Sur quelques fonctionnelles des forêts de branchement multitypes

Thi Ngoc Anh Nguyen

UNIVERSITE BRETAGNE LOIRE

université angers

Thèse de Doctorat

Thi Ngoc Anh NGUYEN

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université d'Angers
sous le sceau de l'Université Bretagne Loire

École doctorale : Sciences et technologies de l'information, et mathématiques

Discipline : Mathématiques et leurs interactions, section CNU 25
Unité de recherche : Laboratoire Angevin de Recherche en Mathématiques (LAREMA)

Soutenue le 15 Juillet 2016
Thèse n° : 77738

Sur quelques fontionnelles des forêts de branchement multitypes

JURY

Examinateurs :	**M. Piotr GRACZYK**, Professeur des universités, Université d'Angers
	M. Rodolphe GARBIT, Maître de conférences, Université d'Angers
	M^me^ Camille CORON, Maître de conférences, Université Paris Sud
Invitée :	**M^me^ Bénédicte HAAS**, Professeur des universités, Université Paris 13
Directeur de thèse :	**M. Loïc CHAUMONT**, Professeur des universités, Université d'Angers

* * *

Rapporteurs :	**M. Jean-François DELMAS**, Professeur des universités, Ecole des Ponts Paris Tech
	M. Götz KERSTING, Professeur, Université de Francfort

Université d'Angers – LAREMA – UMR CNRS 6093

Thèse de Doctorat

Spécialité :

Mathématiques

Sur quelques fonctionnelles des forêts de branchement multitypes

Présentée par :

Thi Ngoc Anh NGUYEN

Table des matières

Remerciements **5**

Avant-propos **6**

1 Introduction **8**

1.1 Notations 12

1.2 Processus de Bienaymé-Galton-Watson multitype 13

1.3 Résultats préliminaires 15

1.3.1 La loi de l'effectif total d'une forêt de branchement multitype . . 15

1.3.2 Représentation de type Lamperti en dimension supérieure 17

1.4 Résumé des travaux 19

1.4.1 Sur les mutations dans le modèle de branchement pour des populations multitypes 19

1.4.2 Suites cycliquement échangeables et énumérations de forêts multitypes 24

1.4.3 Une note sur les sommets ayant un degré donné dans une forêt de branchement multitype 25

1.5 Perspectives 26

2 On mutations in the branching model for multitype populations **27**

2.1 Introduction 27

2.2 Mutations and their asymptotics in discrete multitype forests 28

2.2.1 Preliminaries on discrete multitype forests 28

2.2.2 The total number of mutations and its asymptotics 33
2.3 When continuous time is involved 42
2.3.1 The Lamperti representation 42
2.3.2 Further results on asymptotics of mutations 45
2.3.3 Emergence times of mutations 48

3 Cyclically exchangeable sequences and enumeration of multitype forests 57
3.1 Introduction 57
3.2 The multivariate cyclic lemma 59
3.3 Enumeration of multitype plane forests 61
3.4 Enumeration of multitype labeled forests 65
3.5 The Lagrange-Good inversion formula 69

4 A note on vertices with a given degree in multitype branching forests 73
4.1 Introduction 73
4.2 Preliminaires 74
4.2.1 General notation 74
4.2.2 Multitype branching forest 74
4.3 Multivariate Ballot Theorem 77
4.4 Law of the total number of leaves 78
4.5 Results in some more general scenarios 85
4.5.1 Law of the total number of vertices with a given degree 85
4.5.2 Law of the total number of vertices whose degree is in a given set 86

Bibliographie 88

Remerciements

Je tiens à remercier mon directeur de thèse Loïc Chaumont pour son encadrement, toute sa patience et ses encouragements durant toute ma thèse.

Je tiens à remercier Jean-François Delmas et Götz Kersting qui ont accepté de rapporter cette thèse. Je tiens à remercier également les autres membres du jury : Camille Coron, Piotr Graczyk, et Rodolphe Garbit.

Mes remerciements vont également à tous mes amis au LAREMA, en particulier Delphine Pol, Mohamed Benzerga et Nguyen Le Chi Quyet avec qui j'ai partagé des bons moments.

Avant-propos

Cette thèse est principalement consacrée à l'étude de quelques caractéristiques d'une population à plusieurs types d'individus qui évolue selon un modèle de branchement multitype au cours du temps. Autrement dit, chaque individu vit un certain temps et donne naissance, à la fin de sa vie, à un nombre aléatoire d'individus, suivant une loi de probabilité qui ne dépend que de son type, indépendamment des autres individus. Plus précisément, nous nous intéressons aux aspects statistiques des mutations et des individus ayant une progéniture donnée dans la population en question.

Les problèmes d'énumération de forêts multitypes constituent également une motivation de ce travail de thèse. Des méthodes stochastiques et combinatoires sont appliquées au calcul du nombre de forêts multitypes dont les sommets sont marqués ou non.

La thèse comporte, après une introduction, trois chapitres qui sont issus des trois travaux suivants :

— [CN15] Loïc Chaumont et Thi Ngoc Anh Nguyen : On mutations in the branching model for multitype populations. 2015. arXiv:1510.00845.

— [CLN16]Loïc Chaumont, Rongli Liu et Thi Ngoc Anh Nguyen : Cyclically exchangeable sequences and enumeration of multitype forests. 2016. *À soumettre.*

— [Ngu16] Thi Ngoc Anh Nguyen : A note on vertices with a given degree in multi-

type branching forests. 2016. *À soumettre*.

Chapitre 1

Introduction

L'histoire des processus de branchement remonte au dix-neuvième siècle avec un problème concernant l'extinction des noms de famille en Grande-Bretagne, introduit, formulé et communiqué par Francis Galton à l'*Educational Times* en 1873 [Gal73] :

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,* a_0 *per cent of the adult males have no male children who reach adult life ;* a_1 *have one such male child ;* a_2 *have two ; and so on up to* a_5 *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.*

Ce problème a été traité en premier par Watson de manière incorrecte et ensuite par plusieurs autres mathématiciens durant plusieurs années. Nous renvoyons à [Mod71] pour une introduction plus détaillée à ce sujet. Soulignons également que ce problème avait déjà été soulevé auparavant en 1845 par Bienaymé, un probabiliste et statisticien français, dont le travail était passé inaperçu.

Depuis leur apparition, les processus de branchement se sont révélés être un outil mathématique important dans plusieurs domaines scientifiques comme la biologie, la physique, la chimie, l'épidémiologie et la généalogie. Aujourd'hui ils sont de plus en plus uti-

lisés pour modéliser l'évolution d'une population dont les individus (cellules, molécules, *etc.*) vivent, meurent et donnent naissance indépendamment les uns des autres. L'hypothèse que les individus se multiplient indépendamment les uns des autres est indispensable pour simplifier l'étude des processus de branchement associés. Bien qu'elle apparaisse assez restrictive, celle-ci peut toujours être justifiée dans de nombreux cas, comme pour des populations de molécules ou de bactéries dont la reproduction est asexuée.

Lorsque la population est constituée d'individus de plusieurs types, ce qui est le cas le plus fréquent dans la réalité, une extension du modèle de branchement classique doit être mise en place. Chaque individu de chaque type évolue selon les mêmes hypothèses de branchement que celles décrites précédemment, mais avec une loi propre à son type. Comme attendu, un modèle de branchement multitype est une description de l'évolution d'une population constituée d'un nombre fini de types d'individus ; plus précisément, un tel modèle décrit le nombre d'individus de chaque type à chaque génération dans la population.

Cette thèse s'intéresse principalement à quelques caractéristiques de la dynamique d'une population qui se développe suivant un modèle de branchement multitype. En particulier, nous étudierons l'évolution dans le temps des mutations d'une part et des individus ayant une progéniture donnée d'autre part, dans une population multitype. Outre leur intérêt purement mathématique, ces objets présentent de nombreuses applications en biologie, génétique, médecine, etc.

Il faudrait noter que l'information contenue dans un processus de branchement multitype ne suffit pas à elle seule à décrire certaines fontionnelles telles que les mutations ou les individus ayant une certaine progéniture donnée dans la population sous-jacente. Nous avons donc besoin de toute l'information apportée par *forêt de branchement multitype* qui représente essentiellement la généalogie d'une population à plusieurs types d'individus. Nous en donnerons une définition plus précise dans la section 1.3.1. Grâce à l'étude des forêts de branchement multitypes, nous pouvons obtenir des résultats sur la statistique des mutations ou des individus ayant une certaine progéniture donnée dans

une population à plusieurs types d'individus.

Dans la première partie de ce chapitre, nous rappellons quelques notions de base sur les processus de Bienaymé-Galton-Watson multitypes en temps discret ainsi qu'en temps continu. Nous renvoyons aux ouvrages [AN72], [Har52] et [Mod71] pour plus de détails.

La deuxième partie de ce chapitre est consacrée à la présentation de deux résultats récents importants sur lesquels s'appuie la thèse. Le premier porte sur la loi de l'effectif total d'une population multitype ; il est issu du papier [CL15]. Le deuxième est la représentation de type Lamperti en dimension supérieure d'un processus de branchement multitype donnée dans [Cha15].

En troisième lieu, on résume les travaux de la thèse.

Notre premier travail est consacré aux mutations dans une population qui consiste en un nombre fini de types d'individus. On montre d'abord que la forêt des mutations associée à une forêt de branchement multitype est elle-même une forêt de branchement dont la loi de reproduction peut être calculée explicitement en fonction de la loi de reproduction de la population. Ensuite, on déduit, à partir de ce résultat, la loi du nombre de mutations et son comportement asymptotique lorsque l'effectif total ou le nombre initial d'individus tend vers l'infini. Lorsque le temps est continu, grâce à la représentation de type Lamperti en dimension supérieure, on obtient la limite du nombre de mutations en temps infini dans le cas surcritique. Le temps d'émergence d'une mutation est également étudié sous quelques conditions plus restrictives mais qui restent raisonnables cependant. Il existe plusieurs résultats dans la littérature autour de ce temps d'émergence, par exemple dans Iwasa et al. [Ale13], Serra [Ser06], Durrett [DM10], ou bien plus récemment Alexander [Ale13]. Mais le problème de déterminer sa loi dans une situation générale est toujours ouvert. Dans notre travail, nous tentons de décrire ce temps dans un autre esprit et par une nouvelle approche qui est assez différente de celles des auteurs cités ci-dessus. Cette approche est basée sur la représentation de type Lamperti en dimension supérieure d'un processus de branchement multitype.

Notre deuxième travail est consacré à l'étude des problèmes d'énumération de forêts multitypes. Nous obtenons, comme conséquences du théorème de ballotage obtenu récemment dans [CL15], quelques formules d'énumération des forêts multitypes. Nous retrouvons également quelques résultats récents de dénombrement de forêts planaires marquées, comme par exemple le nombre de telles forêts dont les degrés (multidimensionels) des sommets sont fixés. Une nouvelle preuve de la formule d'inversion (multivariée) de Lagrange-Good est également donnée.

Notre troisième travail s'intéresse aux individus ayant une progéniture donnée dans une population multitype. Ceux-ci correspondent à des sommets ayant un degré donné de la forêt de branchement multitype représentant la généalogie de la population. En particulier, la loi du nombre de feuilles d'une forêt de branchement (monotype) et son comportement asymptotique ont été étudiés par Korchemski dans la première partie de [Kor12]. Notre but est d'étendre ces études au cas multitype.

Finalement, nous présentons quelques perspectives sur notre travail.

1.1 Notations

On utilisera dans la suite les notations suivantes :

- $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$
- $\mathbb{N} = \{1, 2, \ldots\}$
- $\mathbb{R}_+ = [0, +\infty)$
- $[d] := \{1, \ldots, d\}$ pour tout $d \in \mathbb{N}$
- $\mathbf{0} = (0, \ldots, 0),\ \mathbf{1} = (1, \ldots, 1)$
- $\mathbf{e}_i = (0, \ldots, 1, \ldots, 0)$ est le i-ième vecteur de la base canonique de $\mathbb{R}^d$.

Ensuite, pour tout $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^d$, on utilisera les notations et conventions suivantes :

- $\mathbf{x}^{\mathbf{y}} = \prod_{i=1}^d x_i^{y_i}$
- $\mathbf{x} \leq \mathbf{y}$ si, et seulement si $x_i \leq y_i, \forall i \in [d]$
- $\mathbf{x} \cdot \mathbf{y}$ est le produit scalaire dans $\mathbb{R}_+^d$.

Finalement, un indice dans $\mathbb{P}$ ou $\mathbb{E}$ signifie l'état initial du processus associé.

1.2 Processus de Bienaymé-Galton-Watson multitype

Les processus de branchement multitypes les plus simples sont les processus de Bienaymé-Galton-Watson multitypes dont le temps et l'espace d'états sont discrets. Rappellons d'abord la définition suivante qui se trouve dans [AN72] :

Définition 1.2.1 : Un *processus de Bienaymé-Galton-Watson à d types* est une chaîne de Markov $\mathbf{Z} = (\mathbf{Z}_n)_{n\geq 0} = (\mathbf{Z}_n^{(1)}, \ldots, \mathbf{Z}_n^{(d)})_{n\geq 0}$ à valeur dans $\mathbb{Z}_+^d$ qui possède la propriété dite *de branchement* suivante :

$$\mathbb{E}_{\mathbf{i}}\mathbf{s}^{\mathbf{Z}_1} = \prod_{k=1}^{d} (\mathbb{E}_{\mathbf{e}_k}\mathbf{s}^{\mathbf{Z}_1})^{i_k},\ \forall \mathbf{i} = (i_1, \ldots, i_d) \in \mathbb{N}^d, \forall \mathbf{s} \in \mathbb{R}_+^d, \mathbf{s} \leq \mathbf{1}.$$

Intuitivement, la propriété de branchement signifie que le processus $\mathbf{Z}$ partant de $(z_1, \ldots, z_d)$ est la somme de $(z_1 + \cdots + z_d)$ processus indépendants dont z_i d'entre eux partent de $\mathbf{e}_i,\ i \in [d]$.

Un processus de Bienaymé-Galton-Watson à d types $\mathbf{Z}$ peut être vu comme une description de l'évolution d'une population multitype dont chaque individu vit une unité de temps et en fin de vie, reproduit, indépendamment des uns des autres, un nombre aléatoire d'enfants, suivant une certaine loi de probabilité qui ne dépend que du type de l'individu, appelée *loi de reproduction*. À ce titre, $\mathbf{Z}_n^{(i)}$ représente le nombre d'individus de type i à la génération n. Notons $\nu = (\nu_1, \ldots, \nu_d)$, où $(\nu_i,\ i \in [d])$ sont des mesures de probabilités sur $\mathbb{Z}_+^d$ et ν_i est la loi de reproduction d'un individu de type i. On appelle ν la *loi de reproduction du processus*. Les fonctions de transition du processus $\mathbf{Z}$ sont ensuite déterminées entièrement à partir de sa loi de reproduction d'après la relation

$$\mathbb{P}\left(\mathbf{Z}_{n+1} = (j_1, \ldots, j_d) | \mathbf{Z}_n = (i_1, \ldots, i_d)\right) = \nu_1^{*i_1} * \cdots * \nu_d^{*i_d}(j_1, \ldots, j_d)$$

On appelle *matrice de moyenne* la matrice $\mathbf{M} = (m_{ij}; i, j \in [d])$, où $m_{ij} = \sum_{\mathbf{k}\in\mathbb{Z}_+^d} k_j \nu_i(\mathbf{k})$. Pour simplifier, on supposera toujours que $\mathbf{M}$ est *primitive* (au sens où $\forall i, j \in [d],\ m_{ij} < \infty$ et il existe un entier n tel que $\mathbf{M}^n > 0$). On supposera aussi que $\nu_i(\mathbf{e}_i) = 0,\ \forall i \in [d]$ et que ν est *non-singulier*, au sens où il existe un $i \in [d]$ tel que $\sum_{k=1}^{d} \nu_i(\mathbf{e}_k) < 1$, c'est-à-dire

que l'on exclura le cas où chaque individu donne naissance à exactement un individu presque sûrement.

D'après le théorème de Perron-Frobenius, il existe une valeur propre maximale ρ de la matrice $\mathbf{M}$. Les vecteurs propres à gauche et à droite associés à ρ, normalisés par les relations $\mathbf{u} \cdot \mathbf{1} = 1, \mathbf{u} \cdot \mathbf{v} = 1$, seront notés respectivement $\mathbf{v}$ et $\mathbf{u}$. Lorsque $\rho \leq 1$, la population s'éteindra presque sûrement, tandis que si $\rho > 1$, avec une probabilité positive, elle ne s'éteindra jamais. On dit que le processus ou bien la loi de reproduction ν est *sous-critique*, *critique* ou *surcritique* si $\rho < 1, \rho = 1$ or $\rho > 1$, respectivement.

Lorsque les temps de vie des individus d'une population sont aléatoires, le modèle du processus de Bienaymé-Galton-Watson multitype en temps continu nous permettra de décrire de manière plus précise l'évolution de cette population. Néanmoins, les temps de vie des individus doivent suivre une loi exponentielle afin que le processus de branchement associé soit Markovien. Signalons que le fait qu'un individu meure en donnant naissance à un seul individu de même type n'est pas considéré comme un événement, c'est-à-dire qu'il est considéré comme encore vivant dans ce cas. La définition formelle de ces processus est la suivante [AN72, section V.7] :

Définition 1.2.2 : Un *processus de Bienaymé-Galton-Watson à d types en temps continu* est un processus de Markov $\mathbf{Z} = (\mathbf{Z}_t)_{t\geq 0} = (Z_t^{(1)}, \ldots, Z_t^{(d)})_{t\geq 0}$, càdlàg, à valeurs dans $\mathbb{Z}_+^d$ qui possède la propriété dite *de branchement* suivante :

$$\mathbb{E}_{\mathbf{i}}\mathbf{s}^{\mathbf{Z}_t} = \prod_{k=1}^{d} (\mathbb{E}_{\mathbf{e}_k}\mathbf{s}^{\mathbf{Z}_t})^{i_k},$$

$\forall t \geq 0,\ \forall \mathbf{i} \in \mathbb{N}^d$ et $\mathbf{s} \in \mathbb{R}_+^d, \mathbf{s} \leq \mathbf{1}$.

La propriété de branchement en temps continu a le même sens qu'en temps discret et $\mathbf{Z}_t^{(i)}$ représente également le nombre d'individus de type i au temps t. Néanmoins, les fonctions de transition du processus $\mathbf{Z} = (\mathbf{Z}_t)_{t\geq 0} = (Z_t^{(1)}, \ldots, Z_t^{(d)})_{t\geq 0}$ sont cette fois-ci déterminées par sa loi de reproduction $\nu = (\nu_1, \ldots, \nu_d)$ ainsi que par les paramètres des temps de vie des individus qui se distinguent par type ; ces paramètres seront notés $\lambda_i > 0, i \in [d]$. Comme dans le cas discret, on supposera que $\nu_i(\mathbf{e}_i) = 0,\ \forall i \in [d]$.

La matrice de moyenne de la loi de reproduction ν, qui a été définie dans le cas discret, sera toujours notée $\mathbf{M}$. De plus, il existe des matrices de moyenne qui dépendent également du temps t, données par

$$\mathbf{M}(t) := (m_{ij}(t); i, j = 1, \dots, d),$$

où $m_{ij}(t) = \mathbb{E}_{\mathbf{e}_i} Z_t^{(j)}$. Comme dans le cas discret, on fait toujours l'hypothèse de primitivité des matrices de moyenne, c'est-à-dire qu'il existe un nombre $0 < t < \infty$ tel que $m_{ij}(t) > 0, \forall i, j$.

La propriété de branchement nous permet d'assurer l'existence d'une matrice A telle que $\mathbf{M}(t) = e^{tA}$. Notons que A est primitive. Le théorème de Perron-Frobenius appliqué à la matrice A nous donne sa valeur propre maximale notée ρ_1.

Lorsque $\rho_1 \leq 0$, la population s'éteindra presque sûrement, tandis que si $\rho_1 > 0$, avec une probabilité positive, elle ne s'éteindra jamais. On dit alors que le processus ou bien sa loi de reproduction ν est *sous-critique*, *critique* ou *surcritique* si $\rho_1 < 0, \rho_1 = 0$ ou $\rho_1 > 0$, respectivement.

1.3 Résultats préliminaires

Nous présentons dans la suite les résultats antérieurs sur lesquels s'est appuyée la thèse. Ces résultats sont basés sur la notion de *forêt de branchement multitype (avec longueurs d'arc)* qui nous permet de prendre en compte plus d'informations sur l'évolution de la population modélisée que les processus de branchement eux-mêmes. Ces résultats sont obtenus grâce à certains types de codage d'une forêt de branchement multitype.

1.3.1 La loi de l'effectif total d'une forêt de branchement multitype

Nous rappelons d'abord la notion de *forêt de branchement multitype* en suivant le formalisme introduit dans [CL15].

Commençons par définir les *forêts (planaires) multitypes*. Une *forêt (planaire)* $\mathbf{f}$ est un graphe planaire (c'est-à-dire, un graphe qu'on peut dessiner dans le plan sans que ses arêtes ne se touchent, sauf à leurs extrémités), orienté, sans boucle, sur un ensemble non vide de sommets, tel que chaque sommet ait un degré entrant fini et un degré sortant égal à 0 ou 1. Une composante connexe d'une forêt est appelée *arbre (planaire)*. Dans un arbre, il n'y a qu'un seul sommet dont le degré sortant est 0, ce sommet est appelé la *racine* de cet arbre. Les racines des arbres dans $\mathbf{f}$ sont aussi appelées racines de $\mathbf{f}$. Soient u et v deux sommets de $\mathbf{f}$. Si (u, v) est un arc de $\mathbf{t}$ alors on dit que v est le *parent* de u ou bien que u est un *enfant* de v. La *génération* d'un sommet u est définie comme la distance usuelle entre ce sommet et la racine.

À chaque forêt $\mathbf{f}$, on associe une application $c_{\mathbf{f}}$ allant de l'ensemble de ses sommets $\mathbf{v}(\mathbf{f})$ à un ensemble $[d],\ d$ étant un nombre entier supérieur ou égal à 2, telle que $c_{\mathbf{f}}(u) \leq c_{\mathbf{f}}(v)$ pour tous u, v ayant le même parent avec u à gauche de v dans $\mathbf{f}$. On appelle $c_{\mathbf{f}}(u)$ le *type* du sommet u. Le couple $(\mathbf{f}, c_{\mathbf{f}})$ est appelé une *forêt à d types* et sera également notée $\mathbf{f}$ lorsqu'il n'y a pas d'ambiguïté. L'ensemble des forêts à d types est noté $\mathcal{F}_d$.

Une forêt de branchement à d types finie F, de loi de reproduction $\nu = (\nu_1, \ldots, \nu_d)$, est une variable aléatoire à valeurs dans l'ensemble des forêts à d types $\mathcal{F}_d$ telle que

$$\forall \mathbf{f} \in \mathcal{F}_d,\ \mathbb{P}(F = \mathbf{f}) = \prod_{u \in \mathbf{v}(\mathbf{f})} \nu_{c_{\mathbf{f}}(u)}(p_1(u), \ldots, p_d(u)),$$

où pour chaque $u \in \mathbf{v}(\mathbf{f}),\ p_i(u)$ désigne le nombre d'enfants de type i de u.

Étant donnée une forêt de branchement à d types $F,\ d > 1$, de loi de reproduction $\nu = (\nu_1, \ldots, \nu_d)$ sous-critique ou critique, partant de l'état $\mathbf{x}$, on note $\mathbf{Z} = (\mathbf{Z}_n)_{n \geq 0} = (\mathbf{Z}_n^{(1)}, \ldots, \mathbf{Z}_n^{(d)})_{n \geq 0}$ le processus de Bienaymé-Galton-Watson correspondant au sens où pour $i \in [d]$, $\mathbf{Z}_n^{(i)}$ est le nombre d'individus de type i à la génération n. Pour chaque $i \in [d]$, soit N_i l'effectif total du type i, c'est-à-dire, $N_i = \sum_{n \geq 0} \mathbf{Z}_n^{(i)}$. Soit M_{ij} le nombre total d'individus de type j dont le parent est de type $i,\ i \neq j$.

D'après [CL15, Théorème 3.1], toute forêt de branchement à d types critique ou sous-critique dont la loi de reproduction est primitive (c'est-à-dire telle que la matrice de

moyenne soit primitive) est codée par d marches aléatoires indépendantes d dimensionnelles, au travers d'une bijection entre l'ensemble des forêts de branchement à d types et un ensemble particulier de marches aléatoires multidimensionnelles. Plus précisément, la forêt F ci-dessus est codée par d marches aléatoires indépendantes d dimensionnelles

$$X^{(i)} = (X^{i,1}, \dots, X^{i,d}),\ i \in [d]$$

dont la loi de saut $\tilde{\nu_i}$ est définie par

$$\tilde{\nu_i}(k_1, \dots, k_d) := \nu_i(k_1, \dots, k_{i-1}, k_i + 1, k_{i+1}, \dots, k_d)\,, \quad \text{pour tout } (k_1, \dots, k_d) \in \mathbb{Z}_+^d.$$

À partir de ce codage, on peut obtenir la loi de la progéniture totale de la forêt de branchement à d types F.

Théorème 1.3.1 : ([CL15, Théorème 1.2]) Pour tous entiers $x_i, n_i, k_{ij}, i, j \in [d]$, tels que $x_i > 0, x_1 + \dots + x_d \geq 1, k_{ij} \geq 0\ \forall i \neq j, -k_{jj} = x_j + \sum_{i \neq j} k_{ij}$ et $n_i \geq -k_{ii}$,

$$\begin{aligned}&\mathbb{P}_{\mathbf{x}}\left(N_1 = n_1, \dots, N_d = n_d, M_{ij} = k_{ij}, \forall i \neq j\right)\\ &= \frac{\det(K)}{\bar{n}_1 \dots \bar{n}_d} \prod_{i=1}^{d} \nu_i^{*n_i}(k_{i1}, \dots, k_{i(i-1)}, n_i + k_{ii}, k_{i(i+1),\dots,k_{id}}),\end{aligned}$$

où $\mathbf{x} = (x_1, \dots, x_d), \nu_i^{*0} = \delta_0, \bar{n}_i = n_i \vee 1$, K est la matrice $(-k_{ij})_{i,j}$ dont on enlève la i-ième ligne et la i-ième colonne pour tout i tel que $n_i = 0$ et $\mathbb{P}_{\mathbf{x}}$ est la probabilité sous condition $Z_0 = \mathbf{x}$.

1.3.2 Représentation de type Lamperti en dimension supérieure

Nous rappelons d'abord dans cette section la notion de *forêt de branchement multitype avec longueurs d'arc* introduite dans [Cha15] qui nous permet de prendre en compte la continuité des temps de vie des individus.

Commençons par la définition d'une *forêt multitype avec longueurs d'arc*. Une *forêt à d types avec longueurs d'arc*, où $d \in \mathbb{N},\ d > 1$, est un triplet $(\mathbf{f}, c_{\mathbf{f}}, l_{\mathbf{f}})$ qui sera noté $\mathbf{f}$ lorsqu'il n'y a pas ambiguïté, où $(\mathbf{f}, c_{\mathbf{f}}) \in \mathcal{F}_d$ et $l_{\mathbf{f}}$ est une application allant de l'ensemble

des sommets de $\mathbf{f}$ à l'ensemble $(0, +\infty)$. Pour chaque sommet u, $l_{\mathbf{f}}(u)$ est appelé *le temps de vie* de u. L'ensemble des forêts à d types avec longueurs d'arc est noté par $\mathbb{F}_d$.

Une forêt de branchement à d types avec longueurs d'arc, finie, de loi de reproduction $\nu := (\nu_1, \ldots, \nu_d)$ et de paramètres $\lambda_i > 0,\ i \in [d]$, est une variable aléatoire F à valeurs dans l'ensemble des forêts à d types avec longueurs d'arc $\mathbb{F}_d$ telle que le temps de vie de chaque sommet de type $i \in [d]$ est toujours de loi exponentielle de paramètre λ_i, et que son squelette, c'est-à-dire la forêt obtenue en supprimant la longueur des arcs, est une forêt de branchement à d types de loi de reproduction ν. De plus, le squelette est indépendant des longueurs d'arc et les longueurs d'arc sont indépendantes entre elles.

Étant donnée une forêt de branchement à d types avec longueurs d'arc F, $d > 1$, de loi de reproduction ν et de paramètres $\lambda_i > 0,\ i \in [d]$, partant de l'état $\mathbf{x}$, on note $\mathbf{Z} = (\mathbf{Z}_t)_{t\geq 0} = (\mathbf{Z}_t^{(1)}, \ldots, \mathbf{Z}_t^{(d)})_{t\geq 0}$ le processus de Galton-Watson correspondant au sens où pour $i \in [d],\ \mathbf{Z}_t^{(i)}$ est le nombre d'individus de type i au temps t. On introduit ensuite les processus $Z_t^{i,j},\ i, j \in [d]$. Pour $i \neq j,\ Z_t^{i,j}$ est le nombre total des individus de type j dont le parent est de type i qui sont nés avant le temps t. Pour $i = j$, la définition de $Z_t^{i,j}$ est la même, excepté qu'à ce nombre on ajoute le nombre d'individus de type i au temps 0 et on soustrait le nombre d'individus de type i étant morts avant le temps t. Notons que les processus $Z_t^{i,j},\ i, j \in [d]$ portent la partie essentielle de l'information sur l'évolution de la population sous-jacente que contient la forêt F, et que

$$\mathbf{Z}_t^{(i)} = \sum_{j=1}^{d} Z_t^{j,i},\ i \in [d].$$

Le résultat suivant, qui se trouve dans [Cha15], nous donne la représentation de type Lamperti des processus $Z_t^{i,j},\ i, j \in [d]$ et en particulier du processus de branchement multitype $\mathbf{Z}$:

Théorème 1.3.2 : ([Cha15, Théorème 2.4]) Les processus $Z_t^{i,j},\ i, j \in [d]$ admettent la représentation suivante :

$$Z_t^{i,j} = \begin{cases} X^{i,j}_{\int_0^t \mathbf{Z}_s^{(i)} ds}, & i \neq j \\ x_i + X^{i,j}_{\int_0^t \mathbf{Z}_s^{(i)} ds}, & i = j, \end{cases}$$

où pour $i \in [d],\ X^{(i)} = (X^{i,1}, \dots, X^{i,d})$ est un processus de Poisson composé, de loi de saut $\tilde{\nu}_i$ et de paramètre $\lambda_i > 0$. De plus, les processus $X^{(i)},\ i \in [d]$ sont indépendants.

En particulier, le processus $\mathbf{Z}$ admet la représentation suivante :

$$\mathbf{Z}_t^{(i)} = x_i + \sum_{k=1}^{d} X^{k,i}_{\int_0^t \mathbf{Z}_s^{(k)} ds},\ i \in [d],\ t \geq 0.$$

1.4 Résumé des travaux

1.4.1 Sur les mutations dans le modèle de branchement pour des populations multitypes

Notre premier travail concerne les mutations qui apparaissent dans le développement d'une population constituée de plusieurs types d'individus. Par *mutation* nous entendons la naissance d'un individu qui n'est pas du même type que son parent. Nous utilisons le modèle de branchement multitype pour modéliser l'évolution d'une telle population. Plus précisément, nous étudions la forêt de branchement multitype généalogique de la population en question.

On réintroduit, en première étape, la notion de *forêt des mutations* associée à une forêt de branchement multitype, qui a été introduite dans [CL15] sous le nom de *forêt réduite* et qui est en réalité la forêt obtenue en fusionnant tous les sommets connectés d'un même type en un seul sommet. Nous montrons d'abord que la forêt des mutations associée à n'importe quelle forêt de branchement multitype est elle-même une forêt de branchement multitype si et seulement si pour tout $i \in [d]$, l'une des conditions suivantes est satisfaite,

$$(A_i) \qquad m_{ii} \leq 1\,,$$
$$(B_i) \qquad m_{ii} > 1 \text{ et pour tout } j \neq i,\ m_{ij} = 0.$$

De plus, la loi de reproduction de la forêt des mutations peut être calculée en termes de celle de la forêt d'origine. Notre premier résultat est le suivant :

Théorème 1.4.1 : Soient F une forêt de branchement à d types de loi de reproduction $\nu = (\nu_1, \ldots, \nu_d)$ et $\overline{F}$ la forêt de mutations associée. Supposons que pour tout $i \in [d]$, l'une des conditions (A_i) ou (B_i) soit satisfaite. Alors $\overline{F}$ est une forêt de branchement à d types dont la loi de reproduction $\mu = (\mu_1, \ldots, \mu_d)$ est définie sur $\mathcal{S}_i := \{\mathbf{k} \in \mathbb{Z}_+^d : k_i = 0\}$ par :

$$\mu_i(\mathbf{k}) = \sum_{n \geq 1} n^{-1} \nu_i^{*n}(\mathbf{k} + (n-1)\mathbf{e}_i), \quad \mathbf{k} \in \mathcal{S}_i, \tag{1.1}$$

si (A_i) est satisfaite. Si (B_i) est satisfaite, alors μ_i est la mesure de Dirac au point 0. De plus, μ possède les propriétés suivantes :

1. Soient $\overline{\mathbf{M}} = (\bar{m}_{ij})$ la matrice de moyennes de μ et $r \geq 1$. Alors μ_i admet des moments d'ordre r si et seulement si, soit pour tout $j \neq i,\ m_{ij} = 0$, soit ν_i admet des moments d'ordre r et $m_{ii} < 1$. Dans le dernier cas, pour tout i, j tel que $i \neq j,\ \bar{m}_{ij} = \frac{m_{ij}}{1-m_{ii}}$.
2. Supposons que $\bar{m}_{ij} < \infty$, pour tout $i, j \in [d]$. Alors $\overline{\mathbf{M}}$ est irréductible si et seulement si $\mathbf{M}$ est irréductible. Si $\overline{\mathbf{M}}$ est primitive, alors $\mathbf{M}$ l'est. L'inverse n'est pas vrai.
3. Supposons que $\overline{\mathbf{M}}$ soit primitive, alors $\overline{\mathbf{M}}$ est sous-critique (resp. $critique$, resp. $surcritique$) si et seulement si $\mathbf{M}$ est sous-critique (resp. $critique$, resp. $surcritique$).

Si pour un $i \in [d]$, aucune des conditions (A_i) ou (B_i) n'est satisfaite, alors il existe un $j \neq i$ tel que les individus de type i dans $\overline{F}$ donne naissance à une infinité d'enfants de type j avec une probabilité positive. $\overline{F}$ n'est pas donc une forêt de branchement dans notre sens.

Ce résultat nous permet ensuite, en ayant recours au résultat donné dans le *théorème 1.3.1*, d'obtenir la loi du nombre total de mutations. Notons M_i le nombre total de mutations de type i et M_{ij} le nombre d'individus de type j qui ont un parent de type i pour tout $i \neq j$. Nous en déduisons le

Corollaire 1.4.2 : Supposons que l'une des conditions (A_i) ou (B_i) soit satisfaite pour tout $i \in [d]$. Alors pour tous nombres entiers $x_i, n_i, k_{ij}, i, j \in [d]$, tels que $x_i \geq 0$,

$n_i = -k_{ii}$, pour $i \neq j$, $k_{ij} \geq 0$, et pour tout $j \in [d]$, $n_j = x_j + \sum_{i \neq j} k_{ij}$,

$$\begin{aligned}&\mathbb{P}_{\mathbf{x}}\left(M_1 = n_1 - x_1, \ldots, M_d = n_d - x_d, M_{ij} = k_{ij}, \forall i \neq j\right)\\ &= \frac{\det(K)}{\bar{n}_1 \ldots \bar{n}_d} \prod_{i=1}^{d} \mu_i^{*n_i}(k_{i1}, \ldots, k_{i(i-1)}, 0, k_{i(i+1),\ldots,k_{id}}),\end{aligned}$$

où $\mu_i^{*0} = \delta_0, \bar{n}_i = n_i \vee 1$, K est la matrice $(-k_{ij})_{i,j}$ dont on enlève la ligne i et la colonne i pour tout i tel que $n_i = 0$.

On s'intéresse ensuite au comportement asymptotique du nombre total de mutations. Soit N_i l'effectif total de type i, nous pouvons montrer, dans le cas où ν est critique le résultat suivant :

Corollaire 1.4.3 : Soit F une forêt de branchement multitype dont la loi de reproduction ν est critique. Supposons que $i \in [d]$, μ_i admette des moments d'ordre $d+1$. Supposons de plus que $\overline{\mathbf{M}}$ soit primitive et que les matrices de covariance Σ^i, $\overline{\Sigma}^i$ de ν_i et μ_i respectivement soient définies positives. Alors $m_{ii} < 1$, pour tout $i \in [d]$ et il existe des constantes $C_1, C_2 > 0$ telles que pour tout $\mathbf{x}_0 \in \mathbb{Z}_+^d$,

$$\begin{aligned}\lim_{n\to\infty} n^{d/2+1}\mathbb{P}_{\mathbf{x}_0}(M_i = \lfloor n(1-m_{ii})v_i \rfloor,\ i \in [d]) &= C_1 \mathbf{x}_0 \cdot \mathbf{u}\,,\\ \lim_{n\to\infty} n^{d+1}\mathbb{P}_{\mathbf{x}_0}(M_i = \lfloor n(1-m_{ii})v_i \rfloor,\ N_i = \lfloor nv_i \rfloor\,,\ i \in [d]) &= C_2 \mathbf{x}_0 \cdot \mathbf{u}.\end{aligned}$$

Une conclusion que l'on peut tirer de ce résultat est que le nombre de mutations de type i est asymptotiquement proportionnel à $(1 - m_{ii})$ fois le nombre total d'individus de même type.

Une question intéressante est de savoir comment se comporte le nombre total de mutations, le nombre total d'individus et la proportion de mutations dans la population, lorsque le nombre initial d'individus tend vers l'infini en suivant une direction donnée. La réponse est donnée par les résultats suivants, en notant, pour tout $i \in [d], \mathbf{w} \in \mathbb{Z}_+^d \setminus \{0\}$, $N_i(n\mathbf{w})$ (resp. $M_i(n\mathbf{w})$) le nombre total d'individus (resp. mutations) de type i dans une forêt de branchement multitype contenant $n(w_1 + \cdots + w_d)$ arbres dont nw_i arbres de type i.

Théorème 1.4.4 : Soit $\mathbf{w} \in \mathbb{Z}_+^d \setminus \{0\}$.

1. Si ν est critique, alors

$$\lim_{n\to\infty} \frac{N_i(n\mathbf{w})}{n} = \infty \text{ et } \lim_{n\to\infty} \frac{M_i(n\mathbf{w})}{N_i(n\mathbf{w})} = 1 - m_{ii}\,, \text{ en probabilité.}$$

2. Si ν est sous-critique, alors

$$\lim_{n\to\infty} \frac{N_i(n\mathbf{w})}{n} = c_i(w) \text{ et } \lim_{n\to\infty} \frac{M_i(n\mathbf{w})}{n} = w_i + (1 - m_{ii})c_i(\mathbf{w})\,, \text{ en probabilité,}$$

où $c_i(\mathbf{w}) := \sum_{k=1}^d w_k(\mathbf{I} - \mathbf{M})^{-1}_{ki}$ et ν est la loi de reproduction de la forêt en question.

En temps continu, tous les résultat précédents se généralisent. De plus, on peut établir, dans le cas surcritique, le comportement asymptotique du nombre de mutations à un certain instant. Rappelons que d'après un résultat connu (voir par exemple [AN72, Théorème 2]), il existe une variable aléatoire positive ou nulle W telle que pour tout $i \in [d]$,

$$\lim_{t\to\infty} e^{-\rho_1 t} Z_t^{(i)} = v_i W\,, \text{ p.s.,} \tag{1.2}$$

où v_i est la i-ième coordonnée du vecteur propre normalisé associé à ρ_1.

En ce qui nous concerne, nous avons établi le résultat suivant qui montre que le nombre de mutations se comporte de la même manière que le nombre total d'individu, lorsque le temps tend vers l'infini.

Proposition 1.4.5 : Supposons que ν soit surcritique. Pour tout $i \in [d]$, notons $M_{i,t}$ le nombre de mutations de type i qui apparaissent jusqu'au temps t, alors

$$\lim_{t\to\infty} e^{-\rho_1 t} M_{i,t} = K_i W,\ p.s.,$$

où $K_i = v_i(1 + (1 - m_{ii})(\lambda_i\rho_1)^{-1})$.

Le temps d'émergence des mutations dans une classe de modèles plus particuliers est aussi étudié dans cette partie. Dans ce travail, on se restreint aux modèles irréversibles,

autrement dit, les individus de type i ne peuvent donner naissance qu'à des individus de type i ou $i+1$. Quelques hypothèses techniques sur la loi de reproduction sont aussi nécessaires.

Notons $\tau_i := \inf\{t \geq 0 : Z_t^{(i)} \geq 1\}$ le premier temps au bout duquel un individu de type i apparaît dans la population. Notre contribution principale dans cette partie est de donner une approximation de τ_i en terme des processus de Poisson composés qui apparaissaient dans la représentation de Lamperti du processus $\mathbf{Z}$ d'une part, et une approximation de l'espérance de τ_i par la somme des espérances de certaines variables aléatoires élémentaires $\theta_k,\ k = 2, \ldots, i$, lorsque le taux de mutation du type k croît plus vite que celui du type $k-1$ pour tout $k = 2, \ldots, i$, d'autre part. Plus précisément, nous montrons que

Théorème 1.4.6 : Supposons que

$$\begin{cases} \nu_i(\mathbf{k}) > 0 \ \Rightarrow \ k_j = 0, \text{ pour } j \notin \{i, i+1\}, \\ \sum_{\mathbf{k}\in\mathbb{Z}_+^d : k_i = 0} \nu_i(\mathbf{k}) = 0 \ \text{ et } \ \sum_{\mathbf{k}\in\mathbb{Z}_+^d : k_{i+1} = 0} \nu_i(\mathbf{k}) < 1. \end{cases} \tag{1.3}$$

Alors sous $\mathbb{P}_{\mathbf{e}_1}$,

$$\frac{\tau_i}{\sum_{k=2}^{i} \theta_k} \xrightarrow{P} 1\,, \quad \text{lorsque } \frac{\lambda_{k-2,k-1}}{\lambda_{k-1,k}} \to 0, \text{pour } i \geq 3, k = 3, \ldots, i,$$

où $\theta_k = \int_0^{\gamma_k} \frac{1}{X_s^{k-1,k-1}+1}\,\mathrm{d}s$ et γ_k est le premier temps de saut du processus $X^{k-1,k},\ k \geq 2$; et pour $i \geq 3$,

$$\mathbb{E}_{\mathbf{e}_1}(\tau_i) \sim \sum_{k=2}^{i} \mathbb{E}(\theta_k)\,, \quad \text{lorsque } \frac{\lambda_{k-2,k-1}}{\lambda_{k-1,k}} \to 0, \quad \text{pour } k = 3, \ldots, i.$$

Notons que la situation dans laquelle le taux de mutation du type k croît plus vite que celui du type $k-1$ apparait assez souvent dans le développement d'une population. En effet, une fois apparue, une mutation est de plus en plus sensible à de nouvelles mutations qui surviennent de plus en plus rapidement. Un exemple est le développement du cancer.

Finalement, nous donnons un exemple dont la distribution de temps d'émergence peut être estimée assez explicitement.

1.4.2 Suites cycliquement échangeables et énumérations de forêts multitypes

Notre deuxième travail a pour objectif d'obtenir plusieurs résultats anciens et nouveaux sur l'énumération de forêts multitypes ainsi qu'une nouvelle preuve de la formule d'inversion de Lagrange-Good. L'ensemble de ces résultats utilise une version multivariée du théorème de ballotage récemment établie en [CL15], ainsi que ses conséquences sur la loi de l'effectif total des forêts de branchement multitypes.

Nous énonçons ici les deux principales formules d'énumération de forêts planaires multitypes. Celles-ci généralisent les résultats existants sur l'énumération de forêts planaires monotypes. La première formule obtenue est une extension au cas multitype d'un résultat connu en dimension 1 dont le nombre de forêts planaires à n sommets et r arbres est $\frac{r}{n}\binom{2n-r-1}{n-r}$. Notons que pour $r=1$, ce nombre correspond au n-ième nombre de Catalan.

Théorème 1.4.7 : Soit $\mathcal{F}_d^{k_{ij},\mathbf{n}}$ un sous-ensemble des forêts planaires (non marquées) de $\mathcal{F}_d$ ayant n_i sommets de type i, r_i racines de type i et tel que pour $i \neq j$, k_{ij} sommets de type j ayant un parent de type i, alors nous montrons que :

$$\left|\mathcal{F}_d^{k_{ij},\mathbf{n}}\right| = \frac{\det(-k_{ij})}{n_1 n_2 \dots n_d} \prod_{i,j=1}^{d} \binom{n_i + k'_{ij} - 1}{k'_{ij}} ,$$

où $k'_{ii} = n_i + k_{ii}$ et pour $i \neq j$, $k'_{ij} = k_{ij}$.

Nous nous intéressons ensuite au cas des forêts de $\mathcal{F}_d^{k_{ij},\mathbf{n}}$ dont les degrés mutitypes sont donnés.

Théorème 1.4.8 :

$$\left|\mathcal{F}_d^{k_{ij},\mathbf{n}}(\mathbf{O})\right| = \frac{\prod_{j=1}^{d}(n_j-1)!\det(-k_{ij})}{\prod_{i\in[d],\mathbf{u}\in\mathbf{U}}(N_{i,\mathbf{u}})!}.$$

où $\mathcal{F}_d^{k_{ij},\mathbf{n}}(\mathbf{N})$ est un sous-ensemble de $\mathcal{F}_d^{k_{ij},\mathbf{n}}$ constitué de forêts non marquées, ayant $N_{i,\mathbf{u}}$ sommets de type i avec degré $\mathbf{u}$ pour $i \in [d]$, $\mathbf{u} \in \mathbf{U}$, $\mathbf{U} = \prod_{i=1}^{d}\{0,1,\dots,n_i\}$.

Dans une seconde partie de ce travail nous obtenons des preuves directes de certaines formules d'énumération de forêts numérotées récemment établies par Bernardi et Morales [BM14].

Nous présentons à la fin de ce travail une nouvelle preuve de la formule d'inversion de Lagrange-Good qui permet d'obtenir une expression des coefficients de la fonction génératrice d'une loi de probabilité multidimensionnelle.

1.4.3 Une note sur les sommets ayant un degré donné dans une forêt de branchement multitype

Notre troisième travail concerne l'étude d'une autre fonctionnelle d'une forêt de branchement multitype : *le nombre de sommets ayant un degré donné*. Autrement dit, nous nous intéressons au nombre total d'individus ayant une certaine progéniture donnée dans une population à plusieurs types d'individu.

Nous considérons une forêt de branchement à d types représentant la généalogie d'une population d'individus à d types. Dans le but de simplifier, nous travaillons avec les individus n'ayant pas d'enfants ou bien les feuilles de la forêt généalogique. Notons N_i le nombre total d'individus de type i, τ_i le nombre d'individus de type i qui n'ont pas d'enfants et M_{ij} le nombre d'individus de type j dont le parent est de type i.

Dans un premier temps, nous donnons la loi jointe de l'effectif total et du nombre de feuilles. Plus précisément, nous montrons que

Théorème 1.4.9 : Pour tous nombres entiers $q_i, r_i, n_i, k_{ij}, i, j \in [d]$, tels que $\sum_{i=1}^{d} r_i \geq 1, n_i \leq q_i,\ k_{ii} = -r_i - \sum_{j\neq i} k_{ji}$,

$$\begin{aligned}&\mathbb{P}_{\mathbf{r}}\left(N_1 = q_1, \dots, N_d = q_d, \tau_i = n_i, M_{ij} = k_{ij}, \forall i \neq j\right)\\ &= \frac{\det(K)}{\bar{q}_1 \dots \bar{q}_d} \prod_{i=1}^{d} \mathbb{P}\left(S^i_{q_i} = n_i\right) \mathbb{P}\left(W^i_{q_i - n_i} = \mathbf{k}_i + n_i \mathbf{e}_i\right),\end{aligned}$$

où $\mathbf{r} = (r_1, \dots, r_d), \mathbf{k}_i = (k_{i1}, \dots, k_{id}), \bar{q}_i = q_i \vee 1$, K est la matrice $(-k_{ij})$ à laquelle on enlève la ligne i et la colonne i pour tout i tel que $q_i = 0$, S^i est une marche aléatoire dont

la loi de saut est une loi de Bernoulli de paramètre $\nu_i(\mathbf{0})$ et W^i est une marche aléatoire dont les sauts sont distribués selon la loi $\eta_i(\mathbf{z}) = \frac{\nu_i(\mathbf{z}+\mathbf{e}_i)}{1-\nu_i(\mathbf{0})}$ pour tout $\mathbf{z} \in \mathbb{Z}_+^d$.

A partir de ce résultat, nous pouvons déduire un résultat sur le comportement asymptotique du nombre de feuilles. Supposons que la loi de reproduction ν est critique, admet des moments d'ordre $d+1$, et que ses matrices de covariance sont toutes définie-positives. Alors nous montrons que

Corollaire 1.4.10 : Pour tout $\mathbf{r} \in \mathbb{N}^d$,

$$\mathbb{P}_{\mathbf{r}}(\tau_i = \nu_i(\mathbf{0})\lfloor nv_i \rfloor, N_i = \lfloor nv_i \rfloor, i \in [d]) = o\left(n^{-(d+1)}\right).$$

Nous donnons enfin des résultats similaires obtenus dans deux scénarios plus généraux où l'on considère le nombre total d'individus qui ont une certaine progéniture donnée ainsi que le nombre total d'individus dont la progéniture appartient à un certain ensemble (non-vide).

1.5 Perspectives

Pour la suite de notre premier travail présenté au chapitre 2, nous souhaiterions étendre l'étude du temps d'émergence d'une mutation dans une population multitype, de façon à obtenir une expression plus explicite de sa loi, dans des cas plus généraux. Ceci nécessitera un codage approprié des forêts multitypes.

À propos de notre troisième travail sur le nombre de sommets ayant un degré donné d'une forêt de branchement multitype, un défi serait d'étendre au cas multitype le résultat de [Kor12, Théorème 3.1] sur l'estimation asymptotique de la probabilité qu'un arbre de Galton-Watson contienne n feuilles, puis dans un second temps d'étendre les résutats de [Kor12, Théorème 4.5] et de [Kor12, Théorème 5.9] qui portent sur l'étude de la limite d'échelle d'une suite d'arbres de branchement conditionnés au nombre de feuilles.

Chapitre 2

On mutations in the branching model for multitype populations

2.1 Introduction

The homogeneous multitype branching hypothesis provides a relevant model of population growth in the absence of any competitive or environmental constraint. In particular, it is widely used in population genetics, when studying successive mutations whose accumulation leads to the development of cancer. Then determining the statistics of the emergence times of mutations, or evaluating the distribution of the population size of mutant cells at any time become important challenges. In the extensive literature on the subject, let us simply cite [IMKN05], [HIM07], [DM10], and [Dur13].

This work is concerned with the mathematical study of mutations in multitype branching frameworks. We first focus on the problem of the total number of mutations under very general assumptions. This number is not a functional of the associated branching process and its study requires the complete knowledge of the multitype branching structure, that is the underlying plane forest. Then we show that the forest of mutations associated to any multitype forest, is itself a multitype branching forest whose progeny

distribution can be explicitely computed. This result allows us to investigate the asymptotic behaviour of the number of mutations, when either the total population or the initial number of individuals tend to infinity.

When time is continuous, we are mainly interested in emergence times of new mutations in the non reversible case. The characterisation of these times requires a good knowledge of the corresponding multitype branching process and the main tool in this study consists in a recent extention of the Lamperti representation in higher dimensions. Emergence times are then expressed in terms of the underlying multivariate compound Poisson process, which allows us to obtain some accurate approximations of their law.

We start with some preliminaries on the coding of multitype branching forests by multivariate random walks in Section 2.2.1. Then we state and prove our results on the total mutations sizes of branching forests in Sections 2.2.2 and 2.3.2. Results bearing on emergence times are presented in Section 2.3.3.

2.2 Mutations and their asymptotics in discrete multitype forests

2.2.1 Preliminaries on discrete multitype forests

In all this work, we use the notation $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$ and for any positive integer d, we set $[d] = \{1, \ldots, d\}$. We will denote by $\mathbf{e}_i$ is the i-th unit vector of $\mathbb{Z}_+^d$. We define the following partial order on $\mathbb{R}^d$ by setting $x = (x_1, \ldots, x_d) \geq y = (y_1, \ldots, y_d)$, if $x_i \geq y_i$, for all $i \in [d]$. The convention $\inf \emptyset = +\infty$ will be valid all along this paper. Then $(\Omega, \mathcal{F}, P)$ is a reference probability space on which all the stochastic processes involved in this work are defined.

Let us first recall the coding of multitype forests, as it has been defined in [CL15]. A (plane) forest $\mathbf{f}$ is a directed planar graph with no loops on a possibly infinite and non

empty set of vertices $\mathbf{v} = \mathbf{v}(\mathbf{f})$, such that each vertex has a finite inner degree and an outer degree equals to 0 or 1. The connected components of a forest are called the *trees*. In a tree $\mathbf{t}$, the only vertex with outer degree equal to 0 is called the *root* of $\mathbf{t}$. The roots of the connected components of a forest $\mathbf{f}$ are called the roots of $\mathbf{f}$. For two vertices u and v of a forest $\mathbf{f}$, if (u, v) is a directed edge of $\mathbf{f}$, then we say that u is a *child* of v, or that v is the *parent* of u. We first give an order to the trees of the forest $\mathbf{f}$ and denote them by $\mathbf{t}_1(\mathbf{f}), \mathbf{t}_2(\mathbf{f}), \ldots, \mathbf{t}_k(\mathbf{f}), \ldots$ (we will usually write $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_k, \ldots$ if no confusion is possible). Then we rank (a part of) the vertices of $\mathbf{f}$ according to the breadth first search order, by ranking first the vertices of $\mathbf{t}_1$, then the vertices of $\mathbf{t}_2$, and so on, see the labeling of the two forests in Figure 2.2. Note that if $\mathbf{t}_k$, for $k \geq 1$ is the first infinite tree, then the vertices of $\mathbf{t}_{k+1}, \ldots$ have no label according to this procedure.

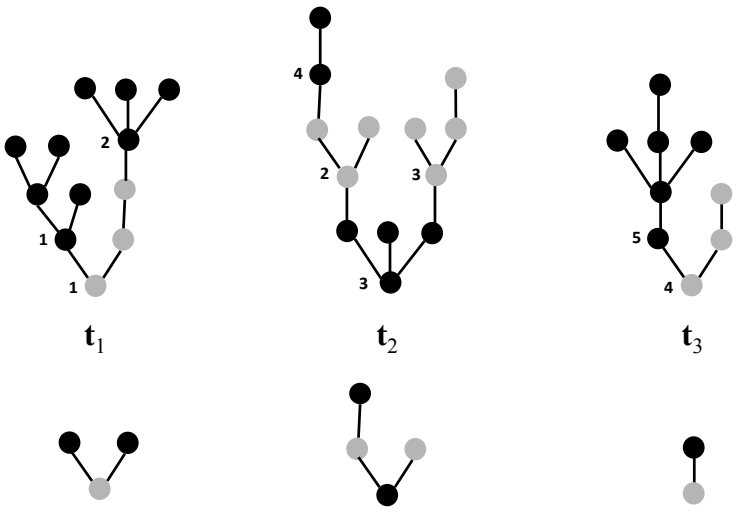

FIGURE 2.1 – On the top, a discrete 2-type forest. Roots of clusters are ranked in the breadth first search order of the forest. The rank is written on the left of these roots. Below, the corresponding forest of mutations.

To each forest $\mathbf{f}$, we associate the application $c_{\mathbf{f}} : \mathbf{v}(\mathbf{f}) \to [d]$ such that if $u_i, u_{i+1}, \ldots, u_{i+j} \in \mathbf{v}(\mathbf{f})$ have the same parent and are placed from left to right, then $c_{\mathbf{f}}(u_i) \leq c_{\mathbf{f}}(u_{i+1}) \leq \cdots \leq c_{\mathbf{f}}(u_{i+j})$. For $v \in \mathbf{v}(\mathbf{f})$, the integer $c_{\mathbf{f}}(v)$ is called the *type* (or the

color) of v. The couple $(\mathbf{f}, c_{\mathbf{f}})$ is called a *d-type forest*. When no confusion is possible, we will simply write $\mathbf{f}$. The set of d-type forests will be denoted by $\mathcal{F}_d$.

A *cluster* or a *subtree of type* $i \in [d]$ of a d-type forest $(\mathbf{f}, c_{\mathbf{f}}) \in \mathcal{F}_d$ is a maximal connected subgraph of $(\mathbf{f}, c_{\mathbf{f}})$ whose all vertices are of type i. Formally, $\mathbf{t}$ is a cluster of type i of $(\mathbf{f}, c_{\mathbf{f}})$, if it is a connected subgraph whose all vertices are of type i and such that either the root of $\mathbf{t}$ has no parent or the type of its parent is different from i. Moreover, if the parent of a vertex $v \in \mathbf{v}(\mathbf{t})^c$ belongs to $\mathbf{v}(\mathbf{t})$, then $c_{\mathbf{f}}(v) \neq i$. Clusters of type i in $\mathbf{t}_1$ are ranked according to the order of their roots in the breadth first search order of $\mathbf{t}_1$, see Figures 2.1 and 2.2. Then if the number of clusters of type i is finite in $\mathbf{t}_1$, we continue by ranking clusters of type i in $\mathbf{t}_2$, and so on. Note that with this procedure, it is possible that clusters of $\mathbf{t}_k, \mathbf{t}_{k+1}, \ldots$, for some k, are not ranked. We denote by $\mathbf{t}_1^{(i)}, \mathbf{t}_2^{(i)}, \ldots, \mathbf{t}_k^{(i)}, \ldots$ the sequence of clusters of type i in $(\mathbf{f}, c_{\mathbf{f}})$. The forest $\mathbf{f}^{(i)} := \{\mathbf{t}_1^{(i)}, \mathbf{t}_2^{(i)}, \ldots, \mathbf{t}_k^{(i)}, \ldots\}$ is called *the subforest of type* i of $(\mathbf{f}, c_{\mathbf{f}})$. We denote by $u_1^{(i)}, u_2^{(i)}, \ldots$ the elements of $\mathbf{v}(\mathbf{f}^{(i)})$, ranked in the breadth first search order of $\mathbf{f}^{(i)}$. The subforests of the 2-type forest given in Figure 2.1 are represented in Figure 2.2.

To any forest $(\mathbf{f}, c_{\mathbf{f}}) \in \mathcal{F}_d$, we associate the *forest of mutations*, denoted by $(\bar{\mathbf{f}}, c_{\bar{\mathbf{f}}}) \in \mathcal{F}_d$, which is the forest of $\mathcal{F}_d$ obtained by aggregating all the vertices of each subtree of $(\mathbf{f}, c_{\mathbf{f}})$ with a given type, in a single vertex with the same type, and preserving an edge between each pair of connected subtrees. An example is given in Figure 2.1.

For a forest $(\mathbf{f}, c_{\mathbf{f}}) \in \mathcal{F}_d$ and $u \in \mathbf{v}(\mathbf{f})$, when no confusion is possible, we denote by $p_i(u)$ the number of children of type i of u. For each $i \in [d]$, let $n_i \in \mathbb{Z}_+ \cup \{\infty\}$ be the number of vertices in the subforest $\mathbf{f}^{(i)}$ of $(\mathbf{f}, c_{\mathbf{f}})$. Then let us define the d-dimensional chain $x^{(i)} = (x^{i,1}, \ldots, x^{i,d})$, with length n_i and whose values belong to the set $\mathbb{Z}^d$, by

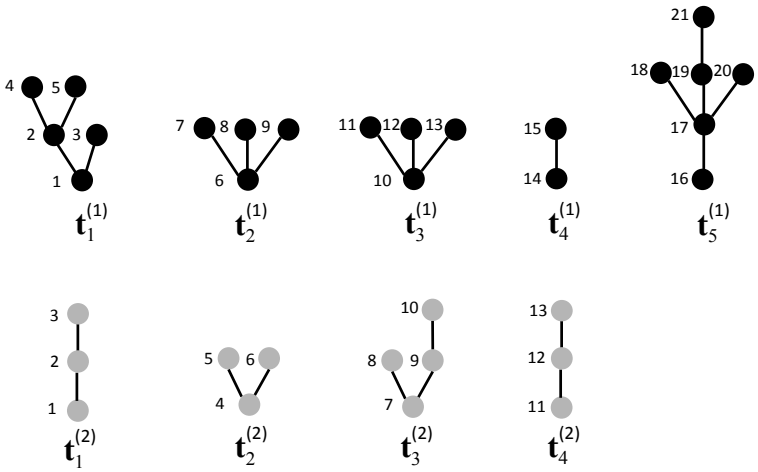

FIGURE 2.2 – The subforests of the 2-type forest given in Figure 2.1 with their deapth first search labeling.

$x_0^{(i)} = 0$ and if $n_i \geq 1$,

$$x_{n+1}^{i,j} - x_n^{i,j} = p_j(u_{n+1}^{(i)}), \quad \text{if } i \neq j \quad \text{and} \quad x_{n+1}^{i,i} - x_n^{i,i} = p_i(u_{n+1}^{(i)}) - 1, \quad 0 \leq n \leq n_i - 1, \tag{2.1}$$

where $(u_n^{(i)})_{n\geq 1}$ is the labeling of the subforest $\mathbf{f}^{(i)}$ in its own breadth first search order. Note that the chains $(x_n^{i,j})$, for $i \neq j$ are nondecreasing whereas $(x_n^{i,i})$ is a downward skip free chain, i.e. $x_{n+1}^{i,i} - x_n^{i,i} \geq -1$, for $0 \leq n \leq n_i - 1$. Besides, if n_i is finite, then $n_i = \min\{n : x_n^{i,i} = \min_{0\leq k\leq n_i} x_k^{i,i}\}$. Let us also mention that from Theorem 2.7 of [CL15], when trees of $(\mathbf{f}, c_\mathbf{f})$ are finite, the data of the chains $x^{(1)}, \ldots, x^{(d)}$ together with the sequence of ranked roots of $(\mathbf{f}, c_\mathbf{f})$, allow us to reconstruct this forest.

Let us now apply this coding to multitype branching forests. Let $\nu := (\nu_1, \ldots, \nu_d)$, where ν_i is some distribution on $\mathbb{Z}_+^d$. We consider a branching process with progeny distribution ν, that is a population of individuals which reproduce independently of each other at each generation. Individuals of type i give birth to n_j children of type $j \in [d]$ with probability $\nu_i(n_1, \ldots, n_d)$. For $i, j \in [d]$, we denote by m_{ij} the mean number of

children of type j, given by an individual of type i, i.e.

$$m_{ij} = \sum_{(n_1,\dots,n_d)\in\mathbb{Z}_+^d} n_j \nu_i(n_1,\dots,n_d)\,.$$

We say that ν is non singular if there is $i \in [d]$ such that $\nu_i(\mathrm{n} : n_1 + \cdots + n_d = 1) < 1$. The matrix $M = (m_{ij})$ is said to be irreducible if for all i, j, $m_{ij} < \infty$ and there exists $n \geq 1$ such that $m_{ij}^{(n)} > 0$, where $m_{ij}^{(n)}$ is the ij entry of the matrix M^n. If moreover the power n does not depend on (i, j), then M is said to be primitive. In the latter case, according to Perron-Frobenius theory, the spectral radius ρ of M is the unique eigenvalue which is positive, simple and with maximal modulus. If $\rho \leq 1$, then the population will become extinct almost surely, whereas if $\rho > 1$, then with positive probability, the population will never become extinct. We say that ν is subcritical if $\rho < 1$, critical if $\rho = 1$ and supercritical if $\rho > 1$. We sometimes say that μ is irreducible, primitive, (sub)critical or supercritical, when this is the case for M.

By multitype branching forest with progeny distribution ν, we mean a sequence with a finite (deterministic) or infinite number of independent multitype branching trees with progeny distribution ν. A multitype branching forest will be considered as a random variable defined on the probability space $(\Omega, \mathcal{F}, P)$ and with values in $\mathcal{F}_d$. To any multitype branching forest F, we associate the random sequences $X = \{X^{(i)}, i \in [d]\}$, where $X^{(i)} = \{(X_n^{i,1}, \dots, X_n^{i,d}), 0 \leq n \leq n_i\}$, which are constructed as in (2.1). It has been proved in [CL15], Theorem 3.1 that if F is a primitive and (sub)critical branching forest with a finite number of trees, then $X^{(i)}$, $i \in [d]$ are independent random walks whose step distribution $\tilde{\nu}_i$ is defined by

$$\tilde{\nu}_i(k_1,\dots,k_d) := \nu_i(k_1,\dots,k_{i-1},k_i+1,k_{i+1},\dots,k_d)\,, \quad \text{for all } (k_1,\dots,k_d) \in \mathbb{Z}_+^d, \tag{2.2}$$

and stopped at the smallest solution $(N_1, \dots, N_d)$ of the system

$$x_j + \sum_{i=1}^{d} X^{i,j}(N_i) = 0\,, \quad j \in [d]\,. \tag{2.3}$$

In this equation, N_i is the total number of vertices of type i in F and x_i is the total number of trees in this forest whose root is of type i. We will say that F is issued from $x = (x_1, \dots, x_d)$. Note that the variables N_i are random, whereas the x_i's are deterministic.

2.2.2 The total number of mutations and its asymptotics

A mutation of type i, is the birth event of an individual of type i from an individual of any type $j \neq i$. The aim of this section is to study the evolution of mutations in a multitype branching forest. Our main result asserts that the forest of mutations, that is the forest obtained by merging together all the vertices of a same cluster, is itself a branching forest if and only if for each $i \in [d]$, one of the following conditions is satisfied,

$$(A_i) \qquad m_{ii} \leq 1\,,$$
$$(B_i) \qquad m_{ii} > 1 \text{ and for all } j \neq i,\, m_{ij} = 0.$$

Moreover, its progeny distribution can be expressed in terms of this of the initial forest. Note that the branching property of the forest of mutations is intuitively clear. In the neutral case, it has been pointed out in [Taï92].

Theorem 2.2.1 : Let F be any multitype branching forest with progeny distribution $\nu = (\nu_1, \dots, \nu_d)$ and denote by $\overline{F}$ the associated forest of mutations. Assume that for all $i \in [d]$, one of the conditions (A_i) or (B_i) holds. Then $\overline{F}$ is a multitype branching forest with progeny distribution $\mu = (\mu_1, \dots, \mu_d)$ on $\mathcal{S}_i := \{\mathbf{k} \in \mathbb{Z}_+^d : k_i = 0\}$, which is defined by

$$\mu_i(\mathbf{k}) = \sum_{n \geq 1} n^{-1} \nu_i^{*n}(\mathbf{k} + (n-1)\mathbf{e}_i)\,, \quad \mathbf{k} \in \mathcal{S}_i\,, \tag{2.4}$$

if (A_i) is satisfied. If (B_i) is satisfied, then μ_i is the Dirac mass at 0. Moreover μ satisfies the following properties :

1. Let $\overline{M} = (\bar{m}_{ij})$ be the mean matrix of μ and let $r \geq 1$. Then μ_i admits moments of order r if and only if either for all $j \neq i$, $m_{ij} = 0$ or ν_i admits moments of order

r and $m_{ii} < 1$. In the latter case, for all i, j such that $i \neq j$, $\bar{m}_{ij} = \frac{m_{ij}}{1-m_{ii}}$.

2. Assume that $\bar{m}_{ij} < \infty$, for all $i, j \in [d]$. Then $\overline{M}$ is irreducible if and only if M is irreducible. If $\overline{M}$ is primitive, then so is M. The converse is not true.
3. Assume that $\overline{M}$ is primitive, then $\overline{M}$ is subcritical (resp. critical, resp. super-critical) if and only if M is subcritical (resp. critical, resp. supercritical).

If for some $i \in [d]$, none of the conditions (A_i) and (B_i) holds, then there is $j \neq i$ such that individuals of type i in $\overline{F}$ give birth to an infinite number of children of type j with positive probability. Therefore $\overline{F}$ is not a branching forest in our sense.

Démonstration. Since the result only bears on the progeny law of forests, we do not loose any generality by assuming that F has an infinite number of trees. Then the stochastic processes $X = \{X^{(i)}, i \in [d]\}$ obtained from F, as in (2.1) are defined on the whole integer line $\{0, 1, \ldots\}$. Note that their definition slighly extends the definition which is given in [CL15]. Indeed, without any more assumption on ν, trees of the forest can be infinite, so that the process X is not necessarily a coding of the forest, that is, if some trees are infinite then it is not possible to reconstruct the whole forest from X and the sequence of its roots. However, it is straightforward to check that $X^{(i)}$, $i \in [d]$ are independent random walks and that the step distribution of $X^{(i)}$ is $\tilde{\nu}_i$, which is defined in (2.2). In particular, the law of X characterizes this of F.

Now, let us consider the forest of mutations $\overline{F}$. By construction, this forest is composed of an infinite number of independent and identically distributed trees. Hence, in order to show that $\overline{F}$ is a branching forest, it suffices to show that its trees are branching trees.

Let us denote by $\{\overline{X}^{(i)}, i \in [d]\}$ the process which is defined from $\overline{F}$ as in (2.1). Let $i \in [d]$ and assume first that (A_i) holds. Then we define the first passage time process of the random walks $X^{i,i}$, $i \in [d]$ by,

$$\tau_k^{(i)} = \inf\{n \geq 0 : X_n^{i,i} = -k\}\,, \quad k \geq 0\,.$$

Since $m_{ii} \leq 1$, then from the law of large numbers, $\liminf_{n\to\infty} X_n^{i,i} = -\infty$, a.s., so that

$\tau_k^{(i)}$ is almost surely finite for all $k \geq 0$ and $\lim_{k\to\infty} \tau_k^{(i)} = \infty$, a.s. Moreover, for all $i, j \in [d]$,

$$\overline{X}_k^{i,j} = X^{i,j}(\tau_k^{(i)}), \quad k \geq 0.$$

Indeed, the effect of the time change by $\tau_k^{(i)}$ is to merge all vertices of a same cluster of type i into a single vertex. Note that $\overline{X}^{(i)}$, $i \in [d]$ are independent random walks. Assume with no loss of generallity that the root of the first tree in $\overline{F}$ has type 1, then a slight extention Theorems 2.7 and 3.1 in [CL15] to any progeny distribution, allows us to show that this first tree is coded by the processes $(\overline{X}_k^{(i)}, 0 \leq k \leq N_i)$, $i \in [d]$, where $(N_1, \ldots, N_d)$ is the smallest solution of the system

$$r_j + \sum_{i=1}^{d} \overline{X}^{i,j}(N_i) = 0, \quad j \in [d], \tag{2.5}$$

and $(r_1, \ldots, r_d) = (1, 0, \ldots, 0)$. Note that in our case, N_i can be infinite. This extended notion of smallest solution is defined in [Cha15], see Lemma 1 therein. This coding result implies that the first tree in $\overline{F}$ can be reconstructed from the processes $(\overline{X}_k^{(i)}, 0 \leq k \leq N_i)$, $i \in [d]$ and applying part 3. of Theorem 3.1 in [CL15], we obtain that this tree is a branching tree whose progeny distribution $\mu = (\mu_i, i \in [d])$ is given by

$$\mu_i(k_1, \ldots, k_d) = P(\overline{X}_1^{(i)} = (k_1, \ldots, k_{i-1}, -1, k_{i+1}, k_d)), \quad (k_1, \ldots, k_d) \in \mathcal{S}_i.$$

Then in order to make this law explicit in terms of ν, we apply the Ballot theorem for cyclically exangeable sequences due to Takács [Tak61]. Since conditionally on $X^{i,j}$, $i \neq j$, $X^{i,i}$ is downward skip free with cyclical exchangeable increments, we have for all $(k_1, \ldots, k_d) \in \mathcal{S}_i$,

$$\begin{aligned} & P(\overline{X}_1^{(i)} = (k_1, \ldots, k_{i-1}, -1, k_{i+1}, \ldots, k_d)) \\ = & \sum_{n\geq 1} P(X_n^{(i)} = (k_1, \ldots, k_{i-1}, -1, k_{i+1}, \ldots, k_d), \tau_1^{(i)} = n) \\ = & \sum_{n\geq 1} \frac{1}{n} P(X_n^{(i)} = (k_1, \ldots, k_{i-1}, -1, k_{i+1}, \ldots, k_d)), \end{aligned}$$

which gives (2.4) from (2.2). If (B_i) holds, then by definition, individuals of type i in $\overline{F}$

are all leaves and hence, $\overline{X}^{i,j} \equiv 0$, for all $j \neq i$ and $\overline{X}^{i,i}_n = -n$, for all $n \geq 0$, see (2.1). In this case, the conclusion follows immediately.

Let us now prove properties 1–3 of μ. First note that for all $i \neq j$, $m_{ij} = 0$ if and only if $\bar{m}_{ij} = 0$. Then let $r \geq 1$, assume that μ_i admits moments of order r and that there is $j \neq i$ such that $m_{ij} = \mathbb{E}(X^{i,j}_1) > 0$. The variable $\tau^{(i)}_1$ is a stopping time in the filtration generated by $X^{(i)}$ to which the increasing random walk $X^{i,j}$ is adapted. Then by applying Theorem 5.4 in [Gut09], we obtain that $\mathbb{E}((X^{i,j}_1)^r) < \infty$ and $\mathbb{E}((\tau^{(i)}_1)^r) < \infty$. In particular $\tau^{(i)}_1 < \infty$, a.s. Now by definition, the random walk $(X^{i,i}_n)$ can be written as $X^{i,i}_n = Y^{i,i}_n - n$, where $(Y^{i,i}_n)$ is an increasing random walk. Since $Y^{i,i}(\tau^{(i)}_1) = \tau^{(i)}_1 - 1$ and $\mathbb{E}((\tau^{(i)}_1)^r) < \infty$, we have $\mathbb{E}\left(|Y^{i,i}(\tau^{(i)}_1)|^r\right) < \infty$ and by applying Theorem 5.4 in [Gut09] again, we obtain that $\mathbb{E}\left(|Y^{i,i}_1|^r\right) < \infty$, and hence $\mathbb{E}\left(|X^{i,i}_1|^r\right) < \infty$. So we have proved that ν admits moments of order r. Then it follows from the definition of $\tau^{(i)}_1$ and from Lemma 3.1 in [KM96] that $\mathbb{E}((\tau^{(i)}_1)^r) < \infty$ implies $\lim_{n\to\infty} X^{i,i}_n = -\infty$, and hence $m_{ii} < 1$, from the law of large numbers.

Conversely, if $m_{ij} = 0$ for all $j \neq i$, then $\bar{m}_{ij} = 0$ for all $j \neq i$ and μ_i is the Dirac mass at 0, so it admits moments of order r.

Now assume that ν_i admits moments of order r and $m_{ii} < 1$. Then it follows directly from Lemma 3.1 in [KM96] that $\mathbb{E}((\tau^{(i)}_1)^r) < \infty$. Moreover from Theorem 5.2 in [Gut09], $\mathbb{E}(X^{i,j}(\tau^{(i)}_1)^r) < \infty$, for all $j \neq i$, which means that μ_i admits moments of order r. If ν_i admits moments of order 1 and $m_{ii} < 1$, then it follows from the optional stopping theorem applied to the martingale $(X^{i,j}_n - nE(X^{i,j}_1))$, that $\mathbb{E}(X^{i,i}(\tau^{(i)}_1)) = -1 = \mathbb{E}(X^{i,i}_1)\mathbb{E}(\tau^{(i)}_1) = (m_{ii} - 1)\mathbb{E}(\tau^{(i)}_1)$, and when $i \neq j$, $\mathbb{E}(X^{i,j}(\tau^{(i)}_1)) = \mathbb{E}(X^{i,j}_1)\mathbb{E}(\tau^{(i)}_1) = \frac{m_{ij}}{1-m_{ii}}$ and part 1 is proved.

If $\overline{\mathbf{M}}$ is irreducible, then for all i, there is $j \neq i$ such that $\bar{m}_{ij} > 0$. From part 1., ν_i admits moments of order 1 and $m_{ii} < 1$, for all i. In this case,

$$\overline{\mathbf{M}} + \Delta_2 = \Delta_1 M\,, \quad \text{where} \quad \Delta_1 = \mathrm{diag}(\frac{1}{1-m_{ii}}) \quad \text{and} \quad \Delta_2 = \mathrm{diag}(\frac{m_{ii}}{1-m_{ii}})\,,$$

and we derive from this identity that $\mathbf{M}$ is irreducible. Conversely if $\mathbf{M}$ is irreducible, then for all i, there is $j \neq i$ such that $m_{ij} > 0$ and hence $\bar{m}_{ij} > 0$. Since by assumption,

$\bar{m}_{ij} < \infty$, for all i, j, then from part 1., $m_{ii} < 1$, and $\overline{\mathbf{M}} + \Delta_2 = \Delta_1 M$ holds. We derive from this identity that $\overline{\mathbf{M}}$ is irreducible.

Now if $\overline{\mathbf{M}}$ is primitive, then it is irreducible and as before, $m_{ii} < 1$ for all $i \in [d]$. Moreover,

$$\mathbf{M} = (\mathbf{I} - \mathrm{diag}(m_{ii}))\overline{\mathbf{M}} + \mathrm{diag}(m_{ii})\,.$$

Therefore $\mathbf{M}$ is primitive. The converse cannot be true since there are nonnegative, irreducible matrices whose main diagonal is zero and which are not primitive. We can find distributions ν such that it is the case for $\overline{\mathbf{M}}$ and hence for $(\mathbf{I} - \mathrm{diag}(m_{ii}))\overline{\mathbf{M}}$. If $m_{ii} > 0$, for all i, then it follows from general theory of nonnegative matrices that $\mathbf{M} = (\mathbf{I} - \mathrm{diag}(m_{ii}))\overline{\mathbf{M}} + \mathrm{diag}(m_{ii})$ becomes primitive, see [Sen06].

Let us now prove 3. Recall that by definition, since $\overline{\mathbf{M}}$ is primitive, μ_i admits moments of order 1 for all $i \in [d]$. Then from the same arguments as in part 2., $\mathbf{M} = (\mathbf{I} - \mathrm{diag}(m_{ii}))\overline{\mathbf{M}} + \mathrm{diag}(m_{ii})$ and $m_{ii} < 1$ for all $i \in [d]$. Assume that $\mathbf{M}$ is surpercritical, then there is a positive vector x such that $Mx > x$. Therefore, $(\mathbf{I} - \mathrm{diag}(m_{ii}))\overline{\mathbf{M}}x > (\mathbf{I} - \mathrm{diag}(m_{ii}))x$ and since $m_{ii} < 1$, we obtain $\overline{\mathbf{M}}x > x$. Hence $\overline{\mathbf{M}}$ is supercritical. Conversely, assume that $\overline{\mathbf{M}}$ is supercritical. Then there is a positive vector x such that $\overline{\mathbf{M}}x > x$, so that $Mx = (\mathbf{I} - \mathrm{diag}(m_{ii}))\overline{\mathbf{M}}x + \mathrm{diag}(m_{ii})x > (\mathbf{I} - \mathrm{diag}(m_{ii}))x + \mathrm{diag}(m_{ii})x = x$ and thus $\mathbf{M}$ is supercritical. Then the identity $\mathbf{M} = (\mathbf{I} - \mathrm{diag}(m_{ii}))\overline{\mathbf{M}} + \mathrm{diag}(m_{ii})$ allows us to derive that $\mathbf{M}$ is critical if and only if this is the case for $\overline{\mathbf{M}}$.

Finally assume that $m_{ii} > 1$ for some $i \in [d]$. If $m_{ij} = 0$, for all $j \neq i$, then it is clear that individuals of type i in $\overline{F}$ are leaves. If $m_{ij} > 0$, for some $j \in [d]$, then since clusters of type i are supercritical, some of them have infinitely many children with positive probability. Conditionally to this event, such a cluster produces almost surely infinitely many children of type j, which is equivalent to say that individuals of type i in $\overline{F}$ give birth to an infinite number of children of type j with positive probability. □

Let us now consider a multitype branching forest F with progeny distribution ν, with a finite number of trees and let $Z_n = (Z_n^{(1)}, \ldots, Z_n^{(d)})$, $n \geq 0$ be the associated branching process, that is for each $i \in [d]$, $Z_n^{(i)}$ is the total number of individuals of type i present in

F at generation n. For $x = (x_1, \dots, x_d) \in \mathbb{Z}_+^d$, we denote by $\mathbb{P}_x$ the law on $(\Omega, \mathcal{F})$ under which F is issued from x. In particular, $\mathbb{P}_x(Z_0 = x) = 1$. Then the next result gives the law of the total number of mutations in the forest F, that is the number of mutations up to the last generation whose rank is the extinction time, $T := \inf\{n : Z_n = 0\}$. For $i, j \in [d]$, let us denote by M_i the total number of mutations of type i in F, up to time T and by M_{ij} the total number of mutations of type j produced by individuals of type i. In particular, $M_{ii} = 0$ and M_i and M_{ij} satisfy the relations

$$M_j = \sum_{i=1}^{d} M_{ij}\,, \quad j \in [d].$$

Note that if ν is primitive and supercritical, then $\mathbb{P}_x(T = \infty) > 0$ for all $x \in \mathbb{Z}_+^d$, so that under $\mathbb{P}_x$, M_i and M_{ij} are infinite with positive probability, for some $i, j \in [d]$. We also emphasize that M_i and M_{ij} are not functionals of the branching process (Z_n).

Corollary 2.2.2 : Assume that (A_i) or (B_i) holds for all $i \in [d]$. Then for all integers $x_i, n_i, k_{ij}, i, j \in [d]$, such that $x_i \geq 0$, $n_i = -k_{ii}$, for $i \neq j$, $k_{ij} \geq 0$, and for all $j \in [d]$, $n_j = x_j + \sum_{i \neq j} k_{ij}$,

$$\begin{aligned}&\mathbb{P}_{\mathbf{x}}\left(M_1 = n_1 - x_1, \dots, M_d = n_d - x_d, M_{ij} = k_{ij},\ i, j \in [d],\ i \neq j\right)\\ &\quad = \frac{\det(K)}{\bar{n}_1 \dots \bar{n}_d} \prod_{i=1}^{d} \mu_i^{*n_i}(k_{i1}, \dots, k_{i(i-1)}, 0, k_{i(i+1),\dots,k_{id}}),\end{aligned}$$

where μ_i is defined in Theorem 2.2.1 and $\mu_i^{*0} = \delta_0$, $\bar{n}_i = n_i \vee 1$, K is the matrix $(-k_{ij})_{i,j}$ to which we removed the line i and the column i for all i such that $n_i = 0$.

Démonstration. This result is a direct consequence of Theorem 1.2 in [CL15] and Theorem 2.2.1 applied to the forest of mutations $\overline{F}$. Indeed, it suffices to note that $x_i + M_i$ corresponds to the total number of individuals of type i in $\overline{F}$. Note however that Theorem 1.2 in [CL15] is proved only in the case where ν is primitive and (sub)critical. But using the coding which is presented in Section 2.2.1 and appyling Lemma 1 in [Cha15], we can check that it is still valid in the general case by following along the lines the proof which is given in [CL15]. □

If for some $i \in [d]$, none of the conditions (A_i) and (B_i) holds, then the definition of the vector of mutation sizes $(M_1, \dots, M_d)$ still makes sense. In this case, it is possible to obtain its law by extending Theorem 2.2.1 to branching forests whose progeny laws give mass to infinity. Note also that Corollary 2.2.2 can be considered as an extension of Theorem 1 in [Ber09], where a similar formula is given in the neutral case.

We now turn our attention to the asymptotic behaviour of the number of mutations, when the total population is growing to infinity. Our first result is concerned with the critical case and is a direct consequence of Proposition 2 in [Pé16] and Theorem 2.2.1. If $\mathbf{M}$ is primitive, then we denote by u and v the unique right and left positive eigenvectors of $\mathbf{M}$ which are associated to the eigenvalue 1 and normalized by $u.1 = u.v = 1$. Recall that, for a multitype branching forest F, when no confusion is possible, N_i denotes the total population of type i in F and M_i denotes the total number of mutations of type i in F. Note also that when ν is primitive and critical, then (A_i) necessarily holds for all $i \in [d]$, so that from Theorem 2.2.1, the forest of mutations $\overline{F}$ associated to F is a branching forest with progeny distribution μ defined by (2.4).

Corollary 2.2.3 : Let F be a branching forest with a non singular, primitive and critical progeny distribution ν. Assume that for all $i \in [d]$, μ_i admits moments of order $d+1$. If moreover $\overline{\mathbf{M}}$ is primitive and the covariance matrices Σ^i, $\overline{\Sigma}^i$ of ν_i and μ_i, respectively are positive definite. Then $m_{ii} < 1$, for all $i \in [d]$ and there are constants $C_1, C_2 > 0$ such that for all $\mathbf{x}_0 \in \mathbb{Z}_+^d$,

$$\begin{aligned} \lim_{n\to\infty} n^{d/2+1}\mathbb{P}_{\mathbf{x}_0}(M_i = \lfloor n(1-m_{ii})v_i \rfloor,\ i \in [d]) &= C_1 \mathbf{x}_0.\mathbf{u}\,, \\ \lim_{n\to\infty} n^{d+1}\mathbb{P}_{\mathbf{x}_0}(M_i = \lfloor n(1-m_{ii})v_i \rfloor,\ N_i = \lfloor nv_i \rfloor,\ i \in [d]) &= C_2 \mathbf{x}_0.\mathbf{u}\,. \end{aligned}$$

Démonstration. Since by assumption, $\overline{\mathbf{M}}$ is primitive, then for all i, there is $j \neq i$ such that $\bar{m}_{ij} > 0$, and hence $m_{ij} > 0$. Therefore, from part 1. of Theorem 2.2.1, $m_{ii} < 1$, for all i. Moreover, from our assumptions and part 3. of Theorem 2.2.1, μ is critical. Besides, it is plain that $\overline{\mathbf{M}}$ is non singular. Then conditions of Proposition 2 in [Pé16]

are satisfied for the multitype branching process associated to $\overline{F}$ and the first assertion follows with $\bar{u}$ and $\bar{v}$, the normalized, positive right and left eigenvectors of $\overline{\mathbf{M}}$ associated to the eigenvalue 1. Then recall from the proof of part 3. of Theorem 2.2.1 that $\mathbf{M} = (\mathbf{I} - \mathrm{diag}(m_{ii}))\overline{\mathbf{M}} + \mathrm{diag}(m_{ii})$. We derive from this identity that $\bar{u} = u$ and $\bar{v} = cv(\mathbf{I} - \mathrm{diag}(m_{ii}))$, where $c = \|u \cdot v(\mathbf{I} - \mathrm{diag}(m_{ii}))\|^{-1}$ and the first assertion follows.

The proof of the second assertion follows the same lines as the proof of Proposition 2 in [Pé16]. In this case, since the number of mutations is taken into account together with the total number of individuals, a $2d$-dimensional random walk is involved in the proof, which explains that the rate of convergence in now $d + 1$. □

Note that the constants C_1 and C_2 can be made explicit in terms of the distributions ν and μ by properly exploiting the proof of Proposition 2 in [Pé16].

Through the next result we focus on the asymptotic behaviour of the number of mutations in a branching forest when the initial number of individuals $x = (x_1, \ldots, x_d)$ tends to infinity along some given direction.

Theorem 2.2.4 : Let $F(\mathbf{x})$ be any family of multitype branching forests defined on the space $(\Omega, \mathcal{F}, P)$, indexed by $\mathbf{x} \in \mathbb{Z}_+^d$ and such that for each $\mathbf{x}$, $F(\mathbf{x})$ has progeny distribution ν and is issued from $\mathbf{x}$. For $i \in [d]$, let $N_i(\mathbf{x})$ (resp. $M_i(\mathbf{x})$) be the total number of individuals (resp. of mutations) of type i in $F(\mathbf{x})$. Assume that ν is primitive and let $\mathbf{w} \in \mathbb{Z}_+^d \setminus \{0\}$.

1. If ν is critical, then

$$\lim_{n\to\infty} \frac{N_i(n\mathbf{w})}{n} = \infty \quad \text{and} \quad \lim_{n\to\infty} \frac{M_i(n\mathbf{w})}{N_i(n\mathbf{w})} = 1 - m_{ii}\,, \quad \text{in probability.}$$

2. If ν is subcritical, then

$$\lim_{n\to\infty} \frac{N_i(n\mathbf{w})}{n} = c_i(\mathbf{w}) \quad \text{and} \quad \lim_{n\to\infty} \frac{M_i(n\mathbf{w})}{n} = w_i + (1 - m_{ii})c_i(\mathbf{w})\,, \quad \text{in probability,}$$

where $c_i(\mathbf{w}) := \sum_{k=1}^{d} w_k(\mathbf{I} - \mathbf{M})^{-1}_{ki}$.

In any case, $m_{ii} < 1$, for all $i \in [d]$.

Démonstration. In order to prove our result, it suffices to construct some particular family of forests $F(\mathbf{x})$, such that for each $\mathbf{x}$, $F(\mathbf{x})$ has progeny distribution ν and is issued from $\mathbf{x} \in \mathbb{Z}_+^d$, and to show that the limits in the statement hold.

Recall the coding of multitype branching forests which is presented at the end of Section 2.2.2 and let $X^{(i)} = \{X^{i,j}, j \in [d]\}$ be d independent random walks whose respective step distributions are $\tilde{\nu}_i$, $i \in [d]$ defined in (2.2). Then for each $\mathbf{x} \in \mathbb{Z}_+^d$, we construct a forest $F(\mathbf{x})$ such that $F(\mathbf{x})$ is encoded by the random walks $X^{(i)}$, $i \in [d]$ and contains exactly x_i trees whose root is of type i. This construction is possible in the primitive, (sub)critical case, thanks to part 3. of Theorem 3.1 in [CL15].

Then $N_i(\mathbf{x})$ and $X^{(i)}$, $i \in [d]$, satisfy identity (2.3). Moreover, for $k \neq i$, the number of mutations of type i issued from an individual of type k is $X^{k,i}(N_k(\mathbf{x}))$, so that the total number of mutations of type i is

$$M_i(\mathbf{x}) = \sum_{k \neq i} X^{k,i}(N_k(\mathbf{x})) = -x_i - X^{i,i}(N_i(\mathbf{x}))\,.$$

We derive from Lemma 2.2 in [CL15], that if $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{Z}_+^d$ are such that $\mathbf{x}_1 \leq \mathbf{x}_2$, then the couple of random variables $(N_i(\mathbf{x}_2) - N_i(\mathbf{x}_1), X^{i,i}(N_i(\mathbf{x}_2) - X^{i,i}(N_i(\mathbf{x}_1)))$ is independent of process $((N_i(\mathbf{x}), X^{i,i}(N_i(\mathbf{x}))), \mathbf{x} \leq \mathbf{x}_1)$ and has the same law as $(N_i(\mathbf{x}_2 - \mathbf{x}_1), X^{i,i}(N_i(\mathbf{x}_2 - \mathbf{x}_1))$. Therefore, for any $\mathbf{w} \in \mathbb{Z}_+^d$, $((N_i(n\mathbf{w}), X^{i,i}(N_i(n\mathbf{w})), n \geq 0)$ is a bivariate random walk whose step distribution is the law of $(N_i(\mathbf{w}), X^{i,i}(N_i(\mathbf{w}))$.

Let $Z = (Z^{(1)}, \ldots, Z^{(d)})$ be the branching process associated to $F(\mathbf{w})$. Then by definition of $N_i(\mathbf{w})$, we have $N_i(\mathbf{w}) = \sum_{n=0}^{\infty} Z_n^{(i)}$. But $\mathbb{E}_{\mathbf{w}}(Z_n) = \mathbf{w}\mathbf{M}^n$, so that $\mathbb{E}_{\mathbf{w}}(Z_n^{(j)}) = \sum_{i=1}^d w_i m_{ij}^{(n)}$ and since ν is primitive, we have from Frobenius Theorem for primitive matrices, $m_{ij}^{(n)} \sim u_i v_j \rho^n$, see Theorem 1, Section V.2 in [AN72]. So we have proved that $\mathbb{E}(N_i(\mathbf{w})) < \infty$ if and only if ν is subcritical. Moreover, if ν is subcritical, then $\mathbf{I} - \mathbf{M}$ is invertible and it follows from the above expressions that $\mathbb{E}(N_i(\mathbf{w})) = \sum_{i=1}^d w_i (\mathbf{I} - \mathbf{M})_{ij}^{-1}$. Then assertions 1. and 2. follow directly from the law of large numbers.

Finally, since ν and μ are primitive, by definition, they admit moments of order 1 and

we derive from part 1. of Theorem 2.2.1 that $m_{ii} < 1$, for all $i \in [d]$. □

2.3 When continuous time is involved

2.3.1 The Lamperti representation

Let us now consider a d type population which is composed at time $t = 0$, of x_i individuals of type $i \in [d]$ and whose dynamics in continuous time behave according to a branching model. More specifically, at any time, all individuals in the population live, give birth and die independently of each other. Once it is born, any individual of type $i \in [d]$ gives birth after an exponential time with parameter $\lambda_i > 0$ to n_j individuals of type $j \in [d]$ with probability $\nu_i(n_1, \ldots, n_d)$. Then this individual dies at the same time it gives birth. We emphasize that in this model, the probability for the population to become extinct does not depend on the rates λ_i.

This model is represented as a plane forest with edge lengths, see Figure 2.3. (In each sibling, we rank individuals of type 1 to the left, then individuals of type 2, and so on.) Such a forest will be called a multitype branching forest with edge lengths issued from $\mathbf{x} = (x_1, \ldots, x_d)$, with progeny distribution $\nu := (\nu_1, \ldots, \nu_d)$ and reproduction rates $(\lambda_1, \ldots, \lambda_d)$. By construction, its discrete time skeleton is a multitype branching (plane) forest, as defined in the previous section, with progeny distribution ν, which is independent from the edge lengths. Edge lengths are independent between themselves and the length of an edge issued from a vertex of type i follows an exponential distribution with parameter λ_i. We emphasize that the total number of individuals and the total number of mutations in a multitype branching forest with edge lengths are the same as in its discrete skeleton. Hence, the results of the previous section can be applied in the present setting.

Given a branching forest with edge lengths, as defined above, we denote by $Z =$

$(Z^{(1)}, \dots, Z^{(d)})$ the corresponding multitype branching process, that is for $t \geq 0$ and $i \in [d]$, $Z_t^{(i)}$ is the number of individuals of type i at time t in the population. (Since no confusion is possible, for the branching process we have kept the same notation as in discrete time.) The process Z is a $\mathbb{Z}_+^d$-valued continuous time Markov process which satisfies the branching property, i.e., for $\lambda \in \mathbb{R}_+^d$, $t \geq 0$ and $\mathbf{x}, \mathbf{y} \in \mathbb{Z}_+^d$,

$$\mathbb{E}_{\mathbf{x}+\mathbf{y}}(e^{-\lambda Z_t}) = \mathbb{E}_{\mathbf{x}}(e^{-\lambda Z_t})\mathbb{E}_{\mathbf{y}}(e^{-\lambda Z_t}),$$

where $\mathbb{P}_{\mathbf{x}}$ is the law under which the forest is issued from $\mathbf{x}$. In particular, $Z_0 = \mathbf{x}$, $\mathbb{P}_{\mathbf{x}}$-a.s. The process Z actually contains much less information than the original branching forest. In order to preserve the essential part of this information, we need to decompose Z as in the following definition.

Definition 2.3.1 : For $i \neq j$, we denote by $Z_t^{i,j}$ the total number of individuals of type j whose parent has type i and who were born before time t. For $i = j$, the definition of $Z_t^{i,i}$ is the same, except that to this number we add the number of individuals of type i at time 0 and we subtract the number of individuals of type i who died before time t.

The processes $Z^{i,j}$ whose definition should be clear from the example given in Figure 2.3 will play a crucial role in our continuous time model. A more formal definition can be found in Section 4.2 of [Cha15]. The interest of these processes is the following straightforward decomposition of the branching process $Z = (Z^{(1)}, \dots, Z^{(d)})$:

$$Z_t^{(j)} = \sum_{i=1}^{d} Z_t^{i,j}, \quad j \in [d]. \tag{2.6}$$

Our model bears on a Lamperti type representation of these processes. According to Lamperti representation, any one dimensional branching process can be expressed as a Lévy process time changed by some integral functional. In this subsection, we will recall from [Cha15] the extension of this transformation to multitype, continuous time, discrete valued branching processes. The latter involves time changed multidimensional compound Poisson processes which we now introduce.

Since our models of evolution are only concerned with mutations, individuals of type i having exactly one child of type i do not present any interest. Hence we can assume without loss of generality that

$$\nu_i(\mathbf{e}_i) = 0, \text{ for all } i \in [d].$$

Then let $X = (X^{(1)}, \dots, X^{(d)})$, where $X^{(i)}$, $i \in [d]$ are d independent $\mathbb{Z}^d$-valued compound Poisson processes. We assume that $X_0^{(i)} = 0$ and that $X^{(i)}$ has rate λ_i and jump distribution $\tilde{\nu}_i$ which has been defined in (2.2). In particular, with the notation $X^{(i)} = (X^{i,1}, \dots, X^{i,d})$, the process $X^{i,i}$ is a $\mathbb{Z}$-valued, downward skip free, compound Poisson process, i.e. $\Delta X_t^{i,i} = X_t^{i,i} - X_{t-}^{i,i} \geq -1, t \geq 0$, with $X_{0-} = 0$ and for all $i \neq j$, the process $X^{i,j}$ is an increasing compound Poisson process. We emphasize that in this definition, some of the processes $X^{i,j}$, $i, j \in [d]$ can be identically equal to 0.

The following extension of the Lamperti representation to multitype branching processes can be found in [Cha15], see also [CPGUB16] for the case of continuous state multitype branching processes.

Theorem 2.3.1 : Let us consider a multitype branching forest with edge lengths issued from $\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{Z}_+^d$, with progeny distribution $\nu := (\nu_1, \dots, \nu_d)$ and reproduction rates $(\lambda_1, \dots, \lambda_d)$. Then the processes $Z^{i,j}$, $i, j \in [d]$ introduced in Definition 2.3.1 admit the following representation :

$$Z_t^{i,j} = \begin{cases} X^{i,j}_{\int_0^t Z_s^{(i)}\, ds}, & t \geq 0\,, \quad \text{if } i \neq j, \\ x_i + X^{i,i}_{\int_0^t Z_s^{(i)}\, ds}, & t \geq 0\,, \quad \text{if } i = j, \end{cases} \tag{2.7}$$

where the processes,

$$X^{(i)} = (X^{i,1}, X^{i,2}, \dots, X^{i,d})\,, \quad i = 1, \dots, d\,,$$

are independent $\mathbb{Z}_+^d$ valued compound Poisson processes, with jump distribution $(\tilde{\nu}_1, \dots, \tilde{\nu}_d)$ and rates $(\lambda_1, \dots, \lambda_d)$. In particular from (2.6) and (2.7), the multitype branching pro-

cess Z admits the following representation,

$$(Z_t^{(1)}, \dots, Z_t^{(d)}) = \mathbf{x} + \left(\sum_{i=1}^{d} X^{i,1}_{\int_0^t Z_s^{(i)}\, ds}, \dots, \sum_{i=1}^{d} X^{i,d}_{\int_0^t Z_s^{(i)}\, ds} \right), \quad t \geq 0\,. \tag{2.8}$$

2.3.2 Further results on asymptotics of mutations

For $i \in [d]$ and $t \geq 0$, we will denote by $M_{i,t}$ the total number of mutations of type i which occured up to time t. The definition of this quantity is illustrated on Figure 2.3. Let us also define a cluster of type i as the subtree corresponding to the descendence of type i of an individual of type i which is either a root or an individual whose parent as a type different from i. Then $x_i + M_{i,t}$ corresponds to the number of clusters of type i in the forest truncated at time t.

In Proposition 2.3.3, we describe the asymptotic behaviour of $M_{i,t}$, as t tends to ∞ in the case where the progeny distribution ν is primitive and supercritical. To this aim, we will need the joint representation of $M_{i,t}$ together with the number $Z_t^{(i)}$ of individuals of type i at time t which is presented in Proposition 2.3.2.

Proposition 2.3.2 : Recall from Section 2.3.1 the definition of the compound Poisson processes $X^{i,j}$, $i, j \in [d]$. Then for any $\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{Z}_+^d$, under $\mathbb{P}_\mathbf{x}$, the stochastic process $\left(Z_t^{(i)}, M_{i,t}\right)$ fulfills the following representation,

$$\left(Z_t^{(i)}, M_{i,t}\right) = \left(x_i + \sum_{k=1}^{d} X^{k,i}_{\int_0^t Z_u^{(k)}\, \mathrm{d}u}, \sum_{k=1, k\neq i}^{d} X^{k,i}_{\int_0^t Z_u^{(k)}\, \mathrm{d}u} \right), \quad t \geq 0.$$

Démonstration. This result is a direct consequence of the representation which is recalled in Theorem 2.3.1. Indeed, recall from Section 2.3.1 the definition of $Z^{i,j}$, then the number of mutations of type i up to time t is

$$M_{i,t} = \sum_{k \neq i} Z_t^{k,i}\,.$$

The result follows from identity (2.7) in Theorem 2.3.1. □

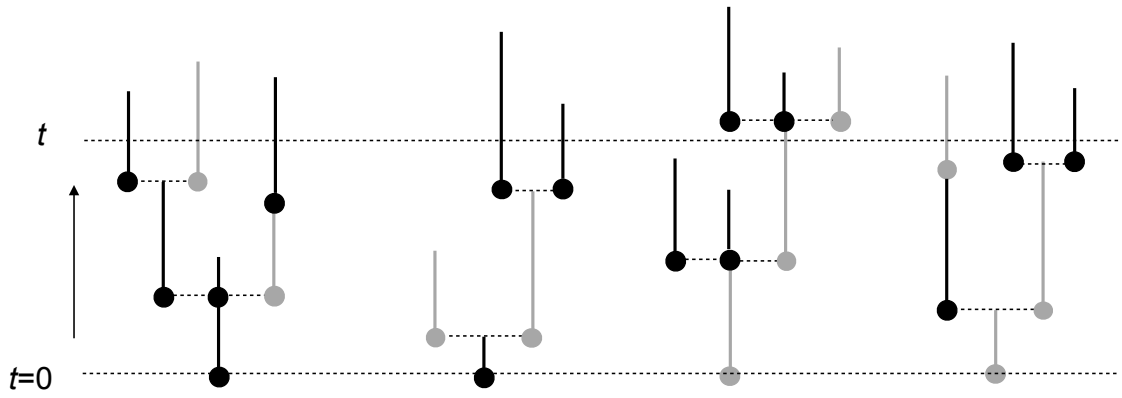

FIGURE 2.3 – A two type forest with edge lengths issued from $\mathbf{x} = (2, 2)$. Vertices of type 1 (resp. 2) are represented in black (resp. grey). At time t,

$$Z_t^{(1)} = 6, \quad Z_t^{(2)} = 3, \quad Z_t^{1,1} = -2, \quad Z_t^{1,2} = 5, \quad Z_t^{2,1} = 8, \quad Z_t^{2,2} = -2,$$

$$\text{and} \quad M_{1,t} = 8, \quad M_{2,t} = 5.$$

Let's us now turn to the limiting behavior of $M_{i,t}$, as t tends to infinity. The next result is concerned with the case where ν is primitive and supercritical. It allows us to evaluate the number of mutations which occured up to time t (or equivalently the number of clusters in the forest truncated at time t), when t is large.

Let us define the matrix $A = \Lambda(\mathbf{M} - \mathbf{I})$, where $\Lambda = \text{diag}(\lambda_i)$. If $\mathbf{M}$ is primitive, then so is A and it follows from Perron-Frobenius theory that the eigenvalues ρ_i, $i \in [d]$ of A can be arranged so that $\rho_1 > \text{Re}(\rho_2) \geq \cdots \geq \text{Re}(\rho_d)$. Moreover, ν is subcritical, critical or supercritical according as $\rho_1 < 0$, $\rho_1 = 0$ or $\rho_1 > 0$. Then a well known result due to [Ath68], see also Theorem 2, p. 206 in [AN72] asserts that when ν is non singular and primitive, there exists a nonnegative random variable W such that for all $i \in [d]$,

$$\lim_{t\to\infty} e^{-\rho_1 t} Z_t^{(i)} = v_i W \ , \quad \text{a.s.}, \tag{2.9}$$

where v_i is the i-th coordinate of the normalized left eigenvector associated with ρ_1.

Proposition 2.3.3 : Assume that ν is non singular, primitive and supercritical. Then for all $i \in [d]$,

$$\lim_{t\to\infty} e^{-\rho_1 t} M_{i,t} = K_i W, \ a.s.,$$

where $K_i = v_i(1 + (1 - m_{ii})(\lambda_i \rho_1)^{-1})$.

Démonstration. We derive from Proposition 2.3.2 that,

$$Z_t^{(i)} - M_{i,t} = X^{i,i}_{\int_0^t Z_u^{(i)} \, \mathrm{d}u}, \quad \text{a.s.}$$

On the other hand, in the supercritical case, ρ_1 is strictly positive. Hence it follows from (2.9) that

$$\int_0^t Z_u^{(i)} \, \mathrm{d}u \sim \rho_1^{-1} W v_i e^{\rho_1 t}, \ a.s., \quad as \ t \to \infty.$$

Then the desired result is a consequence of the latter equivalence and the law of large numbers applied to the compound Poisson process □

Under conditions of Propositon 2.3.3, assume moreover that for some $i \in [d]$, K_i is positive, that is

$$m_{ii} < 1 + \lambda_i \rho_1 \, ,$$

and that for some j, $\mathbb{P}_{\mathbf{e}_j}(W > 0) = 1$. Then using Proposition 2.3.3, we can compare the asymptotic behaviour of the number of mutations prior to t with this of $Z_t^{(i)}$, under $\mathbb{P}_{\mathbf{e}_j}$, that is

$$M_{i,t} \sim K_i Z_t^{(i)} \, , \quad \mathbb{P}_{\mathbf{e}_j}\text{-a.s., as } t \to \infty. \tag{2.10}$$

Regarding the condition $\mathbb{P}_{\mathbf{e}_j}(W > 0) = 1$, note that Theorem 2, p. 206 in [AN72] also asserts that $\mathbb{P}_{\mathbf{e}_k}(W > 0) > 0$, for some (hence for all) $k \in [d]$, if and only if

$$\mathbb{E}(\xi_{ij} \log \xi_{ij}) < \infty, \text{ for all } i, j \in [d], \tag{2.11}$$

where $(\xi_{i1}, \ldots, \xi_{id})$ is a random vector with law ν_i. Moreover, $1 - \mathbb{P}_{\mathbf{e}_k}(W > 0)$ corresponds to the probability of extinction, when the forest is issued from $\mathbf{e}_k$.

2.3.3 Emergence times of mutations

In this section, we shall assume that mutations are not reversible, that is for all $i = 1, \dots, d-1$, individuals of type i can only have children of type i or $i+1$. In particular ν is not irreducible. Moreover when giving birth, individuals of type $i = 1, \dots, d-1$ have at least one child of type i with probability one, and have children of type $i+1$ with positive probability. These conditions can be explicited in terms of the progeny distribution ν_i as follows

$$\begin{cases} \nu_i(\mathbf{k}) > 0 \;\Rightarrow\; k_j = 0, \text{ for } j \notin \{i, i+1\}, \\ \sum_{\mathbf{k}\in\mathbb{Z}_+^d : k_i = 0} \nu_i(\mathbf{k}) = 0 \;\text{ and }\; \sum_{\mathbf{k}\in\mathbb{Z}_+^d : k_{i+1} = 0} \nu_i(\mathbf{k}) < 1. \end{cases} \tag{2.12}$$

We are interested in the waiting time until an individual of type i first emerges in the population, that is

$$\tau_i := \inf\{t \geq 0 : Z_t^{(i)} \geq 1\}\,.$$

The problem of determining a general expression for the law of τ_i is quite challenging. As far as we know, there is no explicit expression for this law in terms of the progeny distribution and the reproduction rates. Various results in this direction can be found in [Ser06], [SH07], [DM10] and [Ale13] for instance. Most of them provide approximations of this law, using martingale convergence theorems [DM10] or through numerical methods for the inversion of the generating function [Ale13]. In Proposition 2.3.4 we first give a relationship between the successive emergence times $\tau_2, \tau_3, \dots$ in terms of the underlying compound Poisson process in the Lamperti representation of Z. We also characterize the joint law under $\mathbb{P}_{\mathbf{e}_{i-1}}$ of the time τ_i and the number of individuals of type $i-1$ at this time. In Theorem 2.3.6 we derive an approximation of the time τ_i, under $\mathbb{P}_{\mathbf{e}_1}$, as the mutation rate of type k increases faster than that of type $k-1$, for all $k = 2, \dots, i$. Then in Corollary 2.3.7 we focus on a case where these law can be explicited.

In the following developments, we use the notation of Section 2.3.1 from which we recall the Lamperti representation of the multitype branching process $Z = (Z^{(1)}, \dots, Z^{(d)})$

in terms of the compound Poisson processes $X^{(i)}$. Let us also introduce a few more notation. For $i, j \in [d]$, we denote by $\lambda_{i,j}$ the parameter of the compound Poisson process $X^{i,j}$, that is

$$\lambda_{i,j} := \lambda_i \left(1 - \textstyle\sum_{\mathbf{k}\in\mathbb{Z}_+^d : k_j = 0} \tilde{\nu}_i(\mathbf{k})\right).$$

Note that from our assumptions (2.12), for all $i = 1, \dots, d-1$, $\lambda_{i,i+1} > 0$ and for $j \notin \{i, i+1\}$, $\lambda_{i,j} = 0$, that is $X^{i,j}$ is identically equal to 0. In particular, $\lambda_i = \lambda_{i,i} + \lambda_{i,i+1}$, for $i \leq d-1$ and $\lambda_d = \lambda_{d,d}$. The parameter $\lambda_{i,i+1}$ will be call the mutation rate of type $i+1$. For $i \geq 2$, let

$$\gamma_i := \inf\{t : X_t^{i-1,i} \geq 1\}$$

be the time of the first jump by the process $X^{i-1,i}$ and note that this time is exponentially distributed with parameter $\lambda_{i-1,i}$.

Proposition 2.3.4 : Assume that (2.12) holds and define $Z^{0,1}$ as the process identically equal to 1 and set $\tau_1 = 0$.

1. For $i = 2, \dots, d$, the emergence time τ_i of type i admits the following representation under $\mathbb{P}_{\mathbf{e}_1}$,

$$\tau_i = \tau_{i-1} + \int_0^{\gamma_i} \frac{1}{X_s^{i-1,i-1} + Z_{\kappa_{i-1}(s)}^{i-2,i-1}} \, \mathrm{d}s, \quad \mathbb{P}_{\mathbf{e}_1}\text{-a.s.}, \tag{2.13}$$

   where κ_{i-1} is the right continuous inverse of the functional $t \mapsto \int_0^t Z_s^{(i-1)} \, \mathrm{d}s$, i.e. $\kappa_{i-1}(t) = \inf\{s > 0 : \int_0^s Z_u^{(i-1)} \, \mathrm{d}u > t\}$.

2. Under $\mathbb{P}_{\mathbf{e}_{i-1}}$, the joint law of the emergence time τ_i of type i together with the number of individuals of type $i-1$ in the population at time τ_i admits the following representation,

$$(\tau_i, Z_{\tau_i}^{(i-1)}) \overset{(\mathrm{d})}{=} \left(\int_0^{\gamma_i} \frac{\mathrm{d}s}{1 + X_s^{i-1,i-1}}, 1 + X_{\gamma_i}^{i-1,i-1}\right). \tag{2.14}$$

3. Let us define $\theta_k = \int_0^{\gamma_k} \frac{1}{X_s^{k-1,k-1}+1} \, \mathrm{d}s$, for $k \geq 2$.Then the random variables θ_k, $k \geq 2$ are independent and for $i = 2, \dots, d$,

$$\mathbb{P}_{\mathbf{e}_1}(\tau_i > t) \leq P\left(\sum_{k=2}^{i} \theta_k > t\right), \quad \text{for all } t > 0\,. \tag{2.15}$$

Démonstration. Since $X^{i,j}$ is identically equal to 0 whenever $j \notin \{i, i+1\}$, then under $\mathbb{P}_{\mathbf{e}_1}$, the representation (2.8) admits the simpler form

$$(Z_t^{(1)}, \ldots, Z_t^{(d)}) = \mathbf{e}_1 + \left(X^{1,1}_{\int_0^t Z_s^{(1)}\, ds}, X^{2,2}_{\int_0^t Z_s^{(2)}\, ds} + X^{1,2}_{\int_0^t Z_s^{(1)}\, ds}, \ldots, X^{d,d}_{\int_0^t Z_s^{(d)}\, ds} + X^{d-1,d}_{\int_0^t Z_s^{(d-1)}\, ds} \right). \tag{2.16}$$

In particular, for $i = 2, \ldots, d$,

$$Z_t^{(i)} = X^{i,i}_{\int_0^t Z_s^{(i)}\, ds} + X^{i-1,i}_{\int_0^t Z_s^{(i-1)}\, ds}, \quad t \geq 0.$$

Since $X_0^{i,i} = 0$, for $i \geq 2$, we see that the time τ_i corresponds to the first hitting time of level 1 by the process $t \mapsto X^{i-1,i}_{\int_0^t Z_s^{(i-1)}\, ds}$, that is

$$\tau_i = \kappa_{i-1}(\gamma_i), \tag{2.17}$$

where γ_i has been defined as the time of the first jump of the process $X^{i-1,i}$. For t such that $\kappa_{i-1}(t) < \infty$, we have $t = \int_0^{\kappa_{i-1}(t)} Z_s^{(i-1)}\, \mathrm{d}s$, so that $\mathrm{d}t = Z^{(i-1)}_{\kappa_{i-1}(t)} \mathrm{d}\kappa_{i-1}(t)$, and since $\kappa_{i-1}(0) = \tau_{i-1}$, we obtain

$$\begin{aligned} \kappa_{i-1}(t) &= \tau_{i-1} + \int_0^t \frac{\mathrm{d}s}{Z^{(i-1)}_{\kappa_{i-1}(s)}} \\ &= \tau_{i-1} + \int_0^t \frac{\mathrm{d}s}{X_s^{i-1,i-1} + X^{i-2,i-1}_{\int_0^{\kappa_{i-1}(s)} Z_u^{(i-2)}\, \mathrm{d}u}}. \end{aligned} \tag{2.18}$$

The latter identity together with (2.17) prove identity (2.13).

The second part of the proposition is easily derived from the same arguments. More specifically, it follows from (2.17) and the following identities

$$Z_t^{(i-1)} = 1 + X^{i-1,i-1}_{\int_0^t Z_s^{(i-1)}\, ds} \quad \text{and} \quad \kappa_{i-1}(t) = \int_0^t \frac{\mathrm{d}s}{1 + X_s^{i-1,i-1}}, \quad t \geq 0,$$

which hold $\mathbb{P}_{\mathbf{e}_{i-1}}$-a.s.

Independence between the variables θ_k, $k \geq 2$ is a direct consequence of the independence between the processes $X^{(i)}$, $i \in [d]$. We derive from the representation of τ_i in part 1. of this proposition that

$$\tau_i = \sum_{k=2}^{i} \int_0^{\gamma_k} \frac{1}{X_s^{k-1,k-1} + X^{k-2,k-1}_{\int_0^{\kappa_{k-1}(s)} Z_u^{(k-2)}\, \mathrm{d}u}}\, \mathrm{d}s, \quad \text{a.s.} \tag{2.19}$$

Note that since $\kappa_{k-1}(0) = \tau_{k-1}$, then from (2.17), for all $k \geq 2$, $\int_0^{\kappa_{k-1}(0)} Z_u^{(k-2)}\,\mathrm{d}u = \gamma_{k-1}$, so that by definition of γ_{k-1},

$$X^{k-2,k-1}_{\int_0^{\kappa_{k-1}(0)} Z_u^{(k-2)}\,\mathrm{d}u} = X^{k-2,k-1}_{\gamma_{k-1}} \geq 1, \quad \text{a.s.} \tag{2.20}$$

Besides, since $s \mapsto X^{k-2,k-1}_{\int_0^{\kappa_{k-1}(s)} Z_u^{(k-2)}\,\mathrm{d}u}$ are increasing processes, then inequality (2.15) is a direct consequence of identities (2.19) and (2.20). □

Note that the law of θ_k or equivalently, the law of τ_k under $\mathbb{P}_{\mathbf{e}_{k-1}}$ can be made explicit in some instances through its Laplace transform, see Corollary 2.3.7 below.

For the remainder of this section we will assume moreover that at each mutation, individuals of type i do not give birth to more than one child of type $i+1$ in a same litter. More specifically, assumptions (2.12) are replaced by,

$$\begin{cases} \nu_i(\mathbf{k}) > 0 \;\Rightarrow\; k_{i+1} = 0 \text{ or } 1 \text{ and } k_j = 0, \text{ for } j \notin \{i, i+1\}, \\ \sum_{\mathbf{k}\in\mathbb{Z}_+^d : k_i = 0} \nu_i(\mathbf{k}) = 0 \;\text{ and }\; \sum_{\mathbf{k}\in\mathbb{Z}_+^d : k_{i+1}=0} \nu_i(\mathbf{k}) < 1\,. \end{cases} \tag{2.21}$$

In particular, under these assumptions, the process $X^{i,i+1}$ is a standard Poisson process. Then we will need the next lemma in order to derive our main result on the estimation of the time τ_i, as the mutation rates $\lambda_{k-1,k}$, $k = 2, \ldots, d$ grow faster.

Lemma 2.3.5 : Assume that (2.21) holds, let $k \geq 3$ and fix $\lambda_{1,2} > 0$, then

$$\mathbb{P}_{\mathbf{e}_1}\big(X^{k-2,k-1}_{\int_0^{\kappa_{k-1}(\gamma_k)} Z_u^{(k-2)}\,\mathrm{d}u} = 1\big) \longrightarrow 1\,, \quad \text{as } \lambda_{n-2,n-1}/\lambda_{n-1,n} \to 0, \text{ for } n = 3, \ldots, k. \tag{2.22}$$

Démonstration. First set $\gamma^{(1)}_{k-1} = \inf\{t > \gamma_{k-1} : X^{k-2,k-1}_t = 2\}$ and note that

$$\begin{aligned} \{X^{k-2,k-1}_{\int_0^{\kappa_{k-1}(\gamma_k)} Z_u^{(k-2)}\,\mathrm{d}u} = 1\} &= \{\textstyle\int_0^{\kappa_{k-1}(\gamma_k)} Z_u^{(k-2)}\,\mathrm{d}u < \gamma^{(1)}_{k-1}\} \\ &= \{\kappa_{k-1}(\gamma_k) < \kappa_{k-2}(\gamma^{(1)}_{k-1})\}. \end{aligned}$$

It is easy to check that $\kappa_{k-2}(\gamma^{(1)}_{k-1}) = \tau^{(1)}_{k-1}$, where

$$\tau^{(1)}_{k-1} := \inf\{t > \tau_{k-1} : Z^{k-2,k-1}_t - Z^{k-2,k-1}_{\tau_{k-1}} = 1\}\,.$$

(Note that from our assumptions $Z^{k-2,k-1}_{\tau_{k-1}} = 1$ and $Z^{k-2,k-1}_{\tau^{(1)}_{k-1}} = 2$, $\mathbb{P}_{\mathbf{e}_1}$-a.s.) So from (2.17), we have showed that

$$\{X^{k-2,k-1}_{\int_0^{\kappa_{k-1}(\gamma_k)} Z^{(k-2)}_u \,\mathrm{d}u} = 1\} = \{\tau_k < \tau^{(1)}_{k-1}\}\,. \tag{2.23}$$

The event $\{\tau_k < \tau^{(1)}_{k-1}\}$ means that before the first time when an individual of type k appears in the population, there has been only one birth of type $k-1$. From the Markov property applied at time τ_{k-1}, we have

$$\mathbb{P}_{\mathbf{e}_1}(\tau_k \leq \tau^{(1)}_{k-1}) = \int \mathbb{P}_{\mathbf{z}}(\tau_k \leq \tau^{(1)}_{k-1})\mathbb{P}_{\mathbf{e}_1}(Z_{\tau_{k-1}} \in d\mathbf{z})\,. \tag{2.24}$$

The support in the integral of (2.24) is included in the set $\{\mathbf{z} : z_{k-1} = 1\}$, so from (2.23), (2.24) and the Lebesgue theorem of dominated convergence, all we need to prove is

$$\mathbb{P}_{\mathbf{z}}(\tau_k \leq \tau^{(1)}_{k-1}) \to 1, \quad \text{as } \lambda_{n-2,n-1}/\lambda_{n-1,n} \to 0, \text{ for } n = 3, \ldots, k, \tag{2.25}$$

for all $\mathbf{z}$ such that $z_{k-1} = 1$. (Note that if $\mathbf{z}$ is such that $z_1 = \cdots = z_{k-2} = 0$, or such that $z_k \geq 1$, then it is clear that $\mathbb{P}_{\mathbf{z}}(\tau_k \leq \tau^{(1)}_{k-1}) = 1$, since in the first case $Z^{k-2,k-1}$ is identically equal to 0, so that $\tau^{(1)}_{k-1} = \infty$, $\mathbb{P}_{\mathbf{z}}$-a.s. and in the second case, $\tau_k = 0$, $\mathbb{P}_{\mathbf{z}}$-a.s.)

Let $\mathbf{z}$ be such that $z_{k-1} = 1$. Without loss of generality we can assume that $z_i \geq 1$, for $i = 1, \ldots k-2$. For $i = 1, \ldots k-1$, let us denote by U_i the first time that the lineage of one of the z_{k-i} initial individuals of type $k-i$ gives birth to an individual of type $k-i+1$. Then from the branching property, under $\mathbb{P}_{\mathbf{z}}$, the r.v.'s U_i are independent and from part 2. of Proposition 2.3.4, U_i has the same law as $\int_0^{\gamma_{k-i+1}} \frac{\mathrm{d}s}{X^{k-i,k-i}_s + z_{k-i}}$. Then set $Y^{(i)}_s := X^{k-i,k-i}_s + z_{k-i}$ and note the inclusions,

$$\left\{\gamma_k \leq \min\left(\gamma_{k-1}/Y^{(2)}_{\gamma_{k-1}}, \ldots, \gamma_2/Y^{(k-1)}_{\gamma_2}\right)\right\} \subset \{U_1 \leq \min(U_2, \ldots, U_{k-1})\} \subset \{\tau_k \leq \tau^{(1)}_{k-1}\}\,,$$

which imply the inequality,

$$P\big(\gamma_k/\gamma_{k-1} \leq \min\left(1/Y^{(2)}_{\gamma_{k-1}}, \gamma_{k-2}/(\gamma_{k-1}Y^{(3)}_{\gamma_{k-2}}), \ldots, \gamma_2/(\gamma_{k-1}Y^{(k-1)}_{\gamma_2})\right)\big) \leq \mathbb{P}_z(\tau_k \leq \tau^{(1)}_{k-1})\,.$$

But when $\lambda_{n-2,n-1}/\lambda_{n-1,n} \to 0$, for $n = 3, \ldots, k$, the parameter $\lambda_{1,2} > 0$ being fixed, we necessarily have $\lim \lambda_{n-1,n} = \infty$, for $n = 3, \ldots, k$. Hence γ_k/γ_{k-1} converges in probability toward 0, $1/Y^{(2)}_{\gamma_{k-1}}$ converges in probability toward $1/z_{k-2}$ and $\gamma_{n-1}/(\gamma_n Y^{(k-n+2)}_{\gamma_{n-1}})$,

for $n = 3, \dots, k-1$ converge in probability toward $+\infty$. Therefore, the left hand side of the above inequality tends to 1, which proves (2.25) and the lemma is proved. □

The following theorem intuitively means that when $\frac{\lambda_{k-1,k}}{\lambda_{k,k+1}} \to 0$, for $k = 3, \dots, i$, the emergence time τ_i, when starting from an individual of type 1, can be approximated by the sum of independent random variables $\tau_{1,2}+\cdots+\tau_{i-1,i}$, where $\tau_{k-1,k}$ is the emergence time of type k, when starting from an individual of type $k-1$. The assumption $\frac{\lambda_{k-1,k}}{\lambda_{k,k+1}} \to 0$ is quite adapted to several biological models such as cancer growth, for instance. Indeed, cancer is often the result of a series of successive mutations, [IMKN05], [DM10], [Dur13]. Each new mutation is itself more unstable than the previous ones, and in particular, the successive mutation rates can increase very fast. It would interesting to study the asymptotic behavior of τ_i, when $\frac{\lambda_{k,k}}{\lambda_{k+1,k+1}} \to 0$, that is when the intrinsic reproduction rates increase very fast. This assumption also fits to the model of cancer, since mutations are always more sensitive to proliferate.

Theorem 2.3.6 : Assume that (2.21) holds. Recall the definition of θ_k in Proposition 2.3.4 and let us fix $\lambda_{1,2} > 0$, then under $\mathbb{P}_{\mathbf{e}_1}$, for $i \geq 3$,

$$\frac{\tau_i}{\sum_{k=2}^{i} \theta_k} \xrightarrow{P} 1\,, \quad \text{as } \frac{\lambda_{k-2,k-1}}{\lambda_{k-1,k}} \to 0, \text{ for } k = 3, \dots, i.$$

Besides, the expectation of τ_i fulfills the following approximation :

$$\mathbb{E}_{\mathbf{e}_1}(\tau_i) \sim \sum_{k=2}^{i} E(\theta_k)\,, \quad \text{as } \frac{\lambda_{k-2,k-1}}{\lambda_{k-1,k}} \to 0, \text{ for } k = 3, \dots, i.$$

Démonstration. Since $s \mapsto X^{k-2,k-1}_{\int_0^{\kappa_{k-1}(s)} Z_u^{(k-2)}\,\mathrm{d}u}$ are increasing processes, then from (2.20), $\mathbb{P}_{\mathbf{e}_1}$-almost surely on the set $\{X^{k-2,k-1}_{\int_0^{\kappa_{k-1}(\gamma_k)} Z_u^{(k-2)}\,\mathrm{d}u} = 1\}$, we have

$$\int_0^{\gamma_k} \frac{1}{X_s^{k-1,k-1} + X^{k-2,k-1}_{\int_0^{\kappa_{k-1}(s)} Z_u^{(k-2)}\,\mathrm{d}u}}\,\mathrm{d}s = \int_0^{\gamma_k} \frac{1}{X_s^{k-1,k-1}+1}\,\mathrm{d}s\,.$$

Hence it follows from Lemma 2.3.5 that for fixed $\lambda_{1,2} > 0$, as $\lambda_{n-2,n-1}/\lambda_{n-1,n} \to 0$, for all $n = 3, \dots, k$,

$$\left(\int_0^{\gamma_k} \frac{1}{X_s^{k-1,k-1}+1}\,\mathrm{d}s\right)^{-1} \int_0^{\gamma_k} \frac{1}{X_s^{k-1,k-1} + X^{k-2,k-1}_{\int_0^{\kappa_{k-1}(s)} Z_u^{(k-2)}\,\mathrm{d}u}}\,\mathrm{d}s \xrightarrow{P} 1\,,$$

and the first part of the theorem is easily derived from this convergence and (2.13) (or equivalently (2.19)).

In order to prove the second part, let us first set

$$H_k := \int_0^{\gamma_k} \frac{1}{X_s^{k-1,k-1} + X^{k-2,k-1}_{\int_0^{\kappa_{k-1}(s)} Z_u^{(k-2)}\,\mathrm{d}u}}\,\mathrm{d}s \quad \text{and} \quad A_k := \{X^{k-2,k-1}_{\int_0^{\kappa_{k-1}(\gamma_k)} Z_u^{(k-2)}\,\mathrm{d}u} = 1\}\,.$$

Then from (2.13), $\mathbb{E}_{\mathbf{e}_1}(\tau_i) = \sum_{k=2}^{i} \mathbb{E}_{\mathbf{e}_1}(H_k)$, so it suffices to prove that for all $k = 2, \dots, i$,

$$\mathbb{E}_{\mathbf{e}_1}(H_k) \sim \mathbb{E}(\theta_k), \quad \text{as } \lambda_{n-2,n-1}/\lambda_{n-1,n} \to 0, \text{ for } n = 3, \dots, k. \tag{2.26}$$

Observe that $\mathbb{E}_{\mathbf{e}_1}(H_k) = \mathbb{E}(\theta_k \mathbf{1}_{A_k}) + \mathbb{E}_{\mathbf{e}_1}(H_k \mathbf{1}_{A_k^c})$. Moreover, $\mathbb{E}_{\mathbf{e}_1}(H_k \mathbf{1}_{A_k^c}) \le \mathbb{E}(\theta_k \mathbf{1}_{A_k^c})$. Then to obtain (2.26), it is enough to prove that

$$\frac{\mathbb{E}(\theta_k \mathbf{1}_{A_k^c})}{\mathbb{E}(\theta_k)} \to 0, \quad \text{as } \lambda_{n-2,n-1}/\lambda_{n-1,n} \to 0, \text{ for } n = 3, \dots, k. \tag{2.27}$$

But for any $p, q \ge 1$, such that $p^{-1}+q^{-1} = 1$, we have from Holder inequality $\mathbb{E}(\theta_k \mathbf{1}_{A_k^c}) \le \mathbb{E}(\theta_k^p)^{1/p} P(A_k^c)^{1/q}$. Moreover, we clearly have $\mathbb{E}(\theta_k^p)^{1/p} \sim 1/\lambda_{k-1,k}$, as $\lambda_{k-1,k} \to \infty$. Hence, (2.27) is satisfied thanks to Lemma 2.3.5. □

We end this section with an example where the distribution of τ_i can be estimated a bit more specifically. We consider the case of binary fission with mutations, where each individual of type i can give birth to either two individuals of type i or one individual of type i and one individual of type $i+1$. In particular, all jumps of $Z^{i,i}$ have size 1 and $X^{i,i}$ is a Poisson process with parameter $\lambda_{i,i}$.

Corollary 2.3.7 : With the above assumtions, the law of τ_i can be specified as follows.

1. Under $\mathbb{P}_{\mathbf{e}_{i-1}}$, the Laplace transform of τ_i is expressed as,

$$\mathbb{E}_{\mathbf{e}_{i-1}}(e^{-\alpha\tau_i}) = \lambda_{i-1,i} \sum_{n\ge 0} \frac{\lambda_{i-1,i-1}^n}{\prod_{k=1}^{n+1}(\alpha_k + \cdots + \alpha_n + \bar{\alpha}_{n+1})}, \quad \alpha \ge 0\,,$$

where $\alpha_k = \frac{\alpha}{k(k+1)}$, $\bar{\alpha}_k = \lambda_{i-1} + \frac{\alpha}{k}$, for $k \ge 1$.

2. The expectation of τ_i is given by $\mathbb{E}_{\mathbf{e}_{i-1}}(\tau_i) = \frac{1}{\lambda_{i-1,i}\lambda_{i-1,i-1}} \ln \frac{\lambda_{i-1}}{\lambda_{i-1,i}}$. In particular, for fixed $\lambda_{1,2} > 0$, under $\mathbb{P}_{\mathbf{e}_1}$, the expectation of τ_i, for $i \geq 3$, fulfills the following approximation :

$$\mathbb{E}_{\mathbf{e}_1}(\tau_i) \sim \sum_{k=2}^{i} \lambda_{k-1,k}^{-2}\,, \quad \text{as } \frac{\lambda_{k-2,k-1}}{\lambda_{k-1,k}} \to 0, \text{ for } k = 3, \ldots, i.$$

Démonstration. From part 2. of Proposition 2.3.4, for all $\alpha \geq 0$,

$$\begin{aligned}\mathbb{E}_{\mathbf{e}_{i-1}}(e^{-\alpha\tau_i}) &= \mathbb{E}\left(e^{-\alpha\int_0^{\gamma_i}\frac{1}{1+X_s^{i-1,i-1}}\mathrm{d}s}\right)\\ &= \lambda_{i-1,i}\int_0^{+\infty}\mathbb{E}\left(e^{-\alpha\int_0^x\frac{1}{1+X_s^{i-1,i-1}}\mathrm{d}s}\right)e^{-\lambda_{i-1,i}x}\mathrm{d}x\,. \qquad (2.28)\end{aligned}$$

Under $\mathbb{P}_{\mathbf{e}_{i-1}}$, $X^{i-1,i-1}$ is a standard Poisson process with parameter $\lambda_{i-1,i-1}$ starting at 0. So if we denote by $(J_n)_{n\geq 1}$ the sequence of jump times of $X^{i-1,i-1}$ and set $J_0 = 0$, then developing the expression $\mathbb{E}\left(e^{-\alpha\int_0^x\frac{1}{1+X_s^{i-1,i-1}}\mathrm{d}\,s}\right)$, we obtain with the convention that $\sum_{k=0}^{-1} = 0$,

$$\begin{aligned}\mathbb{E}\left(e^{-\alpha\int_0^x\frac{1}{1+X_s^{i-1,i-1}}\mathrm{d}\,s}\right) &= \sum_{n\geq 0}\mathbb{E}\left(X_x^{i-1,i-1}=n, e^{-\alpha\left(\frac{x-J_n}{n+1}+\sum_{k=0}^{n-1}\frac{J_{k+1}-J_k}{k+1}\right)}\right)\\ &= e^{-(\alpha+\lambda_{i-1,i-1})x} + \sum_{n\geq 1}e^{-\lambda_{i-1,i-1}x}\frac{(\lambda_{i-1,i-1}x)^n}{n!}\\ &\qquad\times\int_{0\leq x_1\leq\cdots\leq x_n\leq x} e^{-\alpha\left(\frac{x}{n+1}+\sum_{k=1}^n\frac{x_k}{k(k+1)}\right)}\frac{n!}{x^n}\mathrm{d}x_1\ldots\mathrm{d}x_n\\ &= e^{-(\alpha+\lambda_{i-1,i-1})x} + \sum_{n\geq 1}\lambda_{i-1,i-1}^n e^{-\left(\lambda_{i-1,i-1}+\frac{\alpha}{n+1}\right)x}\\ &\qquad\times\int_{0\leq x_1\leq\cdots\leq x_n\leq x} e^{-\alpha\sum_{k=1}^n\frac{x_k}{k(k+1)}}\mathrm{d}x_1\ldots\mathrm{d}x_n.\end{aligned}$$

Then coming back to expression (2.28), we obtain, using the fact that $\lambda_{i-1} = \lambda_{i-1,i} + \lambda_{i-1,i-1}$ and with the convention that $\sum_{k=1}^{0} = 0$,

$$\mathbb{E}_{\mathbf{e}_{i-1}}(e^{-\alpha\tau_i}) = \lambda_{i-1,i}\sum_{n\geq 0}\lambda_{i-1,i-1}^n\int_{0\leq x_1\leq\cdots\leq x_{n+1}} e^{-\sum_{k=1}^{n+1}\alpha_k x_k}\mathrm{d}x_1\ldots\mathrm{d}x_{n+1},$$

where $\alpha_1, \ldots, \alpha_{n+1}$ are defined in the satement. The computation of the integral is easily done.

Then using again part 2. of Proposition 2.3.4, we obtain the expectation of τ_i under $\mathbb{P}_{\mathbf{e}_{i-1}}$, after easy computations,

$$\begin{aligned}\mathbb{E}_{\mathbf{e}_{i-1}}(\tau_i) &= \int_0^{+\infty} \mathrm{d}x \lambda_{i-1,i} e^{-\lambda_{i-1,i}x} \int_0^x e^{-\lambda_{i-1,i-1}s} \sum_{k\geq 0} \frac{(\lambda_{i-1,i-1}s)^k}{(k+1)!} \mathrm{d}s \\ &= \frac{1}{\lambda_{i-1,i}\lambda_{i-1,i-1}} \ln \frac{\lambda_{i-1}}{\lambda_{i-1,i}} .\end{aligned}$$

We conclude from Theorem 2.3.6. □

Chapitre 3

Cyclically exchangeable sequences and enumeration of multitype forests

3.1 Introduction

It has been proved in [CL15] and [Cha15] that any d-type branching forest can be coded by some random walk $(X^{(1)}, \dots, X^{(d)})$ whose coordinates $X^{(i)}$, $i = 1, \dots, d$ are independent. Moreover for each i, $X^{(i)} = (X^{i,j}(k), i, j = 1, \dots, d,\ k \in \mathbb{Z}_+)$ where $X^{i,j}$ are integer valued processes such that $X_0^{i,j} = 0$, $X^{i,j}$ for $i \neq j$ are nondecreasing and $X^{i,i}$ are downward skip free. The special features of such processes allow us to show that if for some $\mathbf{r} = (r_1, \dots, r_d) \in \mathbb{Z}_+^d$, the system of equations

$$r_j + \sum_{i=1}^{d} X^{i,j}(k_i) = 0\,, \quad j = 1, \dots, d \tag{3.1}$$

admits a solution $\mathbf{k} = (k_1, \dots, k_d) \in \mathbb{Z}_+^d$, then there is a solution $\mathbf{n} = (n_1, \dots, n_d) \in \mathbb{Z}_+^d$ of (3.1) such that $n_i \leq k_i'$, for all $i = 1, \dots, d$ and for any solution $\mathbf{k}' = (k_1', \dots, k_d')$. The multi-index $\mathbf{n} = (n_1, \dots, n_d) \in \mathbb{Z}_+^d$ is called the smallest solution of the system $\{\mathbf{r}, X\}$

or the first passage time of the additive random walk $X(\mathbf{k}) = \sum_{i=1}^{d} X^{i,j}(k_i)$ at level $-\mathbf{r}$ and we will denote it by

$$T_{\mathbf{r}} = \inf\{\mathbf{k} : X(\mathbf{k}) = -\mathbf{r}\}\,.$$

The law of this multivariate first passage time is derived from an extension of the well known Ballot theorem which has been proved in [CL15].

Theorem 3.1.1 (Multivariate Ballot Theorem) **:** Let $(X^{(1)}, \ldots, X^{(d)})$ be a random walk which is defined as above. Fix $\mathbf{n} = (n_1, \ldots, n_d) \in \mathbb{N}^d$. Then for any $\mathbf{r} = (r_1, \ldots, r_d) \in \mathbb{Z}_+^d$ such that $\mathbf{r} > 0$ and k_{ij}, $i, j \in [d]$, such that $k_{ij} \in \mathbb{Z}_+$, for $i \neq j$ and $-k_{jj} = r_j + \sum_{i \neq j} k_{ij}$,

$$\mathbb{P}\Big(X_{n_i}^{i,j} = k_{ij},\ i, j \in [d] \text{ and } T_{\mathbf{r}} = \mathbf{n}\Big) = \frac{\det(K)}{n_1 n_2 \ldots n_d} \mathbb{P}\Big(X_{n_i}^{i,j} = k_{ij}\Big), \tag{3.2}$$

where K is the matrix $(-k_{ij})_{i,j}$.

As in the one dimensional case, this theorem is valid in the more general framework of cyclically exchangeable sequences. It is actually the direct consequence of the multivariate cyclic lemma, proved in [CL15] and recalled in the next section.
The multivariate ballot theorem together with the special coding of multitype branching forest from multidimensional random walks leads directly to the following extension of Dwass formula, which gives the law of the total progeny of d-type branching forests.

Theorem 3.1.2 : Let $(\nu_1, \ldots, \nu_d)$ by the progeny law of a d-type branching forest and assume that $(\nu_1, \ldots, \nu_d)$ is irreducible, non-degenarate and critical or subcritical. For $i, j \in [d]$, let O_i be the total number of individuals of type i, up to the extinction time and for $i \neq j$, let A_{ij} be the total number of individuals of type j, whose parent is of type i.

Then for all integers r_i, n_i, k_{ij}, $i, j \in [d]$, such that $r_i \geq 0$, $r_1 + \cdots + r_d \geq 1$, $k_{ij} \geq 0$,

for $i \neq j$, $-k_{jj} = r_j + \sum_{i \neq j} k_{ij}$, and $n_i \geq -k_{ii}$,

$$\begin{aligned} & \mathbb{P}_{\mathbf{r}}\Big(O_1 = n_1, \dots, O_d = n_d, A_{ij} = k_{ij}, i, j \in [d], i \neq j\Big) \\ = \; & \frac{\det(K)}{\bar{n}_1 \bar{n}_2 \dots \bar{n}_d} \prod_{i=1}^{d} \nu_i^{*n_i}(k_{i1}, \dots, k_{i(i-1)}, n_i + k_{ii}, k_{i(i+1)}, \dots, k_{id}) \,, \end{aligned}$$

where $\mathbf{r} = (r_1, \dots, r_d)$ and $\mathbb{P}_{\mathbf{r}}$ is the probability under wich the branching forest has r_i roots of type $i \in [d]$, a.s. Moreover $\nu_i^{*0} = \delta_0$, $\bar{n}_i = n_i \vee 1$ and K is the matrix $(-k_{ij})_{i,j \in [d]}$ to which we removed the line i and the column i, for all i such that $n_i = 0$.

We emphasize that this result obtained in [CL15] may easily be extended to the case where the process is not irreducible.

Then goal of this paper is to derive from the above results some enumeration formulas for multitype forests. In Section 3.3, we will first consider the case of plane forests and give the number of forests with a given number of vertices of type j whose parent has type i. We also provide a formula for the number of forests whose number of vertices of each indegree type is given. Then in Section 3.4, we will consider the case of labeled forests and recover some enumeration formulas which have recently been obtained in [BCLL03] and [BM14]. Finally, we will show in Section 3.5 that the multivariate ballot theorem also provides a direct proof of the Lagrange-Good inversion formula.

3.2 The multivariate cyclic lemma

In what follows, $\mathbb{Z}$ is the set of integers, $\mathbb{Z}_+$ is the set of nonnegative integers and $\mathbb{N}$ is the set of positive integers. We will also use the notation : $[d] = \{1, 2, \dots, d\}$ and $\mathbf{1} = (1, 1, \dots, 1)$. Then for $\mathbf{k} = (k_1, \dots, k_d)$ and $\mathbf{n} = (n_1, \dots, n_d)$ elements of $\mathbb{Z}_+^d$, we write $\mathbf{k} \leq \mathbf{n}$ if $k_i \leq n_i$, for all $i \in [d]$ and we write $\mathbf{k} < \mathbf{n}$ if $k_i \leq n_i$, for all $i \in [d]$ and $k_j < n_j$, for some $j \in [d]$.

The standard cyclic lemma is concerned with integer valued, downward skip free sequences. In order to state its multivariate version, we need the following extension of

such sequences.

Definition 3.2.1 : Let S_d be the set of $\left[\mathbb{Z}^d\right]^d$-valued sequences, $x = (x^{(1)}, x^{(2)}, \ldots, x^{(d)})$, such that for all $i \in [d]$, $x^{(i)} = (x^{i,1}, \ldots, x^{i,d})$ is a $\mathbb{Z}^d$-valued sequence defined on some interval of integers, $\{0, 1, 2, \ldots, n_i\}$, $n_i \in \mathbb{N}$, which satisfies $x_0^{(i)} = 0$ and

(i) for $i \neq j$, the sequence $(x_n^{i,j})_{0 \leq n \leq n_i}$ is nondecreasing,

(ii) for all i, $x_{n+1}^{i,i} - x_n^{i,i} \geq -1$, $0 \leq n \leq n_i - 1$.

A sequence $x \in S_d$ will sometimes be denoted by $x = (x_k^{i,j},\ 0 \leq k \leq n_i,\ i, j \in [d])$ and for more convenience, we will sometimes denote $x_k^{i,j}$ by $x^{i,j}(k)$. The vector $\mathbf{n} = (n_1, \ldots, n_d)$ will be called the length of x.

Let $\mathbf{r} = (r_1, \ldots, r_d) \in \mathbb{Z}_+^d$ be such that $\mathbf{r} > \mathbf{0}$ and $x \in S_d$ with length $\mathbf{n}$. Then we say that a vector $\mathbf{k} = (k_1, \ldots, k_d) \in \mathbb{Z}_+^d$ such that $\mathbf{k} \leq \mathbf{n}$ is a solution of the system $(\mathbf{r}, x)$ if

$$r_j + \sum_{i=1}^{d} x^{i,j}(k_i) = 0\,, \quad j = 1, \ldots, d\,. \tag{3.3}$$

Let $x = (x^{(1)}, x^{(2)}, \ldots, x^{(d)}) \in S_d$. For $i \in [d]$, the n_i-cyclical permutations of $x^{(i)}$ are the n_i applications $x_{q,n_i}^{(i)}$, $q = 0, \ldots, n_i - 1$ which are defined by :

$$x_{q,n}^{(i)}(h) \overset{\text{(def)}}{=} \begin{cases} x^{(i)}(q+h) - x^{(i)}(q) & \text{if } \ 0 \leq h \leq n_i - q\,, \\ x^{(i)}(h - (n_i - q)) + x^{(i)}(n_i) - x^{(i)}(q) & \text{if } \ n_i - q \leq h \leq n_i\,. \end{cases} \tag{3.4}$$

Note that $x_{0,n_i}^{(i)} \equiv x^{(i)}$. The transformation $x^{(i)} \mapsto x_{q,n_i}^{(i)}$ consists in inverting the parts $\{x^{(i)}(h),\ 0 \leq h \leq q\}$ and $\{x^{(i)}(h),\ q \leq h \leq n_i\}$ in such a way that the new application, $x_{q,n_i}^{(i)}$, has the same values as $x^{(i)}$ at 0 and n_i, i.e. $x_{q,n_i}^{(i)}(0) = 0$ and $x_{q,n_i}^{(i)}(n_i) = x^{(i)}(n_i)$. Then the $n_1 \times n_2 \times \cdots \times n_d$ cyclical permutations of x are the applications $x_{\mathbf{q},\mathbf{n}} = (x_{q_1,n_1}^{(1)}, \ldots, x_{q_d,n_d}^{(i)})$, for $\mathbf{q} = (q_1, \ldots q_d) \leq \mathbf{n} - \mathbf{1}$. It is plain that $x_{\mathbf{q},\mathbf{n}} \in S_d$.

Definition 3.2.2 : Let $x \in S_d$, with finite length $\mathbf{n} = (n_1, \ldots, n_d) \in \mathbb{N}^d$. Let $\mathbf{r} = (r_1, \ldots, r_d) \in \mathbb{Z}_+^d$ be such that $\mathbf{r} > \mathbf{0}$ and assume that $\mathbf{n}$ is a solution of the system $(\mathbf{r}, x)$. For $\mathbf{0} \leq \mathbf{q} \leq \mathbf{n} - \mathbf{1}$, we say that $x_{\mathbf{q},\mathbf{n}}$ is a *good* $\mathbf{n}$*-cyclical permutation* of x with

respect to $\mathbf{r}$, if $\mathbf{n}$ is the smallest solution of the system $(\mathbf{r}, x_{\mathbf{q},\mathbf{n}})$, that is if $\mathbf{n}'$ is any solution of $(\mathbf{r}, x_{\mathbf{q},\mathbf{n}})$ then $\mathbf{n} \leq \mathbf{n}'$. When no confusion is possible, we will simply say that $x_{\mathbf{q},\mathbf{n}}$ is a *good cyclical permutation* of x.

The next lemma extends the standard cyclic lemma which asserts that for any integer valued, downward skip free sequence $(s(k), 0 \leq k \leq n)$, that is $s(k) - s(k-1) \geq -1$, $k = 1, \ldots, n$ and $s(0) = 0$, $s(n) = -r < 0$, there are exactly r cyclical permutations which first hits $-r$ at time n.

Lemma 3.2.1 (Multivariate Cyclic Lemma) **:** Let $x \in S_d$, with length $\mathbf{n} = (n_1, \ldots, n_d) \in \mathbb{N}^d$ and let $\mathbf{r} = (r_1, \ldots, r_d) \in \mathbb{Z}_+^d$ be such that $\mathbf{r} > \mathbf{0}$. Assume that $\mathbf{n}$ is a solution of the system $(\mathbf{r}, x)$ such that $x^{i,i}(n_i) < 0$, for all $i \in [d]$. Then the number of good cyclical permutations of x is $\det((-x^{i,j}(n_i))_{i,j\in[d]})$.

Let us stress that the multivariate ballot theorem 3.1.1 actually follows directly from Lemma 3.2.1, see the proof of Theorem 3.4 in [CL15].

3.3 Enumeration of multitype plane forests

We will denote by $\mathfrak{F}$ the set of plane forests. More specifically, an element $\mathbf{f} \in \mathfrak{F}$ is a directed plane graph with no loops on a possibly infinite and non empty set of vertices $\mathbf{v} = \mathbf{v}(\mathbf{f})$, with a *finite* number of connected components, such that each vertex has a finite inner degree and an outer degree equals to 0 or 1. The elements of $\mathfrak{F}$ will simply be called forests.The connected components of a forest are called the *trees*. A forest consisting of a single connected component is also called a tree. In a tree $\mathbf{t}$, the only vertex with outer degree equal to 0 is called the *root* of $\mathbf{t}$. It will be denoted by $r(\mathbf{t})$. The roots of the connected components of a forest $\mathbf{f}$ are called the roots of $\mathbf{f}$. For two vertices u and v of a forest $\mathbf{f}$, if (u, v) is a directed edge of $\mathbf{f}$, then we say that u is a *child* of v, or that v is the *parent* of u.

We first give an order to the trees of the forest $\mathbf{f}$ and denote them by $\mathbf{t}_1(\mathbf{f}), \mathbf{t}_2(\mathbf{f}), \ldots, \mathbf{t}_k(\mathbf{f}), \ldots$

(we will usually write $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_k, \ldots$ if no confusion is possible). Then we rank (a part of) the vertices of $\mathbf{f}$ according to the breadth first search order, by ranking first the vertices of $\mathbf{t}_1$, then the vertices of $\mathbf{t}_2$, and so on.

A d-type forest is a couple $(\mathbf{f}, c_\mathbf{f})$, where $\mathbf{f} \in \mathcal{F}$ and $c_\mathbf{f}$ is an application $c_\mathbf{f} : \mathbf{v}(\mathbf{f}) \to [d]$. For $v \in \mathbf{v}(\mathbf{f})$, the integer $c_\mathbf{f}(v)$ is called the *type* (or the *color*) of v. The set of finite d-type forests will be denoted by $\mathcal{F}_d$. An element $(\mathbf{f}, c_\mathbf{f}) \in \mathcal{F}_d$ will often simply be denoted by $\mathbf{f}$. We assume that for any $\mathbf{f} \in \mathcal{F}_d$, if $u_i, u_{i+1}, \ldots, u_{i+j} \in \mathbf{v}(\mathbf{f})$ have the same parent, then $c_\mathbf{f}(u_i) \leq c_\mathbf{f}(u_{i+1}) \leq \cdots \leq c_\mathbf{f}(u_{i+j})$. Moreover, if $u_1, \ldots, u_k$ are the roots of $\mathbf{f}$, then $c_\mathbf{f}(u_1) \leq \cdots \leq c_\mathbf{f}(u_k)$.

A *cluster* or a *subtree of type* $i \in [d]$ of a d-type forest $(\mathbf{f}, c_\mathbf{f}) \in \mathcal{F}_d$ is a maximal connected subgraph of $(\mathbf{f}, c_\mathbf{f})$ whose all vertices are of type i. Formally, $\mathbf{t}$ is a cluster of type i of $(\mathbf{f}, c_\mathbf{f})$, if it is a connected subgraph whose all vertices are of type i and such that either the root of $\mathbf{t}$ has no parent or the type of its parent is different from i. Moreover, if the parent of a vertex $v \in \mathbf{v}(\mathbf{t})^c$ belongs to $\mathbf{v}(\mathbf{t})$, then $c_\mathbf{f}(v) \neq i$. Clusters of type i in $\mathbf{t}_1$ are ranked according to the order of their roots in the breadth first search order of $\mathbf{t}_1$. Then we continue by ranking clusters of type i in $\mathbf{t}_2$, and so on. We denote by $\mathbf{t}_1^{(i)}, \mathbf{t}_2^{(i)}, \ldots, \mathbf{t}_k^{(i)}, \ldots$ the sequence of clusters of type i in $(\mathbf{f}, c_\mathbf{f})$. The forest $\mathbf{f}^{(i)} := \{\mathbf{t}_1^{(i)}, \mathbf{t}_2^{(i)}, \ldots, \mathbf{t}_k^{(i)}, \ldots\}$ is called *the subforest of type* i of $(\mathbf{f}, c_\mathbf{f})$. The vertices of $\mathbf{f}^{(i)}$ will then be ranked in the breadth first search order. For each $i \in [d]$ we will denote by $u_n^{(i)}$ the n-th vertex of type i of the forest $\mathbf{f}^{(i)}$.

Let $(\mathbf{f}, c_\mathbf{f}) \in \mathcal{F}_d$, $u \in \mathbf{v}(\mathbf{f})$ and denote by $p_i(u)$ the number of children of type i of u. For each $i \in [d]$, let $n_i \geq 0$ be the number of vertices of type i in $\mathbf{v}(\mathbf{f})$. Then let $x = (x^{(1)}, \ldots, x^{(d)})$ be the coding sequence of $\mathbf{f}$. That is, for all $i \in [d]$, $x^{(i)}$ is the d-dimensional chain $x^{(i)} = (x^{i,1}, \ldots, x^{i,d})$, with length n_i, whose values belong to the set $\mathbb{Z}^d$, such that $x_0^{(i)} = 0$ and if $n_i \geq 1$ then

$$x_{n+1}^{i,j} - x_n^{i,j} = p_j(u_{n+1}^{(i)}), \quad \text{if } i \neq j \quad \text{and} \quad x_{n+1}^{i,i} - x_n^{i,i} = p_i(u_{n+1}^{(i)}) - 1, \quad 0 \leq n \leq n_i - 1. \tag{3.5}$$

We know from [CL15] that $(n_1, \ldots, n_d)$ is actually the smallest solution of the system

$(\mathbf{r}, x)$, where r_i is the number of roots of type i of the forest $\mathbf{f}$.

We may now derive from the previous results, some enumeration formulas for multitype forests. In all this subsection, r_i, n_i and k_{ij}, $i, j \in [d]$ will be integers satisfying conditions of Theorem 3.1.1, that is $r_i \geq 0$, $r_1 + \cdots + r_d \geq 1$, $k_{ij} \geq 0$, for $i \neq j$, $-k_{jj} = r_j + \sum_{i \neq j} k_{ij}$ and $n_i \geq -k_{ii}$. We assume moreover that $-k_{ii} > 0$, for all $i \in [d]$.

Our first result is an application of Theorem 3.1.1 which gives the number of plane forests with n_i vertices of type i, r_i roots of type i and such that k_{ij} vertices of type j have a parent of type i. It extends the one dimensional case where, the number of the unlabeled forests with r trees and n vertices is $\frac{r}{n}\binom{2n-r-1}{n-r}$.

Theorem 3.3.1 : Let $\mathcal{F}_d^{k_{ij},\mathbf{n}}$ be the subset of plane forests of $\mathcal{F}_d$, with n_i vertices of type i, r_i roots of type i and such that for $i \neq j$, k_{ij} vertices of type j have a parent of type i, then

$$\left|\mathcal{F}_d^{k_{ij},\mathbf{n}}\right| = \frac{\det(-k_{ij})}{n_1 n_2 \dots n_d} \prod_{i,j=1}^{d} \binom{n_i + k'_{ij} - 1}{k'_{ij}},$$

where $k'_{ii} = n_i + k_{ii}$ and for $i \neq j$, $k'_{ij} = k_{ij}$.

Démonstration. We use the same arguments as in Section 6 of Pitman [Pit98] where the case $d = 1$ is treated. Let F be a d-type branching forest with progeny law ν given by

$$\nu_i(k_1, \cdots, k_d) = \prod_{j=1}^{d} (1 - p_{ij})^{k_j} p_{ij}, \qquad i = 1, 2, \dots d,\ (k_1, \cdots, k_d) \in \mathbb{Z}_+,$$

where $0 < p_{ij} < 1$, $i, j \in [d]$ That is to say, each individual of type i gives birth to children of different types independently, respectively according to the geometric distribution $\mu_{ij}(\cdot)$ with parameter p_{ij}, $j \in [d]$. Then

$$\nu_i(k_1, \cdots, k_d) = \prod_{j=1}^{d} \mu_{ij}(k_j).$$

Let $\mathbf{p}(u) = (p_i(u),\ i \in [d])$ be the increment of the Lukasiewicz-Harris path related to some vertex u of some forest $\mathbf{f}$, as it is defined in (3.5). Recall also that $c(u) = c_{\mathbf{f}}(u) \in$

$[d]$ is the type of the vertex u. Then for any $\mathbf{f} \in \mathcal{F}_d^{k_{ij},\mathbf{n}}$,

$$\mathbb{P}_{\mathbf{r}}\left(\mathrm{F} = \mathbf{f}\right) = \prod_{u\in\mathbf{f}} \nu_{c(u)}(\mathbf{p}(u)) = \prod_{u\in\mathbf{f}}\prod_{j=1}^{d} \left(1-p_{c(u)j}\right)^{p_j(u)} p_{c(u)j} = \prod_{i,j=1}^{d} \left(1-p_{ij}\right)^{k'_{ij}} p_{ij}^{n_i}. \tag{3.6}$$

Since this probability is the same for all the forests $\mathbf{f} \in \mathcal{F}_d^{k_{ij},\mathbf{n}}$, the following conditional distribution is the uniform distribution on $\mathcal{F}_d^{k_{ij},\mathbf{n}}$:

$$\mathbb{P}_{\mathbf{r}}\Big(\mathrm{F} \in \cdot \,\Big|O(F) = \mathbf{n}, A_{ij}(F) = k_{ij}, i,j \in [d], i \neq j\Big).$$

But Theorem 3.1.2 tells us that

$$\begin{aligned}
&\mathbb{P}_{\mathbf{r}}\Big(O_1 = n_1, \ldots, O_d = n_d, A_{ij} = k_{ij}, i,j \in [d], i \neq j\Big)\\
=\;& \frac{\det(-k_{ij})}{n_1 n_2 \ldots n_d} \prod_{i=1}^{d} \nu_i^{*n_i}(k_{i1}, \ldots, k_{i(i-1)}, n_i + k_{ii}, k_{i(i+1)}, \ldots, k_{id})\\
=\;& \frac{\det(-k_{ij})}{n_1 n_2 \ldots n_d} \prod_{i,j=1}^{d} \mu_{ij}^{*n_i}(k'_{ij})\\
=\;& \frac{\det(-k_{ij})}{n_1 n_2 \ldots n_d} \prod_{i,j=1}^{d} \binom{n_i + k'_{ij} - 1}{k'_{ij}} (1-p_{ij})^{k'_{ij}} p_{ij}^{n_i}\\
=\;& \frac{\det(-k_{ij})}{n_1 n_2 \ldots n_d} \prod_{i,j=1}^{d} \binom{n_i + k'_{ij} - 1}{k'_{ij}} \left(\prod_{i,j=1}^{d} (1-p_{ij})^{k'_{ij}} p_{ij}^{n_i}\right).
\end{aligned}$$

Comparing this probability with (3.6), we obtain our result. □

Our second result is concerned with the number of unlabeled forests whose number of vertices of each indegree type is given. We say that a vertex has indegree type $\mathbf{u} = (u_1, u_2, \ldots, u_d)$ if it has u_j children of type j for $j \in [d]$. Let $\mathbf{U} = \prod_{i=1}^{d} \{0, 1, \ldots, n_i\}$, and let $\mathbf{N} = (N_{i,\mathbf{u}})_{i\in[d],\mathbf{u}\in\mathbf{U}}$ be a tuple of nonnegative integers satisfying $n_i = \sum_{\mathbf{u}\in\mathbf{U}} N_{i,\mathbf{u}}$ and $k'_{ij} = \sum_{\mathbf{u}\in\mathbf{U}} u_j N_{i,\mathbf{u}}$ for $i,j \in [d]$.

Theorem 3.3.2 : Define the subset $\mathscr{F}_d^{k_{ij},\mathbf{n}}(\mathbf{N})$ of $\mathscr{F}_d^{k_{ij},\mathbf{n}}$ consisting of unlabeled forests having $N_{i,\mathbf{u}}$ vertices of type i with indegree type $\mathbf{u}$ for $i \in [d]$, $\mathbf{u} \in \mathbf{U}$.

$$\left|\mathscr{F}_d^{k_{ij},\mathbf{n}}(\mathbf{N})\right| = \frac{\prod_{j=1}^{d}(n_j-1)!\det(-k_{ij})}{\prod_{i\in[d],\mathbf{u}\in\mathbf{U}}(N_{i,\mathbf{u}})!}.$$

The proof of this theorem uses an analoguous result for labelled forest. It is postponed to the end of the next section.

3.4 Enumeration of multitype labeled forests

Now we shall recover some enumeration formulas of labeled forests according to the degree of their vertices which have recently been obtained by combinatorical techniques. We provide here a much simpler approach based on the coding of multitype forests and the so called Multivariate Cyclic Lemma recalled at the beginning of the paper.

Again in all this section, r_i, n_i and k_{ij}, $i, j \in [d]$ will be integers satisfying conditions of Theorems 3.1.1 and 3.1.2, that is $r_i \geq 0$, $r_1 + \cdots + r_d \geq 1$, $k_{ij} \geq 0$, for $i \neq j$, $-k_{jj} = r_j + \sum_{i \neq j} k_{ij}$ and $n_i \geq -k_{ii}$. We assume moreover that $-k_{ii} > 0$, for all $i \in [d]$.

Definition 3.4.1 : To each forest $\mathbf{f}$ with n_i vertices of type i and to each vertex of type i in $\mathbf{f}$, we associate an integer in $[n_i]$, which is called its *label*. Then $\mathbf{f}$ is called a *labeled plane forest*. Let $\mathcal{L}$ be the set of labeled plane forests with n_i vertices of type i, r_i roots of type i, in which (k_{ij}) vertices of type j have a parent of type i.

Let $\mathbf{c} = (c_{i,j,k})_{i,j \in [d], k \in [n_i]}$ be a tuple of non-negative integers such that $k'_{ij} = \sum_{k=1}^{n_i} c_{i,j,k}$, where k'_{ij} is defined in Theorem 3.3.1. We will denote by $\mathcal{L}(\mathbf{c})$ the subset of $\mathcal{L}$, of forests in which the vertex of type i with label k has $c_{i,j,k}$ offspring of type j. Then $\mathbf{c}$ is called the *indegree tuple* of the forest $\mathbf{f} \in \mathcal{L}(\mathbf{c})$.

The following result has been obtained in [BM14], see Proposition 11.

Proposition 3.4.1 (Enumeration of labeled plane forests by indegree tuple) **:** For any indegree tuple $\mathbf{c} = (c_{i,j,k})_{i,j \in [d], k \in [n_i]}$, the number of forests with $\mathbf{c}$ as the indegree tuple in $\mathcal{L}$ is

$$|\mathcal{L}(\mathbf{c})| = \frac{\prod_{j=1}^{d} (n_j - 1)!}{\prod_{i \in [d]} r_i! \prod_{i,j \in [d], k \in [n_i]} c_{i,j,k}!} \det(-k_{ij}) \,. \tag{3.7}$$

Démonstration. Let $\mathbf{f} \in \mathcal{L}(\mathbf{c})$ and let x be its coding sequence as defined in 3.5. According to Lemma 3.2.1 there are $\det(-k_{ij})$ good cyclical permutations of x, each one coding

a different forest in $\mathcal{L}(\mathbf{c})$. On the other hand, any two forests $\mathbf{f}$ and $\mathbf{f}'$ of $\mathcal{L}(\mathbf{c})$ are coded through two sequences x and x' such that for each $i \in [d]$, the sequence of increments $(\Delta x_1^{(i)}, \dots, \Delta x_{n_i}^{(i)})$ is a permutation of the sequence of increments $(\Delta x'^{(i)}_1, \dots, \Delta x'^{(i)}_{n_i})$. Then there are $\prod_{j=1}^{d}(n_j-1)!\det(-k_{ij})$ good permutations of x, where by *good permutation*, we mean a sequence x' which codes a forest in $\mathcal{L}(\mathbf{c})$ and such that the sequence of increments $(\Delta x'^{(i)}_1, \dots, \Delta x'^{(i)}_{n_i})$ is a permutation of the sequence of increments $(\Delta x_1^{(i)}, \dots, \Delta x_{n_i}^{(i)})$.

In the enumeration that we have just done, we counted forests $\mathbf{f}, \mathbf{f}' \in \mathcal{L}(\mathbf{c})$ such that $\mathbf{f}'$ can be obtained by permuting in $\mathbf{f}$ the $c_{i,j,k}$ subtrees whose roots are the $c_{i,j,k}$ children of type j of the kth vertex of type i, for some $i, j \in [d]$ and $k \in [n_i]$ or by permuting the trees with the same type roots in the whole forest. But in this case, $\mathbf{f}$ and $\mathbf{f}'$ are the same forest. Therefore, we still have to divide the number $\prod_{j=1}^{d}(n_j-1)!\det(-k_{ij})$ by $\prod_{i\in[d]} r_i! \prod_{i,j\in[d],k\in[n_i]} c_{i,j,k}!$, that is

$$|\mathcal{L}(\mathbf{c})| = \frac{\prod_{j=1}^{d}(n_j-1)!}{\prod_{i\in[d]} r_i! \prod_{i,j\in[d],k\in[n_i]} c_{i,j,k}!}\det(-k_{ij})\,.$$

□

From Proposition 3.4.1, we can also derive the number of the forests in $\mathcal{L}$ which was obtained in [BCLL03], see Proposition 2.

Proposition 3.4.2 : Recall the definition of k'_{ij}, $i, j \in [d]$ from Theorem 3.3.1. Then

$$|\mathcal{L}| = \prod_{i,j=1}^{d}(n_i)^{k'_{ij}} \frac{\prod_{j=1}^{d}(n_j-1)!}{\prod_{j=1}^{d} r_j! \prod_{i,j\in[d]} k'_{ij}!}\det(-k_{ij})\,. \tag{3.8}$$

Démonstration. A tuple $\mathbf{c} = (c_{i,j,k})_{i,j\in[d],k\in[n_i]}$ of non-negative integers is an indegree tuple in $\mathcal{L}$ if and only if $\sum_{k=1}^{n_i} c_{i,j,k} = k'_{ij}$ for $i, j \in [d]$. Define the set of indegree tuples $\mathbf{C} = \{\mathbf{c} = (c_{i,j,k})_{i,j\in[d],k\in[n_i]}; c_{i,j,k} \in \mathbb{Z}_+, \sum_{k=1}^{n_i} c_{i,j,k} = k'_{ij}\}$. Summing all the indegree

tuples $\mathbf{c} \in \mathbf{C}$, from Proposition 3.4.1, we obtain

$$\begin{aligned} |\mathcal{L}| &= \sum_{\mathbf{c}\in\mathbf{C}} |\mathcal{L}(\mathbf{c})| = \sum_{\mathbf{c}\in\mathbf{C}} \frac{\prod_{j=1}^{d}(n_j-1)!}{\prod_{j=1}^{d}\left(r_j!\prod_{i\in[d],k\in[n_i]} c_{i,j,k}!\right)} \det(-k_{ij}) \\ &= \frac{\prod_{j=1}^{d}(n_j-1)!}{\prod_{j=1}^{d}\left(r_j!\prod_{i\in[d]} k'_{ij}!\right)} \det(-k_{ij}) \prod_{i,j=1}^{d}\left(\sum_{(c_{i,j,\cdot})\in\mathbf{C}_{ij}} \frac{k'_{ij}!}{\prod_{k=1}^{n_i} c_{i,j,k}!}\right) \\ &= \frac{\prod_{j=1}^{d}(n_j-1)!}{\prod_{j=1}^{d}\left(r_j!\prod_{i\in[d]} k'_{ij}!\right)} \det(-k_{ij}) \prod_{i,j=1}^{d}(n_i)^{k'_{ij}}, \end{aligned}$$

where $\mathbf{C}_{ij} = \left\{(c_{i,j,k})_{k=1}^{n_i};\ c_{i,j,k}\in\mathbb{Z}_+,\ \sum_{k=1}^{n_i} c_{i,j,k} = k'_{ij}\right\}$ for $i,j\in[d]$. (3.8) is obtained. □

A multitype labeled plane forest is said to be injective if every vertex has at most one child of each type. Let $\mathcal{L}_{inj}$ be the set consisting of injective forests in $\mathcal{L}$. Now we count the number of forests in $\mathcal{L}_{inj}$. The following result was obtained in [BM14], see Proposition 9.

Proposition 3.4.3 (Enumeration of injective forests) **:**

$$|\mathcal{L}_{inj}| = \prod_{i,j=1}^{d}\binom{n_i}{k'_{ij}} \frac{\prod_{j=1}^{d}(n_j-1)!\det(-k_{ij})}{\prod_{j=1}^{d} r_j!}. \tag{3.9}$$

Démonstration. If $\mathbf{c} = (c_{i,j,k})_{i,j\in[d],k\in[n_i]}$ is the indegree tuple for an injective forest $\mathbf{f}$ in $\mathcal{L}_{inj}$, then $c_{i,j,k} = 0$ or 1. Therefore, from Proposition 3.4.1,

$$|\mathcal{L}(\mathbf{c})| = \frac{\prod_{j=1}^{d}(n_j-1)!\det(-k_{ij})}{\prod_{j=1}^{d}\left(r_j!\prod_{i\in[d],k\in[n_i]} c_{i,j,k}!\right)} = \frac{\prod_{j=1}^{d}(n_j-1)!\det(-k_{ij})}{\prod_{j=1}^{d} r_j!}.$$

This number is unrelated to the choice of indegree tuple $\mathbf{c}$. Moreover, the forests in $\mathcal{L}_{inj}$ have $\prod_{i=1}^{d}\prod_{j=1}^{d}\binom{n_i}{k'_{ij}}$ different indegree tuples. Thus the cardinality of $\mathcal{L}_{inj}$ is

$$|\mathcal{L}_{inj}| = \prod_{i,j\in[d]}\binom{n_i}{k'_{ij}} \frac{\prod_{j=1}^{d}(n_j-1)!\det(-k_{ij})}{\prod_{j=1}^{d} r_j!}.$$

□

We end this section with the number of forests in $\mathcal{L}$ obtained by fixing the number of vertices of each indegree type (but without fixing their labels). Recall that $\mathbf{U} = \prod_{i=1}^{d}\{0,1,\dots,n_i\}$ and that $\mathbf{N} = (N_{i,\mathbf{u}})_{i\in[d],\mathbf{u}\in\mathbf{U}}$ denotes a tuple of nonnegative integers satisfying $n_i = \sum_{\mathbf{u}\in\mathbf{U}} N_{i,\mathbf{u}}$ and $k'_{ij} = \sum_{\mathbf{u}\in\mathbf{U}} u_j N_{i,\mathbf{u}}$ for $i,j \in [d]$. Denote by $\mathcal{L}(\mathbf{N})$ the subset of $\mathcal{L}$ in which the forests have $N_{i,\mathbf{u}}$ vertices of type i with indegree type $\mathbf{u}$ for $i \in [d]$, $\mathbf{u} \in \mathbf{U}$. Set $N(k) = \sum_{i,j\in[d],u_j=k} N_{i,\mathbf{u}} + \sum_{i=1}^{d} \mathbf{1}_{\{k\}}(r_i)$, $k = 0,1,\dots$. The following result has been proved in [BM14], see Proposition 12 therein.

Proposition 3.4.4 (Enumeration of labeled plane forests by given the number of vertices of each indegree type) **:** Given a tuple of nonnegative integers $\mathbf{N} = (N_{i,\mathbf{u}})_{i\in[d],\mathbf{u}\in\mathbf{u}}$, satisfying the above assumptions, the number of forests in $\mathcal{L}(\mathbf{N})$ is

$$|\mathcal{L}(\mathbf{N})| = \frac{\prod_{j=1}^{d}(n_j)!(n_j-1)!}{\prod_{i\in[d],\mathbf{u}\in\mathbf{U}}(N_{i,\mathbf{u}})!\prod_{k\geq 0}(k!)^{N(k)}}\det(-k_{ij})\,. \tag{3.10}$$

Démonstration. For any indegree tuple $\mathbf{c}$ such that $\mathcal{L}(\mathbf{c}) \subset \mathcal{L}(\mathbf{N})$, (3.7) can be rewritten as

$$|\mathcal{L}(\mathbf{c})| = \frac{\prod_{j=1}^{d}(n_j-1)!}{\prod_{k\geq 0}(k!)^{N(k)}}\det(-k_{ij})\,.$$

Since there are $\prod_{i\in[d]}(n_i)!/\prod_{i\in[d],\mathbf{u}\in\mathbf{U}}(N_{i,\mathbf{u}})!$ different indegree tuples $\mathbf{c}$ for each given $\mathbf{N}$, the number of forests in $\mathcal{L}(\mathbf{N})$ is the product of the above two numbers, which gives (3.10). □

Proof of Theorem 3.3.2. Recall that $N(k) = \sum_{i,j\in[d],u_j=k} N_{i,\mathbf{u}} + \sum_{i=1}^{d}\mathbf{1}_{\{k\}}(r_i)$, $k = 0,1,\dots$ and define the canonical map $\Psi : \mathcal{L}(\mathbf{N}) \to \mathcal{F}_d^{k_{ij},\mathbf{n}}(\mathbf{N})$, where for $\mathbf{f} \in \mathcal{L}(\mathbf{N})$, $\Psi(\mathbf{f})$ is the plane forest obtained by removing the labels of vertices in $\mathbf{f}$. Let $\mathbf{f}' \in \mathcal{F}_d^{k_{ij},\mathbf{n}}(\mathbf{N})$, and observe that for any $\mathbf{f}_1, \mathbf{f}_2 \in \Psi^{-1}(\mathbf{f}')$ with respective indegree tuples $\mathbf{c}^1 = (c^1_{i,j,k})_{i,j\in[d],k\in[n_i]}$ and $\mathbf{c}^2 = (c^2_{i,j,k})_{i,j\in[d],k\in[n_i]}$, we have $\prod_{i,j\in[d],k\in[n_i]} c^1_{i,j,k}! = \prod_{i,j\in[d],k\in[n_i]} c^2_{i,j,k}!$ Let us denote by $\prod_{i,j\in[d],k\in[n_i]} c_{i,j,k}!$ this common value, then we readily check that

$$\left|\Psi^{-1}(\mathbf{f}')\right| = \frac{\prod_{j=1}^{d} n_j!}{\prod_{i=1}^{d} r_i! \prod_{i,j\in[d],k\in[n_i]} c_{i,j,k}!} = \frac{\prod_{j=1}^{d} n_j!}{\prod_{k\geq 0}(k!)^{N(k)}}.$$

Then we derive from (3.10) that

$$\left|\mathscr{F}_d^{k_{ij},\mathbf{n}}(\mathbf{N})\right| = \frac{|\mathcal{L}(\mathbf{N})|}{|\Psi^{-1}(\mathbf{f}')|} = \frac{\prod_{j=1}^{d}(n_j-1)!\det(-k_{ij})}{\prod_{i\in[d],\mathbf{u}\in\mathbf{U}}(N_{i,\mathbf{u}})!}.$$

□

3.5 The Lagrange-Good inversion formula

Since the original paper by Good [Goo60], the multivariate extension of Lagrange inversion formula has been widely studied by many authors. We refer to [GK97, Ges87, BR98, BCLL03] for different forms of Lagrange-Good inversion formula and proofs. The arborescent form of this result is introduced in [Ges87] and [GK97], and is based on the notion of derivative with respect to a directed graph, see Definition 3.5.1 below. It is then proved to be equivalent to the classical form. Here we will consider the arborescent form of this formula, as it fits properly to our setting. We will show that Theorem 3.1.2 implies the Lagrange-Good inversion formula. Although the latter is applicable for formal power series, here we only set up this formula for generating functions of probability distributions.

Definition 3.5.1 (Definition 1 of [BCLL03]) **:** Let $\mathcal{G}$ be a directed graph having $V = \{0,1,2,,\cdots,d\}$ as set of vertices and $E \subset V\times V$ as set of arcs(= directed edges), with the property that 0 has outdegree $d^+(0)$ equals to 0. Let $\mathbf{g}(x) = (g_0(\mathbf{x}), g_1(\mathbf{x}), g_2(\mathbf{x}), \cdots, g_d(\mathbf{x}))$ be a vector of formal power series in $\mathbf{x} = (x_1,\cdots,x_d)$. We define the derivative of $\mathbf{g}(x)$ according to $\mathcal{G}$ by

$$\frac{\partial \mathbf{g}(x)}{\partial \mathcal{G}} = \prod_{j\in V}\left\{\left(\prod_{(i,j)\in E}\frac{\partial}{\partial x_i}\right)g_j(\mathbf{x})\right\}.$$

An elementary forest is a d-type forest that contains exactly one vertex of each type. In particular, each elementary forest contains exactly d vertices and is coded by the d couples $(j_i,i), i\in[d]$ where j_i is the type of the parent of the vertex of type i. If the vertex of type i is a root then we set $j_i = 0$. Let D be the set of vectors $(j_1,j_2,\cdots,j_d)$,

$0 \leq j_i \leq d$ such that $(j_i, i), i \in [d]$ codes an elementary forest.

Definition 3.5.2 : Denote by $g_i(\mathbf{x})$ a generating function of a probability distribution on $\mathbb{Z}_+^d$, $i = 0, 1, \cdots, d$. Set $\mathbf{g}(\mathbf{x}) = (g_0(\mathbf{x}), g_1(\mathbf{x}), g_2(\mathbf{x}), \cdots, g_d(\mathbf{x}))$. Let $\mathbf{j} = (j_1, j_2, \cdots, j_d)$ be a vector in D. Define the derivative of $\mathbf{g}(\mathbf{x})$ with respect to $\mathbf{j}$ by

$$\frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{j}} = \prod_{k=0}^{d} \left\{ \left(\prod_{j_i=k} \frac{\partial}{\partial x_i} \right) g_k(\mathbf{x}) \right\},$$

where $\left(\prod_{j_i=k} \frac{\partial}{\partial x_i} \right)$ is equal to the identical operator when $\{i;\ j_i = k\} = \emptyset$.

According to the definition of D, there exists a unique directed tree corresponding to any $\mathbf{j} \in D$. As a consequence and from Definitions 3.5.1 and 3.5.2, we see that the derivative of $\mathbf{g}(\mathbf{x})$ with respect to $\mathbf{j}$ is equal to the derivative of $\mathbf{g}(\mathbf{x})$ with respect to the corresponding tree of $\mathbf{j}$.

For example, for $d = 2$, there are three elementary trees : $\mathbf{j}_1 = (0, 0)$, $\mathbf{j}_2 = (2, 0)$, $\mathbf{j}_3 = (0, 1)$. The derivatives of the vector function $\mathbf{g}(x) = (g_0(\mathbf{x}), g_1(\mathbf{x}), g_2(\mathbf{x}))$ according to the vectors or the trees are

$$\frac{\partial \mathbf{g}(x)}{\partial \mathbf{j}_1} = \frac{\partial^2 g_0}{\partial x_1 \partial x_2} \cdot g_1 \cdot g_2, \qquad \frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{j}_2} = \frac{\partial g_0}{\partial x_2} \cdot g_1 \cdot \frac{\partial g_2}{\partial x_1}, \qquad \frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{j}_3} = \frac{\partial g_0}{\partial x_1} \cdot \frac{\partial g_1}{\partial x_2} \cdot g_2.$$

Let $f_i(\mathbf{x})$ denote the generating function of ν_i, $i \in [d]$ and let ν_0 be the Dirac measure on $(r_1, \cdots, r_d)$, that is $\nu_0 = \delta_{(r_1, \cdots, r_d)}$. And set $f_0(\mathbf{x}) = x_1^{r_1} \cdots x_d^{r_d}$, $(x_1, \cdots, x_d) \in \mathbb{R}^d$, which is the generating function of ν_0. For any $d-$dimensional nonnegative integer vector $\mathrm{m} = (m_1, m_2, \cdots, m_d)$, set $\mathbf{x}^{\mathrm{m}} = x_1^{m_1} \cdots x_d^{m_d}$. Then for any formal power series $h(\mathbf{x})$ with respect to $\mathbf{x}$, the coefficient of $\mathbf{x}^{\mathrm{m}}$ is denoted by $[\mathbf{x}^{\mathrm{m}}]h(\mathbf{x})$. In the remaining, without lose of generality, we assume that $d-$dimensional vector $\mathbf{n}$ is a positive integer valued vector. In our special setting, Arborescent Good-Lagrange formula can be stated as following.

Theorem 3.5.1 : Let $f_1(\mathbf{x}), \cdots, f_d(\mathbf{x})$ be given generating functions of offspring distributions and let $g_i(\mathbf{x})$, $i \in [d]$ be the generating functions of the total progeny distributions starting with one ancestor of type i. So that $g_i(\mathbf{x}) = x_i f_i(\mathbf{g})$ for $i \in [d]$, where

$\mathbf{g} = (g_1, g_2, \cdots, g_d)$. Then

$$[\mathbf{x}^{\mathbf{n}}] f_0(\mathbf{g}) = \left(\prod_{i=1}^{d} \frac{1}{n_i} \right) [\mathbf{x}^{\mathbf{n-1}}] \sum_{\mathbf{j} \in D} \frac{\partial \left(f_0(\mathbf{x}), f_1^{n_1}(\mathbf{x}), \ldots, f_d^{n_d}(\mathbf{x}) \right)}{\partial \mathbf{j}},$$

where $\mathbf{n} - \mathbf{1} = (n_1 - 1, \ldots, n_d - 1)$.

Démonstration. Note that the random walks $X^{(1)}, \ldots, X^{(d)}$, which is the coding of our branching forest, have distribution

$$\mathbb{P}(X_{n_i}^{i,j} = k_{ij}, j \in [d]) = \nu_i^{*n_i}(k_{i1}, \ldots, k_{i(i-1)}, n_i + k_{ii}, k_{i(i+1)}, \ldots, k_{id}), \quad i \in [d].$$

Then the identity in Theorem 3.1.1 can be rewritten as

$$\begin{aligned}
&\mathbb{P}\Big(X_{n_i}^{i,j} = k_{ij} \ \text{ and } \ T_{\mathbf{r}} = \mathbf{n} \Big) \\
=&\frac{\sum_{(j_1,\ldots,j_d) \in D} \prod_{i=1}^{d} k_{j_i i}}{n_1 n_2 \ldots n_d} \prod_{j=1}^{d} \nu_j^{*n_j}(k_{j1}, \ldots, k_{j(j-1)}, n_j + k_{jj}, k_{j(j+1)}, \ldots, k_{jd}) \\
=&\left(\prod_{i=1}^{d} \frac{1}{n_i} \right) \sum_{(j_1,\ldots,j_d) \in D} \Bigg\{ \left(\prod_{\{i; j_i = 0\}} r_i \right) \times \\
&\qquad \prod_{j=1}^{d} \left(\prod_{\{i; j_i = j\}} k_{ji} \right) \nu_j^{*n_j}(k_{j1}, \ldots, k_{j(j-1)}, n_j + k_{jj}, k_{j(j+1)}, \ldots, k_{jd}) \Bigg\},
\end{aligned} \tag{3.11}$$

where $\left(\prod_{\{i; j_i = j\}} k_{ji} \right) = 1$ when the product is taken over an empty set. Set the vectors $\mathrm{I}_j = (k_{j1}, \ldots, k_{j(j-1)}, n_j + k_{jj}, k_{j(j+1)}, \ldots, k_{jd})$, $j \in [d]$, $\mathrm{I}_0 = (r_1, \ldots, r_d)$. Then $\prod_{\{i; j_i = 0\}} r_i = [\mathbf{x}^{\mathrm{I}'_0}] \left(\prod_{\{i; j_i = 0\}} \frac{\partial}{\partial x_i} \right) f_0(\mathbf{x})$, and for $j \in [d]$,

$$\left(\prod_{\{i; j_i = j\}} k_{ji} \right) \nu_j^{*n_j}(k_{j1}, \ldots, k_{j(j-1)}, n_j + k_{jj}, k_{j(j+1)}, \ldots, k_{jd}) = [\mathbf{x}^{\mathrm{I}'_j}] \left(\prod_{\{i; j_i = j\}} \frac{\partial}{\partial x_i} \right) f_j^{n_j}(\mathbf{x}),$$

where $I'_{ji} = I_{ji} - 1$ when $i \in \{i; j_i = j\}$, and $I'_{ji} = I_{ji}$ when $i \notin \{i; j_i = j\}$, $j = 0, 1, \ldots, d$. Since for $(j_1, \ldots, j_d) \in D$, each $i \in [d]$ appears exactly once in the sets $\{i; j_i = j\}, j = 0, 1, \ldots, d$, $\sum_{j=0}^{d} \mathrm{I}'_j = \mathbf{n} - \mathbf{1}$. Note that $\{\mathrm{I}_j; j \in [d]\}$ depend on $(k_{ij})_{i,j \in [d]}$.

Now fix $\mathbf{n}$, $\mathbf{r}$ and $\mathbf{j} = (j_1, \dots, j_d) \in D$. And define the set

$$\mathbf{k} = \left\{ (k_{ij})_{i,j\in[d]};\ k_{ij} \in \mathbb{Z}_+, i \neq j, -k_{ii} = \sum_{j\neq i} k_{ji} + r_i \leq n_i \right\}.$$

Take a sum for the term in the brace in (3.11) and express it in terms of characteristic functions :

$$\begin{aligned}
&\sum_{(k_{ij})\in\mathbf{k}} \Big(\prod_{\{i;j_i=0\}} r_i \Big) \prod_{j=1}^{d} \Big(\prod_{\{i;j_i=j\}} k_{ji} \Big) \nu_j^{*n_j}(k_{j1}, \dots, k_{j(j-1)}, n_j + k_{jj}, k_{j(j+1)}, \dots, k_{jd}) \\
&= [\mathbf{x}^{\mathrm{I}'_0}] \Big(\prod_{\{i;j_i=0\}} \frac{\partial}{\partial x_i} \Big) f_0(\mathbf{x}) \sum_{(k_{ij})\in\mathbf{k}} \prod_{j=1}^{d} [\mathbf{x}^{\mathrm{I}'_j}] \Big(\prod_{\{i;j_i=j\}} \frac{\partial}{\partial x_i} \Big) f_j^{n_j}(\mathbf{x}) \\
&= [\mathbf{x}^{\mathrm{I}'_0}] \Big(\prod_{\{i;j_i=0\}} \frac{\partial}{\partial x_i} \Big) f_0(\mathbf{x}) \sum_{\sum_{j=0}^{d} \mathrm{I}'_j = \mathbf{n}-\mathbf{1}} \prod_{j=1}^{d} [\mathbf{x}^{\mathrm{I}'_j}] \Big(\prod_{\{i;j_i=j\}} \frac{\partial}{\partial x_i} \Big) f_j^{n_j}(\mathbf{x}) \\
&= [\mathbf{x}^{\mathbf{n}-\mathbf{1}}] \frac{\partial \left(f_0(\mathbf{x}), f_1^{n_1}(\mathbf{x}), \dots, f_d^{n_d}(\mathbf{x}) \right)}{\partial \mathbf{j}}.
\end{aligned}$$

Then we derive from this last computation that,

$$\begin{aligned}
& [\mathrm{x}^{\mathbf{n}}] f_0(\mathrm{g}) = \mathbb{P}\Big(T_{\mathbf{r}} = \mathbf{n} \Big) \\
= & \sum_{(k_{ij})\in\mathrm{K}} \mathbb{P}\Big(X_{n_i}^{i,j} = k_{ij} \ \text{ and } \ T_{\mathbf{r}} = \mathbf{n} \Big) \\
= & \sum_{(k_{ij})\in\mathbf{k}} \sum_{\mathbf{j}\in D} \frac{\prod_{\{i;j_i=0\}} r_i}{\prod_{i=1}^{d} n_i} \prod_{j=1}^{d} \Big(\prod_{\{i;j_i=j\}} k_{ji} \Big) \nu_j^{*n_j}(k_{j1}, \dots, k_{j(j-1)}, n_j + k_{jj}, k_{j(j+1)}, \dots, k_{jd}) \\
= & \left(\prod_{i=1}^{d} \frac{1}{n_i} \right) \sum_{\mathbf{j}\in D} [\mathbf{x}^{\mathbf{n}-\mathbf{1}}] \frac{\partial \left(f_0(\mathbf{x}), f_1^{n_1}(\mathbf{x}), \dots, f_d^{n_d}(\mathbf{x}) \right)}{\partial \mathbf{j}}.
\end{aligned}$$

□

Chapitre 4

A note on vertices with a given degree in multitype branching forests

4.1 Introduction

A *multitype branching forest* can be thought of as representing the genealogy of a population consisting of individuals of several types where each individual lives a unit of time and gives birth to a random number of children at the end of its life, independently of the others. Following the definition of *multitype branching forests* given in [CL15], an individual in a multitype population having a given offspring corresponds to a vertex with a given inner-degree of the multitype branching forest which describes the genealogy of this population.

In this note, we will consider individuals with a given offspring of a multitype population. However, for the sake of simplicity, we will actually give proofs only in the case where individuals have no child. The latters correspond to the vertices with inner-degree equal to 0 in the genealogical forest and we will call them the *leaves* of the forest. We

first give in section 4.4 the law of the total number of leaves of a multitype branching forest in terms of some mutivariate random walks, which generalizes a result given in [Kor12] in to the multidimensional case. As an application, we obtain an asymptotical result which is given in Corollary 4.4.2. Similar results for the number of vertices with a given inner-degree are presented at the end of our paper.

4.2 Preliminaires

4.2.1 General notation

Let $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$, $\mathbb{N} = \{1, 2, \ldots\}$, $d \in \mathbb{N}$ and set $[d] = \{1, \ldots, d\}$. A vector in $\mathbb{R}^d$ will be denoted by $\mathbf{x} = (x_1, \ldots, x_d)$, and its transpose is denoted by $\mathbf{x}^T$. We denote by $\mathbf{e}_i$ the i-th unit vector in $\mathbb{R}^d$. We also write $\mathbf{0} = (0, \ldots, 0)$ and $\mathbf{1} = (1, \ldots, 1)$. The notation $\mathbf{x} \cdot \mathbf{y}$ stands for the scalar product in $\mathbb{R}^d$, and we will use the partial order : $\mathbf{x} \leq \mathbf{y}$ iff $x_i \leq y_i, \forall i \in [d]$.

4.2.2 Multitype branching forest

We will first recall from [CL15] the definition of *multitype branching forests* and their coding from integer valued ssequences.

A (plane) forest $\mathbf{f}$ is a finite directed planar graph with no loops on a possibly infinite and non empty set of vertices $\mathbf{v} = \mathbf{v}(\mathbf{f})$, such that each vertex has a finite inner-degree and an outer-degree equals to 0 or 1. The connected components of a forest are called the *trees*. In a tree $\mathbf{t}$, the only vertex with outer-degree equal to 0 is called the *root* of $\mathbf{t}$. The roots of the connected components of a forest $\mathbf{f}$ are called the roots of $\mathbf{f}$. For two vertices u and v of a forest $\mathbf{f}$, if (u, v) is a directed edge of $\mathbf{f}$, then we say that u is a *child* of v, or that v is the *parent* of u.

We first give an order to the trees of the forest $\mathbf{f}$ and denote them by $\mathbf{t}_1(\mathbf{f}), \mathbf{t}_2(\mathbf{f}), \ldots, \mathbf{t}_k(\mathbf{f}), \ldots$ (we will usually write $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_k, \ldots$ if no confusion is possible). Then we rank (a part of) the vertices of $\mathbf{f}$ according to the breadth first search order, by ranking first the

vertices of $\mathbf{t}_1$, then the vertices of $\mathbf{t}_2$, and so on, see the labeling of the two forests in Figure 4.2. Note that if $\mathbf{t}_k$, for $k \geq 1$ is the first infinite tree, then the vertices of $\mathbf{t}_{k+1}, \ldots$ have no label according to this procedure.

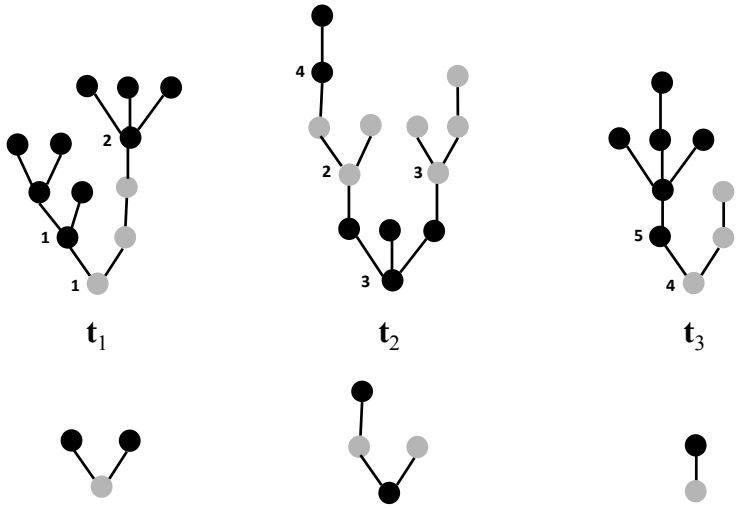

FIGURE 4.1 – On the top, a discrete 2-type forest. Roots of clusters are ranked in the breadth first search order of the forest. The rank is written on the left of these roots.

To each forest $\mathbf{f}$, we associate the application $c_{\mathbf{f}} : \mathbf{v}(\mathbf{f}) \to [d]$ such that if $u_i, u_{i+1}, \ldots, u_{i+j} \in \mathbf{v}(\mathbf{f})$ have the same parent and are placed from left to right, then $c_{\mathbf{f}}(u_i) \leq c_{\mathbf{f}}(u_{i+1}) \leq \cdots \leq c_{\mathbf{f}}(u_{i+j})$. For $v \in \mathbf{v}(\mathbf{f})$, the integer $c_{\mathbf{f}}(v)$ is called the *type* (or the *color*) of v. The couple $(\mathbf{f}, c_{\mathbf{f}})$ is called a *d-type forest*. When no confusion is possible, we will simply write $\mathbf{f}$. From now on, we will refer to *inner-degree* of a vertex $v \in \mathbf{v}(\mathbf{f})$ the d-dimensional vector whose ith-coordinate is the number of i-type children of v. The set of d-type forests will be denoted by $\mathcal{F}_d$.

A *cluster* or a *subtree of type* $i \in [d]$ of a d-type forest $(\mathbf{f}, c_{\mathbf{f}}) \in \mathcal{F}_d$ is a maximal connected subgraph of $(\mathbf{f}, c_{\mathbf{f}})$ whose all vertices are of type i. Formally, $\mathbf{t}$ is a cluster of type i of $(\mathbf{f}, c_{\mathbf{f}})$, if it is a connected subgraph whose all vertices are of type i and such that either the root of $\mathbf{t}$ has no parent or the type of its parent is different from i. Mo-

reover, if the parent of a vertex $v \in \mathbf{v}(\mathbf{t})^c$ belongs to $\mathbf{v}(\mathbf{t})$, then $c_{\mathbf{f}}(v) \neq i$. Clusters of type i in $\mathbf{t}_1$ are ranked according to the order of their roots in the breadth first search order of $\mathbf{t}_1$, see Figures 4.1 and 4.2. Then we continue by ranking clusters of type i in $\mathbf{t}_2$, and so on. We denote by $\mathbf{t}_1^{(i)}, \mathbf{t}_2^{(i)}, \ldots, \mathbf{t}_k^{(i)}, \ldots$ the sequence of clusters of type i in $(\mathbf{f}, c_{\mathbf{f}})$. The forest $\mathbf{f}^{(i)} := \{\mathbf{t}_1^{(i)}, \mathbf{t}_2^{(i)}, \ldots, \mathbf{t}_k^{(i)}, \ldots\}$ is called *the subforest of type* i of $(\mathbf{f}, c_{\mathbf{f}})$. We denote by $u_1^{(i)}, u_2^{(i)}, \ldots$ the elements of $\mathbf{v}(\mathbf{f}^{(i)})$, ranked in the breadth first search order of $\mathbf{f}^{(i)}$. The subforests of the 2-type forest given in Figure 4.1 are represented in Figure 4.2.

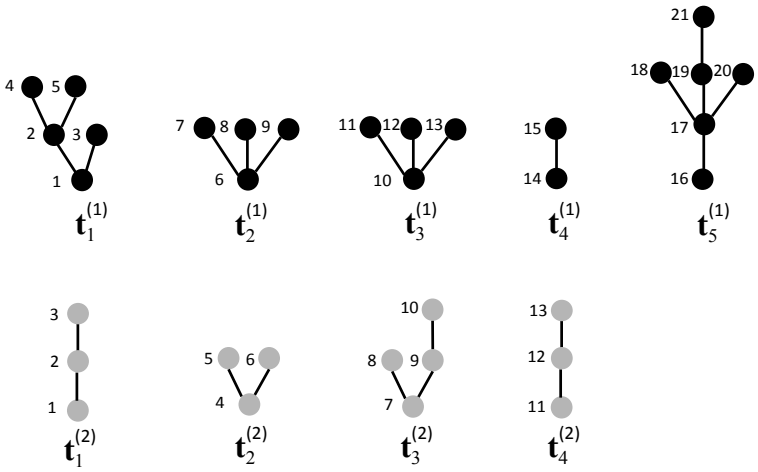

FIGURE 4.2 – The subforests of the 2-type forest given in Figure 4.1 with their deapth first search labeling.

For a forest $(\mathbf{f}, c_{\mathbf{f}}) \in \mathcal{F}_d$ and $u \in \mathbf{v}(\mathbf{f})$, when no confusion is possible, we denote by $p_i(u)$ the number of children of type i of u. For each $i \in [d]$, let $n_i \in \mathbb{Z}_+ \cup \{\infty\}$ be the number of vertices in the subforest $\mathbf{f}^{(i)}$ of $(\mathbf{f}, c_{\mathbf{f}})$. Then let us define the d-dimensional chain $x^{(i)} = (x^{i,1}, \ldots, x^{i,d})$, with length n_i and whose values belong to the set $\mathbb{Z}^d$, by $x_0^{(i)} = 0$ and if $n_i \geq 1$,

$$x_{n+1}^{i,j} - x_n^{i,j} = p_j(u_{n+1}^{(i)}), \quad \text{if } i \neq j \quad \text{and} \quad x_{n+1}^{i,i} - x_n^{i,i} = p_i(u_{n+1}^{(i)}) - 1, \quad 0 \leq n \leq n_i - 1, \tag{4.1}$$

where $(u_n^{(i)})_{n\geq 1}$ is the labeling of the subforest $\mathbf{f}^{(i)}$ in its own breadth first search order. Note that from Theorem 2.7 of [CL15], the data of the chains $x^{(1)}, \ldots, x^{(d)}$ together with the sequence of ranked roots of $(\mathbf{f}, c_{\mathbf{f}})$, allow us to reconstruct this forest.

A multitype branching forest will be considered as a random variable defined on some probability space $(\Omega, \mathcal{G}, P)$ and with values in $\mathcal{F}_d$.

4.3 Multivariate Ballot Theorem

In this section, we will recall the notion of *cyclical permutation* and also the so called *Multivariate Ballot Theorem* which have been given in [CL15]. These knowledges will be necessary to establish our main result.

Let us consider a sequence $x = (x^1, \ldots, x^d)$ where $x^i = (x^{i,1}, \ldots, x^{i,d})$ is a $\mathbb{Z}^d$-valued sequence defined on some interval $\{0, 1, 2, \ldots, n_i\},\ n_i \in \mathbb{N}$, which satisfies $x_0^i = \mathbf{0}$ and if $n_i \geq 1$ then

- for $i \neq j$, the sequence $x_n^{i,j}$ is nondecreasing in n
- $x_{n+1}^{i,i} - x_n^{i,i} \geq -1,\ \forall n, i.$

Then a $\mathbf{n}$-*cyclical permutation* of x is defined by

$$x_{\mathbf{q},\mathbf{n}} := (x^1_{q_1,n_1}, \ldots, x^d_{q_d,n_d}),\ \forall \mathbf{q} = (q_1, \ldots, q_d) \leq \mathbf{n} - \mathbf{1},$$

where $\mathbf{n} = (n_1, \ldots, n_d)$ and for $i \in [d]$

$$x^i_{q_i,n_i}(h) = \begin{cases} x^i(h + q_i) - x^i(h) & \text{if } 0 \leq h \leq n_i - q_i \\ x^i(h + q_i - n_i) + x^i(q_i) - x^i(n_i) & \text{if } n_i - q_i \leq h \leq n_i \\ x^i(h) & \text{if } h \geq n_i. \end{cases}$$

Intuitively, the transformation from x^i to $x^i_{q_i,n_i}$ consists in inverting the parts $\{x^i(h),\ 0 \leq h \leq q_i\}$ and $\{x^i(h),\ q_i \leq h \leq n_i\}$ in such a way that $x^i_{q_i,n_i}$ has the same values as x at 0 and n_i. We now state the *Multivariate Ballot Theorem* as it is given in [CL15, Theorem 3.4].

Theorem 4.3.1 : Let $Y = (Y^1, \dots, Y^d)$ be a stochastic process, with $Y^i = \left(Y^{i,1}, \dots, Y^{i,d}\right)$ and $Y_0^i = \mathbf{0}$. We assume that the coordinates $Y^{i,j}$ for $i \neq j$ are $\mathbb{Z}_+$ valued, non-decreasing and that the coordinates $Y^{i,i}$ are $\mathbb{Z}$ valued and downward skip free. Fix $\mathbf{n} = (n_1, \dots, n_d) \in \mathbb{N}^d$, then we assume further that the process Y is $\mathbf{n}$-cyclically exchangeable, that is for any $\mathbf{q} = (q_1, \dots, q_d) \in \mathbb{Z}_+^d$ such that $\mathbf{q} \leq \mathbf{n} - \mathbf{1}$,

$$Y_{\mathbf{q},\mathbf{n}} \overset{\text{(law)}}{=} Y,$$

where $Y_{\mathbf{q},\mathbf{n}}$ is defined as above for deterministic functions. Then for any $\mathbf{r} = (r_1, \dots, r_d) \in \mathbb{Z}_+^d$ such that $\mathbf{r} > 0$ and $k_{ij},\ i, j \in [d]$ such that $k_{ij} \in \mathbb{Z}_+$ for $i \neq j$ and $k_{ii} = -r_i - \sum_{j \neq i} k_{ji}$,

$$\begin{aligned} &\mathbb{P}\left(Y_{n_i}^{ij} = k_{ij},\ i, j \in [d] \text{ and } \mathbf{n} \text{ is the smallest solution of } (\mathbf{r}, Y)\right) \\ &= \frac{\det(-k_{ij})}{n_1 \dots n_d} \mathbb{P}\left(Y_{n_i}^{ij} = k_{ij},\ i, j \in [d]\right), \end{aligned}$$

where $\mathbf{n}$ is called *the smallest solution* of $(\mathbf{r}, Y)$ if $\sum_{i=1}^d Y_{n_i}^i = -\mathbf{r}$ and $\mathbf{n} \leq \mathbf{n}'$ for all other solutions $\mathbf{n}'$ of this system.

4.4 Law of the total number of leaves

Let us now consider a multitype branching forest F which represents the evolution of a population consisting of individuals of d different types, whose offspring distribution $\nu = (\nu_1, \dots, \nu_d)$ is nonsingular, primitive and (sub)critical .

For $i, j \in [d]$, let N_i be the total number of type-i individuals, τ_i be the number of type-i individuals who have no child (we shall call them *leaves* of the multitype branching forest F) and M_{ij} be the total number of type-j individuals whose parent is of type i, up to the extinction time.

To the forest F, we associate the random sequences $X = \{X^i, i \in [d]\}$, where $X^i = \{(X_n^{i,1}, \dots, X_n^{i,d}), 0 \leq n \leq n_i\}$, which are constructed as in (4.1). It has been proved in [CL15], Theorem 3.1 that $X^{(i)}$, $i \in [d]$ are independent random walks whose

step distribution $\tilde{\nu}_i$ is defined by

$$\tilde{\nu}_i(k_1,\dots,k_d) := \nu_i(k_1,\dots,k_{i-1},k_i+1,k_{i+1},\dots,k_d)\,, \quad \text{for all } (k_1,\dots,k_d)\in\mathbb{Z}_+^d, \tag{4.2}$$

and stopped at the smallest solution $(N_1,\dots,N_d)$ of the system

$$r_j + \sum_{i=1}^{d} X^{i,j}(N_i) = 0\,, \quad j\in[d], \tag{4.3}$$

where r_i is the total number of trees in this forest whose root is of type i.

The following result gives the law of the total number of leaves jointly with the total number of vertices in the multitype branching forest F.

Theorem 4.4.1 : For all non-negative integers $q_i, r_i, n_i, k_{ij},\ i,j\in[d]$ such that $\sum_{i=1}^d r_i \geq 1, n_i \leq q_i$ and $k_{ii} = -r_i - \sum_{j\neq i} k_{ji}$,

$$\begin{aligned}&\mathbb{P}_{\mathbf{r}}\left(N_1=q_1,\dots,N_d=q_d,\tau_i=n_i,M_{ij}=k_{ij},i,j\in[d],i\neq j\right)\\ &=\frac{\det(K)}{\bar{q}_1\dots\bar{q}_d}\prod_{i=1}^{d}\mathbb{P}\left(S^i_{q_i}=n_i\right)\mathbb{P}\left(W^i_{q_i-n_i}=\mathbf{k}_i+n_i\mathbf{e}_i\right),\end{aligned}$$

where $\mathbf{r}=(r_1,\dots,r_d)$, $\mathbf{k}_i=(k_{i1},\dots,k_{id}), \bar{q}_i = q_i\vee 1$, $\mathbf{e}_i$ is the i-th unit vector of $\mathbb{R}^d$, K is the matrix $(-k_{ij})$ to which we removed the line i and the column i for all i such that $q_i=0$, S^i is a random walk whose step distribution is a Bernoulli law of parameter $\nu_i(\mathbf{0})$ and W^i is a random walk with jumps distributed according to $\eta_i(\mathbf{z}) = \frac{\nu_i(\mathbf{z}+\mathbf{e}_i)}{1-\nu_i(\mathbf{0})}$ for every $\mathbf{z}\in\mathbb{N}^d$.

Démonstration. From Theorem 3.1 and Theorem 3.4 in [CL15] recalled in the previous section and the fact that the total number of leaves of each type is invariant under cyclical permutations of $X=(X^1,\dots,X^d)$, we have

$$\begin{aligned}&\mathbb{P}_{\mathbf{r}}\left\{N_1=q_1,\dots,N_d=q_d,\tau_i=n_i,M_{ij}=k_{ij},\forall i\neq j\right\}\\ &=\mathbb{P}\left\{X^i_{q_i}=\mathbf{k}_i \text{ and } \mathbf{q} \text{ is the smallest solution of } (\mathbf{r},X),\ \tau_i=n_i,\ \forall i\in[d]\right\}\\ &=\frac{det(K)}{\bar{q}_1\dots\bar{q}_d}\mathbb{P}\left\{X^i_{q_i}=\mathbf{k}_i,\ \tau_i=n_i,\forall i\in[d]\right\},\end{aligned}$$

where $\mathbf{q} = (q_1, \ldots, q_d)$.

We will now compute $\mathbb{P}\left(X^i_{q_i} = \mathbf{k}_i,\ \tau_i = n_i, \forall i \in [d]\right)$.

Let $Y^i_k := X^i_k - X^i_{k-1}, \forall 1 \leq k \leq q_i,\ i \in [d]$, then we have :

$$\begin{aligned}
&\mathbb{P}\left(X^i_{q_i} = \mathbf{k}_i,\ \tau_i = n_i, \forall i \in [d]\right) \\
&= \prod_{i=1}^{d} \sum_{1 \leq h_1 < \ldots < h_{n_i} \leq q_i} \mathbb{P}\left(Y^i_{h_1} = \cdots = Y^i_{h_{n_i}} = -\mathbf{e_i}\right) \\
&\qquad \times \mathbb{P}\left(\sum_{k \notin \{h_1, \ldots, h_{n_i}\}} Y^i_k = \mathbf{k}_i + n_i \mathbf{e}_i;\ Y^i_k \neq -\mathbf{e_i}, \forall k \notin \{h_1, \ldots, h_{n_i}\}\right) \\
&= \prod_{i=1}^{d} \sum_{1 \leq h_1 < \ldots < h_{n_i} \leq q_i} \mathbb{P}\left(Y^i_{h_1} = \cdots = Y^i_{h_{n_i}} = -\mathbf{e_i};\ Y^i_k \neq -\mathbf{e_i}, \forall k \notin \{h_1, \ldots, h_{n_i}\}\right) \\
&\qquad \times \mathbb{P}\left(\sum_{k \notin \{h_1, \ldots, h_{n_i}\}} Y^i_k = k_i + n_i \mathbf{e}_i \middle| Y^i_k \neq -\mathbf{e_i}, \forall k \notin \{h_1, \ldots, h_{n_i}\}\right)
\end{aligned}$$

It's not hard to show that the last probability is equal to $\mathbb{P}\left(W^i_{q_i - n_i} = k_i + n_i \mathbf{e}_i\right)$, and then we have :

$$\begin{aligned}
&\mathbb{P}\left(X^i_{q_i} = \mathbf{k}_i,\ \tau_i = n_i, \forall i, j = 1, \ldots, d\right) \\
&= \prod_{i=1}^{d} \mathbb{P}\left(W^i_{q_i - n_i} = \mathbf{k}_i + n_i \mathbf{e}_i\right) \\
&\times \sum_{1 \leq h_1 < \ldots < h_{n_i} \leq q_i} \mathbb{P}\left(Y^i_{h_1} = \cdots = Y^i_{h_{n_i}} = -\mathbf{e_i};\ Y^i_k \neq -\mathbf{e_i}, \forall k \notin \{h_1, \ldots, h_{n_i}\}\right) \\
&= \prod_{i=1}^{d} \mathbb{P}\left(W^i_{q_i - n_i} = \mathbf{k}_i + n_i \mathbf{e}_i\right) \mathbb{P}\left(S^i_{q_i} = n_i\right).
\end{aligned}$$

□

Corollary 4.4.2 : Suppose that the offspring distribution ν is critical, admits moments of order $d+1$ and its covariance matrices $\mathbf{\Sigma}^i = (\Sigma^i_{jk})_{1 \leq j,k \leq d},\ i \in [d]$ are all positive-

definite. We also assume that $0 < \nu_i(\mathbf{0}) < 1$ for all $i \in [d]$. Then for all $\mathbf{r} \in \mathbb{Z}_+^d$,

$$\mathbb{P}_{\mathbf{r}}(\tau_i = \lfloor \nu_i(\mathbf{0}) n v_i \rfloor, N_i = \lfloor n v_i \rfloor, i \in [d]) = o\left(n^{-(d+1)}\right).$$

Démonstration. From Theorem 4.4.1, we have :

$$\begin{aligned} & n^{d+1}\mathbb{P}_{\mathbf{r}}(\tau_i = \lfloor \nu_i(\mathbf{0}) n v_i \rfloor, N_i = \lfloor n v_i \rfloor, i \in [d]) \\ & = n^{d+1} \sum_{k_{ij}>0, i\neq j} \frac{\det(K)}{\prod_{i=1}^d \lfloor n v_i \rfloor} \prod_{i=1}^d \mathbb{P}\left(S^i_{\lfloor n v_i \rfloor} = \lfloor \nu_i(\mathbf{0}) n v_i \rfloor\right) \\ & \qquad \times \mathbb{P}\left(W^i_{\lfloor (1-\nu_i(\mathbf{0})) n v_i \rfloor} = \mathbf{k}_i + \lfloor \nu_i(\mathbf{0}) n v_i \rfloor \mathbf{e}_i\right). \end{aligned}$$

A local limit theorem (see [Sto67, Corollary 1]) in the one dimensional case implies in particular that

$$\lim_{n\to+\infty} n^{d/2} \prod_{i=1}^d \mathbb{P}\left(S^i_{\lfloor n v_i \rfloor} = \lfloor \nu_i(\mathbf{0}) n v_i \rfloor\right) = \frac{1}{(2\pi)^{d/2}\left(\prod_{i=1}^d \nu_i(\mathbf{0})(1-\nu_i(\mathbf{0}))v_i\right)^{1/2}}.$$

It remains to prove that the term

$$n^{d/2+1} \sum_{k_{ij}>0, i\neq j} \frac{\det(K)}{\prod_{i=1}^d \lfloor n v_i \rfloor} \prod_{i=1}^d \mathbb{P}\left(W^i_{\lfloor (1-\nu_i(\mathbf{0})) n v_i \rfloor} = \mathbf{k}_i + \lfloor \nu_i(\mathbf{0}) n v_i \rfloor \mathbf{e}_i\right)$$

tends to 0. To do this, we will use some techniques from [Pé16].

First, we can rewrite this term as following

$$
\begin{aligned}
&n^{d/2+1} \sum_{k_{ij}>0, i\neq j} \frac{\det(K)}{\prod_{i=1}^d \lfloor nv_i \rfloor} \prod_{i=1}^d \mathbb{P}\left(W^i_{\lfloor (1-\nu_i(\mathbf{0}))nv_i \rfloor} = \mathbf{k}_i + \lfloor \nu_i(\mathbf{0}) nv_i \rfloor \mathbf{e}_i\right) \\
&= \frac{n^{d/2+1}}{\prod_{i=1}^d \lfloor nv_i \rfloor} \mathbb{E}\Bigg[\det \begin{pmatrix} \nu_1(\mathbf{0})\lfloor nv_1 \rfloor \mathbf{e}_1 - W^1_{\lfloor (1-\nu_1(\mathbf{0}))nv_1 \rfloor} \\ \dots \\ \nu_d(\mathbf{0})\lfloor nv_d \rfloor \mathbf{e}_d - W^d_{\lfloor (1-\nu_d(\mathbf{0}))nv_d \rfloor} \end{pmatrix} \\
&\qquad\qquad \times \mathbb{1}_{\sum_{i=1}^d W^i_{\lfloor (1-\nu_i(\mathbf{0}))nv_i \rfloor} = \sum_{i=1}^d \lfloor \nu_i(\mathbf{0}) nv_i \rfloor \mathbf{e}_i - \mathbf{r}} \Bigg] \\
&= \frac{n^{d/2+1}}{\lfloor nv_d \rfloor} \mathbb{E}\left[\det \begin{pmatrix} \nu_1(\mathbf{0})\mathbf{e}_1 - W^1_{\lfloor (1-\nu_1(\mathbf{0}))nv_1 \rfloor}/\lfloor nv_1 \rfloor \\ \dots \\ \nu_{d-1}(\mathbf{0})\mathbf{e}_{d-1} - W^{d-1}_{\lfloor (1-\nu_{d-1}(\mathbf{0}))nv_{d-1} \rfloor}/\lfloor nv_{d-1} \rfloor \\ \mathbf{r} \end{pmatrix} \mathbb{1}_{A_n} \right],
\end{aligned}
$$

where $A_n = \left\{ \sum_{i=1}^d W^i_{\lfloor (1-\nu_i(\mathbf{0}))nv_i \rfloor} = \sum_{i=1}^d \lfloor \nu_i(\mathbf{0}) nv_i \rfloor \mathbf{e}_i - \mathbf{r} \right\}$.

We will first investigate the asymptotic behavior of $\mathbb{P}(A_n)$. From a mutivariate local limit theorem given in [Sto67, Corollary 1], we have

$$
\mathbb{P}(A_n) = \frac{\exp\left(-\frac{(\sum_{i=1}^d \lfloor \nu_i(\mathbf{0}) nv_i \rfloor \mathbf{e}_i - \mathbf{r} - n\mathbf{m}) \boldsymbol{\Sigma}^{-1} (\sum_{i=1}^d \lfloor \nu_i(\mathbf{0}) nv_i \rfloor \mathbf{e}_i - \mathbf{r} - n\mathbf{m})}{2n} \right)}{(2\pi n)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} + o\left(n^{-d/2}\right),
$$

where $\mathbf{m} = \sum_{i=1}^d v_i(\mathbf{m}^i - \mathbf{e}_i) = \mathbf{0}$ since we are in the critical case, and $|\boldsymbol{\Sigma}|$ denote the determinant of the matrix $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_{i,j}, i, j = 1, \dots, d)$ with $\boldsymbol{\Sigma}_{i,j} = \sum_{k=1}^d v_k \boldsymbol{\Sigma}^k_{i,j}$, for every $i, j \in [d]$.

Note also that

$$\begin{aligned}
&\left(\sum_{i=1}^{d}\lfloor\nu_i(\mathbf{0})nv_i\rfloor\mathbf{e}_i-\mathbf{r}\right)\boldsymbol{\Sigma}^{-1}\left(\sum_{i=1}^{d}\lfloor\nu_i(\mathbf{0})nv_i\rfloor\mathbf{e}_i-\mathbf{r}\right)\\
&=\sum_{i,j=1}^{d}(\lfloor\nu_i(\mathbf{0})nv_i\rfloor-r_i)(\lfloor\nu_j(\mathbf{0})nv_j\rfloor-r_j)\boldsymbol{\Sigma}^{-1}_{i,j}\\
&=n^2\sum_{i,j=1}^{d}\nu_i(\mathbf{0})v_i\nu_j(\mathbf{0})v_j\boldsymbol{\Sigma}^{-1}_{i,j}-\sum_{i,j=1}^{d}\boldsymbol{\Sigma}^{-1}_{i,j}\left(r_ir_j+\lfloor n(r_i\nu_i(\mathbf{0})v_i+r_j\nu_j(\mathbf{0})v_j)\rfloor\right)\\
&=n^2\mathbf{x}\boldsymbol{\Sigma}^{-1}\mathbf{x}-\sum_{i,j=1}^{d}\boldsymbol{\Sigma}^{-1}_{i,j}\left(r_ir_j+\lfloor n(r_i\nu_i(\mathbf{0})v_i+r_j\nu_j(\mathbf{0})v_j)\rfloor\right),
\end{aligned}$$

where $\mathbf{x}=(x_1,\ldots,x_d)$ and $x_i=\nu_i(\mathbf{0})v_i, i\in[d]$. And now remark that since $v_k>0, k\in[d]$ and all the matrices $\boldsymbol{\Sigma}^k, k\in[d]$ are positive definite then so is $\boldsymbol{\Sigma}$ and then $\boldsymbol{\Sigma}^{-1}$, why $\mathbf{x}\boldsymbol{\Sigma}^{-1}\mathbf{x}>0$.

Consequently,

$$\lim_{n\to+\infty}n^{d/2}\mathbb{P}(A_n)=0\qquad.\tag{4.4}$$

Now, let us write

$$\begin{aligned}
&\mathbb{E}\left[\det\begin{pmatrix}\nu_1(\mathbf{0})\mathbf{e}_1-W^1_{\lfloor(1-\nu_1(\mathbf{0}))nv_1\rfloor}/\lfloor nv_1\rfloor\\ \cdots\\ \nu_{d-1}(\mathbf{0})\mathbf{e}_{d-1}-W^{d-1}_{\lfloor(1-\nu_{d-1}(\mathbf{0}))nv_{d-1}\rfloor}/\lfloor nv_{d-1}\rfloor\\ \mathbf{r}\end{pmatrix}\mathbb{1}_{A_n}\right]\\
&=\sum_{\sigma\in S_d}\mathrm{sgn}(\sigma)r_{\sigma(d)}\mathbb{E}\left[\prod_{i=1}^{d-1}\left(\nu_i(\mathbf{0})\delta_{i\sigma(i)}-\frac{W^i_{\lfloor(1-\nu_i(\mathbf{0}))nv_i\rfloor,\sigma(i)}}{\lfloor nv_i\rfloor}\right)\mathbb{1}_{A_n}\right]\\
&=\sum_{I\subset\{1,\ldots,d-1\}}\sum_{\sigma\in S_d}\mathrm{sgn}(\sigma)r_{\sigma(d)}\mathbb{E}\left[\prod_{i\in I}\left((1-\nu_i(\mathbf{0}))m_{i\sigma(i)}-\frac{W^i_{\lfloor(1-\nu_i(\mathbf{0}))nv_i\rfloor,\sigma(i)}}{\lfloor nv_i\rfloor}\right)\mathbb{1}_{A_n}\right]\\
&\qquad\qquad\times\prod_{i\notin I}\left(\nu_i(\mathbf{0})\delta_{i\sigma(i)}-(1-\nu_i(\mathbf{0}))m_{i\sigma(i)}\right),
\end{aligned}$$

where S_d is the set of all permutations of the set $\{1,\ldots,d\}$.

Note that $\frac{W^i_{\lfloor(1-\nu_i(\mathbf{0}))nv_i\rfloor,\sigma(i)}}{\lfloor nv_i\rfloor}$ is bounded on the event A_n, then there exists some constant $A>0$ such that for every $\epsilon>0$ and $i,j\in[d]$,

$$\begin{aligned}
&\mathbb{E}\left(\left|(1-\nu_i(\mathbf{0}))m_{ij}-\frac{W^i_{\lfloor(1-\nu_i(\mathbf{0}))nv_i\rfloor,j}}{\lfloor nv_i\rfloor}\right|\mathbb{1}_{A_n}\right)\\
&\leq \epsilon\mathbb{P}(A_n)+A\mathbb{P}\left(\left|(1-\nu_i(\mathbf{0}))m_{ij}-\frac{W^i_{\lfloor(1-\nu_i(\mathbf{0}))nv_i\rfloor,j}}{\lfloor nv_i\rfloor}\right|>\epsilon\right)\\
&\leq \epsilon\mathbb{P}(A_n)+\frac{A}{\epsilon^{d+1}}\mathbb{E}\left(\left|(1-\nu_i(\mathbf{0}))m_{ij}-\frac{W^i_{\lfloor(1-\nu_i(\mathbf{0}))nv_i\rfloor,j}}{\lfloor nv_i\rfloor}\right|^{d+1}\right)\\
&\leq \epsilon\mathbb{P}(A_n)+\frac{A}{\epsilon^{d+1}}\frac{1}{\lfloor nv_i\rfloor^{\frac{d+1}{2}}}\mathbb{E}\left(\left|(1-\nu_i(\mathbf{0}))m_{ij}-W^i_{1-\nu_i(\mathbf{0}),j}\right|^{d+1}\right),
\end{aligned}$$

where the penultimate inequality comes from Markov's inequality and the last one comes from [DJ69, Theorem 2]. This, together with 4.4, implies that for every non-empty set I,

$$\lim_{n\to+\infty} n^{d/2}\sum_{\sigma\in S_d}\mathrm{sgn}(\sigma)r_{\sigma(d)}\mathbb{E}\left[\prod_{i\in I}\left((1-\nu_i(\mathbf{0}))m_{i\sigma(i)}-\frac{W^i_{\lfloor(1-\nu_i(\mathbf{0}))nv_i\rfloor,\sigma(i)}}{\lfloor nv_i\rfloor}\right)\mathbb{1}_{A_n}\right]$$

$$\times\prod_{i\notin I}\left(\nu_i(\mathbf{0})\delta_{i\sigma(i)}-(1-\nu_i(\mathbf{0}))m_{i\sigma(i)}\right)=0.$$

Hence,

$$\lim_{n\to+\infty}\frac{n^{d/2+1}}{\lfloor nv_d\rfloor}\mathbb{E}\left[\det\begin{pmatrix}\nu_1(\mathbf{0})\mathbf{e}_1-W^1_{\lfloor(1-\nu_1(\mathbf{0}))nv_1\rfloor}/\lfloor nv_1\rfloor\\ \cdots\\ \nu_{d-1}(\mathbf{0})\mathbf{e}_{d-1}-W^{d-1}_{\lfloor(1-\nu_{d-1}(\mathbf{0}))nv_{d-1}\rfloor}/\lfloor nv_{d-1}\rfloor\\ \mathbf{r}\end{pmatrix}\mathbb{1}_{A_n}\right]$$

$$=\frac{1}{v_d}\lim_{n\to+\infty}n^{d/2}\mathbb{P}(A_n)\sum_{\sigma\in S_d}\mathrm{sgn}(\sigma)r_{\sigma(d)}\prod_{i=1}^{d-1}\left(\nu_i(\mathbf{0})\delta_{i\sigma(i)}-(1-\nu_i(\mathbf{0}))m_{i\sigma(i)}\right)$$

$$=0\times\det\begin{pmatrix}\nu_1(\mathbf{0})\mathbf{e}_1-(1-\nu_1(\mathbf{0}))\mathbf{m}^1\\ \cdots\\ \nu_{d-1}(\mathbf{0})\mathbf{e}_{d-1}-(1-\nu_{d-1}(\mathbf{0}))\mathbf{m}^{d-1}\\ \mathbf{r}\end{pmatrix}$$

$$=0$$

□

4.5 Results in some more general scenarios

4.5.1 Law of the total number of vertices with a given degree

Our results can be directly extended to a more general scenario where the leaves of a multitype branching forest are replaced by the vertices with a given inner-degree. In other words, we can easily obtain similar results to Theorem 4.4.1 and Corollary 4.4.2 by considering the number of individuals who have a given offspring. So we state these results without proof in the following.

Theorem 4.5.1 : Let $\tau_{i,\mathbf{l}}$ be the total number of type i-individuals whose offspring is $\mathbf{l}$, $i\in[d]$, $\mathbf{l}\in\mathbb{Z}_+^d$. For $\mathbf{l}\in\mathbb{Z}_+^d$ and all non-negative integers $q_i, r_i, n_{i,\mathbf{l}}, k_{ij},\ i,j\in[d]$ such

that $\sum_{i=1}^{d} r_i \geq 1, n_{i,\mathbf{l}} \leq q_i$ and $k_{ii} = -r_i - \sum_{j \neq i} k_{ji}$,

$$\begin{aligned}&\mathbb{P}_{\mathbf{r}}\left(N_1 = q_1, \ldots, N_d = q_d, \tau_{i,\mathbf{l}} = n_{i,\mathbf{l}}, M_{ij} = k_{ij}, i, j \in [d], i \neq j\right)\\ &= \frac{\det(K)}{\bar{q_1} \ldots \bar{q_d}} \prod_{i=1}^{d} \mathbb{P}\left(S^i_{q_i} = n_{i,\mathbf{l}}\right) \mathbb{P}\left(W^i_{q_i - n_{i,\mathbf{l}}} = \mathbf{k}_i + n_{i,\mathbf{l}}(\mathbf{e}_i - \mathbf{l})\right),\end{aligned}$$

where $\mathbf{r} = (r_1, \ldots, r_d)$, $\mathbf{k}_i = (k_{i1}, \ldots, k_{id}), \bar{q_i} = q_i \vee 1$, $\mathbf{e}_i$ is the i-th unit vector of $\mathbb{R}^d$, K is the matrix $(-k_{ij})$ to which we removed the line i and the column i for all i such that $q_i = 0$, S^i is a random walk whose step distribution is a Bernoulli law of parameter $\nu_i(\mathbf{l})$ and W^i is a random walk with jumps distributed according to $\eta_i(\mathbf{z}) = \frac{\nu_i(\mathbf{z}+\mathbf{e}_i)}{1-\nu_i(\mathbf{l})}$ for every $\mathbf{z} \in \mathbb{N}^d$.

Corollary 4.5.2 : Suppose that the offspring distribution ν is critical, admits moments of order $d+1$ and its covariance matrices are all positive-definite. Then for all $\mathbf{r}, \mathbf{l} \in \mathbb{Z}_+^d$ such that $\nu_i(\mathbf{l}) > 0$ for some $i \in [d]$, we have

$$\mathbb{P}_{\mathbf{r}}(\tau_{i,\mathbf{l}} = \nu_i(\mathbf{l})\lfloor nv_i \rfloor, N_i = \lfloor nv_i \rfloor, i \in [d]) = o\left(n^{-(d+1)}\right),$$

where $\tau_{i,\mathbf{l}}$ is defined in the above theorem.

4.5.2 Law of the total number of vertices whose degree is in a given set

In an even more general situation, we can also obtain similar results on the total number of individuals whose offspring is in a given non-empty subset of $\mathbb{Z}_+^d$. We state here these results which can be proven easily by similar adapted techniques.

Theorem 4.5.3 : Let $\mathcal{A}$ be a non-empty subset of $\mathbb{Z}_+^d$ and $\tau_{i,\mathcal{A}}$ be the total number of type i-individuals whose offspring is in the set $\mathcal{A}$, for $i \in [d]$. Then for all non-negative integers $q_i, r_i, n_{i,\mathcal{A}}, k_{ij},\ i, j \in [d]$ such that $\sum_{i=1}^{d} r_i \geq 1, n_{i,\mathcal{A}} \leq q_i$ and $k_{ii} = -r_i -$

$\sum_{j\neq i} k_{ji}$,

$$\begin{aligned}&\mathbb{P}_{\mathbf{r}}\left(N_1 = q_1, \dots, N_d = q_d, \tau_{i,\mathcal{A}} = n_{i,\mathcal{A}}, M_{ij} = k_{ij}, i, j \in [d], i \neq j\right)\\ &= \frac{\det(K)}{\bar{q}_1 \dots \bar{q}_d} \prod_{i=1}^{d} \mathbb{P}\left(S^i_{q_i} = n_{i,\mathcal{A}}\right) \mathbb{P}\left(W^i_{q_i - n_{i,\mathcal{A}}} = \mathbf{k}_i - U^i_{n_{i,\mathcal{A}}}\right),\end{aligned}$$

where $\mathbf{r} = (r_1, \dots, r_d)$, $\mathbf{k}_i = (k_{i1}, \dots, k_{id}), \bar{q}_i = q_i \vee 1$, $\mathbf{e}_i$ is the i-th unit vector of $\mathbb{R}^d$, K is the matrix $(-k_{ij})$ to which we removed the line i and the column i for all i such that $q_i = 0$. S^i is a random walk whose step distribution is a Bernoulli law of parameter $\nu_i(\mathcal{A})$, W^i is a random walk with jumps distributed according to a distribution η_i and U^i is a random walk with jumps distributed according to a distribution ζ_i, where η_i and ζ_i are defined as follows for $\mathbf{z} \in \mathbb{N}^d$:

$$\eta_i(\mathbf{z}) = \begin{cases} \frac{\nu_i(\mathbf{z}+\mathbf{e}_i)}{1-\nu_i(\mathcal{A})} & \text{if } \mathbf{z} + \mathbf{e}_i \notin \mathcal{A} \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\zeta_i(\mathbf{z}) = \begin{cases} \frac{\nu_i(\mathbf{z}+\mathbf{e}_i)}{\nu_i(\mathcal{A})} & \text{if } \mathbf{z} + \mathbf{e}_i \in \mathcal{A} \\ 0 & \text{otherwise.} \end{cases}$$

Corollary 4.5.4 : Suppose that the offspring distribution ν is critical, admits moments of order $d+1$ and its covariance matrices are all positive-definite. We assume also that $\nu_i(\mathcal{A}) > 0$ for some $i \in [d]$. Then for all $\mathbf{r} \in \mathbb{Z}^d_+$ and $\mathcal{A} \subset \mathbb{Z}^d_+$ such that $\nu_i(\mathcal{A}) > 0$ for some $i \in [d]$, we have

$$\mathbb{P}_{\mathbf{r}}(\tau_{i,\mathcal{A}} = \nu_i(\mathcal{A})\lfloor nv_i \rfloor, N_i = \lfloor nv_i \rfloor, i \in [d]) = o\left(n^{-(d+1)}\right),$$

where $\tau_{i,\mathcal{A}}$ is defined in the above theorem.

Bibliographie

[ADG15] R. Abraham, J.-F. Delmas et H. Guo : Critical multi-type galton-watson trees conditioned to be large. *ArXiv e-prints*, novembre 2015.

[Ale13] H. K. Alexander : Conditional distributions and waiting times in multitype branching processes. *Adv. in Appl. Probab.*, 45(3):692–718, 2013.

[AN72] Krishna B. Athreya et Peter E. Ney : *Branching processes*. Springer-Verlag, New York-Heidelberg, 1972. Die Grundlehren der mathematischen Wissenschaften, Band 196.

[Ath68] Krishna Balasundaram Athreya : Some results on multitype continuous time Markov branching processes. *Ann. Math. Statist.*, 39:347–357, 1968.

[BCLL03] Michel Bousquet, Cedric Chauve, Gilbert Labelle et Pierre Leroux : Two bijective proofs for the arborescent form of the Good-Lagrange formula and some applications to colored rooted trees and cacti. *Theoret. Comput. Sci.*, 307(2):277–302, 2003. Random generation of combinatorial objects and bijective combinatorics.

[Ber09] Jean Bertoin : The structure of the allelic partition of the total population for Galton-Watson processes with neutral mutations. *Ann. Probab.*, 37(4): 1502–1523, 2009.

[BM14] Olivier Bernardi et Alejandro H. Morales : Counting trees using symmetries. *J. Combin. Theory Ser. A*, 123:104–122, 2014.

[BR98] Edward A. Bender et L. Bruce Richmond : A multivariate Lagrange inversion formula for asymptotic calculations. *Electron. J. Combin.*, 5:Research Paper 33, 4 pp. (electronic), 1998.

[Cay97] Arthur Cayley : A theorem on trees. *The Collected Mathematical Papers.*, Vol XIII:26–28., 1897. Cambridge Library Collection - Mathematics. Cambridge : Cambridge University Press.

[Cha15] Loïc Chaumont : Breadth first search coding of multitype forests with application to Lamperti representation. *In In memoriam Marc Yor—Séminaire de Probabilités XLVII*, volume 2137 de *Lecture Notes in Math.*, pages 561–584. Springer, Cham, 2015.

[CL15] Loïc Chaumont et Rongli Liu : Coding multitype forests : Application to the law of the total population of branching forests. *Trans. Amer. Math. Soc.*, 368(4):2723–2747, sep 2015.

[CLN16] Loïc Chaumont, Rongli Liu et Thi Ngoc Anh Nguyen : Cyclically exchangeable sequences and enumeration of multitype forests. 2016. En préparation.

[CN15] Loïc Chaumont et Thi Ngoc Anh Nguyen : On mutations in the branching model for multitype populations. 2015. arXiv:1510.00845.

[CPGUB16] M.E. Caballero, J.L. Pérez Garmendia et G. Uribe Bravo : Affine processes on $\mathbb{R}_+^m \times \mathbb{R}^n$ and multiparameter time changes. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, 2016. A apparaître.

[DJ69] S. W. Dharmadhikari et Kumar Jogdeo : Bounds on moments of certain random variables. *Ann. Math. Statist.*, 40:1506–1509, 1969.

[DM10] Richard Durrett et Stephen Moseley : Evolution of resistance and progression to disease during clonal expansion of cancer. *Theoretical Population Biology*, 77(1):42 – 48, 2010.

[Dur13] Rick Durrett : Population genetics of neutral mutations in exponentially growing cancer cell populations. *Ann. Appl. Probab.*, 23(1):230–250, 2013.

[Dwa69] Meyer Dwass : The total progeny in a branching process and a related random walk. *J. Appl. Probability*, 6:682–686, 1969.

[Gal73] Francis Galton : Problem 4001 : On the extinction of surnames. *Educational Times*, page 17, 1873.

[Ges87] Ira M. Gessel : A combinatorial proof of the multivariable Lagrange inversion formula. *J. Combin. Theory Ser. A*, 45(2):178–195, 1987.

[GK97] I. P. Goulden et D. M. Kulkarni : Multivariable Lagrange inversion, Gessel-Viennot cancellation, and the matrix tree theorem. *J. Combin. Theory Ser. A*, 80(2):295–308, 1997.

[Goo60] I. J. Good : Generalizations to several variables of Lagrange's expansion, with applications to stochastic processes. *Proc. Cambridge Philos. Soc.*, 56: 367–380, 1960.

[Gut09] Allan Gut : *Stopped random walks.* Springer Series in Operations Research and Financial Engineering. Springer, New York, second édition, 2009. Limit theorems and applications.

[Har52] T. E. Harris : First passage and recurrence distributions. *Trans. Amer. Math. Soc.*, 73:471–486, 1952.

[Har02] Theodore E. Harris : *The theory of branching processes.* Dover Phoenix Editions. Dover Publications, Inc., Mineola, NY, 2002. Corrected reprint of the 1963 original [Springer, Berlin ; MR0163361 (29 #664)].

[HIM07] H. Haeno, Y. Iwasa et F. Michor : The evolution of two mutations during clonal expansion. *Genetics*, 177(4):2209–2221, dec 2007.

[IL71] I. A. Ibragimov et Yu. V. Linnik : *Independent and stationary sequences of random variables*. Wolters-Noordhoff Publishing, Groningen, 1971. With a supplementary chapter by I. A. Ibragimov and V. V. Petrov, Translation from the Russian edited by J. F. C. Kingman.

[IMKN05] Yoh Iwasa, Franziska Michor, Natalia L. Komarova et Martin A. Nowak : Population genetics of tumor suppressor genes. *J. Theoret. Biol.*, 233(1):15–23, 2005.

[KM96] Harry Kesten et R. A. Maller : Two renewal theorems for general random walks tending to infinity. *Probab. Theory Related Fields*, 106(1):1–38, 1996.

[Kor12] Igor Kortchemski : Invariance principles for Galton-Watson trees conditioned on the number of leaves. *Stochastic Process. Appl.*, 122(9):3126–3172, 2012.

[LG05] Jean-François Le Gall : Random trees and applications. *Probab. Surv.*, 2: 245–311, 2005.

[Mie08] Grégory Miermont : Invariance principles for spatial multitype Galton-Watson trees. *Ann. Inst. Henri Poincaré Probab. Stat.*, 44(6):1128–1161, 2008.

[Min05] Nariyuki Minami : On the number of vertices with a given degree in a Galton-Watson tree. *Adv. in Appl. Probab.*, 37(1):229–264, 2005.

[Mod71] Charles J. Mode : *Multitype branching processes. Theory and applications*. Modern Analytic and Computational Methods in Science and Mathematics, No. 34. American Elsevier Publishing Co., Inc., New York, 1971.

[Moo94] J. W. Moon : Some determinant expansions and the matrix-tree theorem. *Discrete Math.*, 124(1-3):163–171, 1994. Graphs and combinatorics (Qawra, 1990).

[Nev86] J. Neveu : Arbres et processus de Galton-Watson. *Ann. Inst. H. Poincaré Probab. Statist.*, 22(2):199–207, 1986.

[Ngu16] Thi Ngoc Anh Nguyen : A note on vertices with a given degree in multitype branching forests. 2016. En préparation.

[Ott49] Richard Otter : The multiplicative process. *Ann. Math. Statistics*, 20:206–224, 1949.

[Pit98] Jim Pitman : Enumerations of trees and forests related to branching processes and random walks. *In Microsurveys in discrete probability (Princeton, NJ, 1997)*, volume 41 de *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pages 163–180. Amer. Math. Soc., Providence, RI, 1998.

[Pit06] J. Pitman : *Combinatorial stochastic processes*, volume 1875 de *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard.

[Pé16] S. Pénisson : Beyond the q-process : Various ways of conditioning multitype galton-watson processes. *ALEA, Lat. Am. J. Probab. Math. Stat.*, 13 n°1:223–237, 2016. .

[Sen06] E. Seneta : *Non-negative matrices and Markov chains*. Springer Series in Statistics. Springer, New York, 2006. Revised reprint of the second (1981) edition [Springer-Verlag, New York ; MR0719544].

[Ser06] Maria Conceição Serra : On the waiting time to escape. *J. Appl. Probab.*, 43(1):296–302, 2006.

[SH07] Maria Conceição Serra et Patsy Haccou : Dynamics of escape mutants. *Theoretical Population Biology*, 72(1):167 – 178, 2007.

[Sta12] Richard P. Stanley : *Enumerative combinatorics. Volume 1*, volume 49 de *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, second édition, 2012.

[Sto67] Charles Stone : On local and ratio limit theorems. *In Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. II :*

Contributions to Probability Theory, Part 2, pages 217–224. Univ. California Press, Berkeley, Calif., 1967.

[Taï92] Ziad Taïb : *Branching processes and neutral evolution*, volume 93 de *Lecture Notes in Biomathematics*. Springer-Verlag, Berlin, 1992.

[Tak61] Lajos Takács : The probability law of the busy period for two types of queuing processes. *Operations Res.*, 9:402–407, 1961.

[Tak62] Lajos Takács : A generalization of the ballot problem and its application in the theory of queues. *J. Amer. Statist. Assoc.*, 57:327–337, 1962.

[Tut48] W. T. Tutte : The dissection of equilateral triangles into equilateral triangles. *Proc. Cambridge Philos. Soc.*, 44:463–482, 1948.

UNIVERSITE BRETAGNE LOIRE

université angers

Thèse de Doctorat

Thi Ngoc Anh NGUYEN

Sur quelques fontionnelles des forêts de branchement multitypes

On some functionals of multitype branching forests

Résumé
Cette thèse est principalement consacrée à l'étude de quelques caractéristiques d'une population à plusieurs types d'individus qui évolue selon un modèle de branchement multitype au cours du temps. Autrement dit, chaque individu vit un certain temps et donne naissance, à la fin de sa vie, à un nombre aléatoire d'individus, suivant une loi de probabilité qui ne dépend que de son type, indépendamment des autres individus. Plus précisément, nous nous intéressons aux aspects statistiques des mutations et des individus ayant une progéniture donnée dans la population en question.
Les problèmes d'énumération de forêts multitypes constituent également une motivation de ce travail de thèse.

Mots clés
forêt de branchement multitypes, processus de branchement multitypes, nombre de mutations, temps d'émergence, nombre de sommets ayant un degré donné, énumération des forêts multitypes.

Abstract
This thesis is devoted to the study of some characteristics of a population consisting of individuals of several types which evolve according to a multitype branching model. In other words, each individual lives a certain time and gives birth to a random number of individuals at the end of its life, following a probability law which depends only on the individual's type, independently of the others individuals. More precisely, we are interested in in the statistical aspects of mutations and the individuals having a given offspring in the population of interest.
The problems of enumeration of multitype forests also form a motivation of this thesis's work.

Key Words
multitype branching forests, multitype branching processes, number of mutations, emergence time, number of vertices with a given degree, enumeration of multitype forests