

Master 2 Recherche en Mathématiques Appliquées

Universités de Rennes I et de Toulouse III

---

*Modèles markoviens en biologie*

---

Djalil CHAFAÏ & Florent MALRIEU

Version préliminaire – Mars 2007

<http://www.lsp.ups-tlse.fr/Chafai/>

<http://perso.math.univ-rennes1.fr/florent.malrieu/>



# Table des matières

<b>1 Modèles de transmission de gènes</b>	<b>7</b>
1.1 Loi de Hardy-Weinberg . . . . .	7
1.1.1 La loi . . . . .	7
1.1.2 Un peu de statistique . . . . .	9
1.2 Modèles de Wright-Fisher . . . . .	11
1.2.1 Le modèle simple . . . . .	11
1.2.2 Prise en compte de la mutation . . . . .	14
1.2.3 Prise en compte de la sélection . . . . .	15
1.3 Modèle de Cannings . . . . .	17
1.4 Modèle de Moran . . . . .	18
1.5 Exercices et travaux pratiques . . . . .	20
1.6 Pour aller plus loin . . . . .	20
<b>2 Processus de diffusion</b>	<b>21</b>
2.1 Équations différentielles stochastiques . . . . .	21
2.2 Séjour dans un intervalle . . . . .	23
2.2.1 Deux exemples simples . . . . .	23
2.2.2 Cas général . . . . .	25
2.2.3 Fonction de Green . . . . .	26
2.3 Diffusions à coefficients singuliers . . . . .	30
2.3.1 Définitions . . . . .	30
<b>3 Applications à l'étude de chaînes de Markov</b>	<b>35</b>
3.1 Convergence d'une chaîne de Markov vers une diffusion . . . . .	35
3.1.1 Un peu d'intuition . . . . .	35
3.1.2 Le théorème . . . . .	36
3.1.3 Marche aléatoire simple et mouvement brownien . . . . .	37
3.1.4 Marche aléatoire et mouvement brownien avec dérive . . . . .	37
3.1.5 Modèle d'Ehrenfest et processus d'Ornstein-Uhlenbeck . . . . .	38
3.1.6 Un modèle markovien d'endémie . . . . .	38
3.2 D'autres exemples de convergence de processus . . . . .	40
3.2.1 Le megathéorème de convergence . . . . .	40
3.2.2 Espace d'états fini : du temps discret au temps continu . . . . .	40
3.2.3 File d'attente en temps continu et processus d'Ornstein-Uhlenbeck	41

3.2.4	Modèles de croissance de populations . . . . .	42
3.3	Application aux processus de Wright-Fisher . . . . .	43
3.3.1	Le cas général . . . . .	43
3.3.2	Ni mutation, ni sélection . . . . .	44
3.3.3	Mutation sans sélection . . . . .	44
<b>4</b>	<b>Généalogie</b>	<b>47</b>
4.1	Construction de l'arbre généalogique . . . . .	47
4.2	Le modèle à temps discret . . . . .	48
4.3	Le modèle à temps continu . . . . .	50
4.4	Le passage à la limite . . . . .	50
4.5	Propriétés du processus ancestral . . . . .	52
4.5.1	Expression de la matrice de transition . . . . .	52
4.5.2	Temps d'apparition de l'ACPR . . . . .	52
4.5.3	Longueur de l'arbre . . . . .	53
4.6	Price en compte des mutations . . . . .	54
4.6.1	Loi du nombre d'allèles dans un échantillon . . . . .	55
4.6.2	Loi du nombre de mutations . . . . .	58
4.6.3	Estimation du taux de mutation . . . . .	59
4.6.4	La formule de Ewens . . . . .	60
4.7	Le Coalescent . . . . .	62
<b>5</b>	<b>Processus ponctuels de Poisson</b>	<b>65</b>
5.1	Lois eulériennes et statistique d'ordre uniforme . . . . .	65
5.2	Loi des petits nombres . . . . .	67
5.3	Loi de Poisson et loi multinomiale . . . . .	68
5.4	Processus de Poisson et comptage de tops . . . . .	69
5.5	Processus ponctuels . . . . .	75
5.5.1	Loi et intensité d'un processus ponctuel . . . . .	76
5.5.2	Processus ponctuels simples . . . . .	78
5.6	Processus ponctuels de Poisson . . . . .	78
5.6.1	Stabilité par superposition et restriction . . . . .	81
5.6.2	Construction . . . . .	82
5.6.3	Transformations, marquage, et amincissement . . . . .	84
5.7	Processus ponctuels de Poisson sur $\mathbb{R}_+$ . . . . .	89
5.8	Quelques exemples classiques . . . . .	90
<b>6</b>	<b>Modèles compartimentaux</b>	<b>93</b>
6.1	Introduction . . . . .	93
6.2	Modèles déterministes linéaires . . . . .	93
6.3	Modèles déterministes linéaires, taux constants . . . . .	95
6.3.1	Exemple de modèle à un seul compartiment . . . . .	96
6.3.2	Exemple de modèle à deux compartiments . . . . .	96
6.3.3	Exemple de modèle à trois compartiments . . . . .	97

## TABLE DES MATIÈRES

5

6.4	Modèles stochastiques linéaires . . . . .	97
6.5	Lien entre déterministe et stochastique . . . . .	99
6.5.1	Files d'attente M/M/ $\infty$ en interaction . . . . .	99
6.5.2	Superposition de marches aléatoires . . . . .	100
<b>7</b>	<b>Modèles de maturation-survie</b>	<b>103</b>
7.1	Tapis roulant et maturation/survie . . . . .	103
7.2	Comptage et maturation/survie . . . . .	105
7.3	Tapis roulant et comptage . . . . .	107
	<b>Bibliographie</b>	<b>110</b>



# Chapitre 1

## Modèles de transmission de gènes

### 1.1 Loi de Hardy-Weinberg

En 1908, un mathématicien anglais, G.H. Hardy, et un médecin allemand W. Weinberg ont formulé une loi, connue sous le nom de loi de Hardy-Weinberg, qui a permis de démontrer l'un des premiers arguments anti-Darwin à savoir que dans le modèle darwinien l'hétérogénéité de la population a tendance à disparaître. Nous allons voir qu'il n'en est rien : en l'absence d'influences extérieures, la théorie de Mendel assure (sous certaines hypothèses fortement réductrices) la préservation de la variabilité des génotypes.

#### 1.1.1 La loi

**Hypothèse 1.1.1.** On s'intéresse aux fréquences alléliques pour un gène pouvant s'exprimer sous la forme de deux allèles  $A$  et  $B$  dans une population hermaphrodite<sup>1</sup> diploïde<sup>2</sup> idéale. Nous dirons qu'une population est idéale lorsque

1. Les générations sont séparées.
2. La population est de taille infinie.
3. Les individus s'y unissent aléatoirement, impliquant l'union aléatoire des gamètes. Il n'y a donc pas de choix du conjoint en fonction de son génotype. On dit alors que la population est panmictique.
4. Il n'y a pas de migration. Aucune copie allélique n'est apportée de l'extérieur.
5. Il n'y a pas de mutation.
6. Il n'y a pas de sélection.

**Proposition 1.1.2 (Loi de Hardy-Weinberg).** *Dans une population diploïde idéale, les fréquences alléliques d'un gène s'exprimant sous la forme de deux allèles  $A$  et  $B$  sont constantes au fil des générations. Plus précisément, si  $p \in [0, 1]$  est la proportion*

---

<sup>1</sup>Hermaphrodite : chaque individu peut être mâle ou femelle.

<sup>2</sup>Diploïde : qui possède un double assortiment de chromosomes semblables.

d'allèles  $A$  dans la population initiale, (la proportion de l'allèle  $B$  étant  $1 - p$ ) alors c'est encore encore le cas à chaque génération suivante.

De plus, si  $r$ ,  $s$ ,  $t$  sont les proportions respectives des génotypes  $AA$ ,  $AB$  et  $BB$  dans la population initiale (avec  $r + s + t = 1$ ) alors les fréquences respectives pour toutes les générations suivantes sont égales à  $(r + s/2)^2$ ,  $2(r + s/2)(t + s/2)$  et  $(t + s/2)^2$ .

*Démonstration.* Un individu  $AA$  est issu

- à coup sûr d'un couple  $AA$  et  $AA$  qui a une probabilité  $r^2$  de se former,
- avec probabilité  $1/2$  d'un couple  $AA$  et  $AB$  qui a une probabilité  $2rs$  de se former,
- avec probabilité  $1/4$  d'un couple  $AB$  et  $AB$  qui a une probabilité  $s^2$  de se former.

La fréquence du génotype  $AA$  à la génération 1 est donc donnée par

$$r^2 + rs + \frac{s^2}{4} = \left(r + \frac{s}{2}\right)^2.$$

Par symétrie, la fréquence du génotype  $BB$  est  $(t + s/2)^2$ . La dernière est obtenue par différence. En réappliquant ces formules on voit que les proportions sont constantes dès la première génération. La fréquence de l'allèle  $A$  dans la population initiale est  $2r + s$ . À la génération suivante, elle vaut

$$2\left(r + \frac{s}{2}\right)^2 + 2\left(r + \frac{s}{2}\right)\left(t + \frac{s}{2}\right) = 2r + s.$$

Elle est donc bien constante au cours du temps.  $\square$

**Remarque 1.1.3.** Cette loi est généralisable à un locus<sup>3</sup> avec plusieurs allèles  $A_1$ ,  $A_2, \dots, A_k$ .

### Conséquences

1. Les relations de dominance entre allèles n'ont aucun effet sur l'évolution des fréquences alléliques.
2. La ségrégation mendéienne aléatoire des chromosomes préserve la variabilité génétique des populations.
3. L'évolution étant définie par un changement des fréquences alléliques, une population diploïde idéale n'évolue pas.
4. Seules les violations des propriétés de la population idéale permettent le processus évolutif.

### Validité de la loi de Hardy-Weinberg.

Bien que les propriétés d'une population idéale apparaissent un peu surréalistes, la plupart des populations présentent des fréquences génotypiques en équilibre de Hardy-Weinberg pour une grande majorité des locus. Ceci est dû au fait que cet

---

<sup>3</sup>Locus : localisation précise d'un gène particulier sur un chromosome.

équilibre résulte avant tout de la ségrégation aléatoire des chromosomes qui a lieu à chaque génération.

Dans les populations naturelles, les fréquences alléliques varient constamment d'une génération à l'autre sous l'influence de divers facteurs (sélection, dérive génétique, etc...). Mais l'équilibre de Hardy-Weinberg est rétabli au début de chaque génération.

L'équilibre est avant tout perturbé si les gamètes ne sont pas produites aléatoirement (meiotic drive), ou si il y a choix du conjoint (consanguinité). Notez que la sélection naturelle n'affecte pas l'équilibre de Hardy-Weinberg parmi les nouveaux-nés. Son effet ne devient perceptible que par la suite, au cours du développement.

### 1.1.2 Un peu de statistique

Nous dirons que la population considérée est en équilibre de Hardy-Weinberg de paramètre  $p \in ]0, 1[$  si les trois génotypes  $AA$ ,  $AB$  et  $BB$  sont présents avec les fréquences

$$p_1 = p^2, \quad p_2 = 2p(1-p) \quad \text{et} \quad p_3 = (1-p)^2. \quad (1.1)$$

On effectue un tirage (avec remise) d'un échantillon de taille  $n$  dans cette population et on note  $(N_1, N_2, N_3)$  les effectifs observés des trois génotypes.

Lors d'une expérience menée sur une population de 100 individus, on a mesuré les données suivantes :

$$n = 100, \quad N_1 = 13, \quad N_2 = 49, \quad N_3 = 38.$$

#### Adéquation à la loi de Hardy-Weinberg

La première question que l'on peut se poser est la suivante : est-il raisonnable de dire que la population observée est en équilibre Hardy-Weinberg pour un paramètre  $p$  donné ? Pour répondre à cette question de manière statistique, on peut mettre en place un test du  $\chi^2$  pour tester l'hypothèse

$H_0$  : « la population est en équilibre de Hardy-Weinberg de paramètre  $p$  »,

contre l'hypothèse

$H_1$  : « la population n'est pas en équilibre de Hardy-Weinberg de paramètre  $p$  ».

#### Estimation du paramètre

On suppose à présent que la population considérée est en équilibre de Hardy-Weinberg de paramètre  $p$  inconnu et l'on souhaite à estimer « du mieux possible » ce paramètre  $p$ . Le cadre est le suivant : on dispose d'une réalisation  $(N_1, N_2, N_3)$  de loi multinomiale  $\mathbb{P}_p$  de paramètres  $n$  et  $\nu_p = (p^2, 2p(2-p), (1-p)^2)$  sur  $\{1, 2, 3\}$

(les nombres 1, 2 et 3 représentant les génotypes respectifs  $AA$ ,  $AB$  et  $BB$ ) avec  $p \in ]0, 1[$  inconnu<sup>4</sup>.

On peut penser à estimer  $p$  par l'estimateur du maximum de vraisemblance. Il s'avère que c'est un excellent choix car dans ce cas particulier, il s'agit aussi de l'estimateur sans biais de variance minimale. Écrivons la vraisemblance de l'échantillon, i.e. la densité de la v.a.  $(N_1, N_2, N_3)$  par rapport à la mesure de comptage sur  $\mathbb{N}^3$  :

$$L(p, X) = \frac{n!}{N_1! N_2! N_3!} p_1^{N_1} p_2^{N_2} p_3^{N_3} = \frac{n! 2^{N_2}}{N_1! N_2! N_3!} p^{2N_1+N_2} (1-p)^{N_2+2N_3}.$$

Un estimateur du maximum de vraisemblance est une statistique  $\hat{p}_{MV}$  telle que  $L(\hat{p}_{MV}, N_1, N_2, N_3)$  est maximum (on pourra choisir de maximiser la fonction de log-vraisemblance  $p \mapsto \log L(p, N_1, N_2, N_3)$ ).

**Proposition 1.1.4.** *L'estimateur du maximum de vraisemblance (il est ici unique) s'écrit*

$$\hat{p}_{MV} = \frac{2N_1 + N_2}{2n}.$$

On peut se demander en quoi ce nouvel estimateur est bon, voire optimal. Pour cela, on a recours à la notion d'exhaustivité.

**Définition 1.1.5.** On dira qu'une statistique  $T$  à valeurs dans  $\mathbb{R}$  (i.e. une variable aléatoire de la forme  $T = \phi(N_1, N_2, N_3)$ ) est exhaustive pour  $p$  si la vraisemblance se factorise de la façon suivante :

$$L(p, N_1, N_2, N_3) = h(N_1, N_2, N_3)g(p, T), \quad (1.2)$$

où  $g$  (resp.  $h$ ) est une application de  $]0, 1[ \times \mathbb{R}$  (resp.  $\mathbb{N}^3$ ) dans  $\mathbb{R}_+$ .

On dit de plus que la statistique  $T$  est exhaustive complète pour  $p$  si pour toute fonction mesurable bornée  $f$ ,

$$\forall p \in ]0, 1[, \quad \mathbb{E}_p(f(T)) = 0 \implies \forall p \in ]0, 1[, \quad f = 0 \text{ } \mathbb{P}_p \text{ p.s.}$$

**Proposition 1.1.6.** *La statistique  $T = 2N_1 + N_2$  est exhaustive complète pour  $p$ .*

**Remarque 1.1.7.** La factorisation (1.2) assure que la loi de  $(N_1, N_2, N_3)$  sachant  $T$  ne dépend pas de  $p$ . L'exhaustivité de  $T$  pour  $p$  assure ainsi que toute « l'information pour l'estimation de  $p$  » contenue dans les observations  $(N_1, N_2, N_3)$  est contenue dans la variable aléatoire  $T$ , i.e.. Ainsi, on n'estimera pas mieux  $p$  en connaissant  $(N_1, N_2, N_3)$  qu'en connaissant seulement  $T$ .

On peut alors montrer le résultat suivant.

**Théorème 1.1.8 (Lehmann-Scheffé).** *Soit  $\psi$  un estimateur sans biais de  $p$ . Si  $\phi$  est une statistique exhaustive complète alors la statistique  $\mathbb{E}(\psi|\phi)$  est  $(\mathbb{P}_p)_{p \in ]0, 1[}$ -presque sûrement l'unique estimateur sans biais de variance minimum.*

**Corollaire 1.1.9.** *L'estimateur du maximum de vraisemblance est l'unique estimateur sans biais de variance minimum.*

---

<sup>4</sup>La loi de  $(N_1, N_2, N_3)$  est la loi du nombre de boules dans trois urnes 1, 2 et 3 lorsque l'on a réparti dans ces trois urnes  $n$  boules indépendamment selon la loi  $\nu_p$ .

### Un autre test

Pour finir, on souhaite utiliser l'estimateur du maximum de vraisemblance pour décider si oui ou non, il est raisonnable de considérer la population observée comme étant en équilibre de Hardy-Weinberg. Pour cela, on couple l'estimation de  $p$  et le test du  $\chi^2$ . On peut montrer le théorème suivant :

**Théorème 1.1.10.** *Dans le modèle étudié ( $p_1$ ,  $p_2$  et  $p_3$  s'expriment en fonction de  $p$  d'après (1.1)), si  $D_n$  désigne la statistique du  $\chi^2$  :*

$$D_n(p) = \sum_{i=1}^3 \frac{(N_i - np_i(p))^2}{np_i(p)},$$

alors,

$$D_n(\hat{p}_{MV}) = \sum_{i=1}^3 \frac{(N_i - np_i(\hat{p}_{MV}))^2}{np_i(\hat{p}_{MV})} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \chi^2(1).$$

**Remarque 1.1.11.** Le nombre de degrés de liberté dans le théorème 1.1.10 est égal à 3-1-1, c'est-à-dire le nombre de degrés de liberté d'un test du  $\chi^2$  classique sur trois classes moins la « dimension » de  $p$ .

Ce résultat permet de tester l'hypothèse :

$$H_0 : \text{« la population est en équilibre de Hardy-Weinberg »},$$

contre l'hypothèse

$$H_1 : \text{« la population n'est pas en équilibre de Hardy-Weinberg »}.$$

## 1.2 Modèles de Wright-Fisher

Il s'agit du modèle le plus célèbre (et parmi les plus simples) de l'évolution de la fréquence d'un gène à deux allèles dans une population finie. On suppose que la population est de taille constante, que les générations ne se recouvrent pas et que les unions sont indépendantes du caractère étudié.

### 1.2.1 Le modèle simple

On néglige dans un premier temps les phénomènes de mutation et sélection. On note  $2N$  la taille de la population et  $n = 0, 1, 2, \dots$  les générations successives. Sur le locus étudié, on peut trouver deux allèles différents notés  $A$  et  $B$ . La variable aléatoire  $X_n$  compte le nombre d'allèles  $A$  à la génération  $n$ . La population à la génération  $n+1$  est déduite de celle de la génération  $n$  par un tirage binomial avec remise de  $2N$  gènes disposés dans une urne dont la proportion d'allèles  $A$  est  $X_n/(2N)$ . Plus précisément, sachant que  $X_n = i$ , la probabilité que  $X_{n+1} = j$  est donnée par

$$p_{ij} := \mathbb{P}(X_{n+1} = j | X_n = i) = C_{2N}^j \psi_i^j (1 - \psi_i)^{2N-j} \quad (1.3)$$

pour  $0 \leq i, j \leq 2N$  avec  $\psi_i = i/(2N)$ .

Le processus  $(X_n)_{n \geq 0}$  est une chaîne de Markov homogène d'espace d'états  $\mathcal{S} = \{0, 1, \dots, 2N\}$  et de matrice de transition  $P = (p_{ij})$ . Les états 0 et  $2N$  sont absorbants. Lorsque la population ne contient plus qu'un allèle, on dit qu'il est fixé dans la population. Les questions naturelles qui viennent à l'esprit sont par exemple : quel allèle sera fixé ? au bout de combien de temps ?

Les états 0 et  $2N$  sont absorbants. Tous les autres mènent à  $\{0, 2N\}$ , ils sont donc transients. L'espèce d'états étant fini, le temps d'atteinte de  $\{0, 2N\}$  est fini presque sûrement et même intégrable. La remarque suivante donne une justification de ce point dans le cas particulier de la chaîne de Wright-Fisher.

**Remarque 1.2.1.** Sachant que  $X_n = i \notin \{0, 2N\}$ , la probabilité que  $X_{n+1} \in \{0, 2N\}$  est égale à  $(1 - \psi_i)^{2N} + \psi_i^{2N}$  qui est minimale pour  $\psi_i = 1/2$  et vaut alors  $2^{-2N+1}$ . Ainsi, la chaîne atteindra les états absorbants avant qu'un lanceur de pièces ne fasse *pile* si la pièce est biaisée de telle sorte que *pile* sorte avec probabilité  $2^{-2N+1}$ . Cette borne est extrêmement pessimiste mais elle fournit facilement un contrôle sous géométrique explicite pour le temps d'atteinte étudié et confirme en particulier que le temps d'atteinte de  $\{0, 2N\}$  est fini presque sûrement et même intégrable.

En plus d'être une chaîne de Markov, la suite  $(X_n)_n$  a le bon goût d'être également une martingale.

**Proposition 1.2.2.** *La suite  $(X_n)_n$  est une martingale pour sa filtration naturelle  $(\mathcal{F}_n)_{n \geq 0}$ .*

*Démonstration.* Puisque les transitions sont des lois binomiales, on a

$$\mathbb{E}(X_n | \mathcal{F}_{n-1}) = \mathbb{E}(X_n | X_{n-1}) = 2N \frac{X_{n-1}}{2N} = X_{n-1}.$$

Ainsi,  $(X_n)_n$  est une martingale et, en particulier, l'espérance de  $X_n$  est constante au cours du temps, égale à  $\mathbb{E}(X_0)$ .  $\square$

Cette propriété de  $(X_n)_n$  permet de déterminer la probabilité de fixation (et la probabilité de disparition) de l'allèle  $A$ .

**Corollaire 1.2.3.** *La probabilité sachant que  $X_0 = i$  que  $X$  atteigne  $2N$  (avant 0) est égale à  $i/(2N)$ .*

*Démonstration.* On applique le théorème de convergence des martingales bornées à  $X$ . Notons  $T$  le temps d'atteinte de l'ensemble  $\{0, 2N\}$ . Alors, pour  $i = 0, 1, \dots, 2N$ ,

$$\begin{cases} 1 = \mathbb{P}_i(T < +\infty) = \mathbb{P}_i(X_T = 0) + \mathbb{P}_i(X_T = 2N), \\ i = \mathbb{E}_i(X_0) = \mathbb{E}_i(X_T) = 0 \times \mathbb{P}_i(X_T = 0) + 2N \times \mathbb{P}_i(X_T = 2N), \end{cases}$$

d'où l'on tire  $\mathbb{P}_i(X_T = 2N) = i/(2N)$ .  $\square$

La question suivante est de mesurer la vitesse de disparition de l'un des allèles. Voici quelques résultats dans cette direction.

**Lemme 1.2.4.** Pour tout  $n \geq 1$ ,

$$\mathbb{E}(X_n(2N - X_n)) = \left(1 - \frac{1}{2N}\right)^n \mathbb{E}(X_0(2N - X_0)).$$

*Démonstration.* On a

$$\mathbb{E}(X_n(2N - X_n)) = 2N\mathbb{E}(X_n) - \mathbb{E}(X_n^2) = 2N\mathbb{E}(X_{n-1}) - \mathbb{E}(\mathbb{E}(X_n^2|X_{n-1})).$$

On écrit alors

$$\mathbb{E}(X_n^2|X_{n-1}) = \text{Var}(X_n|X_{n-1}) + (\mathbb{E}(X_n|X_{n-1}))^2 = X_{n-1}(1 - X_{n-1}/(2N)) + X_{n-1}^2.$$

En regroupant les termes, on obtient bien

$$\mathbb{E}(X_n(2N - X_n)) = \left(1 - \frac{1}{2N}\right)\mathbb{E}(X_{n-1}(2N - X_{n-1})).$$

ce qui fournit le résultat.  $\square$

Posons  $\lambda = 1 - 1/(2N)$ . L'hétérozygotie est la probabilité que deux gènes choisis aléatoirement (sans remise) dans la population totale à la génération  $n$  soient représentés par des allèles différents. Elle est donnée par

$$H_n = \frac{2X_n(2N - X_n)}{2N(2N - 1)}.$$

D'après le calcul précédent, l'hétérozygotie moyenne  $h(n)$  vérifie

$$h_n := \mathbb{E}(H_n) = \lambda^n h_0.$$

**Remarque 1.2.5.** De même, la variance de  $X_n$  se calcule par un simple conditionnement :

$$\mathbb{V}(X_n) = \mathbb{E}(\mathbb{V}(X_n|X_{n-1})) + \mathbb{V}(\mathbb{E}(X_n|X_{n-1})).$$

On obtient alors, en posant  $\lambda = 1 - 1/(2N)$ ,

$$\mathbb{V}(X_n) = \mathbb{E}(X_0)(2N - \mathbb{E}(X_0))(1 - \lambda^n) + \lambda^n \mathbb{V}(X_0).$$

Une question naturelle est aussi de mieux comprendre la loi de  $T$  le temps de disparition d'un des deux allèles, par exemple en estimant son espérance. Pour le modèle de Wright-Fisher cette question est délicate : il n'existe pas de formule simple pour tout  $N$ . En effet, si l'on note  $m_i$  l'espérance du temps d'absorption de  $X$  sachant que  $X_0 = i$ , on a  $m_0 = m_{2N} = 0$  et, grâce à la propriété de Markov,

$$m_i = 1 + \sum_{j=0}^{2N} p_{ij} m_j.$$

Ce système à  $2N - 1$  inconnues n'est pas facile à résoudre. Il est toutefois possible de trouver un équivalent de  $m$  lorsque la taille de la population tend vers l'infini. Notons  $p = i/(2N)$  la proportion d'allèles  $A$  dans la population initiale et supposons que l'espérance du temps d'absorption de  $X$  soit proche d'une fonction de classe  $\mathcal{C}^2$  notée  $t(p)$ . On utilise à nouveau la propriété de Markov :

$$\begin{aligned} t(p) &= \sum_{\delta p} \mathbb{P}(p \rightarrow p + \delta p)(t(p + \delta p) + 1) \\ &= \sum_{\delta p} \mathbb{P}(p \rightarrow p + \delta p) \left( t(p) + t'(p)\delta p + \frac{t''(p)}{2}(\delta p)^2 + O((\delta p)^3) + 1 \right) \\ &= t(p) + t'(p)\mathbb{E}(\delta p) + \frac{t''(p)}{2}\mathbb{E}[(\delta p)^2] + O(\mathbb{E}((\delta p)^3)) + 1, \end{aligned}$$

où  $\delta p$  a la loi de  $(Y_{2N} - \mathbb{E}(Y_{2N}))/(\sqrt{2N})$  si  $Y_{2N}$  suit la loi  $\mathcal{B}(2N, p)$ . D'après le théorème limite central,

$$\frac{Y_{2N} - \mathbb{E}(Y_{2N})}{\sqrt{2N}} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} Y, \quad \text{avec } Y \sim \mathcal{N}(0, 1),$$

et, pour tout  $k \in \mathbb{N}$ ,

$$\mathbb{E} \left[ \left( \frac{Y_{2N} - \mathbb{E}(Y_{2N})}{\sqrt{2N}} \right)^k \right] \xrightarrow[n \rightarrow \infty]{} \mathbb{E}(Y^k).$$

En particulier,

$$\mathbb{E}(\delta p) = 0, \quad \mathbb{E}[(\delta p)^2] = \frac{1}{2N}p(1-p) \quad \text{et} \quad \mathbb{E}[(\delta p)^3] = O(N^{-3/2}).$$

Ce calcul heuristique suggère que, si  $N$  est grand,  $t$  est proche de la solution de l'équation différentielle ordinaire suivante :

$$y''(p) = -\frac{4N}{p(1-p)} \quad \text{avec} \quad y(0) = y(1) = 0.$$

On obtient alors

$$t(p) \sim y(p) = -4N(p \ln p + (1-p) \ln(1-p)).$$

Nous justifierons cette approximation au chapitre 2 en montrant (en partie) que la chaîne de Wright-Fisher, convenablement renormalisée en temps et en espace, ressemble à un processus de diffusion pour lequel nous serons capable de calculer l'espérance du temps d'absorption.

### 1.2.2 Prise en compte de la mutation

Supposons en plus des hypothèses du modèle basique précédent que l'allèle  $A$  mute en allèle  $B$  avec probabilité  $u$  et  $B$  en  $A$  avec probabilité  $v$ . Il convient alors

dans la définition de la chaîne (1.3) de remplacer  $\psi_i$  par

$$\psi_i = \frac{i(1-u) + (2N-i)v}{2N}.$$

Le nouveau processus est bien entendu encore une chaîne de Markov mais cette fois-ci, tous les états sont récurrents et apériodiques (et ce n'est plus une martingale). La chaîne admet une unique probabilité invariante, notons-la  $\nu = (\nu_0, \nu_1, \dots, \nu_{2N})$  (identifiée à un vecteur ligne). Elle n'est pas explicitable simplement. On peut toutefois déterminer ses deux premiers moments. Notons  $m$  sa moyenne. Alors

$$m = \nu\xi = \nu P\xi,$$

où  $\xi = (0, 1, \dots, 2N)^T$  (identifié à un vecteur colonne) est l'application identité. Puisque l'on connaît l'espérance de la loi binomiale,

$$(P\xi)_i = 2N\psi_i = i(1-u) + (2N-i)v.$$

Donc

$$\nu P\xi = \sum_{i=0}^{2N} \psi_i \{i(1-u) + (2N-i)v\} = m(1-u) + v(2N-m).$$

On en déduit l'expression de l'espérance de  $\nu$  :

$$m = \frac{2Nv}{u+v}.$$

On peut montrer de même que la variance  $\sigma^2$  de  $\nu$  est de la forme :

$$\sigma^2 = \frac{4N^2uv}{(u+v)^2(4Nu+4Nv+1)} + o(N).$$

### 1.2.3 Prise en compte de la sélection

Les différents allèles procurent à l'individu qui en est doté des capacités plus ou moins grandes dans tous les domaines de son développement et de sa reproduction : viabilité, potentiel attractif, fertilité etc. C'est la fameuse sélection naturelle. Tentons de quantifier ceci. Bien que l'adaptation d'un individu à son environnement soit déterminée par de nombreux facteurs, nous supposerons ici qu'il n'est déterminé que par le locus qui nous intéresse. Nous supposerons de plus que la sélection ne s'opère que sur le critère de viabilité. Supposons que les adaptabilités des trois génotypes  $AA$ ,  $AB$  et  $BB$  soient données par  $1+s$ ,  $1+sh$  et  $1$ . On suppose dans un premier temps qu'il n'y a pas de mutation. On remplace alors dans (1.3)  $\psi_i$  par par

$$\eta_i = \frac{(1+s)i^2 + (1+sh)i(2N-i)}{(1+s)^2i^2 + 2(1+sh)i(2N-i) + (2N-i)^2}.$$

Si de plus, on prend en compte le phénomène de mutation, on remplace alors  $\eta_i$  par

$$\eta_i^* = \eta_i(1 - u) + (1 - \eta_i)v.$$

Peu de résultats quantitatifs peuvent être établis pour ce modèle général. Nous pourrons néanmoins le faire pour son approximation diffusive. Faisons tout de même quelques observations.

### Absence de mutation

Supposons l'absence de mutation ( $u = v = 0$ ). Les états 0 et  $2N$  sont absorbants, les autres sont transients. Supposons de plus que  $s$  d'ordre  $N^{-1}$  et  $h$  d'ordre 1. Posons  $\alpha = 2Ns$  et notons  $(\pi_i)_i$  les probabilités de fixation de la chaîne en  $2N$  (avant 0). La propriété de Markov assure que

$$\pi_i = \sum_{j=0}^{2N} p_{ij}\pi_j, \quad \text{avec } \pi_0 = 0 \text{ et } \pi_{2N} = 1.$$

Appliquons cette formule à  $i = 2Nx$  :

$$\begin{aligned} \pi(x) &= \sum \mathbb{P}(x \rightarrow x + \delta x)\pi(x + \delta x) \\ &= \sum \mathbb{P}(x \rightarrow x + \delta x)(\pi(x) + (\delta x)\pi'(x) + \frac{1}{2}(\delta x)^2\pi''(x) + O((\delta x)^3)). \end{aligned}$$

Dans le modèle présent,

$$\begin{aligned} \mathbb{E}((\delta x)) &= \frac{\alpha x(1-x)(x+h(1-2x))}{2N} + O(N^{-2}), \\ \mathbb{E}((\delta x)^2) &= \frac{\alpha x(1-x)}{2N} + O(N^{-2}), \\ \mathbb{E}((\delta x)^3) &= O(N^{-3/2}). \end{aligned}$$

La fonction  $\pi$  semble donc proche de la solution de l'équation différentielle suivante :

$$2\alpha(x + h(1 - 2x))z'(x) + z''(x) = 0, \quad \text{avec } z(0) = 0 \text{ et } z(1) = 1,$$

c'est-à-dire

$$\pi(x) \sim z(x) = \frac{\int_0^x \psi(u) du}{\int_0^1 \psi(u) du}, \quad \text{avec } \psi(u) = \exp(-\alpha u(2h + u(1 - 2h))).$$

En particulier, si  $h = 1/2$ , c'est-à-dire si l'adaptabilité est proportionnelle au nombre d'allèles  $A$ , ceci revient à

$$\pi(x) \sim z(x) = \frac{1 - e^{-\alpha x}}{1 - e^{-x}}.$$

Supposons que  $N = 10^5$ ,  $s = 10^{-4}$  et  $x = 0,5$ . Alors  $\alpha = 20$  et  $\pi = 0,999955$ . En l'absence de sélection ( $s = 0$ ), on a bien sûr  $\pi(0,5) = 0,5$ . Même le faible avantage 0,0001 (inobservable en laboratoire ou par des mesures statistiques) est pourtant suffisamment grand pour avoir un effet déterminant sur la fixation des allèles. Bien que cet effet soit imperceptible sur une génération, il l'est jusqu'à l'instant de fixation car le temps de fixation est très grand.

## 1.3 Modèle de Cannings

On considère une population de gènes de taille fixe  $2N$  qui se reproduit aux temps entiers. Le processus aléatoire de reproduction est assez général : on impose seulement que la loi des descendants à la génération  $t + 1$  soit échangeable, c'est-à-dire invariante par permutation des coordonnées. Plus précisément, si  $y_i$  est le nombre de enfants de l'individu  $i$ , on impose que  $y_1 + \dots + y_{2N} = 2N$  et que, pour toute permutation  $\tau \in \mathcal{S}_{2N}$ ,

$$(y_1, \dots, y_{2N}) \stackrel{\mathcal{L}}{=} (y_{\tau(1)}, \dots, y_{\tau(2N)}).$$

En particulier tous les individus ont la même loi de reproduction. Celle-ci doit avoir une espérance égale à 1 et sa variance sera notée  $\sigma^2$ .

Le modèle de Wright-Fisher est un cas particulier du modèle de Cannings ; il correspond au cas où la loi de  $(y_1, \dots, y_{2N})$  est loi multinomiale symétrique.

Il est remarquable que, même pour un modèle aussi général, de nombreuses quantités puissent être explicitées. En particulier, il est possible de donner une représentation complète du spectre de la matrice de transition de cette chaîne de Markov.

**Théorème 1.3.1** (Cannings (1974)). *Les valeurs propres de la matrice de transition du modèle de Cannings sont données par*

$$\lambda_0 = 1, \quad \forall j = 1, \dots, 2N, \quad \lambda_j = \mathbb{E}(y_1 y_2 \cdots y_j).$$

*Démonstration.* Notons  $P$  la matrice de transition et définissons la matrice  $Z \in \mathcal{M}_{2N+1}(\mathbb{R})$  par

$$Z = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 1^2 & 1^3 & \cdots & 1^{2N} \\ 1 & 2 & 2^2 & 2^3 & \cdots & 2^{2N} \\ 1 & 3 & 3^2 & 3^3 & \cdots & 3^{2N} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2N & (2N)^2 & (2N)^3 & \cdots & (2N)^{2N} \end{pmatrix}$$

La matrice  $Z$  est inversible (cf Vandermonde) et

$$(PZ)_{ij} = \sum_{k=0}^{2N} P_{ik} Z_{kj} = \sum_{k=0}^{2N} P_{ik} k^j = \mathbb{E}(X(t+1)^j | X(t) = i).$$

Pour tous  $i$ , il existe  $(b_{ij})_j$  tels que

$$\mathbb{E}(X(t+1)^j | X(t) = i) = \mathbb{E}[(y_1 + \dots + y_i)^j] = i^{[j]} \mathbb{E}(y_1 y_2 \cdots y_j) + \sum_{k=0}^{j-1} b_{ik} i^k,$$

où  $i^{[j]} = i(i-1)(i-2)\cdots(i-j+1)$ . Il existe donc une matrice  $A$  de  $\mathcal{M}_{2N+1}(\mathbb{R})$  triangulaire supérieure telle que, pour tous  $i, j$ ,

$$\mathbb{E}(X(t+1)^j | X(t) = i) = \sum_{k=0}^j a_{kj} i^k = (ZA)_{ij}.$$

De plus on a  $a_{jj} = \mathbb{E}(y_1 y_2 \cdots y_j)$ . On a donc écrit  $PZ = ZA$ . Les matrices  $P$  et  $A$  admettent donc les mêmes spectres.  $\square$

**Remarque 1.3.2.** Dans le cas particulier du modèle de Wright-Fisher, on obtient le résultat explicite suivant :

$$\forall j = 1, \dots, 2N, \quad \lambda_j = (2N)(2N-1)\cdots(2N-j+1)/(2N)^j,$$

Puisque  $\mathbb{E}(y_1) = 1$ , 1 est valeur propre double. C'est logique puisque la chaîne possède deux classes de récurrence  $\{0\}$  et  $\{2N\}$ . La deuxième plus grande valeur propre est  $\lambda_2 = \mathbb{E}(y_1 y_2)$ . Puisque  $\sum y_i$  est constante égale à  $2N$ , sa variance est nulle et donc

$$2N\mathbb{V}(y_1) + 2N(2N-1)\text{covar}(y_1, y_2) = 0.$$

On a donc

$$\text{covar}(y_1, y_2) = -\frac{\sigma^2}{2N-1} \quad \text{et} \quad \lambda_2 = 1 - \frac{\sigma^2}{2N-1}.$$

De même, on peut remarquer que

$$\mathbb{V}(X(t+1) | X(t) = i) = \frac{i(2N-i)\sigma^2}{2N-1},$$

et que, si  $x(t) = X(t)/(2N)$ ,

$$\mathbb{V}(x(t+1) | x(t)) = \frac{x(t)(1-x(t))\sigma^2}{2N-1}.$$

## 1.4 Modèle de Moran

Le modèle de Moran ressemble à celui de Wright-Fisher. Il est un peu moins célèbre bien que plus simple à étudier. C'est aussi un cas particulier du modèle de Cannings. Dans le modèle de Moran, la transition se fait de la façon suivante : deux individus sont sélectionnés (avec remise). Le premier a deux fils, le second aucun. Les autres individus ont exactement un descendant.

**Définition 1.4.1.** Le processus de Moran (qui compte le nombre d'allèles  $A$  dans une population de  $N$  gènes) est la chaîne de Markov de matrice de transition

$$\forall i, j \in \{0, \dots, N\}, \quad p_{ij} = \begin{cases} i(N-i)/N^2 & \text{si } |i-j|=1, \\ i^2/N^2 + (N-i)^2/N^2 & \text{si } j=i, \\ 0 & \text{sinon.} \end{cases}$$

Les états 0 et  $N$  sont absorbants et tous les autres états mènent à eux. Notons  $A = \{0, N\}$  l'ensemble des éléments absorbants et

$$T = \inf \{n \geq 0 : X_t \in A\}$$

le temps d'atteinte de  $A$  par  $X$ . Le vecteur  $(m_i)_i$  des espérances de  $T$  pour la chaîne harmonisée issue de  $i$  vérifie les relations suivantes :

$$m_0 = m_N = 0, \quad \forall i = 1, \dots, N-1, \quad m_i = 1 + \sum_{j=0}^N p_{ij} m_j.$$

On obtient la formule de récurrence à trois termes suivante : pour  $i = 1, \dots, N-1$ ,

$$m_{i+1} - 2m_i + m_{i-1} = -\frac{N^2}{i(N-i)}.$$

On en déduit que

$$m_i = N \left( \sum_{j=1}^i \frac{N-i}{N-j} + \sum_{j=i+1}^{N-1} \frac{i}{j} \right).$$

Soit  $p \in ]0, 1[$ . Pour tout  $N \geq 1$ , posons  $i_N = [pN]$ . On a alors

$$\begin{aligned} \frac{k_{i_N}}{N^2} &= \frac{1}{N} \sum_{j=1}^{i_N} \frac{N-i_N}{N-j} + \frac{1}{N} \sum_{j=i_N+1}^{N-1} \frac{i_N}{j} \\ &= \frac{1-i_N/N}{N} \sum_{j=1}^{i_N} \frac{1}{1-j/N} + \frac{i_N/N}{N} \sum_{j=i_N+1}^{N-1} \frac{1}{j/N} \\ &\xrightarrow[N \rightarrow \infty]{} (1-p) \int_0^p \frac{1}{1-x} dx + p \int_p^1 \frac{1}{x} dx = -(1-p) \ln(1-p) - p \ln p. \end{aligned}$$

Ainsi, dans une population de taille  $N$  (avec  $N$  grand), le temps moyen d'absorption partant de  $i = pN$  avec  $p \in ]0, 1[$  est de l'ordre de

$$-N^2(p \ln p + (1-p) \ln(1-p)).$$

**Remarque 1.4.2.** Pour le modèle de Wright-Fisher, l'espérance du temps d'absorption est de l'ordre de  $N$  tandis qu'il est de l'ordre de  $N^2$  pour le modèle de Moran. Cela vient tout simplement du fait que lors d'une transition du modèle de Moran, un seul allèle est modifié tandis que tous sont concernés à chaque transition du modèle de Wright-Fisher.

## 1.5 Exercices et travaux pratiques

## 1.6 Pour aller plus loin

1. **Technologie.** Norris [[Nor98](#)] , Bouleau [[Bou00](#)], Ycart [[Yca02](#)], Brémaud [[Bré99](#)], Ethier-Kurtz [[EK86](#)];
2. **Modèles.** Saint-Flour Tavaré [[Tav04](#)], Durret [[Dur02](#)], Ewens [[Ewe04](#)].

# Chapitre 2

## Processus de diffusion

Dans la section précédente nous avons étudié des processus de Markov dont la dynamique n'a lieu que par des sauts. Nous allons ici adopter le point de vue opposé en nous intéressant aux processus de Markov possédant des trajectoires continues. L'idée intuitive attachée au terme « diffusion » est celle d'une dynamique composée d'un mouvement déterministe de dérive et d'un mouvement aléatoire imprimé via un mouvement brownien, les trajectoires pouvant ainsi être très irrégulières tout en restant continues. Ce chapitre est loin d'être exhaustif sur le sujet. On pourra par exemple consulter [KS91] et [Bas98] pour plus de détails.

### 2.1 Équations différentielles stochastiques

Soit  $\sigma$  et  $b$  deux fonctions régulières (par exemple de classe  $\mathcal{C}^2$  à dérivées bornées) à valeurs respectivement dans  $\mathbb{R}_+$  et  $\mathbb{R}$ . Soit  $(B_t)_{t \geq 0}$  un mouvement brownien standard sur  $\mathbb{R}$  et  $X_0$  une variable aléatoire réelle indépendante de  $B$ . Il existe une unique solution  $(X_t)_{t \geq 0}$  à l'équation différentielle stochastique suivante :

$$\forall t \geq 0, \quad X_t = X_0 + \int_0^t b(X_s) ds + \int_0^t \sigma(X_s) dB_s. \quad (2.1)$$

Le processus  $(M_t)_{t \geq 0}$  défini par

$$\forall t \geq 0, \quad M_t = \int_0^t \sigma(X_s) dB_s$$

est une martingale et la formule d'Itô assure que, pour toute fonction  $h$  de classe  $\mathcal{C}^2$  sur  $\mathbb{R}$ ,

$$\forall t \geq 0, \quad h(X_t) = h(X_0) + \int_0^t Lh(X_s) ds + \int_0^t \sigma(X_s) h'(X_s) dB_s,$$

où

$$Lh(x) = \frac{1}{2}\sigma^2(x)h''(x) + b(x)h'(x),$$

et le processus  $M^h = (M_t^h)_{t \geq 0}$  défini par

$$\forall t \geq 0, \quad M_t^h = \int_0^t \sigma(X_s) h'(X_s) dB_s$$

est une martingale d'intégrale nulle, de carré intégrable avec

$$\mathbb{E} \left[ \left( M_t^f \right)^2 \right] = \int_0^t (\sigma(X_s) h'(X_s))^2 ds.$$

**Remarque 2.1.1.** En particulier, si  $h$  est solution de  $Lh = 0$  alors  $(h(X_t))_t$  est une martingale.

Posons pour  $h$  borélienne bornée sur  $\mathbb{R}^d$ ,

$$P_t h(x) = \mathbb{E}(h(X_t^x)),$$

où  $(X_t^x)_{t \geq 0}$  est la solution de l'équation (2.1) vérifiant  $X_0^x = x$ .

**Théorème 2.1.2.** *La famille  $(P_t)_t$  forme un semi-groupe de Markov de générateur infinitésimal  $L$ , c'est-à-dire que*

- $P_t(\cdot)(x)$  est une mesure de probabilité sur  $\mathbb{R}$ ,
- pour tous  $s, t \geq 0$ ,  $P_{t+s} = P_t \circ P_s$ ,
- pour tout  $t \geq 0$  et toute fonction  $h$  régulière,

$$\partial_t P_t h = \lim_{s \rightarrow 0} \frac{P_{t+s} h - P_t h}{s} = L P_t h = P_t h.$$

On peut reformuler ceci en écrivant que pour toute fonction régulière  $h$ ,

$$P_t h = h + \int_0^t L P_s h ds. \quad (2.2)$$

Ce résultat est le premier pas pour faire le lien entre le processus  $(X_t)_t$  et deux équations aux dérivées partielles. On peut en effet montrer le résultat suivant.

**Théorème 2.1.3.** *Il existe une fonction  $f : \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R} \rightarrow \mathbb{R}$  telle que la mesure  $P_t(\cdot)(x)$  admette pour densité par rapport à la mesure de Lebesgue la fonction  $y \mapsto f(t, x, y)$ . De plus,  $f$  est solution des équations aux dérivées partielles suivantes :*

$$\partial_t f(t, x, y) = \frac{1}{2} a(x) \partial_{xx}^2 f(t, x, y) + b(x) \partial_x f(t, x, y) = L_x f(t, x, y), \quad (2.3)$$

et

$$\partial_t f(t, x, y) = \frac{1}{2} \partial_{yy}^2 (a(y) f(t, x, y)) - \partial_y (b(y) f(t, x, y)) = L_y^* f(t, x, y). \quad (2.4)$$

De manière plus probabiliste, le théorème 2.1.2 assure en fait que la solution d'une équation différentielle stochastique est un processus de Markov fort.

**Théorème 2.1.4.** Soit  $X$  solution de (2.1). Alors, pour tout temps d'arrêt borné  $S$  et toute application  $\mathcal{F}_\infty$ -mesurable bornée  $Y$ ,

$$\mathbb{E}^x(Y \circ \theta_S | \mathcal{F}_S) = \mathbb{E}^{X_S}(Y) \quad p.s.$$

où  $\theta_s$  désigne l'opérateur de translation des trajectoires.

Voici quelques idées de la preuve.

$$\begin{aligned} X_{t+h}^x &= X_t^x + \int_t^{t+h} b(X_s^x) ds + \int_t^{t+h} \sigma(X_s^x) dB_s \\ &= X_t^x + \int_0^h b(X_{t+s}^x) ds + \int_0^h \sigma(X_{t+s}^x) dB_s^t \end{aligned}$$

où  $(B_s^t)_s = (B_{t+s} - B_t)_s$  est un mouvement brownien indépendant de  $(B_u)_{0 \leq u \leq t}$ . Pour tout  $y$  considérons  $Y^y$  la solution de

$$Y_h^y = y + \int_0^h b(Y_s^y) ds + \int_0^h \sigma(Y_s^y) dB_s^t.$$

L'unicité forte de la solution de l'EDS assure que

$$Y_{t+}^{X_t^x} = X_{t+}^x \quad p.s.$$

D'autre part, par indépendance des accroissements du mouvement brownien,  $Y^y$  est indépendant de  $\mathcal{F}_t$ . On a donc, pour toute fonction  $f$  régulière,

$$\begin{aligned} \mathbb{E}(f(X_{t+h}^x) | \mathcal{F}_t) &= \mathbb{E}\left(f\left(Y_h^{X_t^x}\right) | \mathcal{F}_t\right) = \mathbb{E}(f(Y_h^y))_{|y=X_t^x} \\ &= \mathbb{E}\left(f\left(Y_h^{X_t^x}\right) | X_t^x\right) = \mathbb{E}(f(X_{t+h}^x) | X_t^x). \end{aligned}$$

## 2.2 Séjour dans un intervalle

Il est possible de décrire parfaitement le comportement d'une diffusion réelle dans un intervalle donné en fonction des coefficients  $\sigma$  et  $b$ . On se donne un intervalle ouvert  $I = ]a, b[$  borné de  $\mathbb{R}$ .

### 2.2.1 Deux exemples simples

Soit  $B$  un mouvement brownien sur  $\mathbb{R}$ . Soit  $x \in I$ . Notons  $T_x$  le temps que met  $B$  à sortir de  $I$  partant de  $x$ . Puisque les trajectoires de  $B$  sont continues,  $B_{T_x} \in \{a, b\}$ . De plus,  $B$  est une martingale donc le théorème d'arrêt assure que

$$\mathbb{P}_x(B_{T_x} = b) = \frac{x-a}{b-a} \quad \text{et} \quad \mathbb{P}_x(B_{T_x} = a) = \frac{b-x}{b-a}. \quad (2.5)$$

De même,  $((B_t)^2 - t)_t$  est une martingale donc

$$\mathbb{E}^x((B_{t \wedge T})^2 - t \wedge T) = x^2.$$

On fait tendre  $t$  vers l'infini et l'on utilise (2.5) pour obtenir

$$\mathbb{E}^x(T) = (b - x)(x - a).$$

**Remarque 2.2.1.** Pour déterminer la loi des lieux de sortie de  $B$ , on a utilisé que  $B$  est une martingale.

Considérons à présent  $X$  le mouvement brownien avec dérive  $\mu$ , c'est-à-dire que  $X_t = B_t + \mu t$ . Ce n'est plus une martingale. Peut-on trouver une fonction  $S$  telle que  $(S(X_t))_t$  soit une martingale ? La formule d'Itô nous assure que, si  $S$  est de classe  $C^2$ ,

$$\begin{aligned} S(X_t) &= S(X_0) + \int_0^t \left( \frac{1}{2} S''(X_s) + \mu S'(X_s) \right) ds + \int_0^t S'(X_s) dB_s \\ &= S(X_0) + \int_0^t LS(X_s) ds + M_t, \end{aligned}$$

où  $L$  est le générateur infinitésimal de  $X$  et  $M$  est une martingale. Ainsi,  $S$  sera une martingale si  $S$  est solution de  $LS = 0$ . Cette équation est facile à résoudre :

$$S(x) = c_1 + c_2 e^{-2\mu x}.$$

Si l'on reproduit le raisonnement conduisant à (2.5), on a

$$\mathbb{E}^x(S(X_{t \wedge T_x})) = S(x),$$

et, en faisant tendre  $t$  vers l'infini,

$$\mathbb{P}_x(X_T = b) = \frac{e^{-2\mu x} - e^{-2\mu a}}{e^{-2\mu b} - e^{-2\mu a}} \quad \text{et} \quad \mathbb{P}_x(X_T = a) = \frac{e^{-2\mu b} - e^{-2\mu x}}{e^{-2\mu b} - e^{-2\mu a}}.$$

**Remarque 2.2.2.** La valeur des constantes  $c_1$  et  $c_2$  n'intervient nullement dans le résultat final.

**Remarque 2.2.3.** Lorsque  $\mu$  tend vers 0, on retrouve le cas brownien :  $\mathbb{P}_x(X_T = b)$  tend vers  $(x - a)/(x - b)$ .

Qu'en est-il de l'espérance de  $T_x$  ? Supposons que l'on dispose d'une fonction  $M$  régulière telle que, pour tout  $x \in I$ ,  $LM(x) = 1$ . Alors, la formule d'Itô conduit à

$$M(X_t) = M(X_0) + \int_0^t LS(X_s) ds + M_t = M(X_0) + t + M_t.$$

Ceci implique alors

$$\begin{aligned} \mathbb{E}^x(T) &= \mathbb{E}^x(M(X_T)) - M(x) \\ &= (M(b) - M(x))\mathbb{P}_x(X_T = b) + (M(a) - M(x))\mathbb{P}_x(X_T = a). \end{aligned}$$

Pour le mouvement brownien avec dérive, on trouve immédiatement

$$M(x) = c_1 + \frac{x}{\mu} + c_2 e^{-2\mu x} = \frac{x}{\mu} + S(x),$$

Après simplification, il vient,

$$\begin{aligned}\mathbb{E}^x(T) &= \left(\frac{b-x}{\mu}\right)\left(\frac{S(x)-S(a)}{S(b)-S(a)}\right) + \left(\frac{a-x}{\mu}\right)\left(\frac{S(x)-S(a)}{S(b)-S(a)}\right) \\ &= \left(\frac{b-x}{\mu}\right)\left(\frac{e^{-2\mu x}-e^{-2\mu a}}{e^{-2\mu b}-e^{-2\mu a}}\right) + \left(\frac{a-x}{\mu}\right)\left(\frac{e^{-2\mu b}-e^{-2\mu x}}{e^{-2\mu b}-e^{-2\mu a}}\right).\end{aligned}$$

**Remarque 2.2.4.** Lorsque  $\mu$  tend vers 0,  $\mathbb{E}^x(T)$  tend bien vers  $(b-x)(x-a)$ .

## 2.2.2 Cas général

Pour une diffusion générique sur  $I$ , il est possible de reproduire le raisonnement ci-dessus.

**Définition 2.2.5.** On définit la fonction d'échelle (scale function)  $S$  du processus  $(X_t)_{t \geq 0}$  par

$$S(x) = \int_c^x \exp\left(-2 \int_c^y \frac{b(z) dz}{\sigma^2(z)}\right) dy,$$

où  $c \in I$ .

La mesure de vitesse (speed measure)  $m$  est définie par

$$m(dx) = \mathbf{1}_{\{x \in I\}} \frac{2}{S'(x)\sigma^2(x)} dx = \mathbf{1}_{\{x \in I\}} \frac{2}{\sigma^2(x)} \exp\left(2 \int_c^x \frac{b(z) dz}{\sigma^2(z)}\right) dx.$$

**Remarque 2.2.6.** Les définitions de  $S$  et  $m$  dépendent de  $c$  mais ce ne sera plus le cas des formules intéressantes les mettant en jeu.

**Remarque 2.2.7.** La fonction  $S$  ainsi construite est solution de

$$\frac{1}{2}\sigma^2(x)S''(x) + b(x)S'(x) = 0, \quad (2.6)$$

$$S(c) = 0 \quad \text{et} \quad S'(c) = 1.$$

En particulier,  $LS = 0$ . Cette remarque est primordiale car elle montre que  $(S(X_t))_{t \geq 0}$  est une martingale.

**Proposition 2.2.8.** Si  $a < x < b$ ,  $X_0 = x$  et  $T_{a,b} = \inf\{t \geq 0, X_t \notin [a, b]\}$  alors

$$\mathbb{P}(X_{T_{a,b}} = a) = \frac{S(b) - S(x)}{S(b) - S(a)} \quad \text{et} \quad \mathbb{P}(X_{T_{a,b}} = b) = \frac{S(x) - S(a)}{S(b) - S(a)}. \quad (2.7)$$

De plus,

$$\mathbb{E}^x(T_{a,b}) = - \int_a^x (S(x) - S(y))m(dy) + \frac{S(x) - S(a)}{S(b) - S(a)} \int_a^b (S(b) - S(y))m(dy). \quad (2.8)$$

*Démonstration.* Appliquons la formule d'Itô à  $S(X_t)$  : grâce à (2.6), on a

$$S(x) = \mathbb{E}S(X_{T_{a,b}}) = S(a)\mathbb{P}(X_{T_{a,b}} = a) + S(b)\mathbb{P}(X_{T_{a,b}} = b).$$

Combinée avec  $\mathbb{P}(X_{T_{a,b}} = a) + \mathbb{P}(X_{T_{a,b}} = b) = 1$ , la relation précédente entraîne (2.7).

Calculons à présent l'espérance du temps d'atteinte. Posons, pour  $n \in \mathbb{N}^*$ ,

$$\tau_n = \inf \left\{ t \geq; \int_0^t \sigma^2(X_s) ds \geq n \right\}.$$

La fonction

$$M_{a,b}(x) = - \int_a^x (S(x) - S(y))m(dy) + \frac{S(x) - S(a)}{S(b) - S(a)} \int_a^b (S(b) - S(y))m(dy)$$

est solution de l'équation

$$b(x)M'(x) + \frac{1}{2}\sigma^2(x)M''(x) = 1 \quad \text{si } a < x < b,$$

$$M(a) = M(b) = 0.$$

La formule d'Itô appliquée à  $M_{a,b}(X_t)$  assure que

$$M_{a,b}(X_{t \wedge \tau_n \wedge T_{a,b}}) = M_{a,b}(x) - (t \wedge \tau_n \wedge T_{a,b}) + \int_0^{t \wedge \tau_n \wedge T_{a,b}} M'_{a,b}(X_s)\sigma(X_s)dB_s.$$

En prenant l'espérance et en faisant tendre  $n \rightarrow \infty$ , on obtient

$$\mathbb{E}(t \wedge T_{a,b}) = M_{a,b}(x) - \mathbb{E}M_{a,b}(X_{t \wedge T_{a,b}}) \leq M_{a,b}(x) < \infty.$$

Il reste à faire tendre  $t$  vers l'infini pour obtenir que  $\mathbb{E}(T_{a,b})$  est fini. Ceci nous permet de conclure que

$$\lim_{t \rightarrow \infty} \mathbb{E}M_{a,b}(X_{t \wedge T_{a,b}}) = \mathbb{E}M_{a,b}(X_{T_{a,b}}) = 0,$$

puisque  $M_{a,b}(a) = M_{a,b}(b) = 0$ . Ceci fournit (2.8).  $\square$

### 2.2.3 Fonction de Green

Étant donné un intervalle  $I = ]a, b[$  et  $T$  le temps de sortie de  $I$  pour la diffusion issue de  $x \in I$ , que peut-on dire du comportement de la diffusion avant  $T$ ? Nous avons déjà vu dans la proposition 2.2.8 que l'on peut exprimer l'espérance  $T$  en fonction de  $S$  et  $m$ . Plus généralement, soit  $g$  une fonction continue sur  $I$ . On veut trouver une représentation de la fonction

$$\forall x \in I, \quad w(x) = \mathbb{E} \left[ \int_0^T g(X_s) ds | X_0 = x \right]$$

en termes des fonctions  $S$  et  $m$ . La première étape est de remarquer que cette fonction est (la) solution d'une équation différentielle. On la résoudra ensuite.

**Théorème 2.2.9.** Soit  $g$  une fonction continue sur  $I$ . La fonction

$$\forall x \in I, \quad w(x) = \mathbb{E} \left[ \int_0^T g(X_s) ds | X_0 = x \right],$$

est l'unique solution de l'équation différentielle suivante :

$$\forall x \in ]a, b[, \quad Lw = -g(x), \quad w(a) = w(b) = 0. \quad (2.9)$$

*Démonstration.* Supposons que  $\tilde{w}$  soit solution de (2.9). Appliquons la formule d'Itô à  $\tilde{w}(X_t)$  entre les instants 0 et  $t \wedge T$  avec  $X_0 = x$  : il existe une martingale  $M$  telle que

$$\tilde{w}(X_{t \wedge T}) = \tilde{w}(X_0) + \int_0^{t \wedge T} L\tilde{w}(X_{s \wedge T}) ds + M_{t \wedge T} = \tilde{w}(x) - \int_0^{t \wedge T} g(X_s) ds + M_{t \wedge T}.$$

On prend l'espérance et l'on fait tendre  $t$  vers l'infini. Comme  $\tilde{w}$  est nulle au point  $X_T \in \{a, b\}$ , on obtient

$$0 = \tilde{w}(x) - \mathbb{E} \left[ \int_0^T g(X_s) ds | X_0 = x \right].$$

La réciproque est plus délicate à obtenir. Nous supposerons que  $w$  est de classe  $C^2$  sur  $I$ . Soit  $x \in I$  et  $h > 0$  un temps petit et  $\tau$  le temps de sortie de l'intervalle  $[x - \delta, x + \delta]$  où  $\delta$  est choisi de telle sorte que

$$|x - y| \leq \delta \Rightarrow |g(x) - g(y)| + |Lw(x) - Lw(y)| \leq \varepsilon.$$

La propriété de Markov forte assure que

$$\begin{aligned} w(x) &= \mathbb{E}^x \left( \int_0^{h \wedge \tau} g(X_s) ds + \int_{h \wedge \tau}^T g(X_s) ds \right) \\ &= \mathbb{E}^x \left( \int_0^{h \wedge \tau} g(X_s) ds \right) + \mathbb{E}^x(w(X_{h \wedge \tau})). \end{aligned}$$

On a donc

$$\begin{aligned} \left| \mathbb{E}^x \int_0^{h \wedge \tau} g(X_s) ds - g(x)h \right| &\leq \mathbb{E}^x \int_0^{h \wedge \tau} |g(X_s) - g(x)| ds + g(x)\mathbb{E}^x(h - h \wedge \tau) \\ &\leq \varepsilon \mathbb{E}^x(h \wedge \tau) + g(x)\mathbb{E}^x(h - h \wedge \tau) \\ &\leq (\varepsilon + g(x)\mathbb{P}_x(\tau < h))h. \end{aligned}$$

D'autre part, la formule d'Itô appliquée à  $w$  assure que

$$\mathbb{E}^x(w(X_{h \wedge \tau})) = w(x) + \mathbb{E}^x \left( \int_0^{h \wedge \tau} Lw(X_s) ds \right).$$

En faisant le même raisonnement que ci-dessus, on obtient finalement que

$$|w(x) - (g(x)h + w(x) + Lw(x)h)| \leq (\varepsilon + C\mathbb{P}_x(\tau < h))h.$$

La fonction  $w$  vérifie donc  $Lw(x) + g(x) = 0$  pour tout  $x \in I$ . Il est de plus évident par définition que  $w(a) = w(b) = 0$ .  $\square$

En dimension 1, il est possible de résoudre (2.9) ou, plus précisément, d'exprimer  $w$  en fonction de  $S$  et  $m$ . Soit  $c \in ]a, b[$ . L'équation (2.9) s'écrit encore :

$$\frac{d}{dx} \left( \exp \left( \int_c^x \frac{2b(z)}{\sigma^2(z)} dz \right) w'(x) \right) = -\frac{2g(x)}{\sigma^2(x)} \exp \left( \int_c^x \frac{2b(z)}{\sigma^2(z)} dz \right).$$

Par définition de  $S$  et  $m$ , l'égalité ci-dessus s'écrit encore :

$$\frac{d}{dx} \left( \frac{1}{S'(x)} w'(x) \right) = -2g(x)m(x).$$

Il existe  $\alpha, \beta \in \mathbb{R}$  tel que

$$w(x) = -2 \int_a^x S'(u) \left\{ \int_a^u g(v)m(v) dv \right\} du + \beta(S(x) - S(a)) + \alpha.$$

Puisque  $w(a) = 0$ , on a  $\alpha = 0$ . En utilisant le théorème de Fubini, on obtient

$$\begin{aligned} w(x) &= -2 \int_a^x \left\{ \int_v^x S'(u) du \right\} g(v)m(v) dv + \beta(S(x) - S(a)) \\ &= -2 \int_a^x \{S(x) - S(v)\} g(v)m(v) dv + \beta(S(x) - S(a)). \end{aligned}$$

On déduit de  $w(b) = 0$  la valeur de  $\beta$  :

$$\beta = \frac{2}{S(b) - S(a)} \int_a^b (S(b) - S(v)) g(v)m(v) dv.$$

La fonction  $w$  est donc déterminée :

$$\begin{aligned} w(x) &= \frac{2}{S(b) - S(a)} \left\{ (S(x) - S(a)) \int_a^b (S(b) - S(v)) g(v)m(v) dv \right. \\ &\quad \left. - (S(b) - S(a)) \int_a^x (S(x) - S(v)) g(v)m(v) dv \right\} \\ &= \frac{2}{S(b) - S(a)} \left\{ (S(x) - S(a)) \int_x^b (S(b) - S(v)) g(v)m(v) dv \right. \\ &\quad \left. + (S(b) - S(x)) \int_a^x (S(v) - S(a)) g(v)m(v) dv \right\}, \end{aligned}$$

où l'on a coupé la première intégrale  $\int_a^b = \int_a^x + \int_x^b$  et développé les produits.

**Théorème 2.2.10.** *Soit  $g$  une fonction continue sur  $I$ . Alors*

$$\mathbb{E} \left[ \int_0^T g(X_s) ds | X_0 = x \right] = \int_a^b G(x, v) g(v) dv,$$

avec, pour tout  $x \in ]a, b[$ ,

$$G(x, v) = \begin{cases} 2 \frac{S(x) - S(a)}{S(b) - S(a)} (S(b) - S(v)) m(v) & \text{si } x < v < b, \\ 2 \frac{S(b) - S(x)}{S(b) - S(a)} (S(v) - S(a)) m(v) & \text{si } a < v < x. \end{cases}$$

La fonction  $G$  est appelée fonction de Green du processus  $X$ .

**Remarque 2.2.11.** Soit  $a < c < d < b$ . En approchant  $\mathbf{1}_{[c,d]}$  par des fonctions continues, on obtient que

$$\int_c^d G(x, v) dv$$

est le temps de séjour moyen dans l'intervalle  $[c, d]$  de la diffusion  $X$  avant son absorption.

Le théorème 2.2.9 peut être établi pour une diffusion dans un ouvert de  $\mathbb{R}^d$ .

**Théorème 2.2.12.** *Le problème de Dirichlet sur un ouvert borné  $D$ , qui consiste à trouver  $u$  telle que*

$$\begin{cases} Lu(x) + h(x)u(x) + g(x) = 0 & \text{si } x \in D \\ u(x) = f(x) & \text{si } x \in \partial D, \end{cases}$$

admet une unique solution qui se représente sous la forme

$$u(x) = \mathbb{E} \left[ u(X_T) e^{\int_0^T h(X_s) ds} + \int_0^T g(X_s) e^{\int_0^s h(X_r) dr} ds \right].$$

*Éléments de preuve.* Indiquons ici pourquoi une solution du problème de Dirichlet admet la représentation probabiliste ci-dessus (ce qui fournit l'unicité). Soit  $u$  une solution. On applique la formule d'Itô : il existe une martingale  $M$  telle que

$$\begin{aligned} u(X_t) e^{\int_0^t h(X_s) ds} &= u(X_0) + \int_0^t (Lu(X_s) + u(X_s)h(X_s)) e^{\int_0^s h(X_r) dr} ds + M_t \\ &= u(X_0) - \int_0^t g(X_s) e^{\int_0^s h(X_r) dr} ds + M_t. \end{aligned}$$

On applique ceci au temps  $t \wedge T$  où  $T$  est le temps de sortie de  $D$ . En prenant l'espérance sachant que  $X_0 = x$  et en faisant tendre  $t$  vers l'infini, on obtient

$$u(x) = \mathbb{E} \left[ u(X_T) e^{\int_0^T h(X_s) ds} + \int_0^T g(X_s) e^{\int_0^s h(X_r) dr} ds \right].$$

La réciproque est bien plus pénible à montrer. □

## 2.3 Diffusions à coefficients singuliers

On souhaite à présent étudier des processus non plus à valeurs dans  $\mathbb{R}$  tout entier mais plutôt sur un intervalle ouvert

$$I = ]l, r[; \quad -\infty \leq l < r \leq +\infty.$$

Selon les coefficients de l'équation différentielle stochastique

$$dX_t = \sigma(X_t)dB_t + b(X_t)dt, \quad (2.10)$$

le processus restera dans un intervalle donné ou sera absorbé sur son bord. Comme nous le verrons le mouvement brownien ou le processus d'Ornstein-Uhlenbeck sortent presque sûrement de tout intervalle strictement inclus dans  $\mathbb{R}$ . Cela dit, ce n'est pas le cas de tous les processus. Considérons des E.D.S. de la forme

$$dX_t = dB_t + \left( \frac{a}{X_t} - \lambda X_t \right) dt$$

avec  $a > 0$  et  $\lambda \geq 0$ , ou encore

$$dX_t = \sqrt{1 - X_t^2} dB_t - (n - 1)X_t dt.$$

Dans le premier cas, le processus est repoussé de 0 par le terme  $a/x$  et attiré vers 0 par le terme  $-\lambda x$ . Lorsque  $x$  est petit le premier terme est bien sûr prépondérant et le second ne joue aucun rôle tandis que la situation s'inverse lorsque  $x$  est grand. La question est de savoir si le terme répulsif suffit à maintenir le processus dans  $]0, +\infty[$  ou si, partant de  $x > 0$ , il est possible d'atteindre 0 en un temps fini. La réponse que nous allons pouvoir apporter est la suivante : si  $a \geq 1$  alors, avec probabilité 1, le processus n'atteindra pas 0 et il l'atteindra à coup sûr dans le cas contraire.

Dans le deuxième exemple, où  $I = ]-1, 1[$ , le processus subit une force de rappel déterministe vers 0. De plus, lorsqu'il s'approche de 1 (ou -1) la partie diffusive à tendance à disparaître, ce qui permet à la force de rappel de devenir prédominante, à condition que  $n$  soit supérieur ou égal à 2.

### 2.3.1 Définitions

Pour que la solution de (2.10) reste dans l'intervalle  $I$  il semble nécessaire les coefficients soient singuliers aux bords finis de  $I$ . On doit donc définir une notion de solution pour l'EDS qui prenne en compte à la fois le fait que les coefficients peuvent être singuliers et que le processus peut ne pas rester dans  $I$ . L'idée est de considérer un processus qui, tant qu'il est dans  $I$ , évolue selon la dynamique (2.10), et qui est arrêté dès qu'il touche le bord de  $I$ .

**Définition 2.3.1.** Une solution faible dans l'intervalle  $I$  de (2.10) est un triplet

$$((X, B), (\Omega, \mathcal{F}, \mathbb{P}), \{\mathcal{F}_t\}),$$

tel que

1.  $(\Omega, \mathcal{F}, \mathbb{P})$  est un espace probabilisé et  $\{\mathcal{F}_t\}$  une filtration standard,
2.  $X = (X_t, \mathcal{F}_t; 0 \leq t)$  est un processus continu, adapté à valeurs dans  $[l, r]$  avec  $X_0 \in I$  p.s. et  $B = (B_t, \mathcal{F}_t; 0 \leq t)$  est un mouvement brownien standard sur  $\mathbb{R}$ .
3. pour toutes suites strictement monotones  $(l_n)_{n \in \mathbb{N}^*}$  et  $(r_n)_{n \in \mathbb{N}^*}$  telles que  $l < l_n < r_n < r$ ,  $\lim l_n = l$ ,  $\lim r_n = r$ , et

$$\text{pour } n \geq 1, \quad T_n = \inf \{t \geq 0; X_t \notin ]l_n, r_n[\},$$

on a, pour tout  $n \geq 1$ ,

$$\mathbb{P}\left(\forall t \geq 0, X_{t \wedge T_n} = X_0 + \int_0^t b(X_s) \mathbf{1}_{\{s \leq T_n\}} ds + \int_0^t \sigma(X_s) \mathbf{1}_{\{s \leq T_n\}} dB_s\right) = 1.$$

La variable aléatoire

$$T = \inf \{t \geq 0; X_t \notin ]l, r[\} = \lim_{n \rightarrow \infty} T_n$$

est appelé temps de sortie de  $I$ . L'hypothèse  $X_0 \in I$  garantie que  $\mathbb{P}(T > 0) = 1$ .

On supposera dans toute la suite que les coefficients  $\sigma : I \rightarrow \mathbb{R}$  et  $b : I \rightarrow \mathbb{R}$  sont réguliers et vérifient

$$\forall x \in I, \quad \sigma(x) > 0, \tag{ND}$$

$$\forall x \in I, \exists \varepsilon > 0 \text{ tel que } \int_{x-\varepsilon}^{x+\varepsilon} \frac{1 + |b(y)|}{\sigma^2(y)} dy > \infty. \tag{LI}$$

**Remarque 2.3.2.** Le processus  $X$  sort p.s. de tout intervalle compact inclus dans  $I$  en un temps fini et même intégrable.

**Proposition 2.3.3.** Supposons que (ND) et (LI) sont vérifiées et soit  $X$  une solution faible de (2.10) avec condition initiale déterministe  $X_0 = x \in I$ .

1. Si  $S(l+) = -\infty$  et  $S(r-) = +\infty$ , alors

$$\mathbb{P}(T = \infty) = \mathbb{P}\left(\sup_{t \geq 0} X_t = r\right) = \mathbb{P}\left(\inf_{t \geq 0} X_t = l\right) = 1.$$

En particulier, le processus n'explose pas et il est récurrent : pour tout  $y \in I$ ,

$$\mathbb{P}(\exists t > 0, X_t = y) = 1.$$

2. Si  $S(l+) > -\infty$  et  $S(r-) = +\infty$ , alors

$$\mathbb{P}\left(\lim_{t \uparrow T} X_t = l\right) = \mathbb{P}\left(\sup_{0 \leq t < T} X_t < r\right) = 1.$$

3. Si  $S(l+) = -\infty$  et  $S(r-) < +\infty$ , alors

$$\mathbb{P}\left(\lim_{t \uparrow T} X_t = r\right) = \mathbb{P}\left(\sup_{0 \leq t < T} X_t > l\right) = 1.$$

4. Si  $S(l+) > -\infty$  et  $S(r-) < +\infty$ , alors

$$\mathbb{P}\left(\lim_{t \uparrow T} X_t = l\right) = 1 - \mathbb{P}\left(\lim_{t \uparrow T} X_t = r\right) = \frac{S(r-) - S(x)}{S(r-) - S(l+)}.$$

*Démonstration.* D'après (2.7), pour  $l < a < x < b < r$ , on a

$$\mathbb{P}\left(\inf_{0 \leq t < T} X_t \leq a\right) \geq \mathbb{P}(X_{T_{a,b}} = a) = \frac{1 - S(x)/S(b)}{1 - S(a)/S(b)},$$

puisque si  $X$  a atteint  $a$  avant  $b$ , il a atteint  $a$  avant de sortir de  $I$  donc avant  $T$ . En faisant tendre  $b$  vers  $r$ , on obtient que, pour tout  $a < x$ ,  $\inf_{0 \leq t < T} X_t$  est inférieur à  $a$  p.s. Il ne reste plus qu'à écrire

$$\mathbb{P}\left(\inf_{0 \leq t < T} X_t = l\right) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} \left\{ \inf_{0 \leq t < T} X_t \leq l + 1/n \right\}\right) = 1.$$

On raisonne de même pour le supremum de  $X$ . Supposons à présent que  $\mathbb{P}(T < \infty) > 0$ . Alors l'événement

$$\left\{ \lim_{t \rightarrow T} X_t \text{ existe et est égal à } l \text{ ou } r \right\}$$

a une probabilité positive ce qui contredit le fait que les événements

$$\left\{ \inf_{0 \leq t < T} X_t = l \right\} \quad \text{et} \quad \left\{ \sup_{0 \leq t < T} X_t = r \right\}$$

sont de probabilité 1. □

**Remarque 2.3.4.** Le résultat précédent ne répond que partiellement à la question de la non-explosion : dans le cas 1. uniquement, on est assuré que  $T$  est infini. Par exemple, la diffusion associée à  $\sigma(x) = \sigma > 0$  et  $b(x) = \text{sgn}(x)$  entre dans le cas 4. et  $T$  est presque sûrement infini.

Le résultat suivant donne un classement complet des points du bord. On pourra en trouver une preuve dans [GS72].

**Définition 2.3.5.** On dit que  $l$ , extrémité inférieure de  $I$ , est

1. naturel (ou répulsif) si, pour tous  $l < x < b < r$ ,  $X^x$  atteint  $b$  avant  $l$  p.s. ;
2. attractif si, ou bien  $T_{[r,b]} = \infty$  et  $X_t^x$  tend vers  $r$  quand  $t$  tend vers l'infini, ou bien  $T_{[r,b]} < \infty$  et  $X_{T_{[r,b]}}^x = b$  ;
3. absorbant si, pour tout  $l < x < r$ ,  $X^x$  atteint  $l$  en un temps fini avec probabilité non nulle et y reste presque sûrement.
4. régulier si, pour tout  $l < x < r$ ,  $X^x$  atteint  $l$  en un temps fini avec probabilité non nulle et atteint tout point de  $I$  en temps fini presque sûrement.

Les points réguliers ou attractifs sont dits inaccessibles, les autres accessibles (en temps fini).

**Théorème 2.3.6.** *On pose, avec les notations de [GS72], pour  $y \in \bar{I}$ ,*

$$L_1(y) = S(y), \quad L_2(x) = \int_c^x S(y) m(dy) \quad \text{et} \quad L_3(x) = \int_c^x m(dy).$$

*Le point  $y$  du bord est :*

1. *naturel si  $L_1(y) = +\infty$  ;*
2. *attractif si  $L_1(y) < +\infty$  et  $L_2(y) = +\infty$  ;*
3. *absorbant si  $L_1(y) < +\infty$ ,  $L_2(y) < +\infty$  et  $L_3(y) = +\infty$  ;*
4. *régulier si  $L_1(y) < +\infty$ ,  $L_1(y) < +\infty$  et  $L_3(y) < +\infty$ .*

**Exercice 2.3.7.** Dans quels cas entrent les diffusions suivantes :

1. le mouvement brownien avec dérive  $X_t = \mu t + \sigma W_t$  sur  $I = ]-\infty, +\infty[$  ;
2. le mouvement brownien avec dérive  $X_t = \mu t + \sigma W_t$  sur  $I = ]0, +\infty[$  (discuter selon le signe de  $\mu$ ) ;
3. le processus d'Ornstein-Uhlenbeck sur  $\mathbb{R}$  avec  $\sigma(x) = 1$  et  $b(x) = -\lambda x$  ;
4. le processus de Bessel de dimension  $d$  sur  $I = ]0, +\infty[$  avec

$$\sigma(x) = 1 \quad \text{et} \quad b(x) = (d-1)/2x.$$

Discuter selon la valeur de  $d$  et justifier ;

5. le carré de Bessel de dimension  $d$  sur  $I = ]0, +\infty[$  avec

$$\sigma(x) = 1 \quad \text{et} \quad b(x) = (d-1)/2x - \lambda x,$$

où  $\lambda > 0$ . Discuter selon la valeur de  $d$  et justifier.



# Chapitre 3

## Applications à l'étude de chaînes de Markov

### 3.1 Convergence d'une chaîne de Markov vers une diffusion

L'objet de cette section est d'énoncer un théorème permettant très facilement de démontrer la convergence en loi d'une suite de chaînes de Markov convenablement renormalisées en temps vers un processus de diffusion dont les coefficients de diffusion et de dérive sont explicites.

#### 3.1.1 Un peu d'intuition

Le semi-groupe  $(P_t)_{t \geq 0}$  associé au générateur infinitésimal  $A$  est la solution de

$$\forall f, \forall t > 0, \quad P_t f = f + \int_0^t P_s A f \, ds. \quad (3.1)$$

Pour tout  $N \geq 1$ , notons  $K_N$  le noyau de la chaîne  $Y^N$  et  $A_N = K_N - I$ . Soit  $t > 0$ . Alors

$$\begin{aligned} K_N^{[Nt]} f &= f + \sum_{i=0}^{[(N-1)t]} K_N^i A_N f = f + \frac{1}{N} \sum_{i=0}^{[(N-1)t]} K_N^i (N A_N) f \\ &= f + \sum_{i=0}^{[(N-1)t]} \int_{i/N}^{(i+1)/N} K_N^i (N A_N) f \, ds = f + \int_0^t K_N^{[Ns]} (N A_N) f \, ds, \end{aligned}$$

puisque, si  $s \in [i/N, (i+1)/N]$  alors  $[Ns] = i$ . Ainsi, la famille de mesures de probabilité  $(K_N^{[Nt]})_{t \geq 0}$  est-elle solution d'une équation intégrale similaire à (3.1) mais où  $N A_N$  joue le rôle de  $A$ . On peut donc espérer que si la suite de terme général  $N A_N$  converge vers  $A$  alors  $(K_N^{[Nt]})_{t \geq 0}$  converge vers  $(P_t)_{t \geq 0}$ .

Étudions à présent le comportement de  $NA_N$ . Soit  $f$  une fonction régulière.

$$NA_N f(x) = N(K_N f(x) - f(x)) = N\mathbb{E}[(f(X_1^N) - f(x))|X_0^N = x].$$

Soit  $\varepsilon > 0$ .

$$NA_N f(x) = N\mathbb{E}\left[\mathbf{1}_{\{|X_1^N-x|\leqslant\varepsilon\}}(f(X_1^N) - f(x))|X_0^N = x\right] + O(N\mathbb{P}(|X_1^N - x| > \varepsilon)).$$

Enfin, grâce à un développement limité de  $f$  en  $x$  à l'ordre 2, on obtient

$$\begin{aligned} NA_N f(x) = & N\mathbb{E}[X_1^N - x|X_0^N = x]f'(x) + \frac{1}{2}N\mathbb{E}[(X_1^N - x)^2|X_0^N = x]f''(x) \\ & + o(\varepsilon^2) + O(N\mathbb{P}(|X_1^N - x| > \varepsilon)). \end{aligned}$$

Ainsi, en imposant une limite finie aux termes présents devant  $f'(x)$  et  $f''(x)$  et une bonne concentration de  $X_1^N$  autour de  $x$ , on peut espérer la convergence de  $NA_N$  vers le générateur infinitésimal d'une diffusion...

Nous pouvons à présent énoncer le théorème rigoureusement.

### 3.1.2 Le théorème

Le résultat suivant montre que la convergence en loi d'une suite de chaînes de Markov vers un processus de diffusion se lit sur un pas de la chaîne.

**Théorème 3.1.1.** *Soit  $A$  le générateur infinitésimal d'un processus de diffusion, c'est-à-dire que, pour toute fonction régulière  $f$ ,*

$$Af(x) = \frac{1}{2}a(x)f''(x) + b(x),$$

avec  $a$  et  $b$  régulières. Soit  $(\mu_N)_N$  une suite de noyaux de transition sur  $\mathbb{R}$  et posons

$$b_N(x) = N \int_{|y-x|\leqslant 1} (y-x)\mu_N(x, dy) \quad \text{et} \quad a_N(x) = N \int_{|y-x|\leqslant 1} (y-x)^2\mu_N(x, dy).$$

Supposons que pour tous  $r > 0$  et  $\varepsilon > 0$ ,

$$\sup_{|x|\leqslant r} |b_N(x) - b(x)| \rightarrow 0, \quad \text{et} \quad \sup_{|x|\leqslant r} |a_N(x) - a(x)| \rightarrow 0$$

et

$$\sup_{|x|\leqslant r} N\mu_N(x, \{y, |y-x| \geqslant \varepsilon\}) \rightarrow 0.$$

Soit  $(Y^N(n))_{n \geqslant 0}$  la chaîne de Markov de noyau de transition  $\mu_N$  et  $X^N$  le processus défini par  $X_t^N = Y^N([Nt])$ .

Si  $(Y^N(0))_N$  converge en loi vers  $\nu$  alors  $(X^N)$  converge en loi vers la diffusion de générateur  $A$  et de loi initiale  $\nu$ .

**Remarque 3.1.2 (Rien n'est simple).** En pratique, la chaîne de Markov  $Y^N$  est en général à valeurs dans un sous-ensemble  $I_N$  discret (et souvent fini) de  $\mathbb{R}$ . Ainsi, son noyau de transition n'est-il pas défini sur tout  $\mathbb{R}$ . Dans ce cas, il convient d'adapter le théorème ci-dessus en remplaçant  $x$  par une suite  $(x_N)$  telle que  $x_N \in I_N$  et qui converge vers  $x$ .

### 3.1.3 Marche aléatoire simple et mouvement brownien

Soit  $Z$  la marche aléatoire simple sur  $\mathbb{Z}$  définie par

$$Z_n = Z_0 + \sum_{i=1}^n \varepsilon_i$$

où  $(\varepsilon_i)_{i \geq 1}$  est une suite de v.a. i.i.d. (indépendante de  $Z_0$ ) de loi  $(1/2)\delta_1 + (1/2)\delta_{-1}$ . Pour  $N \geq 1$ , on définit  $(Y_n^N)_n$  à partir de  $Z$  en posant  $Y_n^N = Z_n/\sqrt{N}$ . Soit  $x \in \mathbb{R}$ . On pose  $i_N = [x\sqrt{N}]$  pour tout  $N \geq 1$ . On a alors

$$b_N(x_N) = N \left( \frac{1}{2\sqrt{N}} - \frac{1}{2\sqrt{N}} \right) = 0 \quad \text{et} \quad a_N(x_N) = N \left( \frac{1}{2N} + \frac{1}{2N} \right) = 1.$$

L'hypothèse de concentration du noyau  $\mu_N$  est clairement satisfaite puisque la mesure  $\mu_N(x_N, \cdot)$  est à support dans  $\{x_N - 1/\sqrt{N}, x_N + 1/\sqrt{N}\}$ . On en déduit donc que si  $(Y^N(0))_N$  converge en loi vers  $\nu$  alors  $(X^N)$  converge en loi vers le mouvement brownien de loi initiale  $\nu$ .

### 3.1.4 Marche aléatoire et mouvement brownien avec dérive

Soit  $Z$  la marche aléatoire sur  $\mathbb{Z}$  définie par

$$Z_n = Z_0 + \sum_{i=1}^n \varepsilon_i$$

où  $(\varepsilon_i)_{i \geq 1}$  est une suite de v.a. i.i.d. (indépendante de  $Z_0$ ) de loi  $p\delta_1 + q\delta_{-1}$ . On ne peut pas renormaliser cette marche de manière à obtenir à la limite le mouvement brownien avec dérive. En la renormaissant par  $N$ , on obtient uniquement le processus déterministe  $X_t = X_0 + (p - q)t$ . La dérive est trop forte. Il faut l'adoucir au cours du temps. Pour  $N \geq 1$ , on définit  $(Y_n^N)_n$  en posant

$$Y_n^N = Y_0^N + \frac{1}{\sqrt{N}} \sum_{i=1}^n \varepsilon_i^N$$

où  $(\varepsilon_i^N)_{i \geq 1}$  est une suite de v.a. i.i.d. (indépendante de  $Y_0^N$ ) de loi  $p_N\delta_1 + q_N\delta_{-1}$  avec  $p_N = (1 + \delta/\sqrt{N})/2$ . Soit  $x \in \mathbb{R}$ . On pose  $i_N = [x\sqrt{N}]$  pour tout  $N \geq 1$ . On a alors

$$b_N(x_N) = N \left( \frac{1 + \delta/\sqrt{N}}{2\sqrt{N}} - \frac{1 - \delta/\sqrt{N}}{2\sqrt{N}} \right) = \delta \quad \text{et} \quad a_N(x_N) = 1.$$

L'hypothèse de concentration du noyau  $\mu_N$  est encore clairement satisfaite. On en déduit donc que si  $(Y^N(0))_N$  converge en loi vers  $\nu$  alors  $(X^N)$  converge en loi vers le mouvement brownien avec dérive  $\delta$  et de loi initiale  $\nu$ .

### 3.1.5 Modèle d'Ehrenfest et processus d'Ornstein-Uhlenbeck

Soit  $N \geq 1$ . Considérons la chaîne d'Ehrenfest  $Z^N$  à valeurs dans  $\{0, \dots, N\}$  dont la matrice de transition est donnée par

$$\mathbb{P}(Z_1^N = j | Z_0^N = i) = \begin{cases} i/N & \text{si } j = i - 1, \\ 1 - i/N & \text{si } j = i + 1, \\ 0 & \text{sinon.} \end{cases}$$

Cette chaîne de Markov est irréductible récurrente positive et est réversible pour la mesure binomiale  $\mathcal{B}(N, 1/2)$ . Elle modélise l'échange de molécules de gaz entre deux compartiments.

On définit  $Y^N$  en posant  $Y^N = \sqrt{N}(Z^N/N - 1/2)$ . Soit  $x \in \mathbb{R}$ . On pose  $i_N = [x\sqrt{N} + N/2]$  pour tout  $N \geq 1$ . On a alors

$$b_N\left(\sqrt{N}\left(\frac{i_N}{N} - \frac{1}{2}\right)\right) = N\left(-\frac{i_N}{N}\frac{1}{\sqrt{N}} + \frac{N - i_N}{N}\frac{1}{\sqrt{N}}\right) = -2\sqrt{N}\left(\frac{i_N}{N} - \frac{1}{2}\right).$$

Donc  $b_N(x_N)$  tend vers  $b(x) = -2x$  quand  $N$  tend vers  $+\infty$ . De même,

$$a_N\left(\sqrt{N}\left(\frac{i_N}{N} - \frac{1}{2}\right)\right) = N\left(\frac{i_N}{N}\frac{1}{N} + \frac{N - i_N}{N}\frac{1}{N}\right) = 1.$$

L'hypothèse de concentration du noyau  $\mu_N$  est encore clairement satisfaite. On en déduit donc que si  $(Y^N(0))_N$  converge en loi vers  $\nu$  alors  $(X^N)$  converge en loi vers le processus d'Ornstein-Uhlenbeck de loi initiale  $\nu$  solution de l'EDS

$$dX_t = dB_t - 2X_t dt.$$

**Remarque 3.1.3.** La mesure invariante de ce processus est la mesure gaussienne  $\mathcal{N}(0, 1/4)$ , ce qui est cohérent avec le théorème limite central qui assure que si, pour tout  $N \geq 1$ ,  $Y_N \sim \mathcal{B}(N, 1/2)$  alors

$$\sqrt{N}\frac{Y_N}{N} - \frac{1}{2} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1/4).$$

### 3.1.6 Un modèle markovien d'endémie

On cherche à représenter de manière simplifiée l'épidémiologie d'une maladie et de déterminer les (des) conditions sous lesquelles elle peut être considérée comme endémique (c'est-à-dire qu'elle ne s'éteint jamais). Nous supposerons que cette maladie est de courte durée relative et ne confère pas d'immunité après un épisode infectieux. La population est de taille  $N$  supposé fixé (pour l'instant) mais  $N$  a vocation à être grand. Le temps est discret et le pas de temps est de l'ordre du jour. À chaque instant  $n \geq 0$ , la population admet une partition en deux classes aléatoires :  $F_{I,n}$  l'ensemble des individus infectés et  $F_{S,n}$  l'ensemble des individus susceptibles de l'être. On note

$I_n$  et  $S_n$  les cardinaux respectifs de  $F_{I,n}$  et  $S_{I,n}$ . Ces variables aléatoires sont à valeurs dans  $\{0, 1, \dots, N\}$  avec  $I_n + S_n = N$ . De plus, tous les individus infectés au temps  $n$  sont supposés guéris au temps  $n+1$  (la maladie est de courte durée). Un individu ne peut donc être infecté en deux instants consécutifs. Soit  $p \in [0, 1]$  le taux de contact infectieux, c'est-à-dire la probabilité que, entre les instants  $n$  et  $n+1$ , un contact entre deux individus, l'un infecté, l'autre susceptible de l'être, aboutisse à l'infection de l'individu sain. On note  $q = 1 - p$ . On suppose que chaque individu infecté tente de contaminer tout individu sain de indépendamment des autres entre deux instants. Cela revient à dire que si  $I_n$  individus sont infectés au temps  $n$ , la probabilité pour qu'un individu sain soit contaminé est  $1 - q^{I_n}$ .

**Proposition 3.1.4.** *La suite  $(I_n)_n$  est une chaîne de Markov de matrice de transition*

$$\forall i, j = 0, \dots, N, \quad P(i, j) = \begin{cases} C_{N-i}^j (1 - q^i)^j q^{i(N-i-j)} & \text{si } i + j \leq N, \\ 0 & \text{sinon.} \end{cases}$$

De manière plus compacte, on peut donc dire que, pour tout  $i \in \{0, \dots, N\}$ ,

$$\mathcal{L}(I_{n+1}|I_n = i) = \mathcal{B}(N - i, 1 - q^i).$$

En particulier,

$$\mathbb{E}(I_{n+1}|I_n = i) = (N - i)(1 - q^i) \quad \text{et} \quad \mathbb{V}(I_{n+1}|I_n = i) = (N - i)q^i(1 - q^i).$$

On remarque que le point  $N$  est isolé (on ne peut l'atteindre d'un autre point) et il mène à 0 en un coup. D'autre part, 0 est le seul état absorbant et tous les états mènent à 0. Donc  $(I_n)_n$  converge presque sûrement vers 0 et le temps d'atteinte de 0 est fini presque sûrement et même intégrable. Il est de fait possible d'obtenir une formule explicite de la série génératrice du temps d'atteinte de 0 (et donc de ses moments) en fonction de la matrice de transition de la chaîne.

Nous allons plutôt ici supposer, comme pour le processus de Wright-Fisher, que  $N$  tend vers l'infini, renormaliser le nombre d'individus infectés pour en faire une proportion et dilater le temps. Nous allons également faire tendre le taux de contact infection  $p$  vers 0. Pour tout  $N \geq 1$ , on considère la chaîne de Markov  $(I_n^N)_n$  définie ci-dessus mais en remplaçant  $q$  par  $q_N = e^{-A/N}$ .

**Proposition 3.1.5.** *La suite de processus  $(I_{[N]}^N/N)$  converge vers la diffusion sur  $[0, 1]$  solution de*

$$dX_t = \sqrt{\frac{1}{2}(1 - X_t)(1 - e^{-AX_t})e^{-AX_t}} dB_t + (1 - X_t)(1 - e^{-AX_t}) dt.$$

**Exercice 3.1.6.** Que peut-on dire sur ce processus de diffusion ?

## 3.2 D'autres exemples de convergence de processus

La convergence des processus de Markov se voit sur la convergence des générateurs infinitésimaux en toute généralité. Voici le théorème sous sa forme générale (et peu digeste). On en trouvera une démonstration dans [EK86].

### 3.2.1 Le megathéorème de convergence

Soit  $L$  un espace de Banach muni de la norme  $\|\cdot\|$ . Pour tout  $n \geq 1$ , on introduit un espace de Banach  $L_n$  et une application linéaire  $\pi_n$  de  $L_n$  dans  $L$ . On suppose que  $\sup_n \|\pi_n\| < +\infty$ . On note  $f_n \rightarrow f$  si, pour tout  $n \geq 1$ ,  $f_n \in L_n$ ,  $f \in L$  et  $\lim_n \|f_n - \pi_n f\| = 0$ .

**Remarque 3.2.1.** Dans les cas pratiques,  $L$  sera l'ensemble des fonctions continues sur un ensemble  $K$  et  $L_n$  sera l'ensemble des fonctions continues sur un sous-ensemble  $K_n$  de  $K$ . L'application  $\pi_n$  sera alors simplement la restriction à  $K_n$  pour une fonction définie sur  $K$ .

**Théorème 3.2.2.** Pour  $n \in \mathbb{N}^*$ , soit  $T_n$  un endomorphisme sur  $L_n$ . Soit  $(\varepsilon_n)_n$  une suite de réels strictement positifs qui converge vers 0. On pose pour  $n \geq 1$ ,  $A_n = \varepsilon_n^{-1}(T_n - I)$ . Soit  $(T(t))_{t \geq 0}$  un semi-groupe de Feller sur  $L$  de générateur  $A$  et soit  $D$  un ensemble dense dans le domaine de  $A$ . Les propriétés suivantes sont équivalentes :

- Pour toute fonction  $f \in L$ ,  $T_n^{[t/\varepsilon_n]} f$  converge vers  $T(t)f$  pour tout  $t \geq 0$  uniformément sur tout intervalle de temps borné.
- Pour toute fonction  $f \in L$ ,  $T_n^{[t/\varepsilon_n]} f$  converge vers  $T(t)f$  pour tout  $t \geq 0$ .
- Pour toute fonction  $f \in D$ , il existe des fonctions  $f_n \in L_n$  pour tout  $n \geq 1$ , telles que

$$f_n \rightarrow f \quad \text{et} \quad A_n f_n \rightarrow A f.$$

### 3.2.2 Espace d'états fini : du temps discret au temps continu

On considère une chaîne de Markov à espace d'états fini (pour simplifier) de matrice de transition  $P$ . On va examiner ce qui se passe lorsque l'on étire le temps tout en augmentant (en bonne proportion) la probabilité de faire du sur place.

**Corollaire 3.2.3.** Soit  $Z$  une chaîne de Markov à temps discret sur un espace d'états fini  $E$  de noyau de transition  $P$ . Pour tout  $N \geq 1$ , on considère la chaîne à temps discret  $X^N$  de noyau de transition  $(1 - 1/N)I + (1/N)P$ . Alors  $(X_{[Nt]}^N)_{t \geq 0}$  converge en loi vers la chaîne de Markov à temps continu de générateur  $P - I$ .

Ce résultat s'obtient avec cette version du théorème 3.2.2.

**Théorème 3.2.4.** Soit  $Z^N$  une chaîne de Markov de transition  $T_N$  à valeurs dans  $K_N \subset \mathbb{R}$ . S'il existe une suite  $(\varepsilon_N)_N$  de réels strictement positifs et  $G$  le générateur

d'un processus de Markov  $X$  tels que

$$\lim_{N \rightarrow \infty} \sup_{x \in K_N} \left| \frac{(T_N - I)f(x) - Gf(x)}{\varepsilon_N} \right| = 0.$$

pour toute fonction  $f \in D$  où  $D$  est dense dans le domaine  $G$ , alors la suite de processus  $(X^N)_N$  définie par

$$\forall N \geq 1, \quad \forall t \geq 0, \quad X_t^N = Z^N([t/\varepsilon_N]),$$

converge en loi vers le processus  $X$ .

### 3.2.3 File d'attente en temps continu et processus d'Ornstein-Uhlenbeck

On montre ici un exemple de chaîne de Markov à temps continu célèbre qui converge vers un processus de diffusion non moins célèbre. Soit  $\lambda$  et  $\mu$  deux réels strictement positifs. On définit le processus de Markov  $Z$  à temps continu à valeurs dans  $\mathbb{N}$  de générateur infinitésimal  $A$  donné par

$$A(i, j) = \begin{cases} \lambda & \text{si } j = i + 1, \\ i\mu & \text{si } j = i - 1, \\ -(\lambda + i\mu) & \text{si } j = i, \\ 0 & \text{sinon.} \end{cases}$$

Partant de  $i$ , le processus  $Z$  y reste un temps exponentiel de paramètre  $\lambda + i\mu$  puis saute en  $i + 1$  avec probabilité  $\lambda/(\lambda + i\mu)$  ou en  $i - 1$  avec probabilité  $i\mu/(\lambda + i\mu)$ . Ce processus s'appelle la file  $M/M/\infty$  : elle modélise le nombre de clients dans une file d'attente où les personnes arrivent selon un processus de Poisson de paramètre  $\lambda$  et sont servis aussitôt arrivés en un temps exponentiel de paramètre  $\mu$ . On peut se convaincre que

$$\forall k \in \mathbb{N}, \quad \mathcal{L}(Z_t | Z_0 = k) = \mathcal{B}(k, e^{-\mu t}) * \mathcal{P}\left(\lambda \frac{1 - e^{-\mu t}}{\mu}\right).$$

En particulier,  $Z$  admet  $\mathcal{P}(\lambda/\mu)$  pour mesure invariante (symétrique).

**Proposition 3.2.5.** Pour tout  $N \geq 1$ , notons  $Z^N$  le processus  $M/M/\infty(N\lambda, \mu)$  et  $X^N$  le processus défini par

$$X_t^N = \sqrt{N} \left( \frac{Z_t^N}{N} - \frac{\lambda}{\mu} \right) \quad \text{avec} \quad Z_0^N = [N\lambda/\mu + \sqrt{N}x].$$

Alors la suite  $(X^N)$  converge en loi vers le processus de diffusion  $X$  solution de l'équation différentielle stochastique suivante :

$$dX_t = \sqrt{2\lambda} dB_t - \mu X_t dt \quad \text{avec} \quad X_0 = x,$$

c'est-à-dire vers un processus d'Ornstein-Uhlenbeck.

### 3.2.4 Modèles de croissance de populations

Modélisons la taille d'une population par un processus de vie et mort à temps continu dont la dynamique est la suivante. Chaque individu meurt avec un taux  $\mu$  et se dédouble avec un taux  $\lambda$  indépendamment des autres. Nous allons supposer que la population et le temps sont grands (mesurés en unités de taille  $N$ ). Si  $\lambda - \mu$  n'est pas de l'ordre de  $1/N$  cette renormalisation sera triviale (la population devient nulle ou infinie immédiatement). Supposons que  $N(\lambda - \mu) = b$  et  $\lambda + \mu = 2a + O(1/N)$ . Si  $X_t^N$  est le nombre d'individus du processus au temps  $t$ , on considère qu'il y a  $NX_t^N$  dans notre population. Chacun des  $X_t^N$  individus a deux horloges exponentielles qui tournent au dessus de sa tête. L'instant (aléatoire) où un événement (mort ou division) va se produire est donc de loi  $\mathcal{E}(N(\lambda + \mu))$  et il s'agit d'une mort avec probabilité  $\mu/(\lambda + \mu)$ .

Pour le processus  $Y^N$ , chaque individu meurt avec un taux  $\mu_N$  et se dédouble avec un taux  $\lambda_N$  indépendamment des autres et on note  $X^N$  le processus défini par  $X^N(t) = Y^N([tN])/N$ . Le processus  $X^N$  est à valeurs dans  $K_N = (1/N)\mathbb{N}$  et son générateur infinitésimal est défini par

$$\forall i, j \in \mathbb{N}, \quad L_N\left(\frac{i}{N}, \frac{j}{N}\right) = \begin{cases} i\lambda_N & \text{si } j = i+1, \\ i\mu_N & \text{si } j = i-1, \\ -i(\lambda_N + \mu_N) & \text{si } j = i, \\ 0 & \text{sinon.} \end{cases}$$

En particulier, pour toute fonction de classe  $\mathcal{C}^2$  sur  $\mathbb{R}^+$ ,

$$\begin{aligned} L_N f\left(\frac{i}{N}\right) &= i\lambda_N \left[ \left(\frac{i+1}{N}\right) - f\left(\frac{i}{N}\right) \right] + i\mu_N \left( f\left(\frac{i-1}{N}\right) - f\left(\frac{i}{N}\right) \right) \\ &= (\lambda_N - \mu_N) \frac{i}{N} f'\left(\frac{i}{N}\right) + \frac{\lambda_N + \mu_N}{2N} \frac{i}{N} f''\left(\frac{i}{N}\right) + O\left(\frac{(\lambda_N + \mu_N)i}{N^3}\right). \end{aligned}$$

Supposons que  $\lambda_N - \mu_N$  converge vers  $b$  et  $(\lambda_N + \mu_N)/N$  converge vers  $2a$  en choisissant par exemple

$$\lambda_N = aN + \frac{b}{2} \quad \text{et} \quad \mu_N = aN - \frac{b}{2}$$

On obtient alors que si  $(x_N)_N$  est une suite de réels telle que  $x_N \in K_N$  pour tout  $N$  et qui converge vers  $x > 0$  alors

$$L_N f(x_N) \xrightarrow[N \rightarrow \infty]{} Lf(x) := axf''(x) + bx f'(x).$$

La suite de processus de Markov à espaces d'états discrets converge donc vers une diffusion  $X$  solution de l'équation différentielle stochastique suivante

$$dX_t = \sqrt{2aX_t} dB_t + bX_t dt.$$

### 3.3 Application aux processus de Wright-Fisher

On souhaite approcher par un processus de diffusion la chaîne de Wright-Fisher avec mutation et sélection.

#### 3.3.1 Le cas général

Pour tout  $N$  fixé, soit  $(Z_n^N)_n$  la chaîne de Markov sur  $K_N = \{0, 1/(2N), \dots, 1\}$  de noyau de transition

$$\mathbb{P}(Z_{n+1}^N = j/(2N) | Z_n^N = i/(2N)) := C_{2N}^j (\eta_i^*)^j (1 - \eta_i^*)^{2N-j}$$

pour  $0 \leq i, j \leq 2N$  avec  $\eta_i^* = \eta_i(1-u) + (1-\eta_i)v$  et

$$\eta_i = \frac{(1+s)i^2 + (1+sh)i(2N-i)}{(1+s)^2 i^2 + 2(1+sh)i(2N-i) + (2N-i)^2}.$$

On définit également le processus  $X^N$

$$\forall t \geq 0, \quad X^N(t) := Z^N([2Nt]).$$

C'est un processus à temps continu dont les trajectoires sont à valeurs dans  $D_K[0, \infty)$  avec  $K := [0, 1]$ . De fait, ses trajectoires sont constantes par morceaux.

Pour que l'approximation par une diffusion ait un sens, on doit supposer que  $s$ ,  $u$  et  $v$  dépendent de  $N$  et sont de l'ordre de  $N^{-1}$ .

**Théorème 3.3.1.** *Posons*

$$\alpha = 2Ns_N, \quad \beta_1 = 2Nu_N \quad \text{et} \quad \beta_2 = 2Nv_N, \quad (3.2)$$

avec  $\alpha$ ,  $\beta_1$  et  $\beta_2$  d'ordre 1. Alors la suite  $(X_t^N)_{t \geq 0}$  converge en loi vers la solution de l'EDS suivante :

$$dX_t = \sigma(X_t) dB_t + b(X_t) dt.$$

où

$$\sigma^2(x) = x(1-x) \quad \text{et} \quad b(x) = \alpha x(1-x)(x+h(1-2x)) - \beta_1 x + \beta_2(1-x). \quad (3.3)$$

*Démonstration.* Par définition de  $Z^N$ , on a

$$\begin{aligned} 2N\mathbb{E}[Z_1^N - x | Z_0^N = x] &= \alpha x(1-x)(x+h(1-2x)) - \beta_1 x + \beta_2(1-x) + O(N^{-1}), \\ 2N\mathbb{E}[(Z_1^N - x)^2 | Z_0^N = x] &= \alpha x(1-x) + O(N^{-1}), \\ 2N\mathbb{E}[(Z_1^N - x)^3 | Z_0^N = x] &= O(N^{-3/2}) \\ 2N\mathbb{E}[(Z_1^N - x)^4 | Z_0^N = x] &= O(N^{-1}), \end{aligned}$$

uniformément en  $x \in K_N$ . Le théorème 3.1.1 permet de conclure. □

**Remarque 3.3.2.** Pour transférer les estimations de temps d'absorption de la diffusion à la chaîne de Markov, il faudra prendre garde à les multiplier par  $2N$ . Les probabilités d'atteinte ou les profils de mesures invariantes sont pour leurs parts insensibles à ce changement d'échelle temporelle.

Exprimons la fonction d'échelle :

$$S(x) = \int_c^x y^{-2\beta_2} (1-y)^{-2\beta_1} \exp(\alpha(2h-1)y^2 - 2\alpha hy) dy,$$

et la mesure de vitesse :

$$m(x) = \int_c^x y^{2\beta_2-1} (1-y)^{2\beta_1-1} \exp(-\alpha(2h-1)y^2 + 2\alpha hy) dy.$$

**Proposition 3.3.3.** *Le point 1 est régulier (accessible et non absorbant) si  $0 < \beta_1 < 1/2$  inaccessible et non absorbant si  $\beta > 1/2$ .*

*Si  $\beta_1 = 0$  le point 1 est accessible et absorbant.*

Détaillons à présent les divers cas qui peuvent se présenter.

### 3.3.2 Ni mutation, ni sélection

Dans ce cas, la dérive de la diffusion est nulle. Le processus de diffusion est donc une martingale et on obtient directement que la probabilité de sortie du segment  $[0, 1]$  en 1 partant de  $x$  vaut  $x$ .

On a aussi l'expression de l'espérance du temps de fixation :

$$\mathbb{E}(T|X_0 = x) = -2\{x \ln x + (1-x) \ln(1-x)\},$$

et sa variance :

$$\mathbb{V}(T|X_0 = x) = 4 \left( x \int_x^1 \lambda(y) dy - (1-x) \int_0^x \lambda(y) dy \right) - \mathbb{E}(T|X_0 = x)^2,$$

où  $\lambda$  est une primitive de la fonction

$$-2 \left( \frac{\ln y}{1-y} + \frac{\ln(1-y)}{y} \right).$$

**Remarque 3.3.4.** Le temps de fixation pour la chaîne de Wright-Fisher a une espérance de l'ordre de  $2N\mathbb{E}(T|X_0 = x)$  et une variance de l'ordre de  $4N^2\mathbb{V}(T|X_0 = x)$ .

### 3.3.3 Mutation sans sélection

On suppose ici que  $\alpha$  est nul. La suite  $X^N$  converge, lorsque  $N$  tend vers l'infini vers le processus solution de l'EDS suivante

$$dX_t = \sqrt{X_t(1-X_t)} dB_t - \beta_1 X_t dt + \beta_2 (1-X_t) dt.$$

On peut montrer que cette diffusion reste bien dans l'intervalle  $]0, 1[$  et qu'elle admet pour mesure invariante  $\nu$  la loi bêta de paramètres  $2\beta_2$  et  $2\beta_1$ , c'est-à-dire que

$$\nu(y) = \frac{\Gamma(2(\beta_1 + \beta_2))}{\Gamma(2\beta_1)\Gamma(2\beta_2)} y^{2\beta_2-1} (1-y)^{2\beta_1-1}.$$

En particulier, si  $Y$  suit la loi  $\nu$ ,

$$\mathbb{E}(Y) = \frac{\beta_2}{\beta_1 + \beta_2} \quad \text{et} \quad \mathbf{Var}(Y) = \frac{\beta_1\beta_2}{(\beta_1 + \beta_2)^2(2(\beta_1 + \beta_2) + 1)}.$$

**Remarque 3.3.5.** Dans le modèle à  $K$  allèles distincts avec un taux de mutation total  $\theta/2$  et des probabilités respectives de mutation  $(\pi_i)_{1 \leq i \leq K}$ , le processus des proportions de chaque allèle vit sur le simplexe de  $\mathbb{R}^K$  (de dimension  $K - 1$ ) et la densité de sa mesure invariante est

$$\nu(y_1, \dots, y_K) = \frac{\Gamma(\theta)}{\Gamma(\theta\pi_1)\cdots\Gamma(\theta\pi_K)} y_1^{\Gamma(\theta\pi_1)-1} \cdots y_K^{\Gamma(\theta\pi_K)-1},$$

où  $y_i > 0$  pour  $i = 1, \dots, K$  et  $y_1 + \cdots + y_K = 1$ .



# Chapitre 4

## Généalogie

On souhaite ici présenter un modèle d'évolution génétique permettant d'étudier la structure d'arbres phylogéniques. Ces arbres décrivent l'évolution des espèces à partir d'un ancêtre commun, en précisant les instants de différenciation des espèces. Étant donné un groupe de  $k$  individus, on s'intéresse en particulier à la datation du premier ancêtre commun à ce groupe et à la taille de l'arbre généalogique de cet individu.

Cette étude permettra dans un second temps de montrer comment estimer le taux de mutation d'un gène à partir du nombre d'allèles différents portés par  $n$  individus.

### 4.1 Construction de l'arbre généalogique

Rappelons les hypothèses biologiques associées à notre modèle :

- la taille de la population reste constante au cours du temps, égale à  $N$ ,
- les générations ne se chevauchent pas : à chaque instant  $k$ , la  $k$ -ième génération meurt et donne naissance aux  $N$  individus de la  $(k+1)$ -ième génération,
- toutes les combinaisons reliant chacun des enfants à son parent ont une même probabilité,
- la reproduction à l'instant  $k$  est indépendante des reproductions précédentes.

Traduisons en termes mathématiques les hypothèses ci-dessus. À chaque génération, on numérote les individus. Soit l'individu  $i$  à la génération  $k+1$ , on note  $a_i^{k+1}$  son parent à la génération  $k$ .

**Hypothèse 4.1.1.** Les variables aléatoires  $(a_i^{k+1})_{1 \leq i \leq N, k \in \mathbb{N}}$  sont indépendantes et de même loi uniforme sur  $\{1, \dots, N\}$ .

On note  $\nu_i^k$  le nombre de descendants à la génération  $k+1$  de l'individu  $i$  vivant à la génération  $k$ . Les variables aléatoires  $(\nu_i^k)_{1 \leq i \leq N}$  ne sont pas indépendantes puisqu'elles doivent vérifier la relation  $\nu_1^k + \dots + \nu_N^k = N$ .

**Lemme 4.1.2.** *Le  $N$ -uplet  $\nu^k = (\nu_1^k, \dots, \nu_N^k)$ , conditionnellement au fait que la population a une taille constante, suit la loi multinomiale que nous noterons  $\mu_N$  de*

paramètre  $(N, (1/N, \dots, 1/N))$ , c'est-à-dire que

$$\begin{aligned} \mathbb{P}(\nu_1^k = m_1, \dots, \nu_N^k = m_N | \nu_1^k + \dots + \nu_N^k = N) \\ = \frac{N!}{m_1! \dots m_N!} \left( \frac{1}{N} \right)^N \mathbf{I}_{\{m_1 + \dots + m_N = N\}}. \end{aligned} \quad (4.1)$$

De plus, les variables aléatoires  $(\nu_i^k)_k$  sont indépendantes.

*Démonstration.* L'indépendance des vecteurs aléatoires  $(\nu^k)_k$  découle immédiatement de celle des  $(a_i^k)$ . La première partie du lemme correspond à l'interprétation probabiliste de la loi multinomiale : on répartit  $N$  boules dans  $N$  urnes, indépendamment et selon la loi uniforme.  $\square$

**Remarque 4.1.3.** En particulier, chaque individu de la génération  $n+1$  possède un ancêtre à la génération  $n$  choisi indépendamment des autres et selon la loi uniforme sur  $\{1, \dots, N\}$ . Certains membres de la génération  $k$  n'ont donc aucun descendant à la génération  $k+1$ . Ceci va entraîner le phénomène dit de dérive génétique.

**Exercice 4.1.4.** Soit  $Y_1, \dots, Y_N$  des variables aléatoires indépendantes et identiquement distribuées de loi de Poisson de paramètre  $\theta$ . Alors, la loi de la variable aléatoire  $(Y_1, \dots, Y_N)$ , conditionnellement à l'événement  $Y_1 + \dots + Y_N = N$  suit la loi multinomiale  $\mu_N$ . La loi de  $\nu_i^k$  est donc la loi binomiale  $\mathcal{B}(N, 1/N)$ .

**Remarque 4.1.5.** Supposons que chaque individu soit identifié à un allèle  $A$  ou  $B$  d'un gène à deux allèles. Alors, si le nombre d'allèles  $A$  à la génération  $k$  est égal à  $i$ , la loi du nombre d'allèles  $A$  à la génération  $k+1$  suivra une loi binomiale  $\mathcal{B}(N, i/N)$ . On retrouve donc le modèle de Wright-Fisher.

Le modèle de Wright-Fisher sans mutation est connu pour faire apparaître un phénomène dérive génétique, c'est-à-dire qu'un allèle l'emporte sur l'autre et que la population devient homozygote après un nombre aléatoire mais fini (presque sûrement) de générations. De même, après un certain nombre de générations, tous les individus proviendront donc d'un même ancêtre. C'est ce phénomène qu'il s'agit à présent de décrire et quantifier.

## 4.2 Le modèle à temps discret

D'après la remarque 4.1.3, la probabilité que deux individus aient deux parents distincts (à la génération précédente) est égale à  $1 - 1/N$ . Ces parents choisissant eux-mêmes leurs parents indépendamment et aléatoirement, la probabilité pour que le premier ancêtre commun à deux individus donnés remonte à plus de  $r$  générations vaut  $(1 - 1/N)^r$ . Le temps d'apparition de l'ancêtre commun le plus récent suit donc une loi géométrique de paramètre  $1/N$ .

Si l'on considère à présent trois individus, notons  $T'_3$  l'instant d'apparition de l'ACPR pour deux individus au moins (premier temps de coalescence parmi trois

individus). On a

$$\mathbb{P}(T'_3 > 1) = \frac{N(N-1)(N-2)}{N^3} = 1 - \frac{3}{N} + \frac{2}{N^2}.$$

Comme les choix des parents sont indépendants à chaque génération, il vient, pour  $k \geq 1$ ,

$$\mathbb{P}(T'_3 > k) = \left(1 - \frac{3}{N} + \frac{2}{N^2}\right)^k.$$

La loi de  $T'_3$  est donc la loi géométrique de paramètre  $p_3 = 1 - 3/N + 2/N^2$ . On note  $T'_3 + T'_2$  le temps d'apparition d'un ACPR pour les trois individus. Remarquons que si l'ACPR de deux individus est l'ACPR des trois individus alors  $T'_2 = 0$ . Déterminons la loi de  $T'_2$ . La probabilité que trois individus distincts aient le même père est  $1/N^2$ . La probabilité que  $T'_2$  soit nul sachant que  $T'_3 = k$  est donc égale à

$$\mathbb{P}(T'_2 = 0 | T'_3 = k) = \frac{\mathbb{P}(T'_2 = 0, T'_3 = k)}{\mathbb{P}(T'_3 = k)} = \frac{(1-p_3)^{k-1}/N^2}{p_3(1-p_3)^{k-1}} = \frac{1}{3N-2}.$$

Comme les choix des parents sont indépendants d'une génération à l'autre, on déduit de la première partie que la loi de  $T'_2$  sachant que  $T'_3 = k$  et  $T'_2 > 0$  est la loi géométrique de paramètre  $p_2$ . Cela implique que sachant  $T'_3 = k$ ,  $T'_2$  est égal en loi à  $UV$  où  $U$  et  $V$  sont indépendantes de lois respectives  $\mathcal{B}((3N-3)/(3N-2))$  et  $\mathcal{G}(p_2)$ .

Plus généralement, le nombre de parents distincts d'un groupe de  $k$  individus peut être vu comme le nombre d'urnes occupées après que l'on a lancé  $k$  balles, indépendamment et à chaque fois uniformément, dans  $N$  urnes. Pour tout  $j = 1, \dots, k$ , on notera la probabilité pour que ce nombre soit  $j$  de la manière suivante :

$$\begin{aligned} g_{kj}^{(N)} &= \mathbb{P}(k \text{ individus ont } j \text{ parents distincts}) \\ &= N(N-1) \cdots (N-j+1) \mathcal{S}_k^{(j)} N^{-k}, \end{aligned}$$

où  $\mathcal{S}_k^{(j)}$  est le nombre de façons de découper un ensemble à  $k$  éléments en  $j$  ensembles non vides (ces nombres sont parfois appelés nombres de Stirling de deuxième espèce). L'expression de  $g_{kj}$  est obtenue par le raisonnement suivant :  $N(N-1) \cdots (N-j+1)$  est le nombre de façons de choisir  $j$  parents distincts,  $\mathcal{S}_k^{(j)}$  est le nombre de façons d'associer à ces  $j$  parents  $k$  enfants, enfin,  $N^k$  est le nombre de façons d'assigner  $k$  enfants à leurs parents. Définissons à présent le processus ancestral  $(A_n^N(r))_{r \in \mathbb{N}}$  en notant  $A_n^N(r)$  le nombre d'ancêtres distincts à la génération  $r$  pour un groupe de taille  $n$  au temps 0.

**Proposition 4.2.1.** *La suite  $A_n^N(\cdot)$  est une chaîne de Markov sur  $\{1, \dots, n\}$  de matrice de transition  $G_N = (g_{jk}^{(N)})_{j,k}$ .*

Il est évident que 1 est l'unique état absorbant et que les états  $2, \dots, n$  sont transients (puisque tous mènent à 1). Cependant, il est difficile d'étudier les propriétés fines de cette chaîne, comme par exemple des propriétés sur le temps d'atteinte de l'état 1. Nous allons donc remplacer ce modèle à temps discret par un modèle plus simple à temps continu.

### 4.3 Le modèle à temps continu

On considère le processus de Markov  $(A(t))_{t \geq 0}$  sur  $\mathbb{N}^*$  à temps continu de générateur défini par

$$\forall j, k \in \mathbb{N}^*, \quad L(j, k) = \begin{cases} C_n^2 & \text{si } j \geq 2 \text{ et } k = j - 1, \\ -C_n^2 & \text{si } j \geq 2 \text{ et } k = j, \\ 0 & \text{sinon.} \end{cases}$$

C'est un processus de mort pur (le processus ne peut sauter que de  $n$  à  $n - 1$  si  $n \geq 2$ ) pour lequel 1 est absorbant. Le temps de séjour en  $n$  suit la loi exponentielle de paramètre  $C_n^2$ . Ceci s'interprète de la manière suivante : chaque couple d'individus se cherche un père indépendamment des autres et y arrive au bout d'un temps exponentiel de paramètre 1. Pour  $n$  individus,  $C_n^2$  couples distincts se cherchent indépendamment. Or le minimum de  $C_n^2$  variables aléatoires indépendantes et identiquement distribuées de loi  $\mathcal{E}(1)$  est la loi  $\mathcal{E}(C_n^2)$ .

On notera  $(A_n(t))_{t \geq 0}$  le processus issu de  $n$ . Celui-ci est à valeurs dans  $\{1, \dots, n\}$ .

### 4.4 Le passage à la limite

Montrons à présent que le processus à temps discret correctement renormalisé converge vers le processus à temps continu défini ci-dessus. On est intéressé par le cas limite où  $N$  est grand et où l'unité de temps se compte en  $N$  générations. La date d'apparition de l'ACPR de deux individus donnés est donc  $T'_2/N$  où  $T'_2$  suit la loi géométrique de paramètre  $p_2 = \mathcal{G}(1/N)$ .

**Lemme 4.4.1.** *Soit  $(V_n)_n$  une suite de variables aléatoires de loi géométrique de paramètres respectifs  $(\mu_n)_n$  telle que  $n\mu_n$  converge vers  $\mu > 0$  quand  $n \rightarrow +\infty$ . La suite  $(V_n/n)_n$  converge en loi vers une variable aléatoire  $V$  de loi exponentielle de paramètre  $\mu$ .*

*Démonstration.* Il suffit de regarder ce qui se passe pour les fonctions caractéristiques :

$$\varphi_{V_n/n}(t) = \frac{\mu_n e^{it/n}}{1 - (1 - \mu_n)e^{it/n}} = \frac{n\mu_n}{n\mu_n - n(1 - e^{it/n})} \xrightarrow[n \rightarrow +\infty]{} \frac{\mu}{\mu - it} = \varphi_V(t),$$

où  $\mathcal{L}(V) = \mathcal{E}(\mu)$ . □

Dans le cas d'une population comportant  $N$  individu et dans une échelle de temps où l'unité de temps est égale à  $N$  générations, le temps  $T_2$  d'apparition de l'ACPR pour deux individus donnés a approximativement pour loi la loi exponentielle de paramètre 1. En particulier  $\mathbb{E}(T_2) = 1$ . On a donc montré que

$$(A_2^N([Nt]))_{t \geq 0} \xrightarrow[N \rightarrow \infty]{\mathcal{L}} (A_2(t))_{t \geq 0}.$$

Considérons l'arbre généalogique de trois individus. Quand  $N$  tend vers l'infini,  $T'_2/N$  converge en loi vers  $\mathcal{E}(1)$ . De plus, on montre de même que  $(T'_2/N, T'_3/N)$  converge vers  $(T_2, T_3)$  où  $T_2$  et  $T_3$  sont indépendantes de lois exponentielles respectives  $\mathcal{E}(1)$  et  $\mathcal{E}(3)$  :

$$\begin{aligned}\varphi_{(T'_2/N, T'_3/N)}(s, t) &= \mathbb{E}(\exp(isT'_2/N + itT'_3/N)) \\ &= \mathbb{E}[\mathbb{E}(\exp(isT'_2/N)|T'_3)\exp(itT'_3/N)] \\ &= \frac{1}{3N-2}\mathbb{E}[e^{itT'_3/N}] + \frac{3N-3}{3N-2}\mathbb{E}[e^{itV/N}]\mathbb{E}[e^{itT'_3/N}]\end{aligned}$$

Il ne reste plus qu'à utiliser le lemme 4.4.1. On a ainsi montré que  $(A_3^N([Nt]))_{t \geq 0}$  converge en loi vers le processus  $(A_3(t))_{t \geq 0}$  quand  $N$  tend vers l'infini. En d'autres termes, on a le résultat suivant.

**Proposition 4.4.2.** *Lorsque  $N$  tend vers l'infini et que le temps est mesuré en unités de  $N$  générations, le mécanisme d'apparition de l'ACPR d'un groupe composé de trois individus distincts pour le processus limite est le suivant :*

- après un temps exponentiel  $T_3$  de paramètre 3, un ancêtre commun à deux individus apparaît,
- l'ancêtre commun apparaît alors après un temps  $T_2$  de loi exponentielle  $\mathcal{E}(1)$ ,
- les variables aléatoires  $T_2$  et  $T_3$  sont indépendantes.

En particulier, à la limite, la probabilité que le premier ancêtre commun soit commun aux trois individus est nulle.

On peut généraliser le résultat ci-dessus.

**Proposition 4.4.3.** *Pour tout  $n \geq 2$ ,*

$$(A_n^N([Nt]))_{t \geq 0} \xrightarrow[N \rightarrow \infty]{\mathcal{L}} (A_n(t))_{t \geq 0}.$$

*Démonstration.* Les grandes lignes de la preuve sont esquissées ici. Considérons dans un premier temps le comportement de la matrice de transition de  $G_N$  lorsque  $N$  est grand. Puisque  $\mathcal{S}_k^{(k-1)} = C_k^2$ , pour tout  $2 \leq k \leq n$ ,

$$g_{k,k-1}^{(N)} = \mathcal{S}_k^{(k-1)} \frac{N(N-1) \cdots (N-k+1)}{N^k} = C_k^2 \frac{1}{N} + O(N^{-2}).$$

Pour  $j < k-1$ , on a

$$g_{k,j}^{(N)} = S_k^{(j)} \frac{N(N-1) \cdots (N-j+1)}{N^k} = O(N^{-2}).$$

Enfin,

$$g_{k,k}^{(N)} = N^{-k} N(N-1) \cdots (N-k+1) = 1 - C_k^2 \frac{1}{N} + O(N^{-2}).$$

On peut donc écrire  $G_N = I + N^{-1}Q + O(N^{-2})$ .

Reste alors à trouver la limite de  $(G_N)^{[Nt]} \dots$

□

## 4.5 Propriétés du processus ancestral

Le processus  $A_n(\cdot)$  est un processus de mort pur, c'est-à-dire qu'il est à valeurs dans  $\mathbb{N}$ , possède des trajectoires continues par morceaux et décroissantes. Il est issu de  $A_n(0) = n$  et décroît uniquement avec des sauts d'amplitude 1. Pour tout  $k \geq 2$ , sachant que le processus est dans l'état  $k$ , le temps d'attente  $T_k$  avant de passer dans l'état  $k - 1$  (seule transition possible) suit une loi exponentielle de paramètre  $C_k^2$ . De plus, les variables aléatoires  $(T_k)_{2 \leq k \leq n}$  sont indépendantes.

### 4.5.1 Expression de la matrice de transition

Soit  $G(t) = (g_{ij}(t))_{ij}$  la matrice de taille  $n \times n$  définie par

$$G_{ij}(t) = \mathbb{P}(A_i(t) = j).$$

La matrice  $G(t)$  est égale à  $\exp(tQ)$  où  $Q$  est la matrice (de taille  $n$ ) dont les coefficients non nuls sont

$$\text{pour } k = n, n-1, \dots, 2, \quad q_{kk} = -C_k^2, \quad q_{k,k-1} = C_k^2.$$

On peut ici donner une expression explicite la matrice  $G(t)$  (en diagonalisant la matrice  $Q$ ). Le résultat général est le suivant.

**Proposition 4.5.1.** *On a, pour tout  $j = 1 \dots, n$ ,*

$$g_{nj}(t) = \sum_{k=j}^n e^{-k(k-1)t/2} \frac{(2k-1)(-1)^{k-j} j_{(k-1)} n_{[k]}}{j!(k-j)! n_{(k)}},$$

avec les notations

$$a_{(n)} = a(a+1) \cdots (a+n-1), \quad a_{[n]} = a(a-1) \cdots (a-n+1) \quad \text{et} \quad a_{(0)} = a_{[0]} = 1.$$

Le nombre moyen d'ancêtres au temps  $t$  est donné par

$$\mathbb{E} A_n(t) = \sum_{k=1}^n e^{-k(k-1)t/2} (2k-1) \frac{(r+k-2)!}{(r-1)!(k-r)}.$$

### 4.5.2 Temps d'apparition de l'ACPR

Soit  $W_n$  le temps d'apparition de l'ACPR. On a

$$W_n = T_n + \cdots + T_2,$$

où les  $(T_k)$  sont les temps de séjours dans les états  $2, \dots, n$ . En particulier, on a donc

$$\mathbb{E}(W_n) = 2 \left( 1 - \frac{1}{n} \right) \quad \text{et} \quad 1 = \mathbb{V}(W_2) \leqslant \mathbb{V}(W_n) \leqslant \lim_{n \rightarrow \infty} \mathbb{V}(W_n) = \frac{4\pi^2}{3} - 12 \simeq 1.16.$$

Ces calculs suggèrent que la contribution essentielle semble provenir de  $T_2$ .

On peut obtenir une expression explicite de la loi de  $W_n$  en remarquant que :

$$\mathbb{P}(W_n \leq t) = \mathbb{P}(A_n(t) = 1) = \sum_{k=1}^n \frac{(2k-1)(-1)^{k-1} n_{[k]}}{n_{(k)}} e^{-k(k-1)t/2}. \quad (4.2)$$

On peut donc obtenir une expression explicite de la densité de la loi de  $W_n$ .

### 4.5.3 Longueur de l'arbre

On appelle arbre généalogique d'un groupe de  $n$  individus l'ensemble de tous leurs ancêtres toutes générations comprises, eux compris, jusqu'au premier ancêtre commun à tous les individus. On note  $L_n$ , et l'on appelle *longueur de l'arbre généalogique* la variable aléatoire égale à la somme des temps de vie de tous les individus de l'arbre.

**Proposition 4.5.2.** *La longueur de l'arbre  $L_n$  s'exprime en fonction des temps d'apparition des ancêtres commun :*

$$L_n = 2T_2 + \cdots + nT_n.$$

*En particulier,*

$$\mathbb{E}(L_n) \sim 2 \log n \quad \text{et} \quad \mathbb{V}(L_n) \sim \frac{2\pi^2}{3}.$$

*Démonstration.* On utilise le fait que si  $T$  suit la loi exponentielle de paramètre  $\lambda$  alors son espérance vaut  $1/\lambda$ , sa variance  $1/\lambda^2$  et, pour  $\alpha > 0$ ,  $\alpha T$  suit la loi exponentielle de paramètre  $\lambda/\alpha$ . L'indépendance des variables aléatoires  $(T_i)_i$  fait le reste.  $\square$

On peut en fait déterminer complètement la loi de  $L_n$ .

**Proposition 4.5.3.** *La variable aléatoire  $L_n$  suit la loi du maximum de  $n-1$  variables aléatoires indépendantes et de même loi exponentielle de paramètre  $1/2$ .*

*Démonstration.* Pour  $n \geq 2$ , notons  $f_n$  la densité de la loi de  $L_n$  et, pour  $\lambda > 0$ ,  $g_\lambda$  la densité de la loi exponentielle de paramètre  $\lambda$ . Par définition,  $f_2 = g_{1/2}$ . La densité  $f_3$  est obtenue par convolution de  $f_2$  et  $g_1$  : pour tout  $x > 0$ ,

$$f_3(x) = f_2 * g_1(x) = \frac{1}{2} \int_0^x e^{-u/2} e^{-(x-u)} du = e^{-x/2} (1 - e^{-x/2}).$$

Faisons l'hypothèse de récurrence suivante : pour tout  $x > 0$ ,

$$f_n(x) = \frac{n-1}{2} e^{-x/2} (1 - e^{-x/2})^{n-2}.$$

On a alors, pour tout  $x > 0$ ,

$$f_{n+1}(x) = f_n * g_{n/2}(x) = \cdots = \frac{n}{2} e^{-x/2} (1 - e^{-x/2})^{n-1}.$$

D'autre part, puisque la fonction de répartition de la loi exponentielle de paramètre  $1/2$  vaut  $1 - e^{-x/2}$  pour  $x > 0$ , on retrouve bien le fait que  $f_n$  est la densité de la loi du maximum de  $n - 1$  variables aléatoires indépendantes et identiquement distribuées de loi  $\mathcal{E}(1/2)$ .  $\square$

**Corollaire 4.5.4.** *La suite  $(L_n / \ln n)_n$  converge presque sûrement vers 2 et tandis que  $(L_n - 2 \ln n)_n$  converge en loi, quand  $n$  tend vers l'infini, vers la mesure de probabilité de fonction de répartition*

$$F(t) = \exp(-\exp(-t/2)).$$

*Démonstration.* Démontrons la convergence en loi. Il suffit d'écrire la fonction de répartition de  $L_n - 2 \ln n$  en utilisant la proposition 4.5.3 : pour tout  $t \in \mathbb{R}$ ,

$$\mathbb{P}(L_n - 2 \ln n \leq t) = \mathbb{P}(M_{n-1} \leq t + 2 \ln n) = \left(1 - \frac{1}{n} e^{-t/2}\right)^{n-1} \xrightarrow[n \rightarrow \infty]{} F(t).$$

Pour la preuve de la convergence presque sûre, on peut se reporter à Barbe-Ledoux.  $\square$

On peut montrer directement que la convergence presque sûre de  $(L_n / \ln n)_n$  vers 2 grâce au théorème des séries centrées et au lemme de Kronecker.

**Théorème 4.5.5** (Séries centrées). *Soit  $(X_n)_n$  une suite de v.a. indépendantes, centrées et de carré intégrable. Si la série de terme général  $\mathbb{E}(X_n^2)$  converge alors la série (aléatoire) de terme général  $X_n$  converge presque sûrement et dans  $L^2$ .*

**Lemme 4.5.6.** *Soit  $(b_n)_n$  une suite croissante de réels strictement positifs tendant vers  $+\infty$  et  $(x_n)_n$  une suite de réels. Si la série  $\sum_n b_n^{-1} x_n$  converge alors  $b_n^{-1} \sum_{k=1}^n x_k$  converge vers 0.*

## 4.6 Price en compte des mutations

Pour rendre compte de la diversité des individus dans une population, il faut tenir compte des possibilités de mutation de l'ADN, en particulier lors de la réPLICATION. Les mutations entraînent une diversifications des enfants d'un même individu.

Le taux de mutation d'une base est très faible. On suppose pour simplifier que ce taux ne dépend pas de la base concernée, ni de sa position dans l'ADN et qu'il est constant au cours du temps. Ainsi, à chaque génération, un individu a une probabilité  $\mu$  d'avoir une mutation qui le différencie de son parent.

Si l'on considère une séquence de  $m$  bases, la probabilité que deux mutations aient lieu au même endroit est  $1/m$ . Comme les probabilités de mutation sont faibles, cela arrive très rarement. On fera donc l'hypothèse que l'on a une infinité d'allèles possibles et que chaque nouvelle mutation affecte un site différent. Ainsi, **une mutation donne toujours un nouvel allèle**. Le temps d'apparition d'une mutation dans la lignée ancestrale d'un individu est donc une loi géométrique de paramètre  $\mu$ . On

pose  $\theta = 2\mu N$  et on suppose que  $\theta$  est d'ordre 1. Ainsi, lors du passage à la limite en  $N$  tend vers l'infini, le processus de mutation devient un processus de Poisson. Plus exactement, on munit chaque branche de l'arbre d'un processus de Poisson de paramètre  $\theta$  qui comptera les mutations.

Les deux processus de coalescence et mutation sont d'origines aléatoires différentes. Nous les considérerons indépendants.

La question est d'estimer  $\theta$ .

#### 4.6.1 Loi du nombre d'allèles dans un échantillon

On note  $K_n$  le nombre d'allèles distincts dans un groupe de  $n$  personnes. Il est possible de décrire la loi de  $K_n$ .

**Proposition 4.6.1.** *On a*

$$K_n \stackrel{\mathcal{L}}{=} \sum_{k=1}^n \eta_k,$$

avec  $(\eta_k)_{k \geq 1}$  des variables aléatoires indépendantes de loi de Bernoulli de paramètres respectifs  $(\theta/(k-1+\theta))_{k \geq 1}$ . En particulier,  $\eta_1 = 1$  p.s.

*Démonstration.* Pour un groupe de  $n$  individus, on s'intéresse au premier temps (dans le passé) d'apparition d'une coalescence ou d'une mutation. Tout se passe comme si  $U_n$  était le minimum entre  $T_n$ , premier temps de coalescence (de loi  $\mathcal{E}(C_n^2)$ ) et  $R_1, \dots, R_n$  premiers temps de mutation des individus  $1, \dots, n$  de même loi  $\mathcal{E}(\theta/2)$ . Puisqu'elles sont indépendantes, leur minimum est encore une loi exponentielle et son paramètre est égal à

$$C_n^2 + n\theta/2 = n(n-1+\theta)/2.$$

De plus, la probabilité que ce premier phénomène soit une coalescence est

$$\frac{C_n^2}{n(n-1+\theta)/2} = \frac{n-1}{n-1+\theta}.$$

De même, la probabilité pour que ce phénomène soit une mutation est donc

$$1 - \frac{n-1}{n-1+\theta} = \frac{\theta}{n-1+\theta}.$$

On pose  $\eta_n = 1$  si ce premier phénomène est une mutation et  $\eta_n = 0$  si c'est une coalescence. La variable aléatoire  $\eta_n$  suit donc la loi de Bernoulli de paramètre  $\theta/(n-1+\theta)$ . Étudions ce qui se passe après ce temps  $U_n$  :

- Si  $\eta_n = 0$  alors le nombre d'ancêtres à l'instant  $U_n$  est  $n-1$  et le nombre d'allèles distincts dans ce groupe de  $n-1$  individus est  $K_{n-1} = K_n$ .
- Si  $\eta_n = 1$  alors l'individu qui a muté à l'instant  $U_n$  est à l'origine d'un allèle différent. Il correspond, dans la population initiale de  $n$  individus, à la présence d'un allèle distinct de tous les autres. De plus, cet allèle est présent une seule fois dans la population des  $n$  ancêtres. On obtient donc une population de  $n-1$  individus possédant  $K_{n-1} = K_n - 1$  allèles distincts.

Dans tous les cas, on a obtenu  $K_n = K_{n-1} + \eta_n$  et on considère à partir de l'instant  $U_n$  une population de  $n - 1$  individus possédant  $K_{n-1}$  allèles différents. Par récurrence, on obtient bien la décomposition de  $K_n$  en somme de variables aléatoires de Bernoulli. La propriété de Markov du processus de coalescence et l'absence de mémoire du processus de mutation assure l'indépendance des variables aléatoires  $(\eta_k)_k$ .  $\square$

En particulier, on peut calculer les deux premiers moments de  $K_n$ .

$$\mathbb{E}(K_n) = \sum_{k=1}^n \frac{\theta}{k-1+\theta} = \theta \ln n + O(1).$$

En effet, par comparaison série intégrale, on a

$$\int_{\theta}^{n+\theta} \frac{dx}{x} \leq \mathbb{E}(K_n) \frac{1}{\theta} + \int_{\theta}^{n-1+\theta} \frac{dx}{x},$$

et donc

$$\theta \ln n + \theta \ln(1 + \theta/n) - \theta \ln \theta \leq \mathbb{E}(K_n) \leq \theta \ln n + \theta \ln(1 + (\theta - 1)/n) - \theta \ln \theta.$$

Calculons à présent la variance de  $K_n$  : par indépendance des variables aléatoires  $(\eta_k)_k$ ,

$$\begin{aligned} \mathbb{V}(K_n) &= \sum_{k=1}^n \frac{\theta}{k-1+\theta} \left(1 - \frac{\theta}{k-1+\theta}\right) \\ &= \mathbb{E}(K_n) - \theta^2 \sum_{k=1}^n \frac{1}{(k-1+\theta)^2} = \theta \ln n + O(1). \end{aligned}$$

On s'intéresse au comportement de  $K_n$  lorsque  $n$  est grand. L'inégalité de Tchebychev assure alors que

$$\mathbb{P}\left(\left|\frac{K_n}{\ln n} - \theta\right| \geq \varepsilon\right) \leq \frac{\mathbb{V}(K_n) + (\mathbb{E}(K_n) - \theta \ln n)^2}{(\varepsilon^2 \ln n)^2} = \frac{1}{\varepsilon^2} O((\ln n)^{-1}).$$

**Corollaire 4.6.2.** *La suite  $(K_n / \ln n)_n$  converge vers  $\theta$  en probabilité.*

On cherche de plus à contrôler les fluctuations de  $K_n / \ln n$  autour de  $\theta$ . Nous présentons ici deux façons de procéder. La première s'appuie sur une variante du théorème limite central, la seconde consiste à comparer la loi de  $K_n$  à une loi de Poisson.

**Proposition 4.6.3.** *On a la convergence en loi suivante :*

$$\frac{K_n - \mathbb{E}(K_n)}{\sqrt{\mathbb{V}(K_n)}} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1)$$

Première preuve de 4.6.3. On utilise ici la version suivante du théorème limite central.

**Théorème 4.6.4.** Soit  $(X_n)_n$  une suite de variables aléatoires indépendantes centrées admettant toutes un moment d'ordre 3 et de variances respectives  $(\sigma_n^2)_n$ . On suppose que

$$\sum_{k=1}^{\infty} \sigma_k^2 = +\infty \quad \text{et} \quad \lim_{n \rightarrow \infty} \frac{\sum_{k=1}^n \mathbb{E}(|X_k|^3)}{\left(\sum_{k=1}^n \sigma_k^2\right)^{3/2}} = 0.$$

Alors

$$\frac{\sum_{k=1}^n X_k}{\sqrt{\sum_{k=1}^n \sigma_k^2}} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

On pose  $X_k = \eta_k - \mathbb{E}(\eta_k)$ . Bien entendu  $\mathbb{E}(X_k^2) = \mathbb{V}(\eta_k)$  et donc

$$\sum_{k=1}^n \mathbb{V}(X_k) = \mathbb{V}(K_n) = \theta \ln n + O(1).$$

De plus,

$$\mathbb{E}(|X_k|^3) = \mathbb{E}(\eta_k^3) - 3\mathbb{E}(\eta_k^2)\mathbb{E}(\eta_k) + \mathbb{E}(\eta_k)\mathbb{E}(\eta_k)^2 - \mathbb{E}(\eta_k)^3 = \mathbb{E}(\eta_k) - 3\mathbb{E}(\eta_k)^2 + 2\mathbb{E}(\eta_k)^3.$$

En particulier,  $\sum_{k=1}^n \mathbb{E}(X_k^3) = \theta \ln n + O(1)$ . On donc

$$\frac{\sum_{k=1}^n \mathbb{E}(|X_k|^3)}{\left(\sum_{k=1}^n \sigma_k^2\right)^{3/2}} = \frac{\theta \ln n + O(1)}{\left(\theta \ln n + O(1)\right)^{3/2}} \xrightarrow[n \rightarrow \infty]{} 0.$$

□

Seconde preuve de 4.6.3. Elle s'appuie sur le théorème suivant.

**Théorème 4.6.5** (Barbour, Holst, Janson). Soit  $W = \xi_1 + \dots + \xi_n$  la somme de  $n$  variables aléatoires indépendantes et de lois de Bernoulli de paramètres respectifs  $p_1, \dots, p_n$  et  $P$  une variable aléatoire de loi de Poisson de paramètre  $p_1 + \dots + p_n$ . Alors

$$d_{VT}(\mathcal{L}(W), \mathcal{L}(P)) \leq \frac{p_1^2 + \dots + p_n^2}{p_1 + \dots + p_n}.$$

Ainsi, il existe une constante  $c$  telle que

$$d_{VT}(\mathcal{L}(K_n), \mathcal{L}(P_n)) \leq \frac{c}{\ln n},$$

où  $P_n$  suit la loi de Poisson de paramètre  $\mathbb{E}(K_n)$ . On en déduit donc que

$$\frac{K_n - \mathbb{E}K_n}{\sqrt{\mathbb{V}(K_n)}} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

Ce résultat reste vrai en remplaçant l'espérance et la variance de  $K_n$  par  $\theta \ln n$ . □

### 4.6.2 Loi du nombre de mutations

En fait, les données dont on dispose en comparant les séquences d'ADN sont plus riches que la donnée du nombre d'allèles différents. En effet, on dispose, pour un échantillon de  $n$  personnes, du nombre de sites  $S_n$  où ont eu lieu les mutations dans tout l'arbre généalogique. Il y a forcément eu plus de mutations dans l'arbre que d'allèles différents dans la population des  $n$  individus.

Remarquons que dans le modèle initial, le nombre de mutation observées au cours de  $[Nt]$  générations, pour les ancêtres d'un individu, suit la loi binomiale  $\mathcal{B}([Nt], \theta/2N)$ . Quand  $N$  tend vers l'infini, cette loi tend vers la loi de Poisson de paramètre  $\theta t/2$ .

On note  $S(t)$  le nombre de mutations durant la période  $t$  pour la lignée de l'individu. On pose  $R_0 = 0$ . On note  $R_j$  le temps écoulé entre la  $(j-1)$ ième mutation et la  $j$ ième pour  $j \geq 2$ ,  $R_1$  étant le temps d'attente de la première mutation. Les variables aléatoires  $(R_j)_{j \geq 1}$  sont indépendantes et identiquement distribuées de loi  $\mathcal{E}(\theta/2)$ . On a, pour tout  $t \geq 0$ ,

$$S(t) = \inf \left\{ k \geq 0, \sum_{j=1}^{k+1} R_j > t \right\}.$$

**Remarque 4.6.6.** Le processus  $S$  est un processus de Poisson de paramètre  $\theta/2$ .

**Proposition 4.6.7.** Soit  $k \geq 2$ . On note  $Y_k$  le nombre de mutations observées dans une population de  $k$  individus avant le premier temps de coalescence  $T_k$  (de loi exponentielle de paramètre  $C_k^2$ ). La variable  $Y_k$  a même loi que  $G - 1$  où  $G$  suit la loi géométrique de paramètre  $p = (k-1)/(k-1+\theta)$ .

*Démonstration.* Le nombre de mutations  $Y_k$  est la somme des mutations subies par chacun des  $k$  individus :

$$Y_k = Y_k^{(1)} + \cdots + Y_k^{(k)},$$

où les variables aléatoires  $(Y_k^{(m)})_i$  sont indépendantes et identiquement distribuées de même loi que  $S(T_k)$ . Introduisons  $k$  processus de Poisson  $(S^{(m)}(t))_{1 \leq m \leq k}$  indépendants de paramètre  $\theta/2$ . On notera  $(R_j^{(m)})_j$  la suite des temps inter-sauts du processus  $S^{(m)}$ . On a donc

$$\left( Y_k^{(1)}, \dots, Y_k^{(k)} \right) \stackrel{\mathcal{L}}{=} \left( S^{(1)}(T_k), \dots, S^{(k)}(T_k) \right),$$

avec  $T_k$  indépendant des processus  $(S^{(m)})_{1 \leq m \leq k}$ .

On peut alors calculer la fonction caractéristique de  $Y_k$  en utilisant que, pour

tout  $m = 1, \dots, k$ ,  $S^{(m)}(t)$  suit la loi de Poisson de paramètre  $\theta t/2$  :

$$\begin{aligned}\mathbb{E}(e^{iuY_k}) &= \mathbb{E}\left(e^{iu\sum_{m=1}^k S^{(m)}(T_k)}\right) = \int_0^\infty C_k^2 e^{-C_k^2 t} \mathbb{E}\left(e^{iu\sum_{m=1}^k S^{(m)}(t)}\right) dt \\ &= \int_0^\infty C_k^2 e^{-C_k^2 t} \prod_{m=1}^k \mathbb{E}\left(e^{iuS^{(m)}(t)}\right) dt = \int_0^\infty C_k^2 e^{-C_k^2 t} e^{k\theta t(\exp(iu)-1)/2} dt \\ &= \frac{k(k-1)}{2} \frac{1}{k(k-1)/2 - k\theta(\exp(iu)-1)/2} = \frac{p}{1 - (1-p)e^{iu}},\end{aligned}$$

avec  $p = (k-1)/(k-1+\theta)$ . Par ailleurs, il est clair que si  $G$  suit la loi géométrique de paramètre  $p$  alors

$$\mathbb{E}(e^{iu(G-1)}) = \frac{p}{1 - (1-p)e^{iu}},$$

ce qui achève la preuve.  $\square$

Le nombre total de mutations observées sur un échantillon de  $n$  individus est donc

$$S_n \stackrel{\mathcal{L}}{=} Y_2 + \dots + Y_n$$

avec les variables aléatoires  $(Y_i)_i$  indépendantes. On a donc, d'après le lemme 4.6.7,

$$\mathbb{E}(S_n) = \sum_{k=2}^n \mathbb{E}(Y_k) = \sum_{k=2}^n \left( \frac{k-1+\theta}{k-1} - 1 \right) = \theta \sum_{k=1}^{n-1} \frac{1}{k} = \theta \ln n + O(1).$$

De même, puisque la variance d'une variable aléatoire de loi  $\mathcal{G}(p)$  est  $(1-p)/p^2$ , on a

$$\mathbb{V}(S_n) = \sum_{k=2}^n \mathbb{V}(Y_k) = \theta \sum_{k=1}^{n-1} \frac{1}{k} + \theta^2 \sum_{k=1}^{n-1} \frac{1}{k^2} = \theta \ln n + O(1).$$

On peut en déduire, comme au paragraphe précédent, le comportement de  $S_n$  pour  $n$  grand.

### Proposition 4.6.8.

$$\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathbb{V}(S_n)}} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

### 4.6.3 Estimation du taux de mutation

Nous présentons ici différentes façons d'estimer  $\theta$ .

#### Avec le nombre d'allèles

La proposition 4.6.3 permet d'écrire un intervalle de confiance asymptotique pour  $\theta$  : par exemple  $\theta$  appartient à l'intervalle

$$\left[ \frac{K_n}{\ln n} - \frac{1.96\sqrt{K_n}}{\ln n}, \frac{K_n}{\ln n} + \frac{1.96\sqrt{K_n}}{\ln n} \right]$$

avec une probabilité asymptotique 0.95. Comme  $K_n$  est d'ordre  $\ln n$ , l'amplitude de l'intervalle est de l'ordre de  $1/\sqrt{\ln n}$ . Pour diviser cette amplitude par 2, il faut observer  $n^4$  individus au lieu de  $n$ ...

### Avec le nombre de mutations

Ce résultat fournit un intervalle de confiance de confiance asymptotique 0.95 pour  $\theta$  de la forme

$$\left[ \frac{S_n}{\ln n} - \frac{1.96\sqrt{S_n}}{\ln n}, \frac{S_n}{\ln n} + \frac{1.96\sqrt{S_n}}{\ln n} \right].$$

On retrouve un résultat équivalent à la première approche...

#### 4.6.4 La formule de Ewens

On peut en fait décrire plus précisément la structure des allèles d'une population de  $n$  individus. Dans le modèle à nombre d'allèles infini, un échantillon de taille  $n$  peut être représenté par une configuration  $\mathbf{c} = (c_1, \dots, c_n)$  où  $c_i$  est le nombre d'allèles représentés  $i$  fois et  $|\mathbf{c}| = c_1 + 2c_2 + \dots + nc_n = n$ . On notera  $\mathbf{e}_i = (0, 0, \dots, 0, 1, 0, \dots, 0)$  le  $i^{\text{ème}}$  vecteur unitaire. Il s'agit à présent d'établir une équation satisfaite par les probabilités  $(q(\mathbf{c}))$  où  $q(\mathbf{c})$  est la probabilité qu'un échantillon de taille  $|\mathbf{c}|$  tiré sous la probabilité stationnaire ait la configuration  $\mathbf{c}$ . On pose  $q(\mathbf{e}_1) = 1$ .

Supposons que la configuration soit  $\mathbf{c}$ . En remontant la généalogie de l'échantillon jusqu'au premier changement, nous trouvons une mutation ou un ancêtre commun à deux individus. Le premier événement a une probabilité

$$\frac{n\theta/2}{n\theta/2 + n(n-1)/2} = \frac{\theta}{\theta + n - 1}.$$

La configuration  $\mathbf{b}$  qui a conduit à la  $\mathbf{c}$  est

- $\mathbf{b} = \mathbf{c}$  si la mutation a eu lieu sur l'un des  $c_1$  singlets (probabilité  $c_1/n$ ),
- $\mathbf{b} = \mathbf{c} - 2\mathbf{e}_1 + \mathbf{e}_2$  si la mutation a eu lieu sur un des allèles représenté 2 fois (probabilité  $2(c_2 + 1)/n$ ),
- $\mathbf{b} = \mathbf{c} - \mathbf{e}_1 - \mathbf{e}_{j-1} + \mathbf{e}_j$  si la mutation a eu lieu sur un des allèles représenté  $j$  fois (probabilité  $j(c_j + 1)/n$ )

Le second événement a lieu avec une probabilité  $(n-1)/(\theta+n-1)$  et dans ce cas, la configuration précédente était de la forme  $\mathbf{b} = \mathbf{c} + \mathbf{e}_j - \mathbf{e}_{j+1}$  : un allèle présent  $j$  fois (parmi les  $c_j + 1$ ) s'est dédoublé, ramenant de  $c_j + 1$  à  $c_j$  le nombre d'allèles présent  $j$  fois et de  $c_{j+1} - 1$  à  $c_{j+1}$  le nombre d'allèles présent  $j + 1$  fois. Cet événement a une probabilité de  $j(c_j + 1)/(n-1)$ .

En conséquence, on a

$$\begin{aligned} q(\mathbf{c}) &= \frac{\theta}{\theta + n - 1} \left[ \frac{c_1}{n} q(\mathbf{c}) + \sum_{j=2}^n \frac{j(c_j + 1)}{n} q(\mathbf{c} - \mathbf{e}_{j-1} + \mathbf{e}_j) \right] \\ &\quad + \frac{n-1}{\theta + n - 1} \left[ \sum_{j=1}^{n-1} \frac{j(c_j + 1)}{n-1} q(\mathbf{c} - \mathbf{e}_j + \mathbf{e}_{j+1}) \right], \end{aligned}$$

avec la convention  $q(\mathbf{b}) = 0$  si l'un des coefficients de  $\mathbf{b}$  est strictement négatif.

Ewens, en 1972, établit la loi de  $\mathbf{c}$ .

**Théorème 4.6.9.** *Sous la loi stationnaire, un échantillon de taille  $n$  possède la configuration  $\mathbf{c}$  avec la probabilité*

$$q(\mathbf{c}) = \mathbb{P}(C_1(n) = c_1, \dots, C_n(n) = c_n) = \mathbf{1}_{\{\|\mathbf{c}\|=n\}} \frac{n!}{\theta_{(n)}} \prod_{j=1}^n \left( \frac{\theta}{j} \right)^{c_j} \frac{1}{c_j},$$

avec la notation  $x_{(j)} = x(x+1)\cdots(x+j-1)$ .

*Démonstration.* On procède par récurrence sur  $n|\mathbf{c}|$  et  $k = \|\mathbf{c}\| := c_1 + \cdots + c_n$ .  $\square$

On peut retrouver grâce à la formule d'Ewens la loi de  $K_n$ .

*Seconde preuve de la proposition 4.6.1.* Le nombre d'allèles distincts dans l'échantillon s'écrit encore  $K_n = C_1(n) + \cdots + C_n(n)$ . Sa loi peut être obtenue à partir de la FEE :

$$\mathbb{P}(K_n = k) = \sum_{\mathbf{c}, \|\mathbf{c}\|=k} q(\mathbf{c}) = \frac{\theta^k}{\theta_{(n)}} n! \sum_{\mathbf{c}, \|\mathbf{c}\|=k} \left( \frac{1}{c_j} \right)^{c_j} \frac{1}{c_j} = \frac{\theta^k |S_n^k|}{\theta_{(n)}},$$

où  $|S_n^k|$  est le coefficient de  $x^k$  dans  $x(x+1)\cdots(x+n-1)$  (il est appelé nombre de Stirling de première espèce) et la dernière égalité est la formule de Cauchy donnant le nombre de permutations de  $n$  éléments possédant  $k$  cycles distincts.

On peut en déduire une forme plus explicite pour la loi de  $K_n$  en calculant sa fonction génératrice :

$$\begin{aligned} \mathbb{E}(s^{K_n}) &= \sum_{l=1}^n s^l \frac{\theta^l |S_n^l|}{\theta_{(n)}} = \frac{(\theta s)_{(n)}}{\theta_{(n)}} = \frac{\theta s(\theta s + 1)\cdots(\theta s + n - 1)}{\theta(\theta + 1)\cdots(\theta + n - 1)} \\ &= s \left( \frac{1}{\theta + 1} + \frac{\theta}{\theta + 1} s \right) \cdots \left( \frac{n-1}{\theta + n - 1} + \frac{\theta}{\theta + n - 1} s \right) = \prod_{j=1}^n G_{\theta/(\theta+j-1)}(s), \end{aligned}$$

où  $G_p$  est la fonction génératrice de la loi de Bernoulli de paramètre  $p$ . En d'autres termes  $K_n$  a même loi que la somme de variables aléatoires  $(\xi_j)_{1 \leq j \leq n}$  indépendantes et de loi de Bernoulli de paramètres respectifs  $(\theta/(\theta + j - 1))_{1 \leq j \leq n}$ .  $\square$

## 4.7 Le Coalescent

L'objet de cette section est de décrire la structure d'un processus à temps continu appelé coalescent de Kingman, à valeurs dans les relations d'équivalence sur l'ensemble  $[n] = \{1, 2, \dots, n\}$ , qui décrit la généalogie d'une population de taille  $n$ .

Jusqu'à présent, nous avons proposé un modèle pour l'évolution temporelle du nombre d'ancêtres d'une population initiale donnée. Il s'agit à présent d'être plus précis en gardant trace des liens de parentés entre tous les individus. Numérotions les individus de 1 à  $n$ . À un instant  $t$ , on définit la relation d'équivalence (aléatoire)  $\sim$  sur  $[n]$  par  $i \sim j$  si et seulement si les individus  $i$  et  $j$  ont le même ancêtre commun au temps  $t$ . Bien sûr, cette relation d'équivalence s'identifie à une partition de  $[n]$  (en ses classes d'équivalence). Chaque classe d'équivalence correspond à un ancêtre de la population initiale et les éléments de cette classe sont tous les descendants de cet individu. Notons  $C(t)$  cette partition (aléatoire). Quelle est la dynamique de  $(C(t))_{t \geq 0}$ ? Supposons que  $C(0) = \alpha$  pour une certaine partition  $\alpha$  de  $[n]$  et notons  $k$  le nombre de ses classes (on écrira  $|\alpha| = k$ ). Lorsque  $t$  augmente, nous progressons vers le passé et le processus reste constant jusqu'à l'apparition d'un ancêtre commun à deux des ancêtres de chaque classe. Lorsque ceci se produit, les deux ancêtres, et par conséquence tous leurs descendants, partagent cet ancêtre commun. Les deux classes d'équivalence sont donc regroupées (ou coalescent). Le taux d'apparition de cet événement est 1. Il est intuitivement clair que le processus  $C$  est markovien.

Notons  $\mathcal{E}_n$  l'ensemble des relations d'équivalence sur  $[n]$ . Le processus  $(C(t))_{t \geq 0}$  est le processus de Markov à temps continu sur  $\mathcal{E}_n$  d'état initial

$$C(0) = \Delta = \{\{1\}, \{2\}, \dots, \{n\}\},$$

(personne n'est relié à personne), et de générateur  $Q$

$$q_{\alpha\beta} = \begin{cases} -C_k^2 & \text{si } \alpha = \beta \text{ et } |\alpha| = k, \\ 1 & \text{si } \alpha \sim \beta, \\ 0 & \text{sinon,} \end{cases}$$

où la notation  $\alpha \sim \beta$  signifie que la partition  $\beta$  peut être obtenue à partir de la partition  $\alpha$  en fusionnant deux de ses classes d'équivalence.

Le processus  $C$  est appelé coalescent ou  $n$ -coalescent. Pour déterminer sa loi, il faut tout d'abord étudier la chaîne incluse  $(\mathcal{C}_k)_{k=n,n-1,\dots,1}$ . Cette chaîne est issue de  $\Delta$  et admet pour transitions :

$$\mathbb{P}(\mathcal{C}_{k-1} = \beta | \mathcal{C}_k = \alpha) = \begin{cases} C_k^2 & \text{si } \alpha \sim \beta \text{ et } |\alpha| = k, \\ 0 & \text{sinon.} \end{cases}$$

Le processus  $C$  évolue donc selon une suite  $\Delta = \mathcal{C}_n \sim \mathcal{C}_{n-1} \sim \dots \sim \mathcal{C}_1 = \Theta$ , où  $\Theta$  est la partition triviale et passe en la partition  $\mathcal{C}_k$  un temps exponentiel de paramètre  $C_k^2$ . Les taux de transition ne dépendent de la partition qu'au travers de son cardinal et

$$|C(t)| = A_n(t),$$

puisque les classes de  $C(t)$  sont en bijection avec les ancêtres au temps  $t$ . Ainsi, les processus  $(A_n(t))_{t \geq 0}$  et  $(\mathcal{C}_k)_k$  sont indépendants et

$$\forall t \geq 0, \quad C(t) = \mathcal{C}_{A_n(t)}.$$

Plus précisément,

$$\mathbb{P}(C(t) = \alpha) = \mathbb{P}(A_n(t) = |\alpha|) \mathbb{P}(\mathcal{C}_{|\alpha|} = \alpha).$$

on a déjà déterminé la loi de  $A_n$ . Celle de  $\mathcal{C}$  est donnée par le théorème suivant (dû à Kingman).

**Théorème 4.7.1.** *Soit  $1 \leq j \leq n$  et  $\alpha$  une partition de  $[n]$  dont les classes d'équivalence admettent les cardinaux  $\lambda_1, \dots, \lambda_j$ . Alors,*

$$\mathbb{P}(\mathcal{C}_j = \alpha) = \frac{(n-j)!j!(j-1)!}{n!(n-1)!} \lambda_1! \cdots \lambda_j!.$$

*Démonstration.* On procède par récurrence descendante. Le résultat est évident pour  $j = n$  (si  $\alpha$  n'est pas la partition  $\Delta$ , sa probabilité d'apparition est nulle). Supposons que le résultat soit vrai pour  $j \geq 2$ . Alors

$$\begin{aligned} p_{j-1}(\beta) := \mathbb{P}(\mathcal{C}_{j-1} = \beta) &= \sum_{\alpha \in \mathcal{E}_n} p_j(\alpha) \mathbb{P}(\mathcal{C}_{j-1} = \beta | \mathcal{C}_j = \alpha) \\ &= \sum_{\alpha \sim \beta} p_j(\alpha) \frac{2}{j(j-1)}. \end{aligned}$$

Notons  $\lambda_1, \dots, \lambda_{j-1}$  les tailles des classes d'équivalence de  $\beta$ . Celles de  $\alpha$  sont  $\lambda_1, \dots, \lambda_{l-1}, m, \lambda_l - m, \lambda_{l+1}, \dots, \lambda_{j-1}$  pour un certain  $1 \leq l \leq j-1$  et un certain  $1 \leq m \leq \lambda_l - 1$ . Grâce à l'hypothèse de récurrence,

$$\begin{aligned} p_{j-1}(\beta) &= \sum_{l=1}^{j-1} \sum_{m=1}^{\lambda_l-1} \frac{2(n-j)!j!(j-1)!}{j(j-1)n!(n-1)!} \lambda_1! \cdots \lambda_{l-1}! m! (\lambda_l - m)! \lambda_{l+1}! \cdots \lambda_{j-1}! \frac{1}{2} C_{\lambda_m}^m \\ &= \frac{(n-j)!(j-1)!(j-2)!}{n!(n-1)!} \lambda_1! \cdots \lambda_{j-1}! \sum_{l=1}^{j-1} \sum_{m=1}^{\lambda_l-1} 1 \\ &= \frac{(n-j+1)!(j-1)!(j-2)!}{n!(n-1)!} \lambda_1! \cdots \lambda_{j-1}!, \end{aligned}$$

ce qui est le résultat escompté.  $\square$



# Chapitre 5

## Processus ponctuels de Poisson

Les processus ponctuels modélisent des jets de points aléatoires sur des espaces mesurables très généraux. Le processus de Poisson usuel de comptage de tops espacés par durées indépendantes et équidistribuées de loi exponentielle apparaît comme un exemple très particulier de processus ponctuel sur  $\mathbb{R}_+$ . Nous commençons par rappeler les propriétés essentielles des lois eulériennes qui sont intimement reliées à ce processus.

### 5.1 Lois eulériennes et statistique d'ordre uniforme

#### Lois Beta

Pour tout  $a > 0$  et tout  $b > 0$ , la loi Beta( $a, b$ ) est la loi de densité

$$x \mapsto \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} I_{[0,1]}(x).$$

La loi Beta( $1, 1$ ) n'est rien d'autre que la loi uniforme  $\mathcal{U}([0, 1])$ . La loi Beta( $1/2, 1/2$ ) est connue sous le nom de *loi de l'arc-sinus*. La loi Beta est une sorte d'analogie continu de la loi binomiale.

#### Lois Gamma

Pour  $a > 0$  et  $\lambda > 0$ , la loi Gamma( $a, \lambda$ ) est la loi de densité

$$t \in \mathbb{R} \mapsto \frac{\lambda^a}{\Gamma(a)} t^{a-1} e^{-\lambda t} I_{\mathbb{R}_+}(t).$$

Par récurrence : Gamma( $a_1, \lambda$ ) \* ... \* Gamma( $a_n, \lambda$ ) = Gamma( $a_1 + \dots + a_n, \lambda$ ).

La loi Gamma( $1, \lambda$ ) n'est rien d'autre que la loi exponentielle  $\mathcal{E}(\lambda)$ . Si  $E_1, \dots, E_n$  sont indépendantes et de loi  $\mathcal{E}(\lambda)$ , alors  $\mathcal{L}(E_1 + \dots + E_n) = \mathcal{E}(\lambda)^{*n} = \text{Gamma}(n, \lambda)$ .

### Lois de Dirichlet

La loi de Dirichlet  $\mathcal{D}(a_1, \dots, a_n)$  est par définition la loi de

$$\frac{1}{G_1 + \dots + G_n} (G_1, \dots, G_n)$$

où  $G_1, \dots, G_n$  sont des variables aléatoires indépendantes avec  $G_i \sim \text{Gamma}(a_i, 1)$ . Elle est portée par le simplexe  $\{(x_1, \dots, x_n) \in \mathbb{R}_+^n; x_1 + \dots + x_n = 1\}$ , et permet de choisir une loi de probabilité discrète à  $n$  atomes. Le cas  $a_1 = \dots = a_n = 1$  correspond à la loi uniforme sur le simplexe. La loi de Dirichlet n'admet pas de densité car le simplexe est de mesure de Lebesgue nulle. Cependant, si  $(D_1, \dots, D_n)$  suit la loi  $\mathcal{D}(a_1, \dots, a_n)$ , alors le vecteur marginal  $(D_1, \dots, D_{n-1})$  admet pour densité

$$(x_1, \dots, x_{n-1}) \in \mathbb{R}_+^{n-1} \mapsto \frac{\Gamma(a_1 + \dots + a_n)}{\Gamma(a_1) \cdots \Gamma(a_n)} \prod_{i=1}^{n-1} x_i^{a_i-1} \left(1 - \sum_{i=1}^{n-1} x_i\right)^{a_n-1} I_{0 \leq x_1, \dots, x_{n-1} \leq 1}.$$

Pour tout  $1 \leq i \leq n$ , la composante  $D_i$  suit la loi  $\text{Beta}(a_i, a_1 + \dots + a_n - a_i)$ . Si  $I_1, \dots, I_k$  est une partition de  $\{1, \dots, n\}$ , alors

$$\left( \sum_{i \in I_1} D_i, \dots, \sum_{i \in I_k} D_i \right) \sim \mathcal{D} \left( \sum_{i \in I_1} a_i, \dots, \sum_{i \in I_k} a_i \right).$$

En particulier, pour toute partie non vide  $I$  de  $\{1, \dots, n\}$ , on a

$$\sum_{i \in I} D_i \sim \text{Beta} \left( \sum_{i \in I} a_i, \sum_{i \notin I} a_i \right).$$

La loi de Dirichlet est à la loi Beta ce que la loi multinomiale est à la loi binomiale.

### Lien avec la statistique d'ordre uniforme

La *statistique d'ordre uniforme* de taille  $n$  sur  $[0, t]$  est la loi du vecteur

$$(U_{(1)}, \dots, U_{(n)})$$

obtenu en réordonnant de manière croissante  $n$  variables aléatoires  $U_1, \dots, U_n$  indépendantes et de loi uniforme  $\mathcal{U}([0, t])$ . La densité de la *statistique d'ordre uniforme* de taille  $n$  sur  $[0, t]$  est donnée par

$$(t_1, \dots, t_n) \in \mathbb{R}^n \mapsto n! t^{-n} I_{0 < t_1 < \dots < t_n < t} (t_1, \dots, t_n).$$

Il est facile de le voir sans calcul en découplant le cube  $[0, t]^n$  de mesure de Lebesgue  $t^n$  en  $n!$  portions isométriques à  $\{(t_1, \dots, t_n) \in [0, t]^n; t_1 < \dots < t_n\}$ .

Soit  $(E_n)_{n>0}$  une suite de variables aléatoires indépendantes et équidistribuées de loi  $\mathcal{E}(\lambda)$  avec  $\lambda > 0$ . Posons  $T_n := E_1 + \dots + E_n$  pour tout  $n > 0$ . On a  $\mathcal{L}(E_1, \dots, E_n) = \mathcal{E}(\lambda)^{\otimes n}$ , de densité

$$(s_1, \dots, s_n) \mapsto \lambda^n e^{-\lambda(s_1 + \dots + s_n)} I_{\mathbb{R}_+^n}(s_1, \dots, s_n).$$

Le changement de variable linéaire triangulaire

$$(s_1, \dots, s_n) \mapsto (s_1, s_1 + s_2, \dots, s_1 + \dots + s_n)$$

permet d'établir que  $\mathcal{L}(T_1, \dots, T_n)$  a pour densité

$$(t_1, \dots, t_n) \in \mathbb{R}^n \mapsto \lambda^n e^{-\lambda t_n} I_{0 < t_1 < \dots < t_n}(t_1, \dots, t_n).$$

Comme  $\mathcal{L}(T_{n+1}) = \mathcal{E}(\lambda)^{(n+1)} = \text{Gamma}(n+1, \lambda)$ , il en découle par division que la loi conditionnelle  $\mathcal{L}(T_1, \dots, T_n | T_{n+1} = t)$  est la statistique d'ordre de loi uniforme de taille  $n$  sur  $[0, t]$ . Par conséquent, le vecteur aléatoire

$$\frac{1}{T_{n+1}}(T_1, \dots, T_n)$$

est indépendant de  $T_{n+1}$  et suit la statistique d'ordre uniforme de taille  $n$  sur  $[0, 1]$ . On en déduit immédiatement que

$$\mathcal{L}(U_{(1)} - U_{(0)}, \dots, U_{(n+1)} - U_{(n)}) = \mathcal{L}\left(\frac{1}{E_1 + \dots + E_{n+1}}(E_1, \dots, E_{n+1})\right).$$

avec  $U_{(0)} := 0$  et  $U_{(n+1)} := 1$ . Cette loi n'est rien d'autre que la loi de Dirichlet  $D(1, \dots, 1)$  où 1 est répété  $n+1$  fois. Ainsi, les écarts entre des réalisations i.i.d. de loi uniforme suivent des lois Beta, tandis que le vecteur de ces écarts suit une loi de Dirichlet. Dit à l'envers, si le vecteur  $(D_1, \dots, D_{n+1})$  suit la loi  $D(1, \dots, 1)$  alors  $D_1 + \dots + D_{n+1} = 1$  et  $(D_1, D_1 + D_2, \dots, D_1 + \dots + D_n)$  suit la statistique d'ordre uniforme de taille  $n$  sur  $[0, 1]$ .

## 5.2 Loi des petits nombres

La loi binomiale  $\mathcal{B}(n, p)$  est la loi de la somme de  $n$  variables aléatoires de Bernoulli indépendantes et équidistribuées de loi  $\mathcal{B}(p) := p\delta_1 + (1-p)\delta_0$ . Cela s'écrit  $\mathcal{B}(n, p) = \mathcal{B}(p)^{\ast n}$ . La moyenne et la variance de  $\mathcal{B}(n, p)$  valent respectivement  $np$  et  $np(1-p)$ . Si  $np_n \rightarrow \lambda$  lorsque  $n \rightarrow \infty$  pour un réel  $\lambda \geq 0$ , il est bien connu que  $\mathcal{B}(n, p_n) \rightarrow \mathcal{P}(\lambda)$ . Le théorème suivant généralise ce résultat aux convolutions de lois de Bernoulli sans hypothèse d'équidistribution.

**Théorème 5.2.1 (Loi des petits nombres).** Soit  $(B_{n,k})_{1 \leq k \leq n}$  un tableau triangulaire infini de variables aléatoires indépendantes à valeurs dans  $\{0, 1\}$ . Supposons que les moyennes  $(p_{n,k})_{1 \leq k \leq n}$  de ces variables aléatoires de Bernoulli vérifient les deux propriétés suivantes :

1. **stabilisation de la moyenne** :  $\lim_{n \rightarrow \infty} \sum_{k=1}^n p_{n,k} = \lambda < \infty$  ;
2. **dispersion des bits** :  $\lim_{n \rightarrow \infty} \sup_{k=1}^n p_{n,k} = 0$  :  
alors la variable aléatoire  $S_n := \sum_{k=1}^n B_{n,k}$  converge en loi vers  $\mathcal{P}(\lambda)$ .

Lorsque  $p_{n,1} = \dots = p_{n,n} =: p_n$  pour tout entier  $n > 0$ , alors  $S_n \sim \mathcal{B}(n, p_n)$  et on retrouve la condition  $np_n \rightarrow \lambda$ , qui implique  $p_n \rightarrow 0$  (qui est la seconde condition). En substance, le théorème 5.2.1 signifie que les lois de Poisson forment l'adhérence de l'ensemble des convolutions de lois de Bernoulli, pourvu que la moyenne de la convolution se stabilise et que les lois de Bernoulli se dispersent.

*Démonstration.* La fonction génératrice de  $S_n$  s'écrit, pour  $|u| < 1$ ,

$$\mathbb{E}(u^{S_n}) = \prod_{k=1}^n (1 - p_{n,k} + p_{n,k}u) = \exp\left(\sum_{k=1}^n \log(1 - p_{n,k}(1 - u))\right).$$

L'inégalité élémentaire  $-x^2 \leq x + \log(1 - x) \leq 0$  valable pour  $0 \leq x \leq 1/2$  et la seconde hypothèse fournissent, pour  $n$  assez grand :

$$\left| \sum_{k=1}^n \log(1 - p_{n,k}(1 - u)) + \sum_{k=1}^n p_{n,k}(1 - u) \right| \leq \sum_{k=1}^n (p_{n,k})^2 \leq \left( \sup_k p_{n,k} \right) \sum_{k=1}^n p_{n,k}.$$

Lorsque  $n \rightarrow \infty$ , le membre de droite de l'inégalité ci-dessus est équivalent à  $\lambda \sup_k p_{n,k}$  en vertu des hypothèses. Par conséquent,

$$\lim_{n \rightarrow \infty} \sum_{k=1}^n \log(1 - p_{n,k}(1 - u)) = \lim_{n \rightarrow \infty} - \sum_{k=1}^n p_{n,k}(1 - u) = -\lambda(1 - u),$$

et donc  $\lim_{n \rightarrow \infty} \mathbb{E}(u^{S_n}) = \exp(-\lambda(1 - u))$  comme attendu.  $\square$

**Remarque 5.2.2 (Inégalité de Le Cam).** Soit  $\lambda > 0$  un réel fixé. Pour tout entier  $n > 0$ , soit  $S_n$  une variable aléatoire somme de  $n$  variables aléatoires de Bernoulli indépendantes de paramètres respectifs  $p_1, \dots, p_n$ . Si  $p_1 + \dots + p_n = \lambda$  pour tout  $n > 0$ , alors

$$\sum_{k=1}^{\infty} \left| \mathbb{P}(S_n = k) - e^{-\lambda} \frac{\lambda^k}{k!} \right| < 2 \sum_{k=1}^n p_i^2.$$

Il s'agit de l'*inégalité de Le Cam*. Lorsque  $p_1 = \dots = p_n = \lambda/n$ , alors le membre de droite devient  $2\lambda^2/n^2$ , et on retrouve que  $\mathcal{B}(n, \lambda/n) \rightarrow \mathcal{P}(\lambda)$  quand  $n \rightarrow \infty$ .

### 5.3 Loi de Poisson et loi multinomiale

Soit  $d \geq 2$  un entier et  $p_1\delta_1 + \dots + p_d\delta_d$  une loi sur l'ensemble fini  $\{1, \dots, d\}$ . La loi multinomiale  $\mathcal{M}(n, (p_1, \dots, p_d))$  est par définition la loi sur  $\mathbb{N}^d$  donnée par ( $e_1, \dots, e_d$  désigne la base canonique de  $\mathbb{R}^d$ ) :

$$(p_1\delta_{e_1} + \dots + p_d\delta_{e_d})^{*n} = \sum_{n_1+\dots+n_d=n} \frac{n!}{n_1! \cdots n_d!} p_1^{n_1} \cdots p_d^{n_d} \delta_{(n_1, \dots, n_d)}.$$

Elle est portée par  $\{(n_1, \dots, n_d) \in \mathbb{N}^d; n_1 + \dots + n_d = n\}$ . Elle correspond aux effectifs obtenus lors de  $n$  réalisations indépendantes de loi  $p_1\delta_1 + \dots + p_d\delta_d$ . Penser par exemple à  $n$  lancers d'un dé à  $d$  faces. Cette loi est associée aux schémas de Bernoulli à  $d$  issues possibles. Pour  $d = 2$ , la loi multinomiale s'identifie à la loi binomiale d'un schéma de Bernoulli à deux issues possibles (jeu de pile ou face).

Si  $(M_1, \dots, M_d) \sim \mathcal{M}(n, (p_1, \dots, p_d))$  alors pour tout  $1 \leq i \leq d$ , la composante  $M_i$  suit la loi binomiale  $\mathcal{B}(n, p_i)$ . Les composantes ne sont pas indépendantes car reliées par  $M_1 + \dots + M_d = n$ . Si  $I_1, \dots, I_k$  est une partition de  $\{1, \dots, d\}$ , alors

$$\left( \sum_{i \in I_1} M_i, \dots, \sum_{i \in I_k} M_i \right) \sim \mathcal{M}\left( n, \left( \sum_{i \in I_1} p_i, \dots, \sum_{i \in I_k} p_i \right) \right).$$

En particulier, pour tout partie non vide  $I$  de  $\{1, \dots, n\}$ , on a

$$\sum_{i \in I} M_i \sim \mathcal{B}\left( n, \sum_{i \in I} p_i \right).$$

La loi multinomiale est à la loi binomiale ce que la loi de Dirichlet est à la loi Beta. Le théorème suivant montre que les lois multinomiales, et donc également les lois binomiales, apparaissent naturellement en conditionnant des lois de Poisson.

**Théorème 5.3.1** (Poisson et multinomiale). *Si  $P_1, \dots, P_d$  sont des variables aléatoires indépendantes de loi de Poisson de moyennes respectives  $\lambda_1 > 0, \dots, \lambda_d > 0$ , alors pour tout  $n \geq 1$ ,*

$$\mathcal{L}((P_1, \dots, P_d) \mid P_1 + \dots + P_d = n) = \mathcal{M}(n, (p_1, \dots, p_d)),$$

où  $p_i := \lambda_i / (\lambda_1 + \dots + \lambda_d)$ . Lorsque  $\lambda_1 = \dots = \lambda_d = \lambda$ , alors  $p_1 = \dots = p_d = 1/d$  et  $\mathcal{L}((P_1, \dots, P_d) \mid P_1 + \dots + P_d = n)$  ne dépend pas de  $\lambda$ .

*Démonstration.*  $\mathbb{P}(P_1 = n_1, \dots, P_d = n_d) = e^{-\lambda_1 - \dots - \lambda_d} \frac{\lambda_1^{n_1} \dots \lambda_d^{n_d}}{n_1! \dots n_d!}$  pour tout  $d$ -uplet d'entiers  $(n_1, \dots, n_d)$ . D'autre part

$$\mathbb{P}(P_1 + \dots + P_d = n) = e^{-\lambda_1 - \dots - \lambda_d} \frac{(\lambda_1 + \dots + \lambda_d)^n}{n!}$$

car  $P_1 + \dots + P_d$  suit la loi de Poisson de moyenne  $\lambda_1 + \dots + \lambda_d$  (la somme de variables de Poisson indépendantes est encore une variable de Poisson).  $\square$

## 5.4 Processus de Poisson et comptage de tops

Soit  $(E_n)_{n>0}$  une suite de variables aléatoires indépendantes de loi exponentielle  $\mathcal{E}(\lambda)$  de moyenne  $1/\lambda$ , où  $\lambda > 0$  est un nombre réel fixé. Pour tout entier  $n > 0$ , on pose  $T_n := E_1 + \dots + E_n$ . La suite  $(T_n)_{n>0}$  modélise des tops dans l'axe de temps  $\mathbb{R}_+$  espacés par des durées indépendantes et équidistribuées de loi exponentielle.

La suite  $(T_n)_{n>0}$  constitue également une manière de jeter aléatoirement des points sur  $\mathbb{R}_+$ . On note  $X$  l'ensemble aléatoire  $\{T_n; n > 0\}$ . Si  $I$  est un borélien de  $\mathbb{R}_+$ , la variable aléatoire  $\text{Card}(X \cap I)$  à valeurs dans  $\mathbb{N} \cup \{\infty\}$  représente le nombre de tops contenus dans  $I$ .

**Définition 5.4.1** (Processus de Poisson). Le processus  $(N_t)_{t \geq 0}$  défini pour tout  $t \geq 0$  par  $N_t := \text{Card}(X \cap [0, t])$  est appelé *processus de Poisson* sur  $\mathbb{R}_+$ , issu de 0, et d'intensité  $\lambda$ .

Il est clair que  $N_0 = 0$  car les variables aléatoires  $(T_n)_{n>0}$  sont indépendantes et leurs lois n'ont pas d'atomes. Presque sûrement, les trajectoires de  $(N_t)_{t \geq 0}$  sont constantes par morceaux, continues à droite avec limites à gauche. Les sauts valent tous 1, et ont lieu exactement aux instants  $(T_n)$ . On a  $\text{Card}(X \cap [0, t]) = N_t$  pour tout  $t \geq 0$ , et plus généralement, pour tout intervalle  $I$  d'extrémités  $t$  et  $t+s$  avec  $t \geq 0$  et  $s \geq 0$ , l'absence d'atomes dans la loi exponentielle donne

$$\text{Card}(X \cap I) = N_{t+s} - N_t.$$

Le théorème ci-dessous affirme en particulier que les accroissements de  $(N_t)_{t \geq 0}$  sont indépendants et stationnaires, de loi de Poisson.

**Théorème 5.4.2** (Structure par localisation). *Les deux propriétés suivantes ont lieu :*

1. *si  $I_1, \dots, I_d$  sont des intervalles bornés et disjoints de  $\mathbb{R}_+$ , de mesures de Lebesgue respectives  $|I_1|, \dots, |I_d|$ , alors les variables aléatoires*

$$\text{Card}(X \cap I_1), \dots, \text{Card}(X \cap I_d)$$

*sont indépendantes, et suivent respectivement les lois de Poisson*

$$\mathcal{P}(\lambda|I_1|), \dots, \mathcal{P}(\lambda|I_d|).$$

*De plus, pour tout entier  $n > 0$ , et conditionnellement à  $\{\text{Card}(X \cap I) = n\}$  où  $I := I_1 \cup \dots \cup I_d$ , le vecteur aléatoire*

$$(\text{Card}(X \cap I_1), \dots, \text{Card}(X \cap I_d))$$

*suit la loi multinomiale correspondant au vecteur des effectifs de  $n$  réalisations indépendantes de loi discrète  $p_1\delta_1 + \dots + p_d\delta_d$  avec  $p_i := |I_i|/|I|$ . Cette loi ne dépend pas de  $\lambda$  ;*

2. *pour tout intervalle borné et non vide  $I$  de  $\mathbb{R}_+$ , et pour tout entier  $n > 0$ , conditionnellement à  $\{\text{Card}(X \cap I) = n\}$ , les points aléatoires  $X \cap I$  de  $\mathbb{R}_+$  sont distribués comme  $n$  réalisations indépendantes de loi uniforme sur  $I$ .*

*Démonstration.* Pour la première partie du théorème, on se ramène en rajoutant des intervalles intermédiaires au cas où les intervalles  $I_1, \dots, I_d$  sont adjacents, d'extrémités  $0 := t_0 \leq \dots \leq t_d$ . Soient  $k_1, \dots, k_d$  des entiers et  $C_i := \{\text{Card}(X \cap I_i) = k_i\}$ . En notant  $m_0 := 0$  et  $m_i := k_1 + \dots + k_i$  pour tout  $1 \leq i \leq d$ , on a

$$\{\text{Card}(X \cap I_i) = k_i\} = \{t_{i-1} < T_{m_i+1} \leq t_i < T_{m_{i+1}+1}\}.$$

L'expression déjà obtenue de la densité de  $\mathcal{L}(T_1, \dots, T_{m_d+1})$  conduit par récurrence sur  $d$  à la formule

$$\mathbb{P}(\cap_{i=1}^d C_i) = \lambda^{m_d} \prod_{i=1}^d e^{-\lambda(t_i - t_{i-1})} \frac{(t_i - t_{i-1})^{k_i}}{k_i!} = \prod_{i=1}^d e^{-\lambda(t_i - t_{i-1})} \frac{\lambda^{k_i} |I_i|^{k_i}}{k_i!}.$$

Ceci montre que les composantes du vecteur  $(\text{Card}(X \cap I_1), \dots, \text{Card}(X \cap I_d))$  sont indépendantes et de lois de Poisson  $\mathcal{P}(\lambda|I_1|), \dots, \mathcal{P}(\lambda|I_d|)$ . Le résultat multinomial s'obtient ensuite directement en utilisant le théorème 5.3.1.

Pour la seconde partie du théorème. On pose  $I = [t, t+s]$  avec  $t \geq 0$  et  $s > 0$ . Pour tout  $k \in \mathbb{N}$ , l'expression de la densité de  $\mathcal{L}(T_1, \dots, T_{n+k+1})$  montre que sur

$$\{N_t = k; N_{t+s} - N_t = n\} = \{T_k \leq t < T_{k+1}; T_{n+k} \leq t+s < T_{n+k+1}\},$$

la densité de  $\mathcal{L}(T_{k+1}, \dots, T_{n+k})$  s'écrit

$$(t_{k+1}, \dots, t_{n+k}) \in \mathbb{R}^n \mapsto \frac{t^k}{k!} \lambda^{k+n} e^{-\lambda(t+s)} \mathbf{1}_{t < t_{k+1} < \dots < t_{n+k} < t+s}(t_{k+1}, \dots, t_{n+k}).$$

Ainsi, sur  $\{N_t = k; N_{t+s} - N_t = n\}$ , les  $n$  points de  $X \cap I$  ordonnés de manière croissante admettent la densité

$$(u_1, \dots, u_n) \in \mathbb{R}^n \mapsto \frac{t^k}{k!} \lambda^{k+n} e^{-\lambda(t+s)} \mathbf{1}_{t < u_1 < \dots < u_n < t+s}(u_1, \dots, u_n).$$

Ensuite, on somme sur  $k$  puis on divise par  $\mathbb{P}(N_{t+s} - N_t = n) = e^{-\lambda s} (\lambda s)^n / n!$ .  $\square$

La loi multinomiale modélise un schéma de Bernoulli de taille  $n$  à  $d$  issues possibles : penser à  $n$  jets d'un dé à  $d$  faces. Pour  $d = 2$ , on retrouve le jeu de pile ou face et la loi binomiale.

**Remarque 5.4.3 (Intensité).** La première propriété fournie par le théorème 5.4.2 indique que  $\lambda$  joue le rôle d'une *intensité*. Le nombre moyen de points aléatoires dans  $I$  vaut  $\lambda|I|$ , et croît donc linéairement avec  $\lambda$ . Plus précisément, il croît linéairement avec la masse affectée à  $I$  par la mesure  $\Lambda$  définie pour tout borélien  $A$  par  $\Lambda(A) := \lambda|A|$ . La mesure  $\Lambda$  est un multiple de la mesure de Lebesgue.

**Remarque 5.4.4 (Simulation).** La combinaison des deux propriétés fournies par le théorème 5.4.2 permet de simuler une réalisation de l'ensemble  $X \cap I$ , en simulant une réalisation de la loi de Poisson  $\mathcal{P}(\lambda|I|)$ , ce qui fournit un nombre entier  $n$ , puis  $n$  réalisations indépendantes de loi uniforme sur  $I$ . Cette approche est une alternative intéressante à la simulation de la suite  $(E_n)_{n>0}$ .

**Remarque 5.4.5 (Intervalles ou boréliens?).** Dans les résultats précédents, les intervalles peuvent être remplacés par des boréliens quelconques de mesure de Lebesgue finie. En effet, les intervalles bornés engendrent la tribu boréenne, et par conséquent, tout borélien s'obtient par des opérations ensemblistes dénombrables à partir d'intervalles.

**Remarque 5.4.6** (Point de vue markovien). Le processus  $(N_t)_{t \geq 0}$  constitue une chaîne de Markov à temps continu d'espace d'état  $\mathbb{N}$  et de générateur infinitésimal  $L$  donné pour tous  $n$  et  $m$  dans  $\mathbb{N}$  par

$$L(n, m) = \begin{cases} +\lambda & \text{si } m = n + 1 \\ -\lambda & \text{si } m = n \\ 0 & \text{sinon} \end{cases}.$$

Son action sur une fonction  $f : \mathbb{N} \rightarrow \mathbb{R}$  s'écrit, pour tout  $n \in \mathbb{N}$ ,

$$(Lf)(n) = \lambda(f(n+1) - f(n)).$$

La chaîne incluse est déterministe, de noyau  $P : \mathbb{N} \times \mathbb{N} \rightarrow [0, 1]$  donné par  $P(n, n+1) = 1$  pour tout  $n \in \mathbb{N}$ . Les mesures invariantes sont les multiples de la mesure de comptage sur  $\mathbb{N}$ , qui sont de plus symétriques. Il n'y a pas de loi invariante.

**Remarque 5.4.7** (Point de vue mesures ponctuelles aléatoires). L'ensemble aléatoire  $X := \{T_1, T_2, \dots\}$  peut être identifié à la mesure ponctuelles aléatoire :

$$\sum_{x \in X} \delta_x = \sum_{n=1}^{\infty} \delta_{T_n}.$$

Par abus de notation, on note également  $X$  cette mesure ponctuelle aléatoire. Pour tout borélien  $I$  de  $\mathbb{R}_+$ ,

$$X(I) = \int_{\mathbb{R}_+} \mathbf{1}_I(x) dX(x) = \text{Card}(X \cap I).$$

On a également

$$X(I) = \sum_{x \in X} \mathbf{1}_I(x) = \sum_{n=1}^{\infty} \mathbf{1}_I(T_n).$$

Les deux propriétés suivantes sont fondamentales :

1. presque sûrement, la mesure ponctuelle aléatoire  $X$  est *localement finie* : cela signifie que presque sûrement, pour tout compact  $K$  de  $\mathbb{R}_+$ , la variable aléatoire  $X(K)$  est finie. En effet,  $X(K) \sim \mathcal{P}(|K|)$  ;
2. presque sûrement, la mesure ponctuelle aléatoire  $X$  est *simple* : cela signifie que les atomes de  $X$  sont tous différents. En effet, les variables aléatoires  $(E_n)_{n>0}$  qui donnent les espacements des atomes sont i.i.d. et de loi commune à densité.

Ainsi, le processus de Poisson peut être vu comme une variable aléatoire à valeurs dans l'ensemble des mesures ponctuelles simples localement finies.

**Remarque 5.4.8** (Estimation de l'intensité). L'estimation de l'intensité  $\lambda$  peut être menée grâce à la loi des grands nombres (LGN) et au théorème central limite (TCL) suivants :

$$\frac{N_t}{t} \xrightarrow[t \rightarrow +\infty]{\text{p.s.}} \lambda \quad \text{et} \quad \sqrt{t} \left( \frac{N_t}{t} - \lambda \right) \xrightarrow[t \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, \lambda).$$

Le TLC s'obtient par transformée de Fourier tandis que la LGN découle de l'encadrement  $N_{\lfloor t \rfloor} \leq N_t \leq N_{\lfloor t \rfloor + 1}$  et de l'indépendance et stationnarité des accroissements. On dispose également d'une LGN et d'un TLC pour les temps de saut :

$$\frac{T_1 + \cdots + T_n}{n} \xrightarrow[n \rightarrow +\infty]{\text{p.s.}} \frac{1}{\lambda} \quad \text{et} \quad \sqrt{n} \left( \frac{T_1 + \cdots + T_n}{n} - \frac{1}{\lambda} \right) \xrightarrow[t \rightarrow +\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{\lambda^2}\right).$$

Cela fournit au passage une nouvelle preuve de la LGN pour  $(N_t)_{t \geq 0}$  car  $N_{T_n}/T_n = n/T_n$  pour tout entier  $n > 0$ .

**Remarque 5.4.9** (Processus de Poisson composé). Soit  $(N_t)_{t \geq 0}$  un processus de Poisson simple issu de 0 et d'intensité  $\lambda$ . Si  $(Z_n)_{n \in \mathbb{N}}$  est une chaîne de Markov d'espace d'état  $\mathbb{R}^d$ , indépendante de  $(N_t)_{t \geq 0}$ , le processus  $(Z_{N_t})_{t \geq 0}$  est appelé *processus de Poisson composé*. Quels sont ses temps de sauts ? Quel est son générateur ?

**Théorème 5.4.10** (Caractérisation du processus de Poisson simple). *Soit  $(N_t)_{t \geq 0}$  une famille de variables aléatoires à valeurs dans  $\mathbb{N}$  et indexée par le réel  $t \geq 0$ . Supposons que  $N_0 = 0$  et que les trajectoires  $t \mapsto N_t$  sont presque sûrement croissantes et continues à droite (donc avec limites à gauche). Alors les propriétés suivantes sont équivalentes, pour tout réel  $\lambda > 0$ .*

1. **Comptage.**  $(N_t)_{t \geq 0}$  est le processus de comptage de tops espacés par des durées indépendantes et équidistribuées de loi exponentielle  $\mathcal{E}(\lambda)$  ;
2. **Temps de saut et sauts.** Les temps inter-sauts de  $(N_t)_{t \geq 0}$  sont indépendants et équidistribués de loi exponentielle  $\mathcal{E}(\lambda)$ , et les sauts valent tous +1 ;
3. **Structure des accroissements.** Pour tous  $0 \leq t_0 \leq t_1 \leq \cdots \leq t_n$ , les accroissements

$$N_{t_n} - N_{t_{n-1}}, \dots, N_{t_1} - N_{t_0}$$

sont indépendants et de lois

$$\mathcal{P}(\lambda(t_n - t_{n-1})), \dots, \mathcal{P}(\lambda(t_1 - t_0));$$

4. **Propriété infinitésimale.** Les accroissements de  $(N_t)_{t \geq 0}$  sont indépendants, et de plus

$$\sup_{t \geq 0} \mathbb{P}(N_{t+\varepsilon} - N_t = 0) = 1 - \lambda\varepsilon + o(\varepsilon)$$

et

$$\sup_{t \geq 0} \mathbb{P}(N_{t+\varepsilon} - N_t = 1) = \lambda\varepsilon + o(\varepsilon).$$

*Démonstration.* L'équivalence des propriétés 1 et 2 est claire. Montrons que la propriété 3 se déduit de la propriété 2. Soient  $k_1, \dots, k_n$  des entiers et  $C := \{N_1 - N_{t_0} = k_1, \dots, N_{t_n} - N_{t_{n-1}} = k_n\}$ . En notant  $m_0 := 0$  et  $m_i := k_1 + \cdots + k_i$ , on a pour tout  $1 \leq i \leq n$

$$\{N_{t_i} - N_{t_{i-1}} = k_i\} = \{t_{i-1} < T_{m_i+1} \leq t_i < T_{m_{i+1}+1}\}.$$

La loi de  $\mathcal{L}(T_1, \dots, T_{m_n+1})$  donne

$$\mathbb{P}(C) = \lambda^{m_n} \prod_{i=1}^n e^{-(t_i - t_{i-1})} \frac{(t_i - t_{i-1})^{k_i}}{k_i!} = \prod_{i=1}^n e^{-\lambda(t_i - t_{i-1})} \frac{\lambda^{k_i} (t_i - t_{i-1})^{k_i}}{k_i!}.$$

Ainsi, la propriété 2 implique la propriété 3. On montre que la propriété 3 implique la propriété 4 en observant que  $\mathbb{P}(N_{t+\varepsilon} - N_t = 0) = e^{-\lambda\varepsilon}$  et  $\mathbb{P}(N_{t+\varepsilon} - N_t = 0) = e^{-\lambda\varepsilon} \lambda\varepsilon$ . Montrons que la propriété 4 implique la propriété 3. On a  $\sup_{t \geq 0} \mathbb{P}(N_{t+\varepsilon} - N_t = m) = o(\varepsilon)$  pour tout  $m \geq 2$ . Par suite, en notant  $p_k(t) := \mathbb{P}(N_t = k)$ , on obtient pour tout  $m \geq 1$ ,  $p_m(t + \varepsilon) = \sum_{n=0}^m \mathbb{P}(N_{t+\varepsilon} - N_t = n) p_{m-n}(t)$ , et donc

$$\sup_{t \geq 0} \left( \frac{p_m(t + \varepsilon) - p_m(t)}{\varepsilon} + \lambda p_m(t) - \lambda p_{m-1}(t) \right) = O(\varepsilon).$$

En exprimant cette propriété pour  $t - \varepsilon$ , on obtient

$$\frac{p_m(t) - p_m(t - \varepsilon)}{\varepsilon} = -\lambda p_m(t - \varepsilon) + \lambda p_{m-1}(t - \varepsilon) + O(\varepsilon).$$

La fonction  $t \mapsto p_m(t)$  est donc continue et même dérivable, et vérifie l'équation différentielle  $p'_m(t) = -\lambda p_m(t) + \lambda p_{m-1}(t)$ . De la même manière, on montre que  $p'_0(t) = -\lambda p_0(t)$ . Les conditions initiales de ces équations sont imposées par  $N_0 = 0$ . Par récurrence sur  $m$ , on obtient  $p_m(t) = e^{-\lambda t} (\lambda t)^m / m!$ , c'est-à-dire que  $N_t$  suit la loi de Poisson  $\mathcal{P}(\lambda t)$ . Comme  $(N_{s+t} - N_s)_{t \geq 0}$  vérifie également la propriété 4, on obtient bien la propriété 3. Montrons enfin que la propriété 3 implique la propriété 1. La propriété 3 spécifie la loi des trajectoires, et il existe donc un unique processus vérifiant la propriété 3, qui est donc forcément le processus de Poisson simple issu de 0 et d'intensité  $\lambda$  construit par la propriété 1.  $\square$

**Corollaire 5.4.11** (Stabilité par superposition). *Si  $(N_t)_{t \geq 0}$  et  $(N'_t)_{t \geq 0}$  sont deux processus de Poisson simples indépendants, issus de 0 et d'intensités respectives  $\lambda$  et  $\lambda'$ , alors  $(N_t + N'_t)_{t \geq 0}$  est un processus de Poisson simple issu de 0 de d'intensité  $\lambda + \lambda'$ .*

Imaginons que chaque top d'un processus de Poisson soit muni d'une valeur 0 ou 1 de manière indépendante, avec probabilité  $p$ . Alors les tops marqués 0 et les tops marqués 1 constituent deux processus de Poisson indépendants dont les intensités s'additionnent pour donner l'intensité de départ. C'est en substance ce qu'affirme le théorème suivant.

**Théorème 5.4.12** (Propriété de stabilité par marquage et amincissement). *Soit  $(N_t)_{t \geq 0}$  un processus de Poisson simple issu de 0 et d'intensité  $\lambda$ , et de temps de sauts  $(T_n)_{n > 0}$ . Soit  $(Y_n)_{n > 0}$  une suite de variables aléatoires indépendantes de loi de Bernoulli  $\mathcal{B}(p)$ , indépendante de  $(T_n)_{n > 0}$ . Alors les processus de comptage des suites amincies  $\{T_n; t.q. Y_n = 1\}$  et  $\{T_n; t.q. Y_n = 0\}$  sont des processus de Poisson simples indépendants issus de 0 et d'intensités respectives  $p\lambda$  et  $q\lambda$  où  $q := 1 - p$ .*

*Démonstration.* Les processus de comptage  $U := (U_t)_{t \geq 0}$  et  $V := (V_t)_{t \geq 0}$  des deux suites amincies sont donnés par  $U_t := \sum_{k=1}^{N_t} Y_k$  et  $V_t := \sum_{k=1}^{N_t} (1 - Y_k)$ . La transformée de Laplace du couple  $(U_t, V_t)$  se calcule explicitement. En effet, comme

$$B_n := Y_1 + \cdots + Y_n \sim \mathcal{B}(n, p),$$

on a

$$\begin{aligned} \mathbb{E}(e^{\alpha U + \beta V}) &= \mathbb{E}(e^{\beta N_t + (\alpha - \beta) B_{N_t}}) \\ &= \mathbb{E}(\mathbb{E}(e^{\beta N_t} e^{(\alpha - \beta) B_{N_t}} | N_t)) \\ &= \mathbb{E}\left(e^{\beta N_t} (pe^{\alpha - \beta} + q)^{N_t}\right) \\ &= e^{\lambda t} e^{\beta(pe^{\alpha - \beta} + q)} \\ &= e^{p\lambda t(e^\alpha - 1)} e^{q\lambda(e^\beta - 1)} \\ &= \mathbb{E}(e^{\alpha U_t}) \mathbb{E}(e^{\beta V_t}), \end{aligned}$$

où la dernière égalité s'obtient en considérant les cas particuliers  $\alpha = 0$  et  $\beta = 0$ . Ceci montre que  $U_t$  et  $V_t$  sont indépendantes, de lois  $\mathcal{P}(p\lambda t)$  et  $\mathcal{P}(q\lambda t)$ . La même méthode montre que  $U$  et  $V$  sont à accroissements stationnaires, indépendants entre eux et de manière croisée, et de loi de Poisson. Le résultat attendu découle alors du théorème de caractérisation des processus de Poisson simples par la structure des accroissements.  $\square$

**Remarque 5.4.13 (Temps de sauts des processus amincis).** Considérons les données du théorème ci-dessus. Si  $B_n := Y_1 + \cdots + Y_n$  alors  $(B_n)_{n \geq 0}$  est un processus de Bernoulli de paramètre  $p$ , indépendant de  $(N_t)_{t \geq 0}$ . Ses temps inter-sauts sont géométriques et indépendants des temps inter-sauts exponentiels de  $(N_t)_{t \geq 0}$ . Les temps inter-sauts des deux processus amincis sont par conséquent des sommes géométriques de lois Gamma (cela correspond à une stabilité des lois Gamma par mélange géométrique). Le théorème ci-dessus montre que ces temps sont encore exponentiels. Les deux processus amincis constituent des exemples *processus de Poisson composés*, associés à un processus de Bernoulli.

Le processus de Poisson sur  $\mathbb{R}_+$  constitue le processus de comptage de tops espacés par des durées aléatoires indépendantes et de même loi exponentielle. Ce processus constitue également une manière de jeter des points aléatoirement sur l'axe des temps  $\mathbb{R}_+$ . Cette interprétation en terme de processus ponctuel montre que l'orientation de l'axe des temps  $\mathbb{R}_+$  ne joue aucun rôle fondamental. Cela conduit naturellement à la notion de processus ponctuels génériques abordée dans la suite du chapitre.

## 5.5 Processus ponctuels

Cette section est largement inspirée des premières pages du livre [Rob00].

Nous remplaçons à présent  $\mathbb{R}_+$  par un *bon espace*  $E$ . Dans ce chapitre, *bon espace* signifie espace métrique localement compact pour lequel il existe un recouvrement dénombrable par des compacts. C'est le cas de  $\mathbb{R}^d$  et des variétés usuelles. Dans tout ce qui va suivre, on peut en cas de difficulté se contenter de prendre  $E = \mathbb{R}^d$ , ou encore plus simplement  $E = \mathbb{R}_+$ . On munit  $E$  de sa tribu borélienne  $\mathcal{B}(E)$ .

Une mesure de Borel positive  $\mu$  sur  $E$  est *localement finie* lorsque  $\mu(K) < \infty$  pour tout compact  $K$  de  $E$ . Une mesure de Borel positive  $\mu$  sur  $E$  est *ponctuelle* lorsqu'elle s'écrit  $\sum_n \delta_{x_n}$  pour une famille au plus dénombrable  $(x_n)$  de  $E$ .

Une mesure ponctuelle est localement finie si et seulement si elle possède un nombre fini d'atomes dans chaque compact de  $E$ . On note  $\mathbf{A}(E)$  l'ensemble des mesures ponctuelles localement finies.

On muni  $\mathbf{A}(E)$  de la topologie de la convergence faible des mesures : une suite  $(\mu_n)$  de  $\mathbf{A}(E)$  converge vers  $\mu$  dans  $\mathbf{A}(E)$  si seulement si

$$\int_E f \, d\mu_n \xrightarrow{n \rightarrow +\infty} \int_E f \, d\mu$$

pour toute fonction  $f : E \rightarrow \mathbb{R}_+$  continue positive et à support compact. La tribu borélienne de cette topologie est notée  $\mathcal{A}(E)$ . Elle rend mesurable les applications

$$\mu \in \mathbf{A}(E) \mapsto \mu(A), \quad \text{pour tout } A \in \mathcal{B}(E).$$

**Définition 5.5.1** (Processus ponctuels). Un *processus ponctuel*  $X$  sur  $E$  est une variable aléatoire à valeurs dans  $\mathbf{A}(E)$ .

### 5.5.1 Loi et intensité d'un processus ponctuel

La loi d'un processus ponctuel  $X$  sur  $E$  est une loi sur  $(\mathbf{A}(E), \mathcal{A}(E))$ . Une réalisation de  $X$  correspond à un élément de  $\mathbf{A}(E)$ , c'est-à-dire à une mesure ponctuelle localement finie sur  $E$ . Le processus ponctuel  $X$  modélise le jet aléatoire de points dans  $E$ .

**Théorème 5.5.2** (Caractérisation de la loi d'un processus ponctuel). *La loi d'un processus ponctuel  $X$  sur  $E$  est caractérisée par chacune des manières suivantes :*

1. *en spécifiant la loi de la variable aléatoire réelle*

$$\int_E f \, dX = \sum_{x \in X} f(x)$$

*pour toute fonction  $f : E \rightarrow \mathbb{R}$  continue à support compact. Notons que dans le membre de gauche, la notation  $X$  désigne une mesure ponctuelle aléatoire tandis que dans le membre de droite, elle désigne les atomes (aléatoires) de cette mesure. Cette double notation est utilisée très souvent par la suite ;*

2. en spécifiant la loi de la variable aléatoire réelle

$$\int_E f dX$$

pour toute fonction  $f : E \rightarrow \mathbb{R}$  étagée, c'est-à-dire de la forme  $\sum_{i=1}^n \alpha_i I_{A_i}$  où  $n > 0$  est un entier,  $\alpha_1, \dots, \alpha_n$  sont des réels et  $A_1, \dots, A_n$  sont disjoints et dans  $\mathcal{B}(E)$  ;

3. en spécifiant la loi du vecteur aléatoire de  $\mathbb{N}^n$

$$(X(A_1), \dots, X(A_n))$$

pour tout entier  $n > 0$ , et tous  $A_1, \dots, A_n$  disjoints et dans  $\mathcal{B}(E)$  ;

4. en spécifiant la transformée de Laplace de  $X$ , c'est-à-dire la fonctionnelle

$$f \mapsto \mathbb{E}\left(\exp\left(-\int_E f dX\right)\right)$$

où  $f : E \rightarrow \mathbb{R}_+$  est mesurable positive.

Démonstration. Exercice ! □

**Définition 5.5.3** (Mesure d'intensité d'un Processus ponctuel). La mesure d'intensité  $\Lambda$  d'un processus ponctuel  $X$  sur  $E$  est la mesure sur  $E$  définie par

$$\Lambda(A) := \mathbb{E}(X(A)) \quad \text{pour tout } A \in \mathcal{B}(E).$$

**Remarque 5.5.4** (Contre-exemple). Si  $T$  est une variable aléatoire à valeurs dans  $\mathbb{N}$ , non intégrable, alors l'ensemble aléatoire  $\{1, 1/2, \dots, 1/T\}$  constitue un processus ponctuel sur  $\mathbb{R}_+$  dont la mesure d'intensité est infinie sur tout compact de la forme  $[0, t]$  avec  $t > 0$ . Ainsi, la mesure d'intensité d'un processus ponctuel n'est pas forcément localement finie.

**Théorème 5.5.5** (Formule de Campbell). Si  $X$  est un processus ponctuel sur  $E$  de mesure d'intensité  $\Lambda$ , alors pour toute fonction mesurable positive  $f : E \rightarrow \mathbb{R}_+$ ,

$$\mathbb{E}\left(\int_E f(x) dX(x)\right) = \int_E f(x) d\Lambda(x).$$

Démonstration. On se ramène à  $f$  étagée par convergence monotone, puis à  $f = I_A$  avec  $A \in \mathcal{B}(E)$  par linéarité. Dans ce dernier cas, l'identité souhaitée revient à la définition de la mesure d'intensité. □

**Définition 5.5.6** (Restriction). Soit  $X$  un processus ponctuel sur  $E$ , de mesure d'intensité  $\Lambda$ . Soit également  $A \in \mathcal{B}(E)$  tel que  $0 < \Lambda(A) < \infty$ . La restriction de  $X$  à  $A$ , notée  $X_A$ , est obtenue en sélectionnant les atomes de  $X$  qui appartiennent à  $A$ .

**Théorème 5.5.7** (Localisation). Soit  $X$  un processus ponctuel sur  $E$ , de mesure d'intensité  $\Lambda$ . Soit également  $A \in \mathcal{B}(E)$  tel que  $0 < \Lambda(A) < \infty$ . Alors  $X_A$  est un processus ponctuel sur  $E$  de mesure d'intensité  $\Lambda(A \cap \cdot)$ , de densité  $I_A$  par rapport à  $\Lambda$ .

Démonstration. Découle directement des définitions de  $\Lambda$  et  $X_A$  ! □

### 5.5.2 Processus ponctuels simples

Une mesure ponctuelle est *simple* lorsque ses atomes sont tous différents. On note  $\mathbf{S}(E)$  le sous-ensemble de  $\mathbf{A}(E)$  constitué par les mesures ponctuelles simples localement finies. Si  $\mu \in \mathbf{A}(E)$ , alors  $\mu \in \mathbf{S}(E)$  si et seulement si  $\mu(\{a\}) \in \{0, 1\}$  pour tout  $a \in E$ . Si  $\mu \in \mathbf{S}(E)$  et  $A \in \mathcal{B}(E)$ , alors le nombre entier  $\mu(A)$  représente le nombre d'atomes de  $\mu$  contenus dans  $A$ . Cela est faux si  $\mu \in \mathbf{A}(E) \setminus \mathbf{S}(E)$ .

Un sous-ensemble de  $E$  est *localement fini* lorsque son intersection avec tout compact de  $E$  est finie. Comme  $E$  est réunion au plus dénombrable de compacts, un sous-ensemble localement fini de  $E$  est toujours au plus dénombrable. La classe des sous-ensembles localement finis de  $E$  est en bijection avec  $\mathbf{S}(E)$ . Ainsi, un élément  $\mu$  de  $\mathbf{S}(E)$  s'identifie à un ensemble localement fini de  $E$ , et on a  $\mu(A) = \text{Card}(\mu \cap A)$  pour tout  $A \in \mathcal{B}(E)$ .

**Définition 5.5.8** (Processus ponctuels simples). Un *processus ponctuel*  $X$  sur  $E$  est *simple* lorsqu'il prend presque sûrement ses valeurs dans  $\mathbf{S}(E)$ . En d'autres termes, lorsque  $X(\{a\}) \in \{0, 1\}$  avec probabilité 1 pour tout  $a \in E$ .

Pour un processus ponctuel simple  $X$  sur  $E$  de mesure d'intensité  $\Lambda$ , et pour tout  $A \in \mathcal{B}(E)$ , la quantité  $\Lambda(A) = \mathbb{E}(X(A))$  représente le nombre moyens d'atomes de  $X$  contenu dans  $A$ . Cela justifie a posteriori le terme « intensité » utilisé pour  $\Lambda$ .

## 5.6 Processus ponctuels de Poisson

Un processus ponctuel de Poisson possède des propriétés similaires à celles du processus de comptage de tops espacés par des durées aléatoires indépendantes et équidistribuées de loi exponentielle.

**Définition 5.6.1** (Processus ponctuels de Poisson). Un processus ponctuel  $X$  sur  $E$  de mesure d'intensité  $\Lambda$  est un *processus ponctuel de Poisson* lorsque les propriétés suivantes sont vérifiées :

1. la mesure d'intensité  $\Lambda$  est localement finie ;
2. si  $A_1, \dots, A_d \in \mathcal{B}(E)$  sont disjoints, alors  $X(A_1), \dots, X(A_d)$  sont des variables aléatoires indépendantes ;
3. si  $A \in \mathcal{B}(E)$  vérifie  $\Lambda(A) < \infty$  alors la variable aléatoire  $X(A)$  suit la loi de Poisson  $\mathcal{P}(\Lambda(A))$ .

En particulier, pour tout compact  $K$ , la variable aléatoire  $X(K)$  est finie presque sûrement et suit la loi de Poisson  $\mathcal{P}(\Lambda(K))$ .

**Exemple 5.6.2** (Une famille de processus ponctuels de Poisson sur  $\mathbb{R}_+$ ). Considérons une suite  $(E_n)_{n>0}$  de variables aléatoires indépendantes et de même loi exponentielle  $\mathcal{E}(\lambda)$ . Posons  $X := \{T_1, T_2, \dots\}$  où  $T_n := E_1 + \dots + E_n$  pour tout  $n > 0$ . On peut également écrire  $X := \sum_{n=1}^{\infty} \delta_{T_n}$ . En vertu du théorème 5.4.2,  $X$  constitue un processus ponctuel de Poisson sur  $E = \mathbb{R}_+$  dont la mesure d'intensité  $\Lambda$  n'est rien d'autre qu'un multiple de la mesure de Lebesgue :  $\Lambda(A) := \lambda|A|$  pour tout  $A \in \mathcal{B}(E)$ .

**Théorème 5.6.3** (Transformée de Laplace et caractérisation). *Soit  $X$  un processus ponctuel sur  $E$  de mesure d'intensité  $\Lambda$ . Alors  $X$  est de Poisson si et seulement si pour toute fonction mesurable positive  $f : E \rightarrow \mathbb{R}_+$ , la transformée de Laplace de  $X$  en  $f$  est donnée par*

$$\mathbb{E}\left(\exp\left(-\int_E f dX\right)\right) = \exp\left(-\int_E (1 - e^{-f}) d\Lambda\right).$$

*En particulier, la loi d'un processus ponctuel de Poisson est caractérisée par sa mesure d'intensité. De plus, le processus ponctuel de Poisson  $X$  est simple si et seulement si la mesure  $\Lambda$  est diffuse (i.e. n'a pas d'atomes).*

*Démonstration.* Pour l'expression de la transformée de Laplace, il suffit de considérer  $f$  étagée de la forme  $f = \sum_{i=1}^d \alpha_i I_{A_i}$  avec  $d > 0$  entier,  $A_1, \dots, A_d$  disjoints et dans  $\mathcal{B}(E)$ , et  $\alpha_1, \dots, \alpha_d$  dans  $\mathbb{R}$ . Dans ce cas, on a en vertu de la propriété d'indépendance sur les ensembles disjoints :

$$\mathbb{E}\left(\exp\left(-\int_E f dX\right)\right) = \mathbb{E}\left(\exp\left(-\sum_{i=1}^d \alpha_i X(A_i)\right)\right) = \prod_{i=1}^d \mathbb{E}(\exp(-\alpha_i X(A_i))).$$

Le résultat découle alors du fait que  $X(A_i) \sim \mathcal{P}(\Lambda(A_i))$ .

Pour la caractérisation de la simplicité, on constate tout d'abord que si la mesure  $\Lambda$  admet une masse  $\Lambda(\{x\}) > 0$  en  $x \in E$ , la variable aléatoire  $X(\{x\})$  suit la loi de Poisson  $\mathcal{P}(\Lambda(\{x\}))$  et en particulier,  $\mathbb{P}(X(\{x\}) = 2) > 0$ , ce qui entraîne que  $X$  n'est pas simple. Réciproquement, supposons que  $\Lambda$  est diffuse, et soit  $K$  un sous-ensemble compact de  $E$ . Comme  $\Lambda$  est diffuse, pour tout  $\varepsilon > 0$ , il existe une partition finie de  $K$  en boréliens  $A_1, \dots, A_n$  tel que  $\mu(A_i) \leq \varepsilon$  pour tout  $1 \leq i \leq n$ , cf. par exemple [Rob03, Lemma 1.3]. D'autre part, si  $N$  est une variable aléatoire de Poisson de moyenne  $\lambda$ , alors

$$\mathbb{P}(N \geq 2) = e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^k}{k!} = \lambda^2 \sum_{k=0}^{\infty} \frac{\lambda^k}{(k+2)!} \leq \lambda^2.$$

Par conséquent, en utilisant ces deux propriétés,

$$\begin{aligned} \mathbb{P}(X \text{ n'est pas simple sur } K) &\leq \sum_{i=1}^n \mathbb{P}(X(A_i) \geq 2) \\ &\leq \sum_{i=1}^n \Lambda(A_i)^2 \\ &\leq \varepsilon \sum_{i=1}^n \Lambda(A_i) = \varepsilon \Lambda(K). \end{aligned}$$

En faisant tendre  $\varepsilon$  vers 0, on obtient  $\mathbb{P}(X \text{ n'est pas simple sur } K) = 0$ . Il ne reste plus qu'à considérer une suite de compacts qui recouvre le bon espace  $E$  pour conclure par continuité supérieure que  $X$  est simple.  $\square$

Le théorème suivant montre que l'hypothèse de diffusion sur la mesure d'intensité permet d'alléger la définition des processus ponctuels de Poisson simples. L'idée consiste à découper un borélien en tout petits morceaux, sur lequel le processus ponctuel fait apparaître des variables de Bernoulli, indépendantes, puis à utiliser la loi des petits nombres pour faire apparaître la loi de Poisson.

**Théorème 5.6.4 (Intensité diffuse et simplicité).** *Soit  $X$  un processus ponctuel simple sur  $E$  de mesure d'intensité  $\Lambda$  diffuse localement finie, vérifiant la propriété suivante :*

1. *si  $A_1, \dots, A_d \in \mathcal{B}(E)$  sont disjoints, alors  $X(A_1), \dots, X(A_d)$  sont des variables aléatoires indépendantes.*

*Alors  $X$  est un processus ponctuel de Poisson simple.*

*Démonstration.* Comme  $\Lambda$  est diffuse, pour tout sous-ensemble compact  $K$  de  $E$ , et pour tout entier  $n > 0$ , il existe une partition finie  $A_{n,1}, \dots, A_{n,k_n}$  de  $K$  en sous-ensembles relativement compacts tels que pour tout  $1 \leq i \leq k_n$ ,

$$\Lambda(A_{n,i}) < \frac{1}{n} \quad \text{et} \quad A_{n,i} \subset \overline{B}(x_{n,i}, r_{n,i}) \quad \text{avec} \quad r_{n,i} < \frac{1}{n}.$$

Cette propriété est démontrée dans [Rob03, Lemma 1.3]. Cet argument peut être répété pour chacun des  $\overline{A}_{n,i}$ . On obtient ainsi par récurrence une suite de partitions emboîtées,  $(A_{n+1,j})_{j \in I_{n,i}}$  étant une partition de  $A_{n,i}$ . La variable aléatoire  $X(K)$  se décompose alors de la manière suivante :

$$X(K) = \sum_i X(A_{n,i}) = \sum_i I_{\{X(A_{n,i})=1\}} + \sum_i X(A_{n,i})I_{\{X(A_{n,i}) \geq 2\}} = S_n + S'_n.$$

Nous allons montrer que  $S'_n$  est négligeable quand  $n \rightarrow \infty$ . On écrit en utilisant la partition de la partition

$$X(A_{n,i})I_{\{X(A_{n,i}) \geq 2\}} = I_{\{X(A_{n,i}) \geq 2\}} \sum_{j \in I_{n,i}} X(A_{n+1,j}) \geq \sum_{j \in I_{n,i}} X(A_{n+1,j})I_{\{X(A_{n+1,j}) \geq 2\}}.$$

Ainsi, on montre que  $S'_n$  décroît avec  $n$ . Sa limite quand  $n \rightarrow \infty$  est nécessairement nulle car sinon, avec probabilité non nulle, il existerait une suite  $(i_n)$  telle que  $A_{n+1,i_{n+1}} \subset A_{n,i_n}$  et  $X(A_{n,i_n}) \geq 2$  pour tout  $n$ . Comme les  $A_{n,i_n}$  sont inclus dans des boules dont le rayon tend vers 0 quand  $n \rightarrow \infty$ , la limite des  $A_{n,i_n}$  est soit vide soit réduite à un point. La masse de chacun de ces points pour  $X$  doit être au moins 2, ce qui contredit le fait que  $X$  est simple. Il est donc établi que presque sûrement,  $(S'_n)$  tend vers 0. Comme  $S'_n$  est bornée par la variable aléatoire intégrable  $X(K)$ , le théorème de convergence dominée entraîne que  $(S'_n)$  converge dans  $L^1$  vers 0. Par conséquent,

$$\lim_{n \rightarrow \infty} \mathbb{E}(S_n) = \Lambda(K).$$

Or  $\mathbb{E}(I_{\{X(A_{n,i})=1\}}) \leq \mathbb{E}(X(A_{n,i})) = \Lambda(A_{n,i}) < 1/n$ , et donc les variables aléatoires de Bernoulli  $I_{\{A_{n,i}=1\}}$  vérifient les hypothèses de la loi des petits nombres (théorème 5.2.1). Par conséquent,  $(S_n)$  converge en loi vers  $\mathcal{P}(\Lambda(K))$ . Ainsi,  $X(K) \sim \mathcal{P}(\Lambda(K))$ .

Soit à présent un ensemble  $A \in \mathcal{B}(E)$  de mesure finie pour  $\Lambda$ . Comme  $E$  est un bon espace, il existe une suite croissante  $(K_n)$  de sous-ensembles compacts inclus dans  $A$  tels que  $\Lambda(A) = \lim_{n \rightarrow \infty} \Lambda(K_n)$ . Or  $X(K_n) \sim \mathcal{P}(\Lambda(K_n))$  et donc  $X(K_n)$  converge en loi vers  $\mathcal{P}(\Lambda(A))$ . D'autre part,  $\mathbb{E}(|X(A) - X(K_n)|) = \mathbb{E}(X(A \setminus K_n))$  converge vers 0, et donc  $X(A) - X(K_n)$  converge vers 0 dans  $L^1$ . Il en découle que  $X(A)$  suit la loi  $\mathcal{P}(\Lambda(A))$ .  $\square$

**Remarque 5.6.5** (Contre-exemples). Le théorème ci-dessus est faux si  $\Lambda$  n'est pas diffuse. En effet, le processus ponctuel trivial sur  $\mathbb{R}$  constant et égal à  $\delta_0$  est simple et vérifie clairement la propriété d'indépendance mais n'est pas un processus ponctuel de Poisson. D'autre part, le théorème est également faux lorsque  $\Lambda$  est diffuse mais  $X$  n'est pas simple. En effet, si par exemple  $\sum_n \delta_{x_n}$  est un processus ponctuel de Poisson simple d'intensité  $\Lambda/2$  (leur existence est prouvée plus loin) alors le processus ponctuel  $\sum_n n\delta_{x_n}$  a une intensité  $\Lambda$ , vérifie la propriété d'indépendance car  $X$  la vérifie, mais n'est pas un processus ponctuel de Poisson !

**Théorème 5.6.6.** *Soit  $X$  un processus ponctuel de Poisson sur  $E$  de mesure d'intensité  $\Lambda$ . Si  $A \in \mathcal{B}(E)$  vérifie  $\Lambda(A) = \infty$ , alors la variable aléatoire  $X(A)$  est presque sûrement infinie.*

*Démonstration.* Comme  $E$  est un bon espace, il existe une suite croissante de sous-ensembles compacts  $(K_n)$  inclus dans  $A$  tels que  $\lim_n \Lambda(K_n) = \Lambda(A) = \infty$ . Pour tout entier  $n$ , la variable aléatoire  $X(K_n) \sim \mathcal{P}(\Lambda(K_n))$  et donc  $\mathbb{P}(X(K_n) \leq r)$  converge vers 0 quand  $n \rightarrow \infty$  pour tout réel  $r \geq 0$ . Or  $\mathbb{P}(X(A) \leq r) \leq \mathbb{P}(X(K_n) \leq r)$  car  $K_n \subset A$ . Le résultat désiré s'en déduit.  $\square$

### 5.6.1 Stabilité par superposition et restriction

La somme de variables aléatoires de Poisson indépendantes est encore une variable aléatoire de Poisson. Le résultat suivant généralise cette propriété aux processus ponctuels de Poisson.

**Théorème 5.6.7** (Stabilité par superposition). *Si  $(X_n)$  est une suite finie ou infinie dénombrable de processus ponctuels de Poisson  $E$ , indépendants, de mesures d'intensité  $(\Lambda_n)$ , et si la mesure somme  $\Lambda := \sum_n \Lambda_n$  est localement finie, alors la somme  $\sum_n X_n$  est un processus ponctuel de Poisson sur  $E$  de mesure d'intensité  $\Lambda$ .*

*Démonstration.* Découle de la caractérisation de la loi des processus ponctuels de Poisson par leur transformée de Laplace.  $\square$

Le théorème suivant fournit une sorte de réciproque.

**Théorème 5.6.8** (Indépendance par restrictions). *Soit  $X$  un processus ponctuel de Poisson sur  $E$  de mesure d'intensité  $\Lambda$ . Si  $A_1, \dots, A_d$  sont des éléments deux à deux disjoints de  $\mathcal{B}(E)$ , alors les processus restreints*

$$X_{A_1}, \dots, X_{A_d}$$

constituent des processus ponctuels de Poisson indépendants sur  $E$  de mesures d'intensité respectives

$$\Lambda(A_1 \cap \cdot), \dots, \Lambda(A_d \cap \cdot).$$

*Démonstration.* Pour tous  $B_1 \subset A_1, \dots, B_d \subset A_d$  dans  $\mathcal{B}(E)$ , les variables aléatoires

$$X_{A_1}(B_1), \dots, X_{A_d}(B_d)$$

s'écrivent également  $X(B_1), \dots, X(B_d)$ . Elles sont donc indépendantes et suivent des lois de Poisson

$$\mathcal{P}(\Lambda(B_1)), \dots, \mathcal{P}(\Lambda(B_d)).$$

Ceci montre que  $X_{A_1}, \dots, X_{A_d}$  sont des processus ponctuels de Poisson sur  $E$  de mesures d'intensité  $\Lambda(A_1 \cap \cdot), \dots, \Lambda(A_d \cap \cdot)$ . D'autre part, les variables aléatoires

$$\int_E f_1 dX_{A_1}, \dots, \int_E f_d dX_{A_d}$$

sont indépendantes pour toutes fonctions étagées  $f_1, \dots, f_d : E \rightarrow \mathbb{R}_+$ , et ceci montre que  $X_{A_1}, \dots, X_{A_d}$  sont des processus ponctuels indépendants.  $\square$

Les amateurs de partition réinterpréteront les propriétés de stabilité par superposition et restriction des processus ponctuels de Poisson.

### 5.6.2 Construction

Nous savons que les processus ponctuels de Poisson existent sur  $\mathbb{R}_+$  avec une mesure d'intensité proportionnelle à la mesure de Lebesgue. Le théorème suivant généralise considérablement ce résultat.

**Théorème 5.6.9** (Théorème d'existence constructif). *Toute mesure  $\Lambda$  sur  $E$  localement finie est la mesure d'intensité d'un processus ponctuel de Poisson sur  $E$ .*

*Démonstration.* Soit  $A \in \mathcal{B}(E)$  vérifiant  $0 < \Lambda(A) < \infty$ . On note  $\Lambda_A$  la loi de probabilité  $\Lambda(A \cap \cdot)/\Lambda(A)$ . Soit  $(Z_n)$  une suite de variable aléatoires indépendantes et de même loi  $\Lambda_A$ , et  $N$  une variable aléatoire de loi de Poisson  $\mathcal{P}(\Lambda(A))$ , indépendante de  $(Z_n)$ . Posons à présent  $X_A := \delta_{Z_1} + \dots + \delta_{Z_N}$ , et calculons sa transformée de Laplace pour une fonction mesurable positive  $f : E \rightarrow \mathbb{R}_+$  :

$$\begin{aligned} \mathbb{E}\left(\exp\left(-\int_E f dX_A\right)\right) &= \mathbb{E}\left(\mathbb{E}\left(\exp\left(-\int_E f dX_A\right) \mid N\right)\right) \\ &= \mathbb{E}\left(\mathbb{E}\left(\exp\left(-\sum_{i=1}^n f(Z_i)\right) \mid N\right)\right). \end{aligned}$$

Les hypothèses d'indépendance et de loi sur  $(Z_n)$  et  $N$  permettent donc d'écrire

$$\begin{aligned}\mathbb{E}\left(\exp\left(-\int_E f dX_A\right)\right) &= \mathbb{E}\left(\left(\int_E e^{-f} d\Lambda_A\right)^N \middle| N\right) \\ &= e^{-\Lambda(A)} \sum_{n=0}^{\infty} \frac{\Lambda(A)^n}{n!} \left(\int_E e^{-f} d\Lambda_A\right)^n \\ &= \exp\left(-\int_A (1 - e^{-f}) d\Lambda\right).\end{aligned}$$

Ainsi,  $X_A$  est un processus ponctuel de Poisson de mesure d'intensité  $\Lambda(A \cap \cdot)$ . Comme  $E$  est un bon espace et  $\Lambda$  localement finie, on construit une partition au plus dénombrable  $(A_n)$  de  $E$  avec des éléments de  $\mathcal{B}(E)$  qui vérifient  $0 < \Lambda(A_n) < \infty$  pour tout  $n$ . La méthode précédente permet de construire une suite  $(Z_n)$  de processus ponctuels de Poisson sur  $E$ , indépendants, de mesures d'intensité  $(\Lambda(A_n \cap \cdot))$ . Leur somme est un processus ponctuel de Poisson de mesure d'intensité  $\sum_n \Lambda(A_n \cap \cdot) = \Lambda$  en vertu de la propriété de stabilité par superposition.  $\square$

**Corollaire 5.6.10** (Structure par localisation). *Soit  $X$  un processus ponctuel de Poisson sur  $E$ , de mesure d'intensité  $\Lambda$ . Soit également  $A \in \mathcal{B}(E)$  tel que  $0 < \Lambda(A) < \infty$ . Alors la restriction  $X_A$  de  $X$  à  $A$  vérifie les deux propriétés suivantes :*

1.  *$X_A$  est un processus ponctuel de Poisson de mesure d'intensité  $\Lambda(A \cap \cdot)$  ;*
2. *pour tout entier  $n > 0$ , et conditionnellement à  $\{X(A) = n\}$ , les atomes de  $X_A$  ont une loi échangeable, identique à la loi de  $n$  variables aléatoires indépendantes de même loi  $\Lambda(A \cap \cdot)/\Lambda(A)$ .*

*Démonstration.* L'usage de la transformée de Laplace montre que  $X_A$  est un processus ponctuel de Poisson, de mesure d'intensité  $\Lambda(A \cap \cdot)$ . D'après la preuve du théorème précédent, ce processus a la même loi que le processus  $\delta_{Z_1} + \dots + \delta_{Z_N}$  où  $(Z_n)$  est une suite de variables aléatoires indépendantes et équidistribuées de loi  $\Lambda(A \cap \cdot)/\Lambda$  et où  $N$  est une variable aléatoire de Poisson de moyenne  $\Lambda(A)$  indépendante de  $(Z_n)$ . Par conséquent, la loi de  $X_A$  sachant  $\{X(A) = n\}$  est égale à la loi de  $\delta_{Z_1} + \dots + \delta_{Z_N}$  sachant  $\{N = n\}$ , qui est exactement la loi produit  $(\Lambda(A \cap \cdot)/\Lambda(A))^{\otimes n}$ .  $\square$

Dans la preuve du théorème ci-dessus,  $X = X_A + X_{A^c}$ , et cela correspond à la superposition des processus indépendants  $X_A$  et  $X_{A^c}$ .

**Remarque 5.6.11** (Simulation et construction). Si  $X$  est un processus ponctuel de Poisson sur  $E$  de mesure d'intensité  $\Lambda$ , et si  $A \in \mathbf{B}(E)$  est de mesure finie pour  $\Lambda$ , la simulation de  $X_A$  se réduit à la simulation de la loi de Poisson  $\mathcal{P}(\Lambda(A))$ , ce qui fournit un entier  $n$ , puis à la simulation de  $n$  réalisations indépendantes de loi  $\Lambda(A \cap \cdot)/\Lambda(A)$ . Ainsi, il suffit de savoir simuler  $\Lambda$  pour simuler  $X_A$ . La simulation de  $X$  s'obtient ensuite par superposition à partir d'une partition de  $E$  en ensembles mesurables de masse finie pour  $\Lambda$ .

Le théorème suivant découle simplement de la définition et du théorème 5.3.1.

**Théorème 5.6.12** (Structure par localisation et conditionnement). *Soit  $X$  un processus ponctuel de Poisson sur  $E$  de mesure d'intensité  $\Lambda$ . Si  $A_1, \dots, A_d$  sont des éléments deux à deux disjoints de  $\mathcal{B}(E)$ , de mesure finie pour  $\Lambda$ , alors les composantes du vecteur aléatoire*

$$(X(A_1), \dots, X(A_d))$$

*sont indépendantes et suivent des loi de Poisson*

$$\mathcal{P}(\Lambda(A_1)), \dots, \mathcal{P}(\Lambda(A_d)).$$

*De plus, pour tout entier  $n > 0$ , et conditionnellement à  $\{X(A_1) + \dots + X(A_d) = n\}$ , ce vecteur aléatoire suit la loi multinomiale correspondant au vecteurs des effectifs de  $n$  réalisations indépendantes de loi discrète*

$$p_1\delta_1 + \dots + p_d\delta_d$$

*avec  $p_i := \Lambda(A_i)/\Lambda(A)$  et  $A := A_1 \cup \dots \cup A_d$ . Cette loi est invariante par multiplication de  $\Lambda$  par une constante.*

**Exercice 5.6.13** (Estimation de la mesure d'intensité par histogramme). Proposer une méthode d'estimation « histogramme » de la mesure d'intensité d'un processus ponctuel de Poisson  $X$  sur  $E$ , à partir d'une réalisation de  $X$  et d'une partition fixée de  $E$  (joue le rôle des classes).

**Exercice 5.6.14** (Estimation de la mesure d'intensité par lissage). Proposer une méthode d'estimation de la mesure d'intensité d'un processus ponctuel de Poisson  $X$  sur  $E$  par lissage (convolution) d'une réalisation de  $X$  avec une loi de probabilité fixée. On pourra penser aux estimateurs à noyau. Cette approche inclue-t-elle l'estimateur par histogramme ?

**Remarque 5.6.15** (Mesures d'intensité qui sont des lois de probabilité). Soit un  $X$  processus ponctuel de Poisson de mesure d'intensité  $\Lambda$  sur  $E$ . Lorsque  $\Lambda$  est une loi de probabilité, alors les atomes de  $X$  sont des réalisations i.i.d. de loi  $\Lambda$ . Réciproquement, si  $\Lambda$  est une loi de probabilité, et si  $X_1, \dots, X_n$  est une suite de variables aléatoires équidistribuées de loi  $\Lambda$ , alors la mesure ponctuelle aléatoire  $\sum_{k=1}^n \delta_{X_k}$  suit la même loi qu'un processus ponctuel de Poisson sur  $E$  de mesure d'intensité  $\Lambda$  conditionnellement à  $\{X(E) = n\}$ . Grâce à cette correspondance, l'estimation d'une mesure d'intensité finie est équivalente à l'estimation d'une densité de probabilité. Il est utile d'exprimer la vraisemblance associée à une réalisation de  $X$ . Bien entendu, la localisation permet de se ramener à ce cas lorsque la mesure d'intensité n'est pas finie.

### 5.6.3 Transformations, marquage, et amincissement

Les résultats suivants permettent de construire des processus ponctuels de Poisson par transformation, marquage, et amincissement. Ces stabilités généralisent aux

processus ponctuels de Poisson ce qui a déjà été établi pour le processus de Poisson simple de comptage de tops espacés par des durées aléatoires i.i.d. de loi exponentielle.

**Théorème 5.6.16** (Stabilité par transformation). *Soit  $X$  un processus ponctuel de Poisson sur  $E$  de mesure d'intensité  $\Lambda$ . Soit  $T : E \rightarrow F$  est une application mesurable vers un bon espace  $F$ . Si la mesure image de  $\Lambda$  par  $T$  est localement finie, alors la mesure ponctuelle aléatoire  $Y := \sum_{x \in X} \delta_{T(x)}$ , notée également  $T(X)$ , est un processus ponctuel de Poisson sur  $F$  dont la mesure d'intensité est la mesure image de  $\Lambda$  par l'application  $T$ .*

*Démonstration.* Il suffit de calculer la transformée de Laplace de  $Y$ . Pour toute fonction  $f : F \rightarrow \mathbb{R}_+$  mesurable positive,

$$\begin{aligned}\mathbb{E}\left(\exp\left(-\int_F f dY\right)\right) &= \mathbb{E}\left(\exp\left(-\sum_{x \in X} f(T(x))\right)\right) \\ &= \mathbb{E}\left(\exp\left(-\int_E (f \circ T) dX\right)\right) \\ &= \exp\left(-\int_E (1 - e^{-f \circ T}) d\Lambda\right).\end{aligned}$$

□

Notons que si  $T$  n'est pas injective sur le support de  $\Lambda$ , alors  $Y$  n'est pas simple, même si  $X$  l'est. Les transformations les plus utiles sont injectives.

**Remarque 5.6.17** (Homogénéité et isotropie). Un processus ponctuel  $X$  sur  $\mathbb{R}^d$  est *stationnaire* ou *homogène* lorsque  $X$  et  $X + t$  ont la même loi pour tout  $t \in \mathbb{R}^d$ . Il s'agit d'une invariance par l'action du groupe des translations. Ici,  $X + t$  désigne la mesure ponctuelle aléatoire obtenue en translatant les atomes de  $X$  par  $t$ . Le processus ponctuel  $X$  est *isotrope* lorsque  $X$  et  $RX$  ont la même loi pour toute rotation  $R \in \text{SO}(\mathbb{R}^n)$ . Il s'agit d'une invariance par l'action du groupe des rotations. Ici,  $RX$  désigne la mesure ponctuelle aléatoire obtenue en considérant l'image des atomes de  $X$  par la rotation  $R$ . Un processus ponctuel de Poisson sur  $\mathbb{R}^d$  est homogène si et seulement si sa mesure d'intensité est proportionnelle à la mesure de Lebesgue sur  $\mathbb{R}^d$ . Il est alors de plus isotrope. Le multiple en question est parfois qualifié d'*intensité* du processus de Poisson homogène.

Nous savons qu'un processus ponctuel de Poisson sur  $\mathbb{R}_+$  de mesure d'intensité proportionnelle à la mesure de Lebesgue est stable par marquage et amincissement indépendant de Bernoulli. Le résultat suivant généralise considérablement cette propriété de stabilité.

**Théorème 5.6.18** (Stabilité par marquage dépendant et amincissement). *Soit  $X$  un processus ponctuel de Poisson sur  $E$  de mesure d'intensité  $\Lambda$ . Conditionnellement à  $X$ , on associe à chaque atome  $x$  de  $X$  une variable aléatoire  $Y_x$  à valeurs dans un bon espace  $F$ , de sorte que ces variables aléatoires soient conditionnellement*

indépendantes sachant  $X$ . Dans ce cas, les propriétés suivantes ont lieu, en notant  $\nu_x$  la loi de  $Y_x$  :

1. la mesure ponctuelle aléatoire  $Z := \sum_{x \in X} \delta_{(x, Y_x)}$  est un processus ponctuel de Poisson sur  $E \times F$  de mesure d'intensité  $\Theta$  donnée pour toute fonction  $f : E \times F \rightarrow \mathbb{R}$  mesurable positive par

$$\int_{E \times F} f(x, y) d\Theta(x, y) := \int_E \left( \int_F f(x, y) d\nu_x(y) \right) d\Lambda(x);$$

2. pour tous  $A$  et  $B$  disjoints dans  $\mathcal{B}(F)$ , les mesures ponctuelles aléatoires

$$R_A := \sum_{x \in X; Y_x \in A} \delta_x \quad \text{et} \quad R_B := \sum_{x \in X; Y_x \in B} \delta_x$$

constituent des processus ponctuels de Poisson sur  $E$ , indépendants, dont les mesures d'intensité sont absolument continues par rapport à  $\Lambda$ , de densités respectives  $x \mapsto \nu_x(A) = \mathbb{P}(Y_x \in A)$  et  $x \mapsto \nu_x(B) = \mathbb{P}(Y_x \in B)$ .

*Démonstration.* Pour établir la première propriété, on exprime la transformée de Laplace de  $Z$  pour une fonction  $f : E \times F \rightarrow \mathbb{R}$  mesurable positive quelconque :

$$\begin{aligned} \mathbb{E}\left(\exp\left(-\int_{E \times F} f dZ\right)\right) &= \mathbb{E}\left(\mathbb{E}\left(\exp\left(-\int_{E \times F} f dZ\right) \mid X\right)\right) \\ &= \mathbb{E}\left(\mathbb{E}\left(\exp\left(-\sum_{x \in X} f(x, Y_x)\right) \mid X\right)\right) \\ &= \mathbb{E}\left(\mathbb{E}\left(\prod_{x \in X} e^{-f(x, Y_x)} \mid X\right)\right) \\ &= \mathbb{E}\left(\prod_{x \in X} \mathbb{E}(e^{-f(x, Y_x)})\right) \\ &= \mathbb{E}\left(\exp\left(\sum_{x \in X} g(x)\right)\right) \end{aligned}$$

où

$$g(x) := \log(\mathbb{E}(e^{-f(x, Y_x)})) = \log \int_F e^{-f(x, y)} d\nu_x(y).$$

Comme  $X$  est un processus ponctuel de Poisson sur  $E$  d'intensité  $\Lambda$ , on a

$$\mathbb{E}\left(\exp\left(\sum_{x \in X} g(x)\right)\right) = \exp\left(-\int_E (1 - e^{g(x)}) d\Lambda(x)\right).$$

Par conséquent, on obtient

$$\begin{aligned}\mathbb{E}\left(\exp\left(-\int_{E \times F} f dZ\right)\right) &= \exp\left(-\int_E \left(1 - \int_F e^{-f(x,y)} d\nu_x(y)\right) d\Lambda(x)\right) \\ &= \exp\left(-\int_E \left(\int_F (1 - e^{-f(x,y)}) d\nu_x(y)\right) d\Lambda(x)\right) \\ &= \exp\left(-\int_E \left(\int_F (1 - e^{-f(x,y)}) d\Theta(x,y)\right)\right).\end{aligned}$$

Ceci montre que  $Z$  est bien un processus ponctuel de Poisson sur  $E \times F$  de mesure d'intensité  $\Theta$ , comme attendu. La mesure  $\Theta$  est localement finie car  $\Lambda$  est localement finie et  $\nu_x$  est une loi de probabilité pour tout  $x$ .

Pour la seconde propriété, on constate tout d'abord, en vertu de la propriété de stabilité par restriction, que les restrictions du processus  $Z$  aux ensembles *disjoints*  $E \times A$  et  $E \times B$  constituent des processus ponctuels de Poisson sur  $E \times F$ , *indépendants*, de mesures d'intensité  $I_{E \times A}\Theta$  et  $I_{E \times B}\Theta$ . Maintenant, en vertu de la stabilité par transformation, leurs images  $R_A$  et  $R_B$  par la projection  $\pi : (x, y) \in E \times F \mapsto x \in E$  constituent des processus ponctuels de Poisson sur  $E$ , indépendants, dont les mesures d'intensité sont données par les mesures image de  $I_{E \times A}\Theta$  et  $I_{E \times B}\Theta$  par la transformation  $\pi$ , qui ne sont rien d'autre que les mesures absolument continues par rapport à  $\Lambda$  de densités respectives  $x \mapsto \nu_x(A)$  et  $x \mapsto \nu_x(B)$  (ce sont des lois marginales!).  $\square$

Lorsque  $B = A^c$ , on retrouve par la stabilité par superposition le processus  $X$  en superposant les deux processus amincis  $R_A$  et  $R_{A^c}$  (ils sont indépendants). Plus généralement, le théorème 5.6.18 ci-dessus fournit une sorte de partition des processus de Poisson à partir de partition sur l'espace des marques. De point de vue de la simulation, cela revient à exprimer la loi d'un processus de Poisson comme un mélange de processus marqués.

**Remarque 5.6.19** (Marquage de Bernoulli). L'exemple le plus classique correspond au cas où  $F = \{0, 1\}$ . La loi  $\nu_x$  est alors une loi Bernoulli  $\mathcal{B}(p(x)) = p(x)\delta_1 + (1 - p(x))\delta_0$  avec  $p(x) \in [0, 1]$ . Les seuls processus amincis non triviaux correspondent aux choix  $A = \{0\}$  et  $B = \{1\}$ . Lorsque  $x \mapsto p(x)$  est constante, le marquage ne dépend pas de la position spatiale des atomes, et on retrouve ainsi la propriété de stabilité par marquage et amincissement des processus de Poisson de comptage des tops sur  $\mathbb{R}_+$ .

**Remarque 5.6.20** (Le marquage comme coloration). Les marques de Bernoulli correspondent à  $F = \{0, 1\}$ , qui peuvent s'interpréter comme la coloration en blanc ou noir de chacun des atomes du processus ponctuel sous-jacent. Plus généralement, lorsque  $F = \{0, 1, \dots, d\}$ , le marquage correspond à la coloration des atomes en  $d$  couleurs possibles. Plus généralement encore, lorsque  $F = [0, 1]$ , le marquage correspond à la coloration des atomes en une couleur dont la fréquence appartient à l'intervalle  $[0, 1]$ . La distribution de Poisson-Dirichlet et la formule d'Ewens peuvent s'interpréter aux moyen de processus avec marquage continu, voir par exemple [Kin93, chap. 9].

Lorsque le marquage ne dépend pas de la position spatiale des atomes, on obtient le résultat suivant comme cas particulier.

**Corollaire 5.6.21** (Stabilité par marquage indépendant et amincissement). *Soit  $X$  un processus ponctuel de Poisson sur  $E$  de mesure d'intensité  $\Lambda$ . Conditionnellement à  $X$ , on associe à chaque atome  $x$  de  $X$  une variable aléatoire  $Y_x$  sur un bon espace  $F$ , de sorte que ces variables aléatoires soient indépendantes et équidistribuées de loi  $\nu$ . Alors les propriétés suivantes ont lieu :*

1. la mesure ponctuelle aléatoire  $\sum_{x \in X} \delta_{(x, Y_x)}$  est un processus ponctuel de Poisson sur  $E \times F$ , de mesure d'intensité  $\Lambda \otimes \nu$  ;
2. pour tout  $A \in \mathcal{B}(E)$ , la mesure ponctuelle aléatoire  $\sum_{x \in X; Y_x \in A} \delta_x$  est un processus ponctuel de Poisson sur  $E$  de mesure d'intensité  $\nu(A)\Lambda$  ;
3. si  $A, B \in \mathcal{B}(E)$  sont disjoints, alors les processus ponctuels  $\sum_{x \in X; Y_x \in A} \delta_x$  et  $\sum_{x \in X; Y_x \in B} \delta_x$  sont indépendants.

**Remarque 5.6.22** (Vocabulaire). Amincissement se dit « thinning » en anglais.

**Exemple 5.6.23** (File d'attente M/M/ $\infty$ ). Soit  $\lambda$  et  $\mu$  deux réels strictement positifs. Considérons le cadre du corollaire 5.6.21 avec  $E = F = \mathbb{R}_+$  et  $d\Lambda(x) = \lambda I_{\mathbb{R}_+}(x)dx$  où  $dx$  désigne la mesure de Lebesgue sur  $\mathbb{R}$ , et  $\nu = \mathcal{E}(\mu)$ . Les atomes de  $X$  correspondent à des tops sur  $\mathbb{R}_+$  espacés par des durées aléatoires indépendantes de loi  $\mathcal{E}(\lambda)$ . La variable aléatoire  $Y_x$  correspond à la durée de service de l'atome  $x$ . La quantité  $x + Y_x$  correspond au temps de fin de service de l'atome  $x$ . En vertu du corollaire 5.6.21, la mesure ponctuelle aléatoire  $Z := \sum_{x \in X} \delta_{(x, Y_x)}$  constitue un processus ponctuel de Poisson d'intensité  $\Lambda \otimes \nu$ . Soit à présent  $t \in \mathbb{R}_+$  et

$$A_t := \{(u, v) \in \mathbb{R}_+^2; (u, u+v) \in [0, t] \times [t, \infty[\}.$$

La variable aléatoire  $Z(A_t)$  représente le nombre d'atomes en cours de service au temps  $t$ . C'est exactement le nombre de clients en cours de service à l'instant  $t$  dans une file d'attente M/M/ $\infty$  issue de 0, d'intensité d'arrivée  $\lambda$  et d'intensité de service  $\mu$ . La variable aléatoire  $Z(A_t)$  suit la loi de Poisson  $\mathcal{P}((\Lambda \otimes \nu)(A_t))$  de moyenne

$$(\Lambda \otimes \nu)(A_t) = \int_{0 \leqslant u \leqslant t} \int_{u+v \geqslant t} \lambda I_{\mathbb{R}_+}(u) \frac{1}{\mu} e^{-\mu v} I_{\mathbb{R}_+}(v) dudv = \rho(1 - e^{-\mu t}),$$

avec  $\rho := \lambda/\mu$ . On retrouve ainsi la formule  $\mathcal{P}(\rho(1 - e^{-\mu}))$  pour la loi à l'instant  $t \in \mathbb{R}_+$  de la file d'attente M/M/ $\infty$  issue de 0.

**Corollaire 5.6.24** (Stabilité par transformation aléatoire). *Soit  $X$  un processus ponctuel de Poisson sur  $E$  de mesure d'intensité  $\Lambda$ . Conditionnellement à  $X$ , on associe à chaque atome  $x$  de  $X$  une variable aléatoire  $Y_x$  sur un bon espace  $F$  de loi  $\nu_x$  de sorte que ces variables aléatoires soient indépendantes. Alors la mesure ponctuelle aléatoire  $\sum_{x \in X} \delta_{Y_x}$  est un processus ponctuel de Poisson sur  $F$  de mesure d'intensité donnée pour toute fonction mesurable positive  $f : F \rightarrow \mathbb{R}_+$  par*

$$\int_E \int_F f(y) d\nu_x(y) d\Lambda(x).$$

*Démonstration.* Considérer le processus marqué  $\sum_{x \in X} \delta_{(X, Y_x)}$  sur  $E \times F$  puis la transformation  $T : (x, y) \in E \times F \mapsto y$ . La mesure d'intensité du processus  $\sum_{x \in X} \delta_{Y_x}$  est localement finie car  $\nu_x$  est une loi et  $\Lambda$  est localement finie.  $\square$

Lorsque  $E = F$ , le corollaire ci-dessus peut être vu comme un résultat de stabilité des processus ponctuels de Poisson par transformation des atomes au moyen d'un noyau de transition markovien  $\mathbf{P}$  donné par  $\mathbf{P}(x, \cdot) = \nu_x$ .

Supposons que  $E = F = \mathbb{R}^d$ , et que la loi de  $Y_x - x$  ne dépend pas de  $x$ . En la notant  $\nu$ , on a  $\nu_x = \delta_{-x} * \nu$  pour tout  $x$ , de sorte que la mesure d'intensité résultante s'écrit  $\Lambda * \nu$ . Cela conduit au résultat suivant.

**Corollaire 5.6.25** (Stabilité par translation aléatoire). *Soit  $X$  un processus ponctuel de Poisson sur  $E = \mathbb{R}^d$  de mesure d'intensité  $\Lambda$ . Conditionnellement à  $X$ , on associe à chaque atome  $x$  de  $X$  une variable aléatoire  $Y_x$  de  $E$  de sorte que ces variables aléatoires soient indépendantes et équidistribuées de loi  $\nu$ . Alors la mesure ponctuelle aléatoire  $\sum_{x \in X} \delta_{x+Y_x}$  constitue un processus ponctuel de Poisson de mesure d'intensité  $\Lambda * \nu$ .*

**Remarque 5.6.26** (Cas homogène et invariance). Notons que si  $\Lambda$  est un multiple de la mesure de Lebesgue sur  $\mathbb{R}^d$ , alors  $\Lambda * \nu = \Lambda$ . Cela correspond à une invariance des processus ponctuels de Poisson homogènes sur  $\mathbb{R}^d$  par translation aléatoire.

**Remarque 5.6.27** (Observation bruitée). La loi  $\nu$  peut représenter la loi d'un bruit d'observation additif pour l'atome  $x$ . Au lieu d'observer  $x$ , on observe  $x + Y_x$ . Le cas typique est donné par un bruit gaussien :  $\nu = \mathcal{N}(0, \sigma(x)^2)$ . La mesure d'intensité apparente  $\Lambda * \nu$  est obtenue par « lissage » de la mesure d'intensité réelle. Penser au « bluring » : le lissage rend flou.

## 5.7 Processus ponctuels de Poisson sur $\mathbb{R}_+$

Soit  $(E_n)_{n > 0}$  une suite de variables aléatoires indépendantes et loi exponentielle  $\mathcal{E}(1)$ , et soit  $T_n := E_1 + \dots + E_n$ . La mesure ponctuelle aléatoire  $\sum_{n > 0} \delta_{T_n}$  constitue un processus ponctuel de Poisson sur  $\mathbb{R}_+$  dont la mesure d'intensité est la mesure de Lebesgue sur  $\mathbb{R}_+$ .

Soit  $\lambda : \mathbb{R}_+ \rightarrow \mathbb{R}_+^*$  une fonction localement Lebesgue intégrable strictement positive. La fonction  $F : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  définie par  $F(t) := \int_0^t \lambda(u) du$  est strictement croissante, et son inverse  $F^{-1} : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  vérifie  $(F^{-1})' = 1/\lambda \circ F^{-1}$ .

En vertu de la stabilité par transformation des processus ponctuels de Poisson, la mesure ponctuelle aléatoire  $\sum_{n > 0} \delta_{F^{-1}(T_n)}$  constitue un processus ponctuel de Poisson sur  $\mathbb{R}_+$  dont la mesure d'intensité est l'image par  $F^{-1}$  de la mesure de Lebesgue sur  $\mathbb{R}_+$ . Un petit calcul montre que cette mesure d'intensité est absolument continue par rapport à la mesure de Lebesgue sur  $\mathbb{R}_+$ , de densité de Radon-Nikodym  $\lambda$ .

Notons que  $\{F^{-1}(T_n) > t\} = \{T_n > F(t)\} = \{T_n > \int_0^t \lambda(u) du\}$ , pour tout  $n > 0$  et tout  $t \in \mathbb{R}_+$ . Le processus de comptage de la suite  $(F^{-1}(T_n))$ , associé

à la mesure ponctuelle aléatoire  $\sum_{n>0} \delta_{F^{-1}(T_n)}$ , constitue un processus de Markov sur  $\mathbb{N}$ , inhomogène lorsque  $\lambda$  n'est pas constante. Si  $(N_t)_{t \geq 0}$  désigne le processus de comptage de la suite  $(T_n)$ , alors  $(N_{F(t)})_{t \geq 0}$  est le processus de comptage de la suite  $(F^{-1}(T_n))$ . En quelque sorte, le processus inhomogène  $(N_{F(t)})_{t \geq 0}$  s'obtient à partir du processus homogène en distordant le temps par la fonction d'intensité  $\lambda$ .

Lorsque  $\lambda$  est constante, alors  $F(t) = \lambda t$  et  $F^{-1}(t) = t/\lambda$ . Par conséquent, dans ce cas,  $F^{-1}(T_n) = T_n/\lambda = E_1/\lambda + \dots + E_n/\lambda$ , et on retrouve le processus de Poisson d'intensité  $\lambda$  associé au comptage de tops espacés par des durées indépendantes et de même loi  $\mathcal{E}(\lambda)$ .

## 5.8 Quelques exemples classiques

Soit  $X$  un processus ponctuel de Poisson sur  $\mathbb{R}^d$  de mesure d'intensité  $\Lambda$ . Lorsque  $\Lambda$  est absolument continue par rapport à la mesure de Lebesgue sur  $\mathbb{R}^d$ , de densité de Radon-Nikodym  $\lambda : \mathbb{R} \rightarrow \mathbb{R}_+$ , alors on dit par abus de langage que  $X$  est un processus ponctuel de Poisson sur  $\mathbb{R}^d$  d'intensité  $\lambda$ . En particulier, pour tout borélien  $A$  de mesure de Lebesgue finie,

$$X(A) \sim \mathcal{P}\left(\int_A \lambda(x) dx\right).$$

Lorsque  $d = 1$  et  $A$  est un intervalle  $[s, t]$ , cela donne

$$X([s, t]) \sim \mathcal{P}\left(\int_s^t \lambda(u) du\right).$$

Toute fonction localement Lebesgue intégrable est l'intensité d'un processus ponctuel de Poisson. Lorsque  $\Lambda$  est un multiple de la mesure de Lebesgue, il existe une constante  $\lambda > 0$  telle que  $\Lambda(A) = \lambda|A|$  pour tout borélien  $A$ , où  $|A|$  désigne la mesure de Lebesgue de  $A$ . les deux formules ci-dessus s'écrivent alors

$$X(A) \sim \mathcal{P}(\lambda|A|) \quad \text{et} \quad X([s, t]) \sim \mathcal{P}((t - s)\lambda).$$

On trouvera quelques exemples (trafic routier et texture des filtres) dans [Bou00, pages 110-114]. Voici quelques autres exemples classiques de processus ponctuels de Poisson.

- trajectoires des processus de Markov : la suite des excursions d'une chaîne de Markov récurrente forme par exemple un processus de Poisson sur l'espace des boucles. La théorie des excursions des processus de Markov a été développée suite aux travaux d'Itô dans les années 1970, voir par exemple [RY99, Chapitre XII] ;
- géométrie stochastique : il est possible de définir des processus de Poisson sur l'ensemble des parties convexes compactes non-vides de  $\mathbb{R}^d$  muni de la métrique de Hausdorff ;

- géométrie stochastique : idem avec l’ensemble des hyperplans de  $\mathbb{R}^d$  munis de la métrique habituelle ;
- processus auto-exitatifs (Hawkes-Oakes) consistent à mélanger arbres de Galton-Watson sous-critiques et processus ponctuels FIXME: tremblement de terres, amas de galaxies, écarts de gènes dans l’ADN, etc.

**Remarque 5.8.1** (Estimation de l’intensité). L’estimation de l’intensité des processus de Poisson peut être menée de différentes manières. Les approches par histogramme et plus généralement par lissage sont sans doute les plus naturelles. Les livres [Kut98] [Rip88], et [MW04] traitent de certains aspects statistiques. Le contrôle adaptatif des classes de localisation dans l’estimation par histogramme n’est pas facile, comme le montrent par exemple les articles de recherche récents [RB03, RB06] et [BB06].

**Remarque 5.8.2.** Les notions suivantes n’ont pas été abordées : processus de Cox (ponctuels doublement stochastiques : la mesure d’intensité est aléatoire), mesure de Palm des processus de Poisson stationnaires. On trouve ces développements dans les livres [Kin93], [Rob03], [SKM87], [Jac06], et [vL00].



# Chapitre 6

## Modèles compartimentaux

### 6.1 Introduction

Un *modèle compartmental* est avant tout constitué de boîtes appelées *compartiments*. Chaque compartiment contient une quantité de matière. Cette quantité peut varier au cours du temps, selon trois mécanismes :

- **transfert** de matière entre compartments (la matière reste dans le système) ;
- **création** de matière (la matière nouvelle provient de l’extérieur du système) ;
- **destruction** de matière (la matière perdue quitte le système) ;

En l’absence des deux derniers mécanismes, le système est dit *conservatif*. La description précise de la variation de matière au cours du temps dans chacun des compartments constitue la *dynamique du modèle*. Les modèles compartimentaux sont très utilisés en sciences appliquées, comme en biologie, en écologie, en médecine, etc. Voici trois exemples simples :

1. **pharmacologie.** Chaque compartiment est un organe particulier d’un organisme. La quantité de matière correspond à une quantité de médicament ;
2. **cancérologie.** Chaque compartiment correspond à un stade de maturation cellulaire. La quantité de matière correspond à un nombre de cellules ;
3. **épidémiologie.** Chaque compartiment correspond à une zone géographique ou un état de la maladie. La quantité de matière correspond à une taille de population.

Le livre [MK00] propose de nombreux exemples concrets. Les modèles compartimentaux possèdent des connections naturelles avec les modèles de dynamiques de populations, ainsi qu’avec les systèmes de particules en interactions (réseaux de Petri) dont certains aspects sont présentés par exemple dans [Yca02].

### 6.2 Modèles déterministes linéaires

Considérons un système comportant un nombre fini de compartments, indexés par un ensemble  $I$ . La quantité de matière contenue dans le compartiment  $i \in I$  au

temps  $t$  est représentée par un nombre réel positif ou nul noté  $Q_i(t)$ . On note  $Q(t)$  le vecteur  $i \in I \mapsto Q_i(t)$ , et on identifie  $\mathbb{R}^I$  à  $\mathbb{R}^{|I|}$ . On considère ensuite une dynamique linéaire pour  $(Q(t))_{t \geq t_0}$ , décrite par l'ÉDO linéaire suivante :

$$\partial_t Q_i(t) = \lambda_i(t) + \sum_j \rho_{j,i}(t) Q_j(t) - Q_i(t) \left[ \kappa_i(t) + \sum_{j \neq i} \rho_{i,j}(t) \right], \quad (6.1)$$

pour tout  $i \in I$  et tout  $t \geq t_0$ , avec condition initiale  $Q(t_0)$ . Ces équations correspondent à un bilan instantané de matière.

1.  $\lambda_i(t)$  est un taux de création pour le compartiment  $i$ . Il agit dans (6.1) comme un taux de vie ou d'immigration constant ;
2.  $\rho_{i,i}(t)$  est un taux d'autoreproduction du compartiment  $i$ . Il agit dans (6.1) comme un taux de vie ou d'immigration linéaire ;
3.  $\rho_{i,j}(t)$  avec  $i \neq j$  est un taux de transfert du compartiment  $i$  vers le compartiment  $j$ . Il agit dans (6.1) comme un taux de transfert linéaire de population ;
4.  $\kappa_i(t)$  est un taux de destruction pour le compartiment  $j$ . Il agit dans (6.1) comme un taux de mort ou d'émigration linéaire.

Ces taux sont positifs ou nuls, de sorte que si  $Q(0) \in \mathbb{R}_+^I$  alors  $Q(t) \in \mathbb{R}_+^I$  pour tout  $t \geq 0$ . Les termes « constant » et « linéaire » concernent la dépendance en  $Q(t)$  et non pas la dépendance en  $t$ . Dans l'équation (6.1), les taux  $\kappa_i$  et  $\rho_{i,j}$  agissent proportionnellement au contenu du compartiment qui *donne de la matière*. En termes vectoriels, la dynamique de  $Q(t)$  est de la forme

$$\partial_t Q(t) = \mathbf{M}(t)Q(t) + \lambda(t), \quad (6.2)$$

où  $\lambda(t)$  désigne le vecteur  $(\lambda_i(t); i \in I)$ , et où  $\mathbf{M}(t)$  désigne la matrice définie par

$$\mathbf{M}_{i,j}(t) := \begin{cases} \rho_{j,i}(t) & \text{si } i \neq j \\ -\sum_{k \neq i} \rho_{i,k}(t) - \kappa_i(t) & \text{si } i = j \end{cases}$$

pour tous  $i, j \in I$ . La théorie classique des ÉDO linéaires donne

$$Q(t) = \mathbf{R}(t, t_0)Q(t_0) + \int_{t_0}^t \mathbf{R}(t, u)\lambda(u) du$$

pour tout  $t \geq t_0$ , où la résolvante  $\mathbf{R}$  est solution de l'ÉDO matricielle linéaire suivante :

$$\mathbf{R}(t_0, t_0) = \mathbf{I}, \quad \text{et} \quad \partial_t \mathbf{R}(t, t_0) = \mathbf{M}(t)\mathbf{R}(t, t_0) \quad \text{pour tout } t \geq t_0.$$

Dans la littérature, la matrice  $\mathbf{M}$  est parfois appelée *matrice de transfert*. Il n'y a pas de formule explicite pour la résolvante lorsque  $\mathbf{M}$  dépend du temps (car  $\mathbf{M}(u)$  et  $\mathbf{M}(v)$  ne commutent pas forcément quand  $u \neq v$ ).

L'ensemble des couples  $\{i, j\}$  pour lesquels  $\rho_{i,j} \neq 0$  ou  $\rho_{j,i} \neq 0$  constitue un graphe orienté et pondéré dont les sommets sont les compartiments. Ce graphe donne la topologie des interactions entre compartiments.

**Exercice 6.2.1.** Donner des conditions nécessaires et suffisantes sur les taux pour que le système possède une masse totale constante :  $\partial_t \sum_{i \in I} Q_i(t) \equiv 0$ . C'est par exemple le cas lorsque le système est conservatif (les taux de création, d'autoreproduction et de destruction sont nuls).

**Exemple 6.2.2 (Tapis roulant).** Correspond à un système de  $n$  compartiments numérotés par  $I = \{1, \dots, n\}$ , pour lequel  $\lambda_i \equiv 0$  si  $i \neq 1$ , et  $\rho_{i,j} \equiv 0$  si  $j \neq i + 1$ . Dans ce cas, on note  $\lambda$  pour  $\lambda_1$  et  $\rho_i$  pour  $\rho_{i,i+1}$ . Dans un tel système, il n'y a qu'une seule création de matière (premier compartiment). La destruction de matière peut en revanche avoir lieu dans chaque compartiment. Le système ressemble à une sorte de tapis roulant avec perte. La matrice  $\mathbf{M}$  est tridiagonale et triangulaire.

**Remarque 6.2.3 (Quelques questions naturelles).** Le comportement d'un système compartimental dépend à la fois des taux et des valeurs initiales. La trajectoire d'un seul compartiment est appelée *cinétique*. Plusieurs problèmes sont naturellement associés aux modèles compartimentaux. En particulier, nous pouvons dès à présent isoler les problèmes suivants :

1. problème du *comportement en temps long*, en fonction des taux et des valeurs initiales. Ce problème peut être posé pour le système de tous les compartiments, pour un seul compartiment, ou pour groupe de compartiments ;
2. problème de l'*estimation des taux* à partir d'une *observation bruitée d'une trajectoire* du système de tous les compartiments, d'un seul compartiment, ou pour un groupe de compartiments ;
3. problème de l'*estimation de la loi des taux* à partir d'une *observation bruitée d'une population de trajectoires*, pour tout le système, pour un seul compartiment, ou pour un groupe de compartiments.

Les deux dernières questions se posent également lorsque l'observation concerne le système à l'équilibre, c'est-à-dire en temps infini.

**Remarque 6.2.4 (À retenir).** Les systèmes compartimentaux déterministes linéaires constituent des systèmes d'ÉDO linéaires particuliers. Par conséquent, leur étude générale se résume à l'application de la théorie des systèmes d'ÉDO linéaires.

## 6.3 Modèles déterministes linéaires, taux constants

Lorsque  $\mathbf{M}$  ne dépend pas du temps, on a

$$\mathbf{R}(v, u) = e^{(v-u)\mathbf{M}}.$$

Dire que  $\mathbf{M}$  ne dépend pas du temps revient à dire que les taux  $\rho$  et  $\kappa$  ne dépendent pas du temps. En revanche, les taux  $\lambda$  peuvent très bien dépendre du temps. On obtient l'expression suivante :

$$Q(t) = e^{(t-t_0)\mathbf{M}} Q(t_0) + e^{t\mathbf{M}} \int_{t_0}^t e^{-u\mathbf{M}} \lambda(u) du$$

Cette solution se calcule explicitement lorsque  $e^{-u\mathbf{M}}$  est explicite (par exemple lorsque  $\mathbf{M}$  est nilpotente ou diagonalisable).

Lorsque  $\mathbf{M}$  est constante et inversible et que  $\lambda$  est constant, on obtient l'expression simplifiée suivante :

$$Q(t) = e^{(t-t_0)\mathbf{M}}Q(t_0) + (e^{(t-t_0)\mathbf{M}} - \mathbf{I})\mathbf{M}^{-1}\lambda.$$

L'étude détaillée de ces équations se trouve dans les bons livres de calcul différentiel !

### 6.3.1 Exemple de modèle à un seul compartiment

Considérons un modèle à un seul compartiment, c'est-à-dire tel que  $|I| = 1$ . On note  $\lambda_1 = \lambda$ ,  $\kappa_1 = \kappa$ . Si  $\rho_{1,1} = 0$  alors  $\partial_t Q(t) = \lambda(t) - \kappa(t)Q(t)$ . Lorsque  $\lambda$  et  $\kappa$  sont constants, cette équation admet la solution suivante :

$$Q(t) = Q(0)e^{-\kappa t} + \frac{\lambda}{\kappa}(1 - e^{-\kappa t}).$$

En particulier,  $Q(t) - \lambda/\kappa = (Q(0) - \lambda/\kappa)e^{-\kappa t}$ . Ainsi, quelque soit la valeur initiale  $Q(0)$ , on a  $\lim_{t \rightarrow \infty} Q(t) = \lambda/\kappa$ , et cette convergence a lieu à vitesse exponentielle. Le lecteur perspicace remarquera que la formule pour  $Q(t)$  ci-dessus coïncide avec l'expression de la moyenne au temps  $t$  d'une file d'attente M/M/ $\infty$  d'intensité d'arrivée  $\lambda$  et d'intensité de service  $\kappa$ . La suite du chapitre montre que cette observation est pertinente.

### 6.3.2 Exemple de modèle à deux compartiments

Considérons un modèle à deux compartiments numérotés 1 et 2, à taux constants et dont la structure du graphe est donnée par les identités  $\lambda_2 = \kappa_2 = \rho_{1,1} = \rho_{2,2} = 0$ . Le compartiment 2 apparaît ainsi comme un réservoir. Le système ne reçoit et de perd de la matière qu'à travers le compartiment 1. Le système d'ÉDO associé est :

$$\begin{cases} \partial_t Q_1(t) &= -(\rho_{1,2} + \kappa_1)Q_1(t) + \rho_{2,1}Q_2(t) + \lambda_1 \\ \partial_t Q_2(t) &= \rho_{1,2}Q_1(t) - \rho_{2,1}Q_2(t) \end{cases}.$$

La matrice  $\mathbf{M}$  et le vecteur  $\lambda$  sont donc donnés par :

$$\mathbf{M} = \begin{pmatrix} -\rho_{1,2} + \kappa_1 & \rho_{2,1} \\ \rho_{1,2} & -\rho_{2,1} \end{pmatrix} \quad \text{et} \quad \lambda = \begin{pmatrix} \lambda_1 \\ 0 \end{pmatrix}.$$

La résolution du système s'obtient en diagonalisant  $\mathbf{M}$ , et se trouve par exemple dans [MK00, page 105]. Elle est de la forme

$$\begin{cases} \partial_t Q_1(t) &= \alpha_1 e^{a_1 t} + \alpha_2 e^{a_2 t} + \alpha_3 \\ \partial_t Q_2(t) &= \beta_1 e^{a_1 t} + \beta_2 e^{a_2 t} + \beta_3 \end{cases}.$$

où les  $a_i$  sont les valeurs propres de  $\mathbf{M}$  et les  $\alpha_i$  et  $\beta_i$  s'expriment en fonction des taux. Il est possible de vérifier que  $Q_1(\infty) + Q_2(\infty) = 2\lambda_1/\kappa_1$ , ce qui donne la fonction des taux estimable à partir de l'observation d'un système à l'équilibre. Ce modèle compartimental est utilisé par exemple pour l'accumulation des métaux lourds dans les organismes.

De manière générale, un modèle à deux compartiments et à taux constants possède une solution générale et explicite, car l'exponentielle d'une matrice  $2 \times 2$  s'exprime explicitement en fonction de ses coefficients. La formule, un peu lourde, se trouve dans [MK00, section 8.2.2 page 103], et peut également s'obtenir avec un logiciel de calcul symbolique comme Maple.

### 6.3.3 Exemple de modèle à trois compartiments

Un modèle classique utilisé pour l'élimination du calcium dans les organismes est constitué par trois compartiments numérotés 1, 2, 3, et des taux constants vérifiant les identités  $\lambda_1 = \lambda_2 = \lambda_3 = 0$ ,  $\rho_{1,1} = \rho_{2,2} = \rho_{3,3} = 0$ ,  $\kappa_1 = \kappa_3 = 0$ , et enfin  $\rho_{1,3} = 0$ . Le compartiment 2 représente le plasma, le compartiment 3 représente les os, tandis que le compartiment 1 représente les autres tissus. Le compartiment 2 est le seul à perdre de la matière (élimination du calcium). Ce modèle est étudié dans [MK00, section 8.3.2 page 107].

## 6.4 Modèles stochastiques linéaires

Considérons un système comportant un nombre fini de compartiments, indexés par un ensemble  $I$ . Chaque compartiment contient des particules, et on désigne par  $N_t^i$  le nombre de particules dans le compartiment  $i \in I$  au temps  $t \geq t_0$ . Les particules sont indistinguables, et peuvent être créées, détruites, ou passer d'un compartiment à un autre, conformément à une dynamique markovienne du vecteur d'état

$$N_t := (N_t^i; i \in I).$$

Le processus de Markov  $(N_t)_{t \geq t_0}$  a pour espace d'états  $\mathbb{N}^I$ . Ses probabilités de transition sont données pour tous  $x$  et  $y$  dans  $E$  et tout  $t \geq s \geq t_0$  par les quantités

$$p_{s,t}(x, y) := \mathbb{P}(N_t = y | N_s = x).$$

Cela permet de définir l'opérateur  $P_{s,t}$  agissant sur les fonctions bornées  $E \rightarrow \mathbb{R}$  par

$$P_{s,t}(f)(x) := \sum_{y \in E} p_{s,t}(x, y) f(y) = \mathbb{E}(f(N_t) | N_s = x).$$

En particulier, on a  $P_{s,s}(f) = f$ . La markovianité du processus est capturée par l'équation de Chapman-Kolmogorov :

$$P_{s,u}(P_{u,t}(f)) = P_{s,t}(f)$$

pour tous  $t \geq u \geq s \geq t_0$ . La dynamique du processus est décrite par les générateurs infinitésimaux  $(\mathbf{L}_t)_{t \geq t_0}$ , définis par

$$(\mathbf{L}_t f)(x) := \lim_{\varepsilon \rightarrow 0^+} \frac{P_{t,t+\varepsilon}(f)(x) - P_{t,t}(f)(x)}{\varepsilon}.$$

Ici encore, si  $\mathbf{L}_t(x, y) := (\mathbf{L}_t \mathbf{I}_{\{y\}})(x)$ , on dispose de l'écriture matricielle suivante :

$$(\mathbf{L}_t f)(x) := \sum_{y \in E} \mathbf{L}_t(x, y) f(y).$$

Pour un bon choix de dynamique, le nombre moyen  $Q(t) := \mathbb{E}(N_t | N_{t_0} = x)$  de particules par compartiment est solution d'une ÉDO linéaire de condition initiale  $x$ , comme l'exprime le théorème suivant.

**Théorème 6.4.1.** *Si pour tout  $t \geq t_0$ , la fonction*

$$x \in E \mapsto A_t(x) := \sum_{y \in E} y \mathbf{L}_t(x, y)$$

*est affine, alors la quantité*

$$Q(s, t, x) := \mathbb{E}(N_t | N_s = x)$$

*définie pour tout  $x \in E$  et tous  $t \geq s \geq t_0$  est solution de l'ÉDO linéaire*

$$\partial_t Q(s, t, x) = A_t(Q(s, t, x)),$$

*pour tous  $t \geq s \geq t_0$ , avec la condition initiale  $Q(s, s, x) = x$ .*

FIXME: lien avec les martingales.

*Démonstration.* L'équation de Chapman-Kolmogorov, combinée à la définition des générateurs, fournit l'équation suivante, connue sous le nom d'équation « forward » :

$$\partial_t P_{s,t}(f)(x) := \lim_{\varepsilon \rightarrow 0^+} \frac{P_{s,t+\varepsilon}(f)(x) - P_{s,t}(f)(x)}{\varepsilon} = P_{s,t}(\mathbf{L}_t(f))(x).$$

Maintenant, fixons  $i \in I$ , et considérons la fonction  $f$  donnée par  $f(x) = x_i$  pour tout  $x \in E$ . On a alors  $(\mathbf{L}_t f)(x) = A_t(x)_i$ . Puisque  $A_t$  est affine et que  $P_{s,t}$  est markovien, ils commutent, et on obtient  $P_t(\mathbf{L}_t f)(x) = A_t(P_{s,t}(f)(x))_i$ .  $\square$

L'équation de Chapman-Kolmogorov qui apparaît dans la preuve du théorème 6.4.1 fournit également l'équation « backward »

$$\partial_s P_{s,t}(f)(x) = -\mathbf{L}_s(P_{s,t}(f))(x).$$

Cette équation conduit à l'ÉDO linéaire suivante :

$$\partial_s Q(s, t, x) = -\sum_{y \in E} \mathbf{L}_s(x, y) Q(s, t, y)$$

avec condition initiale  $Q(s, s, x) = x$ . Lorsque le processus est homogène en temps, le générateur  $\mathbf{L}_t$  ne dépend pas du temps et  $P_{s,t} = P_{t_0, t_0 + t - s}$ . Ainsi, dans ce cas,

$$P_u(\mathbf{L}(f)) = \mathbf{L}(P_u(f))$$

pour tout  $u \geq t_0$ , et les équations forward et backward sont identiques à un signe près.

## 6.5 Lien entre déterministe et stochastique

Considérons un modèles compartimental stochastique linéaire, dont les générateurs sont donnés pour tous  $x, y \in E$  avec  $x \neq y$  et tout  $t \geq t_0$  par

$$\mathbf{L}_t(x, y) := \begin{cases} \lambda_i(t) & \text{si } y = x + e_i \text{ pour } i \in I \text{ (naissance d'une particule)} \\ x_i \rho_{i,j}(t) & \text{si } y = x - e_i + e_j \text{ pour } i \neq j \text{ dans } I \text{ (transfert)} \\ x_i \kappa_i(t) & \text{si } y = x - e_i \text{ pour } i \in I \text{ (mort d'une particule)} \\ 0 & \text{sinon} \end{cases}. \quad (6.3)$$

Les termes diagonaux de  $\mathbf{L}_t$  sont donnés par

$$\mathbf{L}_t(x, x) := - \sum_{y \neq x} \mathbf{L}_t(x, y).$$

La notation  $e_i$  désigne le vecteur qui vaut 1 pour le compartiment  $i$  et 0 pour les autres compartiments. La transition associée à  $\rho$  peut s'écrire également  $x_j \rho_{j,i}(t)$  si  $y = x - e_j + e_i$  pour  $j \neq i$ . Rappelons que  $N_t$  est un vecteur à  $|I|$  composantes, qui représente le nombre de particules dans chacun des compartiments au temps  $t$ . Un calcul simple montre que pour tout  $x \in E$  et tout  $i \in I$ ,

$$A_t(x)_i := \sum_{y \in E} \mathbf{L}_{t,x,y} y_i = \lambda_i(t) + \sum_{j \in I, j \neq i} x_j \rho_{j,i}(t) - x_i \left[ \kappa_i(t) + \sum_{j \in I, j \neq i} \rho_{i,j}(t) \right]. \quad (6.4)$$

On reconnaît l'équation (6.2) des modèles déterministes linéaires.

### 6.5.1 Files d'attente M/M/ $\infty$ en interaction

Lorsque le système ne comporte qu'un seul compartiment, le générateur (6.3) est exactement celui d'une file d'attente M/M/ $\infty$  inhomogène en temps : pour  $x \neq y$

$$\mathbf{L}_t(x, y) := \begin{cases} \lambda(t) & \text{si } y = x + 1 \text{ (naissance d'une particule)} \\ x \kappa(t) & \text{si } y = x - 1 \text{ (mort d'une particule)} \\ 0 & \text{sinon} \end{cases}.$$

Plus généralement, lorsque le système comporte plusieurs compartiments, le générateur (6.3) suggère d'interpréter le processus  $(N_t)_{t \geq t_0}$  comme des files d'attentes

M/M/ $\infty$  en interaction. La quantité  $N_t^i$  correspond au nombre de clients dans la  $i^{\text{th}}$  file au temps  $t$ . Pour chaque  $i \in I$ , et en absence d'interaction ( $\rho \equiv 0$ ), le taux d'arrivée est  $\lambda_i$  et le taux de service est  $\kappa_i$ . La linéarité du coefficient  $x_i \kappa_i$  correspond à la nature M/M/ $\infty$  des files. L'interaction  $\rho$  permet aux clients de passer de la file  $i$  à la file  $j$  avec un taux  $x_i \rho_{i,j}$ .

Ces files d'attentes en interaction constituent des réseaux de Petri particuliers (également appelés systèmes de particules en interaction) cf. par exemple [Dur95] et [Lig05]. Ces files en interaction ne sont pas en tandem, ne ne constituent pas des réseaux de Jackson. En mécanique statistique, ces files en interaction constituent des processus « zero-range » sur le graphe des compartiments.

### 6.5.2 Superposition de marches aléatoires

Pour simplifier, les taux qui interviennent dans cette section ne dépendent pas du temps. Considérons une marche aléatoire à temps continu, dont l'espace d'états est  $I \cup \{c\}$ , où  $c$  est un état supplémentaire qui va jouer le rôle de d'état absorbant (ou point cimetière). La dynamique de cette marche est décrite par le générateur infinitésimal suivant : pour  $i \neq j$  dans  $I \cup \{c\}$ ,

$$\mathbf{L}_{i,j} := \begin{cases} \rho_{i,j} & \text{si } i \neq c \text{ et } j \neq c \\ \kappa_i & \text{si } j = c \\ 0 & \text{sinon} \end{cases}$$

Faisons apparaître à présent en chaque compartiment  $i \in I$  une copie indépendante de la marche, selon un processus de Poisson d'intensité  $\lambda_i$  (qui donne les temps d'apparition). Les processus de Poisson utilisés pour chacun des compartiments sont choisis indépendants.

Soit  $Z_t^i$  le nombre de marches se trouvant dans le compartiment  $i \in I$  au temps  $t$ . Le vecteur  $Z_t := (Z_t^i; i \in I)$  suit la même loi que le processus  $(N_t)_{t \geq t_0}$  généré par (6.3).

En effet, l'absence mémoire des lois exponentielles et l'interprétation du générateur infinitésimal en terme d'horloges exponentielles en compétition conduit à l'expression (6.3) pour le générateur infinitésimal de  $(Z_t)$ .

Lorsque  $|I| = 1$ , cette construction correspond à la construction de la file d'attente M/M/ $\infty$  à partir de particules créées selon un processus de Poisson, et détruites après une durée de vie exponentielle (ces durées sont indépendantes et indépendantes du processus de Poisson).

Une telle interprétation reste toujours valable lorsque les taux dépendent du temps.

**FIXME:** Ajouter un point cimetière conduit à des formules de Feynman-Kac.

On a donc deux interprétation de la dynamique (6.3) de  $(N_t)$ . Une interprétation horizontale en terme de superposition de marches aléatoires, et une interprétation verticale en terme de files d'attente en interaction.

**Remarque 6.5.1** (Modèles individus centrés). Les modèles compartimentaux décrivent l'évolution temporelle de population d'individus, regroupés en compartiments, plutôt que l'évolution des individus eux-mêmes. A contrario, les modèles individus-centrés décrivent l'évolution des individus. Chaque individu possède d'une part une *position spatiale*, et d'autre part une variable d'état appelée *trait* (par exemple une couleur, l'allèle d'un gène, etc). Les positions et les traits évoluent au cours du temps. Les modèles individus-centrés décrivent l'évolution spatiale des individus, dictée par des équations de diffusions en interaction. Les interactions entre les diffusions spatiales tiennent compte à la fois de la proximité spatiale des individus et de leurs traits. Les traits sont également modifiés par les interactions. Pour résumer, il s'agit de diffusions colorées en interaction. Elles peuvent par exemple être utilisées en épidémiologie et en écologie. En France, ces modèles ont été étudiés sur le plan mathématique par Méléard et ses collaborateurs, et donnent lieu par exemple à des équations aux dérivées partielles non linéaires de type Boltzmann.

**Remarque 6.5.2** (Équations de Fisher-Kolmogorov-Petrovski-Piskounov). FIXME: introduire les équations de Fisher-KPP, parler de densités spatiales de gènes, de propagation de front d'onde.



# Chapitre 7

## Modèles de maturation-survie

Ce chapitre est consacré à l'étude d'un modèle simple de maturation/survie, faisant intervenir des processus ponctuels, des renormalisations en espace continu, et des modèles compartimentaux.

### 7.1 Tapis roulant et maturation/survie

Il s'agit de décrire le mécanisme de maturation d'une particule. Ce mécanisme correspond à la traversée successive d'un nombre fini de stades de maturation, numérotés de 0 à  $n$ . À chaque stade, la particule peut mourir et être placée dans un état absorbant (cimetière)  $c \notin \{0, 1, \dots, n\}$ . Cela correspond à l'étude d'un modèle compartimental de type tapis roulant avec perte.

La maturation est modélisée par un processus de Markov  $(X_t)_{t \geq 0}$  d'espace d'état

$$\mathcal{S} := \{0, 1, \dots, n\} \cup \{c\} \quad \text{avec } c \notin \{0, 1, \dots, n\}$$

et de générateur infinitésimaux  $(\mathbf{L}_t)_{t \geq 0}$  donnés pour tous  $i \neq j$  dans  $\mathcal{S}$  et tout  $t \geq 0$  par

$$(\mathbf{L}_t)_{i,j} = \begin{cases} \rho & \text{si } i \neq c \text{ et } j = i + 1, \\ \kappa(t, i) & \text{si } i \neq c \text{ et } j = c, \\ 0 & \text{sinon.} \end{cases} \quad (7.1)$$

Le taux  $\rho$  est un nombre réel strictement positif, tandis que la fonction

$$\kappa : \mathbb{R} \times \{0, \dots, n\} \rightarrow \mathbb{R}_+$$

est lisse en la première variable (temps). Ce taux de mort dépend à la fois du temps et du stade de maturation. Le processus  $(X_t)_{t \geq 0}$  décrit la position d'une particule, qui passe de l'état 0 à l'état  $n$  en avançant d'une unité. À chaque transition, la particule peut être tuée et placée dans l'état absorbant  $c$ . Lorsque  $\kappa \equiv 0$ , le processus  $(X_t)_{t \geq 0}$  est un processus de Poisson simple d'intensité  $\rho$ , arrêté à l'état  $n$ .

La probabilité de traversée du système de maturation est donnée par

$$p := \mathbb{P}(\exists t \geq 0; X_t = n \mid X_0 = 0).$$

Le temps de maturation est donné par

$$\tau = \inf\{t \geq 0; X_t = n\}$$

Ce temps est aléatoire.

**Remarque 7.1.1** (Formule de Feynman-Kac). Il est possible d'exprimer la loi de  $(X_t)_{t \geq 0}$  avec une formule de Feynman-Kac, en utilisant le processus de Poisson sous-jacent. De manière générale, ce procédé reste valable pour les processus de Markov perturbés par l'adjonction d'un potentiel (fonction  $\kappa$ ) et d'un état cimetière (absorbant).

Nous allons à présent faire tendre  $n$  vers l'infini, de sorte que l'espace d'état devienne  $[0, 1] \cup \{c\}$ . Soit  $X^n$  le processus décrit ci-dessus, d'espace d'état  $\{0, 1, \dots, n\}$ , avec taux  $\rho^n$  et  $\kappa^n$ . On définit ensuite le processus  $(Y^n)$  par

$$Y_t^n := \frac{1}{n} X_t^n.$$

**Théorème 7.1.2.** Pour tout  $x \in [0, 1]$ , si  $Y_0^n = \lfloor nx \rfloor$ , si  $\lim_{n \rightarrow \infty} \rho^n/n = \rho$  et si

$$\lim_{n \rightarrow \infty} \sup_{i \in \{0, \dots, n\}; t \in \mathbb{R}} |\kappa^n(t, i) - \kappa(t, i/n)| = 0$$

pour une fonction continue par morceaux et bornée

$$\kappa : \mathbb{R} \times [0, 1] \rightarrow \mathbb{R}_+,$$

alors le processus  $Y^n$  converge en loi lorsque  $n \rightarrow \infty$  vers un processus de Markov à temps continu  $Y$  à valeurs dans  $[0, 1] \cup \{c\}$ , vérifiant  $Y_0 = x$ , et de générateurs

$$\mathbf{L}_t(f)(x) = \begin{cases} \rho f'(x) + \kappa(t, x)(f(c) - f(x)) & \text{si } x \neq c \\ 0 & \text{si } x = c \end{cases}.$$

*Démonstration.* Il s'agit d'un cas particulier des méthodes de passage à la limite étudiées dans les premiers chapitres. L'ingrédient supplémentaire ici est l'inhomogénéité en temps des processus.  $\square$

**Remarque 7.1.3.** Le processus vérifie pour tous  $t \geq s \geq 0$  et tout  $x \in [0, 1]$ ,

$$\mathcal{L}(Y_t \mid Y_s = x) = \begin{cases} p(s, t, x) \delta_{x(t-s)} + (1 - p(s, t, x)) \delta_c & \text{si } x \neq c \\ \delta_c & \text{si } x = c \end{cases},$$

où  $x(u) := \min(1, x + \rho u)$ ,

$$p(u, v, x) := \exp \left( - \int_u^v \kappa(w, x(w-u)) dw \right),$$

et  $p(u, v, c) := 0$  pour tout  $x \neq c$ . Ces formules constituent un cas particulier des formules de Feynman-Kac, pour une diffusion déterministe de générateur  $f \mapsto \rho f'$  (dérive uniquement). L'aléa provient uniquement du “potentiel”  $\kappa$ . L'état  $c$  est absorbant.

Dans le modèle continu obtenu, la maturation correspond à un mouvement rectiligne uniforme de vitesse  $\rho$  le long de l'intervalle  $[0, 1]$ . À tout moment, la particule peut être tuée et placée dans l'état cimetière (point absorbant)  $c$ .

Pour tout  $t \in \mathbb{R}$  et tout  $x \in [0, 1]$ , posons

$$g(t, x) := \kappa(t, x)I_{[0,1]}(x) \quad \text{et} \quad \mu(t) := \kappa(t, 1)$$

de sorte que

$$\kappa(t, x) = g(t, x)I_{[0,1]}(x) + \mu(t)I_{\{1\}}(x).$$

La fonction  $g$  correspond à l'action de  $\kappa$  pendant la maturation (i.e. sur l'intervalle d'états  $[0, 1]$ ), tandis que  $\mu$  correspond à l'action après maturation (i.e. après arrivée à l'état final 1).

Le temps de maturation est donné par la constante  $\tau := 1/\rho$ . Il n'est pas aléatoire. La probabilité de survie pendant la maturation est donnée par

$$p := \exp \left( - \int_0^\tau g(w, \rho w) dw \right).$$

Bien entendu, si  $g \equiv 0$ , alors  $p = 1$ , et plus généralement,  $p$  décroît quand  $g$  croît.

## 7.2 Comptage et maturation/survie

Considérons un modèle de maturation de particules dans lequel la maturation prend un temps déterministe  $\tau > 0$ . Une particule qui commence sa maturation à l'instant  $t \in \mathbb{R}$  achève sa maturation à l'instant  $t + \tau$ . Supposons également que la particule peut mourir pendant le processus de maturation, avec probabilité  $1 - p(t)$ , où  $p(t) \in [0, 1]$  et où  $t$  désigne le temps de début de maturation de la particule. Dans ce schéma de maturation/survie, une particule qui commence sa maturation au temps  $t \in \mathbb{R}$  peut soit mourir avec probabilité  $1 - p(t)$ , soit survivre avec probabilité  $p(t)$  et achever sa maturation au temps  $t + \tau$ .

**Théorème 7.2.1** (Maturation/survie = translation/amincissement). *Considérons des particules qui commencent leur maturation à des instants répartis selon un processus ponctuel de Poisson sur  $\mathbb{R}$  d'intensité  $\lambda : \mathbb{R} \rightarrow \mathbb{R}_+$ . Si  $\tau$  désigne la durée déterministe de la maturation et si  $p(t)$  désigne la probabilité de survie d'une particule pendant maturation ( sachant qu'elle a commencé sa maturation à l'instant  $t$ ), alors en supposant que les survies sont indépendantes, les temps de fin de maturation des particules survivantes suivent un processus ponctuel de Poisson d'intensité  $\tilde{\lambda} : \mathbb{R} \rightarrow \mathbb{R}_+$  donnée pour tout  $t \in \mathbb{R}$  par*

$$\tilde{\lambda}(t) := \lambda(t - \tau)p(t - \tau).$$

*Démonstration.* Découle immédiatement de la propriété de stabilité par amincissement et translation des processus ponctuels de Poisson.  $\square$

**Remarque 7.2.2** (Durées de vie). Soit  $\mu : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  une fonction localement Lebesgue intégrable. On dit qu'une variable aléatoire  $Z$  à valeurs dans  $\mathbb{R}_+$  est d'intensité  $\mu$  lorsque pour tous  $s \leq t$  dans  $\mathbb{R}_+$ ,

$$\mathbb{P}(Z > t | Z > s) = \exp\left(-\int_s^t \mu(u) du\right).$$

La variable aléatoire  $Z$  suit la même loi que  $F^{-1}(E)$  où  $E$  est une variable aléatoire exponentielle de moyenne 1, et où  $F^{-1}$  est l'inverse généralisée de la fonction continue et croissante  $F : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  définie pour tout  $t \in \mathbb{R}_+$  par

$$F(t) := \int_0^t \mu(u) du.$$

De manière équivalente,  $F(Z)$  suit la loi exponentielle de moyenne 1. Lorsque  $\mu$  est constante, cela signifie tout simplement que  $Z$  suit une loi exponentielle de moyenne  $1/\mu$ . Ces variables aléatoires (ou plutôt ces lois) associées à une fonction d'intensité sont souvent utilisées pour modéliser des durées de vie. Les temps de sauts d'un processus de Poisson simple sur  $\mathbb{R}_+$  sont de cette forme (voir chapitre sur les processus ponctuels).

Supposons à présent qu'après maturation, chaque particule survit pendant une durée aléatoire d'intensité  $\mu : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ , indépendamment de toutes les autres. Le théorème suivant exprime la loi du nombre  $N_t$  de particules matures et encore en vie au temps temps  $t \in \mathbb{R}_+$ .

**Théorème 7.2.3.** Pour tous  $0 \leq s \leq t$  et tout  $m \in \mathbb{N}$ ,

$$\mathcal{L}(N_t | N_s = m) = \mathcal{B}(m, \alpha(s, t)) * \mathcal{P}(\beta(s, t))$$

où

$$\alpha(s, t) := \exp\left(-\int_s^t \mu(w) dw\right) \quad \text{et} \quad \beta(s, t) := \int_s^t \lambda(u - \tau) p(u - \tau) \alpha(u, t) du.$$

En particulier, pour tous  $0 \leq s \leq t$  et tout  $m \in \mathbb{N}$ ,

$$\mathbb{E}(N_t | N_s = m) = m\alpha(s, t) + \beta(s, t).$$

Il y a trois sources d'aléa qui sont indépendantes : le temps de début de maturation (processus de Poisson), le fait de survivre pendant maturation (amincissement Bernoulli), la durée de survie après maturation (amincissement par durée de vie).

*Démonstration.* Lorsque  $\lambda$  et  $\mu$  sont constantes, alors  $(N_t)_{t \geq 0}$  est une file d'attente M/M/ $\infty$  (homogène en temps). Plus généralement, le processus  $(N_t)_{t \geq 0}$  est une file d'attente M/M/ $\infty$  inhomogène en temps. Le résultat découle de la stabilité des processus ponctuels de Poisson par marquage et amincissement, comme pour la file d'attente M/M/ $\infty$  homogène en temps.  $\square$

### 7.3 Tapis roulant et comptage

Considérons le processus de maturation/survie ( $X_t$ ) introduit précédemment, avec vitesse de maturation  $\rho$  et fonction d'élimination  $\kappa : \mathbb{R} \times [0, 1] \rightarrow \mathbb{R}_+$ . La maturation a lieu pendant une durée déterministe  $\tau$ . Une particule qui débute sa maturation au temps  $T$  peut soit mourir avec probabilité  $1 - p(T)$  soit survivre avec probabilité  $p(T)$  et finir sa maturation au temps  $T + \tau$ . Maintenant, considérons des particules indépendantes qui débutent leur maturation aux temps  $T_0, T_1, \dots$ . Ces particules survivent avec probabilités  $p(T_0), p(T_1), \dots$ , et achèvent leur maturation aux temps  $M_0 = T_0 + \tau, M_1 = T_1 + \tau, \dots$  à condition que les probabilités  $p(T_i)$  soient non nulles. Les durées de survie après maturation sont des variables aléatoires indépendantes de taux commun  $\mu : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  où  $\mu(t) := \kappa(t, 1)$ . Supposons que les temps initiaux  $T_0, T_1, \dots$  sont aléatoires et distribuées selon un processus ponctuel de Poisson indépendant sur  $\mathbb{R}$  d'intensité  $\lambda : \mathbb{R} \rightarrow \mathbb{R}_+$ . Pour tout  $t \geq 0$ , soit  $N_t$  le nombre de particules matures encore en vie à l'instant  $t$ . Le résultat suivant est un corollaire des résultats précédents.

**Théorème 7.3.1.** *Pour tout  $0 \leq s \leq t$  et tout  $m \in \mathbb{N}$ ,*

$$\mathcal{L}(N_t | N_s = m) = \mathcal{B}(m, \alpha(s, t)) * \mathcal{P}(\beta(s, t)),$$

où

$$\alpha(s, t) := \exp \left( - \int_s^t \mu(w) dw \right)$$

et

$$\beta(s, t) := \int_{s-\tau}^{t-\tau} \lambda(u) \exp \left( - \int_0^\tau g(u+w, \rho w) dw \right) \alpha(u+\tau, t) du.$$

En particulier, pour tout  $0 \leq s \leq t$  et tout  $m \in \mathbb{N}$ ,

$$\mathbb{E}(N_t | N_s = m) = m\alpha(s, t) + \beta(s, t). \quad (7.2)$$

Le processus  $(N_t)_{t \geq 0}$  n'est rien d'autre qu'une file d'attente  $M/M/\infty$  inhomogène en temps. Lorsque  $g \equiv 0$  et  $\mu$  est constant, on a  $p \equiv 1$ , et  $(N_t)_{t \geq 0}$  devient alors une file d'attente  $M/M/\infty$  d'intensités constantes  $\lambda$  et  $\mu$ . Lorsque de plus  $\lambda$  est constant, la loi de Poisson  $\mathcal{P}(\lambda/\mu)$  est invariante et symétrique (et donc réversible).

Le rôle joué par la file  $M/M/\infty$  dans ce modèle est du au fait que les particules sont indépendantes. L'adjonction d'une interaction entre particules, avant, pendant, ou après maturation conduirait à un processus de comptage différent, dont les taux ne sont pas affine.

**Remarque 7.3.2 (Valeurs négatives de  $\kappa$  et amplification).** L'expression de  $\beta$  donnée ci-dessus fait encore sens lorsque  $\kappa$  prend des valeurs négatives sur l'intervalle  $[0, 1]$ . Dans ce cas, la quantité  $p(u) = \exp(-\int_0^\tau g(u+w, \rho w) dw)$  peut dépasser 1 et ne peut donc plus être interprétée comme une probabilité de survie. De telles valeurs négatives de  $\kappa$  peuvent être vues comme une sorte d'amplification de l'intensité d'arrivée, par opposition à l'élimination, et peuvent être par conséquent incorporées à  $\lambda$ .

### Cas sans élimination après maturation

Supposons que  $\mu \equiv 0$  (pas d'élimination après maturation). Dans ce cas,  $\alpha(s, t) = 0$  pour tout  $0 \leq s \leq t$ , et la formule exprimant  $\beta(s, t)$  devient

$$\beta(s, t) = \int_{s-\tau}^{t-\tau} \lambda(u) \exp\left(-\int_0^\tau g(u+w, \rho w) dw\right) du.$$

Dans ce cas,  $t \in \mathbb{R}_+ \mapsto \beta(s, t)$  est croissante puisque la fonction intégrée ne dépend pas de  $t$ . Cela n'est pas surprenant car les particules ne sont pas tuées après maturation. Ainsi, sur l'événement  $\{N_s = 0\}$ , le processus  $t \in \mathbb{R}_+ \mapsto M_t$  est croissant, et en particulier sa valeur moyenne  $\beta(s, t)$  est croissante également. Lorsque  $\lambda$  est constant et  $g \equiv 0$ , on retrouve le processus de Poisson simple de comptage des tops, d'intensité  $\lambda$ .

### Cas sans élimination pendant maturation

Supposons que  $g \equiv 0$  (pas d'élimination pendant maturation). Dans ce cas, la formule exprimant  $\beta(s, t)$  pour  $0 \leq s \leq t$  devient

$$\beta(s, t) = \int_{s-\tau}^{t-\tau} \lambda(u) \alpha(u + \tau, t) du.$$

Lorsque de plus  $\lambda$  et  $\mu$  sont constants, on retrouve la file d'attente  $M/M/\infty$  homogène en temps d'intensités  $\lambda$  et  $\mu$ , pour laquelle

$$\alpha(s, t) = e^{-(t-s)\mu} \quad \text{and} \quad \beta(s, t) = (1 - e^{-(t-s)\mu}) \frac{\lambda}{\mu}. \quad (7.3)$$

La loi  $\mathcal{P}(\lambda/\mu)$  est invariante et symétrique (et donc réversible) pour  $(N_t)_{t \geq 0}$ .

### Élimination partielle pendant maturation et constante après maturation

Considérons à présent le cas particulier où  $\lambda$  est constant et où  $\kappa$  est de la forme

$$k(t, x) = g(t) I_{[0, \delta)}(x) + \mu I_{\{1\}}(x), \quad (7.4)$$

où  $\delta \in [0, 1]$ ,  $\mu \in \mathbb{R}_+$ , et où  $g : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  est une fonction lisse. Il correspond à une élimination qui dépend du temps avant le stade de maturation  $\delta$ , et à une élimination constante après maturation. Il n'y a pas d'élimination pour les stades de maturation après  $\delta$ . La formule pour  $\beta(s, t)$  lorsque  $0 \leq s \leq t$  devient

$$\beta(s, t) = \lambda \int_{s-\tau}^{t-\tau} \exp\left(-\mu(t - \tau - u) - \int_0^{\delta\tau} g(u+w) dw\right) du. \quad (7.5)$$

Lorsque  $g \equiv 0$ , on retrouve la file d'attente  $M/M/\infty$  homogène en temps. Supposons a contrario que  $g$  ne s'annule qu'à l'infini. Dans ce cas, les deux formules (7.5) et (7.3) pour  $\beta$  se confondent lorsque  $t$  tends vers  $\infty$ . En particulier, la loi de Poisson  $\mathcal{P}(\lambda/\mu)$  est une loi stationnaire de  $(N_t)_{t \geq 0}$ .

### Identifiabilité et invariance par certaines transformations

La dynamique de  $N := (N_t)_{t \geq 0}$  est complètement décrite par le quadruple  $(\tau, \mu, g, \lambda)$ , où  $\tau := 1/\rho$ . Il est naturel de s'intéresser à l'injectivité de la fonctionnelle

$$(\tau, \mu, g, \lambda) \mapsto \mathcal{L}(N | N_0).$$

En vertu du théorème 7.3.1, la loi  $\mathcal{L}(N | N_0)$  est complètement déterminée par le triplet  $(\tau, \mu, \lambda_g)$ , où  $\lambda_g : \mathbb{R} \rightarrow \mathbb{R}_+$  est définie par

$$\lambda_g(t) := \lambda(t)p(t) = \lambda(t) \exp \left( - \int_0^\tau g(t+w, \rho w) dw \right).$$

Par conséquent, le couple  $(\lambda, g)$  ne peut pas être identifié, car il agit sur la loi du processus à travers l'alliage  $\lambda_g$ . L'action de  $g$  peut être compensée par  $\lambda$  et vice versa. Supposons par exemple que  $g$  s'écrive  $g = g_1 + g_2$ , où  $g_1$  et  $g_2$  sont positives. Alors les modèles correspondants respectivement à  $(\tau, \mu, g, \lambda)$  et à  $(\tau, \mu, g_2, \lambda_{g_1})$  sont indistinguables. Le cas extrême correspond à  $(\tau, \mu, 0, \lambda_g)$ , pour lequel l'intégralité de la fonction d'élimination  $g$  est intégrée dans l'intensité d'arrivée  $\lambda$ .

Analysons quelques cas particuliers. Pour tout  $\theta > 1$ , considérons la transformation qui remplace  $(\lambda, g)$  par  $(\lambda^\theta, g^\theta)$  défini par

$$\lambda^\theta := \theta \lambda \quad \text{et} \quad g^\theta := g + \rho \log(\theta).$$

La fonction  $\mu$  et le paramètre  $\rho$  sont inchangés, et on peut vérifier que  $\lambda_{g^\theta}^\theta = \lambda_g$ . Par conséquent, la dynamique est invariante par cette transformation, et les deux modèles correspondant à  $(\tau, \mu, g, \lambda)$  et  $(\tau, \mu, g^\theta, \lambda^\theta)$  sont indistinguables, malgré leur interprétation différente. Une perturbation multiplicative de l'intensité d'arrivée  $\lambda$  correspond à une perturbation additive de la fonction d'élimination pendant maturation  $g$ . Cela suggère de normaliser la paramétrisation en prenant par exemple  $\lambda \equiv 1$ .



# Bibliographie

- [Bas98] R. F. BASS – *Diffusions and elliptic operators*, Probability and its Applications (New York), Springer-Verlag, New York, 1998. [21](#)
- [BB06] Y. BARAUD et L. BIRGÉ – « Estimating the intensity of a random measure by histogram type estimators », prépublication, <http://arxiv.org/abs/math.ST/0608663>, 2006. [91](#)
- [Bou00] N. BOULEAU – *Processus stochastiques et applications*, Hermann, 2000. [20](#), [90](#)
- [Bré99] P. BRÉMAUD – *Markov chains*, Texts in Applied Mathematics, vol. 31, Springer-Verlag, New York, 1999, Gibbs fields, Monte Carlo simulation, and queues. [20](#)
- [Dur95] R. DURRETT – « Ten lectures on particle systems », Lectures on probability theory (Saint-Flour, 1993), Lecture Notes in Math., vol. 1608, Springer, Berlin, 1995, p. 97–201. [100](#)
- [Dur02] R. DURRETT – *Probability models for DNA sequence evolution*, Probability and its Applications (New York), Springer-Verlag, New York, 2002. [20](#)
- [EK86] S. N. ETHIER et T. G. KURTZ – *Markov processes*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Inc., New York, 1986, Characterization and convergence. [20](#), [40](#)
- [Ewe04] W. J. EWENS – *Mathematical population genetics. I*, second éd., Interdisciplinary Applied Mathematics, vol. 27, Springer-Verlag, New York, 2004, Theoretical introduction. [20](#)
- [GS72] I. Ī. GĪHMAN et A. V. SKOROHOD – *Stochastic differential equations*, Springer-Verlag, New York, 1972, Translated from the Russian by Kenneth Wickwire, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 72. [32](#), [33](#)
- [Jac06] M. JACOBSEN – *Point process theory and applications*, Probability and its Applications, Birkhäuser Boston Inc., Boston, MA, 2006, Marked point and piecewise deterministic processes. [91](#)
- [Kin93] J. F. C. KINGMAN – *Poisson processes*, Oxford Studies in Probability, vol. 3, The Clarendon Press Oxford University Press, New York, 1993, , Oxford Science Publications. [87](#), [91](#)

- [KS91] I. KARATZAS et S. E. SHREVE – *Brownian motion and stochastic calculus*, second éd., Graduate Texts in Mathematics, vol. 113, Springer-Verlag, New York, 1991. [21](#)
- [Kut98] Y. A. KUTOYANTS – *Statistical inference for spatial Poisson processes*, Lecture Notes in Statistics, vol. 134, Springer-Verlag, New York, 1998. [91](#)
- [Lig05] T. M. LIGGETT – *Interacting particle systems*, Classics in Mathematics, Springer-Verlag, Berlin, 2005, Reprint of the 1985 original. [100](#)
- [MK00] J. H. MATIS et T. R. KIFFE – *Stochastic population models*, Lecture Notes in Statistics, vol. 145, Springer-Verlag, New York, 2000, A compartmental perspective. [93](#), [96](#), [97](#)
- [MW04] J. MØLLER et R. P. WAAGEPETERSEN – *Statistical inference and simulation for spatial point processes*, Monographs on Statistics and Applied Probability, vol. 100, Chapman & Hall/CRC, Boca Raton, FL, 2004. [91](#)
- [Nor98] J. R. NORRIS – *Markov chains*, Cambridge Series in Statistical and Probabilistic Mathematics, vol. 2, Cambridge University Press, Cambridge, 1998, Reprint of 1997 original. [20](#)
- [RB03] P. REYNAUD-BOURET – « Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities », *Probab. Theory Related Fields* **126** (2003), no. 1, p. 103–153. [91](#)
- [RB06] — , « Penalized projection estimators of the Aalen multiplicative intensity », *Bernoulli* **12** (2006), no. 4, p. 633–661. [91](#)
- [Rip88] B. D. RIPLEY – *Statistical inference for spatial processes*, Cambridge University Press, Cambridge, 1988. [91](#)
- [Rob00] P. ROBERT – *Réseaux et files d'attente: méthodes probabilistes*, Mathématiques & Applications (Berlin) [Mathematics & Applications], vol. 35, Springer-Verlag, Berlin, 2000. [75](#)
- [Rob03] — , *Stochastic networks and queues*, french éd., Applications of Mathematics (New York), vol. 52, Springer-Verlag, Berlin, 2003, Stochastic Modelling and Applied Probability. [79](#), [80](#), [91](#)
- [RY99] D. REVUZ et M. YOR – *Continuous martingales and Brownian motion*, third éd., Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 293, Springer-Verlag, Berlin, 1999. [90](#)
- [SKM87] D. STOYAN, W. S. KENDALL et J. MECKE – *Stochastic geometry and its applications*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons Ltd., Chichester, 1987, With a foreword by D. G. Kendall. [91](#)
- [Tav04] S. TAVARÉ – « Ancestral inference in population genetics », Lectures on probability theory and statistics, Lecture Notes in Math., vol. 1837, Springer, Berlin, 2004, p. 1–188. [20](#)

- [vL00] M. N. M. VAN LIESHOUT – *Markov point processes and their applications*, Imperial College Press, London, 2000. [91](#)
- [Yca02] B. YCART – *Modèles et algorithmes markoviens*, Mathématiques & Applications (Berlin) [Mathematics & Applications], vol. 39, Springer-Verlag, Berlin, 2002. [20](#), [93](#)