

Master 2 Modélisation des systèmes biologiques

Université de Rennes 1

Modèles aléatoires en biologie

Jürgen ANGST

Notes de cours

<http://perso.univ-rennes1.fr/jurgen.angst/>

Table des matières

Introduction	i
I Les marches aléatoires et leurs limites d'échelle	1
1 La marche aléatoire simple et ses variantes	3
1.1 Ce que l'on cherche à modéliser	3
1.1.1 Déplacement d'une particule de pollen	3
1.1.2 Géométrie des polymères	5
1.2 Marche aléatoire simple	7
1.2.1 La marche aléatoire symétrique	7
1.2.2 Comptage de chemins	8
1.2.3 Passage et temps de passage en zéro	9
1.2.4 Récurrence et transience de la marche simple	10
1.3 Marche à boucles effacées et marche auto-évitante	11
1.3.1 La marche à boucles effacées	11
1.3.2 La marche auto-évitante	12
2 Limites d'échelle des marches aléatoires	15
2.1 De la marche aléatoire au mouvement brownien	15
2.1.1 Convergence de la marche renormalisée	15
2.1.2 Mouvement brownien, premières propriétés	17
2.2 Limite d'échelle des autres marches	20
2.2.1 Évolution de Schramm-Loewner	20
2.2.2 Convergence de la marche à boucles effacées	21
2.2.3 Convergence de la marche auto-évitante	22
3 Notions de calcul stochastique	25
3.1 Les équations différentielles stochastiques	25
3.1.1 Rappel sur les équation différentielles ordinaires	25
3.1.2 Équations différentielles stochastiques	26
3.2 Propriétés élémentaires des diffusions	28
3.2.1 Temps de séjour dans un intervalle	28
3.2.2 Diffusions à coefficients singuliers	29

3.3	Simulation de diffusions	32
3.3.1	Simulation de trajectoires d'une diffusion	32
3.3.2	Estimation de fonctionnelles de la diffusion	32
II	Éléments de génétique des populations	33
4	Autour du modèle de Wright-Fisher	35
4.1	Ce que l'on cherche à modéliser	35
4.1.1	La problématique	36
4.1.2	Hypothèses simplificatrices	36
4.2	Modèle de Wright-Fisher et variantes	37
4.2.1	Le modèle idéal de Hardy-Weinberg	37
4.2.2	Modèle de Wright-Fisher	38
4.2.3	Modèle de Wright-Fisher avec sélection	43
4.2.4	Modèle de Wright-Fisher avec mutation	45
4.2.5	Modèle avec mutation et sélection	46
4.3	Approximations continues des modèles discrets	46
4.3.1	Le cas général	46
4.3.2	Ni mutation, ni sélection	48
4.3.3	Mutation sans sélection	48
5	Coalescence	51
5.1	Construction de l'arbre généalogique	51
5.2	Le modèle à temps discret	52
5.3	Le modèle à temps continu	54
5.4	Le passage à la limite	54
5.5	Propriétés du processus ancestral	56
5.5.1	Temps d'apparition de l'ACPR	56
5.5.2	Longueur de l'arbre	56
5.6	Mutations pour un nombre infini d'allèles	57
5.6.1	Loi du nombre d'allèles dans un échantillon	58
5.6.2	Estimation du taux de mutation	60
	Bibliographie	63

Introduction

De la nécessité des modèles aléatoires en biologie

Les termes *biologie* ou *sciences du vivant* ont des spectres très larges : ils recouvrent à la fois l'étude de phénomènes à l'échelle microscopique, macroscopique, ou mésoscopique. Du côté de "l'infiniment petit", on peut citer par exemple la recherche de séquences sur un brin d'ADN, l'étude des échanges moléculaires dans une cellule, l'évolution de tumeurs cancéreuses, l'invasion de parasites dans une cellule. À l'opposé, parmi les problèmes macroscopiques, mentionnons ceux des comportements de grands groupes d'individus et leurs interactions (extinction de populations, équilibre des éco-systèmes, équilibre proies-prédateurs), ou des problèmes de génétique de populations (recherche d'ancêtres communs à plusieurs individus dans une espèce, phylogénie).

Dans la modélisation mathématique de ces différents phénomènes, quelque soit leur échelle, l'introduction d'objets de nature aléatoire est très fréquente. Il y a plusieurs raisons à cela parmi lesquelles les deux suivantes. D'une part, l'aléa peut-être intrinsèque au phénomène à étudier, auquel cas, une approche probabiliste est inévitable. On peut penser par exemple à l'existence ou non d'une mutation aux cours de la réplication de l'ADN.

D'autre part, l'introduction de quantités aléatoires est motivée par le principe général suivant : même si un phénomène biologique semble présenter un certain nombre de caractéristiques déterministes, il résulte le plus souvent d'interactions entre un grand nombre d'individus (molécules, espèces) dont le comportement est complexe et soumis à une grande variabilité. Un modèle déterministe cherchant à décrire l'évolution globale du phénomène au cours du temps et prenant en compte le comportement de chaque individu, résulterait en un gigantesque système d'équations le plus souvent non résoluble en pratique. La démarche probabiliste consiste à supposer que le comportement de chaque individu est aléatoire (la complexité et la variabilité sont alors cachées dans l'aléa) puis d'utiliser la théorie des probabilités (loi des grands nombres, théorème limite central etc.) pour en déduire l'évolution globale du phénomène. Cela permet ainsi, à partir d'une description individuelle / microscopique très précise d'en déduire des comportements globaux / macroscopiques de manière rigoureuse, qu'ils soient déterministes ou aléatoires.

Objectif du cours

L'objet de ce cours est d'illustrer, à travers quelques exemples significatifs, la richesse de l'approche probabiliste dans la modélisation de phénomènes biologiques. Nous chercherons ainsi à comprendre, en utilisant des modèles aléatoires, l'évolution temporelle, ou *dynamique*, de ces différents phénomènes. Ce faisant, nous serons le plus souvent obligé de simplifier le système biologique pour obtenir un modèle plus facile à étudier mathématiquement. Une difficulté de cette démarche est donc d'obtenir un bon compromis entre le réalisme biologique du modèle d'une part et la faisabilité des calculs d'autre part. L'approche probabiliste est de ce point de vue plus riche qu'une simple approche déterministe puisqu'elle prend en compte la variabilité individuelle. Une autre qualité de cette approche est le fait que des problèmes biologiques très différents (par exemple par les échelles de taille : gènes - cellules - bactéries - individus - colonies) peuvent avoir des comportements aléatoires de même type et correspondre à des modèles analogues.

L'existence d'un modèle mathématique permettra de pouvoir quantifier numériquement certains phénomènes et de pouvoir prédire certains comportements : par exemple montrer qu'une certaine population va s'éteindre et calculer l'espérance du temps d'extinction. Évidemment, on peut toujours (et il le faut !) se poser la question de la justification du modèle. Une étape ultérieure, dans le cas où l'on peut obtenir des données observées pour le phénomène qui nous intéresse, est de construire des tests statistiques, qui permettront, ou non, de valider le modèle.

Les modèles que nous étudierons sont élémentaires et dans la mesure du possible, nous chercherons à démontrer rigoureusement la plupart des énoncés les concernant. Cependant, pour aller plus loin et sachant que l'auditoire n'est pas composé d'étudiants en mathématique, certains résultats seront admis. Nous nous efforcerons alors d'expliquer les grandes lignes de leur preuve, ou en quoi ils sont "naturels". De même, nous introduirons certains concepts avancés, par exemple celui d'équation différentielle stochastique, comme des "boîtes noires" : nous utiliserons ces concepts et les propriétés associées sans justification.

Le plus souvent, nous considérerons des modèles discrets (l'espace et le temps sont discrétisés) mais aussi des modèles continus, et nous passerons parfois des uns aux autres. Ainsi, dans un modèle discret, quand la taille d'un système est très grande, il devient difficile de décrire le comportement individuel / microscopique de ses individus. Nous changerons alors d'échelle de taille de population et d'échelle de temps (passage au continu), pour nous ramener à des approximations plus facilement manipulables du point de vue mathématique (résultats théoriques et calculs). Dans certaines échelles, nous obtiendrons ainsi des approximations déterministes, qui ont été historiquement les premières introduites pour décrire ces évolutions. Dans d'autres échelles, nous obtiendrons des approximations aléatoires. C'est ainsi que nous introduirons le mouvement brownien, un des processus les plus célèbres de l'histoire des probabilités.

Première partie

Marches aléatoires, limites d'échelle,
notions de calcul stochastique

Chapitre 1

La marche aléatoire simple et ses variantes

Ce premier chapitre est consacré à l'introduction de la marche aléatoire simple et de deux de ses variantes, la marche aléatoire à boucles effacées et la marche aléatoire auto-évitante. Les marches aléatoires sont des objets centraux en théorie des probabilités et dans ses applications, au sens où elles constituent l'archétype de nombreux processus stochastiques (chaînes de Markov, martingales) et où elles permettent de modéliser un grand nombre d'objets ou de phénomènes naturels. Nous commençons précisément par motiver l'introduction des marches aléatoires via la modélisation de deux phénomènes biologiques. Nous donnons ensuite la définition mathématique de la marche aléatoire simple et de ses variantes, en rappelant un certains nombres de leurs propriétés.

1.1 Ce que l'on cherche à modéliser

Nous nous limiterons ici essentiellement à la modélisation de deux objets ou phénomènes naturels. Nous chercherons tout d'abord à expliquer le phénomène de diffusion d'une particule de pollen dans l'eau, *i.e.* nous expliquerons pourquoi, en l'absence de toute force extérieure, les grains de pollen se meuvent de façon erratique dans l'eau. Nous chercherons ensuite à décrire la géométrie typique d'un polymère, *i.e.* la forme typique d'une longue molécule obtenue par concaténation d'un grand nombre de briques élémentaires.

1.1.1 Déplacement d'une particule de pollen

En 1828, un botaniste du nom de Robert Brown rend compte de ses observations : une particule de pollen (observable au microscope) plongée dans l'eau est animée de mouvements aussi rapides qu'irréguliers. De nombreuses explications plus ou moins farfelues furent proposées pour rendre compte de ce phénomène mais peu à peu une thèse s'imposait : le mouvement observé serait lié aux chocs de la particule de pollen

avec les molécules d'eau. À la fin du dix neuvième siècle, les points suivants avaient établis empiriquement :

1. le mouvement est incessant, il est très irrégulier et la trajectoire semble ne pas admettre de tangente ;
2. deux particules (de pollen) semblent se déplacer indépendamment (si elles ne se choquent pas), même si la distance les séparant ne dépasse pas leur diamètre ;
3. le mouvement est d'autant plus rapide que la particule est petite, que la température est élevée, et d'autant plus lent que le fluide est visqueux ;
4. la composition et la densité de la particule n'ont aucun effet.

Remarque 1.1.1. Le premier point fut admis après une expérience de trente ans et l'observation de liquides prisonniers de quartz vieux de plusieurs milliers d'années.

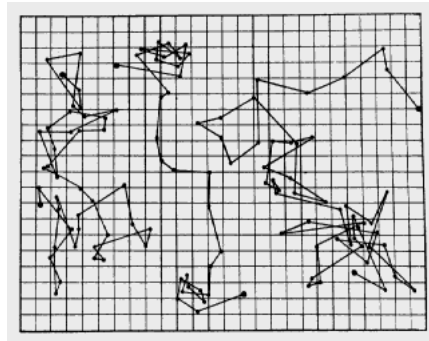


FIG. 1.1 – Trajectoire brownienne empirique observée par Perrin (Les atomes).

L'intuition qui précède la modélisation est la suivante. La particule de pollen est « énorme » par rapport aux molécules d'eau. À chaque instant, d'innombrables chocs ont lieu entre la particule de pollen et les molécules qui l'entourent. Ces chocs se répartissent en moyenne de manière uniforme sur la surface de la particule. Pourtant un instant donné il peut y avoir plus de chocs sur une partie que sur une autre.

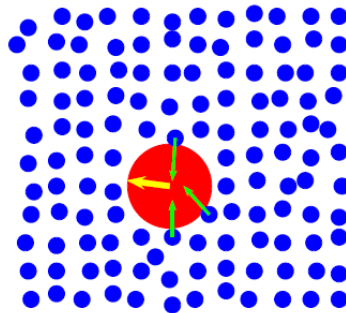


FIG. 1.2 – Chocs entre les molécules d'eau et le grain de pollen.

1.1.2 Géométrie des polymères

Intéressons-nous maintenant à la géométrie des polymères. Un *polymère* est une molécule composée d'une longue chaîne de *monomères* reliés entre eux par des liaisons chimiques. Les monomères peuvent être des unités relativement simples comme le méthylène CH_2 , ou plus complexe comme les paires de base adénine-thymine and cytosine-guanine de l'acide désoxyribonucléique (ADN). Les polymères peuvent être classés en deux grandes famille :

1. les homopolymères, dont tous les monomères sont identiques ;
2. les copolymères, constitués d'au moins deux types de monomères.

Dans ce dernier cas, l'ordre des monomères peut être périodique ou aléatoire.

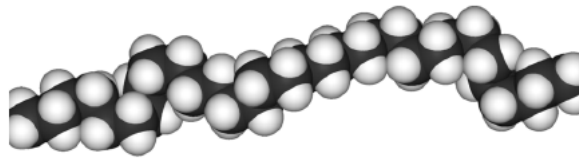


FIG. 1.3 – Exemple de polymère : le polyéthylène.

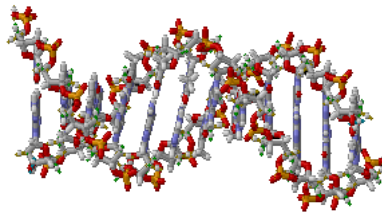


FIG. 1.4 – Structure de la molécule d'ADN.

Les polymères abondent dans la nature, l'une des raisons à cela étant la multi-valence des atomes de type carbone, oxygène, nitrogène ou soufre qui peuvent former de très longues structures par simple concaténation. Le processus chimique, naturel ou artificiel, aboutissant à la formation de polymères à partir des briques élémentaires que sont les monomères est appelé *polymérisation*.

La taille des polymères peut varier de 10^3 à 10^{10} monomères. Par exemple, les polysaccharides sont constitués de 10^3 à 10^4 unités de sucre, la lignine consiste en 10^6 à 10^7 unités de phénylpropane et l'ADN humain est constitué de 10^9 à 10^{10} paires de bases. La taille des monomères varie de $1.5\text{\AA}$ pour le méthylène jusqu'à $20\text{\AA}$ pour les paires de base de l'ADN ($1\text{\AA} = \text{un Angström} = 10^{-10}\text{m}$).

Les liaisons chimiques qui relient les monomères entre eux sont flexibles de sorte que les polymères peuvent présenter des formes très diverses. Plus la chaîne est longue, plus la géométrie de la molécule peut être complexe.

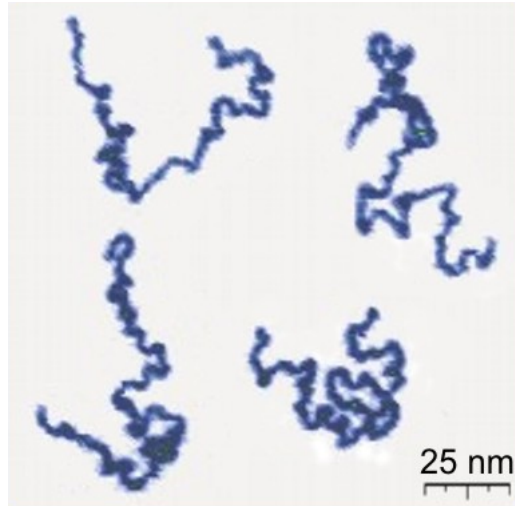


FIG. 1.5 – Polymères observés au microscope électronique à balayage.

Dans la nature, on rencontre ainsi des polymères qui s'enroulent sur eux-mêmes pour former des noeuds, s'allongent à cause de force répulsives ou au contraire se contractent à cause de force attractives (van der Waals), interagissent avec une membrane ou sont soumis à une force à l'une de leur extrémités.

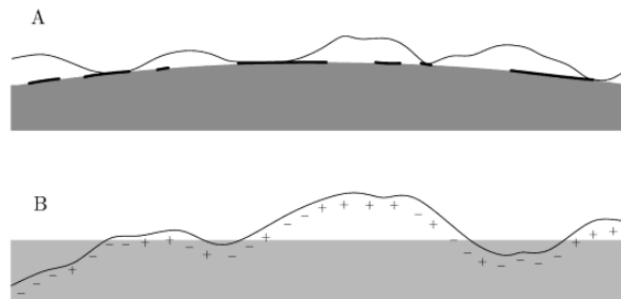


FIG. 1.6 – Polymères en interaction avec une membrane.

À travers la modélisation, on aimerait répondre aux questions suivantes, en particulier lorsque la taille du polymère est grande :

- pour un nombre de monomères fixé, quel est le nombre de configurations géométriques possibles pour le polymère ?
- quelle est la distance typique entre les extrémités du polymère ?
- lorsqu'il y a interaction, quelle est la proportion de monomères absorbés par la surface, la force nécessaire pour arracher la polymère à la surface ?

1.2.2 Comptage de chemins

On considère ici le cas d'une marche aléatoire en dimension un et on cherche à expliciter le nombre de chemins qui relient deux points donnés. On a ainsi la proposition et le théorème suivants :

Proposition 1.2.2. *Soient (m, a) et (n, b) deux couples d'entiers, avec $n > m$*

- *Si $n - m$ et $b - a$ n'ont pas même parité ou si $|b - a| > n - m$, alors il n'existe aucun chemin de (m, a) à (n, b) ;*
- *Si $|b - a| \leq n - m$ et si $n - m$ et $b - a$ ont même parité, alors le nombre de chemins de (m, a) à (n, b) est exactement*

$$\binom{n - m}{\frac{n - m}{2} + \frac{b - a}{2}}.$$

Démonstration. Les deux premiers points ne posent pas de problème : étant donné un chemin joignant les points $(m_0, a_0), (m_1, a_1), \dots, (m_r, a_r)$, il est clair que la parité de m_{i+1} est l'inverse de la parité de m_i , tout comme la parité de a_{i+1} est l'inverse de celle de a_i . Si la suite $m_0, m_1, \dots, m_r$ change un nombre pair de fois de parité, il en va de même pour la suite $a_0, a_1, \dots, a_r$ (idem dans le cas impair), et donc les extrémités (m, a) et (n, b) vérifient la propriété : $n - m$ et $b - a$ ont même parité. De même, si $|b - a| > n - m$, alors pour joindre (m, a) à (n, b) , il faudrait au moins $|b - a|$ montées (ou descentes), en seulement $n - m$ étapes. Enfin, si $|b - a| \leq n - m$ et $n - m$ et $b - a$ n'ont pas même parité, alors tout chemin joignant (m, a) et (n, b) est obtenu en “montant” $\frac{n - m}{2} + \frac{b - a}{2}$ fois et en “descendant” $\frac{n - m}{2} - \frac{b - a}{2}$ fois. $\square$

Théorème 1.2.3 (Principe de réflexion). *Soient a et b deux entiers strictement positifs, et $m < n$ deux entiers. Alors le nombre de chemins joignant (m, a) à (n, b) et touchant l'axe des abscisses est exactement le nombre de chemins joignant (m, a) à $(n, -b)$.*

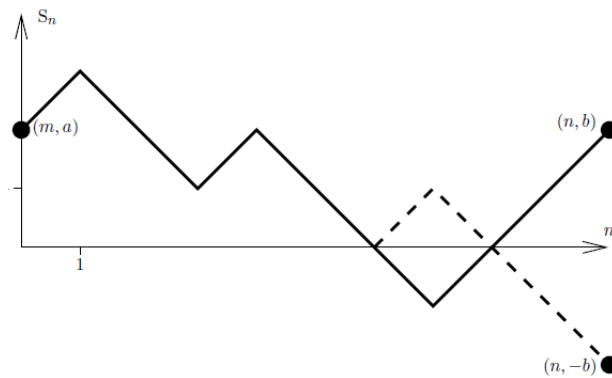


FIG. 1.9 – Le principe de réflexion.

Démonstration. La preuve de ce principe est assez simple : elle repose sur l'idée qu'un chemin joignant (m, a) à $(n, -b)$ passe nécessairement par la valeur 0. On peut alors établir une correspondance bijective entre les chemins de ces deux type par symétrie par rapport à l'axe des abscisses comme l'illustre la figure ci-dessus $\square$

1.2.3 Passage et temps de passage en zéro

Une question naturelle concernant la marche aléatoire est de connaître son comportement lorsque n devient grand. Si l'on pense aux modèles de la section 1.1, cela revient par exemple à se demander où se trouve la particule de pollen si on la laisse diffuser longtemps, ou encore quelle est la distance entre les deux extrémités d'un "grand" polymère. Pour simplifier les expressions, nous supposons ici que les marches aléatoires sont issues du point zéro. Lorsque la marche est asymétrique, par exemple lorsque $p := \mathbb{P}(X_i = 1) > \mathbb{P}(X_i = -1) = 1 - p$, la loi des grands nombres implique que S_n tend presque sûrement vers $+\infty$. En effet, on a

$$\frac{S_n}{n} = \frac{X_1 + \dots + X_n}{n} \xrightarrow{n \rightarrow +\infty} 2p - 1 = \mathbb{E}[X_1] > 0.$$

Dans le cas symétrique, la réponse est moins claire. Un premier pas vers la description du comportement asymptotique de la marche est de savoir si celle-ci passe plusieurs fois par le même point (ici zéro), et si oui, est-ce que cela est fréquent ? Grâce aux résultats du paragraphe précédent, nous sommes en mesure de répondre à ces questions. Une simple application de la proposition 1.2.2 et de la formule de Stirling donne tout d'abord :

Corollaire 1.2.4. *La probabilité pour que S_{2n} soit nul est*

$$\mathbb{P}(S_{2n} = 0) = \frac{\binom{2n}{n}}{2^{2n}} \approx \frac{1}{\sqrt{\pi n}}.$$

On souhaite maintenant préciser le temps qu'il faut attendre pour voir la marche repasser en zéro.

Définition 1.2.5. On définit l'instant de premier retour en 0 comme :

$$T_0 = \begin{cases} \inf\{n \geq 1, S_n = 0\}, & \text{si cet ensemble est fini,} \\ +\infty & \text{sinon.} \end{cases}$$

Grâce au principe de réflexion établi dans le théorème 1.2.3, nous pouvons calculer la probabilité de l'évènement $\{T_0 = 2n\}$:

Théorème 1.2.6. *La probabilité que le premier retour en 0 ait lieu à l'instant $2n$ est :*

$$\mathbb{P}(T_0 = 2n) = \frac{\binom{2n-2}{n-1} - \binom{2n-2}{n-2}}{2^{2n-1}} = \frac{(2n-2)!}{2^{2n-1}n!(n-1)!} \approx \frac{1}{\sqrt{4\pi n^3}}.$$

En outre, on a

$$\mathbb{P}(T_0 = 2n) = \mathbb{P}(S_{2n-2} = 0) - \mathbb{P}(S_{2n} = 0).$$

Démonstration. On raisonne conditionnellement au premier pas de la marche. Si $S_1 = 1$, alors il s'agit de calculer le nombre de chemins de $(1, 1)$ à $(2n, 0)$ qui ne touchent pas l'axe des abscisses avant l'étape $2n$. Mais le dernier pas est alors imposé : on a nécessairement $S_{2n-1} = 1$, puisque l'on cherche des chemins qui aboutissent à la valeur 0 sans avoir jamais touché l'axe des abscisses. Il s'agit donc de dénombrer le nombre de chemins de $(1, 1)$ à $(2n-1, 1)$ ne touchant pas l'axe des abscisses, c'est-à-dire le nombre total de chemins moins le nombre de chemins qui touchent l'axe. D'après le principe de réflexion, ce dernier est égal au nombre de chemins de $(1, 1)$ à $(2n-1, -1)$. Au total, on trouve : $\binom{2n-2}{n-1} - \binom{2n-2}{n-2}$. Le cas où $S_1 = -1$ se traite de façon analogue. $\square$

En sommant les probabilités ci-dessus, on obtient alors

Théorème 1.2.7. *Presque sûrement, la marche aléatoire symétrique unidimensionnelle issue de zéro revient en zéro, c'est-à-dire $\mathbb{P}(T_0 < +\infty) = 1$.*

Démonstration. En effet, la probabilité $\mathbb{P}(T_0 < +\infty)$ est donnée par la somme

$$\sum_{n \geq 1} \mathbb{P}(T_0 = 2n).$$

D'après le dernier point du théorème 1.2.6, cette somme est télescopique et l'on a

$$\mathbb{P}(T_0 < +\infty) = 1 - \lim_{n \rightarrow +\infty} \mathbb{P}(S_{2n} = 0).$$

Or d'après la formule de Stirling, on a

$$\mathbb{P}(S_{2n} = 0) = \frac{\binom{2n}{n}}{2^{2n}} = \frac{(2n)!}{2^{2n}(n!)^2} \approx \frac{2n^{2n}e^{-2n}\sqrt{4\pi n}}{2^{2n}(n^n e^{-n}\sqrt{2\pi n})^2} = \frac{1}{\sqrt{\pi n}} \xrightarrow{n \rightarrow +\infty} 0,$$

d'où le résultat. $\square$

Remarque 1.2.8. L'équivalent du théorème 1.2.6 indique que $\mathbb{E}[T_0]$, le temps d'attente moyen pour voir le premier retour en zéro, est infini. Autrement dit, presque sûrement la marche revient en zéro, mais il faut attendre un temps infini pour voir se réaliser cet événement.

1.2.4 Récurrence et transience de la marche simple

Nous venons de voir qu'en dimension un, la marche aléatoire symétrique issue de zéro revient presque sûrement en zéro. En fait, zéro ne joue pas de rôle particulier et ce résultat est vrai pour tout entier. On qualifie alors la marche de *récurrente*. À l'opposé, nous avons vu que, d'après la loi des grands nombres, une marche asymétrique s'en va presque sûrement vers $\pm\infty$. On dit alors que la marche est *transiente*. Les propriétés de transience et de récurrence de la marche aléatoire symétrique sont intimement liées à la dimension de l'espace. Nous avons ainsi le théorème (assez difficile) suivant.

Théorème 1.2.9. *La marche aléatoire symétrique sur $\mathbb{Z}^d$ est récurrente si la dimension d est égale 1 ou 2. Si $d \geq 3$, la marche aléatoire symétrique est transiente.*

1.3 Marche à boucles effacées et marche auto-évitante

Si l'on considère la marche aléatoire en dimension deux comme un modèle potentiel de polymère, les auto-intersections de la marche ne sont pas réalistes : cela voudrait dire que plusieurs monomères se trouvent en un même point de l'espace. Pour palier à ce défaut, de nombreux autres modèles de courbes simples ont été introduits dans la littérature. Nous citons ici les plus simples de ces modèles : la marche à boucles effacées et la marche auto-évitante.

1.3.1 La marche à boucles effacées

Comme son nom l'indique, la marche aléatoire à boucles effacées est obtenue à partir de la marche aléatoire simple symétrique dans $\mathbb{Z}^2$ introduite précédemment, en effaçant les éventuelles boucles par ordre chronologique d'apparition. Naturellement, la courbe obtenue après l'effacement des boucles est une courbe simple, sans auto-intersection.

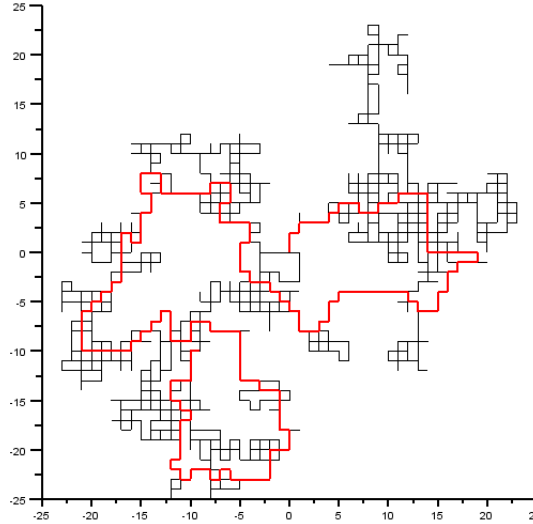


FIG. 1.10 – Exemple de trajectoire de marche aléatoire symétrique bidimensionnelle (en noir) et de sa version à boucles effacées (en rouge).

Voici l'algorithme qui à partir des coordonnées $(S_0, S_1, \dots, S_n)$ de la marche simple initiale fourni les coordonnées de son analogue à boucles effacées :

$$\begin{cases} \text{on initialise } i_0 := \max\{i, S_i = S_0\}, \text{ et tant que } S_{i_j} \neq S_n \\ \text{on pose } i_{j+1} = \max\{i, S_i = S_{i_j}\} + 1. \end{cases}$$

Dès lors, si J est le premier entier tel que $S_{i_J} = S_n$, la marche à boucles effacées est donnée par $LE(S) := (S_{i_0}, S_{i_0}, \dots, S_{i_J})$.

Remarque 1.3.1. En général, étant donnée une marche aléatoire simple S , la marche $LE(S)$ obtenue en effaçant les boucles dans l'ordre chronologique et la marche $\widehat{LE}(S)$ obtenue en effaçant les boucles dans l'ordre inverse diffèrent. Cependant, on a le théorème difficile suivant.

Théorème 1.3.2. *La loi de la marche $LE(S)$ parcourue en sens inverse est la même que la loi de la marche $\widehat{LE}(S)$.*

En travaux pratiques, nous verrons comment simuler une marche aléatoire simple symétrique, une marche à boucles effacées et nous essaierons en particulier de décrire leur rayon typique en fonction du nombre de pas de la marche.

1.3.2 La marche auto-évitante

La marche auto-évitante est une autre marche aléatoire dans $\mathbb{Z}^2$ dont les trajectoires n'ont pas d'auto-intersection. Pour fixer les idées, nous supposons ici que les marches sont issues de zéro. Sur le réseau $\mathbb{Z}^2$, l'ensemble Ω_n des chemins de longueur n , issus de zéro, et qui n'ont pas d'auto-intersection est un ensemble fini. Notons C_n son cardinal. Une réalisation de la marche auto-évitante issue de zéro est simplement un choix d'une telle trajectoire selon la probabilité uniforme. Autrement dit, toutes les trajectoires licites ont une chance $1/C_n$ d'être tirées au sort.

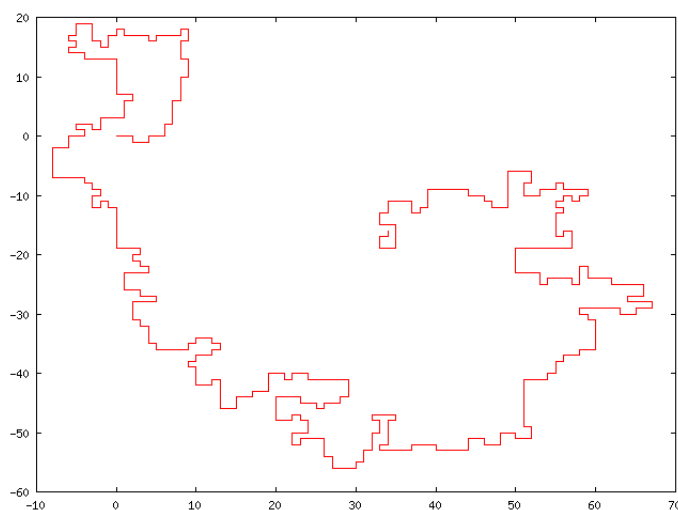


FIG. 1.11 – Exemple de trajectoire de marche auto-évitante ($n = 500$).

Un simple examen des cas où $n = 3, 4, 5, 6$ montre qu'il devient rapidement très difficile de dénombrer les chemins auto-évitant issus de zéro et de longueur n . À titre d'exemple sont regroupés dans le tableau ci-dessous le nombre de chemins auto-évitant sur une grille de $m \times n$ points, qui partent de $(0, 0)$ et finissent en (n, m) , et ce pour de petites valeurs de m et n . Pour $m = n = 6$, il y a déjà plus d'un million de tels chemins !

m / n	2	3	4	5	6
2	2				
3	4	12			
4	8	38	184		
5	16	125	976	8512	
6	32	414	5382	79384	1262816

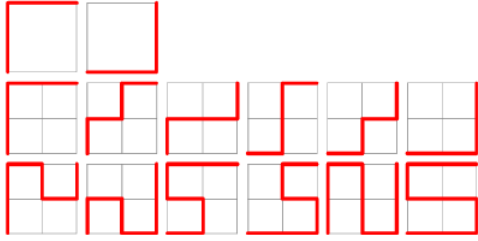


FIG. 1.12 – À droite, trajectoires auto-évitanes de $(0, 0)$ à (n, n) sur la grille carrée de $n \times n$ points, pour $n = 2, 3$.

En fait, on ne connaît pas de formule donnant le nombre C_n pour n fixé. Cependant, on peut montrer la majoration suivante : $C_{n+m} \leq C_n C_m$. En effet, le membre de droite compte le nombre de façons de concaténer une marche auto-évitante de longueur m à une marche auto-évitante de longueur n . Faisant cela, on obtient toutes les marches auto-évitanes de longueur $n + m$, mais aussi des marches qui s'intersectent, d'où l'inégalité. Par un argument sous-additif simple, on en déduit que C_n croît exponentiellement vite avec n , autrement dit la limite $\mu = \lim_{n \rightarrow \infty} C_n^{1/n}$ existe et elle est strictement positive. Cette limite est appelée *constante de connectivité*.

Pour se faire une idée de la difficulté de l'énumération des chemins auto-évitanes, notons que l'on ne connaît même pas la valeur de μ lorsque le réseau sous-jacent est le réseau carré. Dans les années 1980, dans le cas où le réseau sous-jacent est hexagonal (en nid d'abeille), des considérations heuristiques ont amené les physiciens à conjecturer la valeur suivante pour la constante de connectivité : $\mu = \sqrt{2 + \sqrt{2}}$. Des simulations numériques ont ensuite renforcé cette conjecture. Presque trente ans après la formulation de cet énoncé, deux mathématiciens sont récemment parvenu à en donner la preuve :

Théorème 1.3.3 (Duminil-Copin, Smirnov, 2010). *Sur le réseau hexagonal, la constante de connectivité est $\mu = \sqrt{2 + \sqrt{2}}$.*

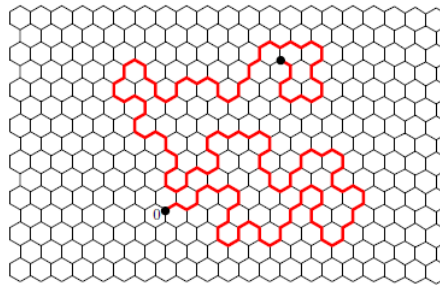


FIG. 1.13 – Marche auto-évitante sur le réseau hexagonal.

Chapitre 2

Limites d'échelle des marches aléatoires

Dans ce chapitre, nous cherchons à décrire la limite d'échelle des marches aléatoires introduites précédemment, c'est-à-dire à décrire la géométrie de très longues marches lorsqu'on les normalise de façon convenable. Il s'agit donc de passer de modèles discrets, définis sur l'ensemble $\mathbb{Z}$ des entiers ou sur le réseau $\mathbb{Z}^2$, à des modèles continus, définis sur la droite réelle $\mathbb{R}$ ou dans le plan $\mathbb{R}^2$. En passant à la limite, nous allons voir ainsi apparaître le mouvement brownien, qui est un objet central dans la théorie des probabilités, ainsi que les évolutions de Schramm-Loewner.

2.1 De la marche aléatoire au mouvement brownien

Considérons une marche aléatoire symétrique sur $\mathbb{Z}^2$ issue de $S_0 = 0$. Elle saute d'une amplitude de l'ordre d'une unité d'espace pendant chaque unité de temps. Si l'on souhaite modéliser le déplacement de particules de pollen la surface de l'eau, ce modèle est inadéquat au sens où les mouvements observés sont très erratiques et semblent varier presque continûment (*i.e.* à tout instant). Afin d'améliorer le modèle initial, nous allons effectuer des changements d'échelles d'espace et de temps qui vont permettre, à partir d'une marche aléatoire, de définir un nouvel objet mathématique décrivant de façon très fidèle de mouvement des particules de pollen.

2.1.1 Convergence de la marche renormalisée

Fixons un intervalle de temps $[0, t]$. Nous allons modifier la marche aléatoire de sorte qu'elle saute tous les temps multiples de $1/n$ au lieu de tous les temps unités. Il y aura donc $\lfloor nt \rfloor$ sauts sur l'intervalle $[0, t]$, où $\lfloor nt \rfloor$ désigne la partie entière de nt . Cela revient simplement à considérer la marche aléatoire $S_{\lfloor nt \rfloor} = \sum_{k=1}^{\lfloor nt \rfloor} X_k$ et notre but est d'en connaître le comportement asymptotique, quand n tend vers l'infini. Remarquons que comme la marche saute tous les temps $1/n$, cette limite rend les changements de position très rapides.

Comme nous l'avons observé en travaux pratiques, une marche aléatoire symétrique issue de zéro se trouve à une distance de l'ordre de $\sqrt{n}$ de l'origine après n pas. Ceci nous indique que le bon changement d'échelle spatial pour la marche consiste à la renormaliser par $\sqrt{n}$. Autrement dit, nous allons considérer la nouvelle marche

$$S_t^n := \frac{1}{\sqrt{n}} S_{[nt]}.$$

Les figures ci-après montrent deux trajectoires de la marche normalisée en temps et en espace, en dimension un et deux.

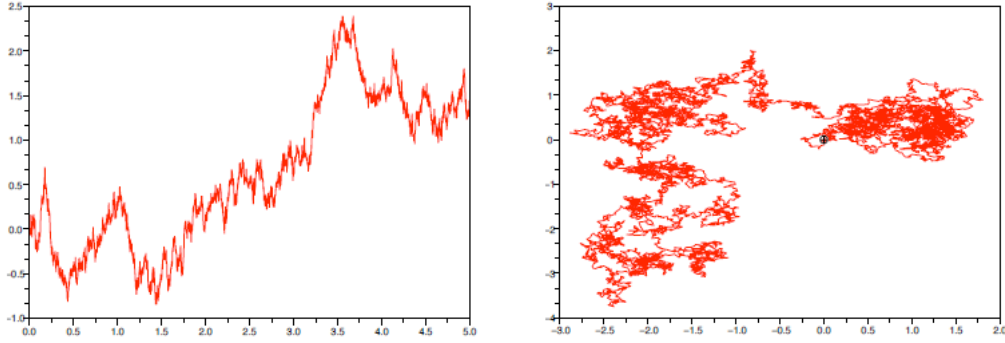


FIG. 2.1 – Trajectoires de la marche normalisée en dimensions un et deux.

La marche renormalisée semble converger vers un processus à trajectoires continues en temps, et très fortement irrégulier. La continuité de la limite de $t \mapsto S_t^n$ n'est pas surprenante puisque l'amplitude des sauts de S^n vaut $1/\sqrt{n}$ et tend donc vers 0 quand n tend vers l'infini. D'après le théorème limite central, pour tout t fixé, on a naturellement :

Proposition 2.1.1.

$$S_t^n := \frac{1}{\sqrt{n}} S_{[nt]} \longrightarrow \mathcal{N}(0, t)$$

où $\mathcal{N}(0, t)$ désigne une variable aléatoire de loi normale centrée et de variance t .

Notons par ailleurs, en reprenant les notations du chapitre précédent, que la marche aléatoire initiale vérifie les deux propriétés fondamentales suivantes :

- pour tous n et k , $S_{n+k} - S_n$ est indépendant de S_n ;
- pour tous n et k , $S_{n+k} - S_n$ a même loi que S_k .

On dit que les accroissements de la marche sont indépendants et stationnaires. On a en fait beaucoup mieux que la convergence ponctuelle, c'est-à-dire t par t , de la marche normalisée. En effet, considérons une suite finie de temps $t_1 < t_2 < \dots < t_p$. Nous savons que les variables $S_{t_1}^n, S_{t_2}^n - S_{t_1}^n, \dots, S_{t_p}^n - S_{t_{p-1}}^n$ sont indépendantes, et de

plus on peut écrire :

$$\begin{aligned} S_{t_j}^n - S_{t_{j-1}}^n &= \frac{1}{\sqrt{n}} (X_{\lfloor nt_{j-1} \rfloor + 1} + \dots + X_{\lfloor nt_j \rfloor}) \\ &= \frac{\sqrt{\lfloor nt_j \rfloor - \lfloor nt_{j-1} \rfloor}}{\sqrt{n}} \times \frac{(X_{\lfloor nt_{j-1} \rfloor + 1} + \dots + X_{\lfloor nt_j \rfloor})}{\sqrt{\lfloor nt_j \rfloor - \lfloor nt_{j-1} \rfloor}}. \end{aligned}$$

Lorsque n tend vers l'infini, le premier terme du membre de droite tend vers la constante $\sqrt{t_j - t_{j-1}}$ et le deuxième vers une variable aléatoire normale centrée réduite $\mathcal{N}(0, 1)$. Le produit converge donc vers une variable $\mathcal{N}(0, t_j - t_{j-1})$. Les accroissements étant indépendants, les variables $S_{t_1}^n, S_{t_2}^n - S_{t_1}^n, \dots, S_{t_p}^n - S_{t_{p-1}}^n$ convergent donc en loi vers des gaussiennes indépendantes $\mathcal{N}(0, t_1), \mathcal{N}(0, t_2 - t_1), \dots, \mathcal{N}(0, t_p - t_{p-1})$ lorsque n tend vers l'infini. Autrement dit, on a le résultat suivant :

Proposition 2.1.2. *Pour tous temps $t_1 < t_2 < \dots < t_p$, lorsque n tend vers l'infini, le vecteur $(S_{t_1}^n, S_{t_2}^n - S_{t_1}^n, \dots, S_{t_p}^n - S_{t_{p-1}}^n)$ converge en loi vers un vecteur aléatoire $(B_{t_1}, B_{t_2} - B_{t_1}, \dots, B_{t_p} - B_{t_{p-1}})$ où pour tout $s < t$, $B_t - B_s$ est une variable aléatoire de loi normale centrée de variance $t - s$ indépendante de B_s .*

La limite de la marche normalisée S_t^n lorsque n tend vers l'infini est donc une famille $(B_t)_{t \geq 0}$ de variables aléatoires, indexée par les temps $t \in \mathbb{R}^+$.

Définition 2.1.3. On appelle processus aléatoire $X = (X_t)_{t \geq 0}$ une famille de variables aléatoires indexée par $\mathbb{R}^+$, toutes ces variables étant définies sur le même espace de probabilités. On suppose ici que chaque X_t est valeurs réelles.

Remarque 2.1.4. Un processus aléatoire peut être vu de manière équivalente comme la donnée d'une variable aléatoire X définie sur un espace de probabilité et à valeurs dans l'ensemble des fonctions $t \mapsto x_t$ de $\mathbb{R}^+$ dans $\mathbb{R}$.

On peut ainsi reformuler la proposition 2.1.2 de la façon suivante : lorsque n tend vers l'infini, la marche normalisée $(S_t^n)_{t \geq 0}$ converge (au sens des marginales de dimension finie) vers un processus aléatoire à accroissements indépendants et stationnaires.

2.1.2 Mouvement brownien, premières propriétés

Le processus $(B_t)_{t \geq 0}$ ci-dessus, obtenu comme limite d'échelle de la marche aléatoire symétrique est appelé mouvement brownien.

Définition 2.1.5. Un processus stochastique $(B_t)_{t \geq 0}$ valeurs dans $\mathbb{R}$ est appelé mouvement brownien (standard) si les trois conditions suivantes sont satisfaites :

1. $B_0 = 0$;
2. $(B_t)_{t \geq 0}$ est accroissements indépendants et stationnaires ;
3. pour tout $t \geq 0$, la variable aléatoire B_t suit la loi normale $\mathcal{N}(0, t)$.

On peut montrer que les trajectoires du mouvement brownien sont continues (la particule ne saute pas) mais qu'elles ne sont dérivables en aucun point (elles sont très irrégulières). Par définition, B_t suit la loi $\mathcal{N}(0, t)$, sa densité est donnée par

$$p(t, x) = \frac{1}{\sqrt{2\pi t}} e^{-x^2/(2t)}.$$

Cette fonction est solution de l'équation aux dérivées partielles appelée équation de la chaleur :

$$\frac{\partial p}{\partial t}(t, x) = \frac{1}{2} \frac{\partial^2}{\partial x^2} p(t, x).$$

La figure 2.2 illustre l'évolution de la loi de B_t au cours du temps.

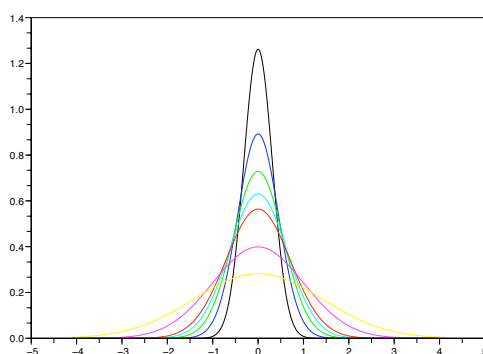


FIG. 2.2 – Evolution de la loi de B_t pour $t = 0.1, 0.2, 0.3, 0.4, 0.5, 1, 2$.

Pour obtenir un mouvement brownien dans $\mathbb{R}^2$ il suffit de prendre deux mouvements browniens indépendants $B^{(1)}$ et $B^{(2)}$ et considérer $B_t = (B_t^{(1)}, B_t^{(2)})$.

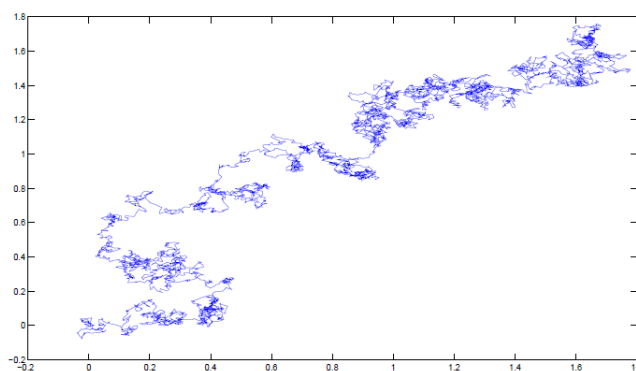


FIG. 2.3 – Trajectoire brownienne en dimension 2.

Remarque 2.1.6. Nous avons défini le mouvement brownien comme la limite de la marche aléatoire simple et symétrique. On montre en fait qu'il est la limite d'une très large classe de marches aléatoires ou processus discrets. À l'image de la loi gaussienne qui est la limite universelle de sommes de variables aléatoires (TLC), le mouvement brownien est la limite universelle de trajectoires discrètes aléatoires.

Proposition 2.1.7 (symétrie, invariance d'échelle, retournement du temps). *Soit $c > 0$. Si $(B_t, t \geq 0)$ est un mouvement brownien, alors les processus*

$$(-B_t)_{t \geq 0}, \quad \left(\frac{1}{c} B_{c^2 t}\right)_{t \geq 0} \quad \text{et} \quad (t B_{1/t})_{t \geq 0}$$

sont aussi des mouvements browniens.

Démonstration. Le deux premiers points sont clairs si l'on pense au mouvement brownien comme la limite d'échelle de la marche aléatoire symétrique : premièrement, la loi de la marche simple est invariante par réflexion le long de l'axe des abscisses, c'est donc aussi le cas de sa limite d'échelle ; deuxièmement, la limite d'échelle est par définition un point fixe du changement d'échelle. Le troisième point est plus délicat, on note cependant qu'à t fixé, $t B_{1/t}$ suit une loi $\mathcal{N}(0, t)$ comme B_t . $\square$

La propriété d'invariance d'échelle, qui est évidente si l'on définit le mouvement brownien comme limite d'échelle, est aussi appelée propriété auto-similarité. Elle justifie l'aspect fractal des trajectoires browniennes.

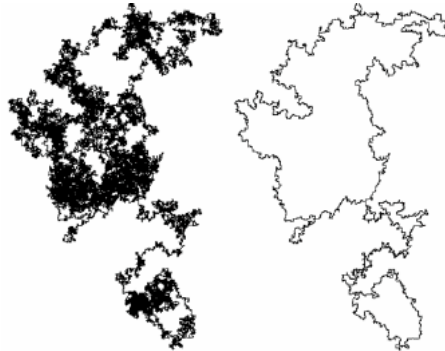


FIG. 2.4 – Aspect fractal de l'enveloppe extérieure de la courbe brownienne.

Une des façons de quantifier l'irrégularité d'une courbe est de calculer sa dimension de Hausdorff. Cette dimension est un nombre entre zéro et un : plus la courbe est irrégulière, plus sa dimension est grande.

Théorème 2.1.8 (Lawler, Schramm, Werner, 2001). *La dimension de Hausdorff de la frontière extérieure du mouvement brownien plan est égale à $4/3$.*

Le mouvement brownien hérite des nombreuses propriétés des marches aléatoires dont il est la limite. En particulier, on peut montrer le théorème suivant, analogue du théorème 1.2.9 du cas discret.

Théorème 2.1.9. *Le mouvement brownien est récurrent en dimension un et deux. En dimension trois et plus, le mouvement brownien est transient.*

2.2 Limite d'échelle des autres marches

Comme nous l'avons vu en travaux pratiques, la marche aléatoire simple symétrique en dimension deux et les marches aléatoires à boucles effacées et auto-évitante n'ont pas le même comportement asymptotique. Par exemple, nous avons constaté que la distance typique entre les extrémités d'une marche à boucles effacées de longueur n est de l'ordre de $n^{4/5}$, alors qu'elle est de l'ordre de $\sqrt{n}$ pour la marche aléatoire simple. Par ailleurs, la courbe brownienne, limite de la marche aléatoire simple renormalisée, a de nombreuses auto-intersections ce qui n'est pas le cas des marches à boucles effacées et auto-évitante. Ceci nous indique les limites d'échelle de ces dernières marches, si elles existent, ne sont pas le mouvement brownien mais d'autres processus aléatoires à préciser. La détermination de ces limites d'échelle est récente : elle remonte au début des années 2000 et aux travaux de Schramm, Lawler, Werner et Smirnov.

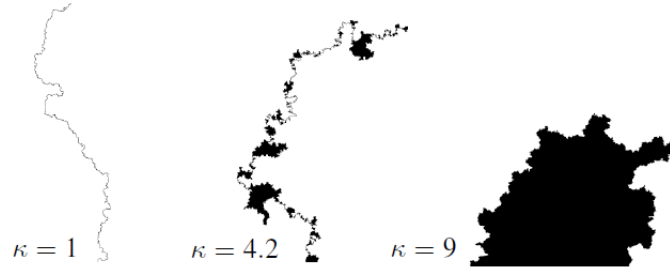
2.2.1 Évolution de Schramm-Loewner

L'objet mathématique qui décrit la limite d'échelle des marches aléatoires à boucles effacées et des marches auto-évitantes porte le nom d'évolution de Schramm-Loewner, que l'on abrège souvent en SLE pour Schramm-Loewner-Evolution. Il s'agit en fait d'une famille à un paramètre de processus SLE_κ , où le paramètre κ est un nombre positif. Pour κ fixé, le processus SLE_κ est une famille croissante $(K_t)_{t \geq 0}$ d'ensembles compacts du plan. On peut montrer que ces ensembles compacts K_t sont engendrés par une courbe aléatoire γ_t elle aussi tracée dans le plan, et appelée *trace*. Plus précisément, le compact K_t peut être vu comme le remplissage de la courbe γ sur l'intervalle $[0, t]$, lorsque celle-ci a des points doubles. Le théorème suivant illustre les liens complexes entre la famille de compacts K_t et la trace γ_t .

Théorème 2.2.1. *Soit $(K_t)_{t \geq 0}$ un SLE_κ et $(\gamma_t)_{t \geq 0}$ sa trace. Alors presque sûrement :*

- Si $0 \leq \kappa \leq 4$, alors la courbe γ est simple et $K_t = \gamma([0, t])$;
- Si $4 < \kappa < 8$, alors la courbe a des points doubles, et $\gamma([0, t]) \subsetneq K_t$;
- Si $\kappa \geq 8$, alors la courbe γ_t remplit l'espace et $K_t = \gamma([0, t])$.

La figure ci-après représente des réalisations de processus SLE_κ , pour un temps t fixé, et différentes valeurs de κ .

FIG. 2.5 – Réalisations de SLE_κ pour $\kappa = 1, 4.2, 9$.

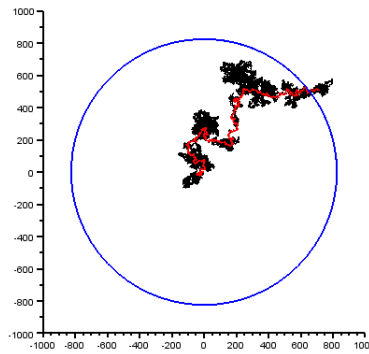
On peut constater sur les images ci-dessus que plus le paramètre κ est grand, plus la courbe γ_t est irrégulière. Comme précédemment, on peut quantifier cette irrégularité grâce à la notion de dimension de Hausdorff. On montre ainsi le théorème (très difficile) suivant :

Théorème 2.2.2 (Beffara, 2004-2008). *Considérons $\kappa > 0$ et (K_t) un SLE_κ de trace $(\gamma_t)_{t \geq 0}$. Si on pose $\mathcal{H} := \gamma([0, +\infty[)$, alors*

$$\dim_H(\mathcal{H}) = \min \left(2, 1 + \frac{\kappa}{8} \right).$$

2.2.2 Convergence de la marche à boucles effacées

Nous avons constaté en travaux pratiques que la taille typique d'une marche à boucles effacées de longueur n est $n^{4/5}$. La figure ci-dessous illustre cette propriété.

FIG. 2.6 – Marche à boucles effacées avec $n = 5000$ et cercle de rayon $n^{4/5}$.

On peut en fait établir rigoureusement cette propriété et montrer bien plus, à savoir que la limite d'échelle de la marche aléatoire à boucles effacées est un SLE_κ , pour la valeur du paramètre κ égale à 2.

Théorème 2.2.3. *Sur le réseau $\mathbb{Z}^2$, la limite d'échelle de la marche aléatoire à boucles effacées est un processus SLE_2*

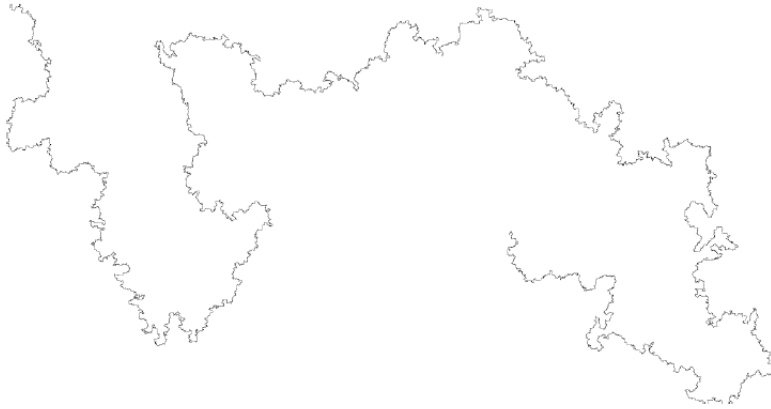


FIG. 2.7 – Simulation d'une trajectoire de SLE_2

2.2.3 Convergence de la marche auto-évitante

En simulant des trajectoires de la marche auto-évitante de longueur n (il existe des algorithmes complexes pour cela), on s'aperçoit que la taille typique d'un tel chemin est $n^{3/4}$. La normalisation à effectuer avant de passer à la limite d'échelle est donc encore différente de celle effectuée pour la marche simple ($\sqrt{n}$) ou pour la marche à boucles effacées ($n^{4/5}$).

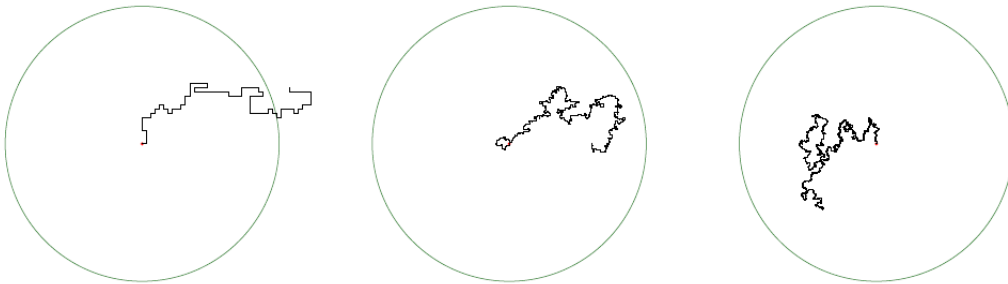


FIG. 2.8 – Simulation de marches auto-évitantes de longueur $n = 100, 1000, 10000$. Le cercle représenté est le cercle centré en l'origine et de rayon $n^{3/4}$.

Pour de nombreuses (et bonnes) raisons, les physiciens théoriciens et les mathématiciens sont convaincus que la limite d'échelle de la marche aléatoire auto-évitante est elle aussi décrite par une évolution de Schramm-Loewner, pour une valeur de

κ égale à $8/3$. Cependant, il n'existe pas à ce jour de preuve rigoureuse d'une telle convergence.

Conjecture 2.2.4. Sur le réseau $\mathbb{Z}^2$, la limite dééchelle de la marche à boucles effacées est un processus $\text{SLE}_{8/3}$

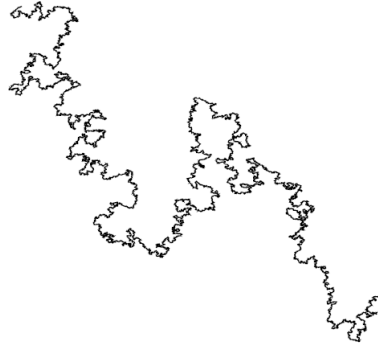


FIG. 2.9 – Simulation d'une très longue marche auto-évitante de longueur $n = 10^6$.

Chapitre 3

Notions de calcul stochastique

De très nombreux phénomènes naturels, au moins en première approximation, peuvent être décrits par des équations différentielles. Cependant, ces mêmes phénomènes sont souvent sujets à d'infimes perturbations aléatoires qui peuvent avoir à plus ou moins long terme, une influence notable. On peut penser par exemple à la trajectoire d'une fusée, soumise aux aléas de la traversée de l'atmosphère terrestre. Au chapitre précédent, nous avons introduit la notion de mouvement brownien à valeurs dans $\mathbb{R}$ ou plus généralement à valeurs dans $\mathbb{R}^d$. Comme nous l'avons déjà mentionné, le mouvement brownien est une fonction aléatoire universelle, une sorte de bruit universel, tout comme les variables gaussiennes jouent le rôle de variables aléatoires universelles via le théorème limite central.

3.1 Les équations différentielles stochastiques

Les équations différentielles stochastiques sont des équations différentielles dans lesquelles on fait intervenir un terme aléatoire : un “bruit” brownien. Elle permettent de prendre en compte, dans la description des phénomènes naturels, les perturbations évoquées plus haut. L'objet de cette section est d'introduire ce type d'équations. Nous donnerons ensuite quelques propriétés élémentaires de leurs solutions.

3.1.1 Rappel sur les équation différentielles ordinaires

Commençons par quelques rappels concernant les équations différentielles ordinaires. Si b est une fonction régulière, l'équation différentielle

$$\begin{cases} X'(t) = b(X(t)) & \text{si } t \geq 0, \\ X(0) = x, \end{cases} \quad i.e. \quad \begin{cases} dX(t) = b(X(t))dt & \text{si } t \geq 0, \\ X(0) = x, \end{cases}$$

admet une unique solution lorsque la valeur initiale x est donnée. La fonction b fait dériver X le long de trajectoires déterministes. Les figures suivantes décrivent par exemple l'allure des solutions lorsque $X(0) = 0$ et $b \equiv 1$ puis $X(0) > 0$ et $b(x) = x$, celles-ci étant données par $X(t) = t$ et $X(t) = X(0)e^t$ respectivement.

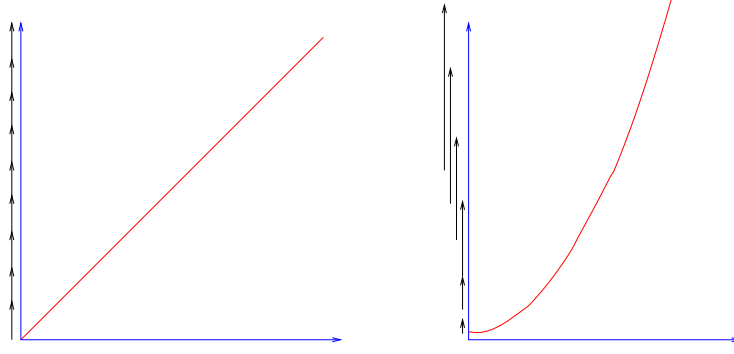


FIG. 3.1 – Flots associés aux équations $X' = b(X)$ lorsque $b = 1$ et $b(x) = x$.

Remarque 3.1.1. L'équation précédente $dX(t) = b(X(t))dt$ signifie qu'au temps t , l'accroissement de $X(t)$ est proportionnel à $b(X(t))$. Plus précisément, si h est un élément infinitésimal, on a

$$X(t+h) - X(t) = b(X(t)) \times h.$$

3.1.2 Équations différentielles stochastiques

Comme annoncé au début du chapitre, afin de décrire des phénomènes naturels soumis à des aléas, on souhaite construire des processus stochastiques qui soient à la fois soumis à une dérive déterministe ainsi qu'à de la diffusion. De tels processus sont les solutions d'équations appelées *équations différentielles stochastiques*, et qui s'écrivent sous la forme :

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t. \quad (3.1)$$

La notation dB_t fait ici référence à l'accroissement du mouvement brownien introduit au chapitre précédent. Attention cependant, dB_t/dt n'a pas de sens a priori puisque la trajectoire brownienne est nulle part dérivable. On peut malgré tout donner un sens mathématique rigoureux à l'équation (3.1) : ceci représente un travail assez conséquent et hors de portée dans le cadre de ce cours. Donnons tout de même quelques intuitions sur l'aspect des trajectoires de ce type de processus :

- en un point x , la solution d'une telle équation subit localement une dérive d'intensité $b(x)$ et une diffusion proportionnelle à $\sigma(x)$;
- plus $b(x)$ est grand par rapport à $\sigma(x)$ et plus la trajectoire est proche de la trajectoire déterministe ;
- plus $\sigma(x)$ est grand par rapport à $b(x)$ et plus la trajectoire est erratique.

Remarque 3.1.2. Moralement, l'équation (3.1) signifie que si h est un élément infinitésimal, on a

$$X(t+h) - X(t) = b(X(t)) \times h + \underbrace{\sigma(X(t)) (B_{t+h} - B_t)}_{\sim \mathcal{N}(0,h)}.$$

Les deux graphiques ci-après, à comparer avec les graphiques de la page précédente, représentent par exemple des trajectoires des solutions respectives des équations

$$dX_t = dt + dB_t, \quad X(0) = 0, \quad \text{et} \quad dX_t = X_t dt + dB_t, \quad X(0) > 0.$$

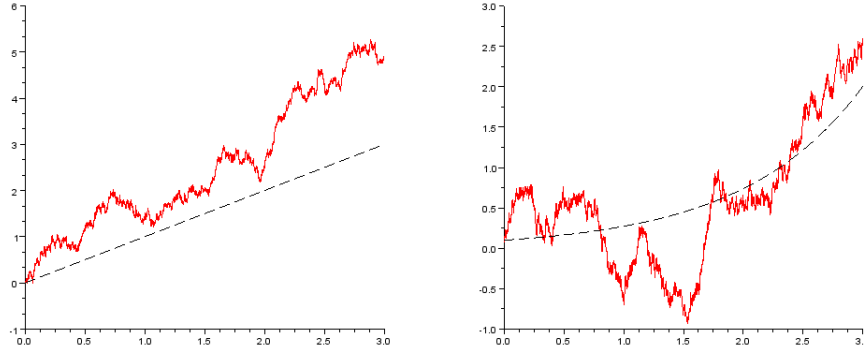


FIG. 3.2 – Trajectoires associées aux équations $dX_t = b(X_t)dt + \sigma(X_t)dB_t$ lorsque $\sigma \equiv 1$ et $b = 1$ puis $b(x) = x$.

Le graphique ci-dessous illustre l'influence de la fonction σ . Il représente en effet la solution de l'équation $dX_t = dt + \sin(X_t)dB_t$, $X(0) = 0$. On note que la courbe devient plus régulière lorsque l'ordonnée est proche d'un multiple de π , c'est-à-dire lorsque $\sin(X_t)$ devient petit par rapport à $b \equiv 1$.

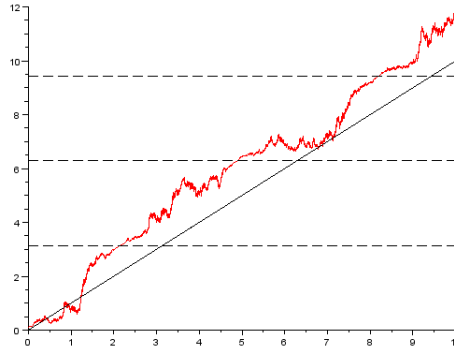


FIG. 3.3 – Trajectoire d'une solution de l'équation $dX_t = dt + \sin(X_t)dB_t$ issue de $X(0) = 0$.

3.2 Propriétés élémentaires des diffusions

Nous donnons à présent quelques propriétés élémentaires des diffusions, *i.e.* des processus stochastiques solutions des équations différentielles stochastiques introduites au paragraphe précédent. La plupart des résultats seront ici admis.

3.2.1 Temps de séjour dans un intervalle

Il est possible de décrire parfaitement le comportement d'une diffusion réelle dans un intervalle donné en fonction des coefficients σ et b . On se donne un intervalle ouvert $I =]l, r[$ de $\mathbb{R}$, l ou r étant éventuellement infinis et une diffusion $(X_t)_{t \geq 0}$ solution de l'équation différentielle stochastique :

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t, \quad X(0) \in I.$$

Définition 3.2.1. On définit la fonction d'échelle 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 m est alors 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 3.2.2. Dans les expressions ci-dessus, la fonction $\mathbf{1}_{\{x \in I\}}$ est la fonction qui vaut 1 si $x \in I$ et 0 sinon. Notons que les définitions de la fonction d'échelle S et de la mesure de vitesse m dépendent du point $c \in I$ choisi de manière arbitraire. Cependant les formules intéressantes qui mettront en jeu ces fonctions ne dépendront pas du choix de c .

Les deux quantités S et m permettent d'exprimer les probabilités de sortie d'un intervalle compact inclus dans I , l'espérance de ce temps ainsi que de nombreuses autres quantités concernant la diffusion.

Théorème 3.2.3. Si $l < a < x < b < r$, $X_0 = x$ et $T_{a,b} = \inf \{t \geq 0, X_t \notin]a, b[\}$ alors

$$\mathbb{E}_x(T_{a,b}) = - \int_a^x (S(x) - S(y))m(dy) + \frac{S(x) - S(a)}{S(b) - S(a)} \int_a^b (S(b) - S(y))m(dy). \quad (3.2)$$

De plus,

$$\mathbb{P}_x(X_{T_{a,b}} = a) = \frac{S(b) - S(x)}{S(b) - S(a)}, \quad \text{et} \quad \mathbb{P}_x(X_{T_{a,b}} = b) = \frac{S(x) - S(a)}{S(b) - S(a)}. \quad (3.3)$$

Exemple 3.2.4. Si $b = 0$ et $\sigma = 1$, X est le mouvement brownien standard. On a alors

$$S(x) = x - c \quad \text{et} \quad m(dx) = 2\mathbf{1}_{\{x \in I\}} dx.$$

On en déduit que $\mathbb{E}_x(T_{a,b}) = (x - a)(b - x)$ et que

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

Remarque 3.2.5. En remplaçant a par zéro, b par N et x par k , le résultat ci-dessus est le parfait analogue de la probabilité de la ruine du joueur envisagée dans le premier exercice de la série 1. Nous avons en effet montré que

$$\mathbb{P}(\text{partant de } k, \text{ la trajectoire touche } 0 \text{ avant } N) = 1 - \frac{k}{N} = \frac{N - k}{N - 0}.$$

3.2.2 Diffusions à coefficients singuliers

Le mouvement brownien unidimensionnel étant récurrent dans $\mathbb{R}$, il sort presque sûrement de tout intervalle borné. D'autres processus de diffusions restent en revanche "bloqués" dans un intervalle $I =]l, r[$, ou sont "absorbés" en un point du bord. Ce type de situation apparaît lorsque les fonctions b et σ ont des singularités, *i.e.* sont irrégulières aux extrémités de l'intervalle I , mais où l'on peut encore donner un sens à l'équation (3.1). Nous souhaitons maintenant décrire ces situations plus complexes. Considérons par exemple les deux équations différentielles stochastiques suivantes :

$$dX_t = dB_t + \left(\frac{a}{X_t} - \lambda X_t \right) dt, \quad X(0) = x > 0,$$

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

$$dX_t = \sqrt{1 - X_t^2} dB_t - (n - 1)X_t dt, \quad X(0) = x \in]-1, 1[.$$

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.

Pour toute la suite, on fixe un intervalle $I =]l, r[$ avec $-\infty \leq l < r \leq +\infty$. On définit la variable aléatoire $T := \inf \{t \geq 0; \quad X_t \notin]l, r[\}$ appelée temps de sortie de I . Sous l'hypothèse $X_0 \in I$, on a $\mathbb{P}(T > 0) = 1$.

On suppose également que les coefficients $\sigma : I \rightarrow \mathbb{R}$ et $b : I \rightarrow \mathbb{R}$ vérifient les conditions suivantes :

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

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

La proposition suivante relie le comportement de la fonction d'échelle aux bords de l'intervalle à celui de la diffusion sous-jacente.

Proposition 3.2.6. *Supposons que (ND) et (LI) sont vérifiées et soit X une solution faible de l'équation (3.1) avec une 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 $y \in I$,

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

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. On donne la preuve du premier point. D'après (3.3), si a et b sont deux réels tels que $l < a < x < b < r$, 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 presque sûrement. 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 évaluer la probabilité mettant en jeu le supremum de X . Si l'on suppose 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 3.2.7. 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.

Nous caractérisons maintenant le comportement de la diffusion (X_t) au voisinage du bord de l'intervalle $I =]l, r[$. Pour cela, on introduit deux nouvelles fonctions qui s'exprime à l'aide de la fonction d'échelle S et de la mesure vitesse m .

Définition 3.2.8. Pour $c \in I$, on définit les fonctions suivantes

$$u(x) := \int_c^x m(u)S'(u) du \quad \text{et} \quad v(x) := \int_c^x m'(u)S(u) du.$$

Remarque 3.2.9. Là encore, les fonctions u et v dépendent du point $c \in I$ choisi de manière arbitraire, mais nous ne sommes intéressé que par leur comportement qualitatif qui lui ne dépend pas du choix de c .

Le comportement de la diffusion $(X_t)_{t \geq 0}$ au voisinage des extrémités de l'intervalle $I =]l, r[$ est dicté par celui des fonctions u et v . On a en effet le théorème suivant :

Théorème 3.2.10. *Les fonctions u et v admettent des limites aux bords de I (éventuellement infinies). De plus, soit s une extrémité de I alors le statut de s est dicté par les limites de u et v en s :*

$u(s)$	$v(s)$	<i>accessible ?</i>	<i>absorbant ?</i>
$< \infty$	$< \infty$	<i>oui</i>	<i>non</i>
$< \infty$	$= \infty$	<i>oui</i>	<i>oui</i>
$= \infty$	$< \infty$	<i>non</i>	<i>non</i>
$= \infty$	$= \infty$	<i>non</i>	<i>oui</i>

3.3 Simulation de diffusions

On souhaite avoir une idée de l'allure des trajectoires de X solution de (3.1) ou plus simplement estimer la quantité $\mathbb{E}(f(X_t))$ où f est une fonction connue. On ne sait en général pas simuler une variable aléatoire de même loi que X_t . L'idée est de procéder comme pour les équations différentielles ordinaires et d'utiliser la remarque 3.1.2.

3.3.1 Simulation de trajectoires d'une diffusion

On introduit ainsi un pas de discrétisation du temps $h > 0$ et on se donne une suite $(W_n)_{n \geq 0}$ de variables aléatoires indépendantes de même loi $\mathcal{N}(0, 1)$. On définit la suite $(Y_n)_{n \geq 0}$ par

$$Y_{n+1} = Y_n + b(Y_n)\alpha + \sigma(Y_n)\sqrt{h}W_n.$$

Lorsque h est assez petit, Y_n a quasiment la même loi que X_{nh} .

3.3.2 Estimation de fonctionnelles de la diffusion

En d'autres termes, $\mathbb{E}(f(X_{nh}))$ est proche de $\mathbb{E}(f(Y_n))$.

Deuxième partie

Éléments de génétique des
populations

Chapitre 4

Autour du modèle de Wright-Fisher

L’objet de ce chapitre et du chapitre suivant est de modéliser des phénomènes complexes intervenant en génétique des populations. Précisément, on s’intéresse à l’évolution, au cours du cycle reproductif, de la fréquence des allèles dans une population de taille finie. Dans ce cadre, l’introduction de modèles aléatoires est motivée d’une part par la grande complexité du cycle reproductif qui fait qu’une modélisation déterministe aboutirait à un système non résoluble en pratique et d’autre part le fait des phénomènes aléatoires sont intrinséquement mis en jeu au cours de la reproduction (par exemple, les mutations des allèles).

4.1 Ce que l’on cherche à modéliser

Comme nous l’avons fait dans la première partie du cours, nous commençons par décrire le problème concret que l’on souhaite modéliser. Pour ce faire, rappelons quelques points de vocabulaire en rapport avec la génétique :

- une cellule biologique est dite *haploïde* lorsque les chromosomes qu’elle contient sont chacun en un seul exemplaire. Le concept est généralement opposé à *diploïde*, terme désignant les cellules avec des chromosomes en double ;
- un *gamète* est une cellule reproductrice de type haploïde. Chez les humains et la plupart des animaux, la reproduction génétique met en jeu une succession de phases diploïdes et haploïde (phase de formation des gamètes) ;
- un *locus* est un emplacement sur le chromosome. Il peut contenir un gène ;
- les *allèles* sont des versions différentes de l’information génétique codée sur un locus. Par exemple, les types “rids” et “lisses” des pois, dans les expériences de Mendel, correspondent à des allèles distincts. Dans le cas du gène déterminant le groupe sanguin, l’un des allèles code le groupe A, un autre pour le groupe B, et un troisième allèle détermine le groupe O ;
- l’*avantage sélectif* d’un allèle est une mesure qui caractérise son aptitude à se transmettre, et qui dépend ainsi de l’aptitude qu’il confère à son porteur à se reproduire, survivre...

4.1.1 La problématique

Nous cherchons ici à modéliser la reproduction des individus dans une population de taille finie, en particulier nous cherchons à décrire la transmission de leurs allèles et l'évolution des fréquences alléliques au cours du temps. Commençons par mettre en évidence quelques difficultés que soulève cette modélisation.

Dans une population haploïde, la proportion des individus portant un allèle donné fournit directement la fréquence de cet allèle dans la population. Mais dans les populations diploïdes, les gènes sont associés par paires. Ainsi, les fréquences génotypiques ne peuvent pas en général se déduire des fréquences alléliques. Supposons par exemple que l'on a deux allèles A et B. Une population avec 50% de gènes A et 50% de gènes B peut-être constituée uniquement d'homozygotes AA et BB, ou uniquement d'hétérozygotes AB, ou de diverses proportions entre ces 3 génotypes. Par exemples, les deux populations ci-après de trois individus, ont les mêmes fréquences alléliques : (AA, AB, BB), (AB, AB, AB).

Dans une situation concrète, le cycle reproductif n'est pas synchronisé : plusieurs générations co-existent et interagissent.

La taille de la population évolue au fur et à mesure de la reproduction, pour calculer les fréquences il faut actualiser le nombre d'individus à chaque étape du processus.

Tout au long du cycle reproductif, des mutations peuvent affecter les allèles. Par ailleurs, il se peut que des allèles bénéficient d'un avantage sélectif et soient "sur-représentés" d'une génération à l'autre.

4.1.2 Hypothèses simplificatrices

Étant la complexité du phénomène à décrire, nous allons faire ici des hypothèses très simplificatrices. Ainsi, nous nous restreindrons à la description de la fréquence allélique d'un seul gène et supposerons que :

- le gène d'intérêt se présente sous deux (et seulement deux) allèles distincts, notés A et B,
- 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,
- chacun des enfants choisit son parent uniformément parmi tous les individus de la génération précédente et indépendamment des autres,
- la reproduction à l'instant k ne dépend pas des reproductions précédentes.

Il semble intuitif qu'un allèle puisse finir par l'emporter sur l'autre car. C'est ce phénomène, appelé dérive génétique, que l'on souhaite en particulier décrire et quantifier.

4.2 Modèle de Wright-Fisher et variantes

Dans cette section, nous introduisons différents modèles, du plus simple au plus complexe permettant de rendre compte du phénomène de dérive génétique.

4.2.1 Le modèle idéal de Hardy-Weinberg

En 1908, Hardy et Weinberg ont publié indépendamment un premier modèle concernant la structure génotypique d'une population. Dans ce modèle, on fait un certain nombre d'hypothèses permettant de simplifier les calculs, et qui le rendent essentiellement déterministe (nombre infini de gamètes, population de taille infinie). Précisément, on suppose que :

- les gamètes s'associent au hasard indépendamment des gènes considérés (hypothèse de panmixie). Cette hypothèse revient à dire que l'on considère une urne avec une infinité de gamètes qui sont appariés au hasard, sans tenir compte du sexe de l'individu ;
- la population a une taille infinie. Par la loi des grands nombres, on remplace la fréquence de chaque allèle par sa probabilité ;
- la fréquence des gènes n'est pas modifiée d'une génération à la suivante par mutation, sélection , migration.

Sous ces hypothèses, supposons qu'en un locus, les probabilités des allèles A et a soient p et $q = 1 - p$. Alors, à la deuxième génération, après appariement d'une gamète mâle et d'un gamète femelle, on a le génotype AA avec probabilité p^2 , le génotype BB avec probabilité q^2 et le génotype AB avec probabilité $2pq$. Cette structure génotypique est connue sous le nom de structure de Hardy-Weinberg. Mais alors, puisque chaque individu a deux copies de chaque gène, la probabilité d'apparition de l'allèle A dans la population, à la deuxième génération sera $\frac{2p^2+2pq}{2} = p^2 + pq = p$. De même, la fréquence de l'allèle B sera q . Ainsi, étant donné les hypothèses faites, on peut énoncer la loi de Hardy-Weinberg :

“Dans une population isolée d'effectif illimité, non soumise à la sélection et dans laquelle il n'y a pas de mutation, les fréquences alléliques restent constantes.”

Remarque 4.2.1. Cette loi 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. Il n'en est rien : en l'absence d'influences extérieures, la théorie de Mendel assure (sous des hypothèses fortement réductrices) la préservation de la variabilité des génotypes.

Si les accouplements sont panmictiques, les fréquences génotypiques se déduisent directement des fréquences alléliques et restent donc constantes, et égales à p^2 ; $2pq$; q^2 . Le modèle se réduit à un calcul de probabilité sur les fréquences alléliques, et l'on peut se ramener à un modèle haploïde.

4.2.2 Modèle de Wright-Fisher

Alors que dans la population de taille infinie les fréquences alléliques sont stables au cours des générations en l'absence de sélection et de mutation (loi des grands nombres), les fréquences alléliques varient aléatoirement dans des populations de taille finie. (Cela est dû à la variabilité dans la distribution des gènes d'une génération à l'autre). Pour permettre un traitement mathématique pas trop compliqué, le modèle de Wright-Fisher modélise la transmission des gènes d'une génération à l'autre de manière très schématique.

Définition du modèle

Il fait l'hypothèse de générations séparées, ce qui est une simplification considérable du cycle de reproduction. Ici, une population de M individus est représentée par un vecteur de $N = 2M$ allèles. On notera par $n \in \mathbb{N}$ les indices de générations. Dans le modèle de Wright-Fisher, le parent de chaque individu de la génération $n+1$ est distribué uniformément dans la n -ième génération. On suppose également que la population d'individus est de taille finie et constante égale à M (On cultive des petits pois et à chaque génération on en garde M). Chaque individu est caractérisé par deux types (ou allèles) **A** et **a** qu'il transmet par hérédité. On a donc $N = 2M$ allèles. Le modèle de Wright-Fisher est un modèle neutre. Cela veut dire qu'il n'y a pas d'avantage sélectif associé à l'un des deux types (qui favoriserait la reproduction d'un des allèles). Il n'y a pas de mutation (modèle simpliste dans la pratique mais très intéressant dans une première approximation).

Nous voulons étudier la fréquence allélique des deux allèles **A** et **B** à un locus donné, au fil des générations. Supposons que l'on connaisse les fréquences alléliques à une certaine génération n . Quelles vont-elles être à la génération $n+1$? On suppose que les individus peuvent se reproduire de manière indépendante les uns des autres, et que la nouvelle génération est formée de N gènes choisis uniformément dans une urne gamétique que l'on suppose infinie (il existe des milliers de gamètes), et dans laquelle la répartition allélique est celle de la génération n . Ainsi, le modèle qui consiste à choisir M individus d'allèles **A** ou **B** à la génération $n+1$ (et donc $N = 2M$ allèles), et à compter ceux d'allèles **A**, devrait suivre une loi hypergéométrique. Mais l'hypothèse d'urne infinie de gamètes permet d'approcher cette loi hypergéométrique (tirages simultanés) par une loi binomiale qui ne prend plus en compte le nombre d'individus mais seulement les proportions des 2 allèles dans l'urne et qui correspond à un tirage avec remise. Du fait de la neutralité du modèle, tous les individus sont échangeables, et seule la répartition des allèles a une importance. Définissons

X_n^N = nombre d'individus de type **A** à la génération n

Proposition 4.2.2. *Pour tout $n \in \mathbb{N}$, pour tous i et j dans $\{0, \dots, N\}$*

$$\mathbb{P}(X_{n+1}^N = j \mid X_n^N = i) = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}.$$

Démonstration. Par définition, pour tout $n \in \mathbb{N}$, la variable X_n^N prend ses valeurs dans $\{0, \dots, N\}$. Il est clair que

$$\mathbb{P}(X_{n+1}^N = 0 \mid X_n^N = 0) = 1, \quad \text{et} \quad \mathbb{P}(X_{n+1}^N = N \mid X_n^N = N) = 1.$$

Plus généralement, soit $2 \leq i \leq N - 1$. Conditionnellement au fait que $X_n^N = i$, la fréquence de l'allèle A la génération n est i/N et celle de l'allèle a est $1 - i/N$. Ainsi, à la génération suivante, chaque individu tire son parent au hasard, sans se préoccuper des autres individus. Cela correspond donc à un tirage avec remise, et le nombre d'individus d'allèle A à la $(n + 1)$ -ième génération, sachant que $X_n^N = i$ suit alors une loi binomiale $\mathcal{B}(N, i/N)$. Ainsi,

$$\mathbb{P}(X_{n+1}^N = j \mid X_n^N = i) = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}.$$

□

Remarque 4.2.3. Dans les modèles non neutres que nous verrons dans la suite (avec mutation, ou sélection), les individus ne seront plus échangeables, et la probabilité de “tirer” un individu d'allèle A connaissant $X_n^N = i$ ne sera plus uniforme.

Remarque 4.2.4. D'après la proposition, si on note $p_i = i/N$, on aura

$$\mathbb{E}(X_{n+1}^N \mid X_n^N = i) = N \times p_i = i$$

et donc

$$\mathbb{E}(X_{n+1}^N - X_n^N \mid X_n^N = i) = 0,$$

$$\text{var}(X_{n+1}^N - X_n^N \mid X_n^N = i) = N \times p_i(1 - p_i) = i \left(1 - \frac{i}{N}\right).$$

Nous nous intéressons à la transmission de l'allèle A. Les questions importantes sont les suivantes : l'allèle A va-t-il envahir toute la population ? Ou l'allèle B ? En combien de temps ? Aura-t-on co-existence ?

Probabilité et temps de fixation

Le premier pas dans la description de dérive génétique, ou encore de la fixation de l'allèle A est le suivant :

Théorème 4.2.5. *La suite $(X_n^N)_{n \geq 0}$ est une chaîne de Markov à espace d'états fini et c'est également une martingale bornée. La matrice de transition P de la chaîne est donnée pour tous i et j dans $\{0, \dots, N\}$ par*

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

Les états 0 et N sont deux états absorbants pour la chaîne.

Démonstration. Par définition, le processus $(X_n^N)_{n \geq 0}$ prend ses valeurs dans l'ensemble $\{0, \dots, N\}$. Il est clair que ce processus définit une chaîne de Markov puisque la loi de X_{n+1}^N conditionnellement à la connaissance de X_n^N est parfaitement défini par le calcul précédent. La forme de la matrice de transition a été calculée dans la proposition précédente. Par ailleurs, si $\mathcal{F}_n$ désigne la filtration engendrée par $X_0^N, \dots, X_n^N$, on a

$$\begin{aligned} \mathbb{E}(X_{n+1}^N \mid \mathcal{F}_n) &= \sum_{j=0}^N j \times \mathbb{P}(X_{n+1}^N = j \mid \mathcal{F}_n) = \sum_{j=0}^N j \times \mathbb{P}(X_{n+1}^N = j \mid X_n^N) \\ &= \sum_{j=0}^N j \times \binom{N}{j} \left(\frac{X_n^N}{N}\right)^j \left(1 - \frac{X_n^N}{N}\right)^{N-j} \\ &= N \times \frac{X_n^N}{N} = X_n^N. \end{aligned}$$

Ainsi, le processus (X_n^N) est une martingale, bornée par zéro et N . Il est clair que les états 0 et N sont absorbants. $\square$

Remarque 4.2.6. Remarquons que comme le processus (X_n^N) est une martingale, la proportion d'allèle **A** est constante en moyenne si et seulement si le modèle est neutre : il n'y a pas d'avantage sélectif pour un allèle. On verra dans une prochaine section comment la sélection influe sur cette dynamique

Nous allons maintenant étudier comment les allèles se fixent dans la population d'individus de taille donnée N , en temps long. Nous allons donc faire tendre n vers l'infini.

Théorème 4.2.7. *Quand le nombre de générations n tend vers l'infini, la suite de variables aléatoires $(X_n^N)_{n \geq 0}$ converge presque-sûrement vers une variable aléatoire X_∞^N à valeurs dans $\{0, N\}$. De plus*

$$\mathbb{P}(X_\infty^N = N \mid X_0^N = i) = \frac{i}{N}$$

Remarquons que $\mathbb{P}(X_\infty^N = N)$ est la probabilité de fixation de l'allèle **A** dans la population et que $\mathbb{P}(X_\infty^N = 0)$ est sa probabilité de disparition, *i.e.* la probabilité de fixation de l'allèle **a**.

Démonstration. La suite (X_n^N) est une martingale bornée, elle converge donc lorsque n tend vers l'infini. Seuls les points 0 et N sont absorbants, les autres états étant transitoires. La limite X_∞^N est donc une variable aléatoire à valeurs dans $\{0, N\}$. Le calcul de la probabilité est immédiat. $\square$

Pour avoir une idée du temps que la fixation met pour avoir lieu, étudions la probabilité que deux copies d'un même locus choisies au hasard (sans remise) portent des allèles différents (hétérozygotie). Soit $h(n)$ cette probabilité, dans le cas où ces copies sont choisies dans la n -ième génération. Calculons tout d'abord la probabilité H_n d'avoir l'hétérozygotie, conditionnellement à la connaissance de X_n^N . Lorsque $X_n^N = i$, on a facilement

$$H_n = \frac{\binom{i}{1} \binom{N-i}{1}}{\binom{N}{2}}$$

Plus généralement, H_n est donc la variable aléatoire donnée par

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

On en déduit alors l'expression de la probabilité $h(n)$:

Proposition 4.2.8. *L'hétérozygotie est donnée par*

$$h(n) = \mathbb{E}(H_n) = \left(1 - \frac{1}{N}\right)^n \mathbb{E}(H_0).$$

Remarque 4.2.9. Remarquons que si l'on part avec une seule copie d'allèle **A**, alors $X_0^N = 1$, et $H_0 = 2/N$. Ainsi, $h(0) = 2/N$, la probabilité qu'à la génération n , deux copies choisies au hasard aient deux allèles différents vaut

$$\frac{2}{N} \left(1 - \frac{1}{N}\right)^n.$$

Cela nous donne la vitesse de décroissance de l'hétérozygotie en fonction du temps.

Une question naturelle est consiste à décrire 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 la suite $(X^n)_{n \geq 0}$ en zéro ou en N , sachant que $X_0 = i$, on a $m_0 = m_N = 0$ et, grâce à la propriété de Markov,

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

Ce système à $N - 1$ inconnues n'est pas facile à résoudre. Il est possible d'utiliser l'ordinateur pour trouver une valeur approchée déterministe de la solution. Cependant, dès que N est un peu grand, des problèmes dus au fait que les coefficients p_{ij} sont très petits risquent de fausser le résultat.

On peut aussi utiliser une méthode probabiliste pour estimer la quantité m_i via une méthode de Monte-Carlo.

Il est toutefois possible de trouver un équivalent des m_i lorsque la taille de la population tend vers l'infini. Notons Z la chaîne de Markov définie par $Z_n := X_n^N/N$. Bien entendu, le temps d'atteinte de $\{0, N\}$ pour la suite $(X_n^N)_{n \geq 0}$ est égal au temps d'atteinte de $\{0, 1\}$ pour la chaîne Z . Pour tout $x \in \{0, 1/N, \dots, 1\}$, notons $t(x)$ l'espérance de T lorsque $Z_0 = x$. Sachant que $Z_0 = x$, Z_1 s'écrit $x + X$ où X est tel que $N(X - x)$ suit la loi binomiale $\mathcal{B}(N, x)$. On a alors, d'après la propriété de Markov et la définition de X^N , on a $t(x) = \mathbb{E}(t(x + X) + 1) = 1 + \mathbb{E}(t(x + X))$.

Supposons que t soit proche d'une fonction de classe $\mathcal{C}^\infty$. Puisque la variable aléatoire X est bornée, on peut écrire

$$t(x + X) = t(x) + t'(x)X + \frac{t''(x)}{2}X^2 + O(|X|^3).$$

Or, il est clair que $\mathbb{E}(X) = 0$, $\mathbb{E}(X^2) = x(1 - x)/N$. D'autre part, en vertu du théorème limite central, la loi de $\sqrt{N}X$ est proche de la loi $\mathcal{N}(0, x(1 - x))$ et ainsi, par exemple, $\mathbb{E}(X^4)$ est de l'ordre de N^{-2} . En résumé, on obtient l'expression suivante :

$$t(x) = 1 + t(x) + \frac{x(1 - x)}{2N}t''(x) + O(N^{-3/2}).$$

On pourrait donc dire que la fonction t doit ressembler à la fonction y solution de

$$\forall x \in]0, 1[, \quad y''(x) = \frac{2N}{x(1 - x)} \quad \text{et} \quad y(0) = y(1) = 0.$$

On obtient alors l'approximation suivante.

Proposition 4.2.10. *Si N est grand et $x = i/N$ alors*

$$m(i) \underset{N}{\sim} -2N(x \ln x + (1 - x) \ln(1 - x)).$$

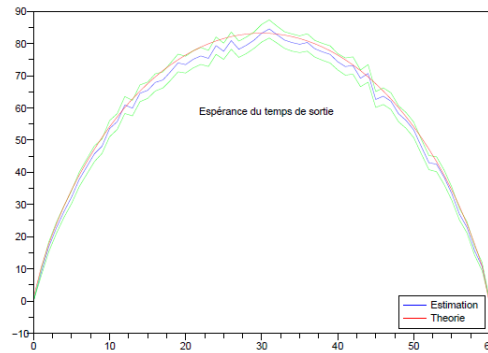


FIG. 4.1 – Estimation du temps de fixation pour $N = 40$.

4.2.3 Modèle de Wright-Fisher avec 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 supposons ici qu'il n'est déterminé que par le locus qui nous intéresse. Nous supposons de plus que la sélection ne s'opère que sur le critère de viabilité.

Supposons que les adaptabilités des deux allèles **A** et **a** soient données par $1 + s$ et 1 , c'est-à-dire que qu'un individu porteur de l'allèle **A** a $1 + s$ fois plus de chances de survivre qu'un individu porteur de l'allèle **B**. Si l'on suppose qu'à la génération n il y a i individus possédant l'allèle **A**, l'avantage sélectif se traduit par le fait que la probabilité d'obtention de l'allèle **A** sera donnée par

$$p_i := \frac{i(1+s)}{i(1+s) + N - i}$$

Comme précédemment, on aura alors de façon formelle

$$\mathbb{P}(X_{n+1}^N = j \mid X_n^N = i) = \binom{N}{j} p_i^j (1 - p_i)^{N-j}$$

Notons $(\pi_i)_i$ les probabilités de fixation de la chane en N (avant 0) lorsque $X_0^N = i$. La propriété de Markov assure que

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

La pression de sélection est très faible en pratique. Supposons que s soit de l'ordre N^{-1} et posons $\alpha = Ns$. Par le mme raisonnement que dans la preuve de la proposition 4.2.10, on peut écrire, pour $x = i/N$

$$\pi(x) = \mathbb{E}(\pi(x + Z)) = \pi(x) + \pi'(x)\mathbb{E}(Z) + \frac{\pi''(x)}{2}\mathbb{E}(Z^2) + O(|Z|^3).$$

En tenant compte de la sélection, on a $Z = Y/N - x$ où Y suit la loi $\mathcal{B}(N, (1+s)x/(sx+1))$ donc

$$\begin{aligned} \mathbb{E}(Z) &= \frac{(1+s)x}{sx+1} - x = \frac{sx(1-x)}{sx+1} = \frac{1}{N}\alpha x(1-x) + O(N^{-2}), \\ \mathbb{E}(Z^2) &= \frac{1}{N^2}\mathbb{V}(Y) + (\mathbb{E}(Y)/N - x)^2 = \frac{1}{N}x(1-x) + O(N^{-2}), \\ \mathbb{E}(|Z|^3) &= O(N^{-3/2}). \end{aligned}$$

La fonction π semble donc proche de la solution de l'équation différentielle suivante :

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

c'est-à-dire

$$\pi(x) \sim z(x) = \frac{1 - e^{-2\alpha x}}{1 - e^{-2\alpha}}.$$

Les figures suivantes proposent une illustration de l'estimation de la probabilité et du temps moyen de fixation de l'allèle A pour $N = 30$ et $\alpha = 2$, et $N = 30$ et $\alpha = 10$.

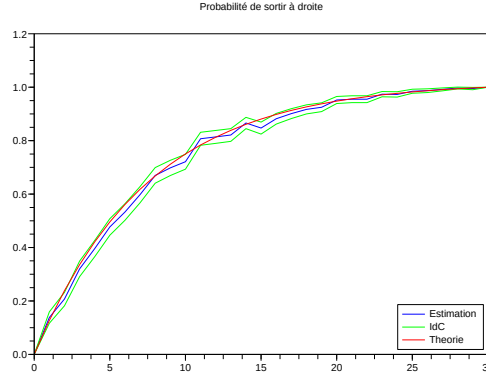


FIG. 4.2 – Probabilité de fixation de A pour $N = 30$ et $\alpha = 2$.

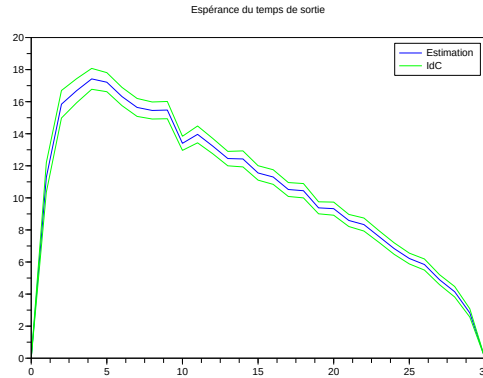


FIG. 4.3 – Espérance du temps de fixation pour $N = 30$ et $\alpha = 10$.

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 $\pi(0,5) = 0,5$. 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.

4.2.4 Modèle de Wright-Fisher avec mutation

Supposons maintenant qu'au cours de la reproduction, on ait des mutations de l'allèle **A** vers l'allèle **a** avec probabilité $\mu_{\mathbf{A}}$ et de l'allèle **a** vers l'allèle **A** avec probabilité $\mu_{\mathbf{a}}$. On suppose toujours que le modèle est neutre, c'est-à-dire que les deux allèles ont le même avantage sélectif. Alors si l'on désigne toujours par $(X_n^N)_{n \geq 0}$ le processus qui décrit le nombre d'allèles **A** à la génération n , on sait que la loi de X_{n+1}^N est une loi binomiale de paramètres qui dépendent de la répartition de l'allèle **A** à la génération n . Si on a i copies de l'allèle **A** à la génération n , cette répartition sera donnée par

$$p_i := \frac{i}{N}(1 - \mu_{\mathbf{A}}) + \left(1 - \frac{i}{N}\right)\mu_{\mathbf{B}}$$

Alors X_{n+1}^N suit une loi binomiale $\mathcal{B}(N, p_i)$, et on a

$$\mathbb{P}(X_{n+1}^N = j \mid X_n^N = i) = \binom{N}{j} p_i^j (1 - p_i)^{N-j}$$

Remarque 4.2.11. Notons que dans ce modèle où les mutations d'un allèle à l'autre sont possibles, les états 0 et N ne sont plus des états absorbants.

La matrice de transition de X^N n'a que des coefficients strictement positifs. La chaîne est donc irréductible, récurrente (positive) et apériodique. Elle admet donc une unique mesure de probabilité invariante $(\rho_k)_{0 \leq k \leq N}$ et, pour toute mesure initiale ν et tout j ,

$$\mathbb{P}_{\nu}(X_n^N = j) \xrightarrow{n \rightarrow \infty} \rho_j.$$

Par raisonnement similaire au cas sans mutation, on a

$$\mathbb{E}(X_{n+1}^N \mid X_n^N = i) = (1 - \mu_{\mathbf{A}})i + \mu_{\mathbf{B}}(N - i) = N\mu_{\mathbf{B}} + (1 - \mu_{\mathbf{A}} - \mu_{\mathbf{B}})i,$$

et donc

$$\mathbb{E}(X_{n+1}^N) = N\mu_{\mathbf{B}} + (1 - \mu_{\mathbf{A}} - \mu_{\mathbf{B}})\mathbb{E}(X_n^N).$$

On peut ainsi calculer l'espérance de X_n^N et de la mesure invariante (par passage à la limite) :

$$\mathbb{E}(X_n^N) = \frac{N\mu_{\mathbf{B}}}{\mu_{\mathbf{A}} + \mu_{\mathbf{B}}} + \left(\mathbb{E}(X_0) - \frac{N\mu_{\mathbf{B}}}{\mu_{\mathbf{A}} + \mu_{\mathbf{B}}} \right) (1 - \mu_{\mathbf{A}} - \mu_{\mathbf{B}})^n \xrightarrow{n \rightarrow \infty} \frac{N\mu_{\mathbf{B}}}{\mu_{\mathbf{A}} + \mu_{\mathbf{B}}}.$$

Toutefois, on ne dispose pas d'expression explicite pour la mesure invariante de la chaîne de Markov.

4.2.5 Modèle avec mutation et sélection

Plus généralement, on peut définir un modèle de Wright-Fisher avec sélection et mutation, où

$$p_i = \frac{(1+s)(i(1-\alpha_1) + (N-i)\alpha_2)}{(1+s)(i(1-\alpha_1) + (N-i)\alpha_2) + i\alpha_1 + (N-i)(1-\alpha_2)}$$

Remarquons que la différence $p_i - i/N$ exprime la différence due à la sélection et à la mutation dans le modèle déterministe de population infinie. Les fluctuations statistiques dues à la taille finie de la population apparaissent à travers le comportement aléatoire de la chaîne de Markov, qui se traduit par les probabilités de transitions P_{ij} . Il est rarement possible, pour une telle chaîne, de calculer les probabilités d'intérêt. Nous allons voir dans le paragraphe suivant que si la taille N de la population est grande, et si les effets individuels des mutations et de la sélection sont faibles, alors le processus peut être approché par un processus de diffusion pour lequel il sera plus facile d'obtenir des résultats quantitatifs.

4.3 Approximations continues des modèles discrets

Lorsque les paramètres m et n sont grands, on souhaite approcher les chaînes de Wright-Fisher introduites précédemment par des processus de diffusion solution d'équations différentielles stochastiques à préciser. Le prochain paragraphe établit précisément la convergence des modèles discrets vers des modèles continus.

4.3.1 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

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

Notons T_N le noyau de transition sur $\mathcal{C}(K_N)$ associé :

$$\forall x \in K_N, \quad T_N f(x) = \mathbb{E}[f(Z_1^N) | Z_0^N = x].$$

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.

On considère enfin le processus $(X_t)_{t \geq 0}$ de semi-groupe $(T(t))_{t \geq 0}$ et de générateur infinitésimal

$$G = \frac{1}{2} \sigma^2(x) \partial_{xx}^2 + b(x) \partial_x,$$

avec

$$\sigma^2(x) = x(1-x) \quad \text{et} \quad b(x) = \alpha x(1-x)(x+h(1-2x)) - \beta_1 x + \beta_2(1-x). \quad (4.1)$$

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

Théorème 4.3.1. *Posons*

$$\alpha = 2Ns, \quad \beta_1 = 2Nu \quad \text{et} \quad \beta_2 = 2Nv, \quad (4.2)$$

avec α , β_1 et β_2 d'ordre 1. Alors, pour toute fonction $f \in \mathcal{C}(K)$ et tout $t_0 > 0$,

$$\lim_{N \rightarrow \infty} \sup_{0 \leq t \leq t_0} \sup_{x \in K} \left| T_N^{[2Nt]} f(x) - T(t)f(x) \right| = 0.$$

Démonstration. Il suffit de montrer que

$$\lim_{N \rightarrow \infty} \sup_{x \in K_N} |2N(T_N - I)f(x) - Gf(x)| = 0.$$

pour toute fonction $f \in \mathcal{C}^2(K)$. Or, par définition de T_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 $x \in K_N$. Un développement limité de f donne alors le résultat. $\square$

Remarque 4.3.2. Pour transférer les estimations de temps d'absorption de la diffusion à la chane 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 4.3.3. *Suivant la valeur de β , trois cas se présentent :*

- si $\beta_1 = 0$, le point 1 est accessible et absorbant ;
- si $0 < \beta_1 < 1/2$, le point 1 est régulier, accessible et non absorbant ;
- si $\beta > 1/2$, le point 1 est inaccessible et non absorbant.

Détaillons à présent les divers cas qui peuvent se présenter.

4.3.2 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 λ est une primitive de la fonction

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

Remarque 4.3.4. Le temps de fixation pour la chane de Wright-Fisher a une espérance de l'ordre de $2N\mathbb{E}(T|X_0 = x)$ et une variance de l'ordre de $4N^2\mathbb{V}(T|X_0 = x)$.

4.3.3 Mutation sans sélection

On rappelle que dans ce cas, la chaîne X^N est ergodique de mesure invariante ρ . On cherche à décrire la mesure ρ lorsque la population est de grande taille (puisque l'on ne dispose pas de formule simple pour N fixé). Pour obtenir une limite non dégénérée, on suppose que les probabilités de mutation satisfont la condition suivante :

$$\lim_{N \rightarrow \infty} 2N\mu_A = \theta_A > 0, \quad \lim_{N \rightarrow \infty} 2N\mu_B = \theta_B > 0.$$

On définit $\theta = \theta_A + \theta_B$ comme la probabilité totale de mutation. Pour obtenir une expression asymptotique de ρ on utilise l'approximation par une diffusion sur $]0, 1[$ dont les coefficients sont donnés, pour $x \in]0, 1[$ par :

$$\mu(x) = -x \frac{\theta_A}{2} + (1 - x) \frac{\theta_B}{2} \quad \text{et} \quad \sigma^2(x) = x(1 - x)$$

En d'autres termes la chane X converge, lorsque N tend vers l'infini vers le processus solution de l'EDS suivante

$$dY_t = \sqrt{Y_t(1 - Y_t)} dB_t - \frac{\theta_A}{2} Y_t dt + \frac{\theta_B}{2} (1 - Y_t) dt.$$

On peut montrer que cette diffusion reste bien dans l'intervalle $]0, 1[$ et qu'elle admet pour mesure invariante ν la loi bta de paramètres θ_B et θ_A , c'est-à-dire que

$$\nu(y) = \frac{\Gamma(\theta)}{\Gamma(\theta_A)\Gamma(\theta_B)} y^{\theta_B-1} (1-y)^{\theta_A-1}.$$

En particulier, si Y suit la loi ν ,

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

Remarque 4.3.5. 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 5

Coalescence

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. Le modèle présenté ici est intimement lié au modèle de Wright-Fisher introduit au chapitre précédent.

5.1 Construction de l'arbre généalogique

Commençons par rappeler les hypothèses biologiques associées aux différents modèles présentés au chapitre précédent :

- la taille de la population observée reste constante au cours du temps, on note N son cardinal ;
- 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 qui relient chacun des enfants à son parent ont une même probabilité ;
- la reproduction à l'instant k est indépendante des reproductions aux instants précédents $k - 1, k - 2, \dots$

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 5.1.1. 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 ν_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 5.1.2. *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 μ_N de paramètre $(N, (1/N, \dots, 1/N))$, c'est-à-dire que*

$$\begin{aligned} \mathbb{P}(\nu_1^k = m_1, \dots, \nu_N^k = m_N | \nu_1^k + \dots + \nu_N^k = N) \\ = \frac{N!}{m_1! \dots m_N!} \left(\frac{1}{N}\right)^N \mathbf{I}_{\{m_1 + \dots + m_N = N\}}. \end{aligned} \quad (5.1)$$

De plus, les variables aléatoires $(\nu^k)_k$ 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 5.1.3. 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 entrainer le phénomène dit de dérive génétique.

Exercice 5.1.4. Soit $Y_1, \dots, Y_N$ des variables aléatoires indépendantes et identiquement distribuées de loi de Poisson de paramètre θ . 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 μ_N . La loi de ν_i^k est donc la loi binomiale $\mathcal{B}(N, 1/N)$.

Remarque 5.1.5. 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 apparatre 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.

5.2 Le modèle à temps discret

D'après la remarque 5.1.3, la probabilité que deux individus aient deux parents distincts (à la génération précédente) est égale à $1 - 1/N$ que nous noterons λ . 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 λ^r . Le temps d'apparition de l'ancêtre commun le plus récent suit donc une loi géométrique de paramètre $1/N$.

Considérons à présent trois individus. La probabilité pour qu'ils aient trois parents distincts vaut

$$\left(1 - \frac{1}{N}\right) \left(1 - \frac{2}{N}\right)$$

et en itérant l'argument, la probabilité pour que les trois individus aient encore trois ancêtres distincts r générations auparavant est donnée par

$$\left[\left(1 - \frac{1}{N}\right) \left(1 - \frac{2}{N}\right) \right]^r = \left(1 - \frac{3}{N} + \frac{2}{N^2}\right)^r.$$

La probabilité pour que trois individus (distincts) aient au plus deux parents distincts est égale à

$$\frac{3(N-1)}{N^2} + \frac{1}{N^2} = \frac{3N-2}{N^2}.$$

Donc, sachant que le premier ancêtre commun (à au moins deux des trois individus) apparait à la génération r , la probabilité conditionnelle que les trois individus aient deux ancêtres distincts à la génération r est égale à

$$\frac{3N-3}{3N-2}.$$

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 faons de découper un ensemble à k éléments en j ensembles non vides (ces nombres sont parfois appelé 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 faons de choisir j parents distincts, $\mathcal{S}_k^{(j)}$ est le nombre de faons d'associer à ces j parents k enfants, enfin, N^k est le nombre de faons 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 5.2.1. *La suite $A_n^N(\cdot)$ est une chane 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 (puisqu'ils mènent tous à 1). Cependant, il est difficile d'étudier les propriétés fines de cette chane, 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.

5.3 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

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

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 $\binom{n}{2}$. 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, il y a $\binom{n}{2}$ couples distincts et le minimum de $\binom{n}{2}$ variables aléatoires indépendantes et identiquement distribuées de loi $\mathcal{E}(1)$ est la loi $\binom{n}{2}$.

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

5.4 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 5.4.1. *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 μ .*

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

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

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 $\mathcal{G}(p_2)$. 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 $\mathcal{E}(3)$.

On a donc montré le résultat suivant.

Proposition 5.4.2. *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 apparait,
- le groupe a conservé deux ancêtres distincts pendant 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.

En d'autres termes, on a 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.

On peut généraliser le résultat ci-dessus.

Proposition 5.4.3. *Pour tout $n \geq 2$,*

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

5.5 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 $\binom{k}{2}$. De plus, les variables aléatoires $(T_k)_{2 \leq k \leq n}$ sont indépendantes.

5.5.1 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

$$\mathbb{E}(W_n) = 2 \left(1 - \frac{1}{n}\right) \quad \text{et} \quad 1 = \mathbb{V}(W_2) \leq \mathbb{V}(W_n) \leq \lim_{n \rightarrow \infty} \mathbb{V}(W_n) = 8 \frac{\pi^2}{6} - 12 \simeq 1.16.$$

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 :

$$\mathbb{P}(W_n \leq t) = \mathbb{P}(A_n(t) = 1) = \sum_{k=1}^n \frac{(2k-1)(-1)^{k-1} n_{[k]}}{n_{(k)}} e^{-k(k-1)t/2}. \quad (5.2)$$

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

5.5.2 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 5.5.1. *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 λ alors son espérance vaut $1/\lambda$, sa variance $1/\lambda^2$ et, pour $\alpha > 0$, αT suit la loi exponentielle de paramètre λ/α . L'indépendance des variables aléatoires $(T_i)_i$ fait le reste. $\square$

On peut en fait déterminer complètement la loi de L_n .

Proposition 5.5.2. *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_λ la densité de la loi exponentielle de paramètre λ . 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 $x > 0$,

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

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 5.5.3. *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 5.5.2 : 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 [BL98]. $\square$

5.6 Mutations pour un nombre infini d'allèles

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 μ 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 μ . On pose $\theta = 2\mu N$ et on suppose que θ est d'ordre 1. Les deux processus de coalescence et mutation sont d'origines aléatoires différentes. Nous les considérerons indépendants.

La question est d'estimer θ .

5.6.1 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 5.6.1. *On a*

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

avec $(\eta_k)_{k \geq 1}$ des variables aléatoires indépendantes de lois 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 $\mathcal{E}(\binom{n}{2})$) 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

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

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

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

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 η_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 5.6.2. *La suite $(K_n/\ln n)_n$ converge vers θ en probabilité.*

On cherche de plus à contrôler les fluctuations de $K_n/\ln n$ autour de θ . 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 5.6.3. *On a la convergence en loi suivante :*

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

Démonstration. On utilise ici la version suivante du théorème limite central.

Théorème 5.6.4. *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*

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

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

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

□

5.6.2 Estimation du taux de mutation

Comment les résultats précédents permettent-ils d'estimer le taux de mutation θ . Le corollaire 5.6.2 suggère que

$$\hat{\theta}_n = \frac{K_n}{\ln n}$$

est un estimateur raisonnable de θ . De plus, la proposition 5.6.3 permet d'écrire un intervalle de confiance asymptotique pour θ : par exemple θ 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 ... La convergence est vraiment très lente.

Bibliographie

[BL98] P. BARBE et M. LEDOUX – *Probabilité*, Belin, 1998. [57](#)