-1,q}/q)^{-1/\beta}\right].$$

Taking t to infinity, it is clear from Lemma 4 that both terms are bounded by

$$l(\varepsilon)t^{-3/2}e^{t\phi_K(\tau)}=\left[2Md(e^{\varepsilon(|\gamma|+\phi_K^+(1))}-e^{-\varepsilon(|\gamma|-\phi_K^-(1))})e^{-\varepsilon\phi_K(\tau)}\right]t^{-3/2}e^{t\phi_K(\tau)} \tag{2.3.22}$$

where ϕ_K^- and ϕ_K^+ are defined in (2.3.2), and $l(\varepsilon)$ goes to 0 when ε decreases. On the other hand, for t large enough

$$(1-2\varepsilon)c_3t^{-3/2}e^{t\phi_K(\tau)}\leq F_{\lfloor qt\rfloor,q}\leq a_F(t)\leq F_{\lfloor qt\rfloor-1,q}\leq(1+2\varepsilon)c_3t^{-3/2}e^{t\phi_K(\tau)},$$

which completes the proof of Theorem 2. □

2.4 Application to a cell division model

When the reproduction law has a finite second moment, the scaling limit of the GW process is a Feller diffusion with growth g and diffusion part σ^2. That is to say, the stable case with $\beta=1$ and additional drift term g. Such a process is also the scaling limit of birth and death processes. It gives a natural model for populations which die and multiply fast, randomly, without interaction. Such a model is considered in [BT11] for parasites growing in dividing cells. The cell divides at constant rate r and a random fraction $\Theta\in(0,1)$ of parasites enters the first daughter cell, whereas the remainder enters the second daughter cell. Following the infection in a cell line, the parasites grow as a Feller diffusion process and undergo a catastrophe when the cell divides. We denote by N_t and N_t^* the numbers of cells and infected cells at time t, respectively. We say that the cell population recovers when the asymptotic proportion of contaminated cells vanishes. If there is one infected cell at time 0, $\mathbb{E}[N_t]=e^{rt}$ and $\mathbb{E}[N_t^*]=e^{rt}\mathbb{P}(Y_t>0)$, where

$$Y_t=1+\int_0^t gY_sds+\int_0^t\sqrt{2\sigma^2Y_s}dB_s+\int_0^t\int_0^1(\theta-1)Y_{s^-}\rho(ds,d\theta). \tag{2.4.1}$$

Here B is a Brownian motion and $\rho(ds,d\theta)$ a Poisson random measure with intensity $2rds\mathbb{P}(\Theta\in d\theta)$. Note that the intensity of ρ is twice the cell division rate. This bias follows from the fact that if we pick an individual at random at time t, we are more likely to choose a lineage in which many division events have occurred. Hence the ancestral lineages from typical individuals at time t have a division rate $2r$.

Corollary 2.1 and Proposition 4 with $\beta=1$, $\psi(\lambda)=-g\lambda+\sigma^2\lambda$ and $\nu(dx)=2r\mathbb{P}(\Theta\in dx)$ imply the following result.

Corollary 2.4. *a/ We assume that* $g<2r\mathbb{E}\left[\log(1/\Theta)\right]$. *Then there exist positive constants* c_1,c_2,c_3 *such that*

(i) If $g < 2r\mathbb{E}\left[\Theta\log(1/\Theta)\right]$, *then*

$$\mathbb{E}\left[N_t^*\right] \sim c_1 e^{gt}, \quad as \quad t \to \infty.$$

(ii) If $g = 2r\mathbb{E}\left[\Theta\log(1/\Theta)\right]$, *then*

$$\mathbb{E}\left[N_t^*\right] \sim c_2 t^{-1/2} e^{gt}, \quad as \quad t \to \infty.$$

(iii) If $g > 2r\mathbb{E}\left[\Theta\log(1/\Theta)\right]$, *then*

$$\mathbb{E}\left[N_t^*\right] \sim c_3 t^{-3/2} e^{\alpha t}, \quad as \quad t \to \infty.$$

where $\alpha = \min_{\lambda\in[0,1]}\{g\lambda + 2r(\mathbb{E}[\Theta^\lambda] - 1/2)\} < g$.

b/ We now assume $g = 2r\mathbb{E}\left[\log(1/\Theta)\right]$, *then there exists* $c_4 > 0$ *such that,*

$$\mathbb{E}\left[N_t^*\right] \sim c_4 t^{-1/2} e^{rt}, \quad as \quad t \to \infty.$$

c/ Finally, if $g > 2r\mathbb{E}\left[\log(1/\Theta)\right]$, *then there exists* $0 < c_5 < 1$ *such that,*

$$\mathbb{E}\left[N_t^*\right] \sim c_5 e^{rt}, \quad as \quad t \to \infty.$$

Hence if $g > 2r\mathbb{E}\left[\log(1/\Theta)\right]$ (supercritical case c/), the mean number of infected cells is equivalent to the mean number of cells. In the critical case (b/), there are somewhat fewer infected cells, owing to the additional square root term. In the strongly subcritical regime (a/(i)), the mean number of infected cells is of the same order as the number of parasites. This suggests that parasites do not accumulate in some infected cells. The asymptotic behavior in the two remaining cases is more complex.

We stress the fact that fixing the growth rate g of parasites and the cell division rate r, but making the law of the repartition Θ vary, it changes the asymptotic behavior of the number of infected cells. For example, if we focus on random variables Θ satisfying $\mathbb{P}(\Theta = \theta) = \mathbb{P}(\Theta = 1-\theta) = 1/2$ for a given $\theta \in]0, 1/2[$, the different regimes can be described easily (see Figure 2.1).

If $g/r > \log 2$, the cell population either recovers or not, depending on the asymmetry of the parasite sharing. If $g/r \leq \log 2/2$, the cell population recovers but the speed of recovery increases with respect to the asymmetry of the parasite sharing, as soon as the weakly subcritical regime is reached. Such phenomena were known in the discrete time, discrete space framework (see [Ban08]), but the boundaries between the regimes are not the same, due to the bias in division rate in the continuous setting. Moreover, we note that if $g/r \in (\log 2/2, \log 2)$, then parasites are in the weakly subcritical regime whatever the distribution of Θ on $]0,1[$. This phenomenon also only occurs in the continuous setting.

A Auxiliary results

This section is devoted to the technical results which are necessary for the previous proofs.

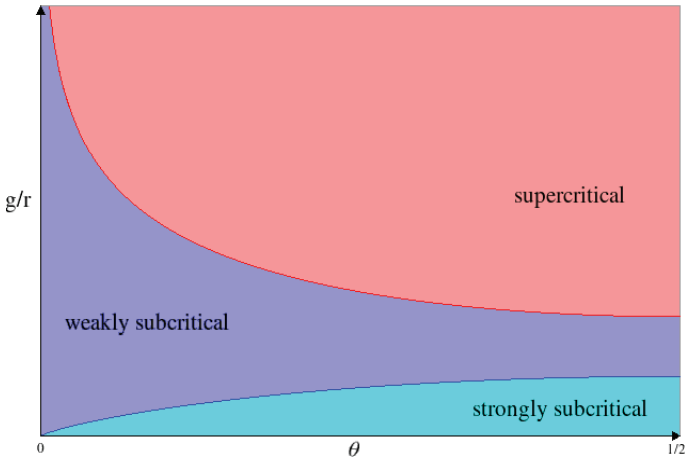

Figure 2.1: Extinction regimes in the case $\mathbb{P}(\Theta=\theta)=\mathbb{P}(\Theta=1-\theta)=1/2$. Boundaries between the different regimes are given by $g/r=-\log(\theta(1-\theta))$ (supercritical and subcritical) and $g/r=-\theta\log\theta-(1-\theta)\log(1-\theta)$ (strongly and weakly subcritical).

Existence and uniqueness of the backward ordinary differential equation

The Laplace exponent of $\widetilde{Z}$ in Theorem 1 is the solution of a backward ODE. The existence and uniqueness of this latter are stated and proved below.

Proposition 5. *Let δ be in $\mathscr{BV}(\mathbb{R}^+)$. Then the backward ordinary differential equation (2.1.5) admits a unique solution.*

The proof relies on a classical approximation of the solution of (2.1.5) and the Cauchy-Lipschitz Theorem. When there is no accumulation of jumps, the latter provides the existence and uniqueness of the solution between two successive jump times of δ. The problem remains on the times where accumulation of jumps occurs. Let us define the family of functions δ^n by deleting the small jumps of δ,

$$\delta^n_t=\delta_t-\sum_{s\le t}\Big(\delta_s-\delta_{s-}\Big)\mathbf{1}_{\{|\delta_s-\delta_{s-}|<1/n\}}.$$

We note that ψ_0 is continuous, and $s\mapsto e^{gs+\delta^n_s}$ is piecewise C^1 on $\mathbb{R}^+$ with a finite number of discontinuities. From the Cauchy-Lipschitz Theorem, for every $n\in\mathbb{N}^*$ we can define a solution $v^n_t(.,\lambda,\delta)$ continuous with càdlàg first derivative of the backward differential equation :

$$\frac{\partial}{\partial s}v^n_t(s,\lambda,\delta)=e^{gs+\delta^n_s}\psi_0\big(e^{-gs-\delta^n_s}v^n_t(s,\lambda,\delta)\big),\quad 0\le s\le t,\qquad v^n_t(t,\lambda,\delta)=\lambda.$$

We want to show that the sequence $(v^n_t(.,\lambda,\delta))_{n\ge1}$ converges to a function $v_t(.,\lambda,\delta)$ which is solution of (2.1.5). This follows from the next result. We fix $t>0$ and define

$$S:=\sup_{s\in[0,t],n\in\mathbb{N}^*}\Big\{e^{gs+\delta^n_s},e^{-gs-\delta^n_s}\Big\}. \tag{A.1}$$

Lemma 8. *For every $\lambda > 0$, there exists a positive finite constant C such that for all $0 \le \eta \le \kappa \le \lambda S$,*

$$0 \le \psi_0(\kappa) - \psi_0(\eta) \le C(\kappa - \eta). \tag{A.2}$$

Démonstration. First, we observe that S is finite and that for all $0 \le \eta < \kappa \le \lambda S$, we have $0 \le e^{-\kappa x} - e^{-\eta x} + (\kappa - \eta)x \le (\kappa - \eta)x$ for $x \ge 0$ since $x \mapsto e^{-x} + x$ is increasing and $e^{-\kappa x} \le e^{-\eta x}$. Moreover

$$0 \le e^{-x} - 1 + x \le x \wedge x^2, \tag{A.3}$$

and combining these inequalities yields

$$\begin{aligned}
&\psi_0(\kappa) - \psi_0(\eta) \\
&\quad = \sigma^2(\kappa^2 - \eta^2) + \int_1^\infty \Big(e^{-\kappa x} - e^{-\eta x} + (\kappa - \eta)x\Big)\mu(dx) \\
&\qquad + (\kappa - \eta)\int_0^1 x(1 - e^{-\eta x})\mu(dx) + \int_0^1 \Big(e^{-(\kappa-\eta)x} - 1 + (\kappa - \eta)x\Big)e^{-\eta x}\mu(dx) \\
&\quad \le \sigma^2(\kappa^2 - \eta^2) + (\kappa - \eta)\int_1^\infty x\mu(dx) + (\kappa - \eta)\eta\int_0^1 x^2\mu(dx) + (\kappa - \eta)^2\int_0^1 x^2\mu(dx) \\
&\quad \le \Big[2\lambda S\sigma^2 + \int_1^\infty x\mu(dx) + \lambda S\int_0^1 x^2\mu(dx)\Big](\kappa - \eta),
\end{aligned}$$

which proves Lemma 8. □

Next, we prove the existence and uniqueness result.

Proof of Proposition 5. We now prove that $(v_t^n(s,\lambda,\delta), s \in [0,t])_{n\ge 0}$ is a Cauchy sequence. For simplicity, we denote $v^n(s) = v_t^n(s,\lambda,\delta)$, and for all $v \ge 0$:

$$\psi^n(s,v) = e^{gs+\delta_s^n}\psi_0\big(e^{-gs-\delta_s^n}v\big) \quad \text{and} \quad \psi^\infty(s,v) = e^{gs+\delta_s}\psi_0\big(e^{-gs-\delta_s}v\big).$$

We have for any $0 \le s \le t$ and $m, n \ge 1$:

$$\begin{aligned}
|v^n(s) - v^m(s)| &= \Big|\int_s^t \psi^n(u, v^n(u))du - \int_s^t \psi^m(u, v^m(u))du\Big| \\
&\le \int_s^t (R^n(u) + R^m(u))du + \int_s^t \Big|\psi^\infty(u, v^n(u)) - \psi^\infty(u, v^m(u))\Big|du,
\end{aligned} \tag{A.4}$$

where for any $u \in [0,t]$,

$$\begin{aligned}
R^n(u) &:= \Big|\psi^n(u, v^n(u)) - \psi^\infty(u, v^n(u))\Big| \\
&\le e^{gu+\delta_u^n}\Big|\psi_0\big(e^{-gu-\delta_u^n}v^n(u)\big) - \psi_0\big(e^{-gu-\delta_u}v^n(u)\big)\Big| + e^{gu}\psi_0\big(e^{-gu-\delta_u}v^n(u)\big)\Big|e^{\delta_u^n} - e^{\delta_u}\Big|.
\end{aligned}$$

Moreover, from (A.1) to (A.2), we obtain

$$\begin{aligned}
R^n(u) &\le SC\lambda\Big|e^{-\delta_u^n} - e^{-\delta_u}\Big| + e^{|g|t}\psi_0(\lambda S)\Big|e^{\delta_u^n} - e^{\delta_u}\Big| \\
&\le \Big(SC\lambda + e^{|g|t}\psi_0(\lambda S)\Big)\sup_{u\in[0,t]}\Big\{\Big|e^{-\delta_u^n} - e^{-\delta_u}\Big|, \Big|e^{\delta_u^n} - e^{\delta_u}\Big|\Big\} := s_n.
\end{aligned}$$

Using similar arguments as above, we get from (A.2),

$$\left|\psi^\infty(u,v^n(u))-\psi^\infty(u,v^m(u))\right| \le CS^2\left|v^n(u)-v^m(u)\right|.$$

From (A.4), we use Gronwall's Lemma (see e.g. Lemma 3.2 in [Dyn91]) with

$$R_{m,n}(s)=\int_s^t R^n(u)du+\int_s^t R^m(u)du,$$

to deduce that for all $0\le s\le t$,

$$|v^n(s)-v^m(s)| \le R_{m,n}(s)+CS^2e^{CS^2(t-s)}\int_s^t R_{m,n}(u)du.$$

Recalling that $R_n(u)\le s_n$ and $\int_s^t R^n(u)du\le ts_n$ for $u\le t$, we get for every $n_0\in\mathbb{N}^*$,

$$\sup_{m,n\ge n_0,s\in[0,t]}|v^n(s)-v^m(s)| \le t\left[1+CS^2e^{CS^2t}t\right]\sup_{m,n\ge n_0}(s_n+s_m).$$

Adding that $s_n\to 0$ ensures that $(v^n(s),s\in[0,t])_{n\ge0}$ is a Cauchy sequence under the uniform norm. Then there exists a continuous function v on $[0,t]$ such that $v^n\to v$, as n goes to ∞.

Next, we prove that v is solution of the Equation (2.1.5). As δ satisfies (A.1), we have for any $s\in[0,t]$ and $n\in\mathbb{N}^*$:

$$\begin{aligned}
&\left|v(s)-\int_s^t\psi^\infty(s,v(s))ds-\lambda\right|\\
&\le\left|v(s)-v^n(s)\right|+\int_s^t\left|\psi^\infty(s,v(s))-\psi^n(s,v(s))\right|ds+\int_s^t\left|\psi^n(s,v(s))-\psi^n(s,v^n(s))\right|ds\\
&\le ts_n+(1+CS^2)\sup\left\{\left|v(s)-v^n(s)\right|,s\in[0,t]\right\},
\end{aligned}$$

so that letting $n\to\infty$ yields $\left|v(s)-\int_s^t\psi^\infty(s,v(s))ds-\lambda\right|=0$. It proves that v is solution of (2.1.5). The uniqueness follows from Gronwall's lemma. □

An upper bound for ψ_0

The study of the Laplace exponent of $\widetilde{Z}$ in Corollary 2.1 requires a fine control of the branching mechanism ψ_0.

Lemma 9. *Assume that the process $(gt+\Delta_t, t\ge0)$ goes to $+\infty$ a.s. There exists a non-negative increasing function k on $\mathbb{R}^+$ such that for every $\lambda\ge0$*

$$\psi_0(\lambda)\le\lambda k(\lambda) \qquad \text{and} \qquad \int_0^\infty k\left(e^{-(gt+\Delta_t)}\right)dt<\infty.$$

Démonstration. The inequality (A.3) implies that for every $\lambda\ge0$,

$$\begin{aligned}
\psi_0(\lambda) &\le \sigma^2\lambda^2+\int_0^\infty\left(\lambda^2z^2\mathbf{1}_{\{\lambda z\le1\}}+\lambda z\mathbf{1}_{\{\lambda z>1\}}\right)\mu(dz)\\
&\le \left(\sigma^2+\int_0^1 z^2\mu(dz)\right)\lambda^2+\lambda^2\mathbf{1}_{\{\lambda<1\}}\int_1^{1/\lambda}z^2\mu(dz)+\lambda\int_{1/\lambda}^\infty z\mu(dz).
\end{aligned}$$

Now, using condition (2.1.10) we obtain the existence of a positive constant c such that

$$\lambda \int_{1/\lambda}^{\infty} z\mu(dz) \leq \lambda \log^{-(1+\varepsilon)}(1+1/\lambda) \int_{1/\lambda}^{\infty} z\log^{1+\varepsilon}(1+z)\mu(dz) \leq c\lambda \log^{-(1+\varepsilon)}(1+1/\lambda).$$

Next, let us introduce the function f, given by

$$f(z) = z^{-1}\log^{1+\varepsilon}(1+z), \quad \text{for } z \in [1,\infty).$$

If we derivate the function f, we deduce that there exists a positive real number $A > 1$ such that f is decreasing on $[A,\infty)$. Therefore, for every $\lambda < 1/A$,

$$\begin{aligned}\int_A^{1/\lambda} \lambda^2 z^2 \mu(dz) &= \lambda \log^{-(1+\varepsilon)}(1+1/\lambda)\, f(1/\lambda) \int_A^{1/\lambda} \frac{z\log^{1+\varepsilon}(1+z)}{f(z)}\mu(dz) \\ &\leq \lambda \log^{-(1+\varepsilon)}(1+1/\lambda) \int_A^{1/\lambda} z\log^{1+\varepsilon}(1+z)\mu(dz).\end{aligned}$$

Adding that $\lambda^2 \int_1^A z^2\mu(dz) \leq \lambda^2 A \int_1^{\infty} z\mu(dz)$ and using again condition (2.1.10), we deduce that there exists a positive constant c' such that for every $\lambda \geq 0$,

$$\psi_0(\lambda) \leq c'\Big(\lambda^2 + \lambda \log^{-(1+\varepsilon)}(1+1/\lambda)\Big).$$

Since λ^2 is negligible with respect to $\lambda\log^{-(1+\varepsilon)}(1+1/\lambda)$ when λ is close enough to 0 or infinity, we conclude that there exists a positive constant c'' such that

$$\psi_0(\lambda) \leq c''\lambda \log^{-(1+\varepsilon)}(1+1/\lambda).$$

Defining the function $k(z) = c''\log^{-(1+\varepsilon)}(1+1/z)$, for $z > 0$, we get that :

$$k\Big(e^{-(gt+\Delta_t)}\Big) \sim c''\log^{-(1+\varepsilon)}(2), \quad (t \to 0),$$

thus the integral of $k(\exp(-gt-\Delta_t))$ is finite in a neighborhood of zero, and

$$0 \leq \int_1^{\infty} k\Big(e^{-(gt+\Delta_t)}\Big)dt \leq c'' \int_1^{\infty} e^{-(gt+\Delta_t)}(gt+\Delta_t)^{-(1+\varepsilon)}dt,$$

which is finite since the process $(gt+\Delta_t, t \geq 0)$ drifts $+\infty$ and has finite first moment. This completes the proof. □

Extinction versus explosion

We now verify that the process $(Y_t)_{t\geq 0}$ can be properly renormalized as $t \to \infty$ on the non-extinction event. We use a classical branching argument.

Lemma 10. *Let Y be a non-negative Markov process satisfying the branching property. We also assume that there exists a positive function a_t such that for every $x_0 > 0$, there exists a non-negative finite random variable W such that*

$$a_t Y_t \xrightarrow[t\to\infty]{} W \quad a.s., \qquad \mathbb{P}_{x_0}(W > 0) > 0, \qquad a_t \overset{t\to\infty}{\longrightarrow} 0.$$

Then

$$\{W = 0\} = \left\{ Y_t \xrightarrow[t\to\infty]{} 0 \right\} \qquad \mathbb{P}_{x_0} \quad a.s.$$

Démonstration. First, we prove that

$$\mathbb{P}_{x_0}(\limsup_{t\to\infty} Y_t = \infty \mid \limsup_{t\to\infty} Y_t > 0) = 1. \tag{A.5}$$

Let $0 < x \le x_0 \le A$ be fixed. Since $a_t \to 0$ and $\mathbb{P}_x(W > 0) > 0$, there exists $t_0 > 0$ such that $\alpha := \mathbb{P}_x(Y_{t_0} \ge A) > 0$. By the branching property, the process is stochastically monotone as a function of its initial value. Thus, for every $y \ge x$ (including $y = x_0$),

$$\mathbb{P}_y(Y_{t_0} \ge A) \ge \alpha > 0.$$

We define recursively the stopping times

$$T_0 := 0, \qquad T_{i+1} = \inf\{t \ge T_i + t_0 : Y_t \ge x\} \qquad (i \ge 0).$$

For any $i \in \mathbb{N}^*$, the strong Markov property implies

$$\mathbb{P}_{x_0}(Y_{T_i+t_0} \ge A \mid (Y_t : t \le T_i),\ T_i < \infty) \ge \alpha.$$

Conditionally on $\{\limsup_{t\to\infty} Y_t > x\}$, the stopping times T_i are finite a.s. and for all $0 < x \le x_0 \le A$,

$$\mathbb{P}_{x_0}(\forall i \ge 0 : Y_{T_i+t_0} < A,\ \limsup_{t\to\infty} Y_t > x) = 0.$$

Then, $\mathbb{P}_{x_0}(\limsup_{t\to\infty} Y_t < \infty,\ \limsup_{t\to\infty} Y_t > x) = 0$. Now since $\{\limsup_{t\to\infty} Y_t > 0\} = \cup_{x\in(0,x_0]}\{\limsup_{t\to\infty} Y_t > x\}$, we get (A.5).

Next, we consider the stopping times $T_n = \inf\{t \ge 0 : Y_t \ge n\}$. The strong Markov property and branching property imply

$$\mathbb{P}_{x_0}(W = 0; T_n < \infty) = \mathbb{E}_{x_0}\left(\mathbf{1}_{T_n<\infty}\mathbb{P}_{Y_{T_n}}(W = 0)\right) \le \mathbb{P}_n(a_t Y_t \xrightarrow[t\to\infty]{} 0) = \mathbb{P}_1(a_t Y_t \xrightarrow[t\to\infty]{} 0)^n,$$

which goes to zero as $n \to \infty$, since $\mathbb{P}_1(a_t Y_t \overset{t\to\infty}{\longrightarrow} 0) = \mathbb{P}_1(W = 0) < 1$. Then,

$$0 = \mathbb{P}_{x_0}(W = 0; \forall n : T_n < \infty) = \mathbb{P}_{x_0}(W = 0, \limsup_{t\to\infty} Y_t = \infty) = \mathbb{P}_{x_0}(W = 0, \limsup_{t\to\infty} Y_t > 0),$$

where the last identity comes from (A.5). This completes the proof. □

A Central limit theorem

We need the following central limit theorem for Lévy processes in Corollary 2.2.

Lemma 11. *Under the assumption (2.1.11) we have*

$$\frac{gt+\Delta_t-\boldsymbol{m}t}{\rho\sqrt{t}} \xrightarrow[t\to\infty]{d} \mathcal{N}(0,1).$$

Démonstration. For simplicity, let η be the image measure of ν under the mapping $x \mapsto e^x$. Hence, assumption (2.1.11) is equivalent to $\int_{|x|\geq 1} x^2\eta(\mathrm{d}x) < \infty$, or $\mathbb{E}[\Delta_1^2] < \infty$.

We define $T(x) = \eta\big((-\infty,-x)\big) + \eta\big((x,\infty)\big)$ and $U(x) = 2\int_0^x yT(y)\mathrm{d}y$, and assume that $T(x) > 0$ for all $x > 0$. According to Theorem 3.5 in Doney and Maller [DM02] there exist two functions $a(t), b(t) > 0$ such that

$$\frac{gt+\Delta_t-a(t)}{b(t)} \xrightarrow[t\to\infty]{d} \mathcal{N}(0,1), \quad \text{if and only if} \quad \frac{U(x)}{x^2T(x)} \xrightarrow[x\to\infty]{} \infty.$$

If the above condition is satisfied, then b is regularly varying with index $1/2$ and it may be chosen to be strictly increasing to ∞ as $t\to\infty$. Moreover $b^2(t) = tU(b(t))$ and $a(t) = tA(b(t))$, where

$$A(x) = g + \int_{\{|z|<1\}} z\eta(\mathrm{d}z) + \eta\big((1,\infty)\big) - \eta\big((-\infty,-1)\big) + \int_1^x \Big(\eta\big((y,\infty)\big) - \eta\big((-\infty,-y)\big)\Big)\mathrm{d}y.$$

Note that under our assumption $x^2T(x)\to 0$, as $x\to\infty$. Moreover, note

$$U(x) = x^2T(x) + \int_{(-x,0)} z^2\eta(\mathrm{d}x) + \int_{(0,x)} z^2\eta(\mathrm{d}x),$$

and

$$A(x) = g + \int_{\{|z|<x\}} z\eta(\mathrm{d}z) + x\Big(\eta\big((x,\infty)\big) - \eta\big((-\infty,-x)\big)\Big).$$

Hence assumption (2.1.11) implies that

$$U(x) \xrightarrow[x\to\infty]{} \int_{(-\infty,\infty)} z^2\eta(\mathrm{d}z) = \rho^2, \quad A(x) \xrightarrow[x\to\infty]{} g + \int_{\mathbb{R}} z\eta(\mathrm{d}z) = \mathbf{m},$$

Therefore, we deduce $U(x)/(x^2T(x))\to\infty$ as $x\to\infty$, $b(t)\sim\rho\sqrt{t}$ and $a(t)\sim\mathbf{m}t$, as $t\to\infty$.

Now assume that $T(x) = 0$, for x large enough. Define

$$\Psi(\lambda,t) = -\log\mathbb{E}\left[\exp\left\{i\lambda\left(\frac{gt+\Delta_t-a(t)}{b(t)}\right)\right\}\right],$$

where the functions $a(t)$ and $b(t)$ are defined as above. Hence since the process $(\Delta_t, t\geq 0)$ is of bounded variation, from the definition of $a(t)$ and the Lévy-Khintchine formula we deduce

$$\begin{aligned}\Psi(\lambda,t) &= -i\lambda\left(\frac{gt}{b(t)} - \frac{a(t)}{b(t)}\right) + t\int_{\mathbb{R}}\Big(1-e^{\frac{i\lambda}{b(t)}x}\Big)\eta(\mathrm{d}x)\\ &= t\int_{\{|x|<b(t)\}}\Big(1-e^{\frac{i\lambda}{b(t)}x} + \frac{i\lambda}{b(t)}x + \frac{(i\lambda)^2}{2b^2(t)}x^2\Big)\eta(\mathrm{d}x) - \frac{t(i\lambda)^2}{2b^2(t)}\int_{\{|x|<b(t)\}} x^2\eta(\mathrm{d}x)\\ &\qquad + t\int_{\{|x|\geq b(t)\}}\Big(1-e^{\frac{i\lambda}{b(t)}x}\Big)\eta(\mathrm{d}x) + i\lambda t\Big(\eta(b(t),\infty) - \eta(-\infty,-b(t))\Big).\end{aligned}$$

Since $T(x) = 0$ for all x large, $b(t) \to \infty$ and $t^{-1}b^2(t) \to \rho^2$, as $t \to \infty$, therefore

$$\Psi(\lambda, t) \xrightarrow[t\to\infty]{} \frac{\lambda^2}{2},$$

which implies the result thanks to Lévy's Theorem. □

A technical Lemma

We now prove a technical lemma that is needed in the proofs of Section 2.3.

Proof of Lemma 2. To obtain (2.3.9), it is enough to choose $\varepsilon \le 1$ as we assume in (3.1.12) that $\varsigma \ge 1$.

In order to prove (2.3.10), we first define the function $\tilde{h} : x \in \mathbb{R}^+ \mapsto (1+x)^{1-\varsigma} h(x)$ and let $0 \le x \le y$. Then,

$$\begin{aligned} \frac{F(x) - F(y)}{C_F} \quad &\le \quad \left((x+1)^{-1/\beta} - (y+1)^{-1/\beta}\right) + (1+y)^{-1/\beta-1}\left|\tilde{h}(x) - \tilde{h}(y)\right| \\ &\quad + \left|\tilde{h}(x)\right|\left((1+x)^{-1/\beta-1} - (1+y)^{-1/\beta-1}\right). \end{aligned} \tag{A.6}$$

We deal with the second term of the right hand side. Denoting by k the Lipschitz constant of $\tilde{h}$ and applying the Mean Value Theorem to $z \in \mathbb{R}_+ \mapsto (z+1)^{-1/\beta}$ on $[x, y]$, we get

$$(1+y)^{-1/\beta-1}\left|\tilde{h}(x) - \tilde{h}(y)\right| \le k(y+1)^{-1/\beta-1}(y-x) \le k\beta\left((x+1)^{-1/\beta} - (y+1)^{-1/\beta}\right).$$

Moreover, as $\beta \in (0, 1]$, we have the following inequalities :

$$\left(\frac{1+y}{1+x}\right)^{1+1/\beta} - 1 \le \left(\left(\frac{1+y}{1+x}\right)^{1/\beta} - 1\right)\left(\frac{1+y}{1+x} - 1\right) \le \left(\left(\frac{y}{x}\right)^{1/\beta} - 1\right) 2\frac{1+y}{1+x}$$

Dividing by $(1+y)^{1/\beta+1}$ and using $(1+y)/[(1+x)(1+y)^{1/\beta+1}] \le y^{-1/\beta}$ yield

$$(1+x)^{-1/\beta-1} - (1+y)^{-1/\beta-1} \le 2\left(x^{-1/\beta} - y^{-1/\beta}\right).$$

Similarly $(1+x)^{-1/\beta} - (1+y)^{-1/\beta} \le x^{-1/\beta} - y^{-1/\beta}$ and equation (A.6) give us

$$0 \le F(x) - F(y) \quad \le \quad C_F(1 + 2[\|h\|_\infty + k\beta])\left(x^{-1/\beta} - y^{-1/\beta}\right).$$

This completes the proof. □

Approximations of the survival probability for $\nu(0,\infty) = \infty$

Finally, we prove Corollary 2.3 in the case when $\nu(0,\infty) = \infty$.

End of the proof of Corollary 2.3. We let $A^{\varepsilon_1,\varepsilon_2} = (0, 1-\varepsilon_1) \cup (1+\varepsilon_2, \infty)$, where $0 < 1-\varepsilon_1 < 1 < 1+\varepsilon_2$ and define the Poisson random measure $N_1^{\varepsilon_1,\varepsilon_2}$ as the restriction of N_1 to $\mathbb{R}^+ \times A^{\varepsilon_1,\varepsilon_2}$. We denote by $\mathrm{d}t\nu^{\varepsilon_1,\varepsilon_2}(\mathrm{d}m)$ for its intensity measure, where $\nu^{\varepsilon_1,\varepsilon_2}(\mathrm{d}m) = \mathbf{1}_{\{m\in A^{\varepsilon_1,\varepsilon_2}\}}\nu(\mathrm{d}m)$, and the corresponding Lévy process $\Delta^{\varepsilon_1,\varepsilon_2}$ is defined by

$$\Delta_t^{\varepsilon_1,\varepsilon_2} = \int_0^t \int_{(0,\infty)} \log m \; N_1^{\varepsilon_1,\varepsilon_2}(\mathrm{d}s, \mathrm{d}m).$$

We also consider the CSBP's $Y^{\varepsilon_1,\varepsilon_2}$ (resp $Y^{\varepsilon_1,\varepsilon_2,-}$ and $Y^{\varepsilon_1,\varepsilon_2,+}$) with branching mechanism ψ (resp. ψ_- and ψ_+) and the same catastrophes $\Delta^{\varepsilon_1,\varepsilon_2}$ via (2.1.3). Since $\nu^{\varepsilon_1,\varepsilon_2}(0,\infty) < \infty$, from the first step we have $u_{+,t}^{\varepsilon_1,\varepsilon_2}(\lambda) \le u^{\varepsilon_1,\varepsilon_2}(t,\lambda) \le u_{-,t}^{\varepsilon_1,\varepsilon_2}(\lambda)$, where as expected $\mathbb{E}[\exp\{-\lambda Y_t^{\varepsilon_1,\varepsilon_2,*}\}] = \exp\{-u_{*,t}^{\varepsilon_1,\varepsilon_2}(\lambda)\}$ for each $* \in \{+, \emptyset, -\}$.

Similarly, let $A^{\varepsilon_1} = (0, 1-\varepsilon_1) \cup (1, \infty)$ and define the Poisson random measure $N_1^{\varepsilon_1}$ as the restriction of N_1 to $\mathbb{R}^+ \times A^{\varepsilon_1}$ with intensity measure $\mathrm{d}t\nu^{\varepsilon_1}(\mathrm{d}m)$, where $\nu^{\varepsilon_1}(\mathrm{d}m) = \mathbf{1}_{\{m\in A^{\varepsilon_1}\}}\nu(\mathrm{d}m)$. Let us fix t in $\mathbb{R}_+^*$, and define Y^{ε_1} as the unique strong solution of

$$\begin{aligned} Y_t^{\varepsilon_1} = Y_0 + \int_0^t g Y_s^{\varepsilon_1}\mathrm{d}s + \int_0^t \sqrt{2\sigma^2 Y_s^{\varepsilon_1}}\mathrm{d}B_s + \int_0^t \int_{[0,\infty)} \int_0^{Y_{s-}^{\varepsilon_1}} z\tilde{N}_0(\mathrm{d}s, \mathrm{d}z, \mathrm{d}u) \\ + \int_0^t \int_{[0,\infty)} \Big(m-1\Big) Y_{s-}^{\varepsilon_1} N_1^{\varepsilon_1}(\mathrm{d}s, \mathrm{d}m). \end{aligned} \tag{A.7}$$

We already know from Theorem 1 that Equation (A.7) has a unique non-negative strong solution. Moreover, from Theorem 5.5 in [FL10] and the fact that $N_1^{\varepsilon_1}$ has the same jumps as $N_1^{\varepsilon_1,\varepsilon_2}$ plus additional jumps greater than one, we conclude

$$Y_t^{\varepsilon_1,\varepsilon_2} \le Y_t^{\varepsilon_1}, \qquad \text{a.s.}$$

Using assumption (2.1.2), we can apply Gronwall's Lemma to the non-negative function $t \mapsto \mathbb{E}[Y_t^{\varepsilon_1} - Y_t^{\varepsilon_1,\varepsilon_2}]$ and obtain

$$\mathbb{E}\Big[|Y_t^{\varepsilon_1,\varepsilon_2} - Y_t^{\varepsilon_1}|\Big] \xrightarrow[\varepsilon_2\to 0]{} 0.$$

Now, since $Y^{\varepsilon_1,\varepsilon_2}$ is decreasing with ε_2, we finally get, $Y_t^{\varepsilon_1,\varepsilon_2} \xrightarrow{a.s.} Y_t^{\varepsilon_1}$, as $\varepsilon_2 \to 0$. Using similar arguments as above for $Y^{\varepsilon_1,\varepsilon_2,+}$ and $Y^{\varepsilon_1,\varepsilon_2,-}$, we deduce

$$u_{+,t}^{\varepsilon_1}(\lambda) \le u^{\varepsilon_1}(t,\lambda) \le u_{-,t}^{\varepsilon_1,}(\lambda).$$

In order to complete the proof, we let ε_1 tend to 0. □

CHAPITRE 3

An Eco-Evolutionary approach of Adaptation and Recombination in a large population of varying size

Introduction

There are at least two different ways of adaptation for a population : selection can either act on a new mutation (hard selective sweep), or on preexisting alleles that become advantageous after an environmental change (soft selective sweep from standing variation). New mutations are sources of diversity, and hard selective sweeps were until recently the only considered way of adaptation. Soft selective sweeps from standing variation allow a faster adaptation to novel environments, and their importance is growing in empirical and theoretical studies (Orr and Betancourt [OB01], Hermisson and Pennings [HP05], Prezeworski, Coop and Wall [PCW05], Barrett and Schluter [BS08], Durand and al [DTR+10]). In particular Messer and Petrov [MP13] review a lot of evidence, from individual case studies as well as from genome-wide scans, that soft sweeps (from standing variation and from recurrent mutations) are common in a broad range of organisms. These distinct selective sweeps entail different genetic signatures in the vicinity of the novely fixed allele, and the multiplication of genetic data available allows one to detect these signatures in current populations as described by Peter, Huerta-Sanchez and Nielsen [PHSN12]. To do this in an effective way, it is necessary to identify accurately the signatures left by these two modes of adaptation. We will not consider in this work the soft selective sweeps from recurrent mutations. For a study of these sweeps we refer to [PH06a, PH06b, HP08].

In this work, we consider a sexual haploid population of varying size, modeled by a birth and death process with density dependent competition. Each individual's ability to survive and reproduce depends on its own genotype and on the population state. More precisely, each individual is characterized by some ecological parameters : birth rate, intrinsic death rate and competition kernel describing the competition with other individuals depending on their genotype. The differential reproductive success of individuals generated by their interactions entails

progressive variations in the number of individuals carrying a given genotype. This process, called natural selection, is a key mechanism of evolution. Such an eco-evolutionary approach has been introduced by Metz and coauthors in [MGM+96] and made rigorous in the seminal paper of Fournier and Méléard [FM04]. Then it has been developed by Champagnat, Méléard and coauthors (see [Cha06, CM11, CJM14] and references therein) for the haploid asexual case and by Collet, Méléard and Metz [CMM11] and Coron and coauthors [Cor13, Cor14] for the diploid sexual case. The recent work of Billiard and coauthors [BFMT13] studies the dynamics of a two-locus model in an haploid asexual population. Following these works, we introduce a parameter K called carrying capacity which scales the population size, and study the limit behavior for large K. But unlike them, we focus on two loci in a sexual haploid population and take into account recombinations : one locus is under selection and has two possible alleles A and a and the second one is neutral with allele b_1 or b_2. When two individuals give birth, either a recombination occurs with probability r_K and the newborn inherits one allele from each parent, or he is the clone of one parent.

We first focus on a soft selective sweep from standing variation occurring after a change in the environment (new pathogen, environmental catastrophe, occupation of a new ecological niche,...). We assume that before the change the alleles A and a were neutral and both represented a positive fraction of the population, and that in the new environment the allele a becomes favorable and goes to fixation. We can divide the selective sweep in two periods : a first one where the population process is well approximated by the solution of a deterministic dynamical system, and a second one where A-individuals are near extinction, the deterministic approximation fails and the fluctuations of the A-population size become predominant. We give the asymptotic value of the final neutral allele proportion as a function of the ecological parameters, recombination probability r_K and solutions of a two-dimensional competitive Lotka-Volterra system.

We then focus on hard selective sweeps. We assume that a mutant a appears in a monomorphic A-population at ecological equilibrium. As stated by Champagnat in [Cha06], the selective sweep is divided in three periods : during the first one, the resident population size stays near its equilibrium value, and the mutant population size grows until it reaches a non-negligible fraction of the total population size. The two other periods are the ones described for the soft selective sweep from standing variation. Moreover, the time needed for the mutant a to fix in the population is of order $\log K$. We prove that the distribution of neutral alleles at the end of the sweep has different shapes according to the order of the recombination probability per reproductive event r_K with respect to $1/\log K$. More precisely, we find two recombination regimes : a strong one where $r_K \log K$ is large, and a weak one where $r_K \log K$ is bounded. In both recombination regimes, we give the asymptotic value of the final neutral allele proportion as a function of the ecological parameters and recombination probability r_K. In the strong recombination regime, the frequent exchanges of neutral alleles between the A and a-populations yield an homogeneous neutral repartition in the two populations and the latter is not modified by the sweep. In the weak recombination regime, the frequency of the neutral allele carried by the first mutant increases because it is linked to the positively selected

allele. This phenomenon has been called genetic hitch-hiking by Maynard Smith and Haigh [SH74].

The first studies of hitch-hiking, initiated by Maynard Smith and Haigh [SH74], have modeled the mutant population size as the solution of a deterministic logistic equation [OK75, KHL89, SWL92, SSL06]. Kaplan and coauthors [KHL89] described the neutral genealogies by a structured coalescent where the background was the frequency of the beneficial allele. Barton [Bar98] was the first to point out the importance of the stochasticity of the mutant population size and the errors made by ignoring it. He divided the sweep in four periods : the two last ones are the analogues of the two last steps described in [Cha06], and the two first ones correspond to the first one in [Cha06]. Following the approaches of [KHL89] and [Bar98], a series of works studied the genealogies of neutral alleles sampled at the end of the sweep and took into account the randomness of the mutant population size during the sweep. In particular, Durrett and Schweinsberg [DS04, SD05], Etheridge and coauthors [EPW06], Pfaffelhuber and Studeny [PS07], and Leocard [Leo09] described the population process by a structured coalescent and finely studied genealogies of neutral alleles during the sweep. Eriksson and coauthors [EFMS08] described a deterministic approximation for the growth of the beneficial allele frequency during a sweep, which leads to more accurate approximation than previous models for large values of the recombination probability. Unlike our model, in all these works, the population size was constant and the individuals' "selective value" did not depend on the population state, but only on the individuals' genotype.

The structure of the paper is the following. In Section 3.1 we describe the model, review some results of [Cha06] about the two-dimensional population process when we do not consider the neutral locus, and present the main results. In Section 3.2 we state a semi-martingale decomposition of neutral proportions, a key tool in the different proofs. Section 3.3 is devoted to the proof for the soft sweep from standing variation. It relies on a comparison of the population process with a four dimensional dynamical system. In Section 3.4 we describe a coupling of the population process with two birth and death processes widely used in Sections 3.5 and 3.6, respectively devoted to the proofs for the strong and the weak recombination regimes of hard sweep. The proof for the weak regime requires a fine study of the genealogies in a structured coalescent process during the first phase of the selective sweep. We use here some ideas developed in [SD05]. Finally in the Appendix we state technical results.

This work stems from the papers of Champagnat [Cha06] and Schweinsberg and Durrett [SD05]. In the sequel, c is used to denote a positive finite constant. Its value can change from line to line but it is always independent of the integer K and the positive real number ε. The set $\mathbb{N} := \{1, 2, ...\}$ denotes the set of positive integers.

3.1 Model and main results

We introduce the sets $\mathscr{A} = \{A, a\}$, $\mathscr{B} = \{b_1, b_2\}$, and $\mathscr{E} = \{A, a\} \times \{b_1, b_2\}$ to describe the genetic background of individuals. The state of the population will be given by the four dimensional

Markov process $N^{(z,K)} = (N^{(z,K)}_{\alpha\beta}(t), (\alpha,\beta) \in \mathscr{E}, t \geq 0)$ where $N^{(z,K)}_{\alpha\beta}(t)$ denotes the number of individuals with alleles (α,β) at time t when the carrying capacity is $K \in \mathbb{N}$ and the initial state is $\lfloor zK \rfloor$ with $z = (z_{\alpha\beta}, (\alpha,\beta) \in \mathscr{E}) \in \mathbb{R}_+^{\mathscr{E}}$. We recall that b_1 and b_2 are neutral, thus ecological parameters only depend on the allele, A or a, carried by the individuals at their first locus. There are the following :

- For $\alpha \in \mathscr{A}$, f_α and D_α denote the birth rate and the intrinsic death rate of an individual carrying allele α.
- For $(\alpha,\alpha') \in \mathscr{A}^2$, $C_{\alpha,\alpha'}$ represents the competitive pressure felt by an individual carrying allele α from an individual carrying allele α'.
- $K \in \mathbb{N}$ is a parameter rescaling the competition between individuals. It can be interpreted as a scale of resources or area available, and is related to the concept of carrying capacity, which is the maximum population size that the environment can sustain indefinitely. In the sequel K will be large.
- r_K is the recombination probability per reproductive event. When two individuals with respective genotypes (α,β) and (α',β') in $\mathscr{E}$ give birth, the newborn individual, either is a clone of one parent and carries alleles (α,β) or (α',β') each with probability $(1-r_K)/2$, or has a mixed genotype (α,β') or (α',β) each with probability $r_K/2$.

We will use, for every $n = (n_{\alpha\beta}, (\alpha,\beta) \in \mathscr{E}) \in \mathbb{Z}_+^{\mathscr{E}}$, and $(\alpha,\beta) \in \mathscr{E}$, the notations

$$n_\alpha = n_{\alpha b_1} + n_{\alpha b_2} \quad \text{and} \quad |n| = n_A + n_a.$$

Let us now give the transition rates of $N^{(z,K)}$ when $N^{(z,K)}(t) = n \in \mathbb{Z}_+^{\mathscr{E}}$. An individual can die either from a natural death or from competition, whose strength depends on the carrying capacity K. Thus, the cumulative death rate of individuals $\alpha\beta$, with $(\alpha,\beta) \in \mathscr{E}$ is given by :

$$d^K_{\alpha\beta}(n) = \left[D_\alpha + C_{\alpha,A} n_A/K + C_{\alpha,a} n_a/K\right] n_{\alpha\beta}. \tag{3.1.1}$$

An individual carrying allele $\alpha \in \mathscr{A}$ produces gametes with rate f_α, thus the relative frequencies of gametes available for reproduction are

$$p_{\alpha\beta}(n) = f_\alpha n_{\alpha\beta}/(f_A n_A + f_a n_a), \quad (\alpha,\beta) \in \mathscr{E}.$$

When an individual gives birth, he chooses his mate uniformly among the gametes available. Then the probability of giving birth to an individual of a given genotype depends on the parents (the couple $((a,b_2),(a,b_1))$ is not able to generate an individual (A,b_1)). We detail the computation of the cumulative birth rate of individuals (A,b_1) :

$$\begin{aligned} b^K_{Ab_1}(n) &= f_A n_{Ab_1}[p_{Ab_1} + p_{Ab_2}/2 + p_{ab_1}/2 + (1-r_K)p_{ab_2}/2] + f_A n_{Ab_2}[p_{Ab_1}/2 + r_K p_{ab_1}/2] \\ &\quad + f_a n_{ab_1}[p_{Ab_1}/2 + r_K p_{Ab_2}/2] + f_a n_{ab_2}(1-r_K)p_{Ab_1}/2 \\ &= f_A n_{Ab_1} + r_K f_A f_a (n_{ab_1} n_{Ab_2} - n_{Ab_1} n_{ab_2})/(f_A n_A + f_a n_a). \end{aligned}$$

If we denote by $\bar{\alpha}$ (resp. $\bar{\beta}$) the complement of α in $\mathscr{A}$ (resp. β in $\mathscr{B}$), we obtain in the same way the cumulative birth rate of individuals (α,β) :

$$b^K_{\alpha\beta}(n) = f_\alpha n_{\alpha\beta} + r_K f_a f_A \frac{n_{\bar{\alpha}\beta} n_{\alpha\bar{\beta}} - n_{\alpha\beta} n_{\bar{\alpha}\bar{\beta}}}{f_A n_A + f_a n_a}, \quad (\alpha,\beta) \in \mathscr{E}. \tag{3.1.2}$$

The definitions of death and birth rates in (3.1.1) and (3.1.2) ensure that the number of jumps is finite on every finite interval, and the population process is well defined.

When we focus on the dynamics of traits under selection A and a, we get the process $(N_A^{(z,K)}, N_a^{(z,K)})$, which is also a birth and death process with competition. It has been studied in [Cha06] and its cumulative death and birth rates, which are direct consequences of (3.1.1) and (3.1.2), satisfy for $\alpha \in \mathscr{A}$:

$$d^K_\alpha(n) = \sum_{\beta\in\mathscr{B}} d^K_{\alpha\beta}(n) = \left[D_\alpha + C_{\alpha,A}\frac{n_A}{K} + C_{\alpha,a}\frac{n_a}{K}\right] n_\alpha, \quad b^K_\alpha(n) = \sum_{\beta\in\mathscr{B}} b^K_{\alpha\beta}(n) = f_\alpha n_\alpha. \tag{3.1.3}$$

It is proven in [Cha06] that when $N_A^{(z,K)}$ and $N_a^{(z,K)}$ are of order K, the rescaled population process $(N_A^{(z,K)}/K, N_a^{(z,K)}/K)$ is well approximated by the dynamical system :

$$\dot{n}^{(z)}_\alpha = (f_\alpha - D_\alpha - C_{\alpha,A} n^{(z)}_A - C_{\alpha,a} n^{(z)}_a) n^{(z)}_\alpha, \quad n^{(z)}_\alpha(0) = z_\alpha, \quad \alpha \in \mathscr{A}. \tag{3.1.4}$$

More precisely Theorem 3 (b) in [Cha06] states that for every compact subset

$$B \subset (\mathbb{R}_+^{A\times\mathscr{B}} \setminus (0,0)) \times (\mathbb{R}_+^{a\times\mathscr{B}} \setminus (0,0))$$

and finite real number T, we have for any $\delta > 0$,

$$\lim_{K\to\infty} \sup_{z\in B} \mathbb{P}\Big(\sup_{0\le t\le T, \alpha\in\mathscr{A}} |N^{(z,K)}_\alpha(t)/K - n^{(z)}_\alpha(t)| \ge \delta \Big) = 0. \tag{3.1.5}$$

Moreover, if we assume

$$f_A > D_A, \quad f_a > D_a, \quad \text{and} \quad f_a - D_a > (f_A - D_A).\sup\left\{C_{a,A}/C_{A,A}, C_{a,a}/C_{A,a}\right\}, \tag{3.1.6}$$

then the dynamical system (3.1.4) has a unique attracting equilibrium $(0,\bar{n}_a)$ for initial condition z satisfying $z_a > 0$, and two unstable steady states $(0,0)$ and $(\bar{n}_A, 0)$, where

$$\bar{n}_\alpha = \frac{f_\alpha - D_\alpha}{C_{\alpha,\alpha}} > 0, \quad \alpha \in \mathscr{A}. \tag{3.1.7}$$

Hence, Assumption (3.1.6) avoids the coexistence of alleles A and a, and $\bar{n}_\alpha$ is the equilibrium density of a monomorphic α-population per unit of carrying capacity. This implies that when K is large, the size of a monomorphic α-population stays near $\bar{n}_\alpha K$ for a long time (Theorem 3 (c) in [Cha06]). Moreover, if we introduce the invasion fitness $S_{\alpha\bar{\alpha}}$ of a mutant α in a population $\bar{\alpha}$,

$$S_{\alpha\bar{\alpha}} = f_\alpha - D_\alpha - C_{\alpha,\bar{\alpha}} \bar{n}_{\bar{\alpha}}, \quad \alpha \in \mathscr{A}, \tag{3.1.8}$$

it corresponds to the per capita growth rate of a mutant α when it appears in a population $\bar{\alpha}$ at its equilibrium density $\bar{n}_{\bar{\alpha}}$. Assumption (3.1.6) is equivalent to

Assumption 1. *Ecological parameters satisfy*

$$\bar{n}_A > 0, \quad \bar{n}_a > 0, \quad \textit{and} \quad S_{Aa} < 0 < S_{aA}.$$

Under Assumption 1, with positive probability, the A-population becomes extinct and the a-population size reaches a vicinity of its equilibrium value $\bar{n}_a K$.

The case we are interested in is referred in population genetics as soft selection [Wal75] : it is both frequency and density dependent. This kind of selection has no influence on the order of the total population size, which has the same order as the carrying capacity K. However, the factor multiplying the carrying capacity can be modified, as the way the individuals use the resources depends on the ecological parameters. We focus on strong selection coefficient, which are caracterized by $S_{aA} \gg 1/K$. In this case the selection outcompetes the genetic drift. However we do not need to assume $S_{aA} \ll 1$ to get approximations unlike [SH74, Bar98, SSL06]. To study the genealogy of the selected allele when the selection coefficient is weak ($S_{aA}K$ moderate or small) we refer to the approach of Neuhauser and Krone [NK97].

Let us now present the main results of this paper. We introduce the extinction time of the A-population, and the fixation event of the a-population. For $(z,K) \in \mathbb{R}_+^{\mathscr{E}} \times \mathbb{N}$:

$$T_{\text{ext}}^{(z,K)} := \inf\Big\{t \geq 0, N_A^{(z,K)}(t) = 0\Big\}, \quad \text{and} \quad \text{Fix}^{(z,K)} := \Big\{T_{\text{ext}}^{(z,K)} < \infty, N_a^{(z,K)}(T_{\text{ext}}^{(z,K)}) > 0\Big\}. \quad (3.1.9)$$

We are interested in the neutral allele proportions. We thus define for $t \geq 0$,

$$P_{\alpha,\beta}^{(z,K)}(t) = \frac{N_{\alpha\beta}^{(z,K)}(t)}{N_\alpha^{(z,K)}(t)}, \quad (\alpha,\beta) \in \mathscr{E}, K \in \mathbb{N}, z \in \mathbb{R}_+^{\mathscr{E}}, \quad (3.1.10)$$

the proportion of alleles β in the α-population at time t. More precisely, we are interested in these proportions at the end of the sweep, that is at time $T_{\text{ext}}^{(z,K)}$ when the last A-individual dies. We then introduce the neutral proportion at this time :

$$\mathscr{P}_{a,b_1}^{(z,K)} = P_{a,b_1}^{(z,K)}(T_{\text{ext}}^{(z,K)}). \quad (3.1.11)$$

We first focus on soft selective sweeps from standing variation. We assume that the alleles A and a were neutral and coexisted in a population with large carrying capacity K. At time 0, an environmental change makes the allele a favorable (in the sense of Assumption 1). Before stating the result, let us introduce the function F, defined for every $(z,r,t) \in (\mathbb{R}_+^{\mathscr{E}})^* \times [0,1] \times \mathbb{R}_+$ by

$$F(z,r,t) = \int_0^t \frac{r f_A f_a n_A^{(z)}(s)}{f_A n_A^{(z)}(s) + f_a n_a^{(z)}(s)} \exp\Big(-r f_A f_a \int_0^s \frac{n_A^{(z)}(u) + n_a^{(z)}(u)}{f_A n_A^{(z)}(u) + f_a n_a^{(z)}(u)} du\Big) ds, \quad (3.1.12)$$

where $(n_A^{(z)}, n_a^{(z)})$ is the solution of the dynamical system (3.1.4). We notice that $F : t \in \mathbb{R}^+ \mapsto F(z,r,t)$ is non-negative and non-decreasing. Moreover, if we introduce the function

$$h : (z,r,t) \in (\mathbb{R}_+^{\mathscr{E}})^* \times [0,1] \times \mathbb{R}_+ \mapsto r f_A f_a \int_0^t n_A^{(z)}(s)/(f_A n_A^{(z)}(s) + f_a n_a^{(z)}(s)) ds$$

non-decreasing in time, then

$$0 \le F(z,r,t) \le \int_0^t \partial_s h(z,r,s) e^{-h(z,r,s)}\, ds = e^{-h(z,r,0)} - e^{-h(z,r,t)} \le 1.$$

Thus $F(z,r,t)$ has a limit in $[0,1]$ when t goes to infinity and we can define

$$F(z,r) := \lim_{t\to\infty} F(z,r,t) \in [0,1]. \tag{3.1.13}$$

Noticing that for every $r \in [0,1]$ and $t \ge 0$,

$$0 \le F(z,r) - F(z,r,t) \le \int_t^\infty \frac{f_A f_a n_A^{(z)}(s)}{f_A n_A^{(z)}(s) + f_a n_a^{(z)}(s)} ds \underset{t\to\infty}{\to} 0,$$

we get that the convergence of $(F(z,r,t), t \ge 0)$ is uniform for $r \in [0,1]$.

In the case of a soft sweep from standing variation, the selected allele gets to fixation with high probability. More precisely, it is proven in [Cha06] that under Assumption 1,

$$\lim_{K\to\infty} \mathbb{P}(\mathrm{Fix}^{(z,K)}) = 1, \quad \forall z \in \mathbb{R}_+^{A\times\mathscr{B}} \times (\mathbb{R}_+^{a\times\mathscr{B}} \setminus (0,0)). \tag{3.1.14}$$

Then recalling (3.1.11) we get the following result whose proof is deferred to Section 3.3 :

Theorem 1. *Let z be in $\mathbb{R}_+^{A\times\mathscr{B}} \times (\mathbb{R}_+^{a\times\mathscr{B}} \setminus (0,0))$ and Assumption 1 hold. Then on the fixation event $\mathrm{Fix}^{(z,K)}$, the proportion of alleles b_1 when the A-population becomes extinct (time $T_{ext}^{(z,K)}$) converges in probability :*

$$\lim_{K\to\infty} \mathbb{P}\Big(\mathbf{1}_{\mathrm{Fix}^{(z,K)}} \Big| \mathscr{P}_{a,b_1}^{(z,K)} - \Big[\frac{z_{Ab_1}}{z_A} F(z,r_K) + \frac{z_{ab_1}}{z_a}(1 - F(z,r_K))\Big]\Big| > \varepsilon\Big) = 0, \quad \forall \varepsilon > 0.$$

The neutral proportion at the end of a soft sweep from standing variation is thus a weighted mean of initial proportions in populations A and a. In particular, a soft sweep from standing variation is responsible for a diminution of the number of neutral alleles with very low or very high proportions in the population, as remarked in [PCW05]. We notice that the weight $F(z,r_K)$ does not depend on the initial neutral proportions. It only depends on r_K and on the dynamical system (3.1.4) with initial condition (z_A, z_a). The proof consists in comparing the population process with the four dimensional dynamical system,

$$\dot{n}_{\alpha\beta}^{(z,K)} = \Big(f_\alpha - D_\alpha - C_{\alpha,A} n_A^{(z,K)} - C_{\alpha,a} n_a^{(z,K)}\Big) n_{\alpha\beta}^{(z,K)} + r f_A f_a \frac{n_{\bar\alpha\beta}^{(z,K)} n_{\alpha\bar\beta}^{(z,K)} - n_{\alpha\beta}^{(z,K)} n_{\bar\alpha\bar\beta}^{(z,K)}}{f_A n_A^{(z,K)} + f_a n_a^{(z,K)}}, \quad (\alpha,\beta) \in \mathscr{E}, \tag{3.1.15}$$

with initial condition $n^{(z,K)}(0) = z \in \mathbb{R}_+^{\mathscr{E}}$. Then by a change of variables, we can study the dynamical system (3.1.15) and prove that

$$\frac{n_{a,b_1}^{(z,K)}(\infty)}{n_a^{(z,K)}(\infty)} = \frac{z_{Ab_1}}{z_A} F(z,r_K) + \frac{z_{ab_1}}{z_a}(1 - F(z,r_K)),$$

which leads to the result.

Now we focus on hard selective sweeps : a mutant a appears in a large population and gets to fixation. We assume that the mutant appears when the A-population is at ecological equilibrium, and carries the neutral allele b_1. In other words, recalling Definition (3.1.7), we assume :

Assumption 2. *There exists* $z_{Ab_1} \in]0, \bar{n}_A[$ *such that* $N^{(z^{(K)},K)}(0) = \lfloor z^{(K)} K \rfloor$ *with*

$$z^{(K)} = (z_{Ab_1}, \bar{n}_A - z_{Ab_1}, K^{-1}, 0).$$

In this case, the selected allele gets to fixation with positive probability. More precisely, it is proven in [Cha06] that under Assumptions 1 and 2,

$$\lim_{K\to\infty} \mathbb{P}\Big(\mathrm{Fix}^{(z^{(K)},K)}\Big) = \frac{S_{aA}}{f_a}. \tag{3.1.16}$$

In the case of a strong selective sweep we will distinguish two different recombination regimes :

Assumption 3. *Strong recombination*

$$\lim_{K\to\infty} r_K \log K = \infty.$$

Assumption 4. *Weak recombination*

$$\limsup_{K\to\infty} r_K \log K < \infty.$$

Recall (3.1.11) and introduce the real number

$$\rho_K := 1 - \exp\Big(-\frac{f_a r_K \log K}{S_{aA}}\Big). \tag{3.1.17}$$

Then we have the following results whose proofs are deferred to Sections 3.5 and 3.6 :

Theorem 2. *Suppose that Assumptions 1 and 2 hold. Then on the fixation event* $\mathrm{Fix}^{(z^{(K)},K)}$ *and under Assumption 3 or 4, the proportion of alleles* b_1 *when the A-population becomes extinct (time* $T_{ext}^{(z^{(K)},K)}$*) converges in probability. More precisely, if Assumption 3 holds,*

$$\lim_{K\to\infty} \mathbb{P}\Big(\mathbf{1}_{\mathrm{Fix}^{(z^{(K)},K)}} \Big| \mathscr{P}_{a,b_1}^{(z^{(K)},K)} - \frac{z_{Ab_1}}{z_A}\Big| > \varepsilon\Big) = 0, \quad \forall \varepsilon > 0,$$

and if Assumption 4 holds,

$$\lim_{K\to\infty} \mathbb{P}\Big(\mathbf{1}_{\mathrm{Fix}^{(z^{(K)},K)}} \Big| \mathscr{P}_{a,b_1}^{(z^{(K)},K)} - \Big[(1-\rho_K) + \rho_K \frac{z_{Ab_1}}{z_A}\Big]\Big| > \varepsilon\Big) = 0, \quad \forall \varepsilon > 0.$$

As stated in [Cha06], the selective sweep has a duration of order $\log K$. Thus, when $r_K \log K$ is large, a lot of recombinations occur during the sweep, and the neutral alleles are constantly exchanged by the populations A and a. Hence in the strong recombination case, the sweep does not modifiy the neutral allele proportion. On the contrary, when r_K is of order $1/\log K$ the number of recombinations undergone by a given lineage does not go to infinity, and the frequency of the neutral allele b_1 carried by the first mutant a increases. More precisely, we will show that the probability for a neutral lineage to undergo a recombination and be descended from an individual of type A alive at the beginning of the sweep is close to ρ_K. Then to know the probability for such an allele to be a b_1 or a b_2, we have to approximate the proportion of alleles b_1 in the A-population when the recombination occurs. We will prove that this proportion stays close to the initial one z_{Ab_1}/z_A during the first phase. With probability $1-\rho_K$, a neutral allele originates from the first mutant. In this case it is necessarily a b_1. This gives the result for the weak recombination regime. In fact the probability for a neutral lineage to undergo no recombination during the first phase is quite intuitive : broadly speaking, the probability to have no recombination at a birth event is $1-r_K$, the birth rate is f_a and the duration of the first phase is $\log K/S_{aA}$. Hence as r_K is small for large K, $1-r_K \sim \exp(-r_K)$ and the probability to undergo no recombination is approximately

$$(1-r_K)^{f_a \log K/S_{aA}} \sim \exp(-r_K f_a \log K/S_{aA}) = 1-\rho_K.$$

Remark 4. *The limits in the two regimes are consistent in the sense that*

$$\lim_{r_K \log K \to \infty} \rho_K = 1.$$

Moreover, let us notice that we can easily extend the results of Theorems 1 and 2 to a finite number of possible alleles b_1, b_2, ..., b_i on the neutral locus.

Remark 5. *As it will appear in the proofs (see Sections 3.5 and 3.6), the final neutral proportion in the a population is already reached at the end of the first phase. In particular, the results are still valid if the sweep is not complete but the allele a only reaches a fraction $0 < p < 1$ of the population at the end of the sweep. The fact that the final neutral proportion is mostly determined by the beginning of the sweep has already been noticed by Coop and Ralph in [CR12].*

3.2 A semi-martingale decomposition

The expression of birth rate in (3.1.2) shows that the effect of recombination depends on the recombination probability r_K but also on the population state via the term $n_{\bar{\alpha}\beta}n_{\alpha\bar{\beta}} - n_{\alpha\beta}n_{\bar{\alpha}\bar{\beta}}$. Proposition 1 states a semi-martingale representation of the neutral allele proportions and makes this interplay more precise.

Proposition 1. *Let (α, z, K) be in $\mathscr{A} \times (\mathbb{R}_+^{\mathscr{E}})^* \times \mathbb{N}$. The process $(P^{(z,K)}_{\alpha,b_1}(t), t \geq 0)$ defined in (3.1.10) is a semi-martingale and we have the following decomposition :*

$$P^{(z,K)}_{\alpha,b_1}(t) = P^{(z,K)}_{\alpha,b_1}(0) + M^{(z,K)}_{\alpha}(t) + r_K f_A f_a \int_0^t \mathbf{1}_{\{N_\alpha(s)\geq 1\}} \frac{N^{(z,K)}_{\tilde{\alpha}b_1}(s) N^{(z,K)}_{\alpha b_2}(s) - N^{(z,K)}_{\alpha b_1}(s) N^{(z,K)}_{\tilde{\alpha}b_2}(s)}{(N^{(z,K)}_{\alpha}(s)+1)(f_A N^{(z,K)}_{A}(s) + f_a N^{(z,K)}_{a}(s))} ds, \quad (3.2.1)$$

where the process $(M^{(z,K)}_{\alpha}(t), t \geq 0)$ is a martingale bounded on every interval $[0,t]$ whose quadratic variation is given by (3.2.7).

To lighten the presentation in remarks and proofs we shall mostly write N instead of $N^{(z,K)}$.

Remark 6. *The process $N_{ab_2}N_{Ab_1} - N_{ab_1}N_{Ab_2}$ will play a major role in the dynamics of neutral proportions. Indeed it is a measure of the neutral proportion disequilibrium between the A and a-populations as it satisfies :*

$$N_A N_a (P_{A,b_1} - P_{a,b_1}) = N_{ab_2}N_{Ab_1} - N_{ab_1}N_{Ab_2}. \quad (3.2.2)$$

This quantity is linked with the linkage disequilibrium of the population, which is the occurrence of some allele combinations more or less often than would be expected from a random formation of haplotypes (see [Dur08] Section 3.3 for an introduction to this notion or [McV07] for a study of its structure around a sweep).

Remark 7. *By taking the expectations in Proposition 1 we can make a comparison with the results of Ohta and Kimura [OK75]. In their work the population size is infinite and the proportion of favorable allele $(y_t, t \geq 0)$ evolves as a deterministic logistic curve :*

$$\frac{dy_t}{dt} = s y_t (1 - y_t).$$

Moreover, x_1 and x_2 denote the neutral proportions of a given allele in the selected and non selected populations respectively, and are modeled by a diffusion. By making the analogies

$$N_e(t) = N_A(t) + N_a(t), \quad y_t = \frac{N_a(t)}{N_A(t)+N_a(t)}, \quad x_1(t) = \frac{N_{ab_1}(t)}{N_a(t)}, \quad x_2(t) = \frac{N_{Ab_1}(t)}{N_A(t)},$$

where N_e is the effective population size, the results of [OK75] can be written

$$\frac{d\mathbb{E}[P_{\alpha,b_1}(t)]}{dt} = r \frac{\mathbb{E}[N_{\tilde{\alpha}b_1}(t)N_{\alpha b_2}(t) - N_{\alpha b_1}(t)N_{\tilde{\alpha}b_2}(t)]}{N_\alpha(t)(N_A(t)+N_a(t))},$$

and

$$\frac{d\mathbb{E}[P^2_{\alpha,b_1}(t)]}{dt} = \frac{\mathbb{E}[P_{\alpha b_1}(t)(1-P_{\alpha b_2}(t))]}{2N_\alpha(t)} + 2r \frac{\mathbb{E}[N_{\alpha b_1}(t)(N_{\tilde{\alpha}b_1}(t)N_{\alpha b_2}(t) - N_{\alpha b_1}(t)N_{\tilde{\alpha}b_2}(t))]}{N^2_\alpha(t)(N_A(t)+N_a(t))}.$$

Hence the dynamics of the first moments are very similar to these that we obtain when we take equal birth rates $f_A = f_a$ and a recombination $r_K = r/f_a$. In contrast, the second moments of neutral proportions are very different in the two models.

Proof of Proposition 1. In the vein of Fournier and Méléard [FM04] we represent the population process in terms of Poisson measure. Let $Q(ds,d\theta)$ be a Poisson random measure on $\mathbb{R}_+^2$ with intensity $dsd\theta$, and $(e_{\alpha\beta},(\alpha,\beta)\in\mathscr{E})$ the canonical basis of $\mathbb{R}^{\mathscr{E}}$. According to (3.1.3) a jump occurs at rate

$$\sum_{(\alpha,\beta)\in\mathscr{E}}(b^K_{\alpha\beta}(N)+d^K_{\alpha\beta}(N))=f_aN_a+d^K_a(N)+f_AN_A+d^K_A(N).$$

We decompose on possible jumps that may occur : births and deaths for a-individuals and births and deaths for A-individuals. It's formula with jumps (see [IW89] p. 66) yields for every function h measurable and bounded on $\mathbb{R}_+^{\mathscr{E}}$:

$$\begin{aligned}h(N(t)) \;=\; & h(N(0))+\int_0^t\int_{R_+}\Big\{\sum_{\alpha\in\mathscr{A}}\Big(h(N(s^-)+e_{\alpha b_1})\mathbf{1}_{0<\theta-\mathbf{1}_{\alpha=A}(f_aN_a(s^-)+d^K_a(N(s^-))\le b^K_{\alpha b_1}(N(s^-))}\\ &+h(N(s^-)+e_{\alpha b_2})\mathbf{1}_{b^K_{\alpha b_1}(N(s^-))<\theta-\mathbf{1}_{\alpha=A}(f_aN_a(s^-)+d^K_a(N(s^-))\le f_\alpha N_\alpha(s^-)}\\ &+h(N(s^-)-e_{\alpha b_1})\mathbf{1}_{0<\theta-f_\alpha N_\alpha(s^-)-\mathbf{1}_{\alpha=A}(f_aN_a(s^-)+d^K_a(N(s^-))\le d^K_{\alpha b_1}(N(s^-))}\\ &+h(N(s^-)-e_{\alpha b_2})\mathbf{1}_{d^K_{\alpha b_1}(N(s^-))<\theta-f_\alpha N_\alpha(s^-)-\mathbf{1}_{\alpha=A}(f_aN_a(s^-)+d^K_a(N(s^-))\le d^K_\alpha(N(s^-))}\Big)\\ &-h(N(s^-))\mathbf{1}_{\theta\le f_aN_a(s^-)+d^K_a(N(s^-))+f_AN_A(s^-)+d^K_A(N(s^-))}\Big\}Q(ds,d\theta).\end{aligned}\tag{3.2.3}$$

Let us introduce the functions μ^α_K defined for $\alpha\in\mathscr{A}$ and (s,θ) in $\mathbb{R}_+\times\mathbb{R}_+$ by,

$$\begin{aligned}\mu^\alpha_K(N,s,\theta) \;=\; & \frac{\mathbf{1}_{N_\alpha(s)\ge1}N_{\alpha b_2}(s)}{(N_\alpha(s)+1)N_\alpha(s)}\mathbf{1}_{0<\theta-\mathbf{1}_{\alpha=A}(f_aN_a(s)+d^K_a(N(s))\le b^K_{\alpha b_1}(N(s))}\\ &-\frac{\mathbf{1}_{N_\alpha(s)\ge1}N_{\alpha b_1}(s)}{(N_\alpha(s)+1)N_\alpha(s)}\mathbf{1}_{b^K_{\alpha b_1}(N(s))<\theta-\mathbf{1}_{\alpha=A}(f_aN_a(s)+d^K_a(N(s))\le f_\alpha N_\alpha(s)}\\ &-\frac{\mathbf{1}_{N_\alpha(s)\ge2}N_{\alpha b_2}(s)}{(N_\alpha(s)-1)N_\alpha(s)}\mathbf{1}_{0<\theta-f_\alpha N_\alpha(s)-\mathbf{1}_{\alpha=A}(f_aN_a(s)+d^K_a(N(s))\le d^K_{\alpha b_1}(N(s))}\\ &+\frac{\mathbf{1}_{N_\alpha(s)\ge2}N_{\alpha b_1}(s)}{(N_\alpha(s)-1)N_\alpha(s)}\mathbf{1}_{d^K_{\alpha b_1}(N(s))<\theta-f_\alpha N_\alpha(s)-\mathbf{1}_{\alpha=A}(f_aN_a(s)+d^K_a(N(s))\le d^K_\alpha(N(s))}.\end{aligned}\tag{3.2.4}$$

Then we can represent the neutral allele proportions P_{α,b_1} as,

$$P_{\alpha,b_1}(t)=P_{\alpha,b_1}(0)+\int_0^t\int_0^\infty\mu^\alpha_K(N,s^-,\theta)Q(ds,d\theta),\quad t\ge0.\tag{3.2.5}$$

A direct calculation gives

$$\int_0^\infty\mu^\alpha_K(N,s,\theta)d\theta=r_Kf_Af_a\mathbf{1}_{\{N_\alpha(s)\ge1\}}\frac{N_{\tilde\alpha b_1}(s)N_{\alpha b_2}(s)-N_{\alpha b_1}(s)N_{\tilde\alpha b_2}(s)}{(N_\alpha(s)+1)(f_AN_A(s)+f_aN_a(s))}.$$

Thus if we introduce the compensated Poisson measure $\tilde Q(ds,d\theta):=Q(ds,d\theta)-dsd\theta$, then

$$\begin{aligned}M_\alpha(t):&=\int_0^t\int_0^\infty\mu^\alpha_K(N,s^-,\theta)\tilde Q(ds,d\theta)\\ &=P_{\alpha,b_1}(t)-P_{\alpha,b_1}(0)-r_Kf_Af_a\int_0^t\mathbf{1}_{\{N_\alpha(s)\ge1\}}\frac{N_{\tilde\alpha b_1}(s)N_{\alpha b_2}(s)-N_{\alpha b_1}(s)N_{\tilde\alpha b_2}(s)}{(N_\alpha(s)+1)(f_AN_A(s)+f_aN_a(s))}ds\end{aligned}$$

is a local martingale. By construction the process P_{α,b_1} has values in $[0,1]$ and as $r_K \le 1$,

$$\sup_{s\le t}\left|r_K f_A f_a \int_0^s \mathbf{1}_{\{N_\alpha\ge 1\}} \frac{N_{\bar{\alpha}b_1}N_{\alpha b_2}-N_{\alpha b_1}N_{\bar{\alpha}b_2}}{(N_\alpha+1)(f_A N_A+f_a N_a)}\right| \le r_K f_\alpha t \le f_\alpha t, \quad t\ge 0. \tag{3.2.6}$$

Thus M_α is a square integrable pure jump martingale bounded on every finite interval with quadratic variation

$$\begin{aligned}\langle M_\alpha\rangle_t &= \int_0^t\int_0^\infty \left(\mu_K^\alpha(N,s,\theta)\right)^2 ds d\theta \\ &= \int_0^t \Big\{P_{\alpha,b_1}(1-P_{\alpha,b_1})\Big[\Big(D_\alpha+\frac{C_{\alpha,A}}{K}N_A+\frac{C_{\alpha,a}}{K}N_a\Big)\frac{\mathbf{1}_{N_\alpha\ge 2}N_\alpha}{(N_\alpha-1)^2} \\ &\quad +\frac{f_\alpha N_\alpha}{(N_\alpha+1)^2}\Big]+r_K f_A f_a \mathbf{1}_{\{N_\alpha\ge 1\}}\frac{(N_{\bar{\alpha}b_1}N_{\alpha b_2}-N_{\alpha b_1}N_{\bar{\alpha}b_2})(1-2P_{\alpha,b_1})}{(N_\alpha+1)^2(f_A N_A+f_a N_a)}\Big\}.\end{aligned} \tag{3.2.7}$$

This ends the proof of Proposition 1. □

Remark 8. *By definition of the functions μ_K^α in (3.2.4) we have for all (s,θ) in $\mathbb{R}_+\times\mathbb{R}_+$,*

$$\mu_K^A(N,s,\theta)\mu_K^a(N,s,\theta)=0. \tag{3.2.8}$$

Lemma 1 states properties of the quadratic variation $\langle M_\alpha\rangle$ widely used in the forthcoming proofs. We introduce a compact interval containing the equilibrium size of the A-population,

$$I_\varepsilon^K := \left[K\Big(\bar{n}_A-2\varepsilon\frac{C_{A,a}}{C_{A,A}}\Big), K\Big(\bar{n}_A+2\varepsilon\frac{C_{A,a}}{C_{A,A}}\Big)\right]\cap\mathbb{N}, \tag{3.2.9}$$

and the stopping times T_ε^K and S_ε^K, which denote respectively the hitting time of $\lfloor\varepsilon K\rfloor$ by the mutant population and the exit time of I_ε^K by the resident population,

$$T_\varepsilon^K := \inf\Big\{t\ge 0, N_a^K(t)=\lfloor\varepsilon K\rfloor\Big\}, \quad S_\varepsilon^K := \inf\Big\{t\ge 0, N_A^K(t)\notin I_\varepsilon^K\Big\}. \tag{3.2.10}$$

Finally we introduce a constant depending on $\alpha\in\mathscr{A}$ and $\nu\in\mathbb{R}_+^*$,

$$C(\alpha,\nu) := 4D_\alpha+2f_\alpha+4(C_{\alpha,A}+C_{\alpha,a})\nu. \tag{3.2.11}$$

Lemma 1. *For $\nu<\infty$ and $t\ge 0$ such that $(N_A^{(z,K)}(t), N_a^{(z,K)}(t))\in[0,\nu K]^2$,*

$$\frac{d}{dt}\langle M_\alpha^{(z,K)}\rangle_t = \int_0^\infty\left(\mu_K^\alpha(N^{(z,K)},t,\theta)\right)^2 d\theta \le C(\alpha,\nu)\frac{\mathbf{1}_{N_\alpha(t)\ge 1}}{N_\alpha(t)}, \quad \alpha\in\mathscr{A}. \tag{3.2.12}$$

Moreover, under Assumptions 1 and 2, there exist $k_0\in\mathbb{N}$, $\varepsilon_0>0$ and a pure jump martingale $\bar{M}$ such that for $\varepsilon\le\varepsilon_0$ and $t\ge 0$,

$$e^{\frac{S_{aA}}{2(k_0+1)}t\wedge T_\varepsilon^K\wedge S_\varepsilon^K}\int_0^\infty\left(\mu_K^a(N^{(z^{(K)},K)}, t\wedge T_\varepsilon^K\wedge S_\varepsilon^K,\theta)\right)^2 d\theta \le (k_0+1)C(a,2\bar{n}_A)\bar{M}_{t\wedge T_\varepsilon^K\wedge S_\varepsilon^K}, \tag{3.2.13}$$

and

$$\mathbb{E}\Big[\bar{M}_{t\wedge T_\varepsilon^K\wedge S_\varepsilon^K}\Big]\le\frac{1}{k_0+1}. \tag{3.2.14}$$

Démonstration. Equation (3.2.12) is a direct consequence of (3.2.7). To prove (3.2.13) and (3.2.14), let us first notice that according to Assumption 1, there exists $k_0 \in \mathbb{N}$ such that for ε small enough and $k \in \mathbb{Z}_+$,

$$\frac{f_a(k_0+k-1)-(D_a+C_{a,A}\bar{n}_A+\varepsilon(C_{a,a}+2C_{A,a}C_{a,A}/C_{A,A}))(k_0+k+1)}{k_0+k-1} \geq \frac{S_{aA}}{2}.$$

This implies in particular that for every $t < T_\varepsilon^K \wedge S_\varepsilon^K$,

$$\frac{f_a N_a(t)(N_a(t)+k_0-1)-d_a^K(N(t))(N_a(t)+k_0+1)}{(N_a(t)+k_0-1)(N_a(t)+k_0+1)} \geq \frac{S_{aA}N_a(t)}{2(N_a(t)+k_0+1)} \geq \frac{S_{aA}\mathbf{1}_{N_a(t)\geq 1}}{2(k_0+1)}, \tag{3.2.15}$$

where the death rate d_a^K has been defined in (3.1.3). For sake of simplicity let us introduce the process X defined as follows :

$$X(t) = \frac{1}{N_a(t)+k_0}\exp\Big(\frac{S_{aA}t}{2(k_0+1)}\Big), \quad \forall t \geq 0.$$

Applying It's formula with jumps we get for every $t \geq 0$:

$$X(t\wedge T_\varepsilon^K \wedge S_\varepsilon^K) = \bar{M}(t\wedge T_\varepsilon^K \wedge S_\varepsilon^K) + \int_0^{t\wedge T_\varepsilon^K\wedge S_\varepsilon^K}\Big(\frac{S_{aA}}{2(k_0+1)} - \frac{f_a N_a(s)(N_a(s)+k_0-1)-d_a^K(N(s))(N_a(s)+k_0+1)}{(N_a(s)+k_0-1)(N_a(s)+k_0+1)}\Big)X(s)ds, \tag{3.2.16}$$

where the martingale $\bar{M}$ has the following expression :

$$\bar{M}(t) = \frac{1}{k_0+1} + \int_0^t\int_{\mathbb{R}_+}\tilde{Q}(ds,d\theta)\exp\Big(\frac{S_{aA}s}{2(k_0+1)}\Big) \Big[\frac{\mathbf{1}_{\theta\leq f_aN_a(s^-)}}{N_a(s^-)+k_0+1} + \frac{\mathbf{1}_{f_aN_a(s^-)<\theta\leq f_aN_a(s^-)+d_a(N(s^-))}}{N_a(s^-)+k_0-1} - \frac{\mathbf{1}_{\theta\leq f_aN_a(s^-)+d_a(N(s^-))}}{N_a(s^-)+k_0}\Big]. \tag{3.2.17}$$

Thanks to (3.2.15) the integral in (3.2.16) is nonpositive. Moreover, according to (3.2.12), for $t \leq T_\varepsilon^K \wedge S_\varepsilon^K$, as $2\varepsilon C_{A,a}/C_{A,A} \leq \bar{n}_A$ for ε small enough,

$$\int_0^\infty\Big(\mu_K^a(N^{(z^{(K)},K)},t,\theta)\Big)^2 d\theta \leq C(a,2\bar{n}_A)\frac{\mathbf{1}_{N_a(t)\geq 1}}{N_a(t)} \leq (k_0+1)C(a,2\bar{n}_A)X(t)\exp\Big(-\frac{S_{aA}t}{2(k_0+1)}\Big), \tag{3.2.18}$$

which ends the proof. □

3.3 Proof of Theorem 1

In this section we suppose that Assumption 1 holds. For $\varepsilon \leq C_{a,a}/C_{a,A} \wedge 2|S_{Aa}|/C_{A,a}$ and z in $\mathbb{R}_+^{A\times\mathscr{B}} \times (\mathbb{R}_+^{a\times\mathscr{B}} \setminus (0,0))$ we introduce a deterministic time $t_\varepsilon(z)$ after which the solution $(n_A^{(z)}, n_a^{(z)})$ of the dynamical system (3.1.4) is close to the stable equilibrium $(0,\bar{n}_a)$:

$$t_\varepsilon(z) := \inf\big\{s \geq 0, \forall t \geq s, (n_A^{(z)}(t), n_a^{(z)}(t)) \in [0,\varepsilon^2/2]\times[\bar{n}_a - \varepsilon/2, \infty)\big\}. \tag{3.3.1}$$

Once $(n_A^{(z)}, n_a^{(z)})$ has reached the set $[0, \varepsilon^2/2] \times [\bar{n}_a - \varepsilon/2, \infty)$ it never escapes from it. Moreover, according to Assumption 1 on the stable equilibrium, $t_\varepsilon(z)$ is finite.

First we compare the population process with the four dimensional dynamical system (3.1.15) on the time interval $[0, t_\varepsilon(z)]$. Then we study this dynamical system and get an approximation of the neutral proportions at time $t_\varepsilon(z)$. Finally, we state that during the A-population extinction period, this proportion stays nearly constant.

Comparison with a four dimensional dynamical system

Recall that $n^{(z,K)} = (n_{\alpha\beta}^{(z,K)}, (\alpha,\beta) \in \mathscr{E})$ is the solution of the dynamical system (3.1.15) with initial condition z. Then we have the following comparison result :

Lemma 2. *Let z be in $\mathbb{R}_+^{\mathscr{E}}$ and ε be in $\mathbb{R}_+^*$. Then*

$$\lim_{K\to\infty} \sup_{s \le t_\varepsilon(z)} \| N^{(z,K)}(s)/K - n^{(z,K)}(s) \| = 0 \quad \textit{in probability} \tag{3.3.2}$$

where $\|.\|$ denotes the L^1-Norm on $\mathbb{R}^{\mathscr{E}}$.

Démonstration. The proof relies on a slight modification of Theorem 2.1 p. 456 in Ethier and Kurtz [EK86]. According to (3.1.1) and (3.1.2), the rescaled birth and death rates

$$\tilde{b}_{\alpha\beta}^K(n) = \frac{1}{K} b_{\alpha\beta}^K(Kn) = f_\alpha n_{\alpha\beta} + r_K f_a f_A \frac{n_{\bar{\alpha}\beta} n_{\alpha\bar{\beta}} - n_{\alpha\beta} n_{\bar{\alpha}\bar{\beta}}}{f_A n_A + f_a n_a}, \quad (\alpha,\beta) \in \mathscr{E}, n \in N^{\mathscr{E}}, \tag{3.3.3}$$

and

$$\tilde{d}_{\alpha\beta}(n) = \frac{1}{K} d_{\alpha\beta}^K(Kn) = \left[D_\alpha + C_{\alpha,A} n_A + C_{\alpha,a} n_a \right] n_{\alpha\beta}, \quad (\alpha,\beta) \in \mathscr{E}, n \in N^{\mathscr{E}}, \tag{3.3.4}$$

are Lipschitz and bounded on every compact subset of $\mathbb{N}^{\mathscr{E}}$. The only difference with [EK86] is that $\tilde{b}_{\alpha\beta}^K$ depends on K via the term r_K. Let $(Y_i^{(\alpha\beta)}, i \in \{1,2\}, (\alpha,\beta) \in \mathscr{E})$ be eight independent standard Poisson processes. From the representation of the population process $N^{(z,K)}$ in (3.2.3) we see that the process $(\bar{N}^{(z,K)}(t), t \ge 0)$ defined by

$$\bar{N}^{(z,K)}(t) = \lfloor zK \rfloor + \sum_{(\alpha,\beta)\in\mathscr{E}} \left[Y_1^{(\alpha\beta)}\Big(\int_0^t b_{\alpha\beta}^K(\bar{N}^{(z,K)}(s)) ds \Big) - Y_2^{(\alpha\beta)}\Big(\int_0^t d_{\alpha\beta}^K(\bar{N}^{(z,K)}(s)) ds \Big) \right],$$

has the same law as $(N^{(z,K)}(t), t \ge 0)$. Applying Definitions (3.3.3) and (3.3.4) we get :

$$\frac{\bar{N}^{(z,K)}(t)}{K} = \frac{\lfloor zK \rfloor}{K} + Mart^{(z,K)}(t) + \int_0^t \sum_{(\alpha,\beta)\in\mathscr{E}} e_{\alpha\beta} \Big(\tilde{b}_{\alpha\beta}^K\Big(\frac{\bar{N}^{(z,K)}(s)}{K} \Big) - \tilde{d}_{\alpha\beta}\Big(\frac{\bar{N}^{(z,K)}(s)}{K} \Big) \Big) ds,$$

where we recall that $(e_{\alpha\beta}, (\alpha,\beta) \in \mathscr{E})$ is the canonical basis of $\mathbb{R}_+^{\mathscr{E}}$ and the martingale $Mart^{(z,K)}$ is defined by

$$Mart^{(z,K)} := \frac{1}{K} \sum_{(\alpha,\beta)\in\mathscr{E}} \left[\tilde{Y}_1^{(\alpha\beta)}\Big(K \int_0^t \tilde{b}_{\alpha\beta}^K\Big(\frac{\bar{N}^{(z,K)}(s)}{K} \Big) ds \Big) - \tilde{Y}_2^{(\alpha\beta)}\Big(K \int_0^t \tilde{d}_{\alpha\beta}\Big(\frac{\bar{N}^{(z,K)}(s)}{K} \Big) ds \Big) \right]$$

and $(\tilde{Y}_i^{(\alpha\beta)}(u) = Y_i^{(\alpha\beta)}(u) - u, u \geq 0, i \in \{1,2\}, (\alpha,\beta) \in \mathscr{E})$ are the Poisson processes centered at their expectation. We also have by definition

$$n^{(z,K)}(t) = z + \int_0^t \sum_{(\alpha,\beta)\in\mathscr{E}} e_{\alpha\beta}\Big(\tilde{b}^K_{\alpha\beta}(n^{(z,K)}(s)) - \tilde{d}_{\alpha\beta}(n^{(z,K)}(s))\Big)ds.$$

Hence, for every $t \leq t_\varepsilon(z)$,

$$\Big|\frac{\bar{N}^{(z,K)}(t)}{K} - n^{(z,K)}(t)\Big| \leq \Big|\frac{\lfloor zK\rfloor}{K} - z\Big| + \Big|Mart^{(z,K)}(t)\Big| + \int_0^t \sum_{(\alpha,\beta)\in\mathscr{E}} \Big|\Big(\tilde{b}^K_{\alpha\beta} - \tilde{d}_{\alpha\beta}\Big)\Big(\frac{\bar{N}^{(z,K)}(s)}{K}\Big) - \Big(\tilde{b}^K_{\alpha\beta} - \tilde{d}_{\alpha\beta}\Big)\Big(n^{(z,K)}(s)\Big)\Big|ds,$$

and there exists a finite constant $\mathscr{K}$ such that

$$\Big|\frac{\bar{N}^{(z,K)}(t)}{K} - n^{(z,K)}(t)\Big| \leq \frac{1}{K} + \Big|Mart^{(z,K)}(t)\Big| + \mathscr{K}\int_0^t \Big|\frac{\bar{N}^{(z,K)}(s)}{K} - n^{(z,K)}(s)\Big|ds.$$

But following Ethier and Kurtz, we get

$$\lim_{K\to\infty} \sup_{s\leq t_\varepsilon(z)} |Mart^{(z,K)}| = 0, \quad \text{a.s.},$$

and using Gronwall's Lemma we finally obtain

$$\lim_{K\to\infty} \sup_{s\leq t_\varepsilon(z)} \|\bar{N}^{(z,K)}(s)/K - n^{(z,K)}(s)\| = 0 \quad \text{a.s.}$$

As the convergence in law to a constant is equivalent to the convergence in probability to the same constant, the result follows. □

Once we know that the rescaled population process is close to the solution of the dynamical system (3.1.15), we can study the dynamical system.

Lemma 3. *Let z be in $\mathbb{R}_+^{\mathscr{E}}$ such that $z_A > 0$ and $z_a > 0$. Then $n_a^{(z,K)}(t)$ and $n_{ab_1}^{(z,K)}(t)$ have a finite limit when t goes to infinity, and there exists a positive constant ε_0 such that for every $\varepsilon \leq \varepsilon_0$,*

$$\Big|\frac{n_{ab_1}^{(z,K)}(\infty)}{n_a^{(z,K)}(\infty)} - \frac{n_{ab_1}^{(z,K)}(t_\varepsilon(z))}{n_a^{(z,K)}(t_\varepsilon(z))}\Big| \leq \frac{2f_a\varepsilon^2}{\bar{n}_A|S_{aA}|}.$$

Démonstration. First notice that by definition of the dynamical systems (3.1.4) and (3.1.15), $n_\alpha^{(z,K)} = n_\alpha^{(z)}$ for $\alpha \in \mathscr{A}$ and $z \in \mathbb{R}^{\mathscr{E}}$. Assumption 1 ensures that $n_a^{(z)}(t)$ goes to $\bar{n}_a$ at infinity. If we define the functions

$$p_{\alpha,b_1}^{(z,K)} = n_{\alpha b_1}^{(z,K)}/n_\alpha^{(z)}, \quad \alpha \in \mathscr{A}, \quad \text{and} \quad g^{(z,K)} = p_{A,b_1}^{(z,K)} - p_{a,b_1}^{(z,K)},$$

we easily check that $\phi : (n^{(z,K)}_{Ab_1}, n^{(z,K)}_{Ab_2}, n^{(z,K)}_{ab_1}, n^{(z,K)}_{ab_2}) \mapsto (n^{(z)}_A, n^{(z)}_a, g^{(z,K)}, p^{(z,K)}_{a,b_1})$ defines a change of variables from $(\mathbb{R}^*_+)^{\mathscr{E}}$ to $\mathbb{R}^{2*}_+ \times]-1,1[\times]0,1[$, and (3.1.15) is equivalent to :

$$\begin{cases} \dot{n}^{(z)}_\alpha = (f_\alpha - (D_\alpha + C_{\alpha,A} n^{(z)}_A + C_{\alpha,a} n^{(z)}_a)) n^{(z)}_\alpha, \quad \alpha \in \mathscr{A} \\ \dot{g}^{(z,K)} = -g^{(z,K)} \Big(r_K f_A f_a (n^{(z)}_A + n^{(z)}_a) / (f_A n^{(z)}_A + f_a n^{(z)}_a) \Big) \\ \dot{p}^{(z,K)}_{a,b_1} = g^{(z,K)} \Big(r_K f_A f_a n^{(z)}_A / (f_A n^{(z)}_A + f_a n^{(z)}_a) \Big), \end{cases} \tag{3.3.5}$$

with initial condition

$$(n^{(z)}_A(0), n^{(z)}_a(0), g^{(z,K)}(0), p^{(z,K)}_{a,b_1}(0)) = (z_A, z_a, z_{Ab_1}/z_A - z_{ab_1}/z_a, z_{ab_1}/z_a).$$

Moreover, a direct integration yields

$$p^{(z,K)}_{a,b_1}(t) = p^{(z,K)}_{a,b_1}(0) - (p^{(z,K)}_{a,b_1}(0) - p^{(z,K)}_{A,b_1}(0)) F(z, r_K, t), \tag{3.3.6}$$

where F has been defined in (3.1.12). According to (3.1.13), $F(z, r_K, t)$ has a finite limit when t goes to infinity. Hence $p^{(z,K)}_{a,b_1}$ also admits a limit at infinity. Let $\varepsilon \le |S_{Aa}|/C_{A,a} \wedge C_{a,a}/C_{A,a} \wedge \bar{n}_a/2$, and $t_\varepsilon(z)$ defined in (3.3.1). Then for $t \ge t_\varepsilon(z)$,

$$\dot{n}^{(z)}_A(t) \le (f_A - D_A - C_{Aa}(\bar{n}_a - \varepsilon/2)) n^{(z)}_A(t) \le S_{Aa} n^{(z)}_A(t)/2 < 0.$$

Recalling that $r_K \le 1$ and $|g(t)| \le 1$ for all $t \ge 0$ we get :

$$\left| p^{(z,K)}_{a,b_1}(\infty) - p^{(z,K)}_{a,b_1}(t_\varepsilon(z)) \right| \le \int_{t_\varepsilon(z)}^\infty \frac{f_A f_a n^{(z)}_A}{f_A n^{(z)}_A + f_a n^{(z)}_a} \le \frac{f_A \varepsilon^2}{\bar{n}_a - \varepsilon/2} \int_0^\infty e^{S_{Aa} s/2} ds \le \frac{2 f_a \varepsilon^2}{(\bar{n}_a - \varepsilon/2)|S_{aA}|}, \tag{3.3.7}$$

which ends the proof. □

A-population extinction

The deterministic approximation (3.1.15) fails when the A-population size becomes too small. We shall compare N_A with birth and death processes to study the last period of the mutant invasion. We show that during this period, the number of A individuals is so small that it has no influence on the neutral proportion in the a-population, which stays nearly constant. Before stating the result, we recall Definition (3.1.9) and introduce the compact set Θ :

$$\Theta := \left\{ z \in \mathbb{R}_+^{A \times \mathscr{B}} \times \mathbb{R}_+^{a \times \mathscr{B}}, z_A \le \varepsilon^2 \quad \text{and} \quad |z_a - \bar{n}_a| \le \varepsilon \right\}, \tag{3.3.8}$$

the constant $M'' = 3 + (f_a + C_{a,A})/C_{a,a}$, and the stopping time :

$$U^K_\varepsilon(z) := \inf \left\{ t \ge 0, N^{(z,K)}_A(t) > \varepsilon K \text{ or } |N^{(z,K)}_a(t) - \bar{n}_a K| > M'' \varepsilon K \right\}. \tag{3.3.9}$$

Lemma 4. *Let z be in Θ. Under Assumption 1, there exist two positive finite constants c and ε_0 such that for $\varepsilon \le \varepsilon_0$,*

$$\limsup_{K \to \infty} \mathbb{P}\Big(\sup_{t \le T^{(z,K)}_{\text{ext}}} \left| P^{(z,K)}_{a,b_1}(t) - P^{(z,K)}_{a,b_1}(0) \right| > \varepsilon \Big) \le c\varepsilon.$$

Démonstration. Let z be in Θ and Z^1 be a birth and death process with individual birth rate f_A, individual death rate $D_A+(\bar{n}_a-M''\varepsilon)C_{A,a}$, and initial state $\lceil\varepsilon^2K\rceil$. Then on $[0,U_\varepsilon^K(z)[$, N_A and Z^1 have the same birth rate, and Z^1 has a smaller death rate than N_A. Thus according to Theorem 2 in [Cha06], we can construct the processes N and Z^1 on the same probability space such that :

$$N_A(t)\le Z_t^1,\quad \forall t<U_\varepsilon^K(z). \tag{3.3.10}$$

Moreover, if we denote by T_0^1 the extinction time of Z^1, $T_0^1:=\inf\{t\ge 0, Z_t^1=0\}$, and recall that

$$f_A-D_A-(\bar{n}_a-M''\varepsilon)C_{A,a}=S_{Aa}+M''C_{A,a}\varepsilon<S_{Aa}/2<0,\quad \forall\varepsilon<|S_{Aa}|/(2M''C_{Aa}), \tag{3.3.11}$$

we get according to (A.10) that for $z\le\varepsilon^2$ and

$$L(\varepsilon,K)=2\log K/|S_{Aa}+M''\varepsilon C_{A,a}|,$$

$$\mathbb{P}_{\lceil zK\rceil}\Big(T_0^1\le L(\varepsilon,K)\Big)\ge\exp\Big(\lceil\varepsilon^2K\rceil\Big[\log(K^2-1)-\log(K^2-f_A(D_A+(\bar{n}_a-M''\varepsilon)C_{A,a})^{-1})\Big]\Big).$$

Thus :

$$\lim_{K\to\infty}\mathbb{P}_{\lceil zK\rceil}\Big(T_0^1<L(\varepsilon,K)\Big)=1. \tag{3.3.12}$$

Moreover, Equation (A.4) ensures the existence of a finite c such that for ε small enough,

$$\mathbb{P}\Big(L(\varepsilon,K)<U_\varepsilon^K(z)\Big)\ge 1-c\varepsilon. \tag{3.3.13}$$

Equations (3.3.12) and (3.3.13) imply

$$\liminf_{K\to\infty}\mathbb{P}\Big(T_0^1<L(\varepsilon,K)<U_\varepsilon^K(z)\Big)\ge 1-c\varepsilon \tag{3.3.14}$$

for a finite c and ε small enough. According to Coupling (3.3.10) we have the inclusion

$$\{T_0^1<L(\varepsilon,K)<U_\varepsilon^K(z)\}\subset\{T_{ext}^K<L(\varepsilon,K)<U_\varepsilon^K(z)\}.$$

Adding (3.3.14) we finally get :

$$\liminf_{K\to\infty}\mathbb{P}(T_{ext}^K<L(\varepsilon,K)<U_\varepsilon^K(z))\ge 1-c\varepsilon. \tag{3.3.15}$$

Recall the martingale decomposition of P_{a,b_1} in (3.2.1). To bound the difference $|P_{a,b_1}(t)-P_{a,b_1}(0)|$ we bound independently the martingale $M_a(t)$ and the integral $|P_{a,b_1}(t)-P_{a,b_1}(0)-M_a(t)|$. On one hand Doob's Maximal Inequality and Equation (3.2.12) imply :

$$\begin{aligned}\mathbb{P}\Big(\sup_{t\le L(\varepsilon,K)\wedge U_\varepsilon^K}|M_a(t)|>\frac{\varepsilon}{2}\Big) &\le \frac{4}{\varepsilon^2}\mathbb{E}\Big[\langle M_a\rangle_{L(\varepsilon,K)\wedge U_\varepsilon^K(z)}\Big]\\ &\le \frac{4C(a,\bar{n}_a+M''\varepsilon)L(\varepsilon,K)}{\varepsilon^2K(\bar{n}_a-M''\varepsilon)}\\ &= \frac{8C(a,\bar{n}_a+M''\varepsilon)\log K}{\varepsilon^2K(\bar{n}_a-M''\varepsilon)|S_{Aa}+M''\varepsilon C_{A,a}|}.\end{aligned} \tag{3.3.16}$$

On the other hand the inequality $|N_{Ab_1}N_{ab_2} - N_{ab_1}N_{Ab_2}| \le N_A N_a$ yields for $t \ge 0$

$$\Big|\int_0^{t\wedge U_\varepsilon^K(z)} \frac{r_K f_A f_a (N_{Ab_1}N_{ab_2} - N_{ab_1}N_{Ab_2})}{(N_a+1)(f_A N_A + f_a N_a)}\Big| \le \int_0^{t\wedge U_\varepsilon^K(z)} \frac{f_A N_A}{(\bar n_a - \varepsilon M'')K}.$$

Hence decomposition (3.2.1), Markov's Inequality, and Equations (3.3.10), (A.8) and (3.3.11) yield

$$\mathbb{P}\Big(\Big|(P_{a,b_1} - M_a)(t\wedge U_\varepsilon^K(z)) - P_{a,b_1}(0)\Big| > \frac{\varepsilon}{2}\Big) \le \frac{2f_A\varepsilon^2}{\varepsilon(\bar n_a - \varepsilon M'')}\int_0^t e^{\frac{S_{Aa}s}{2}}\,ds \le \frac{4f_A\varepsilon}{(\bar n_a - \varepsilon M'')|S_{Aa}|}. \tag{3.3.17}$$

Taking the limit of (3.3.16) when K goes to infinity and adding (3.3.17) end the proof. □

End of the proof of Theorem 1

Recall Definitions (3.1.9) and (3.3.1). We have :

$$\Big|P_{a,b_1}^{(z,K)}(T_{\text{ext}}^{(z,K)}) - p_{a,b_1}^{(z,K)}(\infty)\Big| \le \Big|P_{a,b_1}^{(z,K)}(T_{\text{ext}}^{(z,K)}) - P_{a,b_1}^{(z,K)}(t_\varepsilon(z))\Big| + \Big|P_{a,b_1}^{(z,K)}(t_\varepsilon(z)) - p_{a,b_1}^{(z,K)}(t_\varepsilon(z))\Big| + \Big|p_{a,b_1}^{(z,K)}(t_\varepsilon(z)) - p_{a,b_1}^{(z,K)}(\infty)\Big|.$$

To bound the two last terms we use respectively Lemmas 2 and 3. For the first term of the right hand side, (3.1.5) ensures that with high probability, $N^{(z,K)}(t_\varepsilon(z)) \in \Theta$ and $t_\varepsilon(z) < T_{\text{ext}}^{(z,K)}$. Lemma 4, Equation (3.1.14) and Markov's Inequality allow us to conclude that for ε small enough

$$\limsup_{K\to\infty} \mathbb{P}(\mathbf{1}_{\text{Fix}^{(z,K)}}|P_{a,b_1}^{(z,K)}(T_{\text{ext}}^{(z,K)}) - p_{a,b_1}^{(z,K)}(\infty)| > 3\varepsilon) \le c\varepsilon,$$

for a finite c, which is equivalent to the convergence in probability. Adding (3.3.6) completes the proof.

3.4 A coupling with two birth and death processes

In Sections 3.5 and 3.6 we suppose that Assumptions 1 and 2 hold and we denote by N^K the process $N^{(z^{(K)},K)}$. As it will appear in the proof of Theorem 2 the first period of mutant invasion, which ends at time T_ε^K when the mutant population size hits $\lfloor \varepsilon K\rfloor$, is the most important for the neutral proportion dynamics. Indeed, the neutral proportion in the a-population has already reached its final value at time T_ε^K. Let us describe a coupling of the process N_a^K with two birth and death processes which will be a key argument to control the growing of the population a during the first period. We recall Definition (3.2.10) and define for $\varepsilon < S_{aA}/(2C_{a,A}C_{A,a}/C_{A,A} + C_{a,a})$,

$$s_-(\varepsilon) := \frac{S_{aA}}{f_a} - \varepsilon\frac{2C_{a,A}C_{A,a} + C_{a,a}C_{A,A}}{f_a C_{A,A}}, \quad \text{and} \quad s_+(\varepsilon) := \frac{S_{aA}}{f_a} + 2\varepsilon\frac{C_{a,A}C_{A,a}}{f_a C_{A,A}}. \tag{3.4.1}$$

Definitions (3.1.3) and (3.1.8) ensure that for $t < T_\varepsilon^K \wedge S_\varepsilon^K$,

$$f_a(1-s_+(\varepsilon)) \le \frac{d_a^K(N^K(t))}{N_a^K(t)} = f_a - S_{aA} + \frac{C_{a,A}}{K}(N_A^K(t) - \bar n_A K) + \frac{C_{a,a}}{K}N_a^K(t) \le f_a(1-s_-(\varepsilon)), \tag{3.4.2}$$

and following Theorem 2 in [Cha06], we can construct on the same probability space the processes Z_ε^-, N^K and Z_ε^+ such that almost surely :

$$Z_\varepsilon^-(t) \le N_a^K(t) \le Z_\varepsilon^+(t), \quad \text{for all } t < T_\varepsilon^K \wedge S_\varepsilon^K, \tag{3.4.3}$$

where for $* \in \{-,+\}$, Z_ε^* is a birth and death process with initial state 1, and individual birth and death rates f_a and $f_a(1-s_*(\varepsilon))$.

3.5 Proof of Theorem 2 in the strong recombination regime

In this section, we suppose that Assumptions 1, 2 and 3 hold. We distinguish the three periods of the selective sweep : (i) rare mutants and resident population size near its equilibrium value, (ii) quasi-deterministic period governed by the dynamical system (3.1.4), and (iii) A-population extinction. First we prove that at time T_ε^K proportions of b_1 alleles in the populations A and a are close to z_{Ab_1}/z_A. Once the neutral proportions are the same in the two populations, they do not evolve anymore until the end of the sweep.

Lemma 5. *There exist two positive finite constants c and ε_0 such that for $\varepsilon \le \varepsilon_0$:*

$$\limsup_{K\to\infty} \mathbb{E}\Big[\mathbf{1}_{T_\varepsilon^K \le S_\varepsilon^K}\Big\{\Big|P_{A,b_1}^K(T_\varepsilon^K) - \frac{z_{Ab_1}}{z_A}\Big| + \Big|P_{A,b_1}^K(T_\varepsilon^K) - P_{a,b_1}^K(T_\varepsilon^K)\Big|\Big\}\Big] \le c\varepsilon.$$

Démonstration. First we bound the difference between the neutral proportions in the two populations, $|P_{a,b_1}(t) - P_{A,b_1}(t)|$, then we bound $|P_{A,b_1}(t) - z_{Ab_1}/z_A|$. For sake of simplicity we introduce :

$$G(t) := P_{A,b_1}(t) - P_{a,b_1}(t) = \frac{N_{ab_2}(t)N_{Ab_1}(t) - N_{ab_1}(t)N_{Ab_2}(t)}{N_A(t)N_a(t)}, \quad \forall t \ge 0, \tag{3.5.1}$$

$$Y(t) = G^2(t)e^{r_K f_a t/2}, \quad \forall t \ge 0. \tag{3.5.2}$$

Recalling (3.2.8) and applying It's formula with jumps we get

$$\begin{aligned} Y(t\wedge T_\varepsilon^K \wedge S_\varepsilon^K) = Y(0) + \hat{M}_{t\wedge T_\varepsilon^K \wedge S_\varepsilon^K} + r_K \int_0^t \mathbf{1}_{s<T_\varepsilon^K\wedge S_\varepsilon^K}\Big(f_a/2 - H(s)\Big)Y(s)ds \\ + \int_0^t \mathbf{1}_{s<T_\varepsilon^K\wedge S_\varepsilon^K} e^{r_K f_a s/2} ds \int_{\mathbb{R}_+} \Big[\Big(\mu_A^K(N,s,\theta)\Big)^2 + \Big(\mu_a^K(N,s,\theta)\Big)^2\Big]d\theta, \end{aligned} \tag{3.5.3}$$

where $\hat{M}$ is a martingale with zero mean, and H is defined by

$$H(t) = \frac{2f_a f_A N_A(t)N_a(t)}{f_A N_A(t) + f_a N_a(t)}\Big[\frac{1}{N_A(t)+1} + \frac{1}{N_a(t)+1}\Big] \ge \frac{f_a}{2}, \quad t < T_\varepsilon^K \wedge S_\varepsilon^K, \tag{3.5.4}$$

for ε small enough. In particular the first integral in (3.5.3) is non-positive. Applying Lemma 1 we obtain :

$$\begin{aligned} \mathbb{E}[Y(t\wedge T_\varepsilon^K \wedge S_\varepsilon^K)] &\le 1 + \frac{2C(A,2\bar{n}_A)}{r_K f_a(\bar{n}_A - 2\varepsilon C_{A,a}/C_{A,A})K} e^{\frac{r_K f a t}{2}} \\ &\quad + \int_0^t (k_0+1)C(a,2\bar{n}_A)\mathbb{E}\Big[\tilde{M}_{s\wedge T_\varepsilon^K\wedge S_\varepsilon^K}\Big] e^{(\frac{r_K f a}{2} - \frac{S_{aA}}{2(k_0+1)})s} ds \\ &\le c\Big(1 + \frac{1}{Kr_K} e^{\frac{r_K f a t}{2}} + e^{(\frac{r_K f a}{2} - \frac{S_{aA}}{2(k_0+1)})t}\Big), \end{aligned} \tag{3.5.5}$$

where c is a finite constant which can be chosen independently of ε and K if ε is small enough and K large enough. Combining the semi-martingale decomposition (3.2.1), the Cauchy-Schwarz Inequality, and Equations (3.2.12) and (3.5.5) we get for every $t \geq 0$,

$$\begin{aligned}
&\mathbb{E}\Big[\Big|P_{A,b_1}(t\wedge T_\varepsilon^K \wedge S_\varepsilon^K) - \frac{\lfloor z_{Ab_1}K\rfloor}{\lfloor z_A K\rfloor}\Big|\Big] \\
&\qquad \leq \mathbb{E}\Big[|M_A(t\wedge T_\varepsilon^K \wedge S_\varepsilon^K)|\Big] + \frac{r_K f_a \varepsilon}{\bar{n}_A - 2\varepsilon C_{A,a}/C_{A,A}} \int_0^t \mathbb{E}\Big[\mathbf{1}_{s<T_\varepsilon^K\wedge S_\varepsilon^K}|G(s)|\Big]ds \\
&\qquad \leq \mathbb{E}^{1/2}\Big[\langle M_A\rangle_{t\wedge T_\varepsilon^K\wedge S_\varepsilon^K}\Big] + cr_K\varepsilon\int_0^t \mathbb{E}^{1/2}\Big[Y(s\wedge T_\varepsilon^K\wedge S_\varepsilon^K)\Big]e^{-r_K f_a s/4}ds \\
&\qquad\qquad \leq c\Big(\sqrt{t/K} + \varepsilon r_K \int_0^t \Big(e^{-r_K f_a s/2} + \frac{1}{Kr_K} + e^{-S_{aA}s/2(k_0+1)}\Big)^{1/2} ds\Big),
\end{aligned}$$

where c is finite. A simple integration then yields the existence of a finite c such that :

$$\mathbb{E}\Big[\Big|P_{A,b_1}(t\wedge T_\varepsilon^K \wedge S_\varepsilon^K) - \frac{\lfloor z_{Ab_1}K\rfloor}{\lfloor z_A K\rfloor}\Big|\Big] \leq c\Big(\sqrt{\frac{t}{K}} + \varepsilon\Big(1+\frac{t}{\sqrt{K}}\Big)\Big). \tag{3.5.6}$$

Let us introduce the sequences of times

$$t_K^{(-)} = (1-c_1\varepsilon)\frac{\log K}{S_{aA}}, \quad \text{and} \quad t_K^{(+)} = (1+c_1\varepsilon)\frac{\log K}{S_{aA}},$$

where c_1 is a finite constant. Then according to Coupling (3.4.3) and limit (A.11),

$$\lim_{K\to\infty} \mathbb{P}(T_\varepsilon^K < t_K^{(-)} | T_\varepsilon^K \leq S_\varepsilon^K) = \lim_{K\to\infty} \mathbb{P}(T_\varepsilon^K > t_K^{(+)} | T_\varepsilon^K \leq S_\varepsilon^K) = 0. \tag{3.5.7}$$

Hence applying (3.5.6) at time $t_K^{(+)}$ and using (A.3) and (3.5.7), we bound the first term in the expectation. To bound the second term in the expectation, we introduce the notation

$$A(\varepsilon,K) := \mathbb{E}\Big[\mathbf{1}_{t_K^{(-)}\leq T_\varepsilon^K\leq S_\varepsilon^K\wedge t_K^{(+)}}\Big|P_{A,b_1}^K(T_\varepsilon^K\wedge S_\varepsilon^K) - P_{a,b_1}^K(T_\varepsilon^K\wedge S_\varepsilon^K)\Big|\Big].$$

From (3.5.7) we obtain

$$\limsup_{K\to\infty} \mathbb{E}\Big[\mathbf{1}_{T_\varepsilon^K\leq S_\varepsilon^K}\Big|P_{A,b_1}^K(T_\varepsilon^K) - P_{a,b_1}^K(T_\varepsilon^K)\Big|\Big] = \limsup_{K\to\infty} A(\varepsilon,K), \tag{3.5.8}$$

and by using (3.5.2), the Cauchy-Schwarz Inequality, and (3.5.5) we get

$$\begin{aligned}
A(\varepsilon,K) &\leq \mathbb{E}\Big[\sqrt{Y(t_K^{(+)}\wedge T_\varepsilon^K\wedge S_\varepsilon^K)}\Big]e^{-\frac{r_K f_a}{4}t_K^{(-)}} \\
&\leq \mathbb{E}^{1/2}\Big[Y(t_K^{(+)}\wedge T_\varepsilon^K\wedge S_\varepsilon^K)\Big]e^{-\frac{r_K f_a}{4}t_K^{(-)}} \\
&\leq c\Big(1+\frac{1}{Kr_K}e^{r_K f_a t_K^{(+)}/2} + e^{(\frac{r_K f_a}{2} - \frac{S_{aA}}{2(k_0+1)})t_K^{(+)}}\Big)^{1/2} e^{-\frac{r_K f_a}{4}t_K^{(-)}} \\
&\leq c\Big(e^{-\frac{r_K f_a}{4}t_K^{(-)}} + \frac{1}{\sqrt{Kr_K}}e^{\frac{c_1\varepsilon r_K f_a \log K}{2S_{aA}}} + e^{(\frac{c_1\varepsilon r_K f_a \log K}{2S_{aA}} - \frac{S_{aA}}{4(k_0+1)}t_K^{(+)})}\Big),
\end{aligned}$$

where the value of the constant c can change from line to line. Assumption 3 then yields

$$\limsup_{K\to\infty} A(\varepsilon,K) = 0,$$

and we end the proof of the second bound by applying (3.5.8). □

The following Lemma states that during the second period, the neutral proportion stays constant in the a-population.

Lemma 6. *There exist two positive finite constants c and ε_0 such that for $\varepsilon \le \varepsilon_0$:*

$$\limsup_{K\to\infty} \mathbb{E}\Big[\mathbf{1}_{T_\varepsilon^K \le S_\varepsilon^K}\Big|P^K_{a,b_1}(T_\varepsilon^K + t_\varepsilon(\tfrac{N^K(T_\varepsilon^K)}{K})) - \tfrac{z_{Ab_1}}{z_A}\Big|\Big] \le c\varepsilon.$$

Démonstration. Let us introduce, for $z \in \mathbb{R}^{\mathscr{E}}_+$ and $\varepsilon > 0$ the set Γ and the time t_ε defined as follows :

$$\Gamma := \Big\{z \in \mathbb{R}^{\mathscr{E}}_+, \Big|z_A - \bar n_A\Big| \le 2\varepsilon\frac{C_{A,a}}{C_{A,A}}, \Big|z_a - \varepsilon\Big| \le \frac{\varepsilon}{2}\Big\}, \quad t_\varepsilon := \sup\{t_\varepsilon(z), z\in\Gamma\}, \tag{3.5.9}$$

where $t_\varepsilon(z)$ has been defined in (3.3.1). According to Assumption 1, $t_\varepsilon < \infty$, and

$$I(\Gamma,\varepsilon) := \inf_{z\in\Gamma}\inf_{t\le t_\varepsilon}\{n_A^{(z)}(t), n_a^{(z)}(t)\} > 0,$$

and we can introduce the stopping time

$$L_\varepsilon^K(z) = \inf\big\{t\ge 0, (N_A^{(z,K)}(t), N_a^{(z,K)}(t)) \notin [I(\Gamma,\varepsilon)K/2, (\bar n_A + \bar n_a)K]^2\big\}. \tag{3.5.10}$$

Finally, we denote by $(\mathscr{F}_t^K, t\ge 0)$ the canonical filtration of N^K. Notice that on the event $\{T_\varepsilon^K \le S_\varepsilon^K\}$, $N(T_\varepsilon^K)/K \in \Gamma$, thus $t_\varepsilon(N(T_\varepsilon^K)/K) \le t_\varepsilon$. The semi-martingale decomposition (3.2.1) and the definition of G in (3.5.1) then twice the Strong Markov property and the Cauchy-Schwarz Inequality yield :

$$\begin{aligned}
&\mathbb{E}\Big[\mathbf{1}_{T_\varepsilon^K\le S_\varepsilon^K}\Big|P_{a,b_1}\Big(T_\varepsilon^K + t_\varepsilon(\tfrac{N(T_\varepsilon^K)}{K}) \wedge L_\varepsilon^K(\tfrac{N(T_\varepsilon^K)}{K})\Big) - P_{a,b_1}(T_\varepsilon^K)\Big|\Big]\\
&\le \mathbb{E}\Big[\mathbf{1}_{T_\varepsilon^K\le S_\varepsilon^K}\mathbb{E}\Big[\Big|M_a\Big(T_\varepsilon^K + t_\varepsilon(\tfrac{N(T_\varepsilon^K)}{K}) \wedge L_\varepsilon^K(\tfrac{N(T_\varepsilon^K)}{K}) - M_a(T_\varepsilon^K)\Big| + f_a\int_{T_\varepsilon^K}^{T_\varepsilon^K + t_\varepsilon\wedge L_\varepsilon^K(\frac{N(T_\varepsilon^K)}{K})}|G|\Big|\mathscr{F}_{T_\varepsilon^K}\Big]\Big]\\
&\le \mathbb{E}\Big[\mathbf{1}_{T_\varepsilon^K\le S_\varepsilon^K}\Big\{\mathbb{E}^{1/2}\Big[\langle M_a\rangle_{T_\varepsilon^K + t_\varepsilon\wedge L_\varepsilon^K(\frac{N(T_\varepsilon^K)}{K})} - \langle M_a\rangle_{T_\varepsilon^K}\Big|\mathscr{F}_{T_\varepsilon^K}\Big] + f_a\sqrt{t_\varepsilon}\,\mathbb{E}^{1/2}\Big[\int_{T_\varepsilon^K}^{T_\varepsilon^K + t_\varepsilon\wedge L_\varepsilon^K(\frac{N(T_\varepsilon^K)}{K})} G^2\Big|\mathscr{F}_{T_\varepsilon^K}\Big]\Big\}\Big].
\end{aligned} \tag{3.5.11}$$

To bound the first term of the right hand side we use the Strong Markov Property, Equation (3.2.12) and the definition of L_ε^K in (3.5.10). We get

$$\mathbb{E}\Big[\mathbf{1}_{T_\varepsilon^K\le S_\varepsilon^K}\mathbb{E}^{1/2}\Big[\langle M_a\rangle_{T_\varepsilon^K + t_\varepsilon\wedge L_\varepsilon^K(\frac{N(T_\varepsilon^K)}{K})} - \langle M_a\rangle_{T_\varepsilon^K}\Big|\mathscr{F}_{T_\varepsilon^K}\Big]\Big] \le \sqrt{\frac{2t_\varepsilon C(a,\bar n_A+\bar n_a)}{I(\Gamma,\varepsilon)K}}. \tag{3.5.12}$$

Let us now focus on the second term. It's formula with jumps yields for every $t \geq 0$,

$$\mathbb{E}\Big[G^2\Big(t \wedge L_\varepsilon^K(\tfrac{N(0)}{K})\Big)\Big] \leq \mathbb{E}[G^2(0)] + \mathbb{E}\Big[\langle M_A\rangle_{t\wedge L_\varepsilon^K(\frac{N(0)}{K})}\Big] + \mathbb{E}\Big[\langle M_a\rangle_{t\wedge L_\varepsilon^K(\frac{N(0)}{K})}\Big],$$

and adding the Strong Markov Property we get

$$\begin{aligned}
&\mathbf{1}_{T_\varepsilon^K \leq S_\varepsilon^K} \mathbb{E}^{1/2}\Big[\int_{T_\varepsilon^K}^{T_\varepsilon^K + t_\varepsilon \wedge L_\varepsilon^K(\frac{N(T_\varepsilon^K)}{K})} G^2 \Big| \mathscr{F}_{T_\varepsilon^K}\Big] \\
&\quad \leq \sup_{z\in\Gamma} \mathbb{E}^{1/2}\Big[\int_0^{t_\varepsilon} \Big(G^2\Big(s \wedge L_\varepsilon^K(\tfrac{N(0)}{K})\Big) - G^2(0)\Big) ds \Big| N(0) = \lfloor zK \rfloor\Big] + \mathbf{1}_{T_\varepsilon^K \leq S_\varepsilon^K} \sqrt{t_\varepsilon} |G(T_\varepsilon^K)| \\
&\leq \sup_{z\in\Gamma}\Big[\int_0^{t_\varepsilon} \mathbb{E}\Big[\langle M_A\rangle_{s\wedge L_\varepsilon^K(\frac{N(0)}{K})} + \langle M_a\rangle_{s\wedge L_\varepsilon^K(\frac{N(0)}{K})}\Big] \Big| N(0) = \lfloor zK \rfloor\Big] ds\Big]^{1/2} + \mathbf{1}_{T_\varepsilon^K \leq S_\varepsilon^K} \sqrt{t_\varepsilon} |G(T_\varepsilon^K)|.
\end{aligned}$$

Using again Equation (3.2.12) and the definition of L_ε^K in (3.5.10), and adding Lemma 5 finally lead to

$$\mathbb{E}\Big[\mathbf{1}_{T_\varepsilon^K \leq S_\varepsilon^K} \mathbb{E}^{1/2}\Big[\int_{T_\varepsilon^K}^{T_\varepsilon^K + t_\varepsilon \wedge L_\varepsilon^K(\frac{N(T_\varepsilon^K)}{K})} G^2 \Big| \mathscr{F}_{T_\varepsilon^K}\Big]\Big] \leq c\Big(\frac{1}{\sqrt{K}} + \varepsilon\Big), \tag{3.5.13}$$

for ε small enough and K large enough, where c is a finite constant. Moreover (3.1.5) ensures that

$$\mathbb{P}\Big(T_\varepsilon^K \leq S_\varepsilon^K, L_\varepsilon^K(\tfrac{N(T_\varepsilon^K)}{K}) \leq t_\varepsilon(\tfrac{N(T_\varepsilon^K)}{K})\Big) \leq \mathbb{P}\Big(\tfrac{N(T_\varepsilon^K)}{K} \in \Theta, L_\varepsilon^K(\tfrac{N(T_\varepsilon^K)}{K}) \leq t_\varepsilon(\tfrac{N(T_\varepsilon^K)}{K})\Big) \underset{K\to\infty}{\to} 0,$$

where Θ has been defined in (3.3.8). Adding Equations (3.5.11), (3.5.12), (3.5.13) and Lemma 5, we finally end the proof of Lemma 6. □

Proof of Theorem 2 in the strong recombination regime. Let us focus on the A-population extinction period. We have thanks to the Strong Markov Property :

$$\begin{aligned}
&\mathbb{P}\Big(\mathbf{1}_{N(T_\varepsilon^K + t_\varepsilon(N(T_\varepsilon^K)/K))\in\Theta} \Big| P_{a,b_1}(T_{\text{ext}}^K) - P_{a,b_1}(T_\varepsilon^K + t_\varepsilon(\tfrac{N(T_\varepsilon^K)}{K}))\Big| > \sqrt{\varepsilon}\Big) \\
&\qquad\qquad \leq \sup_{z\in\Theta} \mathbb{P}\Big(|P_{a,b_1}(T_{\text{ext}}^K) - P_{a,b_1}(0)| > \sqrt{\varepsilon} \Big| N(0) = \lfloor zK \rfloor\Big).
\end{aligned} \tag{3.5.14}$$

But Equation (3.1.5) yields $\mathbb{P}(N(T_\varepsilon^K + t_\varepsilon(N(T_\varepsilon^K)/K))/K \in \Theta | N(T_\varepsilon^K)/K \in \Gamma) \to_{K\to\infty} 1$, and $\{T_\varepsilon^K \leq S_\varepsilon^K\} \subset \{N(T_\varepsilon^K)/K \in \Gamma\}$. Adding Equation (A.6) and Lemmas 4 and 6, the triangle inequality allows us to conclude that for ε small enough

$$\limsup_{K\to\infty} \mathbb{P}\Big(\Big|P_{a,b_1}^K(T_{\text{ext}}^K) - \frac{z_{Ab_1}}{z_A}\Big| > \sqrt{\varepsilon} \Big| \text{Fix}^K\Big) \leq c\varepsilon.$$

As $\mathbb{P}(\text{Fix}^K) \to_{K\to\infty} S_{aA}/f_a > 0$, it is equivalent to the claim of Theorem 2 in the strong regime. □

3.6 Proof of Theorem 2 in the weak recombination regime

Coupling with a four dimensional population process and structure of the proof

In this section we suppose that Assumptions 1, 2 and 4 hold. To lighten the proofs of Sections 3.6 to 3.6 we introduce a coupling of the population process N with a process $\tilde{N} = (\tilde{N}_{\alpha\beta}, (\alpha,\beta) \in \mathscr{E})$ defined as follows for every $t \geq 0$:

$$\begin{aligned}\tilde{N}(t) &= \mathbf{1}_{t<S_\varepsilon^K} N(t) + \mathbf{1}_{t\geq S_\varepsilon^K}\Big(e_{Ab_1} N_{Ab_1}((S_\varepsilon^K)^-) + e_{Ab_2} N_{Ab_2}((S_\varepsilon^K)^-) \qquad (3.6.1)\\ &+ \int_0^t \int_{R_+} \Big\{ e_{ab_1} \mathbf{1}_{\theta \leq b^K_{ab_1}(\tilde{N}(s^-))} + e_{ab_2} \mathbf{1}_{b^K_{ab_1}(\tilde{N}(s^-)) < \theta \leq f_a \tilde{N}_a(s^-)} \\ &\qquad - e_{ab_1} \mathbf{1}_{0<\theta - f_a\tilde{N}_a(s^-) \leq d^K_{ab_1}(\tilde{N}(s^-))} \\ &\qquad - e_{ab_2} \mathbf{1}_{d^K_{ab_1}(\tilde{N}(s^-)) < \theta - f_a\tilde{N}_a(s^-) \leq d^K_a(\tilde{N}(s^-))} \Big\} Q(ds, d\theta)\Big), \qquad (3.6.2)\end{aligned}$$

where the Poisson random measure Q has been introduced in (3.2.3). From (A.5) we know that

$$\limsup_{K\to\infty} \mathbb{P}(\{\exists t \leq T_\varepsilon^K, N(t) \neq \tilde{N}(t)\}, T_\varepsilon^K < \infty) \leq c\varepsilon. \qquad (3.6.3)$$

Hence we will study the process $\tilde{N}$ and deduce from this study properties of the dynamics of the process N during the first phase. Moreover, as we want to prove convergences on the fixation event Fix^K, defined in (3.1.9), inequalities (A.6) and (3.6.3) allow us, to study the dynamics of $\tilde{N}$ during the first phase, to restrict our attention to the conditional probability measure :

$$\hat{\mathbb{P}}(.) = \mathbb{P}(.|\tilde{T}_\varepsilon^K < \infty), \qquad (3.6.4)$$

where $\tilde{T}_\varepsilon^K$ is the hitting time of $\lfloor \varepsilon K \rfloor$ by the process $\tilde{N}_a$:

$$\tilde{T}_\varepsilon^K := \inf\Big\{ t \geq 0, \tilde{N}_a^K(t) = \lfloor \varepsilon K \rfloor \Big\}. \qquad (3.6.5)$$

Expectations and variances associated with this probability measure are denoted by $\hat{\mathbb{E}}$ and $\hat{\mathrm{Var}}$ respectively.

Let us notice that, as by definition $\tilde{N}_A(t) \in I_\varepsilon^K$ for all $t \geq 0$, Coupling (3.4.3) with birth and death processes Z_ε^- and Z_ε^+ holds up to time $\tilde{T}_\varepsilon^K$ for the process $\tilde{N}$:

$$Z_\varepsilon^-(t) \leq \tilde{N}_a^K(t) \leq Z_\varepsilon^+(t), \quad \text{for all } t < \tilde{T}_\varepsilon^K. \qquad (3.6.6)$$

The sketch of the proof is the following. We first focus on the neutral proportion in the a population at time $\tilde{T}_\varepsilon^K$. The idea is to consider the neutral alleles of the a individuals at time $\tilde{T}_\varepsilon^K$ and follow their ancestral lines back until the beginning of the sweep, to know whether they are descended from the first mutant or not. Two kinds of events can happen to a neutral lineage : coalescences and m-recombinations (see Section 3.6); we show that we can neglect

the coalescences and the occurrence of several m-recombinations for a lineage during the first period. Therefore, our approximation of the genealogy is the following : two neutral lineages are independent, and each of them undergoes one recombination with an A-individual during the first period with probability ρ_K. If it has undergone a recombination with an A-individual, it can be an allele b_1 or b_2. Otherwise it is descended from the first mutant and is an allele b_1. To get this approximation we follow the approach presented by Schweinsberg and Durrett in [SD05]. In this paper, the authors described the population dynamics by a variation of Moran model with two loci and recombinations. In their model, the population size was constant and each individual has a constant selective advantage, 0 or s. In our model the size is varying and each individual's ability to survive and give birth depends on the population state. After the study of the first period we check that the second and third periods have little influence on the neutral proportion in the a-population.

Coalescence and m-recombination times

Let us introduce the jump times of the stopped Markov process $(\tilde{N}^K(t), t \leq \tilde{T}_\varepsilon^K)$, $0 =: \tau_0^K < \tau_1^K < ... < \tau_{J^K}^K := \tilde{T}_\varepsilon^K$, where J^K denotes the jump number of $\tilde{N}^K$ between 0 and $\tilde{T}_\varepsilon^K$, and the time of the m-th jump is :

$$\tau_m^K = \inf\left\{t > \tau_{m-1}^K, \tilde{N}^K(t) \neq \tilde{N}^K(\tau_{m-1}^K)\right\},\ 1 \leq m \leq J^K.$$

Let us sample two individuals with the a allele uniformly at random at time $\tilde{T}_\varepsilon^K$ and denote by β_p and β_q their neutral alleles. We want to follow their genealogy backward in time and know at each time between 0 and $\tilde{T}_\varepsilon^K$ the types (A or a) of the individuals carrying β_p and β_q.

We say that β_p and β_q coalesce at time τ_m^K if they are carried by two different individuals at time τ_m^K and by the same individual at time τ_{m-1}^K. In other words the individual carrying the allele β_p (or β_q) at time τ_m^K is a newborn and has inherited his neutral allele from the individual carrying allele β_q (or β_p) at time τ_{m-1}^K. The jump number at the coalescence time is denoted by

$$TC^K(\beta_p, \beta_q) := \begin{cases} \sup\{m \leq J^K, \beta_p \text{ and } \beta_q \text{ coalesce at time } \tau_m^K\}, & \text{if } \beta_p \text{ and } \beta_q \text{ coalesce} \\ -\infty, & \text{otherwise.} \end{cases}$$

We say that β_p m-recombines at time τ_m^K if the individual carrying the allele β_p at time τ_m^K is a newborn, carries the allele $\alpha \in \mathscr{A}$, and has inherited his allele β_p from an individual carrying allele $\tilde{\alpha}$. In other words, a m-recombination is a recombination which modifies the selected allele connected to the neutral allele. The jump numbers of the first and second (backward in time) m-recombinations are denoted by :

$$TR_1^K(\beta_p) := \begin{cases} \sup\{m \leq J^K, \beta_p \text{ m-recombines at time } \tau_m^K\}, & \text{if there is at least one m-recombination} \\ -\infty, & \text{otherwise,} \end{cases}$$

$$TR_2^K(\beta_p) := \begin{cases} \sup\{m < TR_1^K(\beta_p), \beta_p \text{ m-recombines at time } \tau_m^K\}, & \text{if there are at least two} \\ & \text{m-recombinations} \\ -\infty, & \text{otherwise.} \end{cases}$$

Let us now focus on the probability for a coalescence to occur conditionally on the state of the process $(\tilde{N}_A, \tilde{N}_a)$ at two successive jump times. We denote by $p^{c_K}_{\alpha_1\alpha_2}(n)$ the probability that the genealogies of two uniformly sampled neutral alleles associated respectively with alleles α_1 and $\alpha_2 \in \mathscr{A}$ at time τ^K_m coalesce at this time conditionally on $(\tilde{N}^K_A(\tau^K_{m-1}), \tilde{N}^K_a(\tau^K_{m-1})) = n \in \mathbb{N}^2$ and on the birth of an individual carrying allele $\alpha_1 \in \mathscr{A}$ at time τ^K_m. Then we have the following result :

Lemma 7. *For every $n = (n_A, n_a) \in \mathbb{N}^2$ and $\alpha \in \mathscr{A}$, we have :*

$$p^{c_K}_{\alpha\alpha}(n) = \frac{2}{n_\alpha(n_\alpha+1)}\left(1 - \frac{r_K f_{\bar{\alpha}} n_{\bar{\alpha}}}{f_A n_A + f_a n_a}\right) \quad \text{and} \quad p^{c_K}_{\alpha\bar{\alpha}}(n) = \frac{r_K f_{\bar{\alpha}}}{(n_\alpha+1)(f_A n_A + f_a n_a)}. \tag{3.6.7}$$

Démonstration. We only state the expression of $p^{c_K}_{\alpha\alpha}(n)$, as the calculations are similar for $p^{c_K}_{\alpha\bar{\alpha}}(n)$. If there is a m-recombination, we cannot have the coalescence of two neutral alleles associated with allele α at time τ^K_m. With probability $1 - r_K f_{\bar{\alpha}} n_{\bar{\alpha}}/(f_A n_A + f_a n_a)$ there is no m-recombination and the parent giving its neutral allele carries the allele α. When there is no m-recombination, two individuals among those who carry allele α also carry a neutral allele which was in the same individual at time τ^K_{m-1}. We have a probability $2/n_\alpha(n_\alpha+1)$ to pick this couple of individuals among the $(n_\alpha+1)$ α-individuals. □

Remark 9. *A m-recombination for a neutral allele associated with an α allele is a coalescence with an $\bar{\alpha}$ individual. Thus if we denote by $p^{r_K}_\alpha(n)$ the probability that an α-individual, chosen uniformly at time τ^K_m, is the newborn and underwent a m-recombination at his birth, conditionally on $(\tilde{N}^K_A(\tau^K_{m-1}), \tilde{N}^K_a(\tau^K_{m-1})) = n \in \mathbb{N}^2$ and on the birth of an individual α at time τ^K_m we get*

$$p^{r_K}_\alpha(n) = n_{\bar{\alpha}} p^{c_K}_{\alpha\bar{\alpha}}(n) = \frac{n_{\bar{\alpha}} r_K f_{\bar{\alpha}}}{(n_\alpha+1)(f_A n_A + f_a n_a)}. \tag{3.6.8}$$

Moreover, if we recall the definition of I^K_ε in (3.2.9), we notice that there exists a finite constant c such that for $k < \lfloor \varepsilon K \rfloor$,

$$(1 - c\varepsilon)\frac{r_K}{k+1} \le \inf_{n_A \in I^K_\varepsilon} p^{r_K}_a(n_A, k) \le \sup_{n_A \in I^K_\varepsilon} p^{r_K}_a(n_A, k) \le \frac{r_K}{k+1}. \tag{3.6.9}$$

Jumps of mutant population during the first period

We want to count the number of coalescences and m-recombinations in the lineages of the two uniformly sampled neutral alleles β_p and β_q. By definition, these events can only occur at a birth time. Thus we need to study the upcrossing number of the process $\tilde{N}^K_a$ before $\tilde{T}^K_\varepsilon$ (Lemma 8). It allows us to prove that the probability that a lineage is affected by two m-recombinations or that two lineages coalesce, and then (backward in time) are affected by a m-recombination is negligible (Lemma 9). Then we obtain an approximation of the probability that a lineage is affected by a m-recombination (Lemma 10), and finally we check that two lineages are approximately independent (Equation (3.6.31)). The last step consists in controlling the neutral proportion in the population A (Lemma 11). Indeed it will give us the probability that a neutral allele which has undergone a m-recombination is a b_1 or a b_2.

Let us denote by ζ_k^K the jump number of last visit to k before the hitting of $\lfloor \varepsilon K \rfloor$,

$$\zeta_k^K := \sup\{m \le J^K, \tilde{N}_a^K(\tau_m^K) = k\}, \quad 1 \le k \le \lfloor \varepsilon K \rfloor. \tag{3.6.10}$$

This allows us to introduce for $0 < j \le k < \lfloor \varepsilon K \rfloor$ the number of upcrossings from k to $k+1$ for the process $\tilde{N}_a^K$ before and after the last visit to j :

$$U_{j,k}^{(K,1)} := \#\{m \in \{0, ..., \zeta_j^K - 1\}, (\tilde{N}_a^K(\tau_m^K), \tilde{N}_a^K(\tau_{m+1}^K)) = (k, k+1)\}, \tag{3.6.11}$$

$$U_{j,k}^{(K,2)} := \#\{m \in \{\zeta_j^K, ..., J^K - 1\}, (\tilde{N}_a^K(\tau_m^K), \tilde{N}_a^K(\tau_{m+1}^K)) = (k, k+1)\}. \tag{3.6.12}$$

We also introduce the number of jumps of the A-population size when there are k a-individuals and the total number of upcrossings from k to $k+1$ before $\tilde{T}_\varepsilon^K$:

$$H_k^K := \#\{m < J^K, \tilde{N}_a^K(\tau_m^K) = \tilde{N}_a^K(\tau_{m+1}^K) = k\}, \tag{3.6.13}$$

$$U_k^K := U_{j,k}^{(K,1)} + U_{j,k}^{(K,2)} = \#\{m < J^K, (\tilde{N}_a^K(\tau_m^K), \tilde{N}_a^K(\tau_{m+1}^K)) = (k, k+1)\}. \tag{3.6.14}$$

The next Lemma states moment properties of these jump numbers. Recall Definition (3.4.1). Then if we define

$$\lambda_\varepsilon := \frac{(1 - s_-(\varepsilon))^3}{(1 - s_+(\varepsilon))^2}, \tag{3.6.15}$$

which belongs to $(0,1)$ for ε small enough, we have

Lemma 8. *There exist two positive and finite constants ε_0 and c such that for $\varepsilon \le \varepsilon_0$, K large enough and $1 \le j \le k < \lfloor \varepsilon K \rfloor$,*

$$\hat{\mathbb{E}}[H_j^K] \le \frac{12 f_A \bar{n}_A K}{s_-^4(\varepsilon) f_a j}, \quad \hat{\mathbb{E}}[(U_{j,k}^{(K,1)})^2] \le \frac{4\lambda_\varepsilon^{k-j}}{s_-^7(\varepsilon)(1 - s_+(\varepsilon))}, \tag{3.6.16}$$

$$\hat{\mathbb{E}}[(U_j^K)^2] \le \frac{2}{s_-^2(\varepsilon)}, \quad \left|\hat{\mathrm{Cov}}(U_{j,k}^{(K,2)}, U_j^K)\right| \le c(\varepsilon + (1 - s_-(\varepsilon))^{k-j}), \tag{3.6.17}$$

and

$$r_K \left| \sum_{k=1}^{\lfloor \varepsilon K \rfloor - 1} \frac{\hat{\mathbb{E}}[U_k^K]}{k+1} - \frac{f_a \log K}{S_{aA}} \right| \le c\varepsilon. \tag{3.6.18}$$

This Lemma is widely used in Sections 3.6 and 3.6. Indeed, we shall decompose on the possible states of the population when a birth occurs, and apply Equations (3.6.7) and (3.6.8) to express the probability of coalescences and m-recombinations at each birth event. The proof of Lemma 8 is quite technical and is postponed to Appendix B.

Negligible events

The next Lemma bounds the probability that two m-recombinations occur in a neutral lineage and the probability that a couple of neutral lineages coalesce and then m-recombine when we consider the genealogy backward in time.

Lemma 9. *There exist two positive finite constants c and ε_0 such that for $K \in \mathbb{N}$ and $\varepsilon \le \varepsilon_0$,*

$$\hat{\mathbb{P}}\Big(TR_2^K(\beta_p) \neq -\infty\Big) \le \frac{c}{\log K}, \quad \textit{and} \quad \hat{\mathbb{P}}\Big(0 \le TR_1^K(\beta_p) \le TC^K(\beta_p, \beta_q)\Big) \le \frac{c}{\log K}.$$

Démonstration. By definition, the neutral allele β_p is associated with an allele a at time $\tilde{T}_\varepsilon^K$. If there are at least two m-recombinations it implies that there exists a time between 0 and $\tilde{T}_\varepsilon^K$ at which β_p has undergone a m-recombination when it was associated with an allele A. We shall work conditionally on the stopped process $((\tilde{N}_A(\tau_m^K), \tilde{N}_a(\tau_m^K)), m \le J^K)$ and decompose according to the a-population size when this m-recombination occurs. We get the inclusion :

$$\{TR_2^K(\beta_p) \neq -\infty\} \subset \bigcup_{k=1}^{\lfloor \varepsilon K \rfloor - 1} \bigcup_{m=1}^{J^K} \Big\{TR_2^K(\beta_p) = m, \tilde{N}_a(\tau_{m-1}^K) = \tilde{N}_a(\tau_m^K) = k\Big\}.$$

We recall the definition of I_ε^K in (3.2.9). Thanks to Equations (3.6.8) and (3.6.16), we get :

$$\hat{\mathbb{P}}(TR_2^K(\beta_p) \neq -\infty) \le \sum_{k=1}^{\lfloor \varepsilon K \rfloor - 1} \sup_{n_A \in I_\varepsilon^K} p_A^{r_K}(n_A, k) \hat{\mathbb{E}}[H_k^K] \le \frac{12 r_K \bar{n}_A \varepsilon}{s_-^4(\varepsilon)(\bar{n}_A - 2\varepsilon C_{A,a}/C_{A,A})^2}.$$

Assumption 4 on weak recombination completes the proof of the first inequality in Lemma 9. The proof of the second one is divided in two steps, presented after introducing, for $(\alpha, \alpha') \in \mathscr{A}^2, m \le J^K$ the notations

$$(\alpha\beta_p)_m := \{\text{the neutral allele } \beta_p \text{ is associated with the allele } \alpha \text{ at time } \tau_m^K\},$$

$$(\alpha\beta_p, \alpha'\beta_q)_m := (\alpha\beta_p)_m \cap (\alpha'\beta_q)_m.$$

First step : We show that the probability that β_p is associated with an allele A at the coalescence time is negligible. We first recall the inclusion,

$$\{TC^K(\beta_p, \beta_q) \neq -\infty, (A\beta_p)_{TC^K(\beta_p,\beta_q)}\} \subset \bigcup_{k=1}^{\lfloor \varepsilon K \rfloor - 1} \bigcup_{m=1}^{J^K} \Big\{TC^K(\beta_p, \beta_q) = m, \tilde{N}_a(\tau_{m-1}^K) = k, (A\beta_p)_m\Big\},$$

and decompose on the possible selected alleles associated with β_q and on the type of the newborn at the coalescence time. Using Lemma 7, Equations (3.6.16) and (3.6.17), and $r_K \le 1$, we get

$$\begin{aligned}
&\hat{\mathbb{P}}(TC^K(\beta_p, \beta_q) \neq -\infty, (A\beta_p)_{TC^K(\beta_p,\beta_q)}) \\
&\le \sum_{k=1}^{\lfloor \varepsilon K \rfloor - 1} \Big[\sup_{n_A \in I_\varepsilon^K} p_{AA}^{c_K}(n_A, k) + \sup_{n_A \in I_\varepsilon^K} p_{Aa}^{c_K}(n_A, k) \Big] \hat{\mathbb{E}}[H_k^K] + \sup_{n_A \in I_\varepsilon^K} p_{aA}^{c_K}(n_A, k) \hat{\mathbb{E}}[U_k^K] \le \frac{c}{K} \sum_{k=1}^{\lfloor \varepsilon K \rfloor - 1} \frac{1}{k},
\end{aligned} \tag{3.6.19}$$

for a finite c, which is of order $\log K/K$.

Second step : Then, we focus on the case where β_p and β_q are associated with an allele a at the coalescence time. The inclusion

$$\left\{\tilde{N}_a\left(\tau^K_{TC^K(\beta_p,\beta_q)-1}\right)=k,(a\beta_p,a\beta_q)_{TC^K(\beta_p,\beta_q)}\right\}\subset\bigcup_{m=1}^{J^K}\left\{TC^K(\beta_p,\beta_q)=m,\tilde{N}_a(\tau^K_{m-1})=k,(a\beta_p,a\beta_q)_m\right\},$$

and Equations (3.6.7) and (3.6.17) yield for every $k\in\{1,...,\lfloor\varepsilon K\rfloor-1\}$:

$$\hat{\mathbb{P}}\left(\tilde{N}_a\left(\tau^K_{TC^K(\beta_p,\beta_q)-1}\right)=k,(a\beta_p,a\beta_q)_{TC^K(\beta_p,\beta_q)}\right)\leq\sup_{n_A\in I^K_\varepsilon}p^{c_K}_{aa}(n_A,k)\hat{\mathbb{E}}[U^K_k]\leq\frac{4}{s^2_-(\varepsilon)k(k+1)}.$$

If β_p and β_q coalesce then undergo their first m-recombination when we look backward in time, and if the a-population has the size k at the coalescence time, it implies that the m-recombination occurs before the ζ^K_k-th jump when we look forward in time. For $k,l<\lfloor\varepsilon K\rfloor$,

$$\begin{aligned}\hat{\mathbb{P}}\left(\tilde{N}_a\left(\tau^K_{TR^K_1(\beta_p)}\right)=l,0\leq TR^K_1(\beta_p)\leq TC^K(\beta_p,\beta_q)\middle|\tilde{N}_a\left(\tau^K_{TC^K(\beta_p,\beta_q)-1}\right)=k,(a\beta_p,a\beta_q)_{TC^K(\beta_p,\beta_q)}\right)\\ \leq\sup_{n_A\in I^K_\varepsilon}p^{r_K}_a(n_A,l)\left(\mathbf{1}_{k>l}\hat{\mathbb{E}}[U^K_l]+\mathbf{1}_{k\leq l}\hat{\mathbb{E}}[U^{(K,1)}_{k,l}]\right)\leq\frac{2r_K}{(l+1)s^2_-(\varepsilon)}\left(\mathbf{1}_{k>l}+\frac{2\mathbf{1}_{k\leq l}\lambda^{l-k}_\varepsilon}{s^5_-(\varepsilon)(1-s_+(\varepsilon))}\right),\end{aligned}$$

where the last inequality is a consequence of (3.6.9), (3.6.16) and (3.6.17). The two last equations finally yield the existence of a finite c such that for every $K\in\mathbb{N}$:

$$\hat{\mathbb{P}}(0\leq TR^K_1(\beta_p)\leq TC^K(\beta_p,\beta_q),(a\beta_p,a\beta_q)_{TC^K(\beta_p,\beta_q)})\leq cr_K\sum_{k,l=1}^{\lfloor\varepsilon K\rfloor}\frac{\mathbf{1}_{k>l}+\mathbf{1}_{k\leq l}\lambda^{l-k}_\varepsilon}{k(k+1)(l+1)}\leq cr_K,$$

which completes the proof of Lemma 9 with Assumption 4. □

Probability to be descended from the first mutant

We want to estimate the probability for the neutral lineage of β_p to undergo no m-recombination. Recall Definition (3.1.17) :

Lemma 10. *There exist two positive finite constants c and ε_0 such that for $\varepsilon\leq\varepsilon_0$:*

$$\limsup_{K\to\infty}\left|\hat{\mathbb{P}}(TR^K_1(\beta_p)=-\infty)-(1-\rho_K)\right|\leq c\varepsilon^{1/2}.$$

Démonstration. We introduce ρ^K_m, the conditional probability that the neutral lineage of β_p m-recombines at time τ^K_m, given $(\tilde{N}_A(\tau^K_n),\tilde{N}_a(\tau^K_n),n\leq J^K)$ and given that it has not m-recombined during the time interval $]\tau^K_m,T^K_\varepsilon]$. The last condition implies that β_p is associated with an allele a at time τ^K_m.

$$\rho^K_m:=\mathbf{1}_{\{\tilde{N}^K_a(\tau^K_m)-\tilde{N}^K_a(\tau^K_{m-1})=1\}}p^{r_K}_a(\tilde{N}^K_A(\tau^K_{m-1}),\tilde{N}^K_a(\tau^K_{m-1})).\tag{3.6.20}$$

We also introduce η^K, the sum of these conditional probabilities for $1 \le m \le J^K$:

$$\eta^K := \sum_{m=1}^{J^K} \rho_m^K.$$

We want to give a rigourous meaning to the sequence of equivalencies :

$$\hat{\mathbb{P}}\Big(TR_1^K(\beta_p) = -\infty \Big| (\tilde{N}_A(\tau_m^K), \tilde{N}_a(\tau_m^K))_{m\le J^K}\Big) = \prod_{m=1}^{J^K}(1-\rho_m^K) \sim \prod_{m=1}^{J^K} e^{-\rho_m^K} \sim e^{-\mathbb{E}[\eta^K]},$$

when K goes to infinity. Jensen's Inequality, the triangle inequality, and the Mean Value Theorem imply

$$\begin{aligned}\hat{\mathbb{E}}\Big|\hat{\mathbb{P}}\Big(TR_1^K(\beta_p) = -\infty \Big| (\tilde{N}_A(\tau_m^K), \tilde{N}_a(\tau_m^K))_{m\le J^K}\Big) - (1-\rho_K)\Big| \le \\ \hat{\mathbb{E}}\Big|\hat{\mathbb{P}}\Big(TR_1^K(\beta_p) = -\infty \Big| (\tilde{N}_A(\tau_m^K), \tilde{N}_a(\tau_m^K))_{m\le J^K}\Big) - e^{-\eta^K}\Big| + \Big|e^{-\hat{\mathbb{E}}\eta^K} - (1-\rho_K)\Big| + \hat{\mathbb{E}}\Big|\eta^K - \hat{\mathbb{E}}\eta^K\Big|.\end{aligned} \tag{3.6.21}$$

We aim to bound the right hand side of (3.6.21). The bounding of the first term follows the method developed in Lemma 3.6 in [SD05]. We refer to this proof, and get the following Poisson approximation

$$\begin{aligned}\hat{\mathbb{E}}\Big|\hat{\mathbb{P}}\Big(TR_1^K(\beta_p) = -\infty \Big| (\tilde{N}_A(\tau_m^K), \tilde{N}_a(\tau_m^K))_{m\le J^K}\Big) - e^{-\eta^K}\Big| &\le \sum_{k=1}^{\lfloor \varepsilon K\rfloor -1} \sup_{n_A\in I_\varepsilon^K} \Big(p_a^{r_K}(n_A,k)\Big)^2 \hat{\mathbb{E}}[U_k^K] \\ &\le \frac{\pi^2 r_K^2}{3s_-^2(\varepsilon)},\end{aligned} \tag{3.6.22}$$

where I_ε^K has been defined in (3.2.9) and the last inequality follows from (3.6.9) and (3.6.17). To bound the second term, we need to estimate $\hat{\mathbb{E}}[\eta^K]$. Inequality (3.6.9) implies

$$(1-c\varepsilon) r_K \sum_{k=1}^{\lfloor \varepsilon K\rfloor -1} \frac{U_k^K}{k+1} \le \eta^K \le r_K \sum_{k=1}^{\lfloor \varepsilon K\rfloor -1} \frac{U_k^K}{k+1}. \tag{3.6.23}$$

Adding (3.6.18) we get that for ε small enough,

$$\limsup_{K\to\infty} \Big|\exp(-\hat{\mathbb{E}}[\eta^K]) - (1-\rho_K)\Big| \le c\varepsilon. \tag{3.6.24}$$

The bounding of the last term of (3.6.21) requires a fine study of dependences between upcrossing numbers before and after the last visit to a given integer by the mutant population size. In particular, we widely use Equation (3.6.17). We observe that $\hat{\mathbb{E}}|\eta^K - \hat{\mathbb{E}}\eta^K| \le (\hat{\mathrm{Var}}\,\eta^K)^{1/2}$, but the variance of η^K is quite involved to study and according to Assumption 4 and Equations (3.6.23) and (3.6.17),

$$\begin{aligned}\Big|\hat{\mathrm{Var}}\,\eta^K - \hat{\mathrm{Var}}\left(r_K \sum_{k=1}^{\lfloor \varepsilon K\rfloor -1} \frac{U_k^K}{k+1}\right)\Big| &\le c\varepsilon \hat{\mathbb{E}}\Big[\Big(r_K \sum_{k=1}^{\lfloor \varepsilon K\rfloor -1} \frac{U_k^K}{k+1}\Big)^2\Big] \\ &\le c\varepsilon r_K^2 \sum_{k,l=1}^{\lfloor \varepsilon K\rfloor -1} \frac{\hat{\mathbb{E}}[(U_k^K)^2] + \hat{\mathbb{E}}[(U_l^K)^2]}{(k+1)(l+1)} \le c\varepsilon,\end{aligned} \tag{3.6.25}$$

for a finite c and K large enough. Let $k \le l < \lfloor \varepsilon K \rfloor$, and recall that by definition, $U_l^K = U_{k,l}^{(K,1)} + U_{k,l}^{(K,2)}$. Then we have

$$\left|\hat{\mathrm{Cov}}(U_k^K, U_l^K)\right| \le \left(\hat{\mathbb{E}}[(U_k^K)^2]\hat{\mathbb{E}}[(U_{k,l}^{(K,1)})^2]\right)^{1/2} + \left|\hat{\mathrm{Cov}}(U_k^K, U_{k,l}^{(K,2)})\right|.$$

Applying Inequalities (3.6.16) and (3.6.17) and noticing that $(1 - s_-(\varepsilon)) < \lambda_\varepsilon^{1/2} < 1$ (recall the definition of λ_ε in (4.5.8)) lead to

$$\left|\hat{\mathrm{Cov}}(U_k^K, U_l^K)\right| \le c(\lambda_\varepsilon^{(l-k)/2} + \varepsilon + (1 - s_-(\varepsilon))^{l-k}) \le c(\lambda_\varepsilon^{(l-k)/2} + \varepsilon) \tag{3.6.26}$$

for a finite c and ε small enough. We finally get :

$$\begin{aligned}\hat{\mathrm{Var}}\Big(r_K \sum_{k=1}^{\lfloor \varepsilon K \rfloor - 1} \frac{U_k^K}{k+1}\Big) &\le 2r_K^2 \sum_{k=1}^{\lfloor \varepsilon K \rfloor - 1} \frac{1}{k+1} \sum_{l=k}^{\lfloor \varepsilon K \rfloor - 1} \frac{\hat{\mathrm{Cov}}(U_k^K, U_l^K)}{l+1} \\ &\le c r_K^2 \sum_{k=1}^{\lfloor \varepsilon K \rfloor - 1} \frac{1}{k+1} \sum_{l=k}^{\lfloor \varepsilon K \rfloor - 1} \frac{\lambda_\varepsilon^{(l-k)/2} + \varepsilon}{l+1} \le c r_K^2 \varepsilon \log^2 K, \end{aligned} \tag{3.6.27}$$

where we used (3.6.26) for the second inequality. Applying Jensen's Inequality to the left hand side of (3.6.21) and adding Equations (3.6.22), (3.6.24), (3.6.25) and (3.6.27) we obtain

$$\begin{aligned}\hat{\mathbb{E}}\Big|\hat{\mathbb{P}}\Big(TR_1^K(\beta_p) = -\infty \Big| (\tilde{N}_A(\tau_m^K), \tilde{N}_a(\tau_m^K))_{m \le J^K}\Big) - (1 - \rho_K)\Big| \\ \le \frac{\pi^2 r_K^2}{3 s_-^2(\varepsilon)} + c\varepsilon + c(\varepsilon + r_K^2 \varepsilon \log^2 K)^{1/2}. \end{aligned} \tag{3.6.28}$$

This completes the proof of Lemma 10. □

We finally focus on the dependence between the genealogies of β_p and β_q, and to this aim follow [SD05] pp. 1622 to 1624 in the case $J = 1$. We define for $m \le J^K$ the random variable

$$K_m = \mathbf{1}_{\{TR_1^K(\beta_p) \ge m\}} + \mathbf{1}_{\{TR_1^K(\beta_q) \ge m\}},$$

which counts the number of neutral lineages which recombine after the m-th jump (forward in time) among the lineages of β_p and β_q. First we will show that for $d \in \{0, 1, 2\}$,

$$\begin{aligned}\Big|\hat{\mathbb{P}}(K_0 = d) - \binom{d}{2} \hat{\mathbb{E}}\Big[\hat{\mathbb{P}}(TR_1^K(\beta_p) \ge 0 | (\tilde{N}_A(\tau_m^K), \tilde{N}_a(\tau_m^K))_{m \le J^K})^d \\ (1 - \hat{\mathbb{P}}(TR_1^K(\beta_p) \ge 0 | (\tilde{N}_A(\tau_m^K), \tilde{N}_a(\tau_m^K))_{m \le J^K})^{2-d}\Big]\Big| \le \frac{c}{\log K}, \end{aligned} \tag{3.6.29}$$

for ε small enough and K large enough, where c is a finite constant. The proof of this inequality can be found in [SD05] pp. 1622-1624 and relies on Equation (A.13). The idea is to couple the process $(K_m, 0 \le m \le J^K)$ with a process $(K'_m, 0 \le m \le J^K)$ satisfying for every $m \le J^K$,

$$\mathscr{L}\Big(K'_{m-1} - K'_m | (\tilde{N}_A(\tau_m^K), \tilde{N}_a(\tau_m^K))_{m \le J^K}, (K'_u)_{m \le u \le J^K}\Big) = Bin(2 - K'_m, \rho_m^K),$$

where $Bin(n,p)$ denotes the binomial distribution with parameters n and p, and ρ_m^K has been defined in (3.6.20). This implies

$$\mathscr{L}\left(K_0' \middle| (\tilde{N}_A(\tau_m^K), \tilde{N}_a(\tau_m^K))_{m\le J^K}\right) = Bin(2, \hat{\mathbb{P}}(TR_1^K(\beta_p) \ge 0 | (\tilde{N}_A(\tau_m^K), \tilde{N}_a(\tau_m^K))_{m\le J^K})),$$

and the coupling yields

$$\hat{\mathbb{P}}(K_m' \neq K_m \text{ for some } 0 \le m \le J^K) \le c/\log K,$$

for ε small enough and K large enough, where c is a finite constant. In particular, the weak dependence between two neutral lineages stated in Lemma 9 is needed in this proof. We now aim at proving that

$$\Big|\hat{\mathbb{E}}[\hat{\mathbb{P}}(TR_1^K(\beta_p) \ge 0 | (\tilde{N}_A(\tau_m^K), \tilde{N}_a(\tau_m^K))_{m\le J^K})^d \\ (1 - \hat{\mathbb{P}}(TR_1^K(\beta_p) \ge 0 | (\tilde{N}_A(\tau_m^K), \tilde{N}_a(\tau_m^K))_{m\le J^K})^{2-d}] - \rho_K^d(1-\rho_K)^{2-d}\Big| \le c\varepsilon^{1/2}, \quad (3.6.30)$$

where we recall the definition of ρ_K in (3.1.17). Equation (A.12) involves

$$\Big|\hat{\mathbb{E}}[\hat{\mathbb{P}}(TR_1^K(\beta_p) \ge 0 | (\tilde{N}_A(\tau_m^K), \tilde{N}_a(\tau_m^K))_{m\le J^K})^d (1 - \hat{\mathbb{P}}(TR_1^K(\beta_p) \ge 0 | (\tilde{N}_A(\tau_m^K), \tilde{N}_a(\tau_m^K))_{m\le J^K})^{2-d}] \\ - \rho_K^d(1-\rho_K)^{2-d}\Big| \le 2\hat{\mathbb{E}}|\hat{\mathbb{P}}(TR_1^K(\beta_p) \ge 0 | (\tilde{N}_A(\tau_m^K), \tilde{N}_a(\tau_m^K))_{m\le J^K}) - \rho_K|.$$

Applying Equation (3.6.28) and adding (3.6.29), we finally get for d in $\{0,1,2\}$,

$$\limsup_{K\to\infty} \left|\hat{\mathbb{P}}(\mathbf{1}_{TR_1^K(\beta_p)\ge 0} + \mathbf{1}_{TR_1^K(\beta_p)\ge 0} = d) - \binom{d}{2}\rho_K^d(1-\rho_K)^{2-d}\right| \le c\varepsilon^{1/2}. \quad (3.6.31)$$

Neutral proportion at time T_ε^K

Let us again focus on the population process N. By abuse of notation, we still use $(TR_i^K(\beta_p), i \in \{1,2\})$ and $TC^K(\beta_p,\beta_q)$ to denote recombination and coalescence times of the neutral genealogies for the process N. According to Lemma 9, Equation (3.6.31), and Coupling (3.6.3),

$$\limsup_{K\to\infty} \mathbb{P}\left(\{TR_2^K(\beta_p) \ge 0\} \cup \{0 \le TR_1^K(\beta_p) \le TC^K(\beta_p,\beta_q)\} \middle| T_\varepsilon^K < \infty\right) \le c\varepsilon,$$

and

$$\limsup_{K\to\infty} \left|\mathbb{P}(\mathbf{1}_{TR_1^K(\beta_p)\ge 0} + \mathbf{1}_{TR_1^K(\beta_p)\ge 0} = d | T_\varepsilon^K < \infty) - \binom{d}{2}\rho_K^d(1-\rho_K)^{2-d}\right| \le c\varepsilon^{1/2},$$

for a finite c and ε small enough. Hence, it is enough to distinguish two cases for the randomly chosen neutral allele β_p : either its lineage has undergone one m-recombination, or no m-recombination. In the second case, β_p is a b_1. In the first one, the probability that β_p is a b_1 depends on the neutral proportion in the A population at the coalescence time. We now state that this proportion stays nearly constant during the first period.

Lemma 11. *There exist two positive finite constants c and ε_0 such that for $\varepsilon \le \varepsilon_0$,*

$$\limsup_{K\to\infty} \mathbb{P}\Big(\sup_{t\le T_\varepsilon^K}\Big|P^K_{A,b_1}(t) - \frac{z_{Ab_1}}{z_A}\Big| > \sqrt{\varepsilon}, T_\varepsilon^K < \infty\Big) \le c\varepsilon.$$

Lemma 11, whose proof is postponed to Appendix B, allows us to state the following lemma.

Lemma 12. *There exist two positive finite constants c and ε_0 such that for $\varepsilon \le \varepsilon_0$,*

$$\limsup_{K\to\infty} \hat{\mathbb{P}}\Big(\Big|P^K_{a,b_2}(T_\varepsilon^K) - \frac{z_{Ab_2}}{z_A}\rho_K\Big| > \varepsilon^{1/6}\Big) \le c\varepsilon^{1/6}.$$

Démonstration. The sequence $(\beta_i, i \le \lfloor \varepsilon K\rfloor)$ denotes the neutral alleles carried by the a-individuals at time T_ε^K and

$$A_2^K(i) := \{\beta_i \text{ has undergone exactly one m-recombination and is an allele } b_2\}.$$

If β_i is a b_2, either its genealogy has undergone one m-recombination with an individual Ab_2, or it has undergone more than two m-recombinations. Thus

$$0 \le N^K_{ab_2}(T_\varepsilon^K) - \sum_{i=1}^{\lfloor \varepsilon K\rfloor} \mathbf{1}_{A_2^K(i)} \le \sum_{i=1}^{\lfloor \varepsilon K\rfloor} \mathbf{1}_{\{TR_2^K(\beta_i)\neq -\infty\}}.$$

Moreover, the probability of $A_2^K(i)$ depends on the neutral proportion in the A-population when β_i m-recombines. For $i \le \lfloor \varepsilon K\rfloor$,

$$\Big|\hat{\mathbb{P}}\Big(A_2^K(i)\Big|TR_1^K(\beta_i)\ge 0, TR_2^K(\beta_i) = -\infty, \sup_{t\le T_\varepsilon^K}\Big|P^K_{A,b_1}(t) - \frac{z_{Ab_1}}{z_A}\Big| \le \sqrt{\varepsilon}\Big) - \Big(1 - \frac{z_{Ab_1}}{z_A}\Big)\Big| \le \sqrt{\varepsilon}. \quad (3.6.32)$$

Lemma 11 and Equation (A.5) ensure that $\limsup_{K\to\infty}\hat{\mathbb{P}}(\sup_{t\le T_\varepsilon^K}|P^K_{A,b_1}(t) - z_{Ab_1}/z_A| > \sqrt{\varepsilon}) \le c\varepsilon$, and Lemmas 9 and 10, and Coupling (3.6.3) that $|\hat{\mathbb{P}}(TR_1^K(\beta_i)\ge 0, TR_2^K(\beta_i) = -\infty) - \rho_K| \le c\varepsilon$. It yields :

$$\Big|\hat{\mathbb{P}}\Big(TR_1^K(\beta_i)\ge 0, TR_2^K(\beta_i) = -\infty, \sup_{t\le T_\varepsilon^K}\Big|P^K_{A,b_1}(t) - \frac{z_{Ab_1}}{z_A}\Big| \le \sqrt{\varepsilon}\Big) - \rho_K\Big| \le c\sqrt{\varepsilon}$$

for a finite c and ε small enough. Adding (3.6.32) we get :

$$\limsup_{K\to\infty}\Big|\hat{\mathbb{E}}[P^K_{a,b_2}(T_\varepsilon^K)] - \rho_K\Big(1 - \frac{z_{Ab_1}}{z_A}\Big)\Big| \le c\sqrt{\varepsilon}. \quad (3.6.33)$$

In the same way, using the weak dependence between lineages stated in (3.6.31) and Coupling (3.6.3), we prove that $\limsup_{K\to\infty}|\hat{\mathbb{E}}[P^K_{a,b_2}(T_\varepsilon^K)^2] - \rho_K^2(1 - z_{Ab_1}/z_A)^2| \le c\sqrt{\varepsilon}$. This implies, adding (3.6.33) that $\limsup_{K\to\infty}\hat{\mathrm{Var}}(P^K_{a,b_2}(T_\varepsilon^K)) \le c\sqrt{\varepsilon}$. We end the proof by using Chebyshev's Inequality. □

Second and third periods

Thanks to Lemma 4 we already know that with high probability the neutral proportion in the a-population stays nearly constant during the third phase. We will prove that this is also the case during the second phase This is due to the short duration of this period, which does not go to infinity with the carrying capacity K.

Lemma 13. *There exist two positive finite constants c and ε_0 such that for $\varepsilon \le \varepsilon_0$,*

$$\limsup_{K\to\infty} \hat{\mathbb{P}}\Big(\Big|P^K_{a,b_1}(T^K_{\text{ext}}) - P^K_{a,b_1}(T^K_\varepsilon)\Big| > \varepsilon^{1/3}\Big) \le c\varepsilon^{1/3}. \tag{3.6.34}$$

Démonstration. Let us introduce the stopping time V^K_ε :

$$V^K_\varepsilon := \inf\Big\{t \le t_\varepsilon, \sup_{\alpha\in\mathscr{A}}\Big|N^K_\alpha(T^K_\varepsilon + t)/K - n_\alpha{}^{(N(T^K_\varepsilon)/K)}(t)\Big| > \varepsilon^3\Big\},$$

where t_ε has been introduced in (3.5.9). Recall that $(\mathscr{F}^K_t, t \ge 0)$ denotes the canonical filtration of N^K. The Strong Markov Property, Doob's Maximal Inequality and Equation (3.2.12) yield :

$$\begin{aligned}
&\mathbb{P}\Big(T^K_\varepsilon \le S^K_\varepsilon, \sup_{t\le t_\varepsilon}|M^K_a(T^K_\varepsilon + t\wedge V^K_\varepsilon) - M^K_a(T^K_\varepsilon)| > \sqrt{\varepsilon}\Big)\\
&\qquad = \mathbb{E}\Big[\mathbf{1}_{T^K_\varepsilon \le S^K_\varepsilon}\mathbb{P}\Big(\sup_{t\le t_\varepsilon}|M^K_a(T^K_\varepsilon + t\wedge V^K_\varepsilon) - M^K_a(T^K_\varepsilon)| > \sqrt{\varepsilon}|\mathscr{F}_{T^K_\varepsilon}\Big)\Big]\\
&\qquad\qquad \le \frac{1}{\varepsilon}\mathbb{E}\Big[\mathbf{1}_{T^K_\varepsilon \le S^K_\varepsilon}\Big(\langle M_a\rangle_{T^K_\varepsilon + t_\varepsilon\wedge V^K_\varepsilon} - \langle M_a\rangle_{T^K_\varepsilon}\Big)\Big] \le \frac{t_\varepsilon C(a, \bar{n}_a + \bar{n}_A)}{\varepsilon(I(\Gamma,\varepsilon) - \varepsilon^3)K},
\end{aligned}$$

for ε small enough, where $I(\Gamma,\varepsilon)$ has been defined in (3.5.9). But according to Equation (3.1.5) with $\delta = \varepsilon^3$, $\limsup_{K\to\infty}\mathbb{P}(V^K_\varepsilon < t_\varepsilon | T^K_\varepsilon \le S^K_\varepsilon) = 0$. Moreover, Equations (3.2.1) and (3.2.6) imply for every $t \ge 0$

$$\sup_{t\le t_\varepsilon}|P^K_{a,b_1}(T^K_\varepsilon + t) - P^K_{a,b_1}(T^K_\varepsilon)| \le \sup_{t\le t_\varepsilon}|M^K_a(T^K_\varepsilon + t) - M^K_a(T^K_\varepsilon)| + r_K t_\varepsilon f_a.$$

As r_K goes to 0 under Assumption 4, we finally get :

$$\limsup_{K\to\infty}\mathbb{P}\Big(\sup_{t\le t_\varepsilon}|P^K_{a,b_1}(T^K_\varepsilon + t) - P^K_{a,b_1}(T^K_\varepsilon)| > \sqrt{\varepsilon}, T^K_\varepsilon \le S^K_\varepsilon\Big) = 0. \tag{3.6.35}$$

Adding Lemma 4 ends the proof of Lemma 13. □

End of the proof of Theorem 2 in the weak recombination regime

Thanks to Lemmas 12 and 13 we get that for ε small enough,

$$\limsup_{K\to\infty}\hat{\mathbb{P}}\Big(\Big|P^K_{a,b_2}(T^K_{\text{ext}}) - \rho_K\frac{z_{Ab_2}}{z_A}\Big| > 2\varepsilon^{1/6}\Big) \le c\varepsilon^{1/6}.$$

Moreover, (A.6) ensures that $\liminf_{K\to\infty}\mathbb{P}(T^K_\varepsilon \le S^K_\varepsilon | \text{Fix}^K) \ge 1 - c\varepsilon$, which implies

$$\limsup_{K\to\infty}\mathbb{P}\Big(\mathbf{1}_{\text{Fix}^K}\Big|P^K_{a,b_2}(T^K_{\text{ext}}) - \rho_K\frac{z_{Ab_2}}{z_A}\Big| > 2\varepsilon^{1/6}\Big) \le c\varepsilon^{1/6}.$$

This is equivalent to the convergence in probability and ends the proof of Theorem 2.

A Technical results

This section is dedicated to technical results needed in the proofs. We first present some results stated in [Cha06]. We recall Definitions (3.1.8), (3.1.9), (3.3.8), (3.3.9), (3.2.10) and (3.5.9) and that the notation $.^K$ refers to the processes that satisfy Assumption 2. Proposition 2 is a direct consequence of Equations (42), (71), (72) and (74) in [Cha06] :

Proposition 2. *There exist two posivite finite constants M_1 and ε_0 such that for every $\varepsilon \leq \varepsilon_0$*

$$\lim_{K\to\infty} \mathbb{P}\left(\left|N_a^K(T_{\text{ext}}^K) - K\bar{n}_a\right| > \varepsilon K \middle| \text{Fix}^K\right) = 0, \quad \textit{and} \quad \limsup_{K\to\infty} \left|\mathbb{P}(T_\varepsilon^K < \infty) - \frac{S_{aA}}{f_a}\right| \leq M_1\varepsilon. \tag{A.1}$$

Moreover there exists $M_2 > 0$ such that for every $\varepsilon \leq \varepsilon_0$, the probability of the event

$$F_\varepsilon^K = \left\{T_\varepsilon^K \leq S_\varepsilon^K, N_A^K(T_\varepsilon^K + t_\varepsilon) < \frac{\varepsilon^2 K}{2}, |N_a^K(T_\varepsilon^K + t_\varepsilon) - \bar{n}_a K| < \frac{\varepsilon K}{2}\right\} \tag{A.2}$$

satisfies

$$\liminf_{K\to\infty} \mathbb{P}(T_\varepsilon^K \leq S_\varepsilon^K) \geq \liminf_{K\to\infty} \mathbb{P}(F_\varepsilon^K) \geq \frac{S_{aA}}{f_a} - M_2\varepsilon, \tag{A.3}$$

and if $z \in \Theta$, then there exist two posivite finite constants V and c such that :

$$\liminf_{K\to\infty} \mathbb{P}(U_\varepsilon^K(z) > e^{VK}) \geq 1 - c\varepsilon. \tag{A.4}$$

Thanks to these results we can state the following Lemma, which motivates the coupling of N and $\tilde{N}$ and allows us to focus on the event $\{\tilde{T}_\varepsilon^K < \infty\}$ rather than on Fix^K in Section 3.6.

Lemma 14. *There exist two posivite finite constants c and ε_0 such that for every $\varepsilon \leq \varepsilon_0$*

$$\limsup_{K\to\infty} \mathbb{P}(T_\varepsilon^K < \infty, T_\varepsilon^K > S_\varepsilon^K) \leq c\varepsilon, \tag{A.5}$$

and

$$\limsup_{K\to\infty} \left[\mathbb{P}(\{T_\varepsilon^K \leq S_\varepsilon^K\} \setminus \text{Fix}^K) + \mathbb{P}(\text{Fix}^K \setminus \{T_\varepsilon^K \leq S_\varepsilon^K\})\right] \leq c\varepsilon. \tag{A.6}$$

Démonstration. We have the following equality

$$\begin{aligned}\mathbb{P}(T_\varepsilon^K < \infty, T_\varepsilon^K > S_\varepsilon^K) &= \mathbb{P}(T_\varepsilon^K < \infty) - \mathbb{P}(T_\varepsilon^K < \infty, T_\varepsilon^K \leq S_\varepsilon^K) \\ &= \mathbb{P}(T_\varepsilon^K < \infty) - \mathbb{P}(T_\varepsilon^K \leq S_\varepsilon^K),\end{aligned}$$

where we used the inclusion $\{T_\varepsilon^K \leq S_\varepsilon^K\} \subset \{T_\varepsilon^K < \infty\}$, as S_ε^K is almost surely finite (a birth and death process with competition has a finite extinction time). Equations (A.1) and (A.3) ends the proof of (A.5). From Equation (A.1), we also have that for $\varepsilon < \bar{n}_a/2$

$$\lim_{K\to\infty} \mathbb{P}(T_\varepsilon^K = \infty | \text{Fix}^K) \leq \lim_{K\to\infty} \mathbb{P}\left(\left|N_a^K(T_{\text{ext}}^K) - K\bar{n}_a\right| > \varepsilon K \middle| \text{Fix}^K\right) = 0. \tag{A.7}$$

This implies that

$$\mathbb{P}(T_\varepsilon^K > S_\varepsilon^K, \text{Fix}^K) \leq \mathbb{P}(T_\varepsilon^K > S_\varepsilon^K, T_\varepsilon^K < \infty) + \mathbb{P}(T_\varepsilon^K = \infty, \text{Fix}^K) \leq c\varepsilon,$$

where we used (A.5). Moreover,

$$\begin{aligned}\mathbb{P}(T_\varepsilon^K \le S_\varepsilon^K, (\mathrm{Fix}^K)^c) &\le \mathbb{P}(T_\varepsilon^K < \infty, (\mathrm{Fix}^K)^c) \\ &= \mathbb{P}(T_\varepsilon^K < \infty) - \mathbb{P}(T_\varepsilon^K < \infty | \mathrm{Fix}^K)\mathbb{P}(\mathrm{Fix}^K) \le c\varepsilon,\end{aligned}$$

where we used (A.1), (A.7) and (3.1.16). □

We also recall some results on birth and death processes whose proofs can be found in Lemma 3.1 in [SD05] and in [AN72] p 109 and 112.

Proposition 3. *Let $Z = (Z_t)_{t\ge 0}$ be a birth and death process with individual birth and death rates b and d. For $i \in \mathbb{Z}^+$, $T_i = \inf\{t \ge 0, Z_t = i\}$ and $\mathbb{P}_i$ (resp. $\mathbb{E}_i$) is the law (resp. expectation) of Z when $Z_0 = i$. Then*

- *For $i \in \mathbb{N}$ and $t \ge 0$,*
$$\mathbb{E}_i[Z_t] = ie^{(b-d)t}. \tag{A.8}$$
- *For $(i,j,k) \in \mathbb{Z}_+^3$ such that $j \in (i,k)$,*
$$\mathbb{P}_j(T_k < T_i) = \frac{1-(d/b)^{j-i}}{1-(d/b)^{k-i}}. \tag{A.9}$$
- *If $d \ne b \in \mathbb{R}_+^*$, for every $i \in \mathbb{Z}_+$ and $t \ge 0$,*
$$\mathbb{P}_i(T_0 \le t) = \left(\frac{d(1-e^{(d-b)t})}{b-de^{(d-b)t}}\right)^i. \tag{A.10}$$
- *If $0 < d < b$, on the non-extinction event of Z, which has probability $1-(d/b)^{Z_0}$, the following convergence holds :*
$$T_N/\log N \underset{N\to\infty}{\to} (b-d)^{-1}, \quad a.s. \tag{A.11}$$

Finally, we recall Lemma 3.4.3 in [Dur08] and Lemma 5.1 in [SD05]. Let $d \in \mathbb{N}$. Then

Lemma 15.
- *Let $a_1, \dots a_d$ and $b_1, \dots, b_d$ be complex numbers of modulus smaller than 1. Then*
$$\left|\prod_{i=1}^d a_i - \prod_{i=1}^d b_i\right| \le \sum_{i=1}^d |a_i - b_i|. \tag{A.12}$$
- *Let V and V' be $\{0,1,\dots,d\}$-valued random variables such that $\mathbb{E}[V] = \mathbb{E}[V']$. Then, there exist random variables $\tilde V$ and $\tilde V'$ on some probability space such that V and $\tilde V$ have the same distribution, V and $\tilde V'$ have the same distribution, and*
$$\mathbb{P}(\tilde V \ne \tilde V') \le d\max\{\mathbb{P}(\tilde V \ge 2), \mathbb{P}(\tilde V' \ge 2)\}. \tag{A.13}$$

For $0 < s < 1$, if $\tilde{Z}^{(s)}$ denotes a random walk with jumps ± 1 where up jumps occur with probability $1/(2-s)$ and down jumps with probability $(1-s)/(2-s)$, we denote by $\mathbb{P}_i^{(s)}$ the law of $\tilde{Z}^{(s)}$ when the initial state is $i \in \mathbb{N}$ and introduce for every $a \in \mathbb{R}_+$ the stopping time

$$\tau_a := \inf\{n \in \mathbb{Z}_+, \tilde{Z}_n^{(s)} = \lfloor a \rfloor\}. \tag{A.14}$$

We also introduce for ε small enough and $0 \le j, k < \lfloor \varepsilon K \rfloor$, the quantities

$$q_{j,k}^{(s_1,s_2)} := \frac{\mathbb{P}_{k+1}^{(s_1)}(\tau_{\varepsilon K} < \tau_k)}{\mathbb{P}_{k+1}^{(s_2)}(\tau_{\varepsilon K} < \tau_j)} = \frac{s_1}{1-(1-s_1)^{\lfloor \varepsilon K \rfloor - k}} \frac{1-(1-s_2)^{\lfloor \varepsilon K \rfloor - j}}{1-(1-s_2)^{k+1-j}}, \quad 0 < s_1, s_2 < 1, \tag{A.15}$$

whose expressions are direct consequences of (A.9). Let us now state a technical result, which helps us to control upcrossing numbers of the process $\tilde{N}_a^K$ before reaching the size $\lfloor \varepsilon K \rfloor$ (see Appendix B).

Lemma 16. *For $a \in]0, 1/2[$, $(s_1, s_2) \in [a, 1-a]^2$, and $0 \le j \le k < l < \lfloor \varepsilon K \rfloor$,*

$$q_{0,k}^{(s_1 \wedge s_2, s_1 \vee s_2)} \ge s_1 \wedge s_2 \quad \textit{and} \quad \left| \frac{1}{q_{k,l}^{(s_1,s_2)}} - \frac{1}{q_{j,l}^{(s_2,s_1)}} \right| \le \frac{4(1+1/s_2)}{ea^2 |\log(1-a)|} |s_2 - s_1| + \frac{(1-s_2)^{l+1-k}}{s_2^3}. \tag{A.16}$$

Démonstration. The first part of (B.3) is a direct consequence of Definiton (A.15). Let a be in $]0, 1/2[$ and consider functions $f_{\alpha,\beta} : x \mapsto (1-x^\alpha)/(1-x^\beta), (\alpha, \beta) \in \mathbb{N}^2, x \in [a, 1-a]$. Then for $x \in [a, 1-a]$,

$$\| f'_{\alpha,\beta} \|_\infty \le 4(ea^2 |\log(1-a)|)^{-1}. \tag{A.17}$$

Indeed, the first derivative of $f_{\alpha,\beta}$ is :

$$f'_{\alpha,\beta}(x) = \frac{\beta x^{\beta-1}(1-x^\alpha) - \alpha x^{\alpha-1}(1-x^\beta)}{(1-x^\beta)^2}.$$

Hence, for $x \in [a, 1-a]$,

$$|f'_{\alpha,\beta}(x)| \le \frac{\beta(1-a)^\beta + \alpha(1-a)^\alpha}{(1-a)a^2} \le 2\frac{\beta(1-a)^\beta + \alpha(1-a)^\alpha}{a^2},$$

where we used that $1 - x^\beta \ge 1 - (1-a)$ and that $1 - a \ge 1/2$. Adding the following inequality

$$\sup_{k \in \mathbb{N}} k(1-a)^k \le \sup_{x \in \mathbb{R}^+} x(1-a)^x = (e|\log(1-a)|)^{-1},$$

completes the proof of (A.17). From (A.9), we get for $0 < s < 1$ and $0 \le j \le k < \lfloor \varepsilon K \rfloor$,

$$\begin{aligned} \left| \mathbb{P}_{l+1}^{(s)}(\tau_{\varepsilon K} < \tau_k) - \mathbb{P}_{l+1}^{(s)}(\tau_{\varepsilon K} < \tau_j) \right| &= \frac{(1-(1-s)^{k-j})((1-s)^{l+1-k} - (1-s)^{\lfloor \varepsilon K \rfloor - k})}{(1-(1-s)^{\lfloor \varepsilon K \rfloor - k})(1-(1-s)^{\lfloor \varepsilon K \rfloor - j})} \\ &\le (1-s)^{l+1-k} s^{-2}. \end{aligned} \tag{A.18}$$

The triangle inequality leads to :

$$\begin{aligned}\left|\frac{1}{q_{k,l}^{(s_1,s_2)}}-\frac{1}{q_{j,l}^{(s_2,s_1)}}\right| &= \left|\frac{\mathbb{P}_{l+1}^{(s_2)}(\tau_{\varepsilon K}<\tau_k)}{\mathbb{P}_{l+1}^{(s_1)}(\tau_{\varepsilon K}<\tau_l)}-\frac{\mathbb{P}_{l+1}^{(s_1)}(\tau_{\varepsilon K}<\tau_j)}{\mathbb{P}_{l+1}^{(s_2)}(\tau_{\varepsilon K}<\tau_l)}\right| \\ &\leq \left|\frac{1}{\mathbb{P}_{l+1}^{(s_1)}(\tau_{\varepsilon K}<\tau_l)}-\frac{1}{\mathbb{P}_{l+1}^{(s_2)}(\tau_{\varepsilon K}<\tau_l)}\right|\mathbb{P}_{l+1}^{(s_2)}(\tau_{\varepsilon K}<\tau_k) \\ &\quad+\frac{1}{\mathbb{P}_{l+1}^{(s_2)}(\tau_{\varepsilon K}<\tau_l)}\left|\mathbb{P}_{l+1}^{(s_2)}(\tau_{\varepsilon K}<\tau_k)-\mathbb{P}_{l+1}^{(s_2)}(\tau_{\varepsilon K}<\tau_j)\right| \\ &\quad+\frac{1}{\mathbb{P}_{l+1}^{(s_2)}(\tau_{\varepsilon K}<\tau_l)}\left|\mathbb{P}_{l+1}^{(s_1)}(\tau_{\varepsilon K}<\tau_j)-\mathbb{P}_{l+1}^{(s_2)}(\tau_{\varepsilon K}<\tau_j)\right|.\end{aligned}$$

Noticing that $\mathbb{P}_{l+1}^{(s_2)}(\tau_{\varepsilon K}<\tau_l)\geq\mathbb{P}_{l+1}^{(s_2)}(\tau_\infty<\tau_l)=\mathbb{P}_1^{(s_2)}(\tau_\infty<\tau_0)=s_2$, and using (A.18) and the Mean Value Theorem with (A.17), we get the second part of (B.3). □

B Proofs of Lemmas 8 and 11

Proof of Equation (3.6.17). In the whole proof, the integer n_A denotes the state of $\tilde{N}_A$ and thus belongs to I_ε^K which has been defined in (3.2.9). $\mathbb{P}_{(n_A,n_a)}$ (resp. $\hat{\mathbb{P}}_{(n_A,n_a)}$) denotes the probability $\mathbb{P}$ (resp. $\hat{\mathbb{P}}$) when $(\tilde{N}_A(0),\tilde{N}_a(0))=(n_A,n_a)\in\mathbb{Z}_+^2$. We introduce for $u\in\mathbb{R}_+$ the hitting time of $\lfloor u\rfloor$ by the process $\tilde{N}_a$:

$$\sigma_u^K:=\inf\{t\geq 0,\tilde{N}_a^K(t)=\lfloor u\rfloor\}. \tag{B.1}$$

Let (i,j,k) be in $\mathbb{Z}_+^3$ with $j<k<\lfloor\varepsilon K\rfloor$. Between jumps ζ_j^K and J^K the process $\tilde{N}_a$ necessarily jumps from k to $k+1$. Then, either it reaches $\lfloor\varepsilon K\rfloor$ before returning to k, either it again jumps from k to $k+1$ and so on. Thus we approximate the probability that there is only one jump from k to $k+1$ by comparing $U_{j,k}^{(K,2)}$ with geometrically distributed random variables. As we do not know the value of $\tilde{N}_A$ when $\tilde{N}_a$ hits $k+1$ for the first time, we take the maximum over all the possible values in I_ε^K. Recall Definition (3.6.4). We get, as $\{\tilde{T}_\varepsilon^K<\sigma_j^K\}\subset\{\tilde{T}_\varepsilon^K<\sigma_k^K\}\subset\{\tilde{T}_\varepsilon^K<\infty\}$:

$$\begin{aligned}\hat{\mathbb{P}}(U_{j,k}^{(K,2)}=1|U_j^K=i) &\leq \sup_{n_A\in I_\varepsilon^K}\hat{\mathbb{P}}_{(n_A,k+1)}(\tilde{T}_\varepsilon^K<\sigma_k^K|\tilde{T}_\varepsilon^K<\sigma_j^K,U_j^K=i) \\ &= \sup_{n_A\in I_\varepsilon^K}\mathbb{P}_{(n_A,k+1)}(\tilde{T}_\varepsilon^K<\sigma_k^K|\tilde{T}_\varepsilon^K<\sigma_j^K,U_j^K=i) \\ &= \sup_{n_A\in I_\varepsilon^K}\frac{\mathbb{P}_{(n_A,k+1)}(\tilde{T}_\varepsilon^K<\sigma_k^K,U_j^K=i)}{\mathbb{P}_{(n_A,k+1)}(\tilde{T}_\varepsilon^K<\sigma_j^K,U_j^K=i)} \\ &= \sup_{n_A\in I_\varepsilon^K}\frac{\mathbb{P}_{(n_A,k+1)}(\tilde{T}_\varepsilon^K<\sigma_k^K)}{\mathbb{P}_{(n_A,k+1)}(\tilde{T}_\varepsilon^K<\sigma_j^K)},\end{aligned}$$

where we used that on the events $\{\tilde{T}_\varepsilon^K<\sigma_j^K\}$ and $\{\tilde{T}_\varepsilon^K<\sigma_k^K\}$ the jumps from j to $j+1$ belong to the past, and Markov Property. Coupling (3.6.6) allows us to compare these conditional

probabilities with the probabilities of the same events under $\mathbb{P}^{(s_-(\varepsilon))}$ and $\mathbb{P}^{(s_+(\varepsilon))}$, and recalling (A.15) we get

$$\hat{\mathbb{P}}(U_{j,k}^{(K,2)} = 1 | U_j^K = i) \leq \frac{\mathbb{P}_{k+1}^{(s_+(\varepsilon))}(\tau_{\varepsilon K} < \tau_k)}{\mathbb{P}_{k+1}^{(s_-(\varepsilon))}(\tau_{\varepsilon K} < \tau_j)} = q_{j,k}^{(s_+(\varepsilon), s_-(\varepsilon))}.$$

In an analogous way we show that $\hat{\mathbb{P}}(U_{j,k}^{(K,2)} = 1 | U_j^K = i) \geq q_{j,k}^{(s_-(\varepsilon), s_+(\varepsilon))}$. We deduce that we can construct two geometrically distributed random variables G_1 and G_2, possibly on an enlarged space, with respective parameters $q_{j,k}^{(s_+(\varepsilon), s_-(\varepsilon))} \wedge 1$ and $q_{j,k}^{(s_-(\varepsilon), s_+(\varepsilon))}$ such that on the event $\{U_j^K = i\}$,

$$G_1 \leq U_{j,k}^{(K,2)} \leq G_2. \tag{B.2}$$

For the same reasons we obtain $q_{j,k}^{(s_-(\varepsilon), s_+(\varepsilon))} \leq \hat{\mathbb{P}}(U_{j,k}^{(K,2)} = 1) \leq q_{j,k}^{(s_+(\varepsilon), s_-(\varepsilon))} \wedge 1$, and again we can construct two random variables $G_1' \overset{d}{=} G_1$ and $G_2' \overset{d}{=} G_2$ such that

$$G_1' \leq U_{j,k}^{(K,2)} \leq G_2'. \tag{B.3}$$

Recall that $U_{0,k}^{(K,2)} = U_k^K$. Hence taking $j = 0$ and adding the first part of Equation (B.3) give the first inequality of (3.6.17). According to Definition (3.4.1), $|s_+(\varepsilon) - s_-(\varepsilon)| \leq c\varepsilon$ for a finite c. Hence Equations (B.2), (B.3) and (B.3) entail the existence of a finite c such that for ε small enough $|\hat{\mathbb{E}}[U_{j,k}^{(K,2)} | U_j^K = i] - \hat{\mathbb{E}}[U_{j,k}^{(K,2)}]| \leq c\varepsilon + (1 - s_-(\varepsilon))^{k+1-j} / s_-^3(\varepsilon)$. Thus according to the first part of Equation (3.6.17),

$$\begin{aligned} \left|\hat{\mathrm{Cov}}(U_{j,k}^{(K,2)}, U_j^K)\right| &\leq \sum_{i \in \mathbb{N}^*} i \hat{\mathbb{P}}(U_j^K = i) \left|\hat{\mathbb{E}}[U_{j,k}^{(K,2)} | U_j^K = i] - \hat{\mathbb{E}}[U_{j,k}^{(K,2)}]\right| \\ &\leq \frac{2}{s_-^2(\varepsilon)} \left(c\varepsilon + \frac{(1 - s_-(\varepsilon))^{k+1-j}}{s_-^3(\varepsilon)} \right), \end{aligned} \tag{B.4}$$

where we use that $U_j^K \leq (U_j^K)^2$. This ends the proof of (3.6.17). □

Proof of Equation (3.6.16). Definitions (3.1.3) and Coupling (3.6.6) ensure that for $n_A \in I_\varepsilon^K$, ε small enough and K large enough,

$$\begin{aligned} \hat{\mathbb{P}}_{(n_A,k)}(\tilde{N}_a(dt) = k+1) &= \frac{\mathbb{P}_{(n_A,k+1)}(\tilde{T}_\varepsilon^K < \infty)}{\mathbb{P}_{(n_A,k)}(\tilde{T}_\varepsilon^K < \infty)} \mathbb{P}_{(n_A,k)}(\tilde{N}_a(dt) = k+1) \\ &\geq \frac{\mathbb{P}_{k+1}^{(s_-(\varepsilon))}(\tau_{\varepsilon K} < \tau_0)}{\mathbb{P}_k^{(s_+(\varepsilon))}(\tau_{\varepsilon K} < \tau_0)} f_a k dt \\ &= \frac{1 - (1 - s_-(\varepsilon))^{k+1}}{1 - (1 - s_-(\varepsilon))^{\lfloor \varepsilon K \rfloor}} \frac{1 - (1 - s_+(\varepsilon))^{\lfloor \varepsilon K \rfloor}}{1 - (1 - s_+(\varepsilon))^k} f_a k dt \\ &\geq s_-^2(\varepsilon) f_a k dt, \end{aligned}$$

and

$$\begin{aligned} \hat{\mathbb{P}}_{(n_A,k)}(\tilde{N}_A(dt) \neq n_A) &\leq \frac{\mathbb{P}_k^{(s_+(\varepsilon))}(\tau_{\varepsilon K} < \tau_0)}{\mathbb{P}_k^{(s_-(\varepsilon))}(\tau_{\varepsilon K} < \tau_0)} \mathbb{P}_{(n_A,k)}(\tilde{N}_A(dt) \neq n_A) \\ &\leq (1 + c\varepsilon) 2 f_A \bar{n}_A K dt \end{aligned}$$

for a finite c, where we use (A.9) and that $D_A + C_{A,A}\bar{n}_A = f_A$. Thus for ε small enough :

$$\hat{\mathbb{P}}(\tilde{N}_a(\tau^K_{m+1}) \neq \tilde{N}_a(\tau^K_m) | \tilde{N}_a(\tau^K_m) = k) \geq \frac{s^2_-(\varepsilon) f_a k}{3 f_A \bar{n}_A K}.$$

If D^K_k denotes the downcrossing number from k to $k-1$ before $\tilde{T}^K_\varepsilon$, then under the probability $\hat{\mathbb{P}}$, we can bound $U^K_k + D^K_k + H^K_k$ by the sum of $U^K_k + D^K_k$ independent geometrically distributed random variables G^K_i with parameter $s^2_-(\varepsilon) f_a k / 3 f_A \bar{n}_A K$ and $H^K_k \leq \sum_{1 \leq i \leq U^K_k + D^K_k} (G^K_i - 1)$. Let us notice that if $k \geq 2$, $D^K_k = U^K_{k-1} - 1$, and $D^K_1 = 0$. Using the first part of (3.6.17) twice we get

$$\hat{\mathbb{E}}[H^K_k] \leq \Big(\frac{4}{s^2_-(\varepsilon)} - 1 \Big) \Big(\frac{3 f_A \bar{n}_A K}{s^2_-(\varepsilon) f_a k} - 1 \Big),$$

which ends the proof of the first inequality in (3.6.16).

As the mutant population size is not Markovian we cannot use symmetry and the Strong Markov Property to control the dependence of jumps before and after the last visit to a given state as in [SD05]. Hence we describe the successive excursions of $\tilde{N}^K_a$ above a given level to get the last inequality in (3.6.16). Let $\tilde{U}^{(i)}_{j,k}$ be the number of jumps from k to $k+1$ during the ith excursion above j. We first bound the expectation $\hat{\mathbb{E}}[(\tilde{U}^{(i)}_{j,k})^2]$. During an excursion above j, $\tilde{N}_a$ hits $j+1$, but we do not know the value of $\tilde{N}_A$ at this time. Thus we take the maximum value for the probability when n_A belongs to I^K_ε, and

$$\hat{\mathbb{P}}(\tilde{U}^{(i)}_{j,k} \geq 1) \leq \sup_{n_A \in I^K_\varepsilon} \hat{\mathbb{P}}_{(j+1,n_A)}(\sigma^K_{k+1} < \sigma^K_j | \sigma^K_j < \tilde{T}^K_\varepsilon).$$

Then using Coupling (3.6.6) and Definition (3.6.4) we obtain

$$\begin{aligned}
\hat{\mathbb{P}}\Big(\tilde{U}^{(i)}_{j,k} \geq 1\Big) &\leq \sup_{n_A \in I^K_\varepsilon} \frac{\mathbb{P}_{(j+1,n_A)}(\sigma^K_{k+1} < \sigma^K_j < \tilde{T}^K_\varepsilon < \infty)}{\mathbb{P}_{(j+1,n_A)}(\sigma^K_j < \tilde{T}^K_\varepsilon < \infty)} \\
&\leq \frac{\mathbb{P}^{(s_+(\varepsilon))}_j(\tau_{\varepsilon K} < \tau_0) \mathbb{P}^{(s_-(\varepsilon))}_{k+1}(\tau_j < \tau_{\varepsilon K}) \mathbb{P}^{(s_+(\varepsilon))}_{j+1}(\tau_{k+1} < \tau_j)}{\mathbb{P}^{(s_-(\varepsilon))}_j(\tau_{\varepsilon K} < \tau_0) \mathbb{P}^{(s_+(\varepsilon))}_{j+1}(\tau_j < \tau_{\varepsilon K})}.
\end{aligned}$$

Adding Equation (A.9) we finally get

$$\hat{\mathbb{P}}\Big(\tilde{U}^{(i)}_{j,k} \geq 1\Big) \leq \frac{(1 - s_-(\varepsilon))^{k+1-j}}{s_-(\varepsilon)(1 - s_+(\varepsilon))}. \tag{B.5}$$

Moreover if $\tilde{U}^{(i)}_{j,k} \geq 1$, $\tilde{N}_a$ necessarily hits k after its first jump from k to $k+1$, and before its return to j. Using the same techniques as before we get :

$$\begin{aligned}
\hat{\mathbb{P}}\Big(\tilde{U}^{(i)}_{j,k} = 1 | \tilde{U}^{(i)}_{j,k} \geq 1\Big) &\geq \inf_{n_A \in I^K_\varepsilon} \hat{\mathbb{P}}_{(n_A,k)}(\sigma^K_j < \sigma^K_{k+1} | \sigma^K_j < \tilde{T}^K_\varepsilon) \\
&\geq \frac{\mathbb{P}^{(s_-(\varepsilon))}_j(\tau_{\varepsilon K} < \tau_0) \mathbb{P}^{(s_+(\varepsilon))}_k(\tau_j < \tau_{k+1})}{\mathbb{P}^{(s_+(\varepsilon))}_j(\tau_{\varepsilon K} < \tau_0) \mathbb{P}^{(s_-(\varepsilon))}_k(\tau_j < \tau_{\varepsilon K})},
\end{aligned}$$

which yields

$$\hat{\mathbb{P}}\Big(\tilde{U}^{(i)}_{j,k} = 1 | \tilde{U}^{(i)}_{j,k} \geq 1\Big) \geq s_-(\varepsilon)s_+(\varepsilon)\Big(\frac{1-s_+(\varepsilon)}{1-s_-(\varepsilon)}\Big)^{k-j} \geq s_-^2(\varepsilon)\Big(\frac{1-s_+(\varepsilon)}{1-s_-(\varepsilon)}\Big)^{k-j} =: q. \tag{B.6}$$

Hence, given that $\tilde{U}^{(i)}_{j,k}$ is non-null, $\tilde{U}^{(i)}_{j,k}$ is smaller than a geometrically distributed random variable with parameter q. In particular,

$$\mathbb{E}\Big[(\tilde{U}^{(i)}_{j,k})^2 | \tilde{U}^{(i)}_{j,k} \geq 1\Big] \leq \frac{2}{q^2} = \frac{2}{s_-^4(\varepsilon)}\Big(\frac{1-s_-(\varepsilon)}{1-s_+(\varepsilon)}\Big)^{2(k-j)}.$$

Adding Equation (B.5) and recalling that $|s_+(\varepsilon) - s_-(\varepsilon)| \leq c\varepsilon$ for a finite c yield

$$\hat{\mathbb{E}}\Big[(\tilde{U}^{(i)}_{j,k})^2\Big] \leq \frac{2\lambda_\varepsilon^{k-j}}{s_-^5(\varepsilon)(1-s_+(\varepsilon))}, \quad \text{where} \quad \lambda_\varepsilon := \frac{(1-s_-(\varepsilon))^3}{(1-s_+(\varepsilon))^2} < 1.$$

Using that for $n \in \mathbb{N}$ and $(x_i, 1 \leq i \leq n) \in \mathbb{R}^n$, $(\sum_{1\leq i\leq n} x_i)^2 \leq n\sum_{1\leq i\leq n} x_i^2$ and that the number of excursions above j before $\tilde{T}^K_\varepsilon$ is $U^K_j - 1$, we get

$$\hat{\mathbb{E}}\Big[(U^{(K,1)}_{j,k})^2\Big] \leq \hat{\mathbb{E}}\Big[U^K_j - 1\Big]\frac{2\lambda_\varepsilon^{k-j}}{s_-^5(\varepsilon)(1-s_+(\varepsilon))} \leq \frac{4\lambda_\varepsilon^{k-j}}{s_-^7(\varepsilon)(1-s_+(\varepsilon))},$$

where we used the first part of Equation (3.6.17). This ends the proof of Equation (3.6.16). □

Proof of Equation (3.6.18). Definition (A.15), Inequality (B.3) and Equation (A.9) yield :

$$r_K \sum_{k=1}^{\lfloor \varepsilon K\rfloor - 1} \frac{\hat{\mathbb{E}}[U^K_k]}{k+1} \geq r_K \sum_{k=1}^{\lfloor \varepsilon K\rfloor - 1} \Big[(k+1)q_{0,k}^{(s_+(\varepsilon),s_-(\varepsilon))}\Big]^{-1} = \frac{r_K(A-B)}{s_+(\varepsilon)(1-(1-s_-(\varepsilon))^{\lfloor \varepsilon K\rfloor})},$$

with

$$A := \sum_{k=1}^{\lfloor \varepsilon K\rfloor - 1} \frac{1-(1-s_-(\varepsilon))^{k+1}}{k+1}, \quad \text{and} \quad B := (1-s_+(\varepsilon))^{\lfloor \varepsilon K\rfloor} \sum_{k=1}^{\lfloor \varepsilon K\rfloor - 1} \frac{1-(1-s_-(\varepsilon))^{k+1}}{(1-s_+(\varepsilon))^k(k+1)}.$$

For large K, $A = \log(\varepsilon K) + O(1)$, and for every $u > 1$ there exists $D(u) < \infty$ such that $\sum_{k=1}^{\lfloor \varepsilon K\rfloor} u^k/(k+1) \leq D(u)u^{\lfloor \varepsilon K\rfloor}/\lfloor \varepsilon K\rfloor$. This implies that $B \leq c/\lfloor \varepsilon K\rfloor$ for a finite c. Finally, by definition, for ε small enough, $|s_+(\varepsilon) - S_{aA}/f_a| \leq c\varepsilon$ for a finite constant c. This yields

$$r_K \sum_{k=1}^{\lfloor \varepsilon K\rfloor - 1} \frac{\hat{\mathbb{E}}[U^K_k]}{k+1} \geq (1-c\varepsilon)\frac{r_K f_a \log K}{S_{aA}}$$

for a finite c and concludes the proof for the lower bound. The upper bound is obtained in the same way. This ends the proof of Lemma 8. □

Proof of Lemma 11. We use Coupling (3.4.3) to control the growing of the mutant population during the first period of invasion, and the semi-martingale decomposition in Proposition 1

to bound the fluctuations of M_A. The hitting time of $\lfloor \varepsilon K \rfloor$ and non-extinction event of Z^*_ε are denoted by :

$$T^{*,K}_\varepsilon = \inf\{t \geq 0, Z^*_\varepsilon(t) = \lfloor \varepsilon K \rfloor\}, \quad \text{and} \quad F^*_\varepsilon = \Big\{ Z^*_\varepsilon(t) \geq 1, \forall\, t \geq 0 \Big\}, \quad * \in \{-,+\}.$$

Let us introduce the difference of probabilities

$$B^K_\varepsilon := \mathbb{P}\Big(\sup_{t \leq T^K_\varepsilon} \Big| P^K_{A,b_1}(t) - \frac{z_{Ab_1}}{z_A} \Big| > \sqrt{\varepsilon}, T^K_\varepsilon < \infty \Big) - \mathbb{P}\Big(\sup_{t \leq T^K_\varepsilon} \Big| P^K_{A,b_1}(t) - \frac{z_{Ab_1}}{z_A} \Big| > \sqrt{\varepsilon}, F^-_\varepsilon, T^K_\varepsilon \leq S^K_\varepsilon \Big).$$

Then B^K_ε is nonnegative and we have

$$\begin{aligned} B^K_\varepsilon &\leq \mathbb{P}(T^K_\varepsilon < \infty) - \mathbb{P}(F^-_\varepsilon, T^K_\varepsilon \leq S^K_\varepsilon) \\ &= \mathbb{P}(T^K_\varepsilon < \infty) - \mathbb{P}(T^K_\varepsilon \leq S^K_\varepsilon) + \mathbb{P}(T^{(+,K)}_\varepsilon < \infty, T^K_\varepsilon \leq S^K_\varepsilon) - \mathbb{P}(F^-_\varepsilon, T^K_\varepsilon \leq S^K_\varepsilon), \end{aligned}$$

where the inequality comes from the inclusion $\{F^-_\varepsilon, T^K_\varepsilon \leq S^K_\varepsilon\} \subset \{T^K_\varepsilon < \infty\}$, as S^K_ε is almost surely finite. The equality is a consequence of Coupling (3.4.3) which ensures that on the event $\{T^K_\varepsilon \leq S^K_\varepsilon\}$, $\{T^{(+,K)}_\varepsilon < \infty\}$ holds. By noticing that

$$\{F^-_\varepsilon, T^K_\varepsilon \leq S^K_\varepsilon\} \subset \{T^{(-,K)}_\varepsilon < \infty, T^K_\varepsilon \leq S^K_\varepsilon\} \subset \{T^{(+,K)}_\varepsilon < \infty, T^K_\varepsilon \leq S^K_\varepsilon\}$$

we get the bound

$$B^K_\varepsilon \leq \mathbb{P}(T^K_\varepsilon < \infty) - \mathbb{P}(T^K_\varepsilon \leq S^K_\varepsilon) + \mathbb{P}(T^{(+,K)}_\varepsilon < \infty) - \mathbb{P}(F^-_\varepsilon). \tag{B.7}$$

The values of the two first probabilities are approximated in (A.1) and (A.3), and (A.9) implies that $\mathbb{P}(T^{+,K}_\varepsilon < \infty) - \mathbb{P}(F^-_\varepsilon) = s_+(\varepsilon)/(1-(1-s_+(\varepsilon))^{\lfloor \varepsilon K \rfloor}) - s_-(\varepsilon)$. Hence

$$\limsup_{K \to \infty} B^K_\varepsilon \leq c\varepsilon, \tag{B.8}$$

where c is finite for ε small enough, which allows us to focus on the intersection with the event $\{F^-_\varepsilon, T^K_\varepsilon \leq S^K_\varepsilon\}$. We recall that $|N_{Ab_1}N_{ab_2} - N_{Ab_2}N_{ab_1}| \leq N_A N_a$, and that Assumption 4 holds. Then (3.2.1) and (3.2.10) imply for ε small enough

$$\sup_{t \leq T^K_\varepsilon \wedge S^K_\varepsilon} \Big| P_{A,b_1}(t) - \frac{z_{Ab_1}}{z_A} - M_A(t) \Big| \leq r_K f_a T^K_\varepsilon \sup_{t \leq T^K_\varepsilon \wedge S^K_\varepsilon} \Big\{ \frac{N_a(t)}{N_A(t)} \Big\} \leq \frac{r_K f_a \varepsilon T^K_\varepsilon}{\bar{n}_A - 2\varepsilon C_{A,a}/C_{A,A}} \leq \frac{c\varepsilon T^K_\varepsilon}{\log K},$$

for a finite c. Moreover, $F^-_\varepsilon \cap \{T^K_\varepsilon \leq S^K_\varepsilon\} \subset F^-_\varepsilon \cap \{T^K_\varepsilon \leq T^{(-,K)}_\varepsilon\}$. Thus we get

$$\mathbb{P}\Big(\sup_{t \leq T^K_\varepsilon} \Big| P_{A,b_1}(t) - \frac{z_{Ab_1}}{z_A} - M_A(t) \Big| > \frac{\sqrt{\varepsilon}}{2}, F^-_\varepsilon, T^K_\varepsilon \leq S^K_\varepsilon \Big) \leq \mathbb{P}\Big(\frac{c\varepsilon T^{(-,K)}_\varepsilon}{\log K} > \sqrt{\varepsilon}/2, F^-_\varepsilon \Big).$$

Finally, Equation (A.11) ensures that $\lim_{K\to\infty} T^{(-,K)}_\varepsilon / \log K = s_-(\varepsilon)^{-1}$ a.s. on the non-extinction event F^-_ε. Thus for ε small enough,

$$\lim_{K \to \infty} \mathbb{P}\Big(\sup_{t \leq T^K_\varepsilon} \Big| P_{A,b_1}(t) - \frac{z_{Ab_1}}{z_A} - M_A(t) \Big| > \frac{\sqrt{\varepsilon}}{2}, F^-_\varepsilon, T^K_\varepsilon \leq S^K_\varepsilon \Big) = 0. \tag{B.9}$$

To control the term $|M_A|$, we introduce the sequence of real numbers $t_K = (2f_a \log K)/S_{aA}$:

$$\mathbb{P}\Big(\sup_{t\le T_\varepsilon^K}|M_A(t)|>\frac{\sqrt{\varepsilon}}{2},F_\varepsilon^-,T_\varepsilon^K\le S_\varepsilon^K\Big)\le\mathbb{P}\Big(\sup_{t\le T_\varepsilon^K}|M_A(t)|>\frac{\sqrt{\varepsilon}}{2},T_\varepsilon^K\le S_\varepsilon^K\wedge t_K\Big)+\mathbb{P}(T_\varepsilon^K>t_K,F_\varepsilon^-).$$

Equation (3.4.1) yields for ε small enough, $t_K.s_-(\varepsilon)/\log K>3/2$. Thus thanks to (A.11) we get,

$$\lim_{K\to\infty}\mathbb{P}(T_\varepsilon^K>t_K,F_\varepsilon^-)\le\lim_{K\to\infty}\mathbb{P}(T_\varepsilon^{-,K}>t_K,F_\varepsilon^-)=0.$$

Applying Doob's maximal inequality to the submartingale $|M_A|$ and (3.2.12) we get :

$$\begin{aligned}\mathbb{P}(\sup_{t\le T_\varepsilon^K}|M_A(t)|>\sqrt{\varepsilon}/2,T_\varepsilon^K\le S_\varepsilon^K\wedge t_K) &\le \mathbb{P}(\sup_{t\le t_K}|M_A(t\wedge T_\varepsilon^K\wedge S_\varepsilon^K)|>\sqrt{\varepsilon}/2)\\ &\le \frac{4}{\varepsilon}\mathbb{E}\Big[\langle M_A\rangle_{t_K\wedge T_\varepsilon^K\wedge S_\varepsilon^K}\Big]\\ &\le \frac{4}{\varepsilon}\frac{C(A,2\bar{n}_A)t_K}{(\bar{n}_A-2\varepsilon C_{A,a}/C_{A,A})K},\end{aligned}$$

which goes to 0 at infinity. Adding Equation (B.9) leads to :

$$\lim_{K\to\infty}\mathbb{P}\Big(\sup_{t\le T_\varepsilon^K}\Big|P_{A,b_1}(t)-\frac{z_{Ab_1}}{z_A}\Big|>\sqrt{\varepsilon},F_\varepsilon^-,T_\varepsilon^K\le S_\varepsilon^K\Big)=0.$$

Finally, Equation (B.8) complete the proof of Lemma 11. □

CHAPITRE 4

Genealogies of two linked neutral loci after a selective sweep in a large population of varying size

Introduction

We study the hitchhiking effect of a beneficial mutation in a sexual haploid population of varying size. We assume that a mutation occurs in one individual of a monomorphic population and that individuals carrying the new allele a are better adapted to the current environment and spread in the population. We suppose that the mutant allele a eventually replaces the resident one, A, and study the influence of this fixation on the neutral gene genealogy of a sample taken at the end of the selective sweep. That is, in each sampled individual we consider the same set of partially linked loci including the locus where the advantageous mutation occurred. We then trace back the ancestral lineages of all loci in the sample until the beginning of the sweep and update the genetic relationships whenever a coalescence or a recombination (see Definition 5) changes the ancestry of one or several loci. Our main result is the derivation of a sampling formula for the ancestral partition of two neutral loci situated in the vicinity of the selected allele. Such a result allows us to derive the expression of the linkage disequilibrium (see Section 4.2) which is more and more used to understand past evolutionary and demographic events [KN04, Sla08].

The first studies of hitchhiking, initiated by Maynard Smith and Haigh [SH74], have modeled the mutant population size as the solution of a deterministic logistic equation [OK75, KHL89, SWL92, SSL06]. Barton [Bar98] was the first to point out the importance of the stochasticity of the mutant population size. Following this paper, a series of works took into account this randomness during the sweep. In [DS04, SD05] Schweinsberg and Durrett based their analysis on a Moran model with selection and recombination, while Etheridge and coauthors [EPW06] worked with the diffusion limit of such discrete population models. Then Brink-Spalink and Sturm [BSS15b], Pfaffelhuber and Studeny [PS07] and Leocard [Leo09] extended the respective findings of these two approaches for the ancestry of one neutral locus to the two-locus (resp. multiple-locus) case.

However, in all these models, the population size was constant and each individual had a "fitness" only dependent on its type and not on the population state. The fundamental idea of Darwin is that the individual traits have an influence on the interactions between individuals, which in turn generate selection on the different traits. In this paper we aim at modeling precisely these interactions by extending the model introduced in Chapter 3 where we only considered one neutral locus. Such an eco-evolutionary approach has been introduced by Metz and coauthors [MGM+96] and made rigorous in the seminal paper of Fournier and Méléard [FM04]. Then they have been developed by Champagnat, Méléard and coauthors (see [Cha06, CM11, CJM14] and references therein) for the haploid asexual case and by Collet, Méléard and Metz [CMM11] and Coron and coauthors [Cor14, Cor13] for the diploid sexual case.

The population dynamics, described in Section 4.1, is a multitype birth and death Markov process with competition. We reflect the carrying capacity of the underlying environment by a scaling parameter $K \in \mathbb{N}$ and state results in the limit for large K. In [Cha06] it was shown that such kind of invasion processes can be divided into three phases (see Figure 4.2) : an initial phase in which the fraction of a-individuals does not exceed a fixed value $\varepsilon > 0$ and where the dynamics of the wild-type population is nearly undisturbed by the invading type. A second phase where both types account for a non-negligible percentage of the population and where the evolution of the population can be well approximated by a deterministic competitive Lotka-Volterra system. And finally a third phase where the roles of the types are interchanged and the wild-type population is near extinction. The durations of the first and third phases of the selective sweep are of order $\log K$ whereas the second phase only lasts an amount of time of order 1.

In Section 4.3 we precisely describe these three phases and introduce two couplings of the population process, key tools to study the dynamics of the A- and a-population dynamics. Section 4.4 is devoted to the proofs of the main theorems on the ancestral partition of the two neutral alleles. Sections 4.5 to 4.7 are dedicated to the proofs of auxiliary statements. Finally, we state technical results needed in the proofs in the Appendix.

4.1 Model and results

We consider a three locus model : one locus under selection, SL, with alleles in $\mathscr{A} := \{A, a\}$ and two neighboring neutral loci $N1$ and $N2$ with alleles in the finite sets $\mathscr{B}$ and $\mathscr{C}$ respectively. We denote by $\mathscr{E} = \mathscr{A} \times \mathscr{B} \times \mathscr{C}$ the type space. Two geometric alignments are possible : either the two neutral loci are adjacent (geometry $SL - N1 - N2$), either they are separated by the selected locus (geometry $N1 - SL - N2$). We introduce the model and notations for the adjacent geometry and their analogues for the separated one can be deduced straightforward.

Whenever a reproduction event takes place, recombinations between SL and $N1$ or between $N1$ and $N2$ occur independently with probabilities r_1 and r_2, respectively. These probabilities depend on the parameter K, representing the environment's carrying capacity, but for the purpose of readability we do not indicate this dependence. We assume a regime of weak

recombination :

$$\limsup_{K\to\infty} r_j \log K < \infty, \quad j = 1,2. \tag{4.1.1}$$

This is motivated by Theorem 2 in Chapter 3 which states that this is the good scale to observe a signature on the neutral allele distribution. If the recombination probabilities are larger (which means that neutral loci are more distant from the selected locus), there are many recombinations and the sweep does not modify the neutral diversity at these sites. Recombinations may lead to a mixing of the parental genetic material in the newborn, and hence, parents with types (α,β,γ) and (α',β',γ') in $\mathscr{E}$ can generate the following offspring :

possible genotype	event	probability
$\alpha\beta\gamma, \alpha'\beta'\gamma'$	no recombination	$(1-r_1)(1-r_2)$
$\alpha\beta'\gamma', \alpha'\beta\gamma$	one recombination between SL and $N1$	$r_1(1-r_2)$
$\alpha\beta\gamma', \alpha'\beta'\gamma$	one recombination between $N1$ and $N2$	$(1-r_1)r_2$
$\alpha\beta'\gamma, \alpha'\beta\gamma'$	two recombinations	$r_1 r_2$

We will see in the sequel that the probability to witness a birth event with two simultaneous recombinations in the neutral genealogy of a uniformly chosen individual is very small.

As we assume the loci $N1$ and $N2$ to be neutral, the ecological parameters of an individual only depend on the allele α at the locus under selection. Let us denote by f_α the fertility of an individual with type α. In the spirit of [CMM11], such an individual gives birth at rate f_α (female role), and has a probability proportional to f_α to be chosen as the father in a given birth event (male role). Addressing the complementary type of the allele α by $\bar{\alpha}$ we get the following result for the birth rate of individuals of type $(\alpha,\beta,\gamma) \in \mathscr{E}$:

$$\begin{aligned} b^K_{\alpha\beta\gamma}(n) \;=\; & (1-r_1)(1-r_2) f_\alpha n_{\alpha\beta\gamma} + r_1(1-r_2) f_\alpha n_\alpha \frac{f_\alpha n_{\alpha\beta\gamma} + f_{\bar{\alpha}} n_{\bar{\alpha}\beta\gamma}}{f_a n_a + f_A n_A} \\ & + (1-r_1) r_2 f_\alpha \frac{\sum_{\beta'\in\mathscr{B}} \sum_{\gamma'\in\mathscr{C}} n_{\alpha\beta\gamma'} (f_\alpha n_{\alpha\beta'\gamma} + f_{\bar{\alpha}} n_{\bar{\alpha}\beta'\gamma})}{f_a n_a + f_A n_A} \\ & + r_1 r_2 f_\alpha \frac{\sum_{\beta'\in\mathscr{B}} \sum_{\gamma'\in\mathscr{C}} n_{\alpha\beta'\gamma} (f_\alpha n_{\alpha\beta\gamma'} + f_{\bar{\alpha}} n_{\bar{\alpha}\gamma'})}{f_a n_a + f_A n_A}, \end{aligned} \tag{4.1.2}$$

where $n_{\alpha\beta\gamma}$ (resp. n_α) denotes the current number of $\alpha\beta\gamma$-individuals (resp. α-individuals) and

$$n = (n^{(A)}, n^{(a)}) = ((n_{A\beta\gamma}, (\beta,\gamma) \in \mathscr{B}\times\mathscr{C}), (n_{a\beta\gamma}, (\beta,\gamma) \in \mathscr{B}\times\mathscr{C})) \tag{4.1.3}$$

is the current state of the population. An α-individual can die either from a natural death (rate D_α), either from type-dependent competition : the parameter $C_{\alpha,\alpha'}$ models the impact an individual of type α' has on an individual of type α, where $(\alpha,\alpha') \in \mathscr{A}^2$. The strength of the competition also depends on the carrying capacity K. This results in the total death rate of individuals carrying the alleles $(\alpha,\beta,\gamma) \in \mathscr{E}$:

$$d^K_{\alpha\beta\gamma}(n) = \left(D_\alpha + \frac{C_{\alpha,A}}{K} n_A + \frac{C_{\alpha,a}}{K} n_a \right) n_{\alpha\beta\gamma}. \tag{4.1.4}$$

Hence the population process

$$N^K = (N^K(t), t \geq 0) = \left((N^K_{\alpha\beta\gamma}(t))_{(\alpha,\beta,\gamma)\in\mathscr{E}}, t \geq 0\right),$$

where $N^K_{\alpha\beta\gamma}(t)$ denotes the number of $\alpha\beta\gamma$-individuals at time t, is a multitype birth and death process with rates given in (4.1.2) and (4.1.4). We will often work with the trait population process $((N^K_A(t), N^K_a(t)), t \geq 0)$, where $N^K_\alpha(t)$ denotes the number of α-individuals at time t. This is also a birth and death process with birth and death rates given by :

$$b^K_\alpha(n) = \sum_{(\beta,\gamma)\in\mathscr{B}\times\mathscr{C}} b^K_{\alpha\beta\gamma}(n) = f_\alpha n_\alpha \quad \text{and} \quad d^K_\alpha(n) = \sum_{(\beta,\gamma)\in\mathscr{B}\times\mathscr{C}} d^K_{\alpha\beta\gamma}(n) = \left(D_\alpha + \frac{C_{\alpha,A}}{K} n_A + \frac{C_{\alpha,a}}{K} n_a\right) n_\alpha.$$

As a quantity summarizing the advantage or disadvantage a mutant with allele type α has in an $\bar{\alpha}$-population at equilibrium, we introduce the so-called invasion fitness $S_{\alpha\bar{\alpha}}$ through

$$S_{\alpha\bar{\alpha}} := f_\alpha - D_\alpha - C_{\alpha,\bar{\alpha}} \bar{n}_{\bar{\alpha}}, \tag{4.1.5}$$

where the equilibrium density $\bar{n}_\alpha$ is defined by

$$\bar{n}_\alpha := \frac{f_\alpha - D_\alpha}{C_{\alpha,\alpha}}.$$

The role of the invasion fitness $S_{\alpha\bar{\alpha}}$ and the definition of the equilibrium density $\bar{n}_\alpha$ follow from the properties of the two-dimensional competitive Lotka-Volterra system :

$$\dot{n}^{(z)}_\alpha = (f_\alpha - D_\alpha - C_{\alpha,A} n^{(z)}_A - C_{\alpha,a} n^{(z)}_a) n^{(z)}_\alpha, \quad z \in \mathbb{R}^{\mathscr{A}}_+, \quad n^{(z)}_\alpha(0) = z_\alpha, \quad \alpha \in \mathscr{A}. \tag{4.1.6}$$

If we assume

$$\bar{n}_A > 0, \quad \bar{n}_a > 0, \quad \text{and} \quad S_{Aa} < 0 < S_{aA}, \tag{4.1.7}$$

then $\bar{n}_\alpha$ is the equilibrium size of a monomorphic α-population and the system (4.1.6) has a unique stable equilibrium $(0, \bar{n}_a)$ and two unstable steady states $(\bar{n}_A, 0)$ and $(0,0)$. Thanks to Theorem 2.1 p. 456 in [EK86] we can prove that if $N^K_A(0)$ and $N^K_a(0)$ are of order K and K is large, the rescaled process $(N^K_A/K, N^K_a/K)$ is very close to the solution of (4.1.6) during any finite time interval. The invasion fitness S_{aA} corresponds to the *per capita* initial growth rate of the mutant a when it appears in a monomorphic population of individuals A at their equilibrium size $\bar{n}_A K$. Hence the dynamics of the allele a is very dependent on the properties of the system (4.1.6) and it is proven in [Cha06] that under Condition (4.1.7) one mutant a has a positive probability to fix in the population and replace a wild-type A. More precisely, if we use the convention

$$\mathbb{P}(.) := \mathbb{P}(.|N^K_A(0) = \lfloor \bar{n}_A K \rfloor, N^K_a(0) = 1), \tag{4.1.8}$$

Equation (39) in [Cha06] states that

$$\lim_{K\to\infty} \mathbb{P}(\text{Fix}^K) = \frac{S_{aA}}{f_a} =: s, \tag{4.1.9}$$

where s is called rescaled invasion fitness, and the extinction time of the A-population and the event of fixation of the a-allele are rigorously defined as follows :

$$T_{\text{ext}}^K := \inf\{t \geq 0 : N_A^K(t) = 0\}, \quad \text{and} \quad \text{Fix}^K := \{T_{\text{ext}}^K < \infty, N_a^K(T_{\text{ext}}^K) > 0\}. \tag{4.1.10}$$

Let d be in $\mathbb{N}$. We aim at quantifying the effect of the selective sweep on the neutral diversity. Our method consists in tracing back the neutral genealogies of d individuals sampled uniformly at the end of the sweep (time T_{ext}^K) until the beginning of the sweep. Two event types (see Definition 5) may affect the relationships of the sampled neutral alleles : coalescences correspond to the merging of the neutral genealogies of two individuals at one or two neutral loci, and recombinations redistribute the selected and neutral alleles of one individual into two groups carried by its two parents. We will represent the neutral genealogies by a partition Θ_d^K which belongs to the set $\mathscr{P}_d^*$ of marked partitions of $\{(i,k), i \in \{1,...,d\}, k \in \{1,2\}\}$ with (at most) one block distinguished by the mark $*$. In this notation $(i,1)$ (resp. $(i,2)$) is the neutral allele at locus $N1$ (resp. $N2$) of the ith sampled individual. Let us define rigorously the random partition Θ_d^K :

Definition 1. *Let d be in $\mathbb{N}$ and sample d individuals uniformly and without replacement at the end of the sweep (time T_{ext}^K). Follow the genealogies of the first and second neutral alleles of the i-th sampled individual, $(i,1)$ and $(i,2)$ for $i \in \{1,...,d\}$. Then the partition $\Theta_d^K \in \mathscr{P}_d^*$ is defined as follows : each block of the partition Θ_d^K is composed of all those neutral alleles which originate from the same given individual alive at the beginning of the sweep ; the block containing the descendants of the mutant a (if such a block exists) is distinguished by the mark $*$.*

We will show in Theorems 1 and 2 that when K is large the partition Θ_d^K belongs with a probability close to one to a subset Δ_d of $\mathscr{P}_d^*$, which is defined as follows :

Definition 2. *Let d be in $\mathbb{N}$. Δ_d is the subset of $\mathscr{P}_d^*$ consisting of those partitions whose unmarked blocks (if there are any) are either singletons either pairs of the form $\{(i,1),(i,2)\}$ for one $i \in \{1,...,d\}$.*

A d-sample whose genealogy is represented by a partition in Δ_d satisfies the following property : two neutral alleles of two distinct sampled individuals cannot originate from the same A-individual alive at the beginning of the sweep.

Example 1. *In the example represented in Figure 4.1, the marked partition $\pi^{(ex)}$ belongs to Δ_d :*

$$\pi^{(ex)} = \Big\{\{(1,1),(1,2),(2,1),(5,2)\}^*, \{(2,2)\}, \{(3,1),(3,2)\}, \{(4,1)\}, \{(4,2)\}, \{(5,1)\}\Big\}.$$

For a partition $\pi \in \mathscr{P}_d^*$, we define for some possible ancestral relationships the number of individuals in the sample whose two neutral loci are related in that particular way :

Definition 3. *Let $d \in \mathbb{N}$ and $\pi \in \mathscr{P}_d^*$. Then we set :*

$|\pi|_1 = \#\{\ 1 \leq i \leq d$ *such that* $(i,1)$ *and* $(i,2)$ *belong to the marked block* $\}$

$|\pi|_2 = \#\{\ 1 \leq i \leq d$ *such that* $(i,1)$ *belongs to the marked block and* $\{(i,2)\}$ *is an unmarked block*$\}$

$|\pi|_3 = \#\{\ 1 \leq i \leq d$ *such that* $(i,2)$ *belongs to the marked block and* $\{(i,1)\}$ *is an unmarked block*$\}$

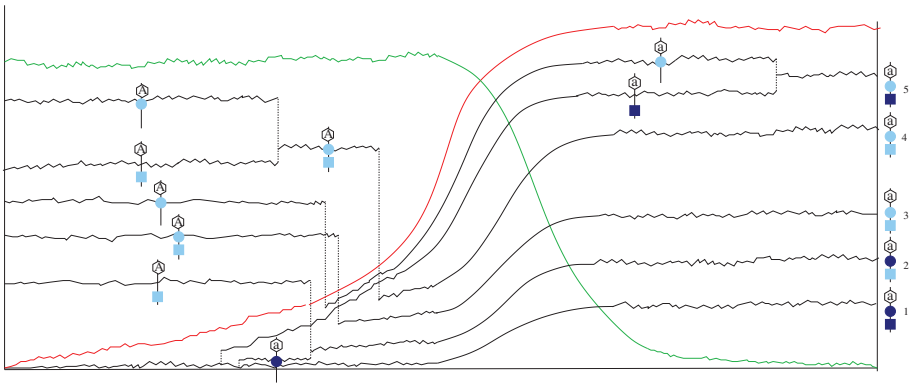

Figure 4.1: Example of genealogy for a 5-sample : dark blue neutral alleles originate from the mutant and light blue ones from an A-individual. We indicate the selected allele, A or a, associated with the neutral alleles during the sweep. It can change when a recombination occurs. Bold lines represent the A(green)- and a(red)-population sizes. In this example, the two neutral alleles of the first individual, the first neutral allele of the second individual and the second neutral allele of the fifth individual originate from the mutant; the two neutral alleles of the third individual originate from the same A-individual, whereas the two neutral alleles of the fourth individual originate from two distinct A-individuals.

$$|\pi|_4 = \#\{\ 1 \le i \le d \text{ such that } \{(i,1),(i,2)\} \text{ is an unmarked block}\}$$

$$|\pi|_5 = \#\{\ 1 \le i \le d \text{ such that } \{(i,1)\} \text{ and } \{(i,2)\} \text{ are two distinct unmarked blocks}\}$$

To express the limit distribution of the partition Θ_d^K we need to introduce :

$$q_1 := e^{-\frac{r_1 \log K}{s}}, \quad q_2 := e^{-\frac{r_2 \log K}{s}}, \quad \bar{q}_2 := e^{-\frac{f_a r_2 \log K}{|S_{Aa}|}} \quad \text{and} \quad q_3 := \frac{r_1}{r_1 + r_2(1 - f_A/f_a)}(q_2^{f_A/f_a} - q_1 q_2), \tag{4.1.11}$$

where S_{Aa} has been defined in (4.1.5) and s in (4.1.9). We did not make any assumption on the sign of $f_a(r_1+r_2) - f_A r_2$, but q_3 can be written in the form $\delta(e^{-\mu} - e^{-\nu})/(\nu - \mu)$ for $(\delta, \mu, \nu) \in \mathbb{R}_+^3$ so that it is well defined and non-negative. It is easy to check that $q_3 \le 1$. We now define five non-negative numbers $(p_k, 1 \le k \le 5)$ which will quantify the law of Θ_d^K for large K in Theorem 1 :

$$p_1 := q_1 q_2[1 - (1 - q_1)(1 - \bar{q}_2)], \quad p_2 := q_1[(1 - q_1 q_2) - q_2 \bar{q}_2(1 - q_1)], \tag{4.1.12}$$
$$p_3 := q_1 q_2(1 - \bar{q}_2)(1 - q_1), \quad p_4 := \bar{q}_2 q_3 \quad \text{and} \quad p_5 := (1 - q_1)(1 - q_1 q_2(1 - \bar{q}_2)) - \bar{q}_2 q_3.$$

Note that $\sum_{1 \le k \le 5} p_k = 1$. Finally, we introduce an assumption which summarizes all the assumptions made in this work :

Assumption 1. $(N_A^K(0), N_a^K(0)) = (\lfloor \bar{n}_A K \rfloor, 1)$ *and Conditions* (4.1.1) *on the recombination probability and* (4.1.7) *on the equilibrium densities and fitness hold.*

With Definitions 1, 2 and 3 in mind, we can now state our main results :

Theorem 1 (Genealogy of a sample, geometry $SL-N1-N2$). *Let d be in $\mathbb{N}$. Then under Assumption 1, we have for every $\pi \in \mathscr{P}_d^*$*

$$\lim_{K\to\infty} \left| \mathbb{P}(\Theta_d^K = \pi | \mathrm{Fix}^K) - \mathbf{1}_{\{\pi\in\Delta_d\}} p_1^{|\pi|_1} p_2^{|\pi|_2} p_3^{|\pi|_3} p_4^{|\pi|_4} p_5^{|\pi|_5} \right| = 0.$$

Notice that when K is large, Θ_d^K belongs to Δ_d with a probability close to one, and that

$$(p_1^{|\pi|_1} p_2^{|\pi|_2} p_3^{|\pi|_3} p_4^{|\pi|_4} p_5^{|\pi|_5}, \pi \in \Delta_d)$$

is a probability on Δ_d (depending on K). Moreover, this result implies that the d sampled individuals have asymptotically independent neutral genealogies. With high probablity, the neutral alleles of a given sampled individual i either originate from the first mutant a and belong to the marked block, or escape the sweep and originate from an A individual. In this case they belong to an unmarked block which is of the form $\{(i,1)\}$, $\{(i,2)\}$ or $\{(i,1),(i,2)\}$, according to Definition 3. As a consequence, if some neutral alleles of two distinct sampled individuals escape the sweep, they originate from distinct A-individuals with high probability. However, the genealogies of the two neutral alleles of a given individual are not independent. For example the probability that $(i,1)$ and $(i,2)$ escape the sweep is p_4+p_5 ; the probability that $(i,1)$ (resp. $(i,2)$) escapes the sweep is $p_3+p_4+p_5$ (resp. $p_2+p_4+p_5$), and for every $K\in\mathbb{N}$ such that $r_1 \neq 0$

$$(p_3+p_4+p_5)(p_2+p_4+p_5) = (1-q_1)(1-q_1q_2) < (1-q_1)(1-q_1q_2+q_1q_2\bar{q}_2) = p_4+p_5.$$

This is due to the fact that if (backwards in time) a recombination first occurs between SL and $N1$, the neutral allele at $N2$, linked to $N1$, also escapes the sweep. As the term $q_1q_2\bar{q}_2$ does not tend to 0 when K goes to infinity under Condition (4.1.1), the only possibility to have an equality in the limit is the case where $r_1 \log K \ll 1$ or in other words when the probability to see a recombination between SL and $N1$ is negligible.

Let us now consider the separated geometry, $N1-SL-N2$:

Theorem 2 (Genealogy of a sample, geometry $N1-SL-N2$). *Let d be in $\mathbb{N}$. Then under Assumption 1, we have for every $\pi \in \mathscr{P}_d^*$*

$$\lim_{K\to\infty} \left| \mathbb{P}(\Theta_d^K = \pi | \mathrm{Fix}^K) - \mathbf{1}_{\{\pi\in\Delta_d\}} [q_1q_2]^{|\pi|_1} [q_1(1-q_2)]^{|\pi|_2} [(1-q_1)q_2]^{|\pi|_3} [(1-q_1)(1-q_2)]^{|\pi|_5} \right| = 0.$$

Again the neutral genealogies of the d sampled individuals are asymptotically independent. Furthermore, we have independence between the neutral loci. Indeed the result stated in Theorem 2 means that a neutral allele at locus Nk escapes the sweep with probability $1-q_k$ independently of all other neutral alleles, including the allele at the other neutral locus of the same individual. This is due to the fact that in the separated geometry a recombination between SL and one neutral locus has no impact on the genetic background of the allele at the other neutral locus. Note in particular that there is no block of the form $\{(i,1),(i,2)\}$ in the limit partition, as the two neutral alleles have a very small probability to recombine at the same time.

4.2 Application and comparison with previous work

From this point onward we will write N_α (resp. $N_{\alpha\beta\gamma}$) instead of N_α^K (resp. $N_{\alpha\beta\gamma}^K$) for sake of readability.

Linkage disequilibrium

The linkage disequilibrium (LD) is the non-uniform association of alleles at several loci. For sake of simplicity we suppose for this application that $\mathscr{B} = \{\beta, \bar{\beta}\}$ and $\mathscr{C} = \{\gamma, \bar{\gamma}\}$. The LD between the two neutral loci in the α-population at time t can be expressed by :

$$D(\alpha, t) := \left| \frac{N_{\alpha\beta\gamma}(t)}{N_\alpha(t)} - \frac{N_{\alpha\beta\gamma}(t) + N_{\alpha\beta\bar{\gamma}}(t)}{N_\alpha(t)} \frac{N_{\alpha\beta\gamma}(t) + N_{\alpha\bar{\beta}\gamma}(t)}{N_\alpha(t)} \right|.$$

In other words it corresponds to the absolute value of the difference between the frequency of the combination of alleles β and γ and the product of frequencies of allele β and of allele γ. We can check that the value of $D(\alpha, t)$ does not depend on the choice of $(\beta, \gamma) \in \mathscr{B} \times \mathscr{C}$. Theoretical and empirical studies have shown that LD is an important signature of selection [Asm86, PMH01, KS02, Prz02, SRH+02, WFF+02]. Initially, biologists only focused on the LD between the selected locus and a neighboring locus. The expected effect of selective sweeps was then to increase the level of LD as the genetic variation is reduced. However, as first observed by Gillespie [Gil97], then followed by [KN04, SSL06, McV07], "linked selection can reduce variation without building up high levels of linkage disequilibrium, contrary to our intuition". Suppose that the LD between the neutral loci in the A-population was zero at the beginning of the sweep ($D(A,0) = 0$, which is a classical assumption, see [TK87] for instance), and recall (4.1.11). Let $D^{(ga)}$ (and $D^{(gs)}$) denote the LD between loci $N1$ and $N2$ for the adjacent ($SL - N1 - N2$) (and separated ($N1 - SL - N2$)) geometry. Then the patterns of LD are the following :

Proposition 1. *Suppose that $D(A,0)^{(ga)} = D(A,0)^{(gs)} = 0$ and that the first mutant is of type $(a, \beta, \gamma) \in \mathscr{E}$. Denote by u_β and u_γ the initial proportions of alleles β and γ in the A-population. Then under Assumption 1*

$$\lim_{K\to\infty} \left| \mathbb{E}[D^{(ga)}(a, T_{ext}^K) | \mathrm{Fix}^K] - (1-u_\beta)(1-u_\gamma)(1-q_1)q_1q_2\bar{q}_2 \right| = \lim_{K\to\infty} \mathbb{E}[D^{(gs)}(a, T_{ext}^K) | \mathrm{Fix}^K] = 0.$$

Hence the LD between two neutral loci is very dependent on the relative position of the selected locus (between the two neutral loci or adjacent to these latter, more or less close to one of them,...). By using this property it is possible to construct statistical tests (see [McV07] for example) to distinguish loci which have undergone a recent selective event, and further the selection strength during this event. We could also express the expectation of the LD when its initial value is not zero, but the result is much more complex in this case.

Previous work

In [SD05] the authors gave an approximate sampling formula for the genealogy of one neutral locus during a selective sweep. In their work, the population evolved as a two-locus modified

Moran model with recombination, selection, and in particular constant population size. They introduced the fitness s^{SD} of the mutant a as follows : when one of the iid exponential clocks of the living individuals rings, one picks two individuals uniformly at random (with replacement), one dies, and the other one gives birth. But a replacement of an a-individual by an A-individual is rejected with probability s^{SD}. In this case, nothing happens. In Chapter 3, the author studied the two-locus version of the here presented model. It was shown that the ancestral relationships in a sample taken at the end of the sweep correspond to the ones derived in [SD05] when we equal the fitness of [SD05] and the rescaled invasion fitness $s^{SD} = S_{aA}/f_a$ and when we have the equality $|S_{Aa}|/f_A = S_{aA}/f_a$ (in this case the first and third phases have the same duration, $S_{aA}\log K/f_a$).

In [BSS15b], the authors generalized the model introduced in [SD05] towards two neutral loci and used similar methods to derive a corresponding statement for the genealogy of a sample taken at the end of the sweep. If we however make the analogous comparison and try to match our result for the adjacent geometry with the statement from [BSS15b], we observe an interesting phenomenon : the probabilities of the different types of ancestry only coincide if the birth rates of a- and A-individuals are the same, that is, if $f_a = f_A$ holds true. In biology, the fitness describes the ability to both survive and reproduce, and can be defined by the average contribution of an individual with a given genotype to the gene pool of the next generation. Hence a mutation which affects the fitness of an individual in a given environment can either act on the fertility (f_α in our model), either on the death rate, intrinsic (D_α) or by competition ($C_{\alpha,\alpha'}$), or on both. Our result is comparable to this of [SD05] if the mutation only affects the death rate (and still if $s^{SD} = S_{aA}/f_a = |S_{Aa}|/f_A$).

In [PS07], instead of a birth and death process, the authors modeled the population with a structured coalescent. It is shown that this process can be approximated by a marked Yule tree where the different marks are realized by Poisson processes and indicate a recombination of one or two loci into the wild-type background. The impact of the third phase is taken into account by a certain refinement prior to the beginning of the coalescent which leads to the same effect of splitting of the two neutral loci as it is seen here. We again find similarities with our results when $f_A = f_a$. Moreover, the techniques used in [PS07] yield that coalescent events with A-individuals cannot be ignored, that is, there are neutral loci of different individuals from the sample which have the same type-A-ancestor. The structure of the sample is therefore different from our results here.

4.3 Dynamics of the sweep and couplings

Description of the three phases

We only need to focus on the trajectories of the population process where the mutant allele a goes to fixation and replaces the resident allele A. Champagnat has described these trajectories in [Cha06] and in particular divided the sweep into three phases with distinct A- and a-population dynamics (see Figure 4.2). In the sequel, ε will be a positive real number

independent of K, as small as needed for the different approximations to hold.

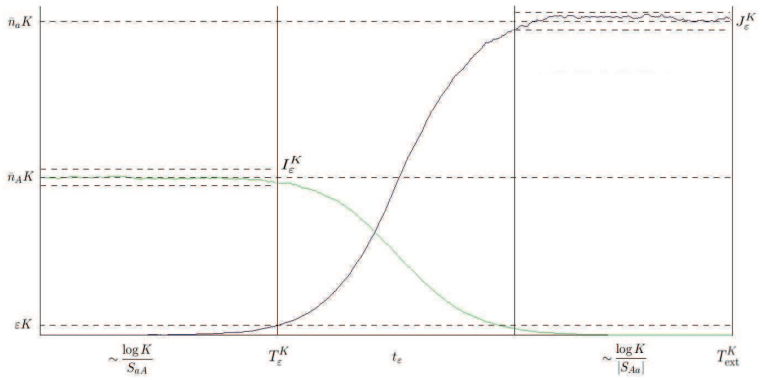

Figure 4.2: The three phases of a selective sweep ; in this simulation, $K = 10\,000$, $\bar{n}_a = 2\bar{n}_A = 2$ and $S_{aA} = |S_{Aa}| = 4$. We have also indicated some of the notations introduced in Section 4.3

First phase

The resident population size stays close to its equilibrium value $\bar{n}_A K$ as long as the mutant population size has not hit $\lfloor \varepsilon K \rfloor$: if we introduce the finite subset of $\mathbb{N}$

$$I_\varepsilon^K := \left[K\left(\bar{n}_A - 2\varepsilon \frac{C_{A,a}}{C_{A,A}}\right), K\left(\bar{n}_A + 2\varepsilon \frac{C_{A,a}}{C_{A,A}}\right)\right] \cap \mathbb{N}, \tag{4.3.1}$$

and the stopping times T_ε^K and S_ε^K, which denote respectively the hitting time of $\lfloor \varepsilon K \rfloor$ by the mutant population size and the exit time of I_ε^K by the resident population size,

$$T_\varepsilon^K := \inf\{t \geq 0, N_a(t) = \lfloor \varepsilon K \rfloor\} \quad \text{and} \quad S_\varepsilon^K := \inf\{t \geq 0, N_A(t) \notin I_\varepsilon^K\}, \tag{4.3.2}$$

then we can deduce from [Cha06] (see Equations (A.5) and (A.6) in Chapter 3 for the details of the derivation) that the events Fix^K, $\{T_\varepsilon^K \leq S_\varepsilon^K\}$ and $\{T_\varepsilon^K < \infty\}$ are very close :

$$\limsup_{K\to\infty} \mathbb{P}(\{T_\varepsilon^K \leq S_\varepsilon^K\} \bigtriangleup \text{Fix}^K) \leq c\varepsilon, \quad \text{and} \quad \limsup_{K\to\infty} \mathbb{P}(\{T_\varepsilon^K < \infty\} \bigtriangleup \text{Fix}^K) \leq c\varepsilon, \tag{4.3.3}$$

for a finite c and ε small enough, where we recall convention (4.1.8). In this context, $\bigtriangleup$ is the symmetric difference : for two sets B and C, $B \bigtriangleup C = (B \cap C^c) \cup (C \cap B^c)$. From this point onwards, "first phase" will denote the time interval $[0, T_\varepsilon^K]$ when the a-population size is smaller than $\lfloor \varepsilon K \rfloor$.

Second phase

When N_A and N_a are of order K, the rescaled population process $(N_A/K, N_a/K)$ is well approximated by the Lotka-Volterra system (4.1.6). Moreover, under Condition (4.1.7) the system (4.1.6) has a unique attracting equilibrium $(0, \bar{n}_a)$ for initial condition z satisfying

$z_a > 0$, where $\bar{n}_a$ has been defined in (4.1.7). In particular, if we introduce for $(n_A, n_a) \in \mathbb{N}^2$ the notation,

$$\mathbb{P}_{(n_A,n_a)}(.) := \mathbb{P}(.|N_A(0) = n_A, N_a(0) = n_a), \tag{4.3.4}$$

then Theorem 3 (b) in [Cha06] implies :

$$\lim_{K\to\infty} \sup_{z\in\Gamma} \mathbb{P}_{(\lfloor z_A K\rfloor, \lfloor z_a K\rfloor)}\Big(\sup_{0\le t\le t_\varepsilon, \alpha\in\mathscr{A}} \Big| \frac{N_a(t)}{K} - n_\alpha^{(z)}(t)\Big| \ge \delta\Big) = 0, \tag{4.3.5}$$

for every $\delta > 0$, where

$$\Gamma := \Big\{ z \in \mathbb{R}_+^{\mathscr{A}}, \lfloor z_A K\rfloor \in I_\varepsilon^K, z_a \in [\varepsilon/2, \varepsilon]\Big\}, \tag{4.3.6}$$

$$t_\varepsilon(z) := \inf\big\{ s \ge 0, \forall\, t \ge s, n_A^{(z)}(t) \in [0, \varepsilon^2/2], n_a^{(z)}(t) \in [\bar{n}_a - \varepsilon/2, \bar{n}_a + \varepsilon/2]\big\}, \tag{4.3.7}$$

and

$$t_\varepsilon := \sup\{t_\varepsilon(z), z \in \Gamma\} < \infty. \tag{4.3.8}$$

In the sequel, "second phase" will denote the time interval $[T_\varepsilon^K, T_\varepsilon^K + t_\varepsilon]$ when the population process is close to the system (4.1.6).

Third phase

Equation (4.3.5) also implies that

$$\lim_{K\to\infty} \mathbb{P}\Big(\frac{N_A(T_\varepsilon^K + t_\varepsilon)}{K} \in [\omega_1, \omega_2], \Big| \frac{N_a(T_\varepsilon^K + t_\varepsilon)}{K} - \bar{n}_a\Big| \le \varepsilon, \ \Big| \Big(\frac{N_A(T_\varepsilon^K)}{K}, \frac{N_a(T_\varepsilon^K)}{K}\Big) \in \Gamma\Big) = 1,$$

where

$$2\omega_1 := \inf\{n_A^{(z)}(t_\varepsilon), z \in \Gamma\} > 0, \quad \text{and} \quad \omega_2/2 := \sup\{n_A^{(z)}(t_\varepsilon), z \in \Gamma\} \le \varepsilon^2/2. \tag{4.3.9}$$

The "third phase", which corresponds to the time interval $[T_\varepsilon^K + t_\varepsilon, T_{\text{ext}}^K]$, can be seen as the symmetric counterpart of the first phase, where the roles of A and a are interchanged : during the extinction of the A-population, the a-population size stays close to its equilibrium value $\bar{n}_a K$.

Let us introduce the positive real number $M'' := 3 + (f_a + C_{a,A})/C_{a,a}$, the finite subset of $\mathbb{N}$

$$J_\varepsilon^K := \Big[K\Big(\bar{n}_a - M''\varepsilon\Big), K\Big(\bar{n}_a + M''\varepsilon\Big)\Big] \cap \mathbb{N}, \tag{4.3.10}$$

and the stopping times $T_u^{(K,A)}$ and $S_\varepsilon^{(K,a)}$, which denote respectively the hitting times of $\lfloor uK\rfloor$ by the A-population for $u \in \mathbb{R}_+$, and the exit time of J_ε^K by the a-population during the third phase,

$$T_u^{(K,A)} := \inf\{t \ge 0, N_A(T_\varepsilon^K + t_\varepsilon + t) = \lfloor uK\rfloor\}, \quad S_\varepsilon^{(K,a)} := \inf\{t \ge 0, N_a(T_\varepsilon^K + t_\varepsilon + t) \notin J_\varepsilon^K\}. \tag{4.3.11}$$

If we define the event

$$\mathscr{N}_\varepsilon^K := \{T_\varepsilon^K \le S_\varepsilon^K\} \cap \Big\{ \frac{N_A(T_\varepsilon^K + t_\varepsilon)}{K} \in [\omega_1, \omega_2], \Big| \frac{N_a(T_\varepsilon^K + t_\varepsilon)}{K} - \bar{n}_a\Big| \le \varepsilon\Big\}, \tag{4.3.12}$$

we get from the proof of Lemma 3 in [Cha06] that for a finite c and ε small enough,

$$\limsup_{K\to\infty} \Big\{ \mathbb{P}(\text{Fix}^K \bigtriangleup [\mathscr{N}_\varepsilon^K \cap \{T_0^{(K,A)} < T_\varepsilon^{(K,A)} \wedge S_\varepsilon^{(K,a)}\}]) + \mathbb{P}(\text{Fix}^K \bigtriangleup [\mathscr{N}_\varepsilon^K \cap \{T_0^{(K,A)} < T_\varepsilon^{(K,A)}\}])\Big\} \le c\varepsilon. \tag{4.3.13}$$

Couplings for the first and third phases

We are interested in the law of the neutral genealogies on the event Fix^K. Equations (4.3.3) and (4.3.13) imply that it is enough to concentrate our attention on the event $\mathcal{N}_\varepsilon^K \cap \{T_0^{(K,A)} < T_\varepsilon^{(K,A)}\}$, but the dynamics of the population process N conditionally on this event is complex to study. Hence we couple N with two jump processes $\tilde{N}$ and $\tilde{\tilde{N}}$, easier to study, and satisfying

$$\limsup_{K\to\infty} \mathbb{P}(\{\exists t \le T_\varepsilon^K, N(t) \neq \tilde{N}(t)\}, T_\varepsilon^K < \infty) \le c\varepsilon. \tag{4.3.14}$$

$$\limsup_{K\to\infty} \mathbb{P}(\{\exists\, 0 \le t \le T_0^{(K,A)}, N(T_\varepsilon^K + t_\varepsilon + t) \neq \tilde{\tilde{N}}(T_\varepsilon^K + t_\varepsilon + t)\}, T_0^{(K,A)} < T_\varepsilon^{(K,A)} | \mathcal{N}_\varepsilon^K < \infty) \le c\varepsilon. \tag{4.3.15}$$

To describe the couplings we need to introduce a version of Moran process with recombination.

Definition 4. *Recall notation* (4.1.3). *Let α be in $\mathscr{A}$ and $n^{(\alpha)}$ be in $N^{\alpha\times\mathscr{B}\times\mathscr{C}}$. We call Moran process of type α with recombination and population size $\sum_{(\beta,\gamma)\in\mathscr{B}\times\mathscr{C}} n^{(\alpha)}_{\alpha\beta\gamma}$ a process $MR^{(n^{(\alpha)})}$ with values in $N^{\alpha\times\mathscr{B}\times\mathscr{C}}$, initial state $n^{(\alpha)}$, and which evolves as follows :*

- *After an exponential time with parameter $f_A \bar{n}_A K$ we pick uniformly and with replacement three individuals and draw a Bernoulli variable R with parameter r_2*
- *The first individual dies, the second one gives birth to an individual carrying its alleles at loci SL and $N1$*
- *If $R = 0$, there is no recombination and the allele at locus $N2$ of the newborn is also inherited from the second individual ; if $R = 1$ there is a recombination and the newborn inherits its second neutral allele from the third individual*
- *We again draw an exponential variable with parameter $f_A \bar{n}_A K$ and restart the procedure*

Let us first describe the coupling with $\tilde{N}$: N and $\tilde{N}$ are equal up to time S_ε^K ; after this time the A individuals in the population process $\tilde{N}$ follow a Moran process with recombination independent of the a-individuals. The a-individuals evolve as if all the A-individuals had the same genotype. More precisely we choose an individual in the A-population and decide that this individual exerts a competition $C_{A,a}\tilde{N}_A/K$ on each a-individual and the probability that an a-individual recombine with this A-individual is $f_A\tilde{N}_A/(f_A\tilde{N}_A + f_a\tilde{N}_a)$ at each birth of an a-individual. To choose the individual A which interacts with the a-population we introduce a total order $\prec$ on the pairs $(\beta,\gamma) \in \mathscr{B}\times\mathscr{C}$ and define for every $t \ge S_\varepsilon^K$,

$$\underline{\beta\gamma}(t) := \{(\beta,\gamma) \in \mathscr{B}\times\mathscr{C}, \tilde{N}_{A\beta\gamma}(t) \neq 0, \forall (\beta',\gamma') \neq (\beta,\gamma) \in \mathscr{B}\times\mathscr{C}, \tilde{N}_{A\beta'\gamma'}(t) \neq 0 \Rightarrow (\beta,\gamma) \prec (\beta',\gamma')\},$$

the minimum label of the A-individuals in the process $\tilde{N}$ at time t. The process $\tilde{N}$ is defined by :

$$\begin{aligned}\tilde{N}(t) \quad = \quad & \mathbf{1}_{t<S_\varepsilon^K} N(t) + \mathbf{1}_{t\geq S_\varepsilon^K}\Big(MR^{(N^{(A)}(S_\varepsilon^K))}(t-S_\varepsilon^K) + \sum_{(\beta,\gamma)\in\mathscr{B}\times\mathscr{C}} e_{a\beta\gamma} \\ & \Big[\int_{S_\varepsilon^K}^t\int_{R_+} Q^{(1)}(ds,d\theta)\mathbf{1}_{\{0<\theta-\sum_{(\beta',\gamma')<(\beta,\gamma)} b^K_{a\beta'\gamma'}(\tilde{N}_A(s^-)e_{A\underline{\beta\gamma}}(s^-),\tilde{N}^{(a)}(s^-))\leq b^K_{a\beta\gamma}(\tilde{N}_A(s^-)e_{A\underline{\beta\gamma}}(s^-),\tilde{N}^{(a)}(s^-))\}} \\ & -\int_{S_\varepsilon^K}^t\int_{R_+} Q^{(2)}(ds,d\theta)\mathbf{1}_{\{0<\theta-\sum_{(\beta',\gamma')<(\beta,\gamma)} d^K_{a\beta'\gamma'}(\tilde{N}_A(s^-)e_{A\underline{\beta\gamma}}(s^-),\tilde{N}^{(a)}(s^-))\leq d^K_{a\beta\gamma}(\tilde{N}_A(s^-)e_{A\underline{\beta\gamma}}(s^-),\tilde{N}^{(a)}(s^-))\}}\Big]\Big),\end{aligned} \tag{4.3.16}$$

where $MR^{(N^{(A)})}$ has been defined in Definition 4, $Q^{(1)}$ and $Q^{(2)}$ are two independent Poisson Point processes with density $dsd\theta$, also independent of $MR^{(N^{(A)})}$, and $(e_{\alpha\beta\gamma},(\alpha,\beta,\gamma)\in\mathscr{E})$ is the canonical basis of $\mathbb{R}^{\mathscr{E}}$. Let us notice that $\tilde{N}$ is a Markov process.

The coupling with the process $\tilde{\tilde{N}}$ is simpler : we assume that $\mathscr{N}_\varepsilon^K$ holds ; N and $\tilde{\tilde{N}}$ are equal from time $T_\varepsilon^K + t_\varepsilon$ and during a time $S_\varepsilon^{(K,a)} \wedge T_\varepsilon^{(K,A)}$. Then the a-individuals in the population process $\tilde{\tilde{N}}$ follow a Moran process with recombination independent of the A-individuals, and each $A\beta\gamma$-population evolves as a birth and death process with individual birth and death rates f_A and $f_A+|S_{Aa}|$, independent of the a-individuals and the $A\beta'\gamma'$-population with $(\beta,\gamma)\neq(\beta',\gamma')$:

$$\begin{aligned}&\tilde{\tilde{N}}(T_\varepsilon^K+t_\varepsilon+t) = \mathbf{1}_{t<S_\varepsilon^{(K,a)}} N(T_\varepsilon^K+t_\varepsilon+t) + \mathbf{1}_{t\geq S_\varepsilon^{(K,a)}}\Big(MR^{(N^{(a)}(S_\varepsilon^{(K,a)}))}(t-S_\varepsilon^{(K,a)}) + \\ &+\sum_{(\beta,\gamma)\in\mathscr{B}\times\mathscr{C}} e_{A\beta\gamma}\Big[\int_{T_\varepsilon^K+t_\varepsilon+S_\varepsilon^{(K,a)}}^{T_\varepsilon^K+t_\varepsilon+t}\int_{R_+} Q_{\beta\gamma}(ds,d\theta)\Big\{\mathbf{1}_{\{0<\theta\leq f_A\tilde{\tilde{N}}_{A\beta\gamma}(s^-)\}} - \mathbf{1}_{\{0<\theta-f_A\tilde{\tilde{N}}_{A\beta\gamma}(s^-)\leq(f_A+|S_{Aa}|)\tilde{\tilde{N}}_{A\beta\gamma}(s^-)\}}\Big\}\Big]\Big),\end{aligned} \tag{4.3.17}$$

where $MR^{(N^{(a)})}$ has been defined in Definition 4 and is independent of the sequence of independent Poisson measures $(Q_{\beta\gamma},(\beta,\gamma)\in\mathscr{B}\times\mathscr{C})$, with intensity $dsd\theta$. The process $\tilde{\tilde{N}}$ is also Markovian.

Inequality (4.3.14) follows from (4.3.3). Moreover, from the proof of Lemma 3 in [Cha06] we know that

$$\liminf_{K\to\infty}\mathbb{P}(T_0^{(K,A)} < T_\varepsilon^{(K,A)}\wedge S_\varepsilon^{(K,a)}|\mathscr{N}_\varepsilon^K) \geq 1-c\varepsilon$$

for a finite c and ε small enough. Adding (4.3.13) we get that (4.3.15) is also satisfied. Hence we will study the processes $\tilde{N}$ and $\tilde{\tilde{N}}$ and deduce properties of the dynamics of the process N during the first and third phases. Let $\tilde{\mathscr{F}}$ (resp. $\tilde{\tilde{\mathscr{F}}}$) be the filtration generated by the process $\tilde{N}$ (resp. $\tilde{\tilde{N}}$) and $\tilde{C}$ (resp. $\tilde{\tilde{C}}$) a $\tilde{\mathscr{F}}$-measurable (resp. $\tilde{\tilde{\mathscr{F}}}$-measurable) event. Then we define

$$\mathbb{P}^{(1)}(\tilde{C}) := \mathbb{P}(\tilde{C}|\tilde{T}_\varepsilon^K<\infty), \quad \text{and} \quad \mathbb{P}^{(3)}(\tilde{\tilde{C}}) := \mathbb{P}(\tilde{\tilde{C}}|\mathscr{N}_\varepsilon^K, \tilde{\tilde{T}}_0^{(K,A)} < \tilde{\tilde{T}}_\varepsilon^{(K,A)}), \tag{4.3.18}$$

where $\tilde{T}_\varepsilon^K$ is the analogues of T_ε^K for the process $\tilde{N}$, and $\tilde{\tilde{T}}_0^{(K,A)}$ and $\tilde{\tilde{T}}_\varepsilon^{(K,A)}$ the analogues of $T_0^{(K,A)}$ and $T_\varepsilon^{(K,A)}$ for the process $\tilde{\tilde{N}}$. Expectations and variances associated with these probability measures are denoted by $\mathbb{E}^{(1)}$, $\mathrm{Var}^{(1)}$, $\mathbb{E}^{(3)}$ and $\mathrm{Var}^{(3)}$ respectively. Equations (4.3.3) and (4.3.13) to (4.3.15) imply that we can concentrate our attention to $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(3)}$ to prove convergences on the fixation event Fix^K.

4.4 Proofs of the main results

Events impacting the genealogies in each phase

Let us now summarize the results on the genealogies for the three successive phases of the sweep that we will derive in Sections 4.6 and 4.7.

First phase : As explained in the previous section, we work with the process $\tilde{N}$ to study the first phase. Let us introduce the jump times of $\tilde{N}$:

$$\tau_0^K = 0 \quad \text{and} \quad \tau_m^K = \inf\{t > \tau_{m-1}^K, \tilde{N}(t) \neq \tilde{N}(\tau_{m-1}^K)\}, \quad m \geq 1. \tag{4.4.1}$$

The number of jumps during the first phase is denoted by $J^K(1)$:

$$J^K(1) := \inf\{m \in \mathbb{N}, \tilde{N}_a(\tau_m^K) = \lfloor \varepsilon K \rfloor\}. \tag{4.4.2}$$

Coalescence and recombination events are defined as follows (see Figure 4.3) :

Definition 5. *We sample two distinct individuals at time τ_m^K and denote by (α, β, γ) and $(\alpha', \beta', \gamma')$ their type.*

We say that β and β' coalesce at time τ_m^K if they are carried by two distinct individuals at time τ_m^K and by the same individual at time τ_{m-1}^K. Seen forwards in time it corresponds to a birth and hence a copy of the neutral allele. Seen backwards in time it corresponds to the fusion of two neutral alleles into one, carried by one parent of the newborn. We define in the same way coalescent events at locus $N2$ (resp. loci $N1$ and $N2$) for alleles γ and γ' (resp. allele pairings (β, γ) and (β', γ')).

We say that β (and/or γ) recombines at time τ_m^K from the α- to the α'-population if the individual carrying the allele β (and/or γ) at time τ_m^K is a newborn, carries the allele α inherited from it first parent, and has inherited its allele β (and/or γ) from a different individual carrying allele α'.

We are interested in recombinations which provoke new associations of alleles. In particular, in the adjacent geometry $SL - N1 - N2$ we will not consider the simultaneous recombinations of a pair (β, γ) within the α-population.

Let us now describe the genealogical scenarios which modify the ancestral relationships between the neutral alleles of one individual and occur with positive probability when K is large. Let us first focus on the first phase and pick uniformly an individual i from the

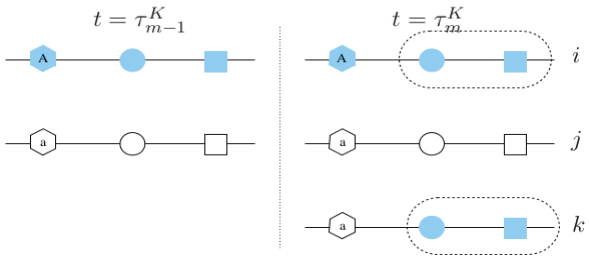

Figure 4.3: Illustration of Definition 5 : the newborn (individual k) has inherited the selected allele from its "white" parent and the two neutral alleles from its "blue" parent; hence the encircled neutral loci (of individuals i and k) coalesce at time τ^K_m. In terms of recombinations, the two neutral loci of the newborn individual recombine at time τ^K_m from the a- to the A-population

a-population at time $\tilde{T}^K_\varepsilon$. We introduce for the adjacent geometry $SL-N1-N2$:

$NR(i)^{(1)}$: there is no recombination into the A-population affecting $(i,1)$ or $(i,2)$ and both neutral loci of the i-individual originate from the first mutant,

$R2(i)^{(1)}$: only the neutral allele $(i,2)$ is affected by a recombination with the A-population, hence $(i,1)$ originates from the first mutant and $(i,2)$ from an A-individual,

$R12(i)^{(1)}$: one recombination between SL and $N1$ from the a- into the A-population occurs and both neutral alleles $(i,1)$ and $(i,2)$ originate from the same A-individual,

$[2,1]^{rec}_{A,i}$: first (backwards in time) $(i,2)$ recombines into the A-population, then $(i,1)$ recombines into the A-population and connects to a different individual than $(i,2)$.

$[12,2]^{rec}_{A,i}$: first (backwards in time) the tuple $\{(i,1),(i,2)\}$ recombines into the A-population, then a second recombination splits the two neutral loci inside the A-population.

$R1|2(i)^{(1,ga)}$: $[2,1]^{rec,1}_{A,i} \cup [12,2]^{rec,1}_{A,i}$ (see Figure 4.4)

Hence, recalling the definition of (q_1,q_2,q_3) in (4.1.11) we will prove in Section 4.6 :

Proposition 2 (Neutral genealogies during the first phase, geometry $SL-N1-N2$). *Let i be an a-individual sampled uniformly at the end of the first phase (time $\tilde{T}^K_\varepsilon$). Under Assumption 1, there exist two finite constants c and ε_0 such that for every $\varepsilon \le \varepsilon_0$,*

$$\limsup_{K\to\infty}\Big\{\Big|\mathbb{P}^{(1)}(NR(i)^{(1)}) - q_1q_2\Big| + \Big|\mathbb{P}^{(1)}(R2(i)^{(1)}) - q_1(1-q_2)\Big| + \Big|\mathbb{P}^{(1)}(R12(i)^{(1)}) - q_3\Big| + \Big|\mathbb{P}^{(1)}(R1|2(i)^{(1,ga)}) - (1-q_1-q_3)\Big|\Big\} \le c\varepsilon.$$

For large K, the sum of the four probabilities of Proposition 2 equals one up to a constant times ε. Hence, in the limit we only observe the events described page 139. The probabilities of the first two events are quite intuitive : broadly speaking, the probability to have no

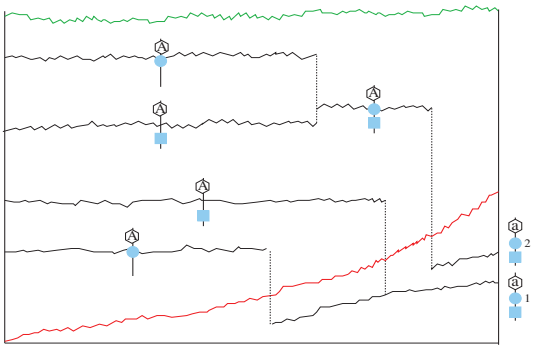

Figure 4.4: Illustration of events $[2,1]^{rec}_{A,i}$ (individual 1) and $[12,2]^{rec}_{A,i}$ (individual 2)

recombination at a birth event is $1-r_1-r_2$, the birth rate is f_a and the duration of the first phase is $\log K/S_{aA}$. Hence under $\mathbb{P}^{(1)}$, the probability of the event $NR(i)^{(1)}$ is approximately

$$(1-(r_1+r_2))^{f_a \log K/S_{aA}} \sim \exp(-(r_1+r_2))^{f_a \log K/S_{aA}} = q_1 q_2.$$

Similarly the probability to have no recombination between SL and $N1$ is close to q_1 and subtracting the probability of $NR(i)^{(1)}$ we get this of $R2(i)^{(1)}$. The probabilities of $R12(i)^{(1)}$ and $R1|2(i)^{(1,ga)}$ are more complex. The proofs rely on a fine study of the different possible scenarios.

Let us now introduce the possible genealogical trajectories for the separated geometry $N1-SL-N2$ during the first phase :

$NR(i)^{(1)}, R2(i)^{(1)}$: defined as for the adjacent geometry $SL-N1-N2$

$R1(i)^{(1)}$: only $(i,1)$ is affected by a recombination with the A-population ; $(i,2)$ originates from the first mutant and $(i,1)$ from an A-individual

$R1|2(i)^{(1,gs)}$: $(i,1)$ and $(i,2)$ are affected by a recombination with the A-population ; they originate from two distinct A-individuals

We will prove in Section 4.6 the following asymptotics for the separated geometry $N1-SL-N2$:

Proposition 3 (Neutral genealogies during the first phase, geometry $N1-SL-N2$). *Let i be an a-individual sampled uniformly at the end of the first phase (time $\tilde{T}^K_\varepsilon$). Under Assumption 1, there exist two finite constants c and ε_0 such that for every $\varepsilon \le \varepsilon_0$,*

$$\limsup_{K\to\infty} \Big\{ \Big|\mathbb{P}^{(1)}(NR(i)^{(1)}) - q_1 q_2\Big| + \Big|\mathbb{P}^{(1)}(R2(i)^{(1)}) - q_1(1-q_2)\Big| + \Big|\mathbb{P}^{(1)}(R1(i)^{(1)}) - (1-q_1)q_2\Big| + \Big|\mathbb{P}^{(1)}(R1|2(i)^{(1,gs)}) - (1-q_1)(1-q_2)\Big| \Big\} \le c\varepsilon.$$

For this geometry, the intuitive interpretation gives the good results. The independence of the two neutral loci follows from the fact that a recombination which affects one neutral locus

does not change the genetic background of the second neutral locus.

Second phase : We work with the process N to study the second phase. The latter one has a duration of order 1, and the recombination probabilities are negligible with respect to one (Condition (4.1.1)). Consequently, no event impacting the genealogies of the neutral loci occurs during the second phase. More precisely, let us sample uniformly two distinct a-individuals i and j at the end of the second phase (time $T_\varepsilon^K + t_\varepsilon$) and introduce the events :

$$\begin{aligned} NR(i)^{(2)} &: \text{ there is no recombination affecting } (i,1) \text{ or } (i,2), \\ NC(i,j)^{(2)} &: \text{ there is no coalescence between the neutral genealogies of } i \text{ and } j. \end{aligned}$$

Then we have the following result, which will be proven in Section 4.7.

Proposition 4 (Neutral genealogies during the second phase for the two geometries). *Let i and j be two distinct a-individuals sampled uniformly at the end of the second phase (time $T_\varepsilon^K + t_\varepsilon$). Then under Assumption 1,*

$$\lim_{K\to\infty} \mathbb{P}(NR(i)^{(2)} \cap NC(i,j)^{(2)} | T_\varepsilon^K \leq S_\varepsilon^K) = 1.$$

Third phase : Finally, we focus on the process $\tilde{\tilde{N}}$. When K is large, there is only one event occurring with positive probability during the third phase which may modify the ancestry of the neutral alleles of an individual i sampled at the end of the sweep in the adjacent geometry :

$$R2(i)^{(3,ga)} : \text{ a recombination between loci } N1 \text{ and } N2 \text{ occurs and separates } (i,1) \text{ and } (i,2) \text{ within the } a\text{-population}, \tag{4.4.3}$$

Indeed, if we also define the events

$$\begin{aligned} NR(i)^{(3)} &: \text{ there is no recombination affecting } (i,1) \text{ or } (i,2) \text{ and they both originate from the same } a\text{-individual at the end of the second phase} \\ NC(i,j)^{(3)} &: \text{ defined as } NC(i,j)^{(2)} \text{ for two distinct individuals sampled uniformly at the end of the sweep.} \end{aligned}$$

Then we will prove in Section 4.7 :

Proposition 5 (Neutral genealogies during the third phase, geometry $SL - N1 - N2$). *Let i and j be two distinct a-individuals sampled uniformly at the end of the sweep. Under Assumption 1, there exist two finite constants c and ε_0 such that for every $\varepsilon \leq \varepsilon_0$,*

$$\limsup_{K\to\infty} \left\{ \left| \mathbb{P}^{(3)}(R2(i)^{(3,ga)}) - (1-\tilde{q}_2) \right| + \left| \mathbb{P}^{(3)}(NR(i)^{(3)}) - \tilde{q}_2 \right| + \mathbb{P}^{(3)}(NC(i,j)^{(3)}) \right\} \leq c\varepsilon.$$

In particular, there is no recombination with the A-population during the third phase. As for Proposition 3 this result is quite intuitive, as the duration of the third phase is close to

$\log K/|S_{Aa}|$.

Independence : Finally we again consider the population process N and state a proposition which enables us to give the statement of Theorem 1 independently for all sampled individuals, that is, jointly for the whole sample. To this aim, let us introduce a partition $\Theta_d^{(K,1)} \in \mathscr{P}_d^*$ which is the analogue of Θ_d^K where the d individuals are sampled at the end of the first phase and not at the end of the sweep. Recall Definitions 2 and 3, and denote by $|R2^{(3,ga)}|_d$ (resp. $|NR^{(3)}|_d$) the number of a-individuals in a d-sample taken at the end of the sweep whose neutral alleles originate from two distinct a-individuals (resp. from the same a-individual) at the beginning of the third phase. Then we have the following result :

Proposition 6. *Let $d \in \mathbb{N}$ and Assumption 1 hold. Then there exist two finite constants c and ε_0 such that for every $\varepsilon \le \varepsilon_0$, the ancestral relationships of a d-sample taken at the end of the first phase (time T_ε^K) satisfy in the adjacent geometry $SL-N1-N2$, for every $(m_k, 1 \le k \le 4) \in \mathbb{Z}_+^4$:*

$$\limsup_{K\to\infty} \Big| \mathbb{P}(|\Theta_d^{(K,1)}|_k = m_k, 1 \le k \le 4 | T_\varepsilon^K \le S_\varepsilon^K) - \mathbf{1}_{\{m_1+m_2+m_3+m_4=d\}} \frac{d!}{m_1!m_2!m_3!m_4!} (q_1 q_2)^{m_1} (q_1(1-q_2))^{m_2} q_3^{m_3} (1-q_1-q_3)^{m_4} \Big| \le c\varepsilon.$$

In the same way, the neutral genealogy of a d-sample taken at the end of the sweep satisfies for every $(m_k, 1 \le k \le 2) \in \mathbb{Z}_+^2$:

$$\limsup_{K\to\infty} \Big| \mathbb{P}((|R2^{(3,ga)}|_d, |NR^{(3)}|_d) = (m_1, m_2) | \mathscr{N}_\varepsilon^K) - \mathbf{1}_{\{m_1+m_2=d\}} \frac{d!}{m_1!m_2!} (1-\bar{q}_2)^{m_1} \bar{q}_2^{m_2} \Big| \le c\varepsilon.$$

Proposition 6 is a key result : we only need to focus on individual neutral genealogies to get general results on the genealogy of a d-sample with respect to the neutral loci. It will be proven in Section 4.8.

Proof of Theorem 1

Let i be an individual sampled uniformly at the end of the sweep. The idea of the proof is the following : in a first step, we list certain compositions of coalescent and recombination events leading to specific ancestral relationships which could be described by blocks of a partition of Δ_d. Then we approximate the probabilities of the described events and finally prove that these probabilities sum to one up to a constant times $\sqrt{\varepsilon}$ for some fixed small ε. This shows that in the limit for large K the neutral genealogy of the individual i belongs to these described page 139 with a probability close to one. In a second step we use Proposition 6 to treat the neutral genealogies of the d sampled individuals independently.

i) We consider two possible trajectories such that the alleles at both neutral loci originate from the mutant : either the two neutral loci separate inside the a-population during the third phase and coalesce during the first phase, or they stay in the a-population and

do not separate during the whole sweep (see individual 1 in Figure 4.1); more rigorously, this corresponds to the event :

$$\Big(R2(i)^{(3,ga)} \cap NR(i1)^{(2)} \cap NR(i2)^{(2)} \cap NC(i1,i2)^{(2)} \cap [NR(i1)^{(1)} \sqcup R2(i1)^{(1)}] \cap NR(i2)^{(1)}\Big) \bigsqcup \Big(NR(i)^{(3)} \cap NR(i)^{(2)} \cap NR(i)^{(1)}\Big),$$

where we denote by $i1$ and $i2$ the labels of the parents of the first and second neutral loci of i, respectively, at the end of the second phase (the way we label the a-individuals has no importance as they are exchangeable).

ii) We consider two possible trajectories such that $(i,1)$ originates from the mutant and $(i,2)$ originates from some A-individual

$$\Big(R2(i)^{(3,ga)} \cap NR(i1)^{(2)} \cap NR(i2)^{(2)} \cap NC(i1,i2)^{(2)} \cap [NR(i1)^{(1)} \cup R2(i1)^{(1)}] \cap [R12(i2)^{(1)} \sqcup R1|2(i2)^{(1)} \sqcup R2(i2)^{(1)}]\Big) \bigsqcup \Big(NR(i)^{(3)} \cap NR(i)^{(2)} \cap R2(i)^{(1)}\Big).$$

Note that the first bracket considers a separation of the two neutral loci during the third phase. As a consequence, the fate of the first neutral locus of individual $i2$ during the first phase has no consequence on the neutral genealogy of i. This is why we consider the event $\{R12(i2)^{(1)} \sqcup R1|2(i2)^{(1)} \sqcup R2(i2)^{(1)}\}$ and not only $\{R2(i2)^{(1)}\}$. The second bracket corresponds to individual 2 in Figure 4.1.

iii) We consider one possible trajectory such that $(i,1)$ originates from some A-individual and $(i,2)$ originates from the mutant (see individual 5 in Figure 4.1)

$$R2(i)^{(3,ga)} \cap NR(i1)^{(2)} \cap NR(i2)^{(2)} \cap NC(i1,i2)^{(2)} \cap [R12(i1)^{(1)} \sqcup R1|2(i1)^{(1,ga)}] \cap NR(i2)^{(1)}$$

iv) We consider one possible trajectory such that $(i,1)$ and $(i,2)$ originate from the same A-individual (see individual 3 in figure 4.1)

$$NR(i)^{(3)} \cap NR(i)^{(2)} \cap R12(i)^{(1)}$$

v) Finally, we consider two possible trajectories such that $(i,1)$ and $(i,2)$ originate from different A-individuals

$$\Big(R2(i)^{(3,ga)} \cap NR(i1)^{(2)} \cap NR(i2)^{(2)} \cap NC(i1,i2)^{(2)} \cap [R12(i1)^{(1)} \sqcup R1|2(i1)^{(1)}] \cap [R12(i2)^{(1)} \sqcup R1|2(i2)^{(1)} \cup R2(i2)^{(1)}]\Big) \bigsqcup \Big(NR(i)^{(3)} \cap NR(i)^{(2)} \cap R1|2(i)^{(1,ga)}\Big)$$

The second bracket corresponds to individual 4 in figure 4.1.

Thanks to (4.3.3), and (4.3.13) to (4.3.15) we know that for every measurable events $C^{(1)}$, $C^{(2)}$ and $C^{(3)}$ occurring during the first, second and third phase respectively,

$$\begin{aligned}\mathbb{P}(C^{(1)} \cap C^{(2)} \cap C^{(3)} | \mathrm{Fix}^K) &= \mathbb{P}(C^{(3)} | C^{(1)} \cap C^{(2)}, \mathcal{N}_\varepsilon^K, T_0^{(K,A)} \leq T_\varepsilon^{(K,A)}) \\ &\qquad \mathbb{P}(C^{(2)} | \{T_\varepsilon^K \leq S_\varepsilon^K\}, C^{(1)}) \mathbb{P}(C^{(1)} | T_\varepsilon^K < \infty) + O_K(\varepsilon) \\ &= \mathbb{P}^{(3)}(\tilde{\tilde{C}}^{(3)}) \mathbb{P}(C^{(2)} | \{T_\varepsilon^K \leq S_\varepsilon^K\}) \mathbb{P}^{(1)}(\tilde{C}^{(1)}) + O_K(\varepsilon), \qquad (4.4.4)\end{aligned}$$

where $\tilde{C}^{(1)}$ (resp. $\tilde{\tilde{C}}^{(3)}$) corresponds to the event $C^{(1)}$ (resp. $C^{(3)}$) expressed in terms of the process $\tilde{N}$ (resp. $\tilde{\tilde{N}}$), and $O_K(\varepsilon)$ is a function of K and ε satisfying

$$\limsup_{K\to\infty} |O_K(\varepsilon)| \leq c\varepsilon, \tag{4.4.5}$$

for $\varepsilon \leq \varepsilon_0$ where ε_0 and c are finite. Then by applying Propositions 2, 4, 5 and 6 we get for these five successive events the value of the probabilities $(p_k, 1 \leq k \leq 5)$ defined in (4.1.12), which sum to one. Let us detail the calculations for the case $i)$: by applying (4.4.4) and Proposition 4, the probability to see one of the two trajectories described in $i)$ is

$$\begin{aligned}\mathscr{P}(i,1) = \mathbb{P}^{(3)}(R2(i)^{(3,ga)})\mathbb{P}^{(1)}([NR(i1)^{(1)} \sqcup R2(i1)^{(1)}] \cap NR(i2)^{(1)}) \\ + \mathbb{P}^{(3)}(NR(i)^{(3)})\mathbb{P}^{(1)}(NR(i)^{(1)}) + O_K(\varepsilon).\end{aligned} \tag{4.4.6}$$

But thanks to Proposition 6 we know that the neutral genealogies of individuals i and j are nearly independent. Hence adding Proposition 2 leads to

$$\mathbb{P}^{(1)}([NR(i1)^{(1)} \sqcup R2(i1)^{(1)}] \cap NR(i2)^{(1)}) = (q_1q_2 + q_1(1-q_2))q_1q_2 + O_K(\varepsilon).$$

Applying Propositions 2 and 5 in (4.4.6) yields

$$\mathscr{P}(i,1) = (1-\bar{q}_2)q_1^2q_2 + \bar{q}_2q_1q_2 + O_K(\sqrt{\varepsilon}) = p_1 + O_K(\sqrt{\varepsilon}),$$

where we recall the definition of p_1 in (4.1.12).

Finally, we get the asymptotic independence of the neutral genealogies of the d sampled individuals during the first and third phases by applying the multinomial version of de Finetti Representation Theorem (see [DP14] Chapter 4 for a simple proof) to the result of Proposition 6. The asymptotic independence during the second phase follows from Proposition 4 as, with high probability, nothing happens.

Proof of Theorem 2

It follows the same ideas that the proof of Theorem 1 : Proposition 3 states that in the separated geometry, $N1-SL-N2$, the two neutral loci recombine independently with the A-population during the first phase. Propositions 4 and 5 state that coalescences and recombinations between A- and a- populations during the second and the third phases are negligible. Hence the ancestral relations are not modified by these two phases and the overall ancestral relations are those stated in Proposition 3. In the separated geometry, the independence between genealogies is even easier to derive as the genetic background of a neutral allele does not depend on the recombinations undergone by the individual's other neutral allele.

4.5 Number of births and deaths during the selective sweep

In this section we derive some results on birth and death numbers of the population processes $\tilde{N}$ and $\tilde{\tilde{N}}$, needed in Sections 4.6 and 4.7 to prove Propositions 2, 4 and 5. This technical part may be skipped on first reading.

Coupling with supercritical birth and death processes during the first phase

We are interested in the dynamics of the process $\tilde{N}_a$ during the first phase, that is, before the time $\tilde{T}_\varepsilon^K$. The idea is to couple this process with two supercritical birth and death processes, and thus deduce its dynamics from well known results on birth and death processes. Recall the definition of the rescaled invasion fitness s in (4.1.9), and for $\varepsilon < S_{aA}/(2C_{a,A}C_{A,a}/C_{A,A} + C_{a,a})$ define the two approximations,

$$s - \frac{2C_{a,A}C_{A,a} + C_{a,a}C_{A,A}}{f_a C_{A,A}}\varepsilon =: s_-(\varepsilon) \le s \le s_+(\varepsilon) := s + 2\frac{C_{a,A}C_{A,a}}{f_a C_{A,A}}\varepsilon. \tag{4.5.1}$$

Then for $t < \tilde{T}_\varepsilon^K \wedge S_\varepsilon^K$ the death rate of a-individuals in the process $\tilde{N}$ equals this of the process N, defined in (4.1.5) and satisfies

$$1 - s_+(\varepsilon) \le \frac{d_a(\tilde{N}(t))}{f_a \tilde{N}_a(t)} = 1 - s + \frac{C_{a,A}}{f_a K}(\tilde{N}_A(t) - \bar{n}_A K) + \frac{C_{a,a}}{f_a K}\tilde{N}_a(t) \le 1 - s_-(\varepsilon). \tag{4.5.2}$$

For $S_\varepsilon^K \le t < \tilde{T}_\varepsilon^K$, the death rate of a-individuals also satisfies

$$1 - s_+(\varepsilon) \le \frac{d_a^K(\tilde{N}_A e_{A\beta\gamma(t)}, \tilde{N}^{(a)}(t))}{f_a \tilde{N}_a(t)} \le 1 - s_-(\varepsilon). \tag{4.5.3}$$

Hence following Theorem 2 in [Cha06], we can construct the processes Z_ε^-, $(\tilde{N}_A, \tilde{N}_a)$ and Z_ε^+ on the same probability space such that almost surely :

$$Z_\varepsilon^-(t) \le \tilde{N}_a(t) \le Z_\varepsilon^+(t), \quad \text{for all } t < \tilde{T}_\varepsilon^K, \tag{4.5.4}$$

where for $* \in \{-, +\}$, Z_ε^* is a birth and death process with initial state 1, and individual birth and death rates f_a and $f_a(1 - s_*(\varepsilon))$.

We end this section with the definition of two stopping times used on several occasions in Section 4.5. Let σ_u^K denote the time of the first hitting of $\lfloor u \rfloor$ by the process $\tilde{N}_a$:

$$\sigma_u^K := \inf\{t \ge 0, \tilde{N}_a(t) = \lfloor u \rfloor\}, \quad u \in \mathbb{R}_+. \tag{4.5.5}$$

If for $0 < s < 1$, $\tilde{Z}^{(s)}$ is a random walk with jumps ± 1 where up-jumps occur with probability $1/(2-s)$ and down-jumps with probability $(1-s)/(2-s)$, we denote by $\mathscr{P}_i^{(s)}$ the law of $\tilde{Z}^{(s)}$ when the initial state is $i \in \mathbb{N}$ and introduce for every $\rho \in \mathbb{R}_+$ the stopping time

$$\tau_\rho := \inf\{n \in \mathbb{Z}_+, \tilde{Z}_n^{(s)} = \lfloor \rho \rfloor\}. \tag{4.5.6}$$

Number of jumps of $\tilde{N}_a$ during the first phase

Expectation of the number of upcrossings

Let us recall Equation (4.4.1) and consider $k < \lfloor \varepsilon K \rfloor$. Then the number of upcrossings from k to $k+1$ during the first phase is :

$$U_k^K(1) := \#\{m, \tau_m^K < \tilde{T}_\varepsilon^K, (\tilde{N}_a(\tau_m^K), \tilde{N}_a(\tau_{m+1}^K)) = (k, k+1)\}. \tag{4.5.7}$$

If we recall Equations (4.3.1) and (4.5.1), and introduce a real number λ_ε

$$\lambda_\varepsilon := (1-s_-(\varepsilon))^3(1-s_+(\varepsilon))^{-2}, \tag{4.5.8}$$

which belongs to $(0,1)$ for ε small enough, then we have the following result :

Lemma 1. *There exist three positive finite constants c, K_0 and ε_0 such that for $K \geq K_0$ and $\varepsilon \leq \varepsilon_0$:*

If $j \leq k < \lfloor \varepsilon K \rfloor$ and $n_A \in I_\varepsilon^K 1$,

$$\left| \mathbb{E}^{(1)}_{(n_A,j)}[U_k^K(1)] - \frac{1-(1-s)^{\lfloor \varepsilon K \rfloor - k} - (1-s)^{k+1}}{s} \right| \leq c\varepsilon. \tag{4.5.9}$$

If $k < j < \lfloor \varepsilon K \rfloor$ and $n_A \in I_\varepsilon^K 1$,

$$\mathbb{E}^{(1)}_{(n_A,j)}[U_k^K(1)] \leq \frac{(1-s_-(\varepsilon))^{j-k}}{s_+(\varepsilon)s_-^2(\varepsilon)}. \tag{4.5.10}$$

If $k' \leq k < \lfloor \varepsilon K \rfloor$ and $n_A \in I_\varepsilon^K 1$,

$$\left| \mathrm{Cov}^{(1)}_{(n_A,j)}(U_k^K(1), U_{k'}^K(1)) \right| \leq c\left(\lambda_\varepsilon^{(k-k')/2} + \varepsilon\right). \tag{4.5.11}$$

Démonstration. The idea, which will be used several times throughout Section 4.5, is to compare the number of upcrossings with geometric random variables. Suppose first that $j \leq k$. Then on the event $\{\tilde{T}_\varepsilon^K < \infty\}$ the process $\tilde{N}_a$ necessarily jumps from k to $k+1$. Being in $k+1$, it either reaches $\lfloor \varepsilon K \rfloor$ before k, or it goes back and again from k to $k+1$ and so on. We first approximate the probability that there is only one jump from k to $k+1$. As we do not know the value of $\tilde{N}_A$ when $\tilde{N}_a$ hits k for the first time, we bound the probability by the extreme values it can take. Recall Definitions (4.5.5) and (4.5.6). The upper bound is derived as follows :

$$\begin{aligned} \mathbb{P}^{(1)}_{(n_A,j)}(U_k^K(1)=1) &\leq \sup_{n_A \in I_\varepsilon^K 1} \mathbb{P}^{(1)}_{(n_A,k+1)}(\tilde{T}_\varepsilon^K < \sigma_k^K) \\ &= \sup_{n_A \in I_\varepsilon^K 1} \frac{\mathbb{P}_{(n_A,k+1)}(\tilde{T}_\varepsilon^K < \sigma_k^K)}{\mathbb{P}_{(n_A,k+1)}(\tilde{T}_\varepsilon^K < \infty)} \leq q_k^{(s_+(\varepsilon),s_-(\varepsilon))}, \end{aligned} \tag{4.5.12}$$

where we use (4.3.18) and for $(s_1,s_2) \in (0,1)^2$

$$q_k^{(s_1,s_2)} := \frac{\mathscr{P}_{k+1}^{(s_1)}(\tau_{\varepsilon K} < \tau_k)}{\mathscr{P}_{k+1}^{(s_2)}(\tau_{\varepsilon K} < \tau_0)}. \tag{4.5.13}$$

Similarly, we show that $\mathbb{P}^{(1)}_{(n_A,j)}(U_k^K(1)=1) \geq q_k^{(s_-(\varepsilon),s_+(\varepsilon))}$. In the same way, we can approximate the probability that there are least three jumps from k to $k+1$ knowing that there are at least two jumps, and so on. We deduce that we can construct two geometric random variables

G_1 and G_2, possibly on an enlarged space, with respective parameters $q_k^{(s_+(\varepsilon),s_-(\varepsilon))} \wedge 1$ and $q_k^{(s_-(\varepsilon),s_+(\varepsilon))}$ such that

$$G_1 \le U_k^K(1) \le G_2, \quad \text{a.s.} \tag{4.5.14}$$

In particular, taking the expectation we get from (B.1)

$$\frac{(1-(1-s_+(\varepsilon))^{\lfloor \varepsilon K\rfloor -k})(1-(1-s_-(\varepsilon))^{k+1})}{s_+(\varepsilon)(1-(1-s_-(\varepsilon))^{\lfloor \varepsilon K\rfloor})} \le \mathbb{E}^{(1)}_{(n_A,j)}[U_k^K(1)] \le \frac{(1-(1-s_-(\varepsilon))^{\lfloor \varepsilon K\rfloor -k})(1-(1-s_+(\varepsilon))^{k+1})}{s_-(\varepsilon)(1-(1-s_+(\varepsilon))^{\lfloor \varepsilon K\rfloor})}. \tag{4.5.15}$$

According to (4.1.9) and (4.5.1), $0<s<1$ and $|s_+(\varepsilon)-s_-(\varepsilon)| \le (4C_{a,A}C_{A,a}+C_{a,a}C_{A,A})\varepsilon/(f_aC_{A,A})$. Hence the last inequality and straightforward calculations lead to (4.5.9).

Let us now assume that $k<j$. Then we have

$$\begin{aligned}
\mathbb{P}^{(1)}_{(n_A,j)}(U_k^K(1)\ge 1) &\le \sup_{n_A\in I_\varepsilon^K 1} \mathbb{P}^{(1)}_{(n_A,j)}(\sigma_k^K < \tilde{T}_\varepsilon^K) \\
&= \sup_{n_A\in I_\varepsilon^K 1} \frac{\mathbb{P}_{(n_A,j)}(\tilde{T}_\varepsilon^K<\infty|\sigma_k^K<\tilde{T}_\varepsilon^K)\mathbb{P}_{(n_A,j)}(\sigma_k^K<\tilde{T}_\varepsilon^K)}{\mathbb{P}_{(n_A,j)}(\tilde{T}_\varepsilon^K<\infty)} \\
&\le \frac{\mathscr{P}_k^{(s_+(\varepsilon))}(\tau_{\varepsilon K}<\tau_0)\mathscr{P}_j^{(s_-(\varepsilon))}(\tau_k<\tau_{\varepsilon K})}{\mathscr{P}_j^{(s_-(\varepsilon))}(\tau_{\varepsilon K}<\tau_0)} \le \frac{(1-s_-(\varepsilon))^{j-k}}{s_+(\varepsilon)s_-(\varepsilon)},
\end{aligned}$$

where we again used (4.3.18) and (B.1). Moreover, the same proof as for (4.5.14) leads to :

$$\mathbb{E}^{(1)}_{(n_A,j)}[U_k^K(1)|U_k^K(1)\ge 1] \le \left(q_k^{(s_-(\varepsilon),s_+(\varepsilon))}\right)^{-1} \le s_-^{-1}(\varepsilon),$$

where we used Equation (B.3). This ends the proof of (4.5.10).

The last inequality, (4.5.11), has been stated in Chapter 3 (Equation (7.26)). □

Remark 10. *Let c, K and ε_0 be defined as in Lemma 1. If we define the total number of downcrossings from k to $k-1$,*

$$D_k^K(1) := \#\{m, \tau_m^K \le \tilde{T}_\varepsilon^K, (\tilde{N}_a(\tau_m^K), \tilde{N}_a(\tau_{m+1}^K)) = (k, k-1)\}, \tag{4.5.16}$$

then by definition of the probability $\mathbb{P}^{(1)}$, for every $2\le k<\lfloor \varepsilon K\rfloor$

$$D_k^K(1) = U_{k-1}^K(1) - 1, \quad \mathbb{P}^{(1)} - a.s., \tag{4.5.17}$$

and for $j<k<\lfloor \varepsilon K\rfloor$ and $n_A\in I_\varepsilon^K 1$,

$$\left|\mathbb{E}^{(1)}_{(n_A,j)}[D_k^K(1)] - \frac{1-s}{s}(1-(1-s)^{\lfloor \varepsilon K\rfloor -k} - (1-s)^{k-1})\right| \le c\varepsilon. \tag{4.5.18}$$

Expectation of hitting numbers

Let us recall (4.5.7), (4.6.12) and (4.5.16), and introduce for $0 < j \leq k < \lfloor \varepsilon K \rfloor$ the number of hittings of the state k by the process $\tilde{N}_a$ before the time $\tilde{T}_\varepsilon^K$:

$$V_k^K(1) := U_{k-1}^K(1) + D_{k+1}^K(1) = \#\{m, \tau_m^K \leq \tilde{T}_\varepsilon^K, \tilde{N}_a(\tau_m^K) = k, \tilde{N}_a(\tau_{m+1}^K) \neq k\}. \tag{4.5.19}$$

Recall the definition of $\lambda_\varepsilon \in (0,1)$ in (4.5.8). We can state the following Lemma, which will be useful to get bounds on the number of upcrossings of the A-population during the first phase (see Lemma 4) :

Lemma 2. *There exist three finite constants c, K_0 and ε_0 such that for $K \geq K_0$, $\varepsilon \leq \varepsilon_0$ and $k' < k < \lfloor \varepsilon K \rfloor$:*

$$\left| \mathbb{E}^{(1)}[V_k^K(1)] - \frac{(2-s)(1-(1-s)^{\lfloor \varepsilon K \rfloor - k} - (1-s)^k)}{s} \right| \leq c\varepsilon, \quad \text{and } |\mathrm{Cov}^{(1)}(V_{k'}^K(1), V_k^K(1))| \leq c(\varepsilon + \lambda_\varepsilon^{(k-k')/2}).$$

Démonstration. Under $\mathbb{P}^{(1)}$ the a-population size goes from 1 to $\lfloor \varepsilon K \rfloor$, thus the number of downcrossings from $k+1$ to k is equal to the number of upcrossings from k to $k+1$ minus 1. Adding (4.5.19) yields

$$V_k^K(1) = U_{k-1}^K(1) + U_k^K(1) - 1, \quad \mathbb{P}^{(1)} - a.s.$$

We get the first part of the Lemma by taking the expectation and applying (4.5.18). The proof of the second part follows that of (4.5.11), and once again we can find the details in the proof of Equation (7.26) in Chapter 3. □

Number of upcrossings during an excursion above or below a given level

We now focus on the number of upcrossings from k to $k+1$ during an excursion above or below l. Let us denote by $\sigma_l^K(1)$ the jump number of the first hitting of l before the end of the first phase : for $l < \lfloor \varepsilon K \rfloor$,

$$\sigma_l^K(1) := \inf\{m, \tau_m^K \leq \tilde{T}_\varepsilon^K, \tilde{N}_a(\tau_m^K) = l\}, \tag{4.5.20}$$

and for $1 \leq k, l < \lfloor \varepsilon K \rfloor$ and $n_A \in I_\varepsilon^K 1$,

$$U_{n_A,l,k}^K(1) := \#\left\{ m < \sigma_l^K(1), (\tilde{N}_a(\tau_m^K), \tilde{N}_a(\tau_{m+1}^K)) = (k, k+1) \right\}. \tag{4.5.21}$$

Then, if we denote by μ_ε the real number

$$\mu_\varepsilon := (1 - s_-(\varepsilon))^2 (1 - s_+(\varepsilon))^{-1}, \tag{4.5.22}$$

which belongs to $(0,1)$ for ε small enough, we can derive the following bounds :

Lemma 3. *There exist three positive finite constants c, K_0 and ε_0 such that for $K \geq K_0$, $\varepsilon \leq \varepsilon_0$, $1 \leq k < l < \lfloor \varepsilon K \rfloor$ and $n_A \in I_\varepsilon^K 1$,*

$$\mathbb{E}_{(n_A,k+1)}^{(1)}[U_{n_A,k,l}^K(1) | \sigma_k^K(1) < \infty] \vee \mathbb{E}_{(n_A,l-1)}^{(1)}[U_{n_A,l,k}^K(1)] \leq c\mu_\varepsilon^{l-k}.$$

Démonstration. Equations (B.5) and (B.6) in Chapter 3 state that for $k < l < \lfloor \varepsilon K \rfloor$ and $n_A \in I_\varepsilon^K 1$,

$$\mathbb{P}^{(1)}_{(n_A,k+1)}(U^K_{n_A,k,l}(1) \geq 1 | \sigma^K_k(1) < \infty) \leq c(1 - s_-(\varepsilon))^{l-k},$$

and

$$\mathbb{P}^{(1)}_{(n_A,k+1)}(U^K_{n_A,k,l}(1) = 1 | U^K_{n_A,k,l}(1) \geq 1, \sigma^K_k(1) < \infty) \geq c\Big(\frac{1 - s_+(\varepsilon)}{1 - s_-(\varepsilon)}\Big)^{l-k}$$

for a finite c. By comparing $U^K_{n_A,k,l}(1)$ with a geometric random variable we get the first inequality. To bound the expectation of upcrossings from k to $k+1$ during an excursion under l we first bound the probability to have at least one jump from k to $k+1$ during an excursion below l. By definition, $\tilde{N}_a$ necessary hits $l-1$ during such an excursion. Recall Definitions (4.3.18), (4.5.5) and (4.5.6). Then for every n_A in $I_\varepsilon^K 1$,

$$\begin{aligned}
\mathbb{P}^{(1)}_{(n_A,l-1)}(\sigma^K_k < \sigma^K_l | \sigma^K_l < \infty) &= \frac{\mathbb{P}_{(n_A,l-1)}(\tilde{T}^K_\varepsilon < \infty | \sigma^K_k < \sigma^K_l)\mathbb{P}_{(n_A,l-1)}(\sigma^K_k < \sigma^K_l)}{\mathbb{P}_{(n_A,l-1)}(\tilde{T}^K_\varepsilon < \infty)} \\
&\leq \frac{\mathscr{P}^{(s_+(\varepsilon))}_k(\tau_{\varepsilon K} < \tau_0)\mathscr{P}^{(s_-(\varepsilon))}_{l-1}(\tau_k < \tau_l)}{\mathscr{P}^{(s_-(\varepsilon))}_{l-1}(\tau_{\varepsilon K} < \tau_0)} \leq \frac{(1 - s_-(\varepsilon))^{l-k-1}}{s_-(\varepsilon)},
\end{aligned}$$

where we used (B.1). The next step consists in bounding the number of upcrossings from k to $k+1$ during the excursion knowing that this number is greater than one : for $n_A \in I_\varepsilon^K 1$,

$$\begin{aligned}
\mathbb{P}^{(1)}_{(n_A,k+1)}(\sigma_l < \sigma_k) &= \frac{\mathbb{P}_{(n_A,k+1)}(\tilde{T}^K_\varepsilon < \infty | \sigma_l < \sigma_k)\mathbb{P}_{(n_A,k+1)}(\sigma_l < \sigma_k)}{\mathbb{P}_{(n_A,k+1)}(\tilde{T}^K_\varepsilon < \infty)} \\
&\geq \frac{\mathscr{P}^{(s_-(\varepsilon))}_l(\tau_{\varepsilon K} < \tau_0)\mathscr{P}^{(s_-(\varepsilon))}_{k+1}(\tau_l < \tau_k)}{\mathscr{P}^{(s_+(\varepsilon))}_{k+1}(\tau_{\varepsilon K} < \tau_0)} \geq s^2_-(\varepsilon),
\end{aligned}$$

where we again used (B.1). Hence on the event $\{U^K_{n_A,l,k}(1) \geq 1\}$, $U^K_{n_A,l,k}(1)$ is smaller than a geometric random variable with parameter $s^2_-(\varepsilon)$ and we get :

$$\mathbb{E}^{(1)}_{(n_A,l-1)}[U^K_{n_A,l,k}(1)] \leq s^{-2}_-(\varepsilon)\mathbb{P}^{(1)}_{(n_A,l-1)}(U^K_{n_A,l,k}(1) \geq 1) \leq \frac{(1 - s_-(\varepsilon))^{l-k-1}}{s^3_-(\varepsilon)},$$

which ends the proof of Lemma 3. □

Number of jumps $\tilde{N}_A$ during the first phase

We introduce for $k < \lfloor \varepsilon K \rfloor$ the number of upcrossings of the A-population when the a-population is of size k :

$$\mathscr{U}^K_k(1) := \#\{m, \tau^K_m \leq \tilde{T}^K_\varepsilon, \tilde{N}_A(\tau^K_{m+1}) - \tilde{N}_A(\tau^K_m) = 1, \tilde{N}_a(\tau^K_m) = k\}. \tag{4.5.23}$$

We are now able to get bounds for the expectations and covariances of these quantities :

Lemma 4. *There exist three finite constants c, K_0 and ε_0 such that for $K \geq K_0$, $\varepsilon \leq \varepsilon_0$ and $k < \lfloor \varepsilon K \rfloor$,*

$$\left| \mathbb{E}^{(1)} \Big[\sum_{i=1}^{k} \mathscr{U}_i^K(1) \Big] - \frac{f_A \bar{n}_A K \log k}{s f_a} \right| \leq cK(1+\varepsilon \log k) \quad \textit{and} \quad \mathrm{Var}^{(1)} \Big(\sum_{i=1}^{k} \mathscr{U}_i^K(1) \Big) \leq cK^2(1+\varepsilon \log^2 k).$$

Démonstration. The proof is based on the comparison of the A- and a-population jump rates. Let us first focus on the a-population. For $k \leq \lfloor \varepsilon K \rfloor$ and $n_A \in I_\varepsilon^K 1$,

$$\begin{aligned}
\mathbb{P}^{(1)}_{(n_A,k)}(\tilde{N}_a(dt) \neq k) &= \sum_{*\in\{+,-\}} \frac{\mathbb{P}_{(n_A,k)}(\tilde{T}_\varepsilon^K < \infty | \tilde{N}_a(dt) = k*1)}{\mathbb{P}_{(n_A,k)}(\tilde{T}_\varepsilon^K < \infty)} \mathbb{P}_{(n_A,k)}(\tilde{N}_a(dt) = k*1) \\
&\leq \sum_{*\in\{+,-\}} \frac{\mathscr{P}_{k*1}^{(s_+(\varepsilon))}(\tilde{T}_\varepsilon^K < \infty)}{\mathscr{P}_k^{(s_-(\varepsilon))}(\tilde{T}_\varepsilon^K < \infty)} \mathbb{P}_{(n_A,k)}(\tilde{N}_a(dt) = k*1) \\
&\leq \frac{(1+c\varepsilon)}{1-(1-s)^k} \Big(\big(1-(1-s)^{k+1}\big) f_a k + (1-s)\big(1-(1-s)^{k-1}\big)(D_a + C_{aA}\bar{n}_A)k \Big) dt \\
&= (1+c\varepsilon) f_a (2-s) k dt, \qquad (4.5.24)
\end{aligned}$$

for a finite constant c and ε small enough, where we used the definition of $\mathbb{P}^{(1)}$ in (4.3.18) for the equality, Coupling (4.5.4) for the first inequality, (B.1) for the second one, and equality $S_{aA} = f_a - D_a - C_{a,A}\bar{n}_A$ for the last one. Reasoning similarly we get :

$$(1-c\varepsilon) f_a (2-s) k dt \leq \mathbb{P}^{(1)}_{(n_A,k)}(\tilde{N}_a(dt) \neq k). \qquad (4.5.25)$$

Let us now focus on the number of upcrossings of the A-population. The definition of $\tilde{N}$ in (4.3.16) and Bayes' Theorem yield

$$(1-c\varepsilon) f_A \bar{n}_A K dt \leq \mathbb{P}^{(1)}_{(n_A,k)}(\tilde{N}_A(dt) = n_A + 1) \leq (1+c\varepsilon) f_A \bar{n}_A K dt, \qquad (4.5.26)$$

for a finite c and ε small enough. Indeed, from Coupling (4.5.4) and Equation (B.1) we get the following bound, independent of n_A in $I_\varepsilon^K 1$:

$$\frac{1-(1-s_-(\varepsilon))^k}{1-(1-s_-(\varepsilon))^{\lfloor \varepsilon K \rfloor}} \leq \mathbb{P}_{(n_A,k)}(\tilde{T}_\varepsilon^K < \infty) \leq \frac{1-(1-s_+(\varepsilon))^k}{1-(1-s_+(\varepsilon))^{\lfloor \varepsilon K \rfloor}}.$$

Hence there exist two finite constants c and ε_0 such that for every $\varepsilon \leq \varepsilon_0$, if we introduce the parameters

$$\frac{1}{q_k^{(1)}(\varepsilon)} := 1 + (1-c\varepsilon)\frac{f_A \bar{n}_A K}{(2-s) f_a k} \leq 1 + (1+c\varepsilon)\frac{f_A \bar{n}_A K}{(2-s) f_a k} =: \frac{1}{q_k^{(2)}(\varepsilon)}, \qquad (4.5.27)$$

we can deduce from (4.5.24) to (4.5.26) that for $k < \lfloor \varepsilon K \rfloor$

$$\sum_{V_k^K(1)} \Big(G^i_{q_k^{(1)}(\varepsilon)} - 1 \Big) \leq \mathscr{U}_k^K(1) \leq \sum_{V_k^K(1)} \Big(G^i_{q_k^{(2)}(\varepsilon)} - 1 \Big), \qquad (4.5.28)$$

where for $j \in \{1,2\}$, $(G^i_{q_k^{(j)}(\varepsilon)}, i \in \mathbb{N})$ is a sequence of geometric random variables with parameter $q_k^{(j)}(\varepsilon)$ independent of $V_l^K(1)$ (defined in (4.5.19)) for all $l < \lfloor \varepsilon K \rfloor$. Hence a direct application of Lemmas 2 and 13 leads to

$$\left| \mathbb{E}^{(1)}\left[\mathscr{U}_k^K(1) \right] - \frac{f_A \bar{n}_A K}{s f_a k} (1 - (1-s)^k - (1-s)^{\lfloor \varepsilon K \rfloor - k}) \right| \leq c\varepsilon \frac{K}{k}, \tag{4.5.29}$$

for a finite c and ε small enough. This implies the first inequality of Lemma 4.

Let us now bound the second moment of $\mathscr{U}_k^K(1)$ and the expectation of $\mathscr{U}_k^K(1)\mathscr{U}_l^K(1)$ for $k \neq l$. The first upper bound follows again from a direct application of Lemmas 2 and 13. We get

$$\mathbb{E}^{(1)}\left[(\mathscr{U}_k^K(1))^2 \right] \leq \mathbb{E}^{(1)}\left[\Big(\sum_{V_k^K(1)} G^i_{q_k^{(2)}(\varepsilon)} \Big)^2 \right] \leq \frac{2(\mathbb{E}^{(1)}[V_k^K(1)])^2}{(q_k^{(2)}(\varepsilon))^2} \leq 2(1+c\varepsilon)\Big(\frac{f_A \bar{n}_A K}{s f_a k} \Big)^2, \tag{4.5.30}$$

for a finite c and ε small enough. A new application of the same Lemmas yields, for $k < l < \lfloor \varepsilon K \rfloor$

$$\mathbb{E}^{(1)}\left[\mathscr{U}_k^K(1)\mathscr{U}_l^K(1) \right] \leq \frac{\mathbb{E}^{(1)}[V_k^K(1)V_l^K(1)]}{q_k^{(2)}(\varepsilon) q_l^{(2)}(\varepsilon)} \leq c(1+\varepsilon+\lambda_\varepsilon^{(l-k)/2}) \frac{(f_A \bar{n}_A K)^2}{(f_a s)^2 k l}, \tag{4.5.31}$$

where we used that $\mathbb{E}^{(1)}[XY] = \mathbb{E}^{(1)}[X]\mathbb{E}^{(1)}[Y] + \mathrm{Cov}^{(1)}(X,Y)$ for any real random variables (X,Y). From (4.5.28) to (4.5.31) and (B.2) we deduce that there exists a finite c such that for ε small enough and $k < \lfloor \varepsilon K \rfloor$,

$$\mathbb{E}^{(1)}\left[\Big(\sum_{i=1}^k \mathscr{U}_i^K(1) \Big)^2 \right] \leq (1+c\varepsilon)\Big(\frac{f_A \bar{n}_A K \log k}{f_a s} \Big)^2 + cK^2. \tag{4.5.32}$$

Reasoning similarly to get the lower bound, we obtain

$$\left| \mathbb{E}^{(1)}\left[\Big(\sum_{i=1}^k \mathscr{U}_i^K(1) \Big)^2 \right] - \Big(\frac{f_A \bar{n}_A K \log k}{f_a s} \Big)^2 \right| \leq cK^2(1+\varepsilon \log^2 k). \tag{4.5.33}$$

Adding the first inequality if Lemma 4 we conclude the proof. □

Coupling with subcritical birth and death processes during the third phase

We couple the process $\tilde{\tilde{N}}_a$ with two subcritical birth and death processes to control its dynamics. We recall the definition of $\mathscr{N}_\varepsilon^K$ in (4.3.12) and introduce

$$\bar{s} := |S_{Aa}|/f_A. \tag{4.5.34}$$

Let us define for ε small enough,

$$\bar{s} - \frac{M'' C_{A,a}}{f_A}\varepsilon =: \bar{s}_-(\varepsilon) < \bar{s} < \bar{s}_+(\varepsilon) := \bar{s} + \frac{C_{A,A} + M'' C_{A,a}}{f_A}\varepsilon, \tag{4.5.35}$$

where M'' has been defined just before Definition (4.3.10). Then according to the definition of $\tilde{\tilde{N}}$ in (4.3.17), we can follow Theorem 2 in [Cha06] and construct the processes Y_ε^+, $\tilde{\tilde{N}}$ and Y_ε^- on the same probability space such that on the event $\mathcal{N}_\varepsilon^K$

$$Y_\varepsilon^+(t) \le \tilde{\tilde{N}}_A(t) \le Y_\varepsilon^-(t), \quad \text{for all } T_\varepsilon^K + t_\varepsilon \le t < T_\varepsilon^K + t_\varepsilon + \tilde{\tilde{T}}_0^{(K,A)}, \quad \text{a.s.,} \tag{4.5.36}$$

where for $* \in \{-,+\}$, Y_ε^* is a birth and death process with initial state $N_A(T_\varepsilon^K + t_\varepsilon)$ and individual birth and death rates f_A and $f_A(1+\bar{s}_*(\varepsilon))$.

If for $0 < s < 1$, $\tilde{Z}^{(s)}$ denotes a random walk with jump ± 1 where up-jumps occur with probability $1/(2+s)$ and down-jumps with probability $(1+s)/(2+s)$, we denote by $\mathcal{Q}_i^{(s)}$ the law of $\tilde{Z}^{(s)}$ when the initial state is $i \in \mathbb{N}$ and introduce for every $\rho \in \mathbb{R}_+$ the stopping time

$$\nu_\rho := \inf\{n \in \mathbb{Z}_+, \tilde{Z}_n^{(s)} = \lfloor \rho \rfloor\}. \tag{4.5.37}$$

Number of jumps of $\tilde{\tilde{N}}_A$ during the third phase

Similarly as in (4.5.19) we introduce for $1 \le k < \lfloor \varepsilon K \rfloor$ the random variable $\mathcal{V}_k^K(3)$ which corresponds to the number of hittings of state k by the process $\tilde{\tilde{N}}_A$ during the third phase. Recall Definitions (4.3.9), (4.3.10) and (4.5.35). We have the following approximations :

Lemma 5. *Let u be in $[\omega_1, \omega_2]$. There exist three finite constants c, K_0 and ε_0 such that for $K \ge K_0$, $\varepsilon \le \varepsilon_0$ and n_a in $J_\varepsilon^K 1$, if $\lfloor uK \rfloor < k < \lfloor \varepsilon K \rfloor$,*

$$\mathbb{E}^{(3)}_{(\lfloor uK \rfloor, n_a)}[\mathcal{V}_k^K(3)] \le (1+c\varepsilon)\frac{2+\bar{s}}{\bar{s}}(1+\bar{s}_-(\varepsilon))^{\lfloor uK \rfloor - k},$$

and if $k \le \lfloor uK \rfloor$,

$$\left| \mathbb{E}^{(3)}_{(\lfloor uK \rfloor, n_a)}[\mathcal{V}_k^K(3)] - \frac{2+\bar{s}}{\bar{s}}(1-(1+\bar{s})^{-k} - (1+\bar{s})^{k-\lfloor \varepsilon K \rfloor}) \right| \le c\varepsilon.$$

Démonstration. The proof is very similar to that of (4.5.9), hence we do not detail all the calculations and refer to the proof of Lemma 1. First we consider $\lfloor uK \rfloor < k < \lfloor \varepsilon K \rfloor$ and approximate under $\mathbb{P}^{(3)}$ the probability for $\tilde{\tilde{N}}_A$ to hit k before the A-population extinction. Indeed, if $k \le \lfloor uK \rfloor$, we know that $\tilde{\tilde{N}}_A$ hits k $\mathbb{P}^{(3)}$-a.s. Let $\lfloor uK \rfloor < k < \lfloor \varepsilon K \rfloor$. Then for every $n_a \in J_\varepsilon^K 1$, Equation (B.1) implies

$$\mathbb{P}^{(3)}_{(\lfloor uK \rfloor, n_a)}(\tilde{\tilde{N}}_A \text{ hits } k) \le \frac{\mathcal{Q}_k^{(\bar{s}_+(\varepsilon))}(\nu_0 < \nu_{\varepsilon K})\mathcal{Q}_{\lfloor uK \rfloor}^{(\bar{s}_-(\varepsilon))}(\nu_k < \nu_0)}{\mathcal{Q}_{\lfloor uK \rfloor}^{(\bar{s}_-(\varepsilon))}(\nu_0 < \nu_{\varepsilon K})} \le \frac{1+c\varepsilon}{(1+\bar{s}_-(\varepsilon))^{k-\lfloor uK \rfloor}}, \tag{4.5.38}$$

for a finite c, ε small enough and K large enough. The second step consists in counting how many times the process $\tilde{\tilde{N}}_A$ hits k during the third phase knowing that it happens at least once. Once again we will compare this number with geometric random variables, by approximating the probability to have only one jump. The following inequality follows the spirit of (4.5.12). The only difference is that in the third phase $\tilde{\tilde{N}}_A$ is coupled with subcritical

birth and death processes, whereas in the first phase $\tilde{N}_a$ was coupled with supercritical birth and death processes. For every $n_a \in J_\varepsilon^K 1$ and $k < \lfloor \varepsilon K \rfloor$,

$$\begin{aligned}\mathbb{P}^{(3)}_{(k,n_a)}(\tilde{\tilde{N}}_A(t) \le k, \forall\, t \ge 0) &\ge \frac{\mathcal{Q}_{k-1}^{(\bar{s}_-(\varepsilon))}(\nu_0 < \nu_k)\mathcal{Q}_k^{(\bar{s}_-(\varepsilon))}(\nu_{k-1} < \nu_{k+1})}{\mathcal{Q}_k^{(\bar{s}_+(\varepsilon))}(\nu_0 < \nu_{\varepsilon K})} \\ &\ge \frac{(1-c\varepsilon)\bar{s}}{(2+\bar{s})(1-(1+\bar{s})^{-k}-(1+\bar{s})^{k-\lfloor \varepsilon K \rfloor})}.\end{aligned}$$

We derive the upper bound similarly and end the proof by comparing the hitting numbers with geometric random variables. For $\lfloor uK \rfloor < k < \lfloor \varepsilon K \rfloor$ we have to multiply the expectation of the geometric random variables by the probability to hit k at least once, approximated in (4.5.38). □

Number of births of a-individuals during the third phase

Let $U_k^K(3)$ be the number of births in the a-population during the third phase when $\tilde{\tilde{N}}_A$ equals $k \le \lfloor \varepsilon K \rfloor$

$$\begin{aligned}U_k^K(3) := \#\{m, T_\varepsilon^K + t_\varepsilon < \tau_m^K \le T_{\text{ext}}^K, \tilde{\tilde{N}}_A(\tau_m^K) = k, \text{ and } \{\{\tilde{\tilde{N}}_a(\tau_{m+1}^K) - \tilde{\tilde{N}}_a(\tau_m^K) = 1\} \\ \text{or } \{\tilde{\tilde{N}}_a(\tau_{m+1}^K) = \tilde{\tilde{N}}_a(\tau_m^K), \tilde{\tilde{N}}^{(a)}(\tau_{m+1}^K) \ne \tilde{\tilde{N}}^{(a)}(\tau_m^K)\}\}. \quad (4.5.39)\end{aligned}$$

We now state an approximation for the expectation of $U_k^K(3)$. We do not prove this result as it is obtained in the same way as Lemma 4 : the birth rate of the a-population is close to $f_a \bar{n}_a K$, the jump rate of the A-population is of order $(2+\bar{s}) f_A k$ when $\tilde{\tilde{N}}_A = k$ and the expectations of the hitting numbers for the A-population are given in Lemma 5. The only difference is that the A-population size can hit values bigger than the initial value of the third phase, $\tilde{\tilde{N}}_A(T_\varepsilon^K + t_\varepsilon)$. However the probabilities to hit such values decrease geometrically (see Lemma 5) and they will have a negligible influence on the final result. Thus we get

Lemma 6. *There exist three finite constants c, ε_0 and K_0 such that for $\varepsilon \le \varepsilon_0$ and $K \ge K_0$,*

$$\left|\mathbb{E}^{(3)}\left[\sum_{i=1}^k U_i^K(3)\right] - \frac{f_a \bar{n}_a K \log k}{\bar{s} f_A}\right| \le cK(1+\varepsilon \log k), \quad \textit{and} \quad \text{Var}^{(3)}\left(\sum_{i=1}^k U_i^K(3)\right) \le cK^2(1+\varepsilon \log^2 K).$$

4.6 First phase

This section is dedicated to the proof of Propositions 2 and 3. We first consider in detail the alignment $SL-N1-N2$ and prove that there are only four different possible ancestral relationships of the two neutral loci, then we calculate the probabilities for the non-negligible possibilities. In Section 4.6 we briefly consider the separated geometry, $N1-SL-N2$.

Coalescence and recombination probabilities, negligible events

Recall Definition 5 and define, for $j \in \{1,2\}$

$$r_j^* := r_1 + \mathbf{1}_{\{j=2\}}(r_2 - r_1 r_2), \quad \text{and} \quad r_{(1,2)}^* := r_1 r_2,$$

which denote the probability to have (only) one (resp. two) recombination(s) somewhere before the locus Nj (resp. before the locus $N2$) at a birth event.

Definition 6. *For $(\alpha,\alpha') \in \mathscr{A}^2$, $j \in \{1,2\}$ and $n \in \mathbb{N}^{\mathscr{A}}$ we define :*

$p_{\alpha\alpha'}^{(c_j)}(n)$ *:= probability that two randomly chosen neutral alleles, located at locus Nj and associated respectively with alleles α and α' at time τ_m^K, coalesce at this time conditionally on $(N_A,N_a)(\tau_{m-1}^K) = n$ and on the birth of an individual carrying allele α at time τ_m^K.*

$p_{\alpha\alpha'}^{(r_j^*)}(n)$ *:= probability to have one (and only one) recombination from the α- into the α'-population before locus Nj conditionally on $(N_A,N_a)(\tau_{m-1}^K) = n$ and on the birth of an individual carrying allele α at time τ_m^K.*

$p_{\alpha\alpha'}^{(r_{(1,2)}^*)}(n)$ *:= probability to have a double recombination under the same conditions*

Then we have the following result :

Lemma 7. *Let $\alpha \in \mathscr{A}$, $n \in \mathbb{N}^{\mathscr{A}}$ such that $n_a \le \lfloor \varepsilon K \rfloor$, $n_A \in I_\varepsilon^K 1$ and $j \in \{1,2\}$. Then there exists a finite c such that,*

$$p_{aa}^{(c_j)}(n) = \frac{2}{n_a(n_a+1)}\Big(1 - \frac{r_j^* f_A n_A}{f_A n_A + f_a n_a}\Big), \quad p_{aA}^{(c_j)}(n) = \frac{r_j^* f_A}{(n_a+1)(f_A n_A + f_a n_a)}, \quad \text{and} \quad p_{A\alpha}^{(c_j)}(n) \le \frac{c}{K^2}.$$

Démonstration. The proof of the two equalities can be found in Chapter 3 (Lemma 7.1) as the expression is the same for $n_A \in I_\varepsilon^K$ or $d(n_A, I_\varepsilon^K) = 1$. The only difference is that we consider two neutral loci and have to exclude the double recombination case. Indeed, if there are simultaneous recombinations the alleles located at SL and N2 in the newborn originate from the same parent. The expressions of $p_{A\alpha}^{(c_j)}(n)$ in the case where $n_A \in I_\varepsilon^K$ are also stated in Chapter 3 (Lemma 7.1), and from the definition of $\tilde{N}$ in (4.3.16) we get that when $d(n_A, I_\varepsilon^K) = 1$, $p_{AA}^{(c_j)}(n) = 2/n_A^2$ and $p_{Aa}^{(c_j)}(n) = 0$. This ends the proof. □

Next we focus on the recombination probabilities :

Lemma 8. *Let $\alpha \in \mathscr{A}$, $n \in \mathbb{N}^{\mathscr{A}}$ such that $n_A \le \lfloor \varepsilon K \rfloor$, $n_A \in I_\varepsilon^K 1$ and $j \in \{1,2,(1,2)\}$. Then there exist two finite constants c and ε_0 such that for every $\varepsilon \le \varepsilon_0$,*

$$p_{aa}^{(r_j^*)}(n) = \frac{r_j^* f_a (n_a - 1)}{(n_a+1)(f_A n_A + f_a n_a)}, \quad p_{aA}^{(r_j^*)}(n) = \frac{r_j^* f_A n_A}{(n_a+1)(f_A n_A + f_a n_a)},$$

$$p_{Aa}^{(r_j^*)}(n) \le \frac{c\varepsilon}{K \log K} \quad \text{and} \quad (1 - c\varepsilon)\frac{r_2}{n_A} \le p_{AA}^{(r_2)}(n_A, k) \le \frac{r_2}{n_A} \tag{4.6.1}$$

Démonstration. The second equality is stated in Chapter 3 Equation (7.2).

Conditional on the birth of an α-individual and the state of the process at the $(m-1)$-th jump, the probability of picking the newborn when choosing an individual at random amongst the a-individuals is equal to $1/(n_a+1)$. A recombination before the locus Nj (or before locus $N1$ and locus $N2$ if $j=12$) happens with probability r_j^*, independent of all other events. Finally, the probability that the second parent is an a-individual but is different from the first parent is equal to $f_a(n_a-1)/(f_An_A+f_an_a)$. This proves the first equality.

When $n_A \in I_\varepsilon^K$ we get similarly that

$$p_{AA}^{(r_j^*)}(n) = \frac{r_j^* f_A(n_A-1)}{(n_A+1)(f_An_A+f_an_a)} \quad \text{and} \quad p_{Aa}^{(r_j^*)}(n) = \frac{r_j^* f_a n_a}{(n_A+1)(f_An_A+f_an_a)},$$

and from the definition of $\tilde{N}$ in (4.3.16) we obtain that when $d(n_A, I_\varepsilon^K)=1$, $p_{AA}^{(r_2^*)}(n) = r_2(n_A-1)/n_A^2$ and $p_{Aa}^{(r_j^*)}(n)=0$. Condition (4.1.1) completes the proof. □

Remark 11. *Let us recall the definition of I_ε^K in (4.3.1). Then there exist three finite constants c, ε_0 and K_0 such that for $\varepsilon \le \varepsilon_0$, $K \ge K_0$, $j \in \{1,2,(1,2)\}$, $n_A \in I_\varepsilon^K 1$ and $k < \lfloor \varepsilon K \rfloor$,*

$$(1-c\varepsilon)\frac{r_j^*}{k+1} \le p_{aA}^{(r_j^*)}(n_A,k) \le \frac{r_j^*}{k+1} \quad \textit{and} \quad p_{aa}^{(r_2)}(n_A,k) \le \frac{f_a}{f_A}\frac{r_2}{n_A} \le \frac{c}{K\log K}. \tag{4.6.2}$$

Recalling the definition of the mth jump time in (4.4.1), we define for $k \in \{1,2,(1,2)\}$ and $m \in \mathbb{N}$,

$$(\alpha i k)_m := \{m \le J^K(1) \text{ and the } k\text{-th locus/loci of the } i\text{-th individual is/are associated to an allele } \alpha \text{ at the } m\text{-th jump time}\}. \tag{4.6.3}$$

The notation $(\alpha i 1)_m, (\alpha' i 2)_m$ here implies that the two neutral loci of individual i are associated to two distinct individuals at the mth jump time, for any $\alpha, \alpha' \in \mathscr{A}$.

To approximate the genealogy of the neutral alleles sampled at the end of the first phase we will focus on the recombinations and coalescences which may happen during this time interval. We first prove that we can neglect some event combinations. Let d be in $\mathbb{N}$. Sample d distinct individuals uniformly at the end of the sweep and define :

aAa : a neutral allele recombines from the a-population to the A-population, and then (backwards in time) back into the a-population

CR : two neutral alleles coalesce in the a-population, and then (backwards in time) recombine into the A-population

CA : two neutral alleles coalesce and at least one of them carries the allele A at the time of coalescence

$2R$: a neutral allele takes part in a double recombination (i.e. a recombination before $N1$ and a recombination before $N2$ at the same birth event)

$R2a$: a recombination separates the two neutral loci of an individual within the a-population

We can bound the probability of these events as follows :

Lemma 9. *There exist three positive finite constants c, K_0 and ε_0 such that for $\varepsilon \le \varepsilon_0$ and $K \ge K_0$*

$$\mathbb{P}^{(1)}(aAa) + \mathbb{P}^{(1)}(CR) + \mathbb{P}^{(1)}(2R) + \mathbb{P}^{(1)}(R2a) \le \frac{c}{\log K}, \quad \textit{and} \quad \mathbb{P}^{(1)}(CA) \le \frac{c \log K}{K}.$$

Démonstration. The probabilities of events aAa, CR and CA are bounded in Chapter 3 Lemma 7.3 and Equation (7.19) for the process N. But according to Lemmas 7 and 8 the coalescence and recombination probabilities for the process $\tilde{N}$ are very close or even smaller when $d(n_A, I_\varepsilon^K) = 1$ than when N and $\tilde{N}$ are equal. Hence we just have to bound the probability of $2R$ and $R2a$. If a neutral allele experiences a double recombination, it happens either when it is associated with an allele a, or with an allele A. From Lemma 8 and the fact that r_1 and r_2 are of order $1/\log K$ we get for $k < \lfloor \varepsilon K \rfloor$:

$$\sup_{n_A \in I_\varepsilon^K 1} \left(p_{aa}^{(r^*_{(1,2)})}(n_A, k) + p_{aA}^{(r^*_{(1,2)})}(n_A, k) \right) \le \frac{c}{(k+1)\log^2 K},$$

and

$$\sup_{n_A \in I_\varepsilon^K 1} \left(p_{Aa}^{(r^*_{(1,2)})}(n_A, k) + p_{AA}^{(r^*_{(1,2)})}(n_A, k) \right) \le \frac{c}{K \log^2 K}.$$

Recall the definitions of $U_k^K(1)$ and $\mathscr{U}_k^K(1)$ in (4.5.7) and (4.5.23) respectively. As a birth of an α-individual is needed to have a recombination from the α- to the α'-population, we can bound the probability to have a double recombination by :

$$\mathbb{P}^{(1)}(2R) \le \frac{c}{\log^2 K} \mathbb{E}^{(1)} \Big[\sum_{k=1}^{\lfloor \varepsilon K \rfloor - 1} \Big(\frac{U_k^K(1)}{k+1} + \frac{\mathscr{U}_k^K(1)}{K} \Big) \Big].$$

By applying inequality (4.5.9) and Lemma 4 we succeed in bounding $\mathbb{P}^{(1)}(2R)$ by a constant over $\log K$. It remains to consider the event $R2a$ of a recombination within the a-population. Define the first time (with respect to the backwards-in-time process) that this event happens :

$$R_{aa}^{(1)}(i) := \sup\{m,\ m \le J^K(1) \text{ and both neutral loci of the } i\text{-th individual are associated to distinct } a\text{-individuals at the } (m-1)\text{th jump}, \tag{4.6.4}$$

where $R^{(1)}_{aa}(i) = -\infty$ if the event does not happen during the first phase of the sweep. Then,

$$\begin{aligned}
\mathbb{P}^{(1)}(R^{(1)}_{aa}(i) \geq 0) &= \sum_{l=1}^{\lfloor \varepsilon K \rfloor - 1} \mathbb{P}^{(1)}(R^{(1)}_{aa}(i) \geq 0, \tilde{N}_a(\tau^K_{R^{(1)}_{aa}(i)}) = l) \\
&= \sum_{l=1}^{\lfloor \varepsilon K \rfloor - 1} \sum_{m<\infty} \mathbb{P}^{(1)}(m \leq J^K(1), \tilde{N}_a(\tau^K_{m-1}) = l, \tilde{N}_a(\tau^K_m) = l+1, \\
&\qquad\qquad (ai1)_m, (ai2)_m, \forall m' > m : (ai12)_{m'}) \\
&\leq \sum_{l=1}^{\lfloor \varepsilon K \rfloor - 1} \sum_{m<\infty} \sup_{n_A \in I^K_\varepsilon 1} \left(p^{(r_2)}_{aa}(n_A, l) \mathbb{P}^{(1)}_{(n_A, l+1)}(\forall m \geq 0 : (ai12)_m) \right) \\
&\qquad\qquad \mathbb{P}^{(1)}(m \leq J^K(1), \tilde{N}_a(\tau^K_{m-1}) = l, \tilde{N}_a(\tau^K_m) = l+1) \\
&\leq \sum_{l=1}^{\lfloor \varepsilon K \rfloor - 1} \frac{c}{K \log K} \mathbb{E}^{(1)}[U^K_l(1)] \leq \frac{c}{\log K},
\end{aligned}$$

by (4.5.9) and (4.6.2). □

To simplify the notations we will denote the union of all negligible events by

$$NE := aAa \cup CR \cup CA \cup 2R \cup R2a. \tag{4.6.5}$$

The two loci of one individual separate within the A-population

Having excluded events of small probability, there are exactly two ways for the neutral alleles of an individual sampled at the end of the first phase to originate from two distinct A-individuals. These two ways are described page 139 and represented in Figure 4.4.

Event $[2,1]^{rec}_{A,i}$

The aim of this section is to prove the following approximation :

Proposition 7. *Let i be an a-individual sampled uniformly at the end of the first phase. There exist two finite constants c and ε_0 such that for $\varepsilon \leq \varepsilon_0$,*

$$\limsup_{K\to\infty} \left| \mathbb{P}^{(1)}([2,1]^{rec}_{A,i}) - \left[\frac{r_2}{r_1+r_2} - e^{-\frac{r_1}{s}\log\lfloor \varepsilon K \rfloor} + \frac{r_1}{r_1+r_2} e^{-\frac{r_1+r_2}{s}\log\lfloor \varepsilon K \rfloor} \right] \right| \leq c\sqrt{\varepsilon}.$$

We first give a preliminary Lemma before proving Proposition 7. Recall (4.4.1) and define for $k \in \{1, 2, (1,2)\}$ and $m \in \mathbb{N}$,

$$\begin{aligned}
R(i,k) := \sup\{m, m \leq J^K(1) \text{ and the } k\text{-th locus/loci of the } i\text{-th individual} \\
\text{is/are associated to an allele } A \text{ at the } (m-1)\text{th jump time}\},
\end{aligned} \tag{4.6.6}$$

the last jump (forwards in time) when the k-th locus/loci of the i-th individual belongs to the A-population (with $\sup \emptyset = -\infty$). To prove Proposition 7 the idea is to decompose the event

$[2,1]^{rec}_{A,i}$ according to the different possible a-population sizes when the first (backwards in time) recombination between $N1$ and $N2$ occurs.

$$\begin{aligned}\mathbb{P}^{(1)}([2,1]^{rec}_{A,i}) &= \mathbb{P}^{(1)}(R(i,2)>R(i,1)\geq 0)\\ &= \sum_{l=1}^{\lfloor \varepsilon K\rfloor}\mathbb{P}^{(1)}(R(i,1)\geq 0, R(i,2)>R(i,1), \tilde{N}_a(\tau^K_{R(i,2)})=l)\\ &= \sum_{l=1}^{\lfloor \varepsilon K\rfloor -1}\mathbb{P}^{(1)}(R(i,2)>R(i,1), \tilde{N}_a(\tau^K_{R(i,2)})=l)\\ &\qquad \mathbb{P}^{(1)}(R(i,1)\geq 0 | R(i,2)>R(i,1), \tilde{N}_a(\tau^K_{R(i,2)})=l). \qquad (4.6.7)\end{aligned}$$

In the following Lemma, which then gives rise to the proof of Proposition 7, we consider separately the two probabilities of the above product :

Lemma 10. *There exist three finite constants c, K_0 and ε_0 such that for $K\geq K_0$, $\varepsilon\leq\varepsilon_0$ and $l<\lfloor\varepsilon K\rfloor$,*

$$\left|\mathbb{P}^{(1)}(R(i,2)>R(i,1), \tilde{N}_a(\tau^K_{R(i,2)})=l) - r_2\frac{1-(1-s)^{\lfloor\varepsilon K\rfloor-l}-(1-s)^{l+1}}{s(l+1)}e^{-\frac{r_1+r_2}{s}\log\frac{\lfloor\varepsilon K\rfloor}{l}}\right|\leq\frac{c\sqrt{\varepsilon}}{l\log K} \qquad (4.6.8)$$

and

$$\left|\mathbb{P}^{(1)}(R(i,1)\geq 0 | R(i,2)>R(i,1), N^K_a(\tau^K_{R(i,2)})=l) - \sum_{k=1}^{l-1}\frac{r_1}{s(k+1)}e^{-\frac{r_1}{s}\log\frac{l-1}{k}}\right|\leq c\sqrt{\varepsilon}. \qquad (4.6.9)$$

Proof of Proposition 7. From Lemma 10 and Equation (4.6.7) we get the existence of a finite c such that for K large enough and ε small enough,

$$\mathbb{P}^{(1)}([2,1]^{rec}_{A,i})\leq\sum_{l=1}^{\lfloor\varepsilon K\rfloor-1}\left[\frac{r_2}{s(l+1)}e^{-\frac{r_1+r_2}{s}\log\frac{\lfloor\varepsilon K\rfloor}{l}}+\frac{c\sqrt{\varepsilon}}{l\log K}\right]\left[\sum_{k=1}^{l-1}\frac{r_1}{s(k+1)}e^{-\frac{r_1}{s}\log\frac{l-1}{k}}+c\sqrt{\varepsilon}\right]. \qquad (4.6.10)$$

Rewritting the second term in brackets and applying Lemma 14 with $c_N/\log N = r_1/s$ yields :

$$\begin{aligned}e^{-\frac{r_1}{s}\log(l-1)}\frac{r_1}{s}\sum_{k=1}^{l-1}\frac{1}{k+1}e^{\frac{r_1}{s}\log k}+c\sqrt{\varepsilon} &\leq e^{-\frac{r_1}{s}\log(l-1)}\left(e^{\frac{r_1}{s}\log l}-1+c\frac{r_1}{s}\right)+c\sqrt{\varepsilon}\\ &\leq 1-e^{-\frac{r_1}{s}\log l}+c\sqrt{\varepsilon},\end{aligned}$$

for K large enough, ε small enough and a finite c, whose value can change from line to line and which can be chosen independently of l. We use in the last inequality Condition (4.1.1) which claims that $\limsup_{K\to\infty} r_1\log K<\infty$. Including the last inequality in (4.6.10) gives

$$\mathbb{P}^{(1)}([2,1]^{rec}_{A,i})\leq\sum_{l=1}^{\lfloor\varepsilon K\rfloor-1}\frac{r_2}{s(l+1)}e^{-\frac{r_1+r_2}{s}\log\lfloor\varepsilon K\rfloor}\left(e^{\frac{r_1+r_2}{s}\log l}-e^{\frac{r_2}{s}\log l}\right)+c\sqrt{\varepsilon},$$

for a finite c, K large enough and ε small enough, where we again use (4.1.1) which ensures that exponential terms are bounded away from zero and infinity in the following sense :

$$\frac{1}{c} \leq \liminf_{K\to\infty} e^{-\frac{r_1+r_2}{s}\log\lfloor \varepsilon K\rfloor} \leq \limsup_{K\to\infty} e^{\frac{r_1+r_2}{s}\log\lfloor \varepsilon K\rfloor} \leq c$$

for a finite c. Applying again Lemma 14, we get :

$$\mathbb{P}^{(1)}([2,1]^{rec}_{A,i}) \leq \left(\frac{r_2}{r_1+r_2} - e^{-\frac{r_1}{s}\log\lfloor \varepsilon K\rfloor} + \frac{r_1}{r_1+r_2} e^{-\frac{r_1+r_2}{s}\log\lfloor \varepsilon K\rfloor}\right) + c\sqrt{\varepsilon}.$$

The lower bound is obtained in the same way. Notice that it is a little bit more involved as we need to use (B.2) in addition. □

The end of this section is devoted to the proof of Lemma 10.

Proof of Equation (4.6.8). We can decompose the event $\{R(i,2) > R(i,1), \tilde{N}_a(\tau^K_{R(i,2)}) = l\}$ according to the jump number of the (backwards in time) first recombination. Recall the definition of $NR(i)^{(1)}$ page 139. We will use this event with a different initial condition for $(\tilde{N}_A, \tilde{N}_a)$, which will not necessarily be $(\lfloor \bar{n}_A K\rfloor, 1)$. It will however still correspond to the absence of any recombination before the end of the first phase. We recall conventions (4.1.8) and (4.3.4). With the definition of $(\alpha i k)_m$ in (4.6.3) we get

$$\begin{aligned}
&\mathbb{P}^{(1)}(R(i,2) > R(i,1), N_a(\tau^K_{R(i,2)}) = l)\\
&= \sum_{m>1} \mathbb{P}^{(1)}(m \leq J^K(1), \tilde{N}_a(\tau^K_{m-1}) = l-1, \tilde{N}_a(\tau^K_m) = l, (ai1)_{m-1}, (Ai2)_{m-1}, \forall m \leq m' \leq J^K(1) : (ai12)_{m'})\\
&\leq \sum_{m>1} \sup_{n_A \in I^K_\varepsilon 1} \left\{ p^{(r_2)}_{aA}(n_A, l-1)\mathbb{P}^{(1)}_{(n_A,l)}(NR(i)^{(1)})\right\} \mathbb{P}^{(1)}(m \leq J^K(1), \tilde{N}_a(\tau^K_{m-1}) = l-1, \tilde{N}_a(\tau^K_m) = l)\\
&= \sup_{n_A \in I^K_\varepsilon 1} \left\{ p^{(r_2)}_{aA}(n_A, l-1)\mathbb{P}^{(1)}_{(n_A,l)}(NR(i)^{(1)})\right\} \mathbb{E}^{(1)}[U^K_{l-1}(1)], \quad (4.6.11)
\end{aligned}$$

and the same expression with the infimum on $n_A \in I^K_\varepsilon 1$ for a lower bound. To enlight the proof the approximation of the second probability, $\mathbb{P}^{(1)}_{(n_A,l)}(NR(i)^{(1)})$, is derived in Lemma 11 in Appendix A. The first probability in the sum, $p^{(r_1)}_{aA}(n_A, l-1)$, can be bounded thanks to (4.6.2). This yields,

$$\begin{aligned}
\mathbb{P}^{(1)}(R(i,2) > R(i,1), \tilde{N}_a(\tau^K_{R(i,2)}) = l) &\leq (1+c\varepsilon)\frac{r_2}{l+1}(e^{-\frac{r_1+r_2}{s}\log\frac{\lfloor \varepsilon K\rfloor}{l}} + c\sqrt{\varepsilon})\mathbb{E}^{(1)}[U^K_{l-1}(1)]\\
&\leq (1+c\sqrt{\varepsilon})\frac{r_2}{l+1} e^{-\frac{r_1+r_2}{s}\log\frac{\lfloor \varepsilon K\rfloor}{l}} \mathbb{E}^{(1)}[U^K_{l-1}(1)],
\end{aligned}$$

for a finite c, ε small enough and K large enough, where we used that $(r_1+r_2)\log K$ is bounded. We similarly get a lower bound and end up the proof of Equation (4.6.8) by applying (4.5.9). □

Proof of Equation (4.6.9). We will decompose the event considered here according to the value of $\tilde{N}_a$ when the first (backwards in time) recombination occurs. Let us denote by $\zeta_k^K(1)$ the jump number of the last hitting of $k \le \lfloor \varepsilon K \rfloor$ by $\tilde{N}_a$ during the first phase,

$$\zeta_k^K(1) := \sup\{m, \tau_m^K \le \tilde{T}_\varepsilon^K, \tilde{N}_a(\tau_m^K) = k\}, \tag{4.6.12}$$

and recall (4.5.20). Then we can define the events

$$\begin{aligned} NR(l,\xi,i) := \{&\text{the first locus of individual } i \text{ sampled at jump time } \tau^K_{\xi_l^K(1)} \\ &\text{does not recombine from the } a\text{- to the } A\text{-population between 0 and } \tau^K_{\xi_l^K(1)}\} \end{aligned} \tag{4.6.13}$$

where $\xi \in \{\zeta, \sigma\}$. Similarly as in (4.6.11), Bayes' rule leads to :

$$\begin{aligned} &\mathbb{P}^{(1)}(R(i,1) \ge 0 \mid R(i,2) > R(i,1), \tilde{N}_a(\tau^K_{R(i,2)}) = l) \\ &= \sum_{k=1}^{\lfloor \varepsilon K \rfloor} \mathbb{P}^{(1)}(R(i,1) \ge 0, \tilde{N}_a(\tau^K_{R(i,1)}) = k \mid R(i,2) > R(i,1), \tilde{N}_a(\tau^K_{R(i,2)}) = l), \\ &\le \sum_{k=1}^{\lfloor \varepsilon K \rfloor} \Big(\sup_{n_A \in I_\varepsilon^K 1} p_{aA}^{(r_1)}(n_A, k-1) \mathbb{P}^{(1)}_{(n_A,k)}(NR(l,\sigma,i)) \Big) \Sigma(k,l), \end{aligned} \tag{4.6.14}$$

where for sake of simplicity we have introduced the notation

$$\Sigma(k,l) := \sum_{m<\infty} \mathbb{P}^{(1)}(m < R(i,2), \tilde{N}_a(\tau^K_{m-1}) = k-1, \tilde{N}_a(\tau^K_m) = k \mid \tilde{N}_a(\tau^K_{R(i,2)}) = l).$$

The lower bound is obtained by taking the infimum for n_A in $I_\varepsilon^K 1$ and replacing σ by ζ. To enlight the proof, the expectations of these two quantities are stated in Lemma 11.

First we prove that with a probability close to one the a-population size is bigger when the (backwards in time) first recombination occurs than when the second, of locus $(i,1)$, occurs. Note that by (4.5.9) and Lemma 3, there exists a finite c such that for every $l < k < \lfloor \varepsilon K \rfloor$:

$$\Sigma(k,l) \le \mathbb{E}^{(1)}[U_l^K(1)] \sup_{n_A \in I_\varepsilon^K 1} \mathbb{E}^{(1)}_{(n_A,l+1)}[U^K_{n_A,l,k-1}(1) | \sigma_l^K(1) < \infty] \le c\mu_\varepsilon^{k-l},$$

where we recall that $\mu_\varepsilon < 1$ for ε small enough, and we used (4.5.9) and Lemma 3. Hence, recalling (4.6.14) and (4.6.2), we obtain for $k > l$

$$\mathbb{P}^{(1)}(R(i,1) \ge 0, \tilde{N}_a^K(\tau^K_{R(i,1)}) \ge l | R(i,2) > R(i,1), \tilde{N}_a(\tau^K_{R(i,2)}) = l) \le c r_1 \sum_{k=l+1}^{\lfloor \varepsilon K \rfloor} \frac{\mu_\varepsilon^{k-l}}{k} \le \frac{c}{\log K},$$

for a finite c and ε small enough. We therefore can ignore all $k > l$ in the sum in (4.6.14) and continue with the case $k \le l$. Here, we can bound the sum as follows :

$$\mathbb{E}^{(1)}[U_{k-1}^K(1)] - \sup_{n_A \in I_\varepsilon^K 1} \mathbb{E}^{(1)}_{(n_A,l-1)}[U^K_{n_A,l,k-1}(1)] \mathbb{E}^{(1)}[U_l^K(1)] \le \Sigma(k,l) \le \mathbb{E}^{(1)}[U_{k-1}^K(1)].$$

Bounding the difference between the two bounds within Equation (4.6.14) then yields

$$\sum_{k=1}^{l} \frac{r_1}{k} \sup_{n_A \in I_\varepsilon^K 1} \mathbb{E}^{(1)}_{(n_A, l-1)}[U^K_{n_A,l,k-1}(1)]\mathbb{E}^{(1)}[U^K_l(1)] \leq c r_1 \sum_{k=1}^{l} \frac{\mu_\varepsilon^{l-k}}{k} \leq \frac{c}{\log K},$$

for a finite c by (4.6.2), (4.5.9) and Lemma 3 and thus we can work with $\mathbb{E}^{(1)}[U^K_{k-1}(1)]$ as an approximation for the sum $\Sigma(k,l)$. Reasoning in the same way to get a lower bound and using (4.6.2) and (A.2) we get the existence of a finite c such that for K large enough and ε small enough,

$$\left|\mathbb{P}^{(1)}(R(i,1) \geq 0 | R(i,2) > R(i,1), N_a^K(\tau^K_{R(i,2)}) = l) - \sum_{k=1}^{l-1} \frac{r_1}{k} e^{-\frac{r_1}{s}\log\frac{l-1}{k}} \mathbb{E}^{(1)}[U^K_k(1)]\right| \leq c\sqrt{\varepsilon}.$$

Applying (4.5.9) and (B.2) yields Equation (4.6.9). Notice that we have replaced $1/k$ by $1/(k+1)$. We used Condition (4.1.1) to do this. □

Event $[12,2]^{rec}_{A,i}$

Recall the definition of $[12,2]^{rec}_{A,i}$ page 139. This section is devoted to the proof of the following result :

Proposition 8. *Let i be an individual sampled uniformly at the end of the first phase. There exist two finite constants c and ε_0 such that for $\varepsilon \leq \varepsilon_0$,*

$$\limsup_{K\to\infty} \left|\mathbb{P}^{(1)}([12,2]^{rec}_{A,i}) - r_1\left[\frac{1-e^{-\frac{r_1+r_2}{s}\log\lfloor\varepsilon K\rfloor}}{r_1+r_2} + \frac{e^{-\frac{r_1+r_2}{s}\log\lfloor\varepsilon K\rfloor} - e^{-\frac{f_A r_2}{f_a s}\log\lfloor\varepsilon K\rfloor}}{r_1 + r_2(1-f_A/f_a)}\right]\right| \leq c\sqrt{\varepsilon}.$$

Démonstration. As the proof is very similar to the proof of Proposition 7 we will be very brief here and only give the ingredients. Let us introduce for $l < \lfloor\varepsilon K\rfloor$ the event :

$$RA(l,i) := \{[12,2]^{rec}_{A,i}, R(i,2) = R(i,1) \geq 0, \tilde{N}_a(\tau^K_{R(i,1)}) = l\}. \tag{4.6.15}$$

Then we can rewrite the probability of $[12,2]^{rec}_{A,i}$ as follows :

$$\mathbb{P}^{(1)}([12,2]^{rec}_{A,i}) = \sum_{l=1}^{\lfloor\varepsilon K\rfloor} \mathbb{P}^{(1)}(RA(l,i))\mathbb{P}^{(1)}(R(i,2) = R(i,1) \geq 0, N_a^K(\tau^K_{R(i,1)}) = l). \tag{4.6.16}$$

Apart from the point of recombination, the second probability in the above sum coincides with the probability studied in Equation (4.6.8) and we obtain for ε small enough and K large enough,

$$\sup_{l\leq\lfloor\varepsilon K\rfloor} l.\Big|\mathbb{P}^{(1)}(R(i,2) = R(i,1) \geq 0, N_a^K(\tau^K_{R(i,1)}) = l) - \frac{r_1(1-(1-s)^{\lfloor\varepsilon K\rfloor - l} - (1-s)^{l+1})}{s(l+1)} e^{-\frac{r_1+r_2}{s}\log\frac{\lfloor\varepsilon K\rfloor}{l}}\Big| \leq c\frac{\sqrt{\varepsilon}}{\log K}, \tag{4.6.17}$$

for a finite c, when substituting r_2 by r_1 in the fraction which mirrors the recombination probability. The probability of $RA(l,i)$ is derived in Lemma 11. Inserting (4.6.17) and (A.3) into (4.6.16) yields

$$\begin{aligned}\mathbb{P}^{(1)}([12,2]^{rec}_{A,i}) &\leq \sum_{l=1}^{\lfloor \varepsilon K\rfloor}(1-e^{-\frac{f_A}{fa}\frac{r_2}{s}\log l})\frac{r_1}{l+1}e^{-\frac{r_1+r_2}{s}\log\frac{\lfloor \varepsilon K\rfloor}{l}}+c\sqrt{\varepsilon}\\ &\leq r_1 e^{-\frac{r_1+r_2}{s}\log\lfloor \varepsilon K\rfloor}\Big[\frac{e^{\frac{r_1+r_2}{s}\log\lfloor \varepsilon K\rfloor}-1}{r_1+r_2}-\frac{e^{\frac{r_1+r_2-f_A r_2/fa}{s}\log\lfloor \varepsilon K\rfloor}-1}{r_1+r_2-f_A r_2/f_a}\Big]+c\sqrt{\varepsilon}\end{aligned}$$

where we again applied Lemma 14 to express the sum in a different way, and used the finitness of $\limsup_{K\to\infty}(r_1+r_2)\log K$ assumed in Condition (4.1.1). Reasoning similarly for the lower bound and rearranging the terms ends the proof of Proposition 8. □

Proof of Proposition 2

Event $R2(i)^{(1)}$ *:* By definition and from Lemma 9,

$$\mathbb{P}^{(1)}(R2(i)^{(1)})=\mathbb{P}^{(1)}(R(i,2)\geq 0)-\mathbb{P}^{(1)}(R(i,1)\geq 0)+O\Big(\frac{\log K}{K}\Big),$$

where $R(i,1)$ and $R(i,2)$ have been defined in (4.6.6). But these probabilities have already been derived in Chapter 3 Lemma 7.4, and we get :

$$\mathbb{P}^{(1)}(R2(i)^{(1)})=(1-q_1q_2)-(1-q_1)++O_K(\varepsilon)=q_1(1-q_2)+O_K(\varepsilon),$$

where $O_K(\varepsilon)$ satisfies (4.4.5).

Event $R1|2(i)^{(1,ga)}$ *:* By definition (see page 139)

$$\mathbb{P}^{(1)}(R1|2(i)^{(1,ga)})=\mathbb{P}^{(1)}([2,1]^{rec}_{A,i})+\mathbb{P}^{(1)}([12,2]^{rec}_{A,i}).$$

The result then follows from Propositions 7 and 8.

Event $R12(i)^{(1)}$ *:* From Definition (4.6.15) and Equation (A.3) we obtain for K large enough,

$$\begin{aligned}\mathbb{P}^{(1)}(R12(i)^{(1)}) &= \sum_{l=1}^{\lfloor \varepsilon K\rfloor}(1-\mathbb{P}^{(1)}(RA(l,i)))\mathbb{P}^{(1)}(R(i,1)=R(i,2)\geq 0,\tilde{N}_a(\tau^K_{R(i,2)})=l)\\ &= r_1\sum_{l=1}^{\lfloor \varepsilon K\rfloor}e^{-\frac{f_A}{fa}\frac{r_2}{s}\log l}\frac{1-(1-s)^{\lfloor \varepsilon K\rfloor-l}-(1-s)^{l+1}}{s(l+1)}e^{-\frac{r_1+r_2}{s}\log\frac{\lfloor \varepsilon K\rfloor}{l}}+O_K(\sqrt{\varepsilon})\\ &= \frac{r_1}{r_1+r_2-f_Ar_2/f_a}\Big(e^{-\frac{r_2}{s}\frac{f_A}{fa}\log\lfloor \varepsilon K\rfloor}-e^{-\frac{r_1+r_2}{s}\log\lfloor \varepsilon K\rfloor}\Big)+O_K(\sqrt{\varepsilon}),\end{aligned}$$

where we again used the statement of Lemma 14 to substitute the sum, and Equation (B.2).

Event $NR(i)^{(1)}$ *:* From Lemma 9,

$$\mathbb{P}^{(1)}(NR(i)^{(1)}) = 1 - \mathbb{P}^{(1)}(R2(i)^{(1)}) - \mathbb{P}^{(1)}(R12(i)^{(1)}) - \mathbb{P}^{(1)}(R2(i)^{(1)}) + O\Big(\frac{\log K}{K}\Big).$$

This ends up the proof of Proposition 2.

Proof of Proposition 3

All results concerning the background birth and death process $(\tilde{N}_A, \tilde{N}_a)$ are independent of the geometry. The statements derived in Section 4.5 and the conclusions on negligible events can be used without modification. Note that we can stick with the recombination probabilities from Lemma 8 but have to keep in mind that the recombination probability $p_{\alpha\alpha'}^{(r_j^*)}(n)$ now refers to the probability to see a recombination between SL and Nj. Following the reasoning in Section 4.6, there are exactly two non-negligible events which lead to two singletons in the A-population : if $k \neq k' \in \{1,2\}$,

- First (i,k) recombines into the A-population, then (i,k') recombines into the A-population (and connects to a different individual than (i,k)). This event will be called $[k,k']^{rec}_{A,i}$.

A close look at the proofs of Lemma 10 shows that the calculations stay the same also in the case of the separated geometry. We therefore have an analogous statement. As both situations are symmetrical, the probability of $[1,2]^{rec}_{A,i}$ is obtained by exchanging r_1 and r_2 in the expression of $\mathbb{P}^{(1)}([2,1]^{rec}_{A,i})$, and the asymptotical probability of $R1|2(i)^{(1,gs)}$ follows by adding the probabilities for the events $[1,2]^{rec}_{A,i}$ and $[2,1]^{rec}_{A,i}$. Due to symmetry reasons we can treat the probabilities of $R1(i)^{(1)}$ and $R2(i)^{(1)}$ in one step. As the proof for $R1(i)^{(1)}$ from Proposition 2 only uses the results from Lemma 10 which still hold true in the second geometric alignment both claims follow at once. Finally, the statement about $NR(i)^{(1)}$ results from the previous observations about negligible events.

4.7 Second and third phases

This section is devoted to the proofs of Propositions 4 and 5.

Proof of Proposition 4

We need to show that two distinct lineages picked uniformly at the end of the second phase coalesce or recombine during that phase only with negligible probability. Let us recall the definition of the jumps τ_m^K in (4.4.1) and denote by $U^K(2)$ the number of upcrossings of the a-population during the second phase :

$$U^K(2) := \#\{m, T_\varepsilon^K < \tau_m^K \leq T_\varepsilon^K + t_\varepsilon, N_a(\tau_{m+1}^K) - N_a(\tau_m^K) = 1\}. \tag{4.7.1}$$

Let us introduce the event C_ε^K :

$$C_\varepsilon^K := \{T_\varepsilon^K \leq S_\varepsilon^K\} \cap \{N_a^K(t) \geq \varepsilon^2 K/4, \forall\ T_\varepsilon^K \leq t \leq T_\varepsilon^K + t_\varepsilon\}.$$

In particular on the event C_ε^K, for $T_\varepsilon^K \le \tau_m^K \le T_\varepsilon^K + t_\varepsilon$ and $j \in \{1,2\}$

$$p_{aA}^{(r_j)}(N(\tau_m^K)) \le \frac{8r_j}{\varepsilon^2 K} \quad \text{and} \quad p_{aa}^{(c_j)}(N(\tau_m^K)) \le \frac{32}{\varepsilon^4 K^2}.$$

Then if we recall the definition of $NR(i)^{(2)}$ page 141 we have for $m \in \mathbb{N}$,

$$\mathbb{P}^{(1)}(NR(i)^{(2)} | U^K(2) = m, C_\varepsilon^K) \ge \left(1 - \frac{8(r_1+r_2)}{\varepsilon^2 K}\right)^m. \tag{4.7.2}$$

But for K large enough, $\log(1 - 8(r_1+r_2)/(\varepsilon^2 K)) \ge -10(r_1+r_2)/(\varepsilon^2 K)$ and hence

$$\begin{aligned}\mathbb{P}^{(1)}(NR(i)^{(2)} | C_\varepsilon^K) &\ge \left(1 - \mathbb{P}^{(1)}(U^K(2) > K\log\log K | C_\varepsilon^K)\right) e^{K\log\log K \log(1 - \frac{8(r_1+r_2)}{\varepsilon^2 K})} \\ &\ge \left(1 - \mathbb{P}^{(1)}(U^K(2) > K\log\log K | C_\varepsilon^K)\right) e^{-\frac{10(r_1+r_2)\log\log K}{\varepsilon^2}}.\end{aligned}$$

But according to Condition (4.1.1) the exponential term is equivalent to 1 when K is large. Moreover, according to (4.3.5), N_a^K is smaller than $2\bar{n}_a K$ on the time interval $[T_\varepsilon^K, T_\varepsilon^K + t_\varepsilon]$ with probability close to 1. When this property holds, we can bound the birth number $U^K(2)$ by the sum of $2\bar{n}_a K$ iid Poisson random variables with parameter $f_a t_\varepsilon$. The strong law of large numbers then yields

$$\lim_{K\to\infty} \mathbb{P}^{(1)}(U^K(2) > K\log\log K | C_\varepsilon^K) = 0.$$

Applying again (4.3.5) to get $\lim_{K\to\infty} \mathbb{P}(C_\varepsilon^K | T_\varepsilon^K < \infty) = 1$ finally gives

$$\lim_{K\to\infty} \mathbb{P}(NR(i)^{(2)} | T_\varepsilon^K < \infty) = 1.$$

We prove in the same way the coalescence part in Proposition 4.

Proof of Proposition 5

The proof of the asymptotic probability of $R2(i)^{(3,ga)}$ is the same as for (A.3), except that the roles of A and a are exchanged. Hence we do not give more details. Note however that it extensively uses Lemma 6. Let us now focus on the event $NR(i)^{(3)}$, and introduce

$$NRA(i)^{(3)} := \{\text{no neutral allele of individual } i \text{ recombines from the } a \text{ to the } A \text{ population}\}.$$

Recall the definitions of $\mathbb{P}^{(3)}$ and $U^K(3)$ in (4.3.18) and (4.5.39) respectively. We decompose the probabilities according to the number of upcrossings of $\tilde{\tilde{N}}_a$ during the third phase and get in the same way as in (4.7.2), for $m \in \mathbb{N}$

$$\mathbb{P}^{(3)}(NRA(i)^{(3)} | U^K(3) = m, \{T_0^{(K,A)} < T_\varepsilon^{(K,A)} \wedge S_\varepsilon^{(K,a)}\}) \ge \left(1 - \frac{f_A(r_1+r_2)\varepsilon}{f_a(\bar{n}_a - M''\varepsilon)^2 K}\right)^m,$$

where $T_0^{(K,A)}$, $T_\varepsilon^{(K,A)}$ and $S_\varepsilon^{(K,a)}$ have been defined in (4.3.11). But for K large enough and ε small enough,

$$\log\left(1 - \frac{f_A(r_1+r_2)\varepsilon}{f_a(\bar{n}_a - M''\varepsilon)^2 K}\right) \ge -2 f_A \frac{(r_1+r_2)\varepsilon}{f_a \bar{n}_a^2 K}.$$

Hence we get for a finite constant c and ε small enough :

$$\begin{aligned}\mathbb{P}^{(3)}(NRA(i)^{(3)}) &\geq \Big(1-\mathbb{P}^{(3)}\Big(U^K(3)>\frac{K\log K}{\sqrt{\varepsilon}}\Big)\Big)\exp\Big(-\frac{2f_A(r_1+r_2)\sqrt{\varepsilon}\log K}{f_a\bar{n}_a^2}\Big)\\ &\geq \Big(1-\frac{\sqrt{\varepsilon}\mathbb{E}^{(3)}[U^K(3)]}{K\log K}\Big)\Big(1-\frac{2f_A(r_1+r_2)\sqrt{\varepsilon}\log K}{f_a\bar{n}_a^2}\Big)\geq(1-c\sqrt{\varepsilon})^2,\end{aligned}$$

where we used Lemma 6 and that $(r_1+r_2)\log K$ is bounded (Condition (4.1.1)).

The proof of the last part of Proposition 5 is very similar to that of Proposition 4. The key arguments are that the expectation of the birth number of a-individuals during the third phase under $\mathbb{P}^{(3)}$ is of order $K\log K$ (Lemma 6), whereas the probability for two neutral alleles associated with an allele a to coalesce is of order $1/K^2$ at each birth of an a-individual (Lemma 7).

4.8 Independence of neutral lineages

This section is dedicated to the proof of Proposition 6. We sample $d\in\mathbb{N}$ distinct individuals uniformly at the end of the first phase. We recall the definitions of the genealogical events for the adjacent geometry $SL-NI-N2$ during the first phase page 139 and introduce :

$$R(1|2):=\sum_{1\leq i\leq d}\mathbf{1}_{R1|2(i)^{(1,ga)}},\quad R(1):=R(1|2)+\sum_{1\leq i\leq d}\mathbf{1}_{R12(i)^{(1)}}\quad\text{and}\quad R(2):=R(1)+\sum_{1\leq i\leq d}\mathbf{1}_{R2(i)^{(1)}}$$

From Proposition 2 we know that $R(1)$, $R(2)$ and $R(1|2)$ are sufficient to describe the neutral genealogies at the end of the first phase up to a probability negligible with respect to one for large K. Let j,k,l be three integers such that $l\leq j$ and $j+k\leq d$. We aim at approximating

$$\begin{aligned}p(j,k,l):&=\mathbb{P}(R(1)=j,R(2)=j+k,R(1,2)=l|T_\varepsilon^K\leq S_\varepsilon^K)\\ &=\mathbb{P}(R(1)=j|T_\varepsilon^K\leq S_\varepsilon^K)\mathbb{P}(R(2)=j+k|T_\varepsilon^K\leq S_\varepsilon^K,R(1)=j)\\ &\quad\mathbb{P}(R(1,2)=l|T_\varepsilon^K\leq S_\varepsilon^K,R(1)=j,R(2)=j+k).\end{aligned}\tag{4.8.1}$$

The approximations of the two first probabilities are direct adaptations of Lemma 5.2 and proof of Proposition 2.6 in [SD05] pp 1623-1624. More precisely, Lemma 7.3 in Chapter 3 which states that with high probability two neutral lineages do not coalesce then recombine (backwards in time) allows us to get an equivalent of Lemma 5.2 (with $J=0$) in [SD05] :

$$\Big|\mathbb{P}(R(1)=j|T_\varepsilon^K\leq S_\varepsilon^K)-\binom{d}{j}\mathbb{E}^{(1)}[F_1^j(1-F_1)^{n-j}]\Big|\leq c\Big(\frac{1}{\log K}+\varepsilon\Big),$$

for ε small enough, where c is a finite constant,

$$F_1:=\mathbb{P}^{(1)}(R(i,1)\geq 0|((N_A,N_a)(\tau_n^K),n\leq J^K(1))),$$

and $R(i,1)$ is defined in (4.6.6). Then Equations (7.21), (7.23), (7.24) and (7.26) of Chapter 3 yield

$$\limsup_{K\to\infty}\Big|\mathbb{E}^{(1)}[F_1^j(1-F_1)^{d-j}]-(1-q_1)^j q_1^{(n-j)}\Big|\leq c\varepsilon,$$

where q_1 has been defined in (4.1.11), which allows to conclude

$$\limsup_{K\to\infty} \left| \mathbb{P}(R(1)=j | T_\varepsilon^K \le S_\varepsilon^K) - \binom{d}{j}(1-q_1)^j q_1^{(d-j)} \right| \le c\varepsilon, \tag{4.8.2}$$

for ε small enough where c is a finite constant.

The derivation of the second probability, $\mathbb{P}(R(2)=j+k | T_\varepsilon^K \le S_\varepsilon^K, R(1)=j)$, follows the same outline. The lineages where $N1$ does not escape the sweep can be seen as lineages where SL and $N1$ are the same locus and the recombination probability between $SL-N1$ and $N2$ is r_2. This is due to the independence of the recombinations between SL and $N1$ and between $N1$ and $N2$. Hence we can rewrite the probability as follows :

$$\mathbb{P}(R(2)=j+k | T_\varepsilon^K \le S_\varepsilon^K, R(1)=j) = \mathbb{P}(R(2)-R(1)=k | T_\varepsilon^K \le S_\varepsilon^K, d-R(1)=d-j).$$

We can then directly apply the result (4.8.2) for the law of $R(1)$ and get :

$$\limsup_{K\to\infty} \left| \mathbb{P}(R(2)=j+k | T_\varepsilon^K \le S_\varepsilon^K, R(1)=j) - \binom{d-j}{k}(1-q_2)^k q_2^{(d-j-k)} \right| \le c\varepsilon, \tag{4.8.3}$$

for ε small enough where c is a finite constant and q_2 has been defined in (4.1.11).

The derivation of the last probability in (4.8.1) is more involved but follows the same spirit. First let us notice that we only have to focus on genealogies where $N1$ escapes the sweep. Hence the derivation of the probability comes down to the derivation of $\mathbb{P}(R(1|2)=l | T_\varepsilon^K \le S_\varepsilon^K, R(1)=j)$. According to Lemma 9, with high probability there is no recombination between SL and $N1$ after (backwards in time) a coalescence among the d sampled individuals. There is no coalescence either after a recombination between SL and $N1$ (this is due to the large number of A-individuals ; similar proof as for the last probability of Proposition 5). Hence if we introduce

$$NC(2,0) := \{\text{there is no coalescence between those of the } d \text{ sampled neutral alleles} \\ \text{located at } N1 \text{ which undergo a recombination between } SL \text{ and } N1\},$$

we get :

$$\mathbb{P}(R(1,2)=l | T_\varepsilon^K \le S_\varepsilon^K, R(1)=j) = \mathbb{P}(R(1,2)=l | T_\varepsilon^K \le S_\varepsilon^K, R(1)=j, NC(2,0)) + O\Big(\frac{\log K}{K}\Big).$$

Let us denote by $(i_1,...,i_j)$ the j sampled lineages which undergo a recombination between SL and $N1$. We will construct the neutral genealogies to get a property of "nearly independence" : first we add the recombinations between SL and $N1$. As we restrict our attention to the event $\{R(1)=j, NC(2,0)\}$, the lineages whose allele at $N1$ escapes the sweep do not coalesce. We add the recombinations between $N1$ and $N2$ in the lineage i_1. If the lineage i_1 do not coalesce at locus $N2$ with a lineage i_k with $1 \le k \le j$ (event called $NC(2,1)$), then we add the recombinations between $N1$ and $N2$ in the lineage i_2. On the event $\{R(1)=j, NC(2,0), NC(2,1)\}$

these recombinations are independent from the recombination between $N1$ and $N2$ in the lineage i_1. Then if the lineage i_2 does not coalesce at locus $N2$ with a lineage i_k with $1 \le k \le j$ (event called $NC(2,2)$), then we add the recombinations between $N1$ and $N2$ in the lineage i_3, and so on. At each step the probability to coalesce with a lineage in $(i_1,...,i_j)$ is negligible thanks to Lemma 9, and hence the lineages whose allele at $N1$ escapes the sweep undergo "independently" a recombination between $N1$ and $N2$ on an event of high probability. More precisely, if we introduce for $1 \le k \le j$ and $\delta \in \{0,1\}$

$$\{r_{i_k} = \delta\} := \{\text{there is } \delta \text{ recombination between } N1 \text{ and } N2 \text{ in the lineage } i_k\},$$

then for $(\delta_1,...,\delta_j) \in \{0,1\}^j$

$$\mathbb{P}(r_{i_k} = \delta_k, 1 \le k \le j | T_\varepsilon^K \le S_\varepsilon^K, R(1) = j) = \prod_{1 \le k \le j} \mathbb{P}(r_{i_k} = \delta_k | T_\varepsilon^K \le S_\varepsilon^K, R(1) = j, NC(2,0), NC(2,1), ..., NC(2,k-1)) + O\Big(\frac{\log K}{K}\Big).$$

Indeed, the probability that the event $NC(2,k)$ is not realized after adding the recombinations between $N1$ and $N2$ in lineage i_k has order $\log K/K$ according to Lemma 9. But for $1 \le k \le j$,

$$\begin{aligned}
&\mathbb{P}(r_{i_k} = \delta_k | T_\varepsilon^K \le S_\varepsilon^K, R(1) = j, NC(2,0), ..., NC(2,k-1)) \\
&\qquad = \frac{\mathbb{P}(r_{i_k} = \delta_k, R(1) = j, NC(2,0), ..., NC(2,k-1) | T_\varepsilon^K \le S_\varepsilon^K)}{\mathbb{P}(R(1) = j, NC(2,0), ..., NC(2,k-1) | T_\varepsilon^K \le S_\varepsilon^K)} \\
&= \frac{\mathbb{P}(r_{i_k} = \delta_k, R(1) = j | T_\varepsilon^K \le S_\varepsilon^K) - \mathbb{P}(r_{i_k} = \delta_k, R(1) = j, (NC(2,0) \cap ... \cap NC(2,k-1))^c | T_\varepsilon^K \le S_\varepsilon^K)}{\mathbb{P}(R(1) = j | T_\varepsilon^K \le S_\varepsilon^K) - \mathbb{P}(R(1) = j, (NC(2,0) \cap ... \cap NC(2,k-1))^c | T_\varepsilon^K \le S_\varepsilon^K)},
\end{aligned} \tag{4.8.4}$$

and according to Lemma 9 and Coupling (4.3.14), there exists a finite c such that for K large enough and ε small enough,

$$\mathbb{P}((NC(2,0) \cap ... \cap NC(2,k-1))^c | T_\varepsilon^K \le S_\varepsilon^K) \le c\Big(\frac{\log K}{K} + \varepsilon\Big).$$

As $\mathbb{P}(r_{i_k} = \delta_k, R(1) = j | T_\varepsilon^K \le S_\varepsilon^K)$ does not go to 0 when K goes to infinity, we get

$$\begin{aligned}
&\mathbb{P}(r_{i_k} = \delta_k | T_\varepsilon^K \le S_\varepsilon^K, R(1) = j, NC(2,0), ..., NC(2,k-1)) = \mathbb{P}(r_{i_k} = \delta_k | T_\varepsilon^K \le S_\varepsilon^K, R(1) = j) + O\Big(\frac{\log K}{K} + \varepsilon\Big) \\
&= \delta_k \frac{\mathbb{P}(R1|2(i_k)^{(1,ga)} | T_\varepsilon^K \le S_\varepsilon^K)}{\mathbb{P}(R(i_k,1) \ge 0 | T_\varepsilon^K \le S_\varepsilon^K)} + (1-\delta_k)\Big(1 - \frac{\mathbb{P}(R1|2(i_k)^{(1,ga)} | T_\varepsilon^K \le S_\varepsilon^K)}{\mathbb{P}(R(i_k,1) \ge 0 | T_\varepsilon^K \le S_\varepsilon^K)}\Big) + O\Big(\frac{\log K}{K} + \varepsilon\Big) \\
&= \delta_k \frac{1-q_1-q_3}{1-q_1} + (1-\delta_k)\Big(1 - \frac{1-q_1-q_3}{1-q_1}\Big) + O\Big(\frac{\log K}{K} + \varepsilon\Big),
\end{aligned}$$

where we recall the definition of $R(i_k,1)$ in (4.6.6), the definition of $R1|2(i_k)^{(1,ga)}$ page 139, and we used Proposition 2. Adding Equations (4.8.2) and (4.8.3) we finally obtain :

$$\begin{aligned}
p(j,k,l) &= \binom{n}{j}(1-q_1)^j q_1^{(n-j)} \binom{n-j}{k}(1-q_2)^k q_2^{(n-j-k)} \binom{j}{l}\Big(1 - \frac{q_3}{1-q_1}\Big)^l \Big(\frac{q_3}{1-q_1}\Big)^{j-l} + O_K(\varepsilon) \\
&= \frac{n!}{l!(j-l)!k!(n-j-k)!}(q_1q_2)^{n-j-k}(q_1(1-q_2))^k q_3^{j-l}(1-q_1-q_3)^l + O_K(\varepsilon).
\end{aligned} \tag{4.8.5}$$

This ends the proof of the independence between genealogies during the first phase.

The derivation of the asymptotical independence of neutral lineages during the third phase is an easy adaptation of Lemma 5.2 and proof of Proposition 2.6 in [SD05] pp 1623-1624 as with high probability two lineages do not coalesce during this phase.

A Lemma 11

Recall the definition of $NR(i)^{(1)}$ page 139, and Definitions (4.6.13) and (4.6.15). Then we have the following approximations for large K.

Lemma 11. *There exist three finite constants c, K_0 and ε_0 such that for every $K \geq K_0$ and $\varepsilon \leq \varepsilon_0$*

$$\sup_{n_A \in I_\varepsilon^K 1, l \leq \lfloor \varepsilon K \rfloor} \left| \mathbb{P}^{(1)}_{(n_A,l)}(NR(i)^{(1)}) - \exp\Big(- \frac{r_1+r_2}{s} \log \frac{\lfloor \varepsilon K \rfloor}{l} \Big) \right| \leq c\sqrt{\varepsilon}, \tag{A.1}$$

$$\sup_{\tau \in \{\zeta,\sigma\}} \sup_{n_A \in I_\varepsilon^K 1, k \leq l \leq \lfloor \varepsilon K \rfloor} \left| \mathbb{P}^{(1)}_{(n_A,k)}(NR(l,\tau,i) - \exp\Big(- \frac{r_1}{s} \log \frac{l-1}{k} \Big) \right| \leq c\sqrt{\varepsilon}, \tag{A.2}$$

$$\sup_{l \leq \lfloor \varepsilon K \rfloor} \left| \mathbb{P}^{(1)}(RA(l,i)) - \Big(1 - \exp\Big(- \frac{f_A}{f_a} \frac{r_2}{s} \log l \Big) \Big) \right| \leq c\sqrt{\varepsilon}. \tag{A.3}$$

Démonstration. Let us introduce the sigma-algebra generated by the trait population process

$$\mathscr{F} := \sigma\Big((\tilde{N}_A, \tilde{N}_a)(\tau_n^K), \tau_n^K \leq \tilde{T}_\varepsilon^K \Big).$$

We use some ideas developed in [SD05]. The proof, although quite technical, can be summarized easily : for $(g,b,c,d,f) \in \mathbb{R}_+^5$, Triangle Inequality and the Mean Value Theorem imply

$$|g - e^{-b}| \leq |g - e^{-c}| + |c-d| + |d-f| + |f-b|.$$

Hence for every random variables $(X_1, X_2) \in \mathbb{R}_+^2$ and measurable event C :

$$\begin{aligned} \left| \mathbb{P}^{(1)}(C|\mathscr{F}) - e^{-\frac{r_1+r_2}{s} \log \frac{\lfloor \varepsilon K \rfloor}{l}} \right| \leq & \left| \mathbb{P}^{(1)}(C|\mathscr{F}) - e^{-X_1} \right| + \left| X_1 - X_2 \right| \\ & + \left| X_2 - \mathbb{E}^{(1)}[X_2] \right| + \left| \mathbb{E}^{(1)}[X_2] - \frac{r_1+r_2}{s} \log \frac{\lfloor \varepsilon K \rfloor}{l} \right|. \end{aligned}$$

By taking the expectation and applying Jensen and Cauchy-Schwarz Inequalities, we obtain :

$$\begin{aligned} \left| \mathbb{P}^{(1)}(C) - e^{-\frac{r_1+r_2}{s} \log \frac{\lfloor \varepsilon K \rfloor}{l}} \right| \leq & \mathbb{E}^{(1)} \left| \mathbb{P}^{(1)}(C|\mathscr{F}) - e^{-X_1} \right| + \mathbb{E}^{(1)} \left| X_1 - X_2 \right| \\ & + \sqrt{\mathrm{Var}(X_2)} + \left| \mathbb{E}^{(1)}[X_2] - \frac{r_1+r_2}{s} \log \frac{\lfloor \varepsilon K \rfloor}{l} \right|. \end{aligned} \tag{A.4}$$

Hence the idea is to find the appropriate random variables $(X_1, X_2) \in \mathbb{R}_+^2$ to get small quantities on the right hand side.

Proof of Equation (A.1) *:* The first step consists in working conditionally on $\mathscr{F}$, describing this probability as a product of conditional probabilities close to one, and derive a Poisson approximation. To this aim, we define for $m \in \mathbb{N}$:

$$\theta^{(r_1+r_2)}(m) := \mathbf{1}_{\{\tau_m^K \le \tilde{T}_\varepsilon^K\}} \mathbf{1}_{\tilde{N}_a(\tau_m^K)-\tilde{N}_a(\tau_{m-1}^K)=1\}} (p_{aA}^{(r_1^*)} + p_{aA}^{(r_2^*)})(\tilde{N}_A, \tilde{N}_a)(\tau_{m-1}^K),$$

where we recall the definition of the $p_{\alpha\alpha'}^{(r_i^*)}$ in Definition 5. Notice that Remark 11 implies that for $\rho \in \{1,2\}$, $n_A \in I_\varepsilon^K 1$ and $l < \lfloor \varepsilon K \rfloor$,

$$(1-c\varepsilon)(r_1+r_2)^\rho \Big(\sum_{k=1}^{l-1} \frac{\mathbb{E}^{(1)}_{(n_A,l)} U_k^K(1)}{(k+1)^\rho} \Big) \le \mathbb{E}^{(1)}_{(n_A,l)} \Big[\sum_{m=1}^{\infty} (\theta^{(r_1+r_2)}(m) \mathbf{1}_{\{\tilde{N}_a(\tau_m^K)<l\}})^\rho \Big] \le (r_1+r_2)^\rho \Big(\sum_{k=1}^{l-1} \frac{\mathbb{E}^{(1)}_{(n_A,l)} U_k^K(1)}{(k+1)^\rho} \Big). \tag{A.5}$$

Then, similarly as in [SD05], we have for $n_A \in I_\varepsilon^K 1$ and $l < \lfloor \varepsilon K \rfloor$

$$\mathbb{P}^{(1)}_{(n_A,l)}(NR(i)^{(1)}|\mathscr{F}) = \prod_{m=1}^{\infty} (1-\theta^{(r_1+r_2)}(m)), \quad \mathbb{P}^{(1)}_{(n_A,l)} - \text{a.s.}$$

If we introduce the variable,

$$\eta^{(12)} := \sum_{m=1}^{\infty} \theta^{(r_1+r_2)}(m),$$

which will play the role of X_1 in (A.4), we get by following the path of Lemma 3.6 in [SD05] :

$$\mathbb{E}^{(1)}_{(n_A,l)} \Big| \prod_{m=1}^{\infty} (1-\theta^{(r_1+r_2)}(m)) - \exp(-\eta^{(12)}) \Big| \le \mathbb{E}^{(1)}_{(n_A,l)} \Big[\sum_{m=1}^{\infty} (\theta^{(r_1+r_2)}(m))^2 \Big] \le \frac{c}{\log^2 K}, \tag{A.6}$$

for K large enough, $n_A \in I_\varepsilon^K 1, l < \lfloor \varepsilon K \rfloor$ and a finite c (which can be chosen independently of l), where we used Equations (4.5.9) (4.5.10) and (A.5), and Condition (4.1.1) for the last inequality. Next we introduce an approximation of the random variable $\eta^{(12)}$, namely

$$\tilde{\eta}^{(12)} := \sum_{m=1}^{\infty} \theta^{(r_1+r_2)}(m) \mathbf{1}_{\{\tilde{N}_a(\tau_m^K) \ge \tilde{N}_a(0)\}}, \tag{A.7}$$

which will play the role of X_2 in (A.4). For $n_A \in I_\varepsilon^K 1$ and $l \le \lfloor \varepsilon K \rfloor$:

$$0 \le \mathbb{E}^{(1)}_{(n_A,l)}[\eta^{(12)} - \tilde{\eta}^{(12)}] = \mathbb{E}^{(1)}_{(n_A,l)} \Big[\sum_{m=1}^{\infty} \theta^{(r_1+r_2)}(m) \mathbf{1}_{\{\tilde{N}_a(\tau_m^K)<l\}} \Big] \tag{A.8}$$

$$\le \frac{r_1+r_2}{s_+(\varepsilon) s_-^2(\varepsilon)} \sum_{k=1}^{l-1} \frac{(1-s_-(\varepsilon))^{l-k}}{k+1} \le c \frac{(r_1+r_2)}{l}, \tag{A.9}$$

for a finite c and ε small enough, where we used (A.5) and (4.5.10) for the first inequality, and (B.2) for the second one. This latter ensures that c can be chosen independently of l. The expected value of $\tilde{\eta}^{(12)}$ can be bounded by using (A.5), (4.5.9) and (B.2)

$$\begin{aligned} \mathbb{E}^{(1)}_{(n_A,l)}[\tilde{\eta}^{(12)}] &\ge (1-c\varepsilon)(r_1+r_2) \sum_{k=l}^{\lfloor \varepsilon K \rfloor - 1} \frac{1}{k+1} \Big(\frac{1-(1-s)^{\lfloor \varepsilon K \rfloor - k} - (1-s)^{k+1}}{s} - c\varepsilon \Big) \\ &\ge (1-c\varepsilon) \frac{r_1+r_2}{s} \log \frac{\lfloor \varepsilon K \rfloor}{l} - \frac{c}{\log K}, \end{aligned} \tag{A.10}$$

for a finite c and ε small enough. For the upper bound we get similarly,

$$\mathbb{E}^{(1)}_{(n_A,l)}[\tilde{\eta}^{(12)}] \leq (1+c\varepsilon)\frac{r_1+r_2}{s}\log\frac{\lfloor \varepsilon K\rfloor}{l}. \tag{A.11}$$

The last step consists in bounding the variance of $\tilde{\eta}^{(12)}$. As the calculation of this variance is quite involved, we introduce an approximation of $\tilde{\eta}^{(12)}$, namely

$$\tilde{\tilde{\eta}}^{(12)} := \sum_{m=1}^{\infty} \mathbf{1}_{\{\tilde{N}_a(\tau^K_{m-1})\geq \tilde{N}_a(0)\}}\mathbf{1}_{\{\tilde{N}_a(\tau^K_m)-\tilde{N}_a(\tau^K_{m-1})=1\}}\frac{r_1+r_2}{\tilde{N}_a(\tau^K_{m-1})+1} = \sum_{k=\tilde{N}_a(0)}^{\lfloor \varepsilon K\rfloor -1}\frac{r_1+r_2}{k+1}U^K_k(1).$$

Equation (4.6.2) yields $(1-c\varepsilon)\tilde{\tilde{\eta}}^{(12)} \leq \tilde{\eta}^{(12)} \leq \tilde{\tilde{\eta}}^{(12)}$ for a finite c and ε small enough. Hence

$$\begin{aligned}\left|\mathrm{Var}^{(1)}_{(n_A,l)}\tilde{\eta}^{(12)} - \mathrm{Var}^{(1)}_{(n_A,l)}\tilde{\tilde{\eta}}^{(12)}\right| &\leq c\varepsilon\mathbb{E}^{(1)}_{(n_A,l)}\left[\left(\tilde{\tilde{\eta}}^{(12)}\right)^2\right] \\ &\leq c\varepsilon(r_1+r_2)^2\sum_{k,k'=l}^{\lfloor \varepsilon K\rfloor -1}\frac{\mathbb{E}^{(1)}[(U^K_k(1))^2]+\mathbb{E}^{(1)}[(U^K_{k'}(1))^2]}{(k+1)(k'+1)} \leq c\varepsilon,\end{aligned} \tag{A.12}$$

where we used (4.5.14) and (B.3) which ensure that $U^K_k(1)$ is smaller than a geometric random variable with parameter $q_k^{(s_-(\varepsilon),s_+(\varepsilon))} \geq s_-(\varepsilon)$. Thus it is enough to bound $\mathrm{Var}^{(1)}_{(n_A,l)}\tilde{\tilde{\eta}}^{(12)}$. Thanks to (4.5.11) and Condition (4.1.1) we get :

$$\begin{aligned}\mathrm{Var}^{(1)}_{(n_A,l)}\tilde{\tilde{\eta}}^{(12)} &= (r_1+r_2)^2\sum_{k,k'=l}^{\lfloor \varepsilon K\rfloor -1}\frac{\mathrm{Cov}^{(1)}_{(n_A,l)}(U^K_k(1),U^K_{k'}(1))}{(k+1)(k'+1)} \\ &\leq 2(r_1+r_2)^2\sum_{k\leq k'=l}^{\lfloor \varepsilon K\rfloor -1}\frac{\lambda_\varepsilon^{(k'-k)/2}+\varepsilon}{(k+1)(k'+1)} \leq c\frac{\log\lfloor \varepsilon K\rfloor}{\log^2 K}(c+\varepsilon\log\lfloor \varepsilon K\rfloor).\end{aligned}$$

Recalling (A.12) and again Condition (4.1.1), we finally obtain

$$\limsup_{K\to\infty}\mathrm{Var}^{(1)}_{(n_A,l)}\tilde{\eta}^{(12)} \leq c\varepsilon, \tag{A.13}$$

for a finite c independent of l and ε small enough. Applying (A.4) with $X_1 = \eta^{(12)}$ and $X_2 = \tilde{\eta}^{(12)}$ yields

$$\begin{aligned}\left|\mathbb{P}^{(1)}_{(n_A,l)}(NR(i)^{(1)}) - e^{-\frac{r_1+r_2}{s}\log\frac{\lfloor \varepsilon K\rfloor}{l}}\right| &\leq \mathbb{E}^{(1)}_{(n_A,l)}\left|\prod_{m=1}^{\infty}(1-\theta^{(r_1+r_2)}(m)) - \exp(-\eta^{(12)})\right| \\ &+ \mathbb{E}^{(1)}_{(n_A,l)}[\eta^{(12)}-\tilde{\eta}^{(12)}] + \sqrt{\mathrm{Var}^{(1)}_{(n_A,l)}\tilde{\eta}^{(12)}} + \left|\mathbb{E}^{(1)}_{(n_A,l)}[\tilde{\eta}^{(12)}] - \frac{r_1+r_2}{s}\log\frac{\lfloor \varepsilon K\rfloor}{l}\right|.\end{aligned}$$

We end the proof of Equation (A.1) with Inequalities (A.6), (A.8), (A.13), (A.10) and (A.11).

Proof of (A.2) *:* There is a supplementary difficulty due to the randomness of $\tilde{N}_a(\tau^K_{R(i,2)})$. In the previous case we were interested in an event before the first hitting of $\lfloor \varepsilon K\rfloor$, while in the current case, the conditioning on the value of $\tilde{N}_a(\tau^K_{R(i,2)})$ does not tell us how many times $\tilde{N}_a$

has hit this value before. This is why we have introduced $NR(l,\sigma,i)$ and $NR(l,\zeta,i)$ in (4.6.13). Define for $m \geq 1$,

$$\theta^{(r_1)}(m) := \mathbf{1}_{\{\tau_m^K \leq \tilde{T}_\varepsilon^K\}} \mathbf{1}_{\{\tilde{N}_a(\tau_m^K) - \tilde{N}_a(\tau_{m-1}^K) = 1\}} p_{aA}^{(r_1)}((\tilde{N}_A, \tilde{N}_a)(\tau_m^K)).$$

We again condition on the trait population process and get for $n_A \in I_\varepsilon^K 1$ and $k \leq l < \lfloor \varepsilon K \rfloor$,

$$\mathbb{P}^{(1)}_{(n_A,k)}(NR(l,\sigma,i)|\mathscr{F}) = \prod_{m=1}^{\sigma_l^K(1)} (1 - \theta^{(r_1)}(m)), \ \mathbb{P}^{(1)}_{(n_A,k)} - \text{a.s.}, \tag{A.14}$$

and the same expression with σ replacing ζ. We define the corresponding parameters for the Poisson approximation as follows :

$$\eta_l^{(1),-} := \sum_{m=1}^{\sigma_l^K(1)} \theta^{(r_1)}(m), \ \text{ and } \ \eta_l^{(1),+} := \sum_{m=1}^{\zeta_l^K(1)} \theta^{(r_1)}(m).$$

They will play the role of X_1 in (A.4). We will show that both can be approximated by :

$$\tilde{\eta}_l^{(1)} := \sum_{m=1}^{\zeta_l^K(1)} \theta^{(r_1)}(m) \mathbf{1}_{\{\tilde{N}_a(0) \leq \tilde{N}_a(\tau_m^K) \leq l\}}, \tag{A.15}$$

which will play the role of X_2 in (A.4). Recall Definitions (4.5.7), (4.5.16) and (4.5.21). On the one hand, for $n_A \in I_\varepsilon^K 1$ and $k < \lfloor \varepsilon K \rfloor$,

$$\begin{aligned}
\mathbb{E}^{(1)}_{(n_A,k)}[\eta_l^{(1),+} - \tilde{\eta}_l^{(1)}] &= \mathbb{E}^{(1)}_{(n_A,k)}\Big[\sum_{m=1}^{\zeta_l^K(1)} \theta^{(r_1)}(m)(\mathbf{1}_{\{N_a^K(\tau_m^K) < k\}} + \mathbf{1}_{\{N_a^K(\tau_m^K) > l\}}) \Big] \\
&\leq \mathbb{E}^{(1)}[D_k^K(1)] \sum_{j=1}^{k-1} \sup_{n_A \in I_\varepsilon^K} p_{aA}^{(r_1)}(n_A, j) \sup_{n_A \in I_\varepsilon^K 1} \mathbb{E}^{(1)}_{(n_A,k-1)}[U^K_{n_A,k,j}(1)] \\
&+ \mathbb{E}^{(1)}[U_l^K(1)] \sum_{j=l+1}^{\lfloor \varepsilon K \rfloor} \sup_{n_A \in I_\varepsilon^K} p_{aA}^{(r_1)}(n_A, j) \sup_{n_A \in I_\varepsilon^K 1} \mathbb{E}^{(1)}_{(n_A,l+1)}[U^K_{n_A,l,j}(1) | \sigma_l^K(1) < \infty],
\end{aligned} \tag{A.16}$$

where we used that in the first phase, under $\mathbb{P}^{(1)}$, the number of excursions below k (resp. above l) is equal to $D_k^K(1)$ (resp. $U_l^K(1) - 1$). Applying Inequalities (4.5.9), (4.5.18), Lemma 3, and Equation (4.6.2), we get the existence of a finite c such that for ε small enough :

$$\mathbb{E}^{(1)}_{(n_A,k)}[\eta_l^{(1),+} - \tilde{\eta}_l^{(1)}] \leq c r_1 \sum_{j=1}^{\lfloor \varepsilon K \rfloor} \frac{\mu_\varepsilon^{|j-l|}}{j+1} \leq \frac{c}{\log K},$$

as $\mu_\varepsilon \in (0,1)$ for ε small enough and by Condition (4.1.1). On the other hand, by using the same results as in (A.16), we get

$$\begin{aligned}
\mathbb{E}^{(1)}_{(n_A,k)}[|\eta_l^{(1),-} - \tilde{\eta}_l^{(1)}|] &\leq \mathbb{E}^{(1)}_{(n_A,k)}\Big[\sum_{m=1}^{\sigma_l^K(1)} \theta^{(r_1)}(m) \mathbf{1}_{\{\tilde{N}_a(\tau_m^K) < k\}} + \sum_{m=\sigma_l^K(1)+1}^{\zeta_l^K(1)} \theta^{(r_1)}(m) \mathbf{1}_{\{k \leq \tilde{N}_a(\tau_m^K) \leq l\}} \Big] \\
&\leq c r_1 \Big(\sum_{j=1}^{k-1} \frac{\mu_\varepsilon^{k-j}}{j+1} + \sum_{j=k}^{l-1} \frac{\mu_\varepsilon^{l-j}}{j+1} \Big) \leq \frac{c}{\log K}.
\end{aligned}$$

This shows that it is enough to use $\tilde{\eta}_l^{(1)}$ for the Poisson approximation. From (A.6) we deduce that this approximation holds true up to terms of order $1/\log^2 K$. Recalling once again (A.4), we see that it only remains to calculate the expected value of $\tilde{\eta}_l^{(1)}$ and to bound its variance. The expectation can be approximated in the same way as the expected value of $\tilde{\eta}_l^{(12)}$ from the previous part in (A.10) and (A.11) :

$$(1-c\varepsilon)\frac{r_1}{s}\log\frac{l-1}{k}-\frac{c}{\log K}\leq \mathbb{E}^{(1)}_{(n_A,k)}[\tilde{\eta}_l^{(1)}]\leq (1+c\varepsilon)\frac{r_1}{s}\log\frac{l-1}{k}. \tag{A.17}$$

A comparison of the definitions of $\tilde{\eta}_l^{(1)}$ in (A.15) and $\tilde{\eta}^{(12)}$ in (A.7) shows that the variance of $\tilde{\eta}_l^{(1)}$ can be bounded by the same expression, that is, a constant times ε. This ends the proof of Equation (A.2).

Proof of Equation (A.3) It can be done in a similar way as for Equations (A.1) and (A.2). We have the following lower and upper bounds :

$$\prod_{m=1}^{\zeta_l^K(1)}\left[1-p_{AA}^{(r_2)}(\tilde{N}_A,\tilde{N}_a)(\tau_m^K)\right]\leq 1-\mathbb{P}^{(1)}(RA(l,i)|\mathscr{F})\leq \prod_{m=1}^{\sigma_l^K(1)}\left[1-p_{AA}^{(r_2)}(\tilde{N}_A,\tilde{N}_a)(\tau_m^K)\right]. \tag{A.18}$$

Once again we aim at deriving a Poisson approximation. As a birth event in the A-population is needed to see a recombination within the A-population, bounds on the expected number of jumps will concern the process $\tilde{N}_A$ and we have to use Lemma 4. □

B Technical results

This section is dedicated to technical results needed in the proofs. First we recall a well known result on the hitting times of birth and death processes which can be found in [AN72] :

Proposition 9. *Let $Z=(Z_t)_{t\geq 0}$ be a birth and death process with individual birth and death rates b and d. For $i\in\mathbb{Z}^+$, $T_i=\inf\{t\geq 0, Z_t=i\}$ and $\mathbb{P}_i$ is the law of Z when $Z_0=i$. Then for $(i,j,k)\in\mathbb{Z}_+^3$ such that $j\in(i,k)$,*

$$\mathbb{P}_j(T_k<T_i)=\frac{1-(d/b)^{j-i}}{1-(d/b)^{k-i}}. \tag{B.1}$$

We also recall Lemma 3.5 in [SD05] and the first part of Equation (A.16) in Chapter 3 which are used several times :

Lemma 12.

- *If $a>1$ there is a C such that for every $N\in\mathbb{N}$,*

$$\sum_{j=1}^{N}\frac{a^j}{j}\leq\frac{Ca^N}{N}. \tag{B.2}$$

- *Recall Definition* (4.5.13). *Then for* $(s_1, s_2) \in (0,1)^2$ *and* $k < \lfloor \varepsilon K \rfloor$,

$$q_k^{(s_1 \wedge s_2, s_1 \vee s_2)} \geq s_1 \wedge s_2 \tag{B.3}$$

Finally, we state two technical results. The first one can be proven by using characteristic functions, the proof of the second Lemma is given below :

Lemma 13. *Let* V *be a geometric random variable with parameter* p_1 *and* $(G^i, i \in \mathbb{N})$ *a sequence of independent geometric random variables with parameter* p_2*, independent of* V*. Then the random variable :*

$$Z := \sum_{i \leq V} G^i$$

is geometrically distributed with parameter $p_1 p_2$*.*

Lemma 14. *Let* $(c_N, N \in \mathbb{N})$ *be a bounded sequence of* $\mathbb{R}$*. Then there exists a finite constant* c *such that*

$$\limsup_{N \to \infty} \sup_{k \leq N} \Big| \sum_{l=1}^{k-1} \frac{e^{\frac{c_N}{\log N} \log l}}{l+1} - \frac{\log N}{c_N} (e^{\frac{c_N}{\log N} \log k} - 1) \Big| \leq c.$$

Démonstration. We prove the Lemma for a sequence $(c_N, N \in \mathbb{N})$ in $\mathbb{R}^*$ and extend the result by using the convention

$$\Big(\frac{\log N}{c_N} (e^{\frac{c_N}{\log N} \log k} - 1) \Big)_{|c_N = 0} = \log k.$$

The idea is to compare the sum with the integral

$$\int_1^k x^{\frac{c_N}{\log N} - 1} dx = \frac{\log N}{c_N} (e^{\frac{c_N}{\log N} \log k} - 1).$$

Let l be in $\{1, ..., N-1\}$. Then we have

$$\begin{aligned} \int_l^{l+1} x^{\frac{c_N}{\log N} - 1} dx - \frac{l^{\frac{c_N}{\log N}}}{l+1} &= \frac{\log N}{c_N} \Big((l+1)^{\frac{c_N}{\log N}} - l^{\frac{c_N}{\log N}} - \frac{c_N}{\log N} \frac{l^{\frac{c_N}{\log N}}}{l+1} \Big) \\ &= \frac{\log N}{c_N} l^{\frac{c_N}{\log N}} \Big(\Big(1 + \frac{1}{l} \Big)^{\frac{c_N}{\log N}} - 1 - \frac{c_N}{(l+1) \log N} \Big). \end{aligned}$$

An application of the Taylor-Lagrange formula yields that

$$\Big(1 + \frac{1}{l} \Big)^{\frac{c_N}{\log N}} - 1 = \frac{c_N}{l \log N} + \frac{c_N}{\log N} \Big(\frac{c_N}{\log N} - 1 \Big) \frac{1}{2 l^2} \Big(1 + x \Big)^{\frac{c_N}{\log N} - 2}$$

where x belongs to $[0, 1/l]$. As the sequence $(c_N, N \in \mathbb{N})$ is bounded, we deduce that there exists a finite constant c such that

$$\Big| \int_l^{l+1} x^{\frac{c_N}{\log N} - 1} dx - \frac{l^{\frac{c_N}{\log N}}}{l+1} \Big| \leq \frac{c}{l^2}.$$

This ends up the proof of Lemma 14. □

Bibliographie

[AD99] L Alio and RA Doney, *Wiener–hopf factorization revisited and some applications*, Stochastics : An International Journal of Probability and Stochastic Processes **66** (1999), no. 1-2, 87–102.

[AGKV05] Valery I. Afanasyev, Jochen Geiger, Goetz. Kersting, and Vladimir A. Vatutin, *Criticality for branching processes in random environment*, Ann. Probab. **33** (2005), no. 2, 645–673.

[AK71a] Krishna B Athreya and Samuel Karlin, *Branching processes with random environments : 2 : limit theorems*, The Annals of Mathematical Statistics (1971), 1843–1858.

[AK+71b] Krishna B Athreya, Samuel Karlin, et al., *On branching processes with random environments : I : Extinction probabilities*, The Annals of Mathematical Statistics **42** (1971), no. 5, 1499–1520.

[AN72] Krishna B Athreya and Peter E Ney, *Branching processes*, vol. 28, Springer-Verlag Berlin, 1972.

[Asm86] Marjorie A Asmussen, *The dynamics of interlocus associations in the three-locus hitchhiking model*, Journal of mathematical biology **24** (1986), no. 4, 361–380.

[Ban08] Vincent Bansaye, *Proliferating parasites in dividing cells : Kimmel's branching model revisited*, Ann. Appl. Probab. **18** (2008), no. 3, 967–996.

[Ban09] ———, *Surviving particles for subcritical branching processes in random environment*, Stochastic Process. Appl. **119** (2009), no. 8, 2436–2464.

[Bar98] Nicholas H Barton, *The effect of hitch-hiking on neutral genealogies*, Genetical Research **72** (1998), no. 2, 123–133.

[Bar00] ———, *Genetic hitchhiking*, Philosophical Transactions of the Royal Society of London. Series B : Biological Sciences **355** (2000), no. 1403, 1553–1562.

[BB11] Vincent Bansaye and Christian Böinghoff, *Upper large deviations for branching processes in random environment with heavy tails*, Electron. J. Probab **16** (2011), 1900–1933.

[BB13] ———, *Lower large deviations for supercritical branching processes in random environment*, Proceedings of the Steklov Institute of Mathematics **282** (2013), no. 1, 15–34.

[BEKV13] NH Barton, AM Etheridge, J Kelleher, and A Veber, *Genetic hitchhiking in spatially extended populations*, Theoretical population biology **87** (2013), 75–89.

[BFMT13] Sylvain Billiard, Régis Ferrière, Sylvie Méléard, and Viet Chi Tran, *Stochastic dynamics of adaptive trait and neutral marker driven by eco-evolutionary feedbacks*, arXiv preprint arXiv :1310.6274 (2013).

[BGK05] Matthias Birkner, Jochen Geiger, and Götz Kersting, *Branching processes in random environment : a view on critical and subcritical cases*, Interacting stochastic systems, Springer, 2005, pp. 269–291.

[BH12] Christian Boeinghoff and Martin Hutzenthaler, *Branching diffusions in random environment*, Markov Proc. Rel. Fields **18** (2012), no. 2, 269–310.

[BHK+95] John M Braverman, Richard R Hudson, Norman L Kaplan, Charles H Langley, and Wolfgang Stephan, *The hitchhiking effect on the site frequency spectrum of dna polymorphisms.*, Genetics **140** (1995), no. 2, 783–796.

[Bie45] Irénée-Jules Bienaymé, *De la loi de multiplication et de la durée des familles*, Soc. Philomat. Paris Extraits, Sér ***5*** (1845), 37–39.

[Bin76] Nicholas H. Bingham, *Continuous branching processes and spectral positivity*, Stochastic Processes Appl. **4** (1976), no. 3, 217–242.

[BLG00] Jean Bertoin and Jean-François Le Gall, *The Bolthausen-Sznitman coalescent and the genealogy of continuous-state branching processes*, Probab. Theory Related Fields **117** (2000), no. 2, 249–266.

[Böi14] Christian Böinghoff, *Limit theorems for strongly and intermediately supercritical branching processes in random environment with linear fractional offspring distributions*, Stochastic Processes and their Applications (2014).

[BP99] Benjamin M Bolker and Stephen W Pacala, *Spatial moment equations for plant competition : understanding spatial strategies and the advantages of short dispersal*, The American Naturalist **153** (1999), no. 6, 575–602.

[BPS13] Vincent Bansaye, Juan Carlos Pardo, and Charline Smadi, *On the extinction of continuous state branching processes with catastrophes*, Electron. J. Probab **18** (2013), no. 106, 1–31.

[Bro85] Peter J. Brockwell, *The extinction time of a birth, death and catastrophe process and of a related diffusion model*, Adv. in Appl. Probab. **17** (1985), no. 2, 42–52.

[BS08] Rowan DH Barrett and Dolph Schluter, *Adaptation from standing genetic variation*, Trends in Ecology & Evolution **23** (2008), no. 1, 38–44.

[BS11] Vincent Bansaye and Florian Simatos, *On the scaling limits of galton watson processes in varying environment*, arXiv preprint arXiv :1112.2547 (2011).

[BSS15a] Rebekka Brink-Spalink and Charline Smadi, *Genealogies of two neutral loci after a selective sweep in a large population of varying size*, in preparation (2015).

[BSS15b] Rebekka Brink-Spalink and Anja Sturm, *An approximate sampling distribuion for multiple neutral loci after a selective sweep*, in preparation (2015).

[BT11] Vincent Bansaye and Viet Chi Tran, *Branching Feller diffusion for cell division with parasite infection*, ALEA Lat. Am. J. Probab. Math. Stat. **8** (2011), 95–127.

[BY05] Jean Bertoin and Marc Yor, *Exponential functionals of Lévy processes*, Probab. Surv. **2** (2005), 191–212.

[CB+08] Luis-Miguel Chevin, Sylvain Billiard, et al., *Hitchhiking both ways : effect of two interfering selective sweeps on linked neutral variation*, Genetics **180** (2008), no. 1, 301–316.

[CFM06] Nicolas Champagnat, Régis Ferrière, and Sylvie Méléard, *Unifying evolutionary dynamics : from individual stochastic processes to macroscopic models*, Theoretical population biology **69** (2006), no. 3, 297–321.

[Cha06] N. Champagnat, *A microscopic interpretation for adaptive dynamics trait substitution sequence models*, Stochastic Processes and their Applications **116** (2006), no. 8, 1127–1160.

[CJM14] Nicolas Champagnat, Pierre-Emmanuel Jabin, and Sylvie Méléard, *Adaptation in a stochastic multi-resources chemostat model*, Journal de Mathématiques Pures et Appliquées **101** (2014), no. 6, 755–788.

[CM11] Nicolas Champagnat and Sylvie Méléard, *Polymorphic evolution sequence and evolutionary branching*, Probability Theory and Related Fields **151** (2011), no. 1-2, 45–94.

[CMM11] Pierre Collet, Sylvie Méléard, and Johan AJ Metz, *A rigorous model study of the adaptive dynamics of mendelian diploids*, Journal of Mathematical Biology (2011), 1–39.

[CMPR13] Camille Coron, Sylvie Méléard, Emmanuelle Porcher, and Alexandre Robert, *Quantifying the mutational meltdown in diploid populations*, The American Naturalist **181** (2013), no. 5, 623–636.

[Cor13] Camille Coron, *Slow-fast stochastic diffusion dynamics and quasi-stationary distributions for diploid populations*, arXiv preprint arXiv :1309.3405 (2013).

[Cor14] ———, *Stochastic modeling of density-dependent diploid populations and the extinction vortex*, Advances in Applied Probability **46** (2014), no. 2, 446–477.

[CPY97] Philippe Carmona, Frédérique Petit, and Marc Yor, *On the distribution and asymptotic results for exponential functionals of Lévy processes*, Exponential functionals and principal values related to Brownian motion, Bibl. Rev. Mat. Iberoamericana, Rev. Mat. Iberoamericana, Madrid, 1997, pp. 73–130.

[CR12] Graham Coop and Peter Ralph, *Patterns of neutral diversity under general models of selective sweeps*, Genetics **192** (2012), no. 1, 205–224.

[Dek87] FM Dekking, *On the survival probability of a branching process in a finite state iid environment*, Stochastic processes and their applications **27** (1987), 151–157.

[DGC08] Etienne Danchin, Luc-Alain Giraldeau, and Frank Cézilly, *Behavioural ecology : An evolutionary perspective on behaviour*, Oxford University Press, 2008.

[DH97] JC D'Souza and BM Hambly, *On the survival probability of a branching process in a random environment*, Advances in Applied Probability (1997), 38–55.

[DM02] Ron A. Doney and Ross A. Maller, *Stability and attraction to normality for Lévy processes at zero and at infinity*, J. Theoret. Probab. **15** (2002), no. 3, 751–792.

[DMM08] Michel Durinx, JAJ Hans Metz, and Géza Meszéna, *Adaptive dynamics for physiologically structured population models*, Journal of mathematical biology **56** (2008), no. 5, 673–742.

[DP14] Marcio A Diniz and Adriano Polpo, *A simple proof for the multinomial version of the representation theorem*, The Contribution of Young Researchers to Bayesian Statistics, Springer, 2014, pp. 15–18.

[DS04] Richard Durrett and Jason Schweinsberg, *Approximating selective sweeps*, Theoretical population biology **66** (2004), no. 2, 129–138.

[DTR+10] Eléonore Durand, Maud I Tenaillon, Céline Ridel, Denis Coubriche, Philippe Jamin, Sophie Jouanne, Adrienne Ressayre, Alain Charcosset, and Christineh Dillmann, *Standing variation and new mutations both contribute to a fast response to selection for flowering time in maize inbreds*, BMC evolutionary biology **10** (2010), no. 1, 2.

[Dur08] Richard Durrett, *Probability models for dna sequence evolution*, Springer, 2008.

[DVS11] Elena Dyakonova, Vladimir Vatutin, and Serik Sagitov, *Survival of branching processes in random environments*, arXiv preprint arXiv :1110.6139 (2011).

[Dyn91] Eugene B. Dynkin, *Branching particle systems and superprocesses*, Ann. Probab. **19** (1991), no. 3, 1157–1194.

[EFMS08] Anders Eriksson, Pontus Fernström, Bernhard Mehlig, and Serik Sagitov, *An accurate model for genetic hitchhiking*, Genetics **178** (2008), no. 1, 439–451.

[EK86] S.N. Ethier and T.G. Kurtz, *Markov processes : characterization and convergence*, Wiley, 1986.

[Épp79] MS Éppel', *A local limit theorem for first passage time*, Siberian Mathematical Journal **20** (1979), no. 1, 130–138.

[EPW06] Alison Etheridge, Peter Pfaffelhuber, and Anton Wakolbinger, *An approximate sampling formula under genetic hitchhiking*, The Annals of Applied Probability **16** (2006), no. 2, 685–729.

[Erl29] A.K. Erlang, *Opgave nr. 15.*, Mat. Tidsskr. B **36** (1929).

[Fel71] Willliam Feller, *An introduction to probability theory and its applications*, vol. 2, John Wiley & Sons, 1971.

[fig] *http ://mousse.lescigales.org/multiplication.php.*

[Fis30] R.A. Fisher, *Genetical theory of natural selection*, Osmania University Library, 1930.

[FL10] Zongfei Fu and Zenghu Li, *Stochastic equations of non-negative processes with jumps*, Stochastic Processes and their Applications **120** (2010), no. 3, 306–330.

[FM04] Nicolas Fournier and Sylvie Méléard, *A microscopic probabilistic description of a locally regulated population and macroscopic approximations*, The Annals of Applied Probability **14** (2004), no. 4, 1880–1919.

[FW05] Justin C Fay and Chung-I Wu, *Detecting hitchhiking from patterns of dna polymorphism*, Selective Sweep, Springer, 2005, pp. 65–77.

[Gal73] Francis Galton, *Problem 4001*, Educational Times **1** (1873), 17.

[GASK95] P Guttorp, K Albertsen, JF Steffensen, and E Kristensen, *Three papers on the history of branching processes*, International statistical review **63** (1995), no. 2, 233–245.

[Gil97] John H Gillespie, *Junk ain't what junk does : neutral alleles in a selected context*, Gene **205** (1997), no. 1, 291–299.

[GK00] Jochen Geiger and Götz Kersting, *The survival probability of a critical branching process in random environment*, Teor. Veroyatnost. i Primenen. **45** (2000), no. 3, 607–615.

[GKV03] Jochen Geiger, Götz Kersting, and Vladimir A. Vatutin, *Limit theorems for subcritical branching processes in random environment*, Ann. Inst. H. Poincaré Probab. Statist. **39** (2003), no. 4, 593–620.

[GL01] Yves Guivarc'h and Quansheng Liu, *Propriétés asymptotiques des processus de branchement en environnement aléatoire*, C. R. Acad. Sci. Paris Sér. I Math. **332** (2001), no. 4, 339–344.

[GMKM97] Stefan AH Geritz, Johan AJ Metz, Éva Kisdi, and Géza Meszéna, *Dynamics of adaptation and evolutionary branching*, Physical Review Letters **78** (1997), no. 10.

[Gre74] David R. Grey, *Asymptotic behaviour of continuous time, continuous state-space branching processes*, J. Appl. Probability **11** (1974), 669–677.

[Gri74] Anders Grimvall, *On the convergence of sequences of branching processes*, Ann. Probability **2** (1974), 1027–1045.

[Hir98] Katsuhiro Hirano, *Determination of the limiting coefficient for exponential functionals of random walks with positive drift*, J. Math. Sci. Univ. Tokyo **5** (1998), no. 2, 299–332.

[HP05] Joachim Hermisson and Pleuni S Pennings, *Soft sweeps molecular population genetics of adaptation from standing genetic variation*, Genetics **169** (2005), no. 4, 2335–2352.

[HP08] Joachim Hermisson and Peter Pfaffelhuber, *The pattern of genetic hitchhiking under recurrent mutation*, Electron J Probab **13** (2008), no. 68, 2069–2106.

[HS90] Josef Hofbauer and Karl Sigmund, *Adaptive dynamics and evolutionary stability*, Applied Mathematics Letters **3** (1990), no. 4, 75–79.

[Hut11] Martin Hutzenthaler, *Supercritical branching diffusions in random environment*, Electron. Commun. Probab. **16** (2011), no. 69, 781–791.

[IW89] Nobuyuki Ikeda and Shinzo Watanabe, *Stochastic differential equations and diffusion processes, 2nd ed.*, North-Holland, 1989.

[Jir58] Miloslav Jirina, *Stochastic branching processes with continuous state space*, Czechoslovak Math. J. **8 (83)** (1958), 292–313.

[JO00] *Journal officiel, comission générale de terminologie et de néologie*, Tech. report, 2000.

[K+74] Norman Kaplan et al., *A note on the supercritical branching processes with random environments*, The Annals of Probability **2** (1974), no. 3, 509–514.

[Kei75] Niels Keiding, *Extinction and exponential growth in random environments*, Theoretical Population Biology **8** (1975), no. 1, 49 – 63.

[Ken66] David G Kendall, *Branching processes since 1873*, Journal of the London Mathematical Society **1** (1966), no. 1, 385–406.

[KHL89] Norman L Kaplan, RR Hudson, and CH Langley, *The" hitchhiking effect" revisited.*, Genetics **123** (1989), no. 4, 887–899.

[KN04] Yuseob Kim and Rasmus Nielsen, *Linkage disequilibrium as a signature of selective sweeps*, Genetics **167** (2004), no. 3, 1513–1524.

[Koz76] Mikhail V. Kozlov, *On the asymptotic behavior of the probability of non-extinction for critical branching processes in a random environment*, Theory of Probability and Its Applications **21** (1976), 791–804.

[KP08] Andreas E Kyprianou and Juan-Carlos Pardo, *Continuous-state branching processes and self-similarity*, Journal of Applied Probability (2008), 1140–1160.

[KS02] Yuseob Kim and Wolfgang Stephan, *Detecting a local signature of genetic hitchhiking along a recombining chromosome*, Genetics **160** (2002), no. 2, 765–777.

[Kur78] Thomas G Kurtz, *Diffusion approximations for branching processes*, Branching processes (Conf., Saint Hippolyte, Que., 1976), vol. 5, 1978, pp. 269–292.

[Kyp06] Andreas E. Kyprianou, *Introductory lectures on fluctuations of Lévy processes with applications*, Universitext, Springer-Verlag, Berlin, 2006.

[Lam67a] John Lamperti, *Continuous state branching processes*, Bull. Amer. Math. Soc. **73** (1967), 382–386.

[Lam67b] ———, *The limit of a sequence of branching processes*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **7** (1967), 271–288.

[Lam07] Amaury Lambert, *Quasi-stationary distributions and the continuous-state branching process conditioned to be never extinct*, Electron. J. Probab. **12** (2007), no. 14, 420–446.

[Lam08] ———, *Population dynamics and random genealogies*, Stoch. Models **24** (2008), no. suppl. 1, 45–163.

[Lan93] R. Lande, *Risks of population extinction from demographic and environmental stochasticity and random catastrophes*, American Naturalist (1993), 911–927.

[LBD11] Gregory I Lang, David Botstein, and Michael M Desai, *Genetic variation and the fate of beneficial mutations in asexual populations*, Genetics **188** (2011), no. 3, 647–661.

[LD00] Richard Law and Ulf Dieckmann, *Moment approximations of individual-based models*, The geometry of ecological interactions : simplifying spatial complexity (2000), 252–270.

[Leo09] Stephanie Leocard, *Selective sweep and the size of the hitchhiking set*, Advances in Applied Probability **41** (2009), no. 3, 731–764.

[LES03] Russell Lande, Steinar Engen, and Bernt-Erik Saether, *Stochastic population dynamics in ecology and conservation*, Oxford University Press, 2003.

[Li10] Zenghu Li, *Measure-valued branching markov processes*, Springer, 2010.

[LPP97] Émile Le Page and Marc Peigné, *A local limit theorem on the semi-direct product of R^{*+} and R^d*, Ann. Inst. H. Poincaré Probab. Statist. **33** (1997), no. 2, 223–252.

[LRH+13] Gregory I Lang, Daniel P Rice, Mark J Hickman, Erica Sodergren, George M Weinstock, David Botstein, and Michael M Desai, *Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations*, Nature **500** (2013), no. 7464, 571–574.

[McV07] Gil McVean, *The structure of linkage disequilibrium around a selective sweep*, Genetics **175** (2007), no. 3, 1395–1406.

[mei] *http ://bio1151b.nicerweb.net/locked/media/ch13/*.

[MGM+96] Johan AJ Metz, Stefan AH Geritz, Géza Meszéna, Frans JA Jacobs, and JS Van Heerwaarden, *Adaptive dynamics, a geometrical study of the consequences of nearly faithful reproduction*, Stochastic and spatial structures of dynamical systems **45** (1996), 183–231.

[MP13] Philipp W Messer and Dmitri A Petrov, *Population genomics of rapid adaptation by soft selective sweeps*, Trends in ecology & evolution **28** (2013), no. 11, 659–669.

[MV12] Sylvie Méléard and Denis Villemonais, *Quasi-stationary distributions and population processes*, Probability surveys **9** (2012), 340–410.

[NHH+07] Rasmus Nielsen, Ines Hellmann, Melissa Hubisz, Carlos Bustamante, and Andrew G Clark, *Recent and ongoing selection in the human genome*, Nature Reviews Genetics **8** (2007), no. 11, 857–868.

[NK97] Claudia Neuhauser and Stephen M Krone, *The genealogy of samples in models with selection*, Genetics **145** (1997), no. 2, 519–534.

[OB01] H Allen Orr and Andrea J Betancourt, *Haldane's sieve and adaptation from the standing genetic variation*, Genetics **157** (2001), no. 2, 875–884.

[OK75] Tomoko Ohta and Motoo Kimura, *The effect of selected linked locus on heterozygosity of neutral alleles (the hitch-hiking effect)*, Genetical research **25** (1975), no. 03, 313–325.

[PCW05] Molly Prezeworski, Graham Coop, and Jeffrey D Wall, *The signature of positive selection on standing genetic variation*, Evolution **59** (2005), no. 11, 2312–2323.

[PH06a] Pleuni S Pennings and Joachim Hermisson, *Soft sweeps ii-molecular population genetics of adaptation from recurrent mutation or migration*, Molecular biology and evolution **23** (2006), no. 5, 1076–1084.

[PH06b] _____, *Soft sweeps iii : the signature of positive selection from recurrent mutation*, PLoS genetics **2** (2006), no. 12, e186.

[PHSN12] Benjamin M Peter, Emilia Huerta-Sanchez, and Rasmus Nielsen, *Distinguishing between selective sweeps from standing variation and from a de novo mutation*, PLoS genetics **8** (2012), no. 10.

[PMH01] John Parsch, Colin D Meiklejohn, and Daniel L Hartl, *Patterns of dna sequence variation suggest the recent action of positive selection in the janus-ocnus region of drosophila simulans*, Genetics **159** (2001), no. 2, 647–657.

[Prz02] Molly Przeworski, *The signature of positive selection at randomly chosen loci*, Genetics **160** (2002), no. 3, 1179–1189.

[PS07] P Pfaffelhuber and A Studeny, *Approximating genealogies for partially linked neutral loci under a selective sweep*, Journal of mathematical biology **55** (2007), no. 3, 299–330.

[RLF+13] Nicolas O Rode, Eva JP Lievens, Elodie Flaven, Adeline Segard, Roula Jabbour-Zahab, Marta I Sanchez, and Thomas Lenormand, *Why join groups ? lessons from parasite-manipulated artemia*, Ecology letters (2013).

[SD05] J. Schweinsberg and R. Durrett, *Random partitions approximating the coalescence of lineages during a selective sweep*, The Annals of Applied Probability **15** (2005), no. 3, 1591–1651.

[SH74] J Maynard Smith and John Haigh, *The hitch-hiking effect of a favourable gene*, Genet Res **23** (1974), no. 1, 23–35.

[Sil68] Martin L. Silverstein, *A new approach to local times*, J. Math. Mech. **17** (1967/1968), 1023–1054.

[Sla08] Montgomery Slatkin, *Linkage disequilibrium-understanding the evolutionary past and mapping the medical future*, Nature Reviews Genetics **9** (2008), no. 6, 477–485.

[Sma14] Charline Smadi, *An eco-evolutionary approach of adaptation and recombination in a large population of varying size*, arXiv preprint arXiv :1402.4104 (2014).

[SRH+02] Pardis C Sabeti, David E Reich, John M Higgins, Haninah ZP Levine, Daniel J Richter, Stephen F Schaffner, Stacey B Gabriel, Jill V Platko, Nick J Patterson, Gavin J McDonald, et al., *Detecting recent positive selection in the human genome from haplotype structure*, Nature **419** (2002), no. 6909, 832–837.

[SSF+06] PC Sabeti, SF Schaffner, B Fry, J Lohmueller, P Varilly, O Shamovsky, A Palma, TS Mikkelsen, D Altshuler, and ES Lander, *Positive natural selection in the human lineage*, science **312** (2006), no. 5780, 1614–1620.

[SSL06] Wolfgang Stephan, Yun S Song, and Charles H Langley, *The hitchhiking effect on linkage disequilibrium between linked neutral loci*, Genetics **172** (2006), no. 4, 2647–2663.

[SW69] Walter L Smith and William E Wilkinson, *On branching processes in random environments*, The Annals of Mathematical Statistics (1969), 814–827.

[SWL92] Wolfgang Stephan, Thomas HE Wiehe, and Marcus W Lenz, *The effect of strongly selected substitutions on neutral polymorphism : analytical results based on diffusion theory*, Theoretical Population Biology **41** (1992), no. 2, 237–254.

[TK87] Glenys Thomson and William Klitz, *Disequilibrium pattern analysis. i. theory*, Genetics **116** (1987), no. 4, 623–632.

[Wal75] Bruce Wallace, *Hard and soft selection revisited*, Evolution (1975), 465–473.

[WFF+02] John C Wootton, Xiaorong Feng, Michael T Ferdig, Roland A Cooper, Jianbing Mu, Dror I Baruch, Alan J Magill, and Xin-zhuan Su, *Genetic diversity and chloroquine selective sweeps in plasmodium falciparum*, Nature **418** (2002), no. 6895, 320–323.

[WG75] Henry William Watson and Francis Galton, *On the probability of the extinction of families*, The Journal of the Anthropological Institute of Great Britain and Ireland **4** (1875), 138–144.

Résumé. Cette thèse porte sur l'étude probabiliste des réponses démographique et génétique de populations à certains événements ponctuels. Dans une première partie, nous étudions l'impact de catastrophes tuant une fraction de la population et survenant de manière répétée, sur le comportement en temps long d'une population modélisée par un processus de branchement. Dans un premier temps nous construisons une nouvelle classe de processus, les processus de branchement à états continus avec catastrophes, en les réalisant comme l'unique solution forte d'une équation différentielle stochastique. Nous déterminons ensuite les conditions d'extinction de la population. Enfin, dans les cas d'absorption presque sûre nous calculons la vitesse d'absorption asymptotique du processus. Ce dernier résultat a une application directe à la détermination du nombre de cellules infectées dans un modèle d'infection de cellules par des parasites. En effet, la quantité de parasites dans une lignée cellulaire suit dans ce modèle un processus de branchement, et les "catastrophes" surviennent lorsque la quantité de parasites est partagée entre les deux cellules filles lors des divisions cellulaires. Dans une seconde partie, nous nous intéressons à la signature génétique laissée par un balayage sélectif. Le matériel génétique d'un individu détermine (pour une grande partie) son phénotype et en particulier certains traits quantitatifs comme les taux de naissance et de mort intrinsèque, ou sa capacité d'interaction avec les autres individus. Mais son génotype seul ne détermine pas son "adaptation" dans le milieu dans lequel il vit : l'espérance de vie d'un humain par exemple est très dépendante de l'environnement dans lequel il vit (accès à l'eau potable, à des infrastructures médicales,...). L'approche éco-évolutive cherche à prendre en compte l'environnement en modélisant les interactions entre les individus. Lorsqu'une mutation ou une modification de l'environnement survient, des allèles peuvent envahir la population au détriment des autres allèles : c'est le phénomène de balayage sélectif. Ces événements évolutifs laissent des traces dans la diversité neutre au voisinage du locus auquel l'allèle s'est fixé. En effet ce dernier "emmène" avec lui des allèles qui se trouvent sur les loci physiquement liés au locus sous sélection. La seule possibilité pour un locus de ne pas être "emmené" est l'occurence d'une recombination génétique, qui l'associe à un autre haplotype dans la population. Nous quantifions la signature laissée par un tel balayage sélectif sur la diversité neutre. Nous nous concentrons dans un premier temps sur la variation des proportions neutres dans les loci voisins du locus sous sélection sous différents scénarios de balayages. Nous montrons que ces différents scenari évolutifs laissent des traces bien distinctes sur la diversité neutre, qui peuvent permettre de les discriminer. Dans un deuxième temps, nous nous intéressons aux généalogies jointes de deux loci neutres au voisinage du locus sous sélection. Cela nous permet en particulier de quantifier des statistiques attendues sous certains scenari de sélection, qui sont utilisées à l'heure actuelle pour détecter des événements de sélection dans l'histoire évolutive de populations à partir de données génétiques actuelles. Dans ces travaux, la population évolue suivant un processus de naissance et mort multitype avec compétition. Si un tel modèle est plus réaliste que les processus de branchement, la non-linéarité introduite par les compétitions entre individus en rend l'étude plus complexe.

Abstract. This thesis is devoted to the probabilistic study of demographic and genetical responses of a population to some pointwise events. In a first part, we are interested in the effect of random catastrophes, which kill a fraction of the population and occur repeatedly, in populations modeled by branching processes. First we construct a new class of processes, the continuous state branching processes with catastrophes, as the unique strong solution of a stochastic differential equation. Then we describe the conditions for the population extinction. Finally, in the case of almost sure absorption, we state the asymptotical rate of absorption. This last result has a direct application to the determination of the number of infected cells in a model of cell infection by parasites. Indeed, the parasite population size in a lineage follows in this model a branching process, and catastrophes correspond to the sharing of the parasites between the two daughter cells when a division occurs. In a second part, we focus on the genetic signature of selective sweeps. The genetic material of an individual (mostly) determines its phenotype and in particular some quantitative traits, as birth and intrinsic death rates, and interactions with others individuals. But genotype is not sufficient to determine "adaptation" in a given environment : for example the life expectancy of a human being is very dependent on his environment (access to drinking water, to medical infrastructures,...). The eco-evolutive approach aims at taking into account the environment by modeling interactions between individuals. When a mutation or an environmental modification occurs, some alleles can invade the population to the detriment of other alleles : this phenomenon is called a selective sweep and leaves signatures in the neutral diversity in the vicinity of the locus where the allele fixates. Indeed, this latter "hitchhikes" alleles situated on loci linked to the selected locus. The only possibility for an allele to escape this "hitchhiking" is the occurrence of a genetical recombination, which associates it to another haplotype in the population. We quantify the signature left by such a selective sweep on the neutral diversity. We first focus on neutral proportion variation in loci partially linked with the selected locus, under different scenari of selective sweeps. We prove that these different scenari leave distinct signatures on neutral diversity, which can allow to discriminate them. Then we focus on the linked genealogies of two neutral alleles situated in the vicinity of the selected locus. In particular, we quantify some statistics under different scenari of selective sweeps, which are currently used to detect recent selective events in current population genetic data. In these works the population evolves as a multitype birth and death process with competition. If such a model is more realistic than branching processes, the non-linearity caused by competitions makes its study more complex.