Le comportement critique de la quasiespèce

Maxime Berger

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL
Préparée à l'Université Paris-Saclay

Le Comportement Critique de la Quasiespèce

Soutenue par
Maxime Berger
Le 21 Juin 2021

Ecole doctorale n° 574
Ecole Doctorale Jacques Hadamard

Spécialité
Mathématiques

Composition du jury :

Nicolas Champagnat Inria Nancy - Grand Est	*Rapporteur*
Michel Benaïm Université de Neuchâtel	*Rapporteur*
Amandine Veber CNRS - Université de Paris	*Présidente du jury*
Bertrand Maury Université Paris-Sud	*Examinateur*
Amaury Lambert Collège de France	*Examinateur*
Raphaël Cerf Université Paris-Sud	*Directeur de thèse*

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

ENS | PSL

Ce manuscrit signe la fin de trois ans de travail, trois belles années passées au DMA. Je tiens à remercier Raphael Cerf, qui a siégé au premier plan de ce travail de manière constante et bienveillante. Ses cours et sa façon de faire de la recherche resteront dans mes souvenirs et continueront à m'inspirer pour les années à venir. Je tiens également à exprimer ma plus sincère gratitude envers Nicolas Champagnat et Michel Benaim pour leur relecture très précise. Leurs remarques ont contribué à améliorer de manière significative ce manuscrit. Merci à Bertrand Maury, il a été un professeur très inspirant et m'a fait comprendre que les modèles mathématiques les plus simples sont capables de contenir l'essence de problèmes complexes. Merci à Amandine Veber pour nos discussions éclairantes. Je tiens à vous remercier, ainsi qu'Amaury Lambert pour avoir accepté de faire partie de mon jury.

Merci à vous, camarades doctorants pour ces moments complices. Zaina, Fabienne, les moments passés avec vous font partie des meilleurs.

Ce travail contient une part de chacun de vous.

Table des matières

Introduction

1 Motivation

L'objet de ce manuscrit est de décrire mathématiquement l'état d'équilibre d'une population d'individus. Ces individus se reproduisent et transmettent leurs gènes à leurs descendants. Cependant, le matériel génétique d'un individu issu d'une reproduction n'est pas une copie exacte de celui de son parent : quelques erreurs vont venir se glisser. Ces modifications, que nous appellerons mutations, sont nécessaires à l'évolution, sans elles la population serait condamnée à ne jamais découvrir de nouveaux comportements. La plupart du temps ces mutations s'expriment peu, c'est-à-dire qu'elles n'ont pas d'influence sur le comportement des individus, voire s'expriment en défaveur de l'individu. Cependant, lors d'un événement rare, des mutations peuvent permettre à l'individu un meilleur accès aux ressources. Dans ce cas, il laissera plus de descendants, et ces descendants porteront eux-mêmes ce trait avantageux. Dans la population, nous observerons de plus en plus d'individus possédant ce trait, c'est la sélection naturelle déjà comprise par Darwin :

Comme il naît beaucoup plus d'individus de chaque espèce qu'il n'en peut survivre et que par conséquent il se produit souvent une lutte pour la vie, il s'ensuit que tout être qui varie, même légèrement, d'une façon qui lui est profitable, dans les conditions complexes et quelquefois variables de la vie, a une plus grande chance de survivre. Cet être est ainsi l'objet d'une sélection naturelle. En vertu du principe si puissant de l'hérédité, toute variété ainsi choisie aura tendance à se multiplier sous sa nouvelle forme modifiée.

Darwin, extrait de l'origine des espèces

A quelle fréquence ces mutations se produisent-elles ? Pour pouvoir transmettre ses gènes à ses descendants, il ne faut pas que des mutations se produisent trop souvent. D'un autre côté, il faut que des mutations se produisent pour que des individus puissent naître avec un comportement inédit. Tout est donc question d'équilibre pour avoir la meilleure trajectoire évolutive possible. En réalité, sélection et mutation sont deux ingrédients parmi d'autres dans le phénomène complexe qu'est l'évolution. Nous pouvons par exemple mentionner les phénomènes de recombinaison [Fel74], drift, de compétition inter-espèce [Prü97], de contraintes géographiques, voire de coopération entre individus [Now06], ou encore d'accès aux ressources limitées [BB11]. Nous allons cependant considérer seulement ces deux ingrédients et travailler sur un modèle simplifié de l'évolution. Les premiers travaux mathématiques sur la génétique des populations ont été écrits dans les années 1920 par les

trois pionniers du domaine : Haldane [Hal27], Fisher [Fis58] et Wright [Wri49]. Pour une synthèse des travaux et des résultats, se référer à l'article d'Ellen Baake [BG00].

2 Le modèle d'Eigen

2.1 Histoire et importance du modèle

Le modèle que nous allons étudier est une variation des modèles de mutation-selection, déjà présents chez Crow et Kimura dans les années 1965 [CK70] [CK65]. Ces modèles et leurs implications sont extensivement décrits dans la synthèse [Bür98] de Bürger. Comme nous avons pu le voir, un élément fondamental du modèle est de présenter un seuil d'erreur : une valeur critique du paramètre de mutation qui sépare deux régimes distincts. Si les mutations sont inférieures à ce seuil, les individus les plus adaptés à leur environnement vont prospérer dans la population, si les mutations sont supérieures, ces individus vont complètement disparaître et la population sera constituée d'individus au matériel génétique complètement aléatoire.

Le modèle que nous allons décrire dans la prochaine section est un modèle très général qui peut s'appliquer dans de nombreuses situations. Il suffit qu'un ensemble d'entités puissent se reproduire, et que ces reproductions induisent des erreurs. La première application à laquelle on peut penser est de considérer des populations d'entités vivantes. Il a été constaté [Dom+78] [DSP12] que certains virus se reproduisent à un taux de mutation proche de leur seuil d'erreur [Dom02] [AF10]. Une approche thérapeutique en découle : elle consiste à augmenter leur taux de mutation pour les faire se reproduire au dessus du seuil d'erreur afin qu'ils se détruisent ainsi eux-mêmes [CCA01], [ADL04] [Cad16]. De telles thérapies sont envisagées pour le virus du VIH notamment [Tri+12], [HF15], mais aussi pour lutter contre l'hépatite C [Gòm+99], voire contre certains cancers [SD04], [BMS06]. On observe également ce comportement chez certaines enzymes [OF10]. Cependant, augmenter le taux de mutation présente des risques car le virus risque alors de découvrir de meilleures armes [MDL10]. L'actualité nous fournit un excellent exemple : le coronavirus mute constamment et de nouveaux "variants" apparaissent en continu. A certains moments, des variants qui se reproduisent plus que les autres, c'est-à-dire qui sont plus contagieux, sont découverts par le virus. Ceux qu'on appelle variant britannique ou sud africain ont seulement quelques gènes qui diffèrent de la souche d'origine. Bien sûr, la pandémie possède des composantes supplémentaires très complexes. Nous avons exposé un modèle qui ne présente aucune caractéristique géographique, or il a été montré [WAE10] que les phénomènes de seuil d'erreur présentent une grande sensibilité aux phénomènes de migration.

Le modèle révèle aussi tout son intérêt lorsqu'il est appliqué à des entités qui ne sont pas vivantes. Le modèle a par exemple été appliqué pour comprendre l'évolution du langage [Now02]. Bien sûr, une application que nous pouvons citer est précisément ce pourquoi le modèle a été introduit. Ce que Manfred Eigen avait en tête, c'était de pouvoir expliquer l'origine de la vie sur Terre [Luq03], [CN12]. A partir du moment où les conditions ont été propices au développement de la vie et à l'apparition de la cellule, qui constitue l'intégralité du vivant aujourd'hui, il s'est écoulé un milliard

d'années. Durant cette période, tous les composants se sont mélangés et les essais de construction de structures biologiques ont été nombreux. Eigen prétend même que maintenant que la cellule est bien installée, nous ne serons plus jamais témoins d'une telle créativité. Pour cette interprétation, il faut considérer que ce que nous avions appelé le matériel génétique des individus, les chaînes dans l'espace $\{0,1\}^\ell$, est en correspondance avec les différentes façons d'agencer les éléments de la soupe primitive et que la cellule correspond à la master sequence. Cette façon de voir le modèle conduit aussi à des algorithmes d'optimisations, qu'on appelle algorithmes génétiques. L'idée est que nous avons une fonction à optimiser, mais que nous ne pouvons pas calculer son image pour tous les points de l'espace de départ. Nous allons donc calculer l'image de quelques points, choisis initialement au hasard et considérer ces points comme des individus qui peuvent se reproduire. La fitness qui leur est attribuée est par exemple la valeur de la fonction en ce point. Nous faisons ensuite se reproduire ces individus, puis nous les faisons muter pour explorer peu à peu l'espace de départ de la fonction à optimiser et espérer découvrir la chaîne qui maximisera la fitness. La question de l'ajustement des paramètres dans ces algorithmes est cruciale et beaucoup de travaux s'inspirent du seuil d'erreur pour régler au mieux ces paramètres [OHB] [OHB99], [EHM99]. Il n'y a même pas besoin que la fonction soit définie sur l'hypercube pour cela, voir [FZB11] pour une synthèse. Souvent, pour ces algorithmes, les individus sont plutôt considérés diploïdes et des phénomènes de recombinaison appelés cross-over sont intégrées [OH97]. Des résultats théoriques [Fur97], [HM91] ont été obtenus, mais les implémentations pratiques se font souvent au cas par cas, en utilisant au maximum ce qu'on pense savoir sur la fonction à optimiser.

Le modèle possède également une structure mathématique profonde et de nombreux liens avec d'autres domaines ont été creusés. Les premières analogies avec la physique statistique [Hig95] datent des articles de Leuthäuser [Leu86], qui ont fait le lien entre le modèle d'Eigen et le modèle d'Ising. Le phénomène de transition de phase emprunte d'ailleurs son appellation à ces analogies [FP97]. Des liens avec des processus de branchements se sont aussi révélés extrêmement fructueux [DSS85], [BG07].

2.2 Le formalisme mathématique

Le comportement des individus est entièrement caractérisé par une chaîne finie de zéros et de uns, représentant leur génome. Nous considèrerons seulement un modèle génotypique, et nous ne nous préoccuperons pas de l'expression phénotypique des gènes comme a pu le faire Schuster dans [Sch97]. Dans la vie, le matériel génétique est plus souvent composé d'un alphabet à 4 lettres : A, T, G et C par exemple. Cependant, les difficultés mathématiques apparaissent déjà avec un alphabet à 2 lettres et les méthodes que nous emploierons seront complètement généralisables à un alphabet fini quelconque. Dans la suite, la longueur des chaînes est la même pour tous les individus, elle sera notée ℓ. Le comportement de chaque individu est donc déterminé par un point de l'hypercube $\{0,1\}^\ell$. De telles chaînes seront souvent notées dans la suite par les lettres u, v. D'autres ensembles pour l'espace des chaînes ont été étudiés, par exemple dans [KS02], ou [HM+14]. La reproduction

ne nécessite qu'un seul parent, on dit que nos individus sont haploïdes, mais il est aussi possible d'adapter la théorie qui va suivre avec des individus diploïdes, voir [JN06]. Les mutations se produisent lors des reproductions : le génome du nouvel individu est une copie de celui de son parent où chaque digit a été changé avec un petit paramètre q indépendamment des autres. Pour connaître la proportion d'individus possédant la chaîne v parmi tous les enfants d'un individu possédant la chaîne u, il suffit de connaître le nombre de différences entre les deux chaînes u et v. Si nous notons $u(1), \cdots, u(\ell)$ les ℓ digits de la chaîne u, et de même pour la chaîne v, nous définissons la distance de Hamming entre les chaînes u et v par la quantité

$$d_H(u,v) = \mathrm{Card}\big\{k \in \{1 \cdots \ell\} \ : \ u(k) \neq v(k)\big\}\,.$$

Cette quantité compte le nombre de coordonnées qui sont différentes dans les chaînes u et v. Avec cette notation, nous pouvons définir la quantité M_{uv} comme

$$M_{uv} = q^{d_H(v,u)}(1-q)^{\ell - d_H(v,u)}\,.$$

Cette quantité pourra être interprétée de la manière suivante : parmi tous les enfants d'un individu possédant la chaîne u, la proportion d'individus possédant la chaîne v est M_{uv}. En fonction du modèle que nous étudierons, cette quantité pourra correspondre à une vitesse de réaction, une probabilité, voire un taux de mutation. En particulier, la quantité M_{uu} est égale à $(1-q)^\ell$ pour toute chaîne u, cela correspond à la proportion des enfants de la chaîne u qui n'ont subi aucune mutation. Maintenant que nous avons décrit le phénomène de mutation, introduisons le deuxième ingrédient fondamental : la sélection.

Des individus avec un génome différent vont avoir un comportement différent au sein du même environnement. Une compétition va donc s'opérer entre les différentes chaînes. Nous allons attribuer à chaque chaîne un nombre réel positif, que nous appellerons fitness, ce nombre peut être interprété comme le nombre moyen de descendants des individus qui possèdent cette chaîne. L'article [Pel96] discute les interprétations possibles de cette quantité. Nous appellerons paysage de fitness l'ensemble de ces nombres : un paysage de fitness est une fonction f de $\{0,1\}^\ell$ dans $\mathbb{R}_+$. Il serait réaliste de considérer que des chaînes peuvent avoir une fitness nulle comme dans [BK98], ce qui correspondrait à une chaîne qui ne se reproduit pas. Les phénomènes de seuil d'erreur que nous allons décrire se généralisent dans ce cas [TH07]. Cependant, nous imposons que les fitness de toutes les chaînes sont supérieures ou égales à 1, et que la fitness n'est pas la même pour toutes les chaînes. Au cours du premier chapitre, nous arriverons à démontrer quelques formules nouvelles pour un paysage de fitness général, cependant pour simplifier, nous aurons besoin de contraindre un peu plus le choix de la fonction de fitness. Nous nous concentrons sur le paysage suivant, souvent nommé paysage à un pic : toutes les chaînes ont une fitness égale à un, sauf une qui possède une fitness σ strictement supérieure à 1. Cette chaîne particulière, nous la noterons w^* et l'appellerons la master sequence. Pour des paysages de fitness généraux, Wiehe [Wie97] prétend qu'il existe toujours un seuil d'erreur mais qu'il faudrait considérer un taux de mutation qui ne serait plus proportionnel à l'inverse de la longueur des chaînes. Les individus possédant ce génome engendreront plus de descendants que les autres. Sans perdre de généralité,

nous pouvons supposer que la master sequence est la chaîne $0\cdots0$. Prenons $\sigma > 1$, la fonction de fitness f pour le paysage à un pic est donnée par

$$\forall\, u \in \{0,1\}^\ell \qquad f(u) \;=\; \begin{cases} \sigma & \text{si} \quad u = w^*\,, \\ 1 & \text{si} \quad u \neq w^*\,. \end{cases} \tag{1}$$

D'autres paysages de fitness sont envisagés dans la littérature, se référer par exemple à [SH06], il est par exemple possible de prendre un paysage de fitness aléatoire [Gil83], gaussien pour [Xia+07] pour modéliser la fluctuation naturelle due à l'environnement. Il est aussi possible de faire varier le paysage avec le temps pour simuler les variations dans l'environnement [NS02]. D'autres généralisations sont possibles : la fitness pourrait dépendre du nombre de uns dans la chaîne [CD16a], ou bien ne dépendre que du premier digit [TS04]. Ce dernier cas est souvent comparé au paysage à un pic, et la question de savoir lequel des deux est le meilleur reste à trancher, on parle de survival of the flattest [Ast+13], contre survival of the fittest pour le paysage à un pic. Il est aussi possible d'envisager plusieurs pics [Saa+06].

2.3 Les équations d'évolution

Nous disposons maintenant d'une description mathématique des deux ingrédients fondamentaux et nous pouvons décrire les équations d'évolution d'Eigen. Manfred Eigen décrit dans son article fondateur [Eig71] un système d'équations de réactions chimiques qui se produisent toutes en même temps à des vitesses différentes. Son but était de comprendre l'origine de la vie sur Terre, il modélise pour cela l'évolution d'une population infinie de macromolécules dans une soupe primitive. Dans la création de la chaîne u, toutes les chaînes v entrent en jeu proportionnellement à leur concentration dans la population, à leur fitness et au taux de mutation M_{vu}. Nous noterons x_u la proportion d'individus possédant la chaîne u pour génome. Un terme de destruction, noté $\lambda(t)$, est nécessaire pour s'assurer que la somme des concentrations reste égale à 1. L'équation d'évolution de la proportion d'individus ayant la chaîne u est

$$x'_u(t) = \sum_v f(v)x_v(t)M_{vu} - x_u(t)\lambda(t)\,. \tag{2}$$

Nous pouvons exprimer le terme de destruction $\lambda(t)$ en fonction des quantités que nous connaissons déjà : sommer ces équations sur toutes les chaînes u conduit à

$$\Big(\sum_u x_u(t)\Big)' = \sum_v f(v)x_v(t)\Big(\sum_u M_{vu}\Big) - \Big(\sum_u x_u\Big)\lambda(t)\,.$$

Comme la somme des concentrations est constante égale à 1, nous pourrons déterminer $\lambda(t)$: c'est la fitness moyenne de la population :

$$\lambda(t) = \sum_v f(v)x_v(t)\,. \tag{3}$$

Cette quantité dépend donc de toutes les concentrations $x_v(t)$. Dans le cas du paysage à un pic, la fitness moyenne et la proportion de master sequence sont directement reliées. Le système (2) n'est pas linéaire, ce qui le rend difficile à étudier. Des

solutions exactes dans des cadres différents ont été obtenues dans [TM74], [JER76], ou encore [Gal97]. Pour une synthèse des résultats, voir [EMS88]. Plusieurs travaux [BK83], [Jon77], [SS82] ont montré la convergence vers une unique solution stationnaire, [Mar+12] calcule le temps pour que la population atteigne cette limite. A partir de maintenant et dans toute cette thèse, nous nous plaçons à l'équilibre. Nous supposerons que toutes les quantités ont atteint leur limite et nous noterons simplement x_u pour $x_u(t)$ et λ pour $\lambda(t)$.

Voilà donc le cadre général, les master sequences engendrent plus de descendants, mais à chaque reproduction, les mutations entrent en jeu et peuvent modifier certains zéros en uns.

• Si les mutations ne sont pas trop fortes, la population à l'équilibre est constituée de master sequences et d'individus avec peu de uns dans leur génome. On dit que ces individus qui gravitent autour de la master sequence forment le nuage de mutants. Cette structure est appelée la quasiespèce, ce terme est apparu pour la première fois dans le cylce d'articles "The Hypercycle" écrit par M. Eigen et Peter Schuster [ES78], pour une description récente et une synthèse des résultats mathématiques, voir [BNS19]. Il est fondamental de comprendre cette structure : les master sequences alimentent le nuage de mutants, car ce sont elles qui engendrent la plupart des nouveaux individus. D'un autre côté, comme le nuage de mutants est constitué d'individus proches de la master sequence, il est possible que certains mutants subissent les bonnes mutations et engendrent des master sequences. Même si ces mutations sont peu nombreuses, elles participent à la stabilité de la quasiespèce. Beaucoup des résultats de ce manuscrit sont dédiés à essayer de comprendre l'importance de ces "mutations de retour". Dans ce cas, Nous parlerons de régime ordonné.

• En revanche, si les mutations sont trop fortes, les master sequences vont engendrer beaucoup d'individus mais peu de nouvelles master sequences et elles finiront par disparaître de la population. Dans ce cas, la population à l'équilibre est constituée d'individus dont les chaînes sont uniformément distribuées sur $\{0,1\}^\ell$, on parle souvent de régime de chaos.

Nous allons pouvoir caractériser mathématiquement l'équilibre que nous avons décrit au début de ce chapitre. Pour déterminer si les mutations sont trop fortes, nous allons simplement regarder s'il reste des master sequences dans la population à l'équilibre. C'est ce phénomène qu'Eigen a baptisé le "seuil d'erreur", une transition de phase qui s'observe sur la composition de la population à l'équilibre. Cependant avant de développer, nous devons donner une précision sur les paramètres du modèle. Si les paramètres ℓ et q restent finis, il y a peu d'espoir de pouvoir montrer des résultats généraux. Nous travaillerons toujours dans un régime asymptotique défini par la convergence simultanée des paramètres

$$\ell \to \infty, \qquad q \to 0\,.$$

Nous allons ajouter une hypothèse supplémentaire, c'est que le produit de ces deux quantités reste borné et tend vers un nombre réel que nous noterons a.

$$\ell q \to a \in]0, \infty[\,.$$

Le produit ℓq représente le nombre moyen de mutations à chaque événement de reproduction. Cette dernière hypothèse se comprend assez bien en considérant la quantité $M_{uu} = (1-q)^\ell$, la proportion d'enfants qui ont le même génome que leur parent. Pour que cette proportion tende vers un nombre réel fini, il faut que le produit ℓq converge. Nous pouvons maintenant énoncer le seuil d'erreur : la valeur critique du paramètre de mutation qui sépare les deux régimes que nous avons décrits précédemment. Le critère pour déterminer l'allure de la population à l'équilibre est de comparer la limite du produit ℓq avec le logarithme de la fitness maximale σ. Si a est supérieur à $\ln \sigma$, alors la concentration de master sequences à l'équilibre converge vers le rapport

$$\frac{\sigma e^{-a} - 1}{\sigma - 1},$$

pour plus de détails sur les limites des différentes quantités du modèle, voir [Dal18]. Si a est inférieur à $\ln \sigma$, alors à l'équilibre la concentration de master sequences tend vers 0. Cette dichotomie est souvent interprétée sur le paramètre de mutation q et résumée par la définition du paramètre critique q^* :

$$q^* = \frac{\ln \sigma}{\ell}.$$

Cette égalité est valable asymptotiquement et elle correspond au premier ordre d'un développement asymptotique de q^*. L'enjeu du premier chapitre est de pouvoir caractériser l'allure de la population à l'équilibre si le paramètre de mutation vérifie

$$q = \frac{\ln \sigma}{\ell} + \frac{c}{\ell^\alpha},$$

pour certaines valeurs de c et α. Pour cela, nous allons utiliser l'hypothèse d'indépendance sur l'évolution des digits des chaînes pour démontrer la formule suivante vérifiée par la fitness moyenne λ, déjà présente dans [KS02] ou [SBN14],

$$\frac{1}{2^\ell} \sum_{k=0}^{\ell} \binom{\ell}{k} \frac{1}{\frac{\lambda}{(1-2q)^k} - 1} = \frac{1}{\sigma - 1}.$$

Cette formule correspond à l'équation (1.12), les sections 2 et 3 du premier chapitre en constituent la preuve. A partir de cette formule nous démontrerons rigoureusement le phénomène de transition de phase décrit ci-dessus à la section 4. Elle nous permettra également, à la section 7, de prolonger le développement du point critique et nous montrerons que

$$q^* = \frac{\ln \sigma}{\ell} + \frac{\sigma - 3}{2(\sigma - 1)} \frac{\ln^2 \sigma}{\ell^2} + o\Big(\frac{1}{\ell^2}\Big). \tag{4}$$

Nous nous servirons aussi des calculs pour démontrer au passage une formule sur la distance moyenne à la chaîne w^* dans un paysage de fitness général et valable pour toutes les valeurs de ℓ et q. Si nous notons $\overline{H}$ la distance moyenne de la population à la master sequence donnée par

$$\overline{H} = \sum_u d_H(u, w^*) x_u,$$

et que nous notons $\overline{FH}$ la moyenne du produit de la distance à la master sequence et de la fitness,

$$\overline{FH} = \sum_u f(u) d_H(u, w^*) x_u \,,$$

alors ces deux quantités et la fitness moyenne sont reliées à travers l'égalité qui constitue notre théorème 3 :

$$\lambda \overline{H} = \ell q \lambda + (1 - 2q) \overline{FH} \,.$$

Dans toute la suite, nous nous placerons toujours dans le paysage à un pic et au point critique $a = \ln \sigma$. Notre but va être de comprendre comment les master sequences sont créées. Tout d'abord, elles sont créées par des reproduction d'autres master sequences avec un certain taux que nous appellerons r_0, où nous notons

$$r_0 = \frac{\sigma(1-q)^\ell - 1}{\sigma - 1} \,.$$

Mais elles sont aussi créées par des reproductions de mutants qui subissent les bonnes mutations, cette contribution est difficile à estimer car pour la calculer, il est nécessaire de connaître la structure du nuage de mutants. Il nous faudrait connaître la proportion d'individus qui sont à distance 1 de la master sequence, la proportion des individus à distance 2, etc...Manfred Eigen parle déjà du rôle du nuage de mutant dans la stabilité de la quasiespèce et c'est ce rôle que nous voulons comprendre. D'ailleurs, remarquons que le seuil d'erreur (4) correspond à une situation où les master sequences ne sont pas stables par reproduction, r_0 est négatif mais les mutations de retour compensent exactement cette défaillance. Avant de rentrer dans ces détails, nous allons simplifier un peu l'espace des chaînes possibles en rassemblant toutes les chaînes qui ont la même classe de Hamming, c'est-à-dire les chaînes qui ont le même nombre de uns. Pour une chaîne u, nous noterons $C(u)$ sa classe de Hamming,

$$C(u) = d_H(u, w^*) \,.$$

Nous notons x_i la proportion d'individus dont la classe de Hamming est i,

$$x_i = \sum_{u : C(u) = i} x_u \,.$$

Les équations différentielles d'Eigen s'écrivent aussi pour les proportions des individus dans chaque classe de Hamming. Nous montrerons au chapitre 2 que les proportions x_i à l'équilibre sont régies par le système

$$0 = -(\sigma - 1) x_i x_0 + \Big(f(i) M_{ii} - 1 \Big) x_i + \sum_{k=0,\, k \neq i}^{\ell} f(k) x_k M_{ki} \,, \tag{5}$$

où nous avons noté M_{ij} la proportion des enfants de classe j qu'engendre un individu de classe i. A partir de ce système, nous allons construire des encadrements progressifs pour les proportions des individus dans chaque classe d'après le système (5). La proportion de master sequence à l'équilibre x_0 vérifie l'équation

$$x_0^2 = r_0 \, x_0 + \frac{1}{\sigma - 1} \sum_{k \geq 1}^{\ell} x_k M_{k0} \,.$$

Nous distinguons deux termes dans cette équation. Le premier, $r_0 x_0$, correspond aux master sequences qui se reproduisent sans mutations. Le terme avec la somme correspond, lui, aux mutations de retour, c'est ce terme qui représente la possibilité pour les mutants de recréer des master sequences. Cette quantité est difficile à estimer car elle dépend des proportions des individus dans les autres classes x_k. Notre stratégie est la suivante : nous allons trouver une façon de minorer et majorer ce terme pour pouvoir encadrer la proportion de master sequence. Tout d'abord, nous obtenons moins de master sequences si nous remplaçons ce terme par 0. Effectuer cette minoration consiste à oublier toutes les mutations de retour, nous en déduirons un minorant de la proportion x_0. Pour obtenir un majorant de cette proportion, nous dirons que tous les mutants sont dans la première classe Hamming et qu'il leur suffit donc d'une seule bonne mutation pour engendrer une master sequence. Nous pourrons calculer ce majorant car nous savons que les mutants constituent une fraction $1 - x_0$ de la population. De cette façon, nous pourrons déduire un encadrement sommaire sur la proportion de master sequences. Mais nous n'allons pas nous arrêter là, si nous considérons maintenant la proportion d'individus dans la classe de Hamming 1, cette proportion est régie par l'équation issue du système (5) :

$$\Big(1 - M_{11} + (\sigma - 1)x_0\Big)x_1 = \sigma M_{01} x_0 + \sum_{k=2}^{\ell} x_k M_{k1}\,.$$

Nous retrouvons deux termes dans la création d'individus de classe 1, les master sequences qui subissent une mutation et engendrent des individus de la classe 1, mais aussi tous les individus des classes 2 et supérieures qui subissent les bonnes mutations. Nous allons traiter ce dernier terme de la même façon que précédemment : en négligeant ces mutations de retour, nous obtiendrons un minorant de la proportion x_1, en disant que tous les mutants peuvent engendrer des individus de classe 1 avec une seule bonne mutation, nous obtiendrons un majorant de x_1. En utilisant l'encadrement précédent sur la proportion de master sequence x_0, nous en déduirons donc un encadrement de la proportion x_1. Mais alors, si nous revenons maintenant sur la proportion de master sequences, nous pouvons mieux encadrer le terme qui contient les mutations de retour, puisque nous savons encadrer la proportion x_1. Nous allons donc pouvoir obtenir un meilleur encadrement sur la proportion de master sequences en écrivant l'équation qui régit x_0 comme

$$x_0^2 = r_0\, x_0 + x_1 \frac{M_{10}}{\sigma - 1} + \frac{1}{\sigma - 1} \sum_{k=2}^{\ell} x_k\, M_{k0}\,.$$

Nous pouvons alors majorer et minorer le terme avec la somme de la même façon que précédemment : d'un côté, nous disons que les individus des classes 2 et supérieures ne redonnent jamais de master sequences, d'un autre côté, que ces individus peuvent engendrer des master sequences comme s'ils étaient tous des individus de la classe 2. Cela va nous permettre d'obtenir un nouvel encadrement sur la proportion de master sequences, celui-ci va se révéler bien meilleur que le premier. Nous allons donc pouvoir recommencer le procédé pour obtenir un meilleur encadrement sur la proportion d'individus dans la classe 1. Par le même raisonnement, nous obtiendrons

un encadrement encore meilleur sur la proportion de master sequences. Malheureusement, nos deux bornes ne seront toujours pas du même ordre de grandeur après ces opérations. Nous allons devoir réaliser ces étapes une infinité de fois et nous intéresser à la limite des bornes que nous allons obtenir. Nous pourrons en déduire à la section 4 les premiers termes du développement de la proportion de master sequences à l'équilibre :

$$x_0 \sim r_0 + \frac{\sigma}{\sigma - 1} \frac{M_{01} M_{10}}{1 - M_{11}} ,$$

si cette quantité est positive. Nous obtiendrons aussi

$$x_1 \sim \frac{M_{01} x_0}{1 - M_{11}} .$$

Nous allons pouvoir continuer le développement en estimant de la même façon la proportion de master sequences et des individus des classes 1 et 2 : les minorer en ignorant les mutations de retour, les majorer en faisant comme si le reste des mutants se trouvait dans la classe 3. Cela va nous permettre de prolonger le développement précédent en prenant à nouveau la limite de ces processus encadrants. Nous effectuerons ces opérations pour un nombre arbitraire fixé de classes de Hamming. Grâce à ces processus, nous allons pouvoir identifier les contributions de chaque chaîne dans la création de master sequences et déterminer le développement asymptotique des concentrations de chaque classe. Nous notons

$$D_{ij} = \frac{M_{ij}}{1 - M_{ii} + (\sigma - 1) x_0} ,$$

puis C_{ij} la somme sur les chemins croissants $i < i_1 < \cdots < i_n < j$ des coefficients $D_{i_1, \cdots, i_n}$:

$$C_{ij} = \sum_{i = i_1 < \cdots < i_n = j} D_{i_1 i_2} \cdots D i_{n-1} i_n ,$$

et enfin $S_{ij}(k)$ la somme des coefficients D sur les chemins partant de i, montant au plus jusqu'à j puis redescendant à k :

$$S_{ij}(k) = \sum_{\substack{i = i_1 < \cdots < i_n \leq j \\ i_n > i_{n+1} > \cdots > i_m = k}} D_{i_1 \cdots i_n} .$$

Avec ces notations, nous pourrons écrire à la section 5 pour tout entier L négligeable devant ℓ dans le régime asymptotique,

$$x_k \sim \sigma x_0 S_{0L}(k) ,$$

et

$$x_0 = r_0 + \frac{\sigma}{\sigma - 1} \left(C_{01} M_{10} + C_{02} \Big(C_{21} M_{10} + M_{20} \Big) + C_{03} \Big(C_{31} M_{10} + C_{32} M_{20} + M_{30} \Big) + \cdots \right) . \tag{6}$$

Précisons que chaque terme à l'intérieur de la parenthèse est d'ordre une puissance de q différente puisque M_{ij} est d'ordre q si $j = i - 1$, en revanche, si $j = i + 1$,

la quantité M_{ij} est une constante dans le régime asymptotique. Cela s'interprète par le fait qu'il faut subir une mutation sur un 1 de la chaîne pour engendrer un individu dans une classe de Hamming inférieure et qu'il y a peu de 1 dans ces chaînes. Cette expansion nous permet de retrouver le même développement de la fitness moyenne que nous obtenons au premier chapitre sous une autre forme. Nous pensons que cette formule peut se généraliser à un paysage de fitness général. Nous ne pourrions plus regrouper les individus dans les classes de Hamming, mais en appelant master sequence la chaîne qui a la plus grande fitness, nous pourrions développer la proportion x_0 et l'écrire comme une somme de produits de coefficients M_{uv} sur toutes les chaînes intermédiaires qui mènent à la master sequence.

Jusqu'à maintenant, nous avons supposé que nous avions un nombre infini d'individus et nous avons travaillé avec leur proportion. Cependant, les populations réelles ne sont pas infinies, en particulier, le nombre d'individus est bien inférieur au nombre de chaînes possibles. C'est pourquoi il est important d'étudier des modèles qui présentent les mêmes ingrédients mais dans lesquels la population est finie.

3 Pour une population finie

De nombreux travaux ont argumenté qu'il était possible de généraliser le modèle d'Eigen dans le cadre d'une population finie : Wilke [Wil05], mais aussi [SDH12], [DSV12], ou encore [Zha97], [AF96]. Il est même possible de retrouver les mêmes phénomènes en considérant une population diploïde [AF97], [PMD10], [FOC96]. Cependant, plusieurs distinctions fondamentales vont en découler. Dans le modèle d'Eigen, les individus se reproduisent en continu, chacun engendrant un nombre infini d'enfants à chaque instant. Pour les modèles où la population est finie, cela n'est plus le cas, le nombre d'enfants d'un individu est une variable aléatoire. Une autre différence majeure est que, contrairement au modèle d'Eigen, il sera possible de ne plus avoir aucune master sequence dans la population. Dans ce cas, il faudra attendre qu'un individu subisse les bonnes mutations pour pouvoir la retrouver. La question du temps que cela prendra est centrale dans la compréhension de la dynamique de la population. Cette durée joue un rôle crucial [CD18] pour comprendre le comportement d'équilibre du modèle de quasiespèce d'Eigen dans le contexte de populations finies comme nous allons le voir dans la suite.

3.1 Un seul individu

Nous allons commencer par examiner un cas simple dans le chapitre 3 : une population composée d'un seul individu. Le modèle est discret et nous allons suivre les génomes de la lignée issue d'un individu. A chaque pas de temps, l'unique individu de la population se reproduit et donne naissance à un nouvel individu qui le remplace. Le génôme de ce nouvel individu est une copie de celui de son parent soumise aux mutations : chaque digit est changé indépendamment avec une probabilité q. Comme la population est constituée d'un unique individu, il n'y a pas de compétition entre les individus, ce modèle ne comporte donc pas de sélection. Supposons qu'au départ, l'individu n'est pas une master sequence. Nous allons déterminer le nombre moyen

de générations nécessaires pour découvrir la master sequence pour la première fois. De manière très surprenante, de nombreuses similitudes vont apparaître avec un modèle bien connu : le modèle des urnes d'Ehrenfest que nous décrivons maintenant. Considérons deux urnes et ℓ boules. Initialement, toutes les boules sont dans la seconde urne. A chaque pas de temps, une boule est choisie au hasard et déplacée dans l'autre urne. La question centrale du modèle est la suivante : En moyenne, combien de temps faut-il attendre pour que le système revienne à son état initial ? Dans [Kac47], Mark Kać répond à cette question et nous allons nous inspirer de sa méthode. Notons Y_n la classe de Hamming du n-ième individu et définissons τ^* le temps de découverte de la master sequence :

$$\tau^* \;=\; \inf\left\{\, n \geq 1 : Y_n = 0 \,\right\}.$$

Nous allons calculer l'espérance de cette variable aléatoire τ^*. Il est important que nous comprenions le temps nécessaire pour découvrir la master sequence en partant de n'importe quelle classe k. A la section 3 du chapitre 3, nous encadrerons ce temps comme

$$E\big(\tau^* \,\big|\, Y_0 = 0\big) \;\leq\; E\big(\tau^* \,\big|\, Y_0 = k\big) \;\leq\; E\big(\tau^* \,\big|\, Y_0 = \ell\big)\,. \tag{7}$$

En effet, le temps de découverte de la master sequence est majoré par le temps qu'il faut pour atteindre la classe 0 en partant de la classe ℓ. Cela correspond à la situation où, commençant avec uniquement des uns nous attendons jusqu'à obtenir une chaîne avec uniquement des zéros. Nous montrons que

$$E\big(\tau^* \,\big|\, Y_0 = \ell\big) \;=\; \sum_{k=1}^{\ell} \binom{\ell}{k} \frac{1-(-1)^k}{1-(1-2q)^k}\,.$$

Ce temps est aussi minoré par le temps de retour moyen à 0 en partant de la classe 0, donné par

$$E\big(\tau^* \,\big|\, Y_0 = 0\big) \;=\; 2^\ell\,.$$

En revenant à notre encadrement, nous obtenons, pour toute classe initiale k,

$$2^\ell \;\leq\; E\big(\tau^* \,\big|\, Y_0 = k\big) \;\leq\; \sum_{k=1}^{\ell} \binom{\ell}{k} \frac{1-(-1)^k}{1-(1-2q)^k} \;\leq\; \frac{2^\ell}{q}\,.$$

Ces inégalités montrent que le temps de découverte de la master sequence est d'ordre 2^ℓ. Cela signifie qu'en moyenne, il faut attendre autant de générations que le nombre total de chaînes pour retrouver la master sequence. Pour $0 \leq j \leq \ell$, nous définissons aussi le temps d'atteinte de la classe de Hamming j par

$$\tau_j \;=\; \inf\left\{\, n \geq 1 : Y_n = j \,\right\}.$$

A la section 5, nous montrons que, pour $1 \leq j \leq \ell$, le temps de retour moyen à la classe j est

$$E\big(\tau_j \,\big|\, Y_0 = j\big) \;=\; \frac{2^\ell}{\binom{\ell}{j}}\,.$$

Ces formules sont exactement les mêmes que celles que Mark Kać a obtenu pour le modèle d'Ehrenfest. Nous allons employer la même méthode que celle utilisée par Mark Kać : nous allons calculer les quantités $E\big(\tau_j \,\big|\, Y_0 = i\big)$ pour $0 \leq i, j \leq \ell$. Pour cela, nous allons calculer les fonctions génératrices des événements $\{\, Y_n = j \,\}$ et de la variable aléatoire τ_j, et nous allons les relier à travers une équation fonctionnelle bien connue.

3.2 Un nombre d'individu arbitraire

Il est maintenant temps de travailler dans un modèle général, où la taille de la population est arbitraire. Nous ajoutons un nouveau paramètre, que nous noterons m et nous considérons une population constituée de m individus. Nous travaillerons toujours avec des modèles discrets, la population ne change qu'en des instants entiers et nous parlons de générations. A chaque génération, la population est soumise à des événements de mutation et de sélection. La façon de passer d'une génération à une autre va devoir généraliser les mécanismes du modèle d'Eigen. Elle pourrait être implémentée de différentes façons, cependant Dalmau a montré dans [Dal16] que différents modèles discrets semblent liés à travers une sorte d'universalité. Dalmau a en effet démontré les mêmes formules limites pour les modèles discrets de Wright-Fisher, de Moran et de Galton-Watson. Décrivons le modèle avec lequel nous travaillerons : le modèle de Moran, introduit par Moran [Mor58] en 1958. A chaque pas de temps un individu est choisi dans la population pour être un parent, ce choix n'est pas uniforme car les master sequences ont un avantage sélectif. C'est lors de ce choix qu'intervient la sélection dans le modèle : les master sequences ont σ fois plus de chances d'être choisies que toutes les autres chaînes, avec $\sigma > 1$. Si nous notons $\lambda(t)$ la fitness moyenne de la population au temps t, une master séquence est choisie avec probabilité $\sigma/\lambda(t)$, un individu qui n'est pas une master sequences est choisi avec probabilité $1/\lambda(t)$. Les master sequences ont donc de meilleures chances d'engendrer des descendants. L'individu choisi est ensuite dupliqué, mais cette copie est sujette aux mutations : chaque digit de son génôme est changé indépendamment avec probabilité q. Une nouvelle population est alors formée à partir de ce nouvel individu et de tous les individus de la génération précédente sauf un, choisi uniformément au hasard. De cette façon, la taille de la population reste constante au cours du temps, égale à m. Les relations entre la quasiespèce et le modèle de Moran ont été creusées par Cerf et Dalmau dans [CD16b].

Les ingrédients fondamentaux du modèle d'Eigen sont présents dans ce modèle, et de fait, ce modèle converge vers celui d'Eigen dans un régime adéquat, comme l'a montré Dalmau [Dal14]. La question qui arrive naturellement est donc celle du seuil d'erreur, et en effet un seuil d'erreur est également présent dans ce modèle, comme l'a montré Cerf dans [Cer15]. Comment la transition de phase va-t-elle se traduire dans ce modèle ? Rappelons-nous que, dans le modèle d'Eigen, le seuil d'erreur n'est défini que dans un régime asymptotique des paramètres. Si les paramètres sont finis, le seuil d'erreur n'est pas bien défini. Il va se passer le même phénomène dans le modèle de Moran : si les paramètres m, ℓ et q sont finis, on ne peut pas déterminer clairement de paramètre critique. Nous nous placerons donc à nouveau dans un

certain régime asymptotique défini par la convergence simultanée des paramètres

$$\ell \to \infty, \qquad q \to 0, \qquad m \to \infty\,, \tag{8}$$

et nous nous placerons au point critique

$$\ell q \to \ln \sigma\,.$$

Dans le modèle d'Eigen, le paramètre critique séparait un régime ordonné d'un régime de chaos et cette distinction pouvait se lire dans la proportion de master sequence à l'équilibre. La première chose à remarquer est que nous nous placions à l'équilibre : toutes les quantités avaient atteint leur limite et les proportions n'évoluaient plus. Ici, les choses vont être différentes car des quantités aléatoires interviennent à chaque génération : la composition de la population ne restera jamais constante. Il va donc nous falloir être précis quant à la définition de notre paramètre critique, car plusieurs critères seront pertinents, et conduiront à des seuils distincts. Examinons de plus près la dynamique de la population.

Commençons, au temps $t = 0$, par une population composée d'une seule master sequence et de $m - 1$ autres individus qui ne sont pas des master sequences. Définissons N_t comme le nombre de master sequences dans la population à l'instant t et suivons son évolution au cours du temps. Lorsque des master sequences sont présentes dans la population, le rôle de la sélection est de les conserver. Les individus master sequences sont plus souvent choisis pour être parent, et si aucune mutation n'intervient ils donnent naissance à une nouvelle master sequence pour la génération suivante. Cette phase où les master sequences sont présentes dans la population est appelée la **phase quasiespèce** : les master sequences occupent en moyenne une proportion significative de la population, et sont accompagnées d'un nuage de mutants constitué d'individus génétiquement proches de la master sequence, à l'image de la quasiespèce d'Eigen. A un moment donné, en raison de fluctuations aléatoires, les master sequences vont disparaître. Cet instant que nous appellerons le **temps de persistance** τ_0 est défini comme

$$\tau_0 = \inf\left\{\, t \in \mathbb{N} \,\middle|\, N_t = 0 \,\right\}.$$

Au temps τ_0, la population ne contient plus aucune master sequence et entre dans une nouvelle phase : la **phase neutre**. Les individus continuent de se reproduire, ils sont toujours soumis aux mutations mais la sélection ne joue plus aucun rôle, jusqu'à ce qu'un évènement particulier se produise : un individu qui subit les bonnes mutations donne naissance à une master sequence. Le nombre de générations que dure la phase neutre que nous appellerons le **temps de découverte**, est défini par

$$\tau^* = \inf\left\{\, t \geq \tau_0 \,\middle|\, N_t \neq 0 \,\right\} - \tau_0\,.$$

A cet instant, nous entrons dans une nouvelle phase quasiespèce, et le schéma se répète. Nous n'avons donc pas de situation d'équilibre comme nous l'avions dans le modèle d'Eigen, mais les deux phases vont alterner à l'infini. Nous verrons que le paramètre de mutation q va avoir un grand impact sur cette dynamique. L'augmentation du taux de mutation va réduire la stabilité de la phase quasiespèce car

les descendants des master sequences sont alors moins susceptibles d'être des master sequences. La phase neutre, elle, est à peine modifiée, comme nous l'avons vu dans le chapitre précédent. Dans le régime asymptotique (8), Cerf a prouvé dans [Cer15] au chapitre 10.5, que l'espérance du temps τ^* peut être estimé par 2^ℓ, ce qui correspond au nombre total de chaînes. Plus précisément, nous avons l'asymptotique suivante :

$$\lim \frac{1}{\ell} \ln E(\tau^*) = \ln 2 \,.$$

Nous avons déjà parlé du temps de découverte de la master sequence pour le cas $m = 1$. La formule que nous avions obtenue est encore valable dans le cas où m est quelconque. Cela signifie simplement qu'une population composée de m individus ne va pas significativement plus vite pour découvrir la master sequence. Pour définir un seuil d'erreur, nous allons nous appuyer sur la comparaison des durées de ces deux phases. Notre but principal est donc ici d'estimer l'espérance du temps de persistance τ_0 dans le régime asymptotique. Cette quantité est intéressante en elle-même, comme le souligne [EMS07] : il est crucial de pouvoir estimer le temps de vie de la quasiespèce pour espérer comprendre le modèle d'Eigen. Nos deux derniers chapitres vont permettre d'estimer cette espérance. L'idée générale est de se ramener à des processus de vie et de mort pour lesquels nous connaissons les probabilités de transition. Pour ces processus, il existe une formule explicite permettant de calculer le temps d'atteinte de n'importe quel point. Comme la génération $t+1$ diffère d'un seul individu de la génération t, le processus N_t évolue effectivement selon un processus de vie et de mort. Attribuons une notation à ses probabilités de transition. Pour k entre 0 et $m-1$, nous notons δ_k la probabilité que N passe en une étape de k à $k+1$:

$$\forall t \geq 0 \quad \forall k \in \{0, \cdots, m-1\} \quad \delta_k = P\Big(N_{t+1} = k+1 \,|\, N_t = k\Big) \,.$$

Nous notons de même γ_k la probabilité que N passe en une étape de k à $k-1$:

$$\forall t \geq 0 \quad \forall k \in \{1, \cdots, m\} \quad \gamma_k = P\Big(N_{t+1} = k-1 \,|\, N_t = k\Big) \,.$$

Nous notons $\pi_0 = 1$ et

$$\forall i \in \{1, \cdots, m\} \qquad \pi_i = \frac{\delta_1 \cdots \delta_i}{\gamma_1 \cdots \gamma_i} \,.$$

Alors nous pouvons énoncer une formule explicite pour exprimer l'espérance du temps de persistance τ_0 en partant de $N_0 = 1$:

$$E\big(\tau_0\big) = \sum_{i=1}^{m} \frac{1}{\delta_i} \pi_i \,. \tag{9}$$

Cependant, les probabilités de transition du processus (N_t) dépendent de la structure du nuage de mutants. Nous allons suivre la stratégie de Nowak et Schuster dans [NS89] pour simplifier le processus initial en l'encadrant entre deux processus de vie et de mort. Nous allons donc à nouveau séparer l'espace des chaînes en

deux classes. La première classe T_0 regroupe toutes les master sequences, toutes les autres séquences sont mises dans la seconde classe T_1, c'est le nuage de mutants. Ecrivons P_{ij} pour la probabilité qu'un individu de type T_i donne naissance à un individu de type T_j, pour $i, j \in \{0, 1\}$. Certaines de ces probabilités peuvent être calculées immédiatement, par exemple, P_{00} est la probabilité qu'une master sequence donne naissance à une master sequence, pas un seul digit ne doit être changé, donc $P_{00} = M_{00} = (1-q)^\ell$, et bien sûr, $P_{01} = 1 - M_{00}$. Cependant, la probabilité pour un individu de type T_1 de donner naissance à une master sequence dépend du nombre de digits de son génome qui sont différents de 0. Cette probabilité est donc hors de portée si l'on ne fait aucune hypothèse sur la répartition de la population dans les différentes classes de Hamming. C'est pour évaluer cette quantité que Nowak et Schuster ont supposé que les chaînes étaient uniformément réparties sur l'ensemble $\{0, 1\}^\ell$, en prenant

$$P_{10} = \sum_{k=1}^{\ell} \frac{\binom{\ell}{k}}{2^\ell - 1} q^k (1-q)^{\ell-k} = \frac{1 - (1-q)^\ell}{2^\ell - 1} \,.$$

En réalité, les individus qui ne sont pas des master sequences sont génétiquement proches des master sequences. L'hypothèse de Nowak et Schuster sur P_{10} est très forte car elle sous-estime largement la probabilité qu'un individu qui n'est pas une master sequence donne naissance à une master sequence. Ils rendent cette probabilité d'ordre $1/2^\ell$, or, d'après les développements que nous avons obtenus, elle semblerait plutôt être d'ordre $r_0 q$, ce qui est bien plus grand. Dans notre travail, nous ne ferons aucune hypothèse simplificatrice de ce type, nous allons majorer puis minorer la probabilité P_{10} de la même manière que nous le faisons dans la première étape du chapitre 2 : pour la minoration en ignorant les mutations de retour, pour la majoration en considérant tous les mutants comme des individus de classe 1. Cela nous conduira à définir deux processus qui conduiront à la même estimation pour les premiers termes de l'espérance du temps de persistance. Notre méthode consiste à manipuler la formule exacte (9) pour écrire l'espérance du temps sous la forme

$$E\big(\tau_0\big) = K \sum_{i=1}^{m-1} \exp\left(mF\Big(\frac{i}{m}\Big) + G\Big(\frac{i}{m}\Big)\right).$$

Nous allons ensuite implémenter une méthode de Laplace en majorant uniformément la fonction G. Les contributions principales dans cette somme proviennent des termes dont les indices sont proches du maximum de la fonction F, qui se trouve être exactement le taux de reproduction des master sequences r_0. Pour pouvoir effectuer cette méthode de Laplace, nous aurons besoin que $r_0 > 0$. Une fois que nous aurons calculé la somme, nous arriverons à l'expression (4.45) du temps :

$$E\big(\tau_0\big) = \exp\left(m\varphi\big(M_{00}\big) + O\left(\big(1+mq\big)\ln m + \frac{mq^2}{\sigma - 2 + r_0}\right)\right),$$

avec

$$\varphi(x) = \frac{\sigma\big(1-x\big)\ln\frac{\sigma(1-x)}{\sigma-1} + \ln(\sigma x)}{1 - \sigma(1-x)} \,.$$

Cette fonction φ apparaît déjà dans [Cer15] comme la limite

$$\lim \frac{1}{m} \ln E\big(\tau_0\big) = \varphi(e^{-a})\,.$$

En effectuant un développement limité de la fonction φ, nous obtenons que le temps de persistance admet le développement suivant :

$$E\big(\tau_0\big) = \exp\left(\frac{\sigma-1}{2}\, m{r_0}^2 + O\Big((1+mq)\ln m + m{r_0}^3 + \frac{mq^2}{\sigma-2+r_0}\Big)\right). \tag{10}$$

Notre borne inférieure et notre borne supérieure diffèrent d'un facteur mq, nos résultats seront donc beaucoup plus précis si nous ajoutons l'hypothèse

$$\frac{m}{\ell} \to 0\,.$$

Sous cette hypothèse et sous la condition que $\sigma \neq 2$, l'espérance du temps de persistance est d'ordre

$$E\big(\tau_0\big) \sim P(m) \exp\left(\frac{m(\ell q - \ln\sigma)^2}{2(\sigma-1)}\right), \tag{11}$$

où $P(m)$ est un terme qui croît au plus comme un polynôme en m. Cette expression permettra de caractériser certains paramètres critiques, comme nous le verrons à la section 4.

3.3 Vers la mesure stationnaire ?

Dans le chapitre 5, nous proposons une piste pour améliorer l'estimation sur le temps de persistence. Cependant, quelques points mathématiques restent encore à expliciter, et nous espérons concrétiser cette stratégie dans un travail futur. Les processus encadrants que nous avons décrits dans le cadre du modèle d'Eigen pourraient également s'appliquer dans le modèle de Moran : Encadrer le nombre de master sequence sans aucune autre approximation, puis le nombre d'individus dans la première classe avant de revenir sur le nombre de master sequences. Le chapitre 4 ne constitue en fait que la première étape de ces encadrements itératifs, et il est raisonnable de penser que nous pouvons améliorer le développement du temps de persistence en implémentant toutes les étapes. Pour des processus de vie et de mort, il existe aussi une formule explicite pour calculer la mesure invariante en fonction des probabilités de transition :

$$\begin{aligned} \forall i \in \{1, \cdots, m\} \qquad \mu(i) &= \delta_0 \frac{\pi_i}{\delta_i}\,, \\ \mu(0) &= 1\,. \end{aligned} \tag{12}$$

Comme nous l'avons fait pour le temps de persistance, nous allons pouvoir écrire cette quantité sous la forme

$$\forall i \in \{1, \cdots, m-1\} \qquad \mu(i) = \delta_0\, T\, g\Big(\frac{i}{m}\Big)\, e^{mF(i/m)}\,.$$

La méthode de Laplace pourrait encore nous servir pour estimer la masse de cette mesure afin de déterminer la mesure de probabilité invariante.

La stratégie est la suivante. Nous commencerions par établir une formule générale pour une classe de processus de vie et de mort que nous pourrions ensuite appliquer aux proportions d'individus dans les différentes classes de Hamming. Nous l'appliquerions en premier lieu sur les processus encadrants de la proportion de master sequences obtenus comme ci-dessus, d'une part en négligeant les mutations de retour, d'autre part en faisant comme si tous les mutants faisaient partie de la classe de Hamming 1. Nous obtiendrions ainsi une première approximation pour la mesure invariante ν_0 de la proportion de master sequences. Nous travaillerions ensuite sur la proportion d'individus dans la classe de Hamming 1 en supposant que la proportion de master sequences est distribuée selon l'une des deux mesures obtenues précédemment. Après avoir appliqué notre formule générale pour encadrer la mesure invariante de la proportion d'individus dans la classe de Hamming 1, nous devrions intégrer ces mesures contre la mesure ν_0. Le problème auquel nous faisons face est le suivant : La mesure invariante du processus de Markov pour les deux premières classes n'est pas nécéssairement invariante pour le processus du nombre de master sequences quand elle est conditionnée au nombre d'individus dans la classe 1. Ces processus encadrants itératifs n'ont donc a priori aucune raison d'être reliés à la mesure invariante du processus global. Pour réussir à l'approcher, il faudrait montrer que le nombre d'individus dans la classe 1 s'approche beaucoup plus vite de son équilibre que le nombre de master sequences. Nous pensons que c'est le cas : le processus des deux classes devrait suivre une dynamique lent-rapide, le nombre de master sequences étant la composante lente. Si nous arrivons à approcher la mesure invariante du processus, nous aurons accès à la masse qu'elle attribue au singleton $\{0\}$, qui est exactement l'inverse du temps de disparition des master sequences.

Ce projet n'est donc pas achevé et il reste encore du travail pour faire aboutir ce programme. Ce que nous proposons dans ce chapitre est une stratégie possible pour approcher la mesure invariante.

4 Plusieurs critères pour définir un seuil d'erreur

Définir rigoureusement un paramètre critique est un problème délicat, comme l'écrit Ellen Baake dans [BG00]. Dans le modèle d'Eigen, où la population est infinie, le développement du seuil d'erreur fait intervenir seulement les paramètres ℓ, q et σ. Nous nous attendons à ce que l'ajout de la contrainte sur la taille de la population intervienne dans le développement de ce seuil d'erreur et nous nous posons la question de l'ordre de grandeur du terme correctif dû à cette nouvelle contrainte. Parmi les travaux qui ont tenté de généraliser le seuil d'erreur à une population finie, nous pouvons citer les articles de Campos et Fontanari [CF99] [CF98], ou encore celui de Nowak et Schuster [NS89] dont nous aurons l'occasion de reparler. Nous allons voir trois critères possibles pour définir un seuil d'erreur et ces définitions conduiront à un développement du paramètre critique différent. Le premier consistera à comparer les durées des deux phases : un point critique du modèle est atteint lorsque ces deux phases ont approximativement la même durée. Nous considèrerons ensuite un

critère qui sera défini seulement à partir de la phase quasiespèce et qui consistera à observer l'ordre de grandeur du temps de persistance. Le dernier critère consistera à regarder la mesure stationnaire du processus de la proportion de master sequences et ses points critiques.

4.1 En considérant l'équilibre

Ce premier critère ne se repose pas sur la dynamique, mais plutôt sur l'équilibre de la population. Il consiste à comparer la durée des deux phases. Si le rapport $E(\tau^*)/E(\tau_0)$ tend vers zéro, alors la phase quasiespèce durera beaucoup plus longtemps que la phase neutre. Au bout d'un temps très long, la population aura passé quasiment tout son temps dans la phase quasiespèce. A l'inverse, si ce rapport tend vers l'infini, c'est dans la phase neutre que la population passera la majorité de son temps. Pour que cette approche soit significative, il faut que la population passe de nombreuses fois de la phase neutre à la phase quasiespèce et inversement. En comparant la durée de la phase neutre avec notre estimation (11), nous pouvons déduire le point où les deux phases sont tout aussi stables : pour que les deux temps soient du même ordre, il faut que

$$\big(\ell q - \ln\sigma\big)^2 = \frac{2(\sigma-1)\ell\ln 2}{m}\,,$$

ce qui conduit au point critique

$$q^* = \frac{\ln\sigma}{\ell} - \frac{\sqrt{2(\sigma-1)\ln 2}}{\sqrt{\ell m}}\,. \tag{13}$$

4.2 En considérant la dynamique

La phase quasiespèce est assez stable pour des petites valeurs du paramètre q, et le temps de persistance croît de façon exponentielle en fonction du paramètre m. Cependant, il existe un point q^* au-dessus duquel l'espérance de ce temps croît seulement comme un polynôme en m. A partir de la formule (11), nous déduisons que si q^* admet le développement asymptotique

$$q^* = \frac{\ln\sigma}{\ell} - \frac{C}{m^\alpha \ell^\beta}\,,$$

alors, le temps de persistance est d'ordre

$$E\big(\tau_0\big) = P(m)\exp\left(\frac{C^2}{2(\sigma-1)}\, m^{1-2\alpha}\,\ell^{2(1-\beta)}\right).$$

Ainsi, dès que $\alpha \geq 1/2$ et $\beta \geq 1$, ce temps croît moins vite qu'un polynôme. Le point suivant peut donc être considéré comme critique :

$$q^* = \frac{\ln\sigma}{\ell} - \frac{C}{\ell\sqrt{m}}\,. \tag{14}$$

4.3 Les variations de la mesure stationnaire

Dans [NS89], Nowak et Schuster ont cherché un paramètre critique dans une version modifiée du modèle de Moran. Dans leur cadre, le temps est continu et ils travaillent avec les générateurs infinitésimaux des probabilités de transition, qui correspondent à

$$\delta_k = \Big(1 - \frac{k}{m}\Big)\Big(P_{10} + \sigma P_{00}\frac{k}{m}\Big)\,,$$

et

$$\gamma_k = \frac{k}{m}\Big(P_{11} + P_{01}\frac{k}{m}\Big)\,.$$

Ils s'intéressent aux points extrêmaux de la mesure stationnaire de la proportion des master sequences. Comme nous l'avons vu dans la formule (12), elle dépend des quantités π_k qui sont telles que

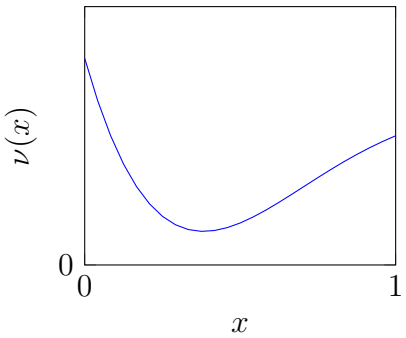
$$\pi_k = \pi_{k-1}\frac{\delta_{k-1}}{\gamma_k}\,.$$

Afin de savoir si π_k croît ou décroît, nous devons savoir laquelle des deux probabilités δ_{k-1} ou γ_k est la plus grande. Nowak et Schuster ont donc étudié la fonction ζ, avec

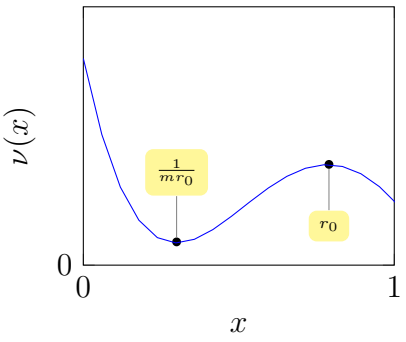
$$\zeta\Big(\frac{k}{m}\Big) = \delta_{k-1} - \gamma_k\,.$$

Les zéros de la fonction ζ nous conduiront aux points extrémaux de la mesure stationnaire. Si ζ est positif, alors la mesure croît, et elle décroît lorsque ζ est négatif. Dans leur modèle à temps continu, la fonction ζ est un polynôme de degré 2 en la variable $(1-q)^\ell$, qui peut être facilement résolu. Il conduit à deux zéros, l'un d'ordre $1/mr_0$, l'autre d'ordre r_0, donc les deux racines sont du même ordre de grandeur lorsque ${mr_0}^2$ est borné.

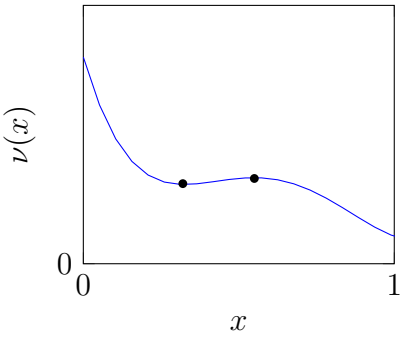
Nous avons également effectué les calculs analogues dans notre modèle à temps discret à la section 6.2, ils conduisent à un polynôme du troisième degré, que nous avons pu résoudre en suivant [Nic93]. Nous avons trouvé la même asymptotique pour les racines de la fonction ζ, ainsi qu'une racine négative au point $-1/(\sigma-1)$. Ce point critique apparaît lorsque nous considérons la mesure stationnaire de la proportion de master sequences. Si nous supprimons toutes les mutations en fixant $q = 0$, deux états sont absorbants : un état sans aucune master sequence, et un autre avec uniquement des master sequences. Lorsque le paramètre de mutation augmente, la mesure invariante présente toujours un maximum autour de 0, due à la phase neutre. Nous pouvons représenter la mesure invariante ν comme ceci :



Si le paramètre q augmente encore, le maximum correspondant à la phase quasiespèce diminue, ce maximum est situé au point r_0. Entre ces deux maxima se trouve un minimum, situé au point $1/mr_0$, comme nous le représentons sur la figure suivante. Nous discutons de ce fait dans la section 4.3.



Il existe une valeur du paramètre q qui fait coalescer la maximum et le minimum.



Ce point q^* est tel que ${mr_0}^2$ tend vers 1. Puisque

$$mr_0^2 \sim \frac{m\big(\ln\sigma - \ell q\big)^2}{(\sigma-1)^2}\,,$$

nous en déduisons que ce paramètre admet le développement

$$q^* = \frac{\ln\sigma}{\ell} - \frac{\sigma-1}{\ell\sqrt{m}}\,. \tag{15}$$

4.4 Comparaison des trois critères

Les trois développements (13), (14), et (15) admettent la même puissance de m dans le terme qui suit $\ln\sigma/\ell$, plus précisément $1/\sqrt{m}$, cependant, la puissance du paramètre ℓ diffère. Deux des développements précédents ont la même asymptotique avec le terme $1/\ell\sqrt{m}$. Ces deux points critiques ont une chose en commun : ils sont définis uniquement à l'aide de la phase quasiespèce. Comme nous l'avons vu à la section 2.1, dans beaucoup d'applications, dès que les master sequences disparaissent, la population meurt parce que le temps pour les retrouver est très long. Par exemple, si une population de virus perd la master sequence, le système immunitaire détruira

le virus très vite. Pour les algorithmes génétiques, perdre la master sequence signifie plus de calculs pour la retrouver. Un développement en

$$q^* = \frac{\ln \sigma}{\ell} - \frac{C}{\ell\sqrt{m}}\,,$$

semble être pertinent pour ces applications en génétique ou en informatique.

La seconde asymptotique est inférieure, sa définition demande que les deux phases alternent beaucoup pour que la population soit proche de son équilibre. Ce paramètre critique semble pertinent si l'environnement ne présente aucune menace pour la population sans master sequence.

Première partie

Le modèle d'Eigen

Chapitre 1

Le modèle d'Eigen

Nous allons ici définir le modèle d'Eigen, et démontrer une équation vérifiée par la fitness moyenne de la population. Au passage, nous montrerons une formule exacte reliant la fitness moyenne, la classe de Hamming moyenne et le produit moyen de ces deux quantités valable pour tous les paysage de fitness. Nous déduirons également de l'équation caractérisant la fitness moyenne la suite du développement du seuil d'erreur d'Eigen en puissances de $1/\ell$:

$$q^* = \frac{\ln \sigma}{\ell} + \cdots ?$$

1 Le modèle

Comme nous l'avons vu dans l'introduction, les individus que nous allons suivre seront décrits par des chaînes finies de 0 et de 1. De telles chaînes seront souvent notées dans la suite par les lettres u, v. Ces chaînes possèdent ℓ digits notés $u(1), \cdots, u(\ell)$. Il est naturel de comparer deux chaînes u et v en calculant leur distance de Hamming :

$$d_H(u,v) = \mathrm{Card}\big\{k \in \{1 \cdots \ell\} \,:\, u(k) \neq v(k)\big\}\,.$$

La chaîne composée uniquement de 0 jouera un rôle particulier, et nous regrouperons souvent toutes les chaînes qui ont la même distance de Hamming à la master sequence en classe de Hamming. Nous noterons $C(u)$ l'indice de la classe de Hamming de la chaîne u, c'est-à-dire

$$C(u) = d_H(u, w^*)\,.$$

A chaque chaîne u est associée une fitness $f(u)$ supérieure à 1, qui dépendra du paysage de fitness que nous aurons choisi, comme nous l'avons vu dans l'introduction. Nous pouvons choisir un paysage quelconque mais nous n'associerons pas la fitness 1 à toutes les chaînes. Chaque digit évolue indépendamment des autre : la chaîne v mute vers la chaîne u, en lien avec la quantité M_{vu} qui ne dépend que de la classe de Hamming entre v et u,

$$M_{vu} = q^{d_H(v,u)}(1-q)^{\ell - d_H(v,u)}\,.$$

En fonction du modèle que nous étudierons, cette quantité pourra s'interpréter comme une vitesse de réaction, une probabilité, voire un taux de mutation. Lors d'un événement de mutation, chaque digit est changé avec la probabilité q indépendamment des autres, ainsi, nous pouvons interpréter M_{vu} comme la probabilité que la chaîne v se transforme en la chaîne u au cours d'un événement de mutation. Avec cette interprétation, nous déduisons que

$$\sum_u M_{vu} = 1\,. \tag{1.1}$$

Pour pouvoir écrire le noyau de mutation comme une matrice, nous devons choisir un ordre sur l'espace des chaînes. Nous allons ordonner les chaînes suivant l'ordre anti-lexicographique, comme si nous comptions en base 2 en lisant de droite à gauche :

$$\begin{array}{c} 000\cdots0 \\ 100\cdots0 \\ 010\cdots0 \\ \vdots \\ 111\cdots1\,. \end{array}$$

La chaîne $d_1 \cdots , d_\ell$ est donc la $(d_1 + 2d_2 + \cdots + 2^{\ell-1}d_\ell)$-ième chaîne. Nous écrivons le noyau de mutation M_{uv} comme une matrice de taille 2^ℓ, et nous la notons M :

$$M = \begin{pmatrix} (1-q)^\ell & q(1-q)^{\ell-1} & \cdots & q^\ell \\ q(1-q)^{\ell-1} & q^2(1-q)^{\ell-2} & \cdots & q^{\ell-1}(1-q) \\ \vdots & \vdots & \vdots & \vdots \\ q^\ell & (1-q)q^{\ell-1} & \cdots & (1-q)^\ell \end{pmatrix},$$

Dans toute la suite, nous identifierons une chaîne avec son numéro dans la liste ci-dessus : Lorsque nous parlerons de la u-ième colonne de la matrice M, que nous noterons M_u, nous réfèrerons à la colonne associée à la chaîne u. La matrice M étant symétrique, la transposée du vecteur colonne M_u est également la u-ième ligne de la matrice M. Remarquons de plus que la matrice M est stochastique et donc bistochastique.

Dans cette première partie, notre population est infinie et nous travaillons avec des proportions : nous noterons x_u la proportion d'individus possédant la chaîne u pour génome. Nous regroupons toutes ces proportions dans le vecteur x, le vecteur colonne dont les coordonnées sont les x_u :

$$x = \begin{pmatrix} \vdots \\ x_u \\ \vdots \end{pmatrix}_u .$$

2 Les équations d'Eigen

Manfred Eigen décrit un système d'équations de réactions chimiques, qui se produisent toutes en même temps à différentes vitesses. Dans la création de la chaîne u,

toutes les chaînes v entrent en jeu proportionnellement à leur concentration, à leur fitness et au taux de mutation M_{vu}. Un terme de destruction, noté $\lambda(t)$, est nécessaire pour s'assurer que la somme des concentrations reste égale à 1. L'équation d'évolution de la proportion d'individus ayant la chaîne u est

$$x'_u(t) = \sum_v f(v)x_v(t)M_{vu} - x_u(t)\lambda(t)\,. \tag{1.2}$$

Sommer ces équations sur toutes les chaînes conduit à

$$\Big(\sum_u x_u(t)\Big)' = \sum_v f(v)x_v(t)\Big(\sum_u M_{vu}\Big) - \Big(\sum_u x_u\Big)\lambda(t)\,.$$

Comme la somme des concentrations est constante égale à 1 et d'après (1.1), nous pouvons déterminer $\lambda(t)$: c'est la fitness moyenne de la population.

$$\lambda(t) = \sum_v f(v)x_v(t)\,. \tag{1.3}$$

Cette quantité dépend donc de toutes les concentrations $x_u(t)$. Le système (1.2) n'est pas linéaire, ce qui le rend difficile à étudier. Ce système a été beaucoup étudié, par exemple dans [BK83], [Jon77] ou encore [TM74]. En particulier, ce système admet une unique solution stationnaire, et pour toute condition initiale, la solution du système converge vers cette solution stationnaire. Cette solution est ce qu'Eigen a appelé la Quasiespèce.

A partir de maintenant et dans toute cette thèse, nous nous plaçons à l'équilibre. Nous noterons simplement x_u pour $x_u(t)$ et λ pour $\lambda(t)$. Remarquons que toutes les concentrations x_u sont strictement positives. Cela implique que le fitness moyenne est strictement comprise entre la fitness minimale et la fitness maximale, car nous avons supposé la fonction de fitness non constante. En particulier,

$$\lambda > 1\,. \tag{1.4}$$

Nous considèrerons que la fitness moyenne λ joue le rôle d'une nouvelle inconnue dans le système, ce chapitre a pour but de déterminer λ. Les manipulations suivantes vont nous permettre de réécrire le système différemment afin de pouvoir isoler une équation sur la fitness moyenne. Nous pouvons réécrire le système (1.2) comme

$$\sum_v f(v)x_v M_{vu} = \lambda x_u\,. \tag{1.5}$$

Nous allons pouvoir isoler la matrice M en retranchant des deux côtés de l'équation le produit Mx. Le système s'écrit à l'aide de la v-ième ligne M_v de la matrice M :

$$\lambda x - Mx = \sum_v \big(f(v) - 1\big)x_v\, M_v\,.$$

Après avoir divisé par λ, le système s'écrit

$$\Big(\mathrm{Id} - \frac{M}{\lambda}\Big)x = \sum_v \frac{f(v)-1}{\lambda}x_v M_v\,. \tag{1.6}$$

Comme la matrice M est stochastique et que λ est strictement supérieur à 1 d'après (1.4), la matrice M/λ possède un rayon spectral strictement inférieur à 1. Nous allons pouvoir inverser la matrice $I - M/\lambda$ grâce à la formule

$$\Big(\mathrm{Id} - \frac{M}{\lambda}\Big)^{-1} = \sum_{n\geq 0} \frac{M^n}{\lambda^n}\,.$$

En multipliant par cette matrice dans l'égalité (1.6), nous obtenons

$$x = \sum_v \frac{f(v)-1}{\lambda} x_v \sum_{n\geq 0} \frac{1}{\lambda^n} M^n\, M_v\,. \tag{1.7}$$

Il est possible d'interpréter cette équation de plusieurs manières. Tout d'abord, la matrice M étant stochastique, nous pouvons donc définir une chaîne de Markov dont M serait la matrice de transition et interpréter M_n comme le n-ième pas de la chaîne. Nous la traiterons différemment : la matrice M est symétrique réelle, nous allons donc pouvoir la diagonaliser et calculer les puissances M^n. Mais voyons d'abord comment cette équation se simplifie pour le paysage à un pic.

2.1 Le paysage à un pic

Seule la master sequence possède une fitness différente de 1 : $f(w^*) = \sigma > 1$ et à toutes les autres chaînes v est associée une fitness 1. La fitness moyenne est donc directement reliée à la proportion de master sequences :

$$\lambda = 1 + (\sigma - 1)x_0\,.$$

Nous déduisons de cette égalité que

$$\lambda < \sigma\,.$$

Nous pourrons ainsi remplacer x_0 dans ce qui suit par

$$x_0 = \frac{\lambda - 1}{\sigma - 1}\,.$$

En reprenant l'équation (1.6), le système d'Eigen s'écrit sous la forme

$$\big(\lambda \mathrm{Id} - M\big)x = (\sigma - 1)x_0 M_0 = (\lambda - 1)M_0\,,$$

ou encore, par un raisonnement analogue

$$x = \frac{\lambda - 1}{\lambda} \sum_{n\geq 0} \frac{1}{\lambda^n} M^n M_0\,.$$

Cette formule est assez simple, et son membre de droite ne dépend que de λ et ne dépend pas des concentrations des différentes chaînes. En considérant l'équation régissant la proportion de master sequences, nous allons obtenir une équation où la seule inconnue du système est la fitness moyenne λ. C'est grâce à cette équation que nous pourrons effectuer le développement de λ. Nous verrons aussi que cette formule contient déjà toutes les informations sur le seuil d'erreur.

3 L'indépendance des digits

La matrice M possède une certaine structure particulière qui découle directement du fait que les digits évoluent de façon indépendante. Commençons par décrire un produit sur les matrices qui est souvent appelé produit tensoriel ou produit de Kronecker. Si A est la matrice des (a_{ij}), et B une matrice sans aucunes contraintes sur la taille de ces matrices, nous notons

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \ddots & \ddots & \vdots \\ a_{n1}B & \cdots & \cdots & a_{nn}B \end{pmatrix}.$$

Nous allons en particulier pouvoir considérer les matrices $A^{\otimes 1} = A$ et ses itérées

$$A^{\otimes n+1} = A \otimes A^{\otimes n}.$$

Si la matrice A est carrée, le nombre de lignes de la matrice $A \otimes A$ est le carré du nombre de lignes de la matrice A, de même pour le nombre de colonnes. Ce produit n'est pas commutatif mais il est compatible avec le produit matriciel :

$$\big(AB\big)^{\otimes n} = A^{\otimes n} B^{\otimes n}.$$

Pour toutes matrices A, B, C, D avec les bonnes dimensions, nous avons également

$$\big(A \otimes B\big)\big(C \otimes D\big) = \big(AC\big) \otimes \big(BD\big).$$

Quel rapport avec la matrice M ? Nous avons ordonné les chaînes de façon à ce que les $2^{\ell-1}$ premières chaînes finissent par un 0 et les $2^{\ell-1}$ dernières finissent par un 1. L'évolution des digits étant indépendante, nous pouvons écrire la matrice de mutation à l'aide d'un produit tensoriel. Par exemple, dans le cas $\ell = 2$, la matrice de mutation s'écrit

$$M = \begin{pmatrix} (1-q)\begin{pmatrix} 1-q & q \\ q & 1-q \end{pmatrix} & q\begin{pmatrix} 1-q & q \\ q & 1-q \end{pmatrix} \\ q\begin{pmatrix} 1-q & q \\ q & 1-q \end{pmatrix} & (1-q)\begin{pmatrix} 1-q & q \\ q & 1-q \end{pmatrix} \end{pmatrix},$$

qui est exactement le produit

$$M = \begin{pmatrix} 1-q & q \\ q & 1-q \end{pmatrix}^{\otimes 2}.$$

Dans ce produit de deux matrices, la première correspond à l'évolution du dernier digit et la seconde correspond au premier digit. Dans le cas général, nous obtenons par récurrence

$$M = \begin{pmatrix} 1-q & q \\ q & 1-q \end{pmatrix}^{\otimes \ell}.$$

Dans cette écriture, le premier facteur correspond au dernier digit et le dernier facteur au premier digit. Le produit tensoriel et le produit matriciel étant compatibles, il suffit de diagonaliser la matrice 2×2 pour en déduire la diagonalisation de la matrice M. La matrice 2×2 admet comme valeurs propres 1 et $(1-2q)$ relativement aux vecteurs propres $(1,1)$ et $(1,-1)$. Ainsi, nous avons que

$$\begin{pmatrix} 1-q & q \\ q & 1-q \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1-2q \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} .$$

La seconde valeur propre est une quantité clé pour la suite : notons

$$\mu = 1 - 2q \,. \tag{1.8}$$

Nous en déduisons une expression des puissances de la matrice M,

$$M^n = \frac{1}{2^\ell} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}^{\otimes \ell} \begin{pmatrix} 1 & 0 \\ 0 & \mu^n \end{pmatrix}^{\otimes \ell} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}^{\otimes \ell} .$$

Cherchons maintenant à exprimer M_v, la v-ième colonne de M. Attribuons une notation aux deux colonnes de la matrice de mutation pour un digit :

$$\varepsilon_0 = \begin{pmatrix} 1-q \\ q \end{pmatrix} , \qquad \varepsilon_1 = \begin{pmatrix} q \\ 1-q \end{pmatrix} .$$

Si $v = d_1 \cdots d_\ell$, alors le vecteur colonne M_v s'écrit aussi avec un produit tensoriel :

$$M_v = \varepsilon_{d_\ell} \otimes \cdots \otimes \varepsilon_{d_1} .$$

L'ordre dans lequel sont pris ces produits tensoriels se comprend bien sur un exemple. Prenons la chaîne, $v_0 = 0 \cdots 01$ qui finit par un 1. Comme les $2^{\ell-1}$ premières chaînes finissent par un 0, il faut qu'un facteur q apparaisse sur les $2^{\ell-1}$ premières coordonnées du vecteur colonne M_v. C'est bien le cas dans le produit précédent. La quantité qui apparaît dans le système (1.7) est $M^n M_v$, avec nos notations, ce produit devient

$$M^n M_v = \frac{1}{2^\ell} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}^{\otimes \ell} \begin{pmatrix} 1 & 0 \\ 0 & \mu^n \end{pmatrix}^{\otimes \ell} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}^{\otimes \ell} \varepsilon_{d_\ell} \otimes \cdots \otimes \varepsilon_{d_1} .$$

Calculons tout d'abord le dernier produit :

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}^{\otimes \ell} \varepsilon_{d_\ell} \otimes \cdots \otimes \varepsilon_{d_1} = \left(\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \varepsilon_{d_\ell} \right) \otimes \cdots \otimes \left(\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \varepsilon_{d_1} \right) .$$

Chacun des facteurs ci-dessus est un des deux produits suivants :

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \varepsilon_0 = \begin{pmatrix} 1 \\ \mu \end{pmatrix} , \qquad \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \varepsilon_1 = \begin{pmatrix} 1 \\ -\mu \end{pmatrix} .$$

Nous pouvons résumer ces deux cas par

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \varepsilon_d = \begin{pmatrix} 1 \\ (-1)^d \mu \end{pmatrix} .$$

Ainsi,

$$M^n M_v = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}^{\otimes \ell} \begin{pmatrix} 1 \\ (-1)^{d_\ell}\mu^{n+1} \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} 1 \\ (-1)^{d_1}\mu^{n+1} \end{pmatrix}$$
$$= \frac{1}{2^\ell}\begin{pmatrix} 1+(-1)^{d_\ell}\mu^{n+1} \\ 1-(-1)^{d_\ell}\mu^{n+1} \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} 1+(-1)^{d_1}\mu^{n+1} \\ 1-(-1)^{d_1}\mu^{n+1} \end{pmatrix}.$$

Chaque coordonnée de ce vecteur colonne est donc un produit de ℓ facteurs qui sont soit $1+\mu^{n+1}$ soit $1-\mu^{n+1}$, il nous suffit de compter combien de fois chacun de ces termes apparaît à chaque ligne pour exprimer les coordonnées de ce vecteur. Considérons la coordonnée associée à la chaîne $u = b_1 \cdots b_\ell$. La valeur de chaque digit b_k va déterminer quel facteur apparaîtra dans le produit final :
• Si $b_k = 0$, le k-ième facteur sera $1+(-1)^{d_k}\mu^{n+1}$.
• Si $b_k = 1$, le k-ième facteur est $1-(-1)^{d_k}\mu^{n+1}$.
Dans les deux cas, si d_k et b_k sont différents, le facteur qui apparaît est $1+\mu^{n+1}$, et s'ils sont égaux le k-ième facteur est $1-\mu^{n+1}$. Il nous suffit alors de compter le nombre de différences entre les chaînes u et v pour exprimer ce vecteur colonne :

$$M^n\, M_v = \begin{pmatrix} \vdots \\ \left(\frac{1-\mu^{n+1}}{2}\right)^{d(u,v)} \left(\frac{1+\mu^{n+1}}{2}\right)^{\ell-d(u,v)} \\ \vdots \end{pmatrix}_u .$$

Les quantités ci-dessus apparaîtront souvent, notons

$$p_n = \frac{1-\mu^{n+1}}{2} \qquad 1-p_n = \frac{1+\mu^{n+1}}{2}. \tag{1.9}$$

Nous avons donc obtenu une expression exacte qui lie entre elles les proportions de chaque individu. Reprenons le système (1.7) :

$$x = \sum_v \frac{f(v)-1}{\lambda} x_v \sum_{n\geq 0} \frac{1}{\lambda^n} \begin{pmatrix} \vdots \\ p_n^{d(u,v)}(1-p_n)^{\ell-d(u,v)} \\ \vdots \end{pmatrix}_u . \tag{1.10}$$

Dans le cas du paysage à un pic, la formule devient

$$x = \frac{\lambda-1}{\lambda} \sum_{n\geq 0} \frac{1}{\lambda^n} \begin{pmatrix} \vdots \\ p_n^{C(u)}(1-p_n)^{\ell-C(u)} \\ \vdots \end{pmatrix}_u , \tag{1.11}$$

où $C(u)$ est la classe de Hamming de la chaîne u. Cette dernière formule exprime toutes les proportions en fonction de la fitness moyenne de la population λ. En considérant la première équation du système, nous pourrons déduire une équation vérifiée par λ, c'est ce que nous ferons à la section suivante. Nous déduirons ensuite une expression assez jolie de la distance moyenne de Hamming et enfin un développement asymptotique de la fitness moyenne λ.

4 La fitness moyenne pour le paysage à un pic

Plaçons-nous dans le cadre du paysage à un pic. En regardant la première ligne du système (1.11), celle qui régit le nombre de master sequences, nous obtenons

$$x_0 = \frac{\lambda - 1}{\lambda} \sum_{n \geq 0} \frac{1}{\lambda^n} \left(\frac{1 + \mu^{n+1}}{2} \right)^{\ell} .$$

En reprenant la définition de λ pour ce paysage de fitness, il vient

$$x_0 = \frac{(\sigma - 1)x_0}{\lambda} \sum_{n \geq 0} \frac{1}{\lambda^n} \left(\frac{1 + \mu^{n+1}}{2} \right)^{\ell} .$$

Après avoir simplifié les x_0 et appliqué la formule du binôme de Newton pour développer la puissance, nous obtenons

$$\frac{1}{\sigma - 1} = \sum_{n \geq 0} \frac{1}{\lambda^{n+1}} \frac{1}{2^{\ell}} \sum_{k=0}^{\ell} \binom{\ell}{k} (\mu^k)^{n+1} .$$

Nous intervertissons les deux sommes, nous calculons la série géométrique et nous obtenons une formule qui apparaît déjà dans les articles de Bratus, Novozhilov et Semenov [SBN14], bien qu'ils semblent en avoir sous estimé le potentiel. Leur preuve est sensiblement constituée des mêmes ingrédients que la notre bien qu'elle soit formulée de manière assez différente.

$$\frac{1}{\sigma - 1} = \sum_{k=0}^{\ell} \frac{1}{2^{\ell}} \binom{\ell}{k} \frac{1}{\frac{\lambda}{\mu^k} - 1} . \tag{1.12}$$

Nous pouvons également interpréter cette formule de façon probabiliste à travers l'espérance d'une certaine fonction d'une variable aléatoire binomiale comme nous l'énonçons dans le théorème suivant.

Théorème 1. *Si la variable aléatoire S_ℓ suit une loi binomiale de paramètres ℓ et $1/2$, la fitness moyenne λ est caractérisée par l'équation*

$$\frac{1}{\sigma - 1} = E\left(\frac{1}{\frac{\lambda}{\mu^{S_\ell}} - 1} \right) .$$

A partir de cette formule, nous pouvons démontrer le phénomène de transition de phase, caractéristique du modèle de la quasiespèce.

4.1 Une preuve de la transition de phase

Tout ce que nous avons fait jusque maintenant est valable pour toute valeur de ℓ et q. Pour cette section, nous nous plaçons dans le régime asymptotique défini par la convergence simultanée des paramètres

$$\ell \to \infty, \qquad q \to 0 ,$$

avec

$$\ell q \to a\,.$$

Pour prouver ce phénomène de transition de phase, nous aurons seulement besoin de l'inégalité de Chebyshev et de simples estimées uniformes. Reprenons un instant la formule sur λ :

$$\frac{1}{\sigma-1} = \sum_{k=0}^{\ell} \frac{1}{2^\ell}\binom{\ell}{k}\frac{1}{\frac{\lambda}{\mu^k}-1}\,.$$

Certains des termes vont dicter le comportement de la somme. Le terme pour $k=0$ ainsi que les termes correspondant aux indices proches de $\ell/2$ constitueront les termes principaux. Séparons donc cette somme en 2, introduisons un réel ε que nous choisirons plus tard et écrivons

$$\sum_{k=0}^{\ell} = \sum_{|k-\ell/2|>\varepsilon\ell} + \sum_{|k-\ell/2|\leq\varepsilon\ell}\,.$$

Comme μ est strictement inférieur à 1, la fonction suivante est décroissante :

$$x \to \frac{1}{\frac{\lambda}{\mu^x}-1}\,,$$

cela conduit à un encadrement uniforme de la première somme

$$0 \leq \sum_{|k-\frac{\ell}{2}|>\varepsilon} \frac{1}{2^\ell}\binom{\ell}{k}\frac{1}{\frac{\lambda}{\mu^k}-1} \leq \frac{P\Big(\big|S_\ell - \frac{\ell}{2}\big| > \varepsilon\ell\Big)}{\lambda-1}\,. \tag{1.13}$$

Pour la seconde somme, nous pouvons encadrer

$$\frac{P\Big(\big|S_\ell - \frac{\ell}{2}\big| \leq \varepsilon\ell\Big)}{\lambda\mu^{-\frac{\ell}{2}-\varepsilon\ell}-1} \leq \sum_{|k-\ell/2|<\varepsilon\ell} \frac{1}{2^\ell}\binom{\ell}{k}\frac{1}{\frac{\lambda}{\mu^k}-1} \leq \frac{P\Big(\big|S_\ell - \frac{\ell}{2}\big| \leq \varepsilon\ell\Big)}{\lambda\mu^{-\frac{\ell}{2}+\varepsilon\ell}-1}$$

La variable S_ℓ suit une loi binomiale, quand ℓ tend vers l'infini, cette variable devient très proche de son espérance. L'inégalité de Chebyshev donne

$$P\Big(\big|S_\ell - \frac{\ell}{2}\big| > \varepsilon\ell\Big) \leq \frac{\mathrm{Var}(S_\ell)}{\varepsilon^2\ell^2}\,.$$

En choisissant $\varepsilon = \sqrt{\frac{\ln\ell}{\ell}}$, nous obtenons

$$P\Big(\big|S_\ell - \frac{\ell}{2}\big| > \varepsilon\ell\Big) \leq \frac{1}{4\ln\ell}\,.$$

Cette probabilité tend donc vers 0 et la probabilité complémentaire

$$P\Big(\big|S_\ell - \frac{\ell}{2}\big| \leq \varepsilon\ell\Big) = 1 - P\Big(\big|S_\ell - \frac{\ell}{2}\big| > \varepsilon\ell\Big)$$

tend vers 1. Regardons maintenant les puissances de μ :

$$\mu^{\frac{\ell}{2}\pm\varepsilon\ell} = \exp\left(\frac{\ell}{2}\ln(1-2q) \pm \varepsilon\ell\ln(1-2q)\right).$$

Comme ε tend vers 0, les quantités $\mu^{\frac{\ell}{2}+\varepsilon\ell}$ et $\mu^{\frac{\ell}{2}-\varepsilon\ell}$ tendent vers e^{-a}. D'après (1.13), nous pouvons obtenir une première inégalité sur la fitness moyenne :

$$\frac{P\Big(\big|S_\ell - \frac{\ell}{2}\big| \leq \varepsilon\ell\Big)}{\lambda\mu^{-\frac{\ell}{2}-\varepsilon\ell}-1} \leq \frac{1}{\sigma-1}\,.$$

Ainsi, nous en déduisons

$$\lambda \geq \frac{(\sigma-1)P\Big(\big|S_\ell - \frac{\ell}{2}\big| \leq \varepsilon\ell\Big)+1}{\mu^{-\frac{\ell}{2}-\varepsilon\ell}}\,,$$

comme ce minorant tend vers σe^{-a}, nous en déduisons que

$$\liminf \lambda \geq \sigma e^{-a}\,.$$

Nous pouvons maintenant distinguer les deux régimes.
• Si $a < \ln\sigma$, alors $\sigma e^{-a} > 1$ et donc $\lambda > 1$ d'après ce qui précède, la somme des termes qui sont loin de $\ell/2$ tend donc vers 0. Pour que la somme complète tende vers $1/(\sigma-1)$, il faut donc que λ tende vers σe^{-a}. C'est le régime quasiespèce, dans ce cas, la proportion de master sequences vérifie

$$x_0 \to \frac{\sigma e^{-a}-1}{\sigma-1} > 0\,.$$

• Si $a > \ln\sigma$, la somme centrale est plus petite que $1/(e^a-1)$ qui est inférieur à $1/(\sigma-1)$, il faut donc que la somme extérieure ne tende pas vers 0, c'est-à-dire que λ tende vers 1 d'après (1.13). C'est le régime neutre, la proportion de master sequences tend vers 0.

Nous avons finalement démontré le théorème suivant, énonçant la transition de phase.

Théorème 2. *La fitness moyenne λ de la population converge dans le régime asymptotique*

$$\lambda \to \max(1,\, \sigma e^{-a})\,.$$

La transition de phase correspond au moment où $1 = \sigma e^{-a}$, c'est-à-dire lorsque

$$a = \ln\sigma\,.$$

5 Un joli couplage

Cette section est purement esthétique, elle vise simplement à éclairer différemment le théorème 1. Plaçons-nous dans le cadre du paysage à un pic, la proportion d'individus associés à la chaîne u vérifie

$$x_u = \frac{\lambda-1}{\lambda}\sum_{n\geq 0}\frac{1}{\lambda^n}\,p_n^{C(u)}(1-p_n)^{\ell-C(u)}\,,$$

où $C(u)$ est l'indice de la classe de Hamming de u, et

$$p_n = \frac{1-\mu^{n+1}}{2}\,.$$

Dans la section précédente, il a été fructueux de reconnaître dans ces formules une loi Binomiale de paramètres ℓ et p_n. Nous allons ici pousser l'analogie.

Définissons $\varepsilon_1^{(n)}, \cdots, \varepsilon_\ell^{(n)}$ des variables indépendantes et de même loi de Bernoulli de paramètre p_n. Construisons maintenant la variable aléatoire X_n, qui consiste en la concaténation de $\varepsilon_1^{(n)}, \cdots, \varepsilon_\ell^{(n)}$. La variable aléatoire X_n est donc une chaîne aléatoire de ℓ digits qui sont tous des zéros ou des uns. Avec ces notations, nous pouvons calculer la probabilité que la chaîne aléatoire X_n soit égale à une chaîne fixée u :

$$P\big(\,X_n = u\,\big) = p_n^{C(u)}(1-p_n)^{\ell-C(u)}\,.$$

Remarquons que la suite des probabilités p_n croît avec n, puisque $\mu = 1-2q$ est strictement inférieure à 1.

5.1 Le couplage

Nous allons maintenant construire toutes ces variables sur le même espace de probabilité. Commençons par définir ℓ variables aléatoires $U_1, \cdots, U_\ell$ indépendantes uniformément distribuées sur l'intervalle $[0,1]$. Définissons maintenant pour tout entier n, et pour tout entier k entre 1 et ℓ la variable aléatoire de Bernoulli $\varepsilon_k^{(n)}$ en fonction de U_k comme suit :

- Si $U_k \leq p_n$, alors $\varepsilon_k^{(n)} = 1$.
- Si $U_k > p_n$, alors $\varepsilon_k^{(n)} = 0$.

Nous définissons ainsi ℓ variables aléatoires de Bernoulli $\varepsilon_1^{(n)}, \cdots, \varepsilon_\ell^{(n)}$ indépendantes et de même loi de Bernoulli de paramètre p_n. Avec cette construction, la suite $\big(\varepsilon_k^{(n)}\big)_n$ est soumise à une contrainte forte : comme la suite p_n est croissante, si pour un certain n_0, $\varepsilon_k^{(n_0)}$ vaut 1, alors tous les termes suivants $\varepsilon_k^{(n)}$, pour $n > n_0$ vaudront aussi 1. Ainsi, pour une réalisation des variables aléatoires $U_1, \cdots, U_\ell$, la suite $(\varepsilon_k^{(n)})_n$ commence à 0 jusqu'à un certain entier où elle devient égale à 1. Nous définissons finalement X_n comme la concaténation $\varepsilon_1^{(n)} \cdots \varepsilon_\ell^{(n)}$. Avec cette définition, nous avons bien que

$$P\big(X_n = u\big) = p_n^{C(u)}(1-p_n)^{\ell-C(u)}\,.$$

Décrivons une réalisation typique de la suite (X_n). Voyons que $p_0 = q$ est très petit, il est donc possible que $X_0 = 0\cdots 0$, de même pour X_1, X_2, jusqu'à ce qu'un 1 apparaisse dans la suite. Par ce que nous avons dit précédemment, si X_n contient un 1 à la k-ième place, toutes les chaînes suivantes $X_{n+1}, \ldots$ contiendront aussi un 1 à la k-ième place. Autrement dit, la chaîne $0\cdots 0$ n'apparaît que dans les premiers termes de cette suite.

5.2 De nouvelles formules

Nous pouvons nous demander ce que deviennent les formules issues du système d'Eigen. La formule pour la chaîne u donne d'après (1.11)

$$x_u = \frac{\lambda - 1}{\lambda} \sum_n E\left(\frac{1_{\{X_n = u\}}}{\lambda^n} \right). \tag{1.14}$$

Avec le couplage, l'ensemble des entiers n tels que $X_n = u$ est un intervalle de $\mathbb{N}$. Nous pouvons donc définir les variables aléatoires N_1 et N_2 comme le premier et dernier instant où la chaîne X_n est égale à u

$$N_1 = \inf \Big\{ n \geq 0 \,|\, X_n = u \Big\},$$

et

$$N_2 = \sup \Big\{ n \geq 0 \,|\, X_n = u \Big\}.$$

Ainsi, la formule (1.14) peut s'écrire

$$x_u = \frac{\lambda - 1}{\lambda} E\left(\sum_{n=N_1}^{N_2} \frac{1}{\lambda^n} \right).$$

Pour chaque réalisation des variables aléatoires $U_1, \cdots, U_\ell$, nous avons partitionné $\mathbb{N}$ en intervalles sur lesquels la suite X_n est constante. Nous pouvons aussi calculer la somme géométrique et nous obtenons

$$x_u = E\left(\frac{1}{\lambda^{N_1}} - \frac{1}{\lambda^{N_2+1}} \right). \tag{1.15}$$

5.3 Une autre équation pour la fitness moyenne ?

Pour l'équation sur la chaîne $u = 0 \cdots 0$, la variable N_1 vaut 0 ou $+\infty$, de plus les lois des variables N_1 et N_2 ne dépendent que de la variable $\min\big(U_1, \cdots, U_\ell\big)$. La fonction de répartition F de cette variable aléatoire est facile à calculer :

$$\forall x \in [0,1], \quad F(x) = 1 - (1-x)^\ell.$$

La variable N_1 est égale à 0 si et seulement si toutes les variables U_k sont supérieures à p_0, c'est-à-dire

$$P\big(N_1 = 0\big) = 1 - F(p_0) = (1-p_0)^\ell.$$

Quant à la seconde variable, N_2 est égale à j si toutes les variables U_k sont supérieures à p_{j-1} sauf une, qui doit être inférieure à p_j, autrement dit

$$P\big(N_2 = j\big) = F(p_j) - F(p_{j-1}) = -(1-p_j)^\ell + (1-p_{j-1})^\ell.$$

En calculant l'espérance de la formule (1.15) pour $u = 0 \cdots 0$, nous retrouvons bien la formule

$$x_0 = \frac{\lambda - 1}{\lambda} \sum_{n \geq 0} \frac{(1-p_n)^\ell}{\lambda^n}.$$

6 La distance de Hamming moyenne

Nous allons continuer à exploiter le système (1.10). Nous en tirons une expression de chaque concentration x_u, nous allons calculer la distance de Hamming moyenne que nous noterons $\overline{H}$ dans la suite

$$\overline{H} = \sum_u C(u) x_u \,.$$

Cette quantité nous permettra de mieux comprendre la répartition du nuage de mutants. Commençons par le cas plus simple du paysage à un pic.

6.1 Pour le paysage à un pic

En sommant toutes les lignes de (1.11) pondérées par la classe de Hamming associée, nous obtenons

$$\overline{H} = \frac{\lambda - 1}{\lambda} \sum_{n\geq 0} \frac{1}{\lambda^n} \sum_u C(u) p_n^{C(u)} (1 - p_n)^{\ell - C(u)} \,.$$

Certaines chaînes font partie de la même classe de Hamming k, il y en a exactement $\binom{\ell}{k}$: il suffit de choisir l'emplacement des uns dans la chaîne nous nous ramenons donc à

$$\overline{H} = \frac{\lambda - 1}{\lambda} \sum_{n\geq 0} \frac{1}{\lambda^n} \sum_{k=0}^{\ell} \binom{\ell}{k} k p_n^k (1 - p_n)^{\ell - k} \,.$$

Nous reconnaissons dans cette formule l'espérance d'une variable binomiale de paramètres ℓ et p_n :

$$\sum_{k=0}^{\ell} k \binom{\ell}{k} p_n^k (1 - p_n)^{\ell - k} = \ell p_n \,.$$

Ainsi, en remplaçant p_n par son expression (1.9), nous obtenons

$$\overline{H} = \frac{\lambda - 1}{\lambda} \sum_{n\geq 0} \frac{1}{\lambda^n} \frac{\ell}{2} \big(1 - \mu^{n+1}\big) \,.$$

En développant, nous en déduisons

$$\overline{H} = \frac{\ell}{2} \frac{\lambda - 1}{\lambda} \left(\sum_{n\geq 0} \frac{1}{\lambda^n} - \mu \sum_{n\geq 0} \frac{\mu^n}{\lambda^n} \right) .$$

En calculant les séries géométriques, nous obtenons

$$\overline{H} = \frac{\ell}{2} \Big(1 - \frac{\mu(\lambda - 1)}{\lambda - \mu}\Big) \,.$$

Nous avons finalement

$$\overline{H} = \ell q \frac{\lambda}{\lambda - \mu} \,. \tag{1.16}$$

Cette formule est remarquable de simplicité : nous pouvons en déduire la moyenne du nuage de mutants dans chaque régime.
• Dans le régime neutre, λ tend vers 1 et la distance moyenne est $\overline{H} = \ell/2$.
• Dans le régime quasiespèce, λ tend vers σe^{-a} et la distance moyenne est donc bornée.

$$\overline{H} = \frac{\sigma a e^{-a}}{\sigma e^{-a} - 1}\,.$$

Nous allons également pouvoir déduire des résultats dans le cas critique, que nous développerons à la fin de ce chapitre.

6.2 Pour un paysage de fitness général

Revenons à un paysage de fitness général et essayons d'utiliser le même raisonnement. Cette fois, nous ne pouvons plus regrouper toutes les chaînes qui ont la même classe car elles ont potentiellement des fitness différentes. Nous avons

$$\overline{H} = \sum_v \frac{f(v)-1}{\lambda} x_v \sum_{n\geq 0} \frac{1}{\lambda^n} \sum_u C(u) p_n^{d(u,v)} (1-p_n)^{\ell - d(u,v)}\,. \tag{1.17}$$

Intéressons nous à une quantité intermédiaire, notons $S_n(v)$ la somme

$$S_n(v) = \sum_u C(u) p_n^{d(u,v)} (1-p_n)^{\ell - d(u,v)}\,. \tag{1.18}$$

Nous allons conditionner cette quantité par rapport à la distance $d(u,v)$, nous obtenons

$$S_n(v) = \sum_{k=0}^{\ell} \left(\sum_{u\,:\,d(u,v)=k} C(u) \right) p_n^k (1-p_n)^{\ell-k}\,.$$

La somme qui apparaît et qu'il nous faut calculer est

$$\sum_{u\,:\,d(u,v)=k} C(u)\,.$$

Cette quantité est la somme des indices des classes des chaînes qui peuvent être atteintes à partir de la chaîne v avec exactement k erreurs. Il nous faut donc compter le nombre de chaînes u dans chaque classe qui diffèrent de k digits de la chaîne v. Notons r le nombre d'erreurs parmi les k qui arrivent sur des digits 1 de la chaîne v. Dans ce cas, $k-r$ zéros sont changés en 1, dans la nouvelle chaîne, il y a donc exactement $C(v) + (k-r) - r$ uns et la somme devient

$$\sum_{u\,:\,d(u,v)=k} C(u) = \sum_{r=0}^{k} \Big(C(v) + k - 2r \Big) \binom{C(v)}{r} \binom{\ell - C(v)}{k-r}\,.$$

Nous pouvons simplifier un peu cette somme, en utilisant le fait que

$$\sum_{r=0}^{k} \binom{C(v)}{r} \binom{\ell - C(v)}{k-r} = \binom{\ell}{k}\,,$$

nous obtenons

$$\sum_{u\,:\,d(u,v)=k} C(u) = \Big(C(v)+k\Big)\binom{\ell}{k} - 2\sum_{r=0}^{k} r\binom{C(v)}{r}\binom{\ell - C(v)}{k-r}\,.$$

Reprenons maintenant l'expression (1.18) de la somme $S_n(v)$. Cette somme se décompose en deux sommes que nous allons calculer séparément. Pour la première, nous reconnaissons la loi d'une variable binomiale de paramètres ℓ, p_n ce qui nous permet de calculer :

$$\sum_{k=0}^{\ell} \Big(C(v)+k\Big)\binom{\ell}{k} p_n^k (1-p_n)^{\ell-k} = C(v) + \ell p_n\,.$$

Pour la seconde il nous faut calculer la somme

$$\sum_{k=0}^{\ell}\sum_{r=0}^{k} r\binom{C(v)}{r}\binom{\ell - C(v)}{k-r} p_n^k (1-p_n)^{\ell-k}$$

Après avoir interverti les deux sommes, nous allons changer d'indice dans la seconde somme et poser $k' = k - r$, cela nous permet d'écrire

$$\begin{aligned}\sum_{r=0}^{C(v)}\sum_{k=r}^{\ell} & r\binom{C(v)}{r}\binom{\ell - C(v)}{k-r} p_n^k (1-p_n)^{\ell-k} \\ &= \sum_{r=0}^{C(v)} r\binom{C(v)}{r} p_n^r (1-p_n)^{C(v)-r} \sum_{k=0}^{\ell-r}\binom{\ell - C(v)}{k} p_n^k (1-p_n)^{\ell-C(v)-k}\,.\end{aligned}$$

Comme r est inférieur à $C(v)$, $\ell - r$ est supérieur à $\ell - C(v)$, la dernière somme est donc égale à 1. Il reste alors l'espérance d'une binomiale de paramètres $C(v)$, p_n. Finalement,

$$S_n(v) = C(v) + \ell p_n - 2C(v)p_n\,.$$

Reprenons l'expression (1.17) de $\overline{H}$ et écrivons

$$\overline{H} = \sum_v \frac{f(v)-1}{\lambda} x_v \sum_{n\geq 0} \frac{1}{\lambda^n}\Big(C(v) + \ell p_n - 2C(v)p_n\Big)\,.$$

En remplaçant p_n par son expression (1.9) et en calculant les séries géométriques, nous obtenons

$$\overline{H} = \ell q \frac{\lambda}{\lambda - \mu} + \frac{\mu}{\lambda - \mu}\sum_v \big(f(v)-1\big) x_v C(v)\,.$$

Si nous notons $\overline{FH}$ la moyenne du produit de la fitness par l'indice de la classe de Hamming

$$\overline{FH} \;= \sum_u f(u) C(u) x_u\,,$$

alors nous obtenons le théorème suivant.

Théorème 3. *Pour tout paysage de fitness, la fitness moyenne de la population, la classe de Hamming moyenne et la moyenne du produit de ces deux quantités sont liées par la relation*

$$\lambda\,\overline{H} = \ell q\lambda + \mu\,\overline{FH}\,.$$

Cette formule permet de retrouver le cas particulier du paysage à un pic et la formule (1.16), dans ce cas, comme la seule valeur de fitness différente de 1 est associée à des individus dont la classe est nulle,

$$\overline{FH} = \overline{H}\,.$$

Le théorème 3 est vrai pour tout paysage de fitness, par exemple, si nous fixons la fitness de tous les individus à une certaine valeur f_0, alors nous obtenons $\overline{H} = \ell/2$. Cela est normal : dans le régime neutre, les individus sont uniformément répartis sur toutes les chaînes.

7 Le développement de la fitness moyenne au point critique

Dans cette partie, nous nous intéressons au cas critique $a = \ln\sigma$. De la même façon que pour prouver l'existence d'une transition de phase nous allons reprendre le théorème 1 et séparer la somme en plusieurs parties. Nous allons isoler le terme pour $k = 0$, qui apportera sa contribution dans la somme globale. Nous écrivons donc, pour un certain ε,

$$\frac{1}{\sigma-1} = \frac{1}{2^\ell(\lambda-1)} + E\Big(\frac{1}{\frac{\lambda}{\mu^{S_\ell}}-1}1_{|S_\ell-\frac{\ell}{2}|\leq\varepsilon\ell}\Big) + E\Big(\frac{1}{\frac{\lambda}{\mu^{S_\ell}}-1}1_{|S_\ell-\frac{\ell}{2}|>\varepsilon\ell,\,S_\ell\neq 0}\Big)\,. \quad (1.19)$$

Comme précédemment, le dernier terme de cette expression pourra être majoré uniformément. Nous avons

$$E\Big(\frac{1}{\frac{\lambda}{\mu^k}-1}1_{|S_\ell-\frac{\ell}{2}|>\varepsilon\ell,\,S_\ell\neq 0}\Big) \leq \frac{1}{\frac{\lambda}{\mu}-1}P\Big(\big|S_\ell-\frac{\ell}{2}\big|>\varepsilon\ell\Big)\,.$$

Comme λ est supérieur à 1,

$$\frac{1}{\frac{\lambda}{\mu}-1} = \frac{\mu}{\lambda-\mu} \leq \frac{1}{2q}\,.$$

L'inégalité de Chebyshev ne suffira pas cette fois, l'inégalité de Chernoff pour des variables binomiales donne

$$P\Big(\big|S_\ell-\frac{\ell}{2}\big|>\varepsilon\ell\Big) \leq e^{-2\varepsilon^2\ell}\,.$$

Ainsi, en choisissant

$$\varepsilon = \sqrt{\frac{\kappa\ln\ell}{\ell}}\,, \quad (1.20)$$

avec κ assez grand, nous pouvons rendre le dernier terme de (1.19) comme une puissance de $1/\ell$ aussi petite que l'on veut. Nous nous ramenons donc à l'équation

$$\frac{1}{\sigma-1}=\frac{1}{2^\ell(\lambda-1)}+E\Big(\frac{1}{\frac{\lambda}{\mu^{S_\ell}}-1}1_{|S_\ell-\frac{\ell}{2}|\leq\varepsilon\ell}\Big)+O\Big(\frac{1}{\ell^\kappa}\Big)\,. \tag{1.21}$$

Intéressons-nous maintenant au second terme de cette équation. Pour la fin de ce chapitre, nous nous contenterons de déterminer un équivalent de $\lambda-1$. Commençons par travailler sur les termes à l'intérieur de l'espérance :

$$\frac{\lambda}{\mu^{S_\ell}}=\lambda\mu^{-\ell/2}\bigg(1+\frac{1}{\mu^{S_\ell-\ell/2}}-1\bigg)\,.$$

En retranchant 1 et en prenant l'inverse, nous obtenons

$$\frac{1}{\frac{\lambda}{\mu^{S_\ell}}-1}=\frac{1}{\lambda\mu^{-\ell/2}-1}\bigg(1+\frac{\lambda\mu^{-\ell/2}}{\lambda\mu^{-\ell/2}-1}\Big(\frac{1}{\mu^{S_\ell-\ell/2}}-1\Big)\bigg)^{-1}\,. \tag{1.22}$$

Remarquons que la fonction indicatrice dans l'espérance va faire que la quantité $S_\ell-\ell/2$ est d'ordre au plus $\varepsilon\ell$, de plus comme $\ln\mu$ est d'ordre q, cela implique que le produit $(S_\ell-\frac{\ell}{2})\ln\mu$ est d'ordre au plus ε. Ainsi,

$$\frac{1}{\mu^{S_\ell-\ell/2}}1_{|S_\ell-\frac{\ell}{2}|\leq\varepsilon\ell}=\exp\Big(-(S_\ell-\ell/2)\ln\mu\Big)1_{|S_\ell-\frac{\ell}{2}|\leq\varepsilon\ell}=\Big(1+O(\varepsilon)\Big)1_{|S_\ell-\frac{\ell}{2}|\leq\varepsilon\ell}\,.$$

Comme le produit $\lambda\mu^{-\ell/2}$ tend vers σ, nous en déduisons que le terme

$$\frac{1}{\mu^{S_\ell-\ell/2}}-1$$

apparraissant dans la formule (1.22) tend vers 0. Nous allons effectuer un développement de la quantité

$$\bigg(1+\frac{\lambda\mu^{-\ell/2}}{\lambda\mu^{-\ell/2}-1}\Big(\frac{1}{\mu^{S_\ell-\ell/2}}-1\Big)\bigg)^{-1}\,.$$

En reprenant l'espérance et l'indicatrice de la formule (1.21), nous obtenons après avoir inversé la somme et l'espérance,

$$E\Big(\frac{1}{\frac{\lambda}{\mu^{S_\ell}}-1}1_{|S_\ell-\frac{\ell}{2}|\leq\varepsilon\ell}\Big)=\frac{1}{\lambda\mu^{-\ell/2}-1}\sum_{p\geq0}\Big(\frac{-\lambda\mu^{-\ell/2}}{\lambda\mu^{-\ell/2}-1}\Big)^pE\bigg(\Big(\frac{1}{\mu^{S_\ell-\ell/2}}-1\Big)^p1_{|S_\ell-\frac{\ell}{2}|\leq\varepsilon\ell}\bigg)\,. \tag{1.23}$$

Remarquons que le terme général dans la somme est d'ordre au plus ε^p, pour la suite, nous nous limiterons aux deux premiers termes de cette équation. Occupons-nous maintenant du terme à l'intérieur des espérances,

$$\begin{aligned}&\frac{1}{\mu^{S_\ell-\ell/2}}1_{|S_\ell-\frac{\ell}{2}|\leq\varepsilon\ell}=\\&\bigg(1-\Big(S_\ell-\frac{\ell}{2}\Big)\ln\mu+\frac{1}{2}\Big(S_\ell-\frac{\ell}{2}\Big)^2\ln^2\mu-\frac{1}{3!}\Big(S_\ell-\frac{\ell}{2}\Big)^3\ln^3\mu+o(\varepsilon^3)\bigg)1_{|S_\ell-\frac{\ell}{2}|\leq\varepsilon\ell}\,.\end{aligned} \tag{1.24}$$

Prenons maintenant l'espérance, comme $S_\ell - \ell/2$ est une variable aléatoire centrée, ses moments d'ordre impair sont nuls

$$E\Big(S_\ell - \frac{\ell}{2}\Big) = 0\,,$$

$$E\Big(\big(S_\ell - \frac{\ell}{2}\big)^2 1_{|S_\ell - \frac{\ell}{2}|\le\varepsilon\ell}\Big) = E\Big(\big(S_\ell - \frac{\ell}{2}\big)^2\Big) + O(\varepsilon\ell) = \mathrm{Var}(S_\ell) + O(\varepsilon\ell) = \frac{\ell}{4} + O(\varepsilon\ell)\,,$$

$$E\Big(\big(S_\ell - \frac{\ell}{2}\big)^3\Big) = 0\,.$$

Comme de plus $\ln\mu$ est d'ordre $-2q$ et ℓq converge vers $\ln\sigma$, nous obtenons

$$E\Big(\big(S_\ell - \frac{\ell}{2}\big)^2\Big)\ln\mu^2 = q\ln\sigma + o(q)\,.$$

Pour le premier terme de la série, nous avons, comme ε^3 est négligeable devant q,

$$E\Big((\frac{1}{\mu^{S_\ell - \ell/2}} - 1)1_{|S_\ell - \frac{\ell}{2}|\le\varepsilon\ell}\Big) = \frac{q\ln\sigma}{2}\,1_{|S_\ell - \frac{\ell}{2}|\le\varepsilon\ell} + o(q)\,.$$

Aussi, en nous ne conservons que le premier terme du développement (1.24), nous écrivons

$$E\Big((\frac{1}{\mu^{S_\ell - \ell/2}} - 1)^2 1_{|S_\ell - \frac{\ell}{2}|\le\varepsilon\ell}\Big) = q\ln\sigma\,1_{|S_\ell - \frac{\ell}{2}|\le\varepsilon\ell} + o(q)\,.$$

en reprenant l'expression (1.23), nous écrivons

$$E\Big(\frac{1}{\frac{\lambda}{\mu^{S_\ell}} - 1}1_{|S_\ell - \frac{\ell}{2}|\le\varepsilon\ell}\Big) = \frac{1_{|S_\ell - \frac{\ell}{2}|\le\varepsilon\ell}}{\lambda\mu^{-\ell/2} - 1}\Big(1 - \frac{\lambda\mu^{-\ell/2}}{\lambda\mu^{-\ell/2} - 1}\frac{q\ln\sigma}{2} + \Big(\frac{\lambda\mu^{-\ell/2}}{\lambda\mu^{-\ell/2} - 1}\Big)^2 q\ln\sigma\Big) + o(q)\,. \tag{1.25}$$

Comme λ tend vers 1 et $\mu^{-\ell/2}$ tend vers σ, nous avons

$$E\Big(\frac{1}{\frac{\lambda}{\mu^{S_\ell}} - 1}1_{|S_\ell - \frac{\ell}{2}|\le\varepsilon\ell}\Big) = \frac{1}{\lambda\mu^{-\ell/2} - 1} + \Big(-\frac{\sigma}{2(\sigma-1)^2} + \frac{\sigma^2}{(\sigma-1)^3}\Big)q\ln\sigma + o(q)\,.$$

En rassemblant les dénominateurs, nous obtenons

$$E\Big(\frac{1}{\frac{\lambda}{\mu^{S_\ell}} - 1}1_{|S_\ell - \frac{\ell}{2}|\le\varepsilon\ell}\Big) = \frac{1}{\lambda\mu^{-\ell/2} - 1} + \frac{\sigma(\sigma+1)}{2(\sigma-1)^3}q\ln\sigma + o(q)\,. \tag{1.26}$$

Cherchons maintenant un développement du premier terme de cette équation

$$\mu^{-\ell/2} = \exp\Big(-\frac{\ell}{2}\ln(1-2q)\Big) = \exp\Big(\ell q + \ell q^2 + o(q)\Big) = \sigma\exp\Big((\ell q - \ln\sigma) + q\ln\sigma + o(q)\Big)\,.$$

Ainsi, comme λ tend vers 1,

$$\lambda\mu^{-\ell/2} = \sigma\lambda + \sigma(\ell q - \ln\sigma) + \sigma q\ln\sigma + o\Big(q + (\ell q - \ln\sigma)\Big)\,.$$

En retranchant 1, nous obtenons

$$\lambda\mu^{-\ell/2} - 1 = \big(\sigma\lambda - 1\big)\left(1 + \frac{\sigma}{\sigma-1}(\ell q - \ln\sigma) + \frac{\sigma}{\sigma-1} q \ln\sigma + o\Big(q + (\ell q - \ln\sigma)\Big)\right).$$

En développant, nous obtenons

$$\frac{1}{\lambda\mu^{-\ell/2} - 1} = \frac{1}{\sigma\lambda - 1} - \frac{\sigma}{(\sigma-1)^2}(\ell q - \ln\sigma) - \frac{\sigma}{(\sigma-1)^2} q \ln\sigma + o\Big(q + (\ell q - \ln\sigma)\Big). \tag{1.27}$$

Le premier terme du membre de droite tend vers $1/(\sigma-1)$, nous allons le regrouper avec le membre de gauche de l'égalité (1.21), nous obtenons

$$\frac{1}{\sigma\lambda - 1} - \frac{1}{\sigma - 1} = -\frac{\sigma}{(\sigma-1)^2}(\lambda - 1) + o(\lambda - 1),$$

En reprenant l'expression (1.21) et en remplaçant ses termes à l'aide de la formule précédente et des équations (1.26), (1.27), nous obtenons finalement

$$\begin{aligned}&\frac{1}{2^\ell(\lambda-1)} - \frac{\sigma}{\big(\sigma-1\big)^2}\,(\lambda - 1)\\ &\quad= \frac{\sigma}{(\sigma-1)^2}\big(\ell q - \ln\sigma\big) + \frac{\sigma(\sigma-3)}{2(\sigma-1)^3} q \ln\sigma + o\Big(q + (\ell q - \ln\sigma) + (\lambda - 1)\Big).\end{aligned} \tag{1.28}$$

7.1 Le développement du paramètre de mutation

Si nous choisissons maintenant un développement du paramètre q en fonction de ℓ,

$$q = \frac{\ln\sigma}{\ell} + \frac{c}{\ell^\alpha},$$

pour des réels c et $\alpha > 1$. Alors, nous savons caractériser si nous nous trouvons dans le régime quasiespèce ou dans le régime neutre. La distinction entre les deux régimes n'est pas rigoureusement définie, nous dirons que nous nous trouvons dans le régime neutre si l'ordre de grandeur de la proportion de master sequences est proche de $1/2^\ell$, la concentration uniforme typique du régime neutre. En revanche si la concentration de master sequence est plutôt d'ordre une puissance de $1/\ell$, nous dirons que nous nous trouvons dans le régime quasiespèce. Comme $\lambda > 1$, il nous suffit de connaître le signe du terme de droite dans l'égalité (1.28) pour déduire lequel des deux termes de gauche lui est équivalent.

Le cas où $\alpha < 2$

Dans ce cas, la différence $\ell q - \ln\sigma$ domine le paramètre q, l'égalité (1.28) devient

$$\frac{(\sigma-1)^2}{2^\ell\sigma(\lambda-1)} - (\lambda - 1) = \frac{c}{\ell^{\alpha-1}} + o\Big((\ell q - \ln\sigma) + (\lambda - 1)\Big).$$

Dans ce cas, le régime dépend du signe de c. Si c est positif, alors $\lambda - 1$ est équivalent à

$$\lambda - 1 \sim \frac{(\sigma-1)^2}{\sigma c}\,\frac{\ell^{\alpha-1}}{2^\ell},$$

le régime est alors neutre. Si c est négatif, alors le régime est celui de la quasiespèce et

$$\lambda - 1 \sim -\frac{c}{\ell^{\alpha-1}}\,.$$

Le cas où $\alpha = 2$

Cette fois, la différence $\ell q - \ln\sigma$ est du même ordre que q, l'égalité (1.28) devient

$$\frac{(\sigma-1)^2}{2^\ell\sigma(\lambda-1)} - (\lambda-1) = \Big(\frac{c}{\ln\sigma} + \frac{\sigma-3}{2(\sigma-1)}\ln\sigma\Big)q + o\Big(q + (\lambda-1)\Big)\,.$$

Le régime va donc dépendre du signe de $c + \frac{\sigma-3}{2(\sigma-1)}\ln^2\sigma$. Dans le cas où

$$c < -\frac{\sigma-3}{2(\sigma-1)}\ln^2\sigma\,,$$

alors, le régime est celui de la quasiespèce et

$$\lambda - 1 \sim -\Big(\frac{c}{\ln\sigma} + \frac{\sigma-3}{2(\sigma-1)}\ln\sigma\Big)q\,.$$

Dans le cas où

$$c > -\frac{\sigma-3}{2(\sigma-1)}\ln^2\sigma\,,$$

le régime est neutre, la fitness moyenne est d'ordre

$$\lambda - 1 \sim \frac{(\sigma-1)^2}{\sigma\big(\frac{c}{\ln\sigma} + \frac{\sigma-3}{2(\sigma-1)}\ln\sigma\big)}\frac{1}{q\,2^\ell}\,.$$

Nous avons finalement démontré le théorème suivant :

Théorème 4. *Le seuil d'erreur d'Eigen est caractérisé par le développement*

$$q^* = \frac{\ln\sigma}{\ell} + \frac{\sigma-3}{2(\sigma-1)}\frac{\ln^2\sigma}{\ell^2} + o\Big(\frac{1}{\ell^2}\Big)\,,$$

dans le sens que nous précisons maintenant. Si le paramètre de mutation est tel que

$$q = \frac{\ln\sigma}{\ell} + \frac{c}{\ell}\,,$$

et

• *si la constante c est inférieure à $\frac{\sigma-3}{2(\sigma-1)}$, alors la concentration de master sequences x_0 est équivalente à*

$$x_0 \sim f_1(c,\sigma)\,\frac{1}{\ell}\,,$$

où la fonction f_1 ne dépend que de la constante σ et de c.

• *si la constante c est supérieure à $\frac{\sigma-3}{2(\sigma-1)}$, alors la concentration de master sequences est équivalente à*

$$x_0 \sim\ f_2(c,\sigma)\,\frac{\ell}{2^\ell}\,,$$

où la fonction f_2 ne dépend que de la constante σ et de c.

Chapitre 2

Processus encadrants pour le modèle d'Eigen

Nous implémentons un procédé itératif permettant d'encadrer le point stationnaire du système d'Eigen. Nous en déduisons un développement de la proportion d'individus dans les premières classes de Hamming.

1 Introduction

1.1 Schéma général

Dans toute la suite de cette thèse, nous nous placerons dans le cadre du paysage à un pic. Nous notons $\sigma > 1$ la fitness de la master sequence, qui est la chaîne $0 \cdots 0$. Nous allons dans ce chapitre et dans les suivants exploiter la structure de l'espace des chaînes $\{0,1\}^\ell$ en les regroupant dans les classes de Hamming. L'idée est naturelle : pour pouvoir estimer la proportion de master sequences, il nous suffit de comprendre les proportions d'individus dans chaque classe.

Dans l'étude de la proportion de master sequences, la difficulté provient des "mutations de retour", c'est-à-dire de la possibilité que les individus qui ne sont pas master sequence peuvent, au hasard d'une mutation engendrer des master sequences. Pour quantifier exactement cette possibilité, il est nécessaire de connaître toute la population. Dans son article original, Eigen parle déjà du rôle important que joue le nuage de mutants dans la stabilité de la quasiespèce. Au cours de ce chapitre, nous allons regrouper les individus dans leur classe de Hamming et estimer progressivement les proportions des individus dans chaque classe, en commençant par ceux qui sont le plus proche de la master sequence. Les individus de la classe 1 sont les plus proches de la master sequence et n'ont besoin que d'une mutation bien placée pour donner naissance à une master sequence. Plus généralement, les individus constituant la classe k ont besoin de k mutations bien placées pour donner naissance à des master sequences, ce qui se fait à taux

$$M_{k0} = q^k (1-q)^{\ell-k}\,.$$

Cette quantité tend vers 0 comme q^k, il peut donc sembler naturel de considérer seulement la première classe de Hamming pour obtenir une première approxima-

tion. Cependant, comme nous allons le voir, il nous faudra prendre en compte les individus jusqu'à la seconde classe de Hamming pour obtenir les premiers termes du développement de la proportion de master sequences. Notre plan est de commencer par estimer la proportion de master sequence sans aucune hypothèse supplémentaire sur la répartition de la population. Nous nous servirons ensuite de cette estimation pour encadrer la proportion d'individus dans la classe 1, et cet encadrement nous sera utile pour améliorer notre estimation sur le nombre de master sequences. Nous prendrons ensuite en compte les master sequences, la classe 1 et la classe 2 pour améliorer encore cette estimation. Et ainsi de suite jusqu'à une classe L qui restera loin de ℓ, cela nous conduira à un développement de la proportion de master sequence. Contrairement au chapitre précédent, nous ne pourrons pas obtenir le développement dans le régime neutre, mais seulement dans le régime quasiespèce. Cependant nous allons obtenir exactement le même développement dans ce régime.

1.2 Le système d'Eigen pour les classes

Nous travaillons avec les classes de Hamming, nous notons x_i la proportion d'individus dont la classe de Hamming est i,

$$x_i = \sum_{u:C(u)=i} x_u \,. \tag{2.1}$$

Les équations différentielles d'Eigen s'écrivent aussi pour les proportions des classes de Hamming, nous nous placerons toujours à l'équilibre. En sommant les équations (2), nous obtenons tout d'abord

$$0 = \sum_{u:C(u)=i} \left(\sum_v f(v) x_v M_{vu} - x_u \sum_v f(v) x_v \right).$$

Commençons par isoler les master sequences dans les sommes. Bien sûr, la classe 0 est uniquement constituée de la master sequence. Comme la somme de toutes les proportions vaut 1,

$$0 = \sum_{u:C(u)=i} \left(\sigma x_0 M_{0u} + \sum_{v\neq 0} x_v M_{vu} - \sigma x_u x_0 - x_u(1-x_0) \right).$$

Nous pouvons maintenant développer, nous obtenons, d'après (2.1),

$$0 = \sigma x_0 \sum_{u:C(u)=i} M_{0u} + \sum_{v\neq 0} x_v \sum_{u:C(u)=i} M_{vu} - \sigma x_i x_0 - x_i(1-x_0)\,.$$

Constatons maintenant que la quantité

$$\sum_{u:C(u)=i} M_{vu}$$

ne dépend que de la classe de Hamming de la chaîne v. En effet, pour qu'un individu v engendre un enfant de classe i, tout ce qui compte est le nombre de bits à changer

dans la chaîne v. Si la chaîne v est de classe k nous notons cette quantité M_{ki} . Nous pouvons alors réécrire l'équation comme

$$0 = \sigma x_0 M_{0i} + \sum_{k=1}^{\ell} x_k M_{ki} - \sigma x_i x_0 - x_i(1 - x_0)\,.$$

En réordonnant les termes, nous obtenons que la proportion x_i d'individus dans la classe i est régie par l'équation

$$0 = -(\sigma - 1)x_i x_0 + \Big(f(i)M_{ii} - 1\Big)x_i + \sum_{k=0,\, k\neq i}^{\ell} f(k)x_k M_{ki}\,. \tag{2.2}$$

Disons un mot sur les quantités M_{ij}. Il existe une formule exacte permettant de les calculer, que l'on peut trouver par exemple au chapitre 6 de [Cer15] :

$$M_{ij} = \sum_{\substack{0\leq k\leq \ell-i \\ 0\leq l\leq i \\ k-l=j-i}} \binom{\ell-i}{k}\binom{i}{l} q^k(1-q)^{\ell-i-k}\, q^l(1-q)^{i-l}\,.$$

Dans cette somme, l'indice l représente les mutations arrivant sur des digits uns, et l'indice k les mutations sur des digits zéro. Nous allons considérer seulement les premières classes, c'est-à-dire que les indices i et j ne tendent pas vers l'infini avec ℓ, cela correspond à la situation où les chaînes comportent peu de uns. Pour pouvoir manipuler correctement nos termes de reste dans la suite, nous considérons deux cas :

• Si $i \leq j$, alors le terme d'indices $k = j - i$ et $l = 0$ converge vers une constante strictement positive dans le régime asymptotique donc M_{ij} aussi.

• Si $i > j$, dans ce cas, certains uns de la chaîne doivent être changés en zéros. Le terme prédominant dans la somme est le terme d'indices $k = 0$ et $l = i - j$, il est d'ordre q^{i-j} ce qui est donc également le cas de M_{ij}.

2 Les master sequences uniquement

Dans cette section, nous allons encadrer la proportion de master sequences en ne faisant aucune hypothèse sur le nuage de mutants. Si le nuage de redonne jamais de master sequences, cela conduira à moins de master sequences, donc à un minorant de la proportion de master sequence. D'un autre côté, si tous les individus qui ne sont pas des master sequences en sont très proches, nous obtiendrons un majorant. L'équation qui régit la proportion de master sequences à l'équilibre s'écrit, d'après (2.2),

$$0 = -(\sigma - 1)x_0^2 + \big(\sigma M_{00} - 1\big)x_0 + \sum_{k\geq 1}^{\ell} x_k M_{k0}\,. \tag{2.3}$$

Le dernier terme dans cette équation dépend des proportions x_k d'individus dans chaque classe. C'est le terme de création de master sequences à partir d'individus qui ne sont pas master sequence. Nous allons majorer et minorer ce terme pour pouvoir calculer la solution de l'équation et en déduire un encadrement de la proportion de master sequences.

2.1 Une borne inférieure

Nous obtenons moins de master sequences si le nuage de mutants ne récrée jamais de master sequences, cela revient à minorer dans l'équation

$$\sum_{k=1}^{\ell} x_k \, M_{k0} \geq 0 \, .$$

Cette minoration conduit à l'équation

$$-(\sigma - 1)\, X^2 + (\sigma M_{00} - 1)\, X = 0 \, . \tag{2.4}$$

La proportion de master sequence est supérieure à la racine positive de ce trinôme, c'est-à-dire que

$$x_0 \geq \Big(\frac{\sigma M_{00} - 1}{\sigma - 1} \Big)_+ \, ,$$

où nous désignons par $(x)_+$ la partie positive de x :

$$(x)_+ = \frac{x + |x|}{2} = \max(x,\, 0) \, . \tag{2.5}$$

Au chapitre précédent, nous avons plutôt rencontré la quantité $(1 - 2q)^{\ell/2}$. Cependant, dans ce chapitre et dans toute la suite, nous rencontrerons plutôt la quantité $(1 - q)^{\ell}$ au travers de l'expression suivante qui apparaîtra de nombreuses fois jusqu'au dernier chapitre, et que nous notons r_0 :

$$r_0 = \frac{\sigma(1 - q)^{\ell} - 1}{\sigma - 1} \, .$$

Avec cette notation, nous écrivons

$$x_0 \geq (r_0)_+ \, . \tag{2.6}$$

2.2 Une borne supérieure

Pour le majorant, nous allons faire comme si tous les individus de la population étaient aussi proches des master sequences que les individus de la classe 1. Tous les individus n'ont donc besoin que d'une seule mutation bien placée pour engendrer une master sequence. Cela revient à majorer

$$\sum_{k=1}^{\ell} x_k \, M_{k0} \leq M_{10} \, ,$$

que nous pouvons encore simplifier

$$\sum_{k=1}^{\ell} x_k \, M_{k0} \leq q \, .$$

L'équation (2.3) conduit alors à la racine positive du trinôme

$$P(X) = X^2 - r_0\, X - \frac{q}{\sigma - 1}\,.$$

La fonction P est un trinôme du second degré qui tend vers $+\infty$ en $+\infty$ et dont les racines sont de signes opposés. Pour majorer sa racine positive, il suffit de trouver une quantité dont l'image par P est positive. Calculons l'image de $(r_0)_+ + \sqrt{\frac{q}{\sigma-1}}$ par P :

$$P\left((r_0)_+ + \sqrt{\frac{q}{\sigma-1}}\right) = (r_0)_+^2 + 2(r_0)_+\sqrt{\frac{q}{\sigma-1}} - r_0(r_0)_+ - r_0\sqrt{\frac{q}{\sigma-1}}\,.$$

Après simplification, et d'après (2.5), nous obtenons

$$P\left((r_0)_+ + \sqrt{\frac{q}{\sigma-1}}\right) = |r_0|\sqrt{\frac{q}{\sigma-1}}\,.$$

Cette quantité est positive, nous en déduisons que

$$x_0 \leq (r_0)_+ + \sqrt{\frac{q}{\sigma-1}}\,. \tag{2.7}$$

A moins de connaître la nature du rapport q/r_0^2, nous ne pouvons rien dire de plus. Bien que r_0 soit défini à partir de ℓ et q, tous les régimes sont possibles pour le rapport r_0/q. Nous avons en effet

$$\begin{aligned}(\sigma-1)r_0 &= \sigma \exp\Big(\ell \ln(1-q)\Big) - 1\\ &= \sigma e^{-a} \exp\Big(-(\ell q - a) - \frac{aq}{2} + o(q)\Big) - 1\,.\end{aligned}$$

Nous nous placerons dans le cas critique $a = \ln\sigma$, ainsi r_0 tend vers 0 et

$$(\sigma-1)r_0 = -(\ell q - a) - \frac{aq}{2} + o\Big((\ell q - a) + q\Big)\,. \tag{2.8}$$

La vitesse à laquelle r_0 tend vers 0 dépend donc de la vitesse à laquelle le produit ℓq tend vers sa limite.

Les deux bornes que nous avons ainsi obtenues ne coïncident donc pas asymptotiquement, cependant elles situent grossièrement la proportion de master sequence et c'est tout ce dont nous aurons besoin pour la suite. Avec ce premier encadrement, nous avons montré que

$$(r_0)_+ \leq x_0 \leq (r_0)_+ + \sqrt{\frac{q}{\sigma-1}}\,. \tag{2.9}$$

Cette partie positive va intervenir tout le long de nos encadrements, elle sera positive dans le régime quasiespèce et nous pourrons alors dire quelque chose sur la proportion de master sequences, si elle est nulle, c'est que nous serons dans le régime neutre et nous obtiendrons seulement un terme de reste.

3 Les master sequences et la classe 1

3.1 La classe 1

Intéressons-nous à l'équation régissant la proportion d'individus de la classe 1. Comme nous avons un encadrement de la proportion de master sequences, nous écrivons l'équation (2.2) pour la classe 1 comme

$$0 = -\Big(1 - M_{11} + (\sigma - 1)x_0\Big)x_1 + \sigma x_0 M_{01} + \sum_{k=2}^{\ell} x_k M_{k1}\,. \tag{2.10}$$

De la même manière que précédemment, le dernier terme est compliqué puisqu'il représente la possibilité que n'importe quel individu puisse donner naissance à un individu de classe 1 s'il subit les bonnes mutations. Il nous faut à nouveau trouver une façon de majorer et de minorer ce terme pour qu'il ne dépende que de quantités qui nous sont accessibles, comme la proportion de master sequences.

Une borne inférieure

Nous allons dire que les individus des classes 2 ou plus ne redonnent jamais d'individus de la classe 1. Seules les master sequences peuvent devenir des individus de classe 1, cela revient à minorer dans l'équation

$$\sum_{k=2}^{\ell} x_k M_{k1} \geq 0\,.$$

Le point stationnaire de l'équation minorante va donc dépendre de la proportion de master sequences x_0,

$$0 = -\Big(1 - M_{11} + (\sigma - 1)x_0\Big)X + \sigma x_0 M_{01}\,. \tag{2.11}$$

Nous allons maintenant transférer les estimées que nous avons obtenues sur la proportion de master sequences, nous avons en effet, d'après (2.9),

$$x_1 \geq \frac{\sigma M_{01}(r_0)_+}{1 - M_{11} + (\sigma - 1)\Big((r_0)_+ + \sqrt{\frac{q}{\sigma - 1}}\Big)}\,. \tag{2.12}$$

Le dénominateur ne s'annule jamais, il tend en effet vers $1 - \frac{1}{\sigma}$ dans le régime asymptotique.

Une borne supérieure

Tous les individus des classes 2 et plus peuvent muter vers la classe 1 comme s'ils étaient de la classe 2. Cela revient à la majoration suivante :

$$\sum_{k=2}^{\ell} x_k M_{k1} \leq M_{21}\,.$$

L'équation devient

$$0 = -\Big(1 - M_{11} + (\sigma-1)x_0\Big)X + \sigma x_0 M_{01} + M_{21}\,. \tag{2.13}$$

Cette équation conduit à

$$x_1 \leq \frac{\sigma M_{01}\Big((r_0)_+ + \sqrt{\frac{q}{\sigma-1}}\Big) + M_{21}}{1 - M_{11} + (\sigma-1)(r_0)_+}\,. \tag{2.14}$$

3.2 Retour sur la proportion de master sequences

Nous savons donc encadrer la proportion d'individus dans la classe 1, nous allons en déduire une estimation plus précise de la proportion de master sequences. Nous réécrivons l'équation (2.3) en estimant plus précisément la proportion d'individus dans la classe 1 :

$$0 = -(\sigma-1)x_0^2 + (\sigma-1)r_0\,x_0 + x_1 M_{10} + \sum_{k=2}^{\ell} x_k\,M_{k0}\,. \tag{2.15}$$

Avant d'encadrer à nouveau le dernier terme, revenons un moment sur nos notations. Nous allons obtenir une nouvelle estimation sur la proportion de master sequences. Attribuons une notation à nos encadrements pour la proportion d'individus de la classe 1 : posons en accord avec (2.12) et (2.14)

$$\underline{\rho_1} = \frac{\sigma M_{01}(r_0)_+}{1 - M_{11} + (\sigma-1)\Big((r_0)_+ + \sqrt{\frac{q}{\sigma-1}}\Big)}\,.$$

$$\overline{\rho_1} = \frac{\sigma M_{01}\Big((r_0)_+ + \sqrt{\frac{q}{\sigma-1}}\Big) + 2q}{1 - M_{11} + (\sigma-1)(r_0)_+}\,.$$

Avec ces notations, nous pouvons écrire

$$\underline{\rho_1} \leq x_1 \leq \overline{\rho_1}\,.$$

Une borne inférieure

Cette fois, la minoration correspond à négliger les retours des classes 2 et plus :

$$\sum_{k=2}^{\ell} x_k\,M_{k0} \geq 0\,.$$

Pour la borne inférieure, nous obtenons l'équation

$$0 = -(\sigma-1)X^2 + (\sigma-1)r_0 X \;\; + x_1 M_{10}\,. \tag{2.16}$$

Définissons alors la fonction P par

$$P(X) = X^2 - r_0 X - \frac{M_{10}}{\sigma-1}\underline{\rho_1}\,.$$

Pour obtenir un minorant de la proportion de master sequence, il suffit de trouver une quantité dont l'image par la fonction P est négative. Calculons :

$$P\left((r_0)_+ + \frac{M_{10}\,\underline{\rho_1}}{(\sigma-1)r_0}\right) = \left(\frac{(r_0)_+}{r_0} - 1\right)\frac{2M_{10}\,\underline{\rho_1}}{\sigma-1}\,,$$

Cette quantité est toujours négative, nous en déduisons donc

$$x_0 \geq (r_0)_+ + \frac{M_{10}\,\underline{\rho_1}}{(\sigma-1)r_0}\,.$$

Une borne supérieure

La majoration consiste à faire comme si tous les individus de classe 2 ou plus pouvaient devenir des master sequences comme des classes 2. Cela revient à majorer

$$\sum_{k=2}^{\ell} x_k\, M_{k0} \leq M_{20} \leq q^2\,.$$

Nous obtenons alors l'équation

$$0 = -(\sigma-1)X^2 + (\sigma-1)r_0\,X + M_{10}\overline{\rho_1} + q^2\,.$$

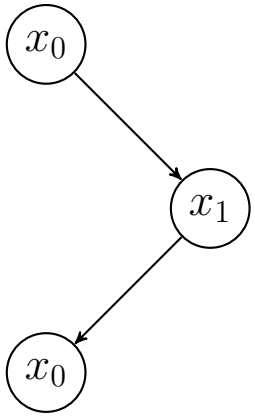
De la même manière que nous l'avons fait précédemment, nous obtenons un majorant de la racine positive du trinôme avec l'expression analogue :

$$x_0 \leq (r_0)_+ + \sqrt{\frac{M_{10}\overline{\rho_1} + q^2}{\sigma-1}}\,.$$

Ces deux nouveaux encadrements améliorent les encadrements précédents et dépendent des encadrements $\underline{\rho_1}$ et $\overline{\rho_1}$. Nous avions obtenu ces bornes à partir de notre encadrement initial sur la proportion de master sequences. Avec cette nouvelle estimation, nous pourrons donc obtenir une estimation encore meilleure de la proportion de master sequence. Ainsi, nous allons itérer ce procédé dans la suite.

3.3 Répéter le procédé

Non seulement nous allons répéter ce procédé, mais nous allons le faire une infinité de fois. La limite de ces estimations encadrera très bien la proportion de master sequences. Le schéma suivant illustre une étape du processus.



Introduisons des notations, à la première étape, qui correspond à la section 2, nous avions

$$\underline{\rho_0}^{(0)} \leq x_0 \leq \overline{\rho_0}^{(0)} ,$$

avec

$$\underline{\rho_0}^{(0)} = (r_0)_+ ,$$

et

$$\overline{\rho_0}^{(0)} = (r_0)_+ + \sqrt{\frac{q}{\sigma - 1}} .$$

Ensuite, dans la section 3.2, nous avons amélioré ces bornes

$$\underline{\rho_0}^{(0)} \leq \underline{\rho_0}^{(1)} \leq x_0 \leq \overline{\rho_0}^{(1)} \leq \overline{\rho_0}^{(0)} ,$$

avec

$$\underline{\rho_0}^{(1)} = (r_0)_+ + \frac{M_{10}\underline{\rho_1}}{(\sigma - 1)r_0} ,$$

et

$$\overline{\rho_0}^{(1)} = (r_0)_+ + \sqrt{\frac{M_{10}\overline{\rho_1} + q^2}{\sigma - 1}} .$$

Voyons maintenant la n-ième étape.

Une minoration

Supposons que nous ayons un minorant de la proportion de master sequence $\underline{\rho_0}^{(n)}$. Alors la proportion d'individus dans la classe 1 est minorée par

$$x_1 \geq \frac{\sigma M_{01}\underline{\rho_0}^{(n)}}{1 - M_{11} + (\sigma - 1)\overline{\rho_0}^{(0)}} .$$

Pour simplifier, nous allons toujours encadrer les termes x_0 présents aux dénominateurs avec nos premières estimations $\underline{\rho_0}^{(0)}$ et $\overline{\rho_0}^{(0)}$. Afin de simplifier les futures expressions, introduisons la notation suivante

$$\underline{D_{01}} = \frac{M_{01}}{1 - M_{11} + (\sigma - 1)\overline{\rho_0}^{(0)}} ,$$

nous pouvons alors écrire que la proportion d'individus de classe 1 vérifie

$$x_1 \geq \sigma \underline{D_{01}}\, \underline{\rho_0}^{(n)} .$$

Notons

$$\underline{\rho_1}^{(n)} = \sigma \underline{D_{01}}\, \underline{\rho_0}^{(n)} . \tag{2.17}$$

Reprenons maintenant l'équation (2.16) donnant le minorant de la proportion de master sequence,

$$-(\sigma - 1)X^2 + (\sigma - 1)r_0 X + M_{10}x_1 = 0 .$$

Un minorant de la proportion de master sequence est donné par la solution de l'équation (2.15) :

$$-(\sigma-1)X^2+(\sigma-1)r_0X+M_{10}\underline{\rho_1}^{(n)}=0\,. \tag{2.18}$$

En remplaçant $\underline{\rho_1}^{(n)}$ par son expression (2.17), nous nous ramenons à l'équation

$$-(\sigma-1)X^2+(\sigma-1)r_0X+\sigma M_{10}\underline{D_{01}}\,\underline{\rho_0}^{(n)}=0\,. \tag{2.19}$$

Arrêtons-nous un instant pour observer cette formule : la quantité $\underline{\rho_0}^{(n+1)}$ est définie comme la racine positive de cette équation. Enonçons un lemme général sur ce genre de suite récurrente qui nous servira plusieurs fois.

Lemme 5. *Soient a,b,c,d des réels positifs tels que $-a+b+c+d<0$ et (u_n) une suite récurrente définie par $u_0\in[0,1]$ et pour tout entier n, u_{n+1} est la racine positive du trinôme $-aX^2+bX+c+du_n$.*

Alors, la suite $(u_n)_{n\in\mathbb{N}}$ converge vers la racine positive du trinôme

$$-aX^2+(b+d)X+c\,.$$

Démonstration. Considérons la fonction f qui à u_n associe u_{n+1} : pour tout réel x positif, le point $f(x)$ est la racine positive de l'équation

$$-a\big(f(x)\big)^2+bf(x)+c+dx=0\,.$$

Cette fonction est croissante : prenons deux réels positifs $\alpha<\beta$, comme d est positif, la courbe définie par

$$y=-ax^2+bx+c+d\alpha\,,$$

est toujours en dessous de la courbe définie par

$$y=-ax^2+bx+c+d\beta\,.$$

Ainsi, $f(\alpha)$ est inférieur à $f(\beta)$ et f est bien croissante. De plus, l'intervalle $[0,1]$ est stable par f, en effet, la racine du trinôme $-aX^2+bX+c+d$ est inférieure à 1 puisque l'image de 1 par cette fonction est négative par hypothèse. La suite $(u_n)_{n\in\mathbb{N}}$ est donc monotone et encadrée entre 0 et 1, donc cette suite converge. Sa limite est un point fixe λ de la fonction f. Le point λ, limite de la suite $(u_n)_{n\in\mathbb{N}}$ est donc la solution positive de l'équation

$$-a\lambda^2+(b+d)\lambda+c=0\,,$$

ce que nous voulions montrer. □

Le lemme 5 nous donne immédiatement la convergence et la limite de la suite $(\underline{\rho_0}^{(n)})_{n\in\mathbb{N}}$ de minorants de la proportion de master sequence. Cette limite est encore un minorant de cette proportion, elle est la solution positive du trinôme

$$(\sigma-1)X^2-\Big((\sigma-1)r_0+\sigma M_{10}\underline{D_{01}}\Big)X=0\,. \tag{2.20}$$

Ainsi nous obtenons

$$x_0\geq\Big(r_0+\frac{\sigma M_{10}\underline{D_{01}}}{\sigma-1}\Big)_+\,. \tag{2.21}$$

Une majoration

Supposons que nous ayons une borne supérieure sur la proportion de master sequence. En reprenant (2.13), la proportion d'individus dans la classe 1 est majorée par

$$x_1 \leq \frac{\sigma M_{01}\overline{\rho_0}^{(n)} + M_{21}}{1 - M_{11} + (\sigma - 1)\underline{\rho_0}^{(0)}}\,.$$

Définissons des notations analogues que pour la borne inférieure :

$$\overline{D_{01}} = \frac{M_{01}}{1 - M_{11} + (\sigma - 1)\underline{\rho_0}^{(0)}}\,,$$

et

$$\overline{D_{21}} = \frac{M_{21}}{1 - M_{11} + (\sigma - 1)\underline{\rho_0}^{(0)}}\,.$$

Nous pouvons écrire que la proportion d'individus dans la classe 1 est majorée par

$$x_1 \leq \sigma\overline{D_{01}}\ \overline{\rho_0}^{(n)} + \overline{D_{21}}\,,$$

et nous posons

$$\overline{\rho_1}^{(n)} = \sigma\overline{D_{01}}\ \overline{\rho_0}^{(n)} + \overline{D_{21}}\,.$$

Un majorant de la proportion de master sequence est alors donné par la solution de l'équation (2.15) :

$$-(\sigma - 1)X^2 + (\sigma - 1)r_0X + \overline{\rho_1}^{(n)}M_{10} + q^2 = 0\,. \tag{2.22}$$

En remplaçant $\underline{\rho_1}^{(n)}$ par son expression, nous obtenons que la proportion de master sequence est inférieure à la solution de

$$-(\sigma - 1)X^2 + (\sigma - 1)r_0X + \sigma M_{10}\overline{D_{01}}\ \overline{\rho_0}^{(n)} + \overline{D_{21}}M_{10} + q^2 = 0\,. \tag{2.23}$$

La suite $\left(\overline{\rho_0}^{(n)}\right)_{n\geq 0}$ est donc une suite du même type que celle du lemme 5, cette fois, nous obtenons une suite décroissante de majorants de la proportion de master sequence à l'équilibre dont la limite est encore un majorant de cette proportion. Cette limite est la solution positive du trinôme

$$(\sigma - 1)X^2 - \Big((\sigma - 1)r_0 + \sigma M_{10}\overline{D_{01}}\Big)X + \overline{D_{21}}M_{10} + q^2 = 0\,. \tag{2.24}$$

Nous rencontrerons souvent ce genre de trinôme du second degré dans les estimations des solutions stationnaires de nos processus. Afin d'éviter des expressions avec des racines carrées compliquées, arrêtons-nous un instant pour énoncer un lemme élémentaire qui nous sera bien utile. Dans toutes les équations du second degré que nous rencontrerons, le terme constant sera très souvent le terme de reste.

Lemme 6. *Soit P un trinôme du second degré unitaire*

$$P(X) = X^2 - bX - c\,,$$

avec c positif et b non nul. Alors la racine positive r_+ de P est encadrée par

$$(b)_+ \leq r_+ \leq (b)_+ + \frac{c}{|b|}\,.$$

Démonstration. Les deux racines de P sont réelles et de signes opposés. Le polynôme P prend donc des valeurs négatives sur $[0, r_+]$ et positives sur $[r_+, \infty]$. Pour montrer les inégalités annoncées, il suffit de calculer l'image de P sur les deux bornes :

$$P\big((b)_+\big) = (b)_+^2 - b(b)_+ - c = -c\,.$$

Comme $-c$ est négatif, nous en déduisons que $(b)_+ \leq r_+$. D'autre part

$$\begin{aligned} P\Big((b)_+ + \frac{c}{|b|}\Big) &= \Big((b)_+ + \frac{c}{|b|}\Big)^2 - b\Big((b)_+ + \frac{c}{|b|}\Big) - c \\ &= \Big(\frac{c}{b}\Big)^2 + \frac{2(b)_+ - b - |b|}{|b|}\,c\,. \end{aligned}$$

D'après (2.5), nous obtenons

$$P\Big((b)_+ + \frac{c}{|b|}\Big) = \Big(\frac{c}{b}\Big)^2,$$

cette image est bien positive, ce que nous voulions démontrer. □

Ainsi nous obtenons

$$x_0 \leq \Big(r_0 + \frac{\sigma M_{10}\overline{D_{01}}}{\sigma - 1}\Big)_+ + \frac{q^2 + \overline{D_{21}}M_{10}}{\big|r_0 + \frac{\sigma}{\sigma-1}M_{10}\overline{D_{01}}\big|}\,. \tag{2.25}$$

3.4 Développement à l'ordre 1

En gardant les termes communs aux deux bornes, nous écrivons

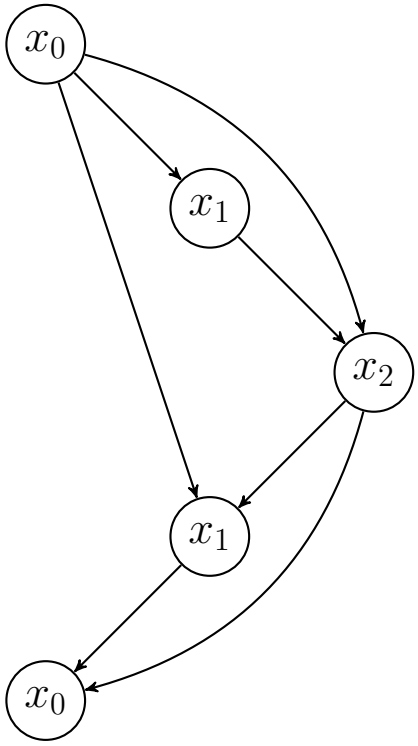
$$x_0 = \left(r_0 + \frac{\sigma M_{01}M_{10}}{(\sigma-1)(1-M_{11})}\right)_+ + O\Big(\frac{q^2}{r_0+q}\Big)\,. \tag{2.26}$$

• Si r_0 est très grand devant q, ce terme de reste tend vers 0 et est toujours négligeable devant le terme dans la partie positive.
• Si r_0 est très petit devant q, ce terme de reste est d'ordre q, dans ce cas, tous les termes sont potentiellement des termes de reste.

En fonction de la nature du rapport q/r_0, nous ne connaissons pas encore le premier terme du développement, il faut donc aller plus loin et prendre en compte les individus de la classe 2.

4 Les master sequences et les classes 1 et 2

Le schéma général est le suivant, une flèche de x_i vers x_j signifie que l'on utilise les estimations sur le point x_i pour encadrer le point d'équilibre x_j.



Repartons de nos approximations précédentes (2.21) et (2.25), nous notons

$$\underline{\rho_0} = \Big(r_0 + \frac{\sigma M_{10}\underline{D_{01}}}{\sigma - 1}\Big)_+ ,$$

et

$$\overline{\rho_0} = \Big(r_0 + \frac{\sigma M_{10}\overline{D_{01}}}{\sigma - 1}\Big)_+ + \frac{q^2 + \overline{D_{21}}M_{10}}{\big|r_0 + \frac{\sigma}{\sigma-1}\overline{D_{01}}M_{10}\big|} .$$

Nous pouvons donc écrire

$$\underline{\rho_0} \leq x_0 \leq \overline{\rho_0} .$$

Encadrement des classes 1

De la même façon que nous l'avons fait à la section 3.1, nous pouvons reprendre l'équation (2.10) et nous obtenons de nouvelles bornes sur la proportion d'individus dans la classe 1.

$$\frac{\sigma M_{01}\underline{\rho_0}}{1 - M_{11} + (\sigma - 1)\overline{\rho_0}} \leq x_1 \leq \frac{\sigma M_{01}\overline{\rho_0} + M_{21}}{1 - M_{11} + (\sigma - 1)\underline{\rho_0}} . \tag{2.27}$$

Encadrement des classes 2

L'équation (2.2) pour la classe 2 conduit à

$$0 = -\Big(1 - M_{22} + (\sigma - 1)x_0\Big)x_2 + \sigma M_{02}x_0 + M_{12}x_1 + \sum_{k\geq 3}^{\ell} x_k M_{k2} .$$

Nous pouvons encadrer le terme qui dépend du reste de la population avec les mêmes idées : nous obtenons une minoration si les individus des classes 3 et supérieures n'engendrent jamais d'individu de classe 2. Nous obtenons une majoration si ces même individus engendrent des classes 2 comme le font les individus de la classe 3. Cela revient à encadrer dans l'équation :

$$0 \leq \sum_{k\geq 3}^{\ell} x_k M_{k2} \leq M_{32}\,.$$

Reprenons les estimations que nous avons obtenues pour la proportion de master sequence et des individus de classe 1. Nous obtenons ainsi d'après (2.27), pour le minorant

$$x_2 \geq \frac{\sigma M_{02}\underline{\rho_0}}{1-M_{22}+(\sigma-1)\overline{\rho_0}} + \frac{\sigma M_{12}M_{01}\underline{\rho_0}}{\big(1-M_{22}+(\sigma-1)\overline{\rho_0}\big)\big(1-M_{11}+(\sigma-1)\overline{\rho_0}\big)}\,.$$

Nous étendons la définition de D_{ij} pour des indices i et j quelconques, en posant

$$\overline{D_{ij}} = \frac{M_{ij}}{1-M_{jj}+(\sigma-1)\underline{\rho_0}}\,,$$

et

$$\underline{D_{ij}} = \frac{M_{ij}}{1-M_{jj}+(\sigma-1)\overline{\rho_0}}\,.$$

Avec ces notations, nous obtenons

$$x_2 \geq \sigma\underline{\rho_0}\Big(\underline{D_{02}} + \underline{D_{01}}\,\underline{D_{12}}\Big)\,,$$

et pour le majorant

$$x_2 \leq \sigma\overline{\rho_0}\Big(\overline{D_{02}} + \overline{D_{01}}\,\overline{D_{12}}\Big) + \overline{D_{32}} + \overline{D_{12}D_{21}}\,.$$

Pour alléger les formules, nous généralisons la notation précédente et nous notons, pour tout $i_1, \cdots, i_n$ entiers entre 1 et ℓ,

$$D_{i_1,\cdots,i_n} = \prod_{k=2}^{n} D_{i_{k-1}i_k}\,.$$

Maintenant que nous avons une estimation de la proportion d'individus dans la classe 2, revenons sur la proportion d'individus dans la classe 1.

$$\sigma\underline{\rho_0}\Big(\underline{D_{02}} + \underline{D_{012}}\Big) \leq x_2 \leq \sigma\overline{\rho_0}\Big(\overline{D_{02}} + \overline{D_{012}}\Big) + \overline{D_{32}} + \overline{D_{12}D_{21}}\,. \tag{2.28}$$

Attribuons une nouvelle notation à ces bornes :

$$\underline{\rho_2} = \sigma\underline{\rho_0}\Big(\underline{D_{02}} + \underline{D_{012}}\Big)\,, \tag{2.29}$$

et pour le majorant

$$\overline{\rho_2} = \sigma\overline{\rho_0}\Big(\overline{D_{02}} + \overline{D_{012}}\Big) + \overline{D_{32}} + \overline{D_{12}D_{21}}\,.$$

Encadrement des classes 1

Nous pouvons maintenant reprendre notre estimation sur la proportion d'individus dans la classe 1, en considérant l'encadrement de la proportion d'individus dans la classe 2 :

$$0 = -\Big(1 - M_{11} + (\sigma - 1)x_0\Big)x_1 + \sigma M_{01}x_0 + M_{21}x_2 + \sum_{k=3}^{\ell} x_k M_{k1}\,. \tag{2.30}$$

Nous pouvons nous servir de cette estimation pour avoir un meilleur encadrement sur l'équation (2.10) et intégrer nos estimations sur la proportion d'individus de la classe 2 en encadrant

$$0 \leq \sum_{k=3}^{\ell} x_k\, M_{k1} \leq M_{31}\,.$$

Avec ces notations, nous obtenons

$$\sigma \underline{D_{01}}\ \underline{\rho_0} + \underline{D_{21}}\ \underline{\rho_2}\ \leq\ x_1 \leq \sigma\overline{D_{01}}\ \overline{\rho_0} + \overline{D_{21}}\ \overline{\rho_2} + \overline{D_{31}}\,.$$

Nous utilisons l'encadrement (2.28) pour factoriser par la proportion de master sequences

$$\sigma\underline{\rho_0}\Big(\underline{D_{01}} + \underline{D_{021}} + \underline{D_{0121}}\Big)\ \leq\ x_1 \leq \sigma\overline{\rho_0}\Big(\overline{D_{01}} + \overline{D_{021}} + \overline{D_{0121}}\Big) + \overline{D_{31}} + \overline{D_{321}} + \overline{D_{121}D_{21}}\,.$$

Remarquons que nos termes vont s'écrire avec un produit de coefficients D. Nous allons donc maintenant introduire de nouvelles notations pour les sommes de produits de nos coefficients D. Définissons C_{ij}, pour $i < j$, comme la somme sur les chemins croissants $i < i_1 < \cdots < i_n < j$ des coefficients $D_{i_1,\cdots,i_n}$:

$$C_{ij} = \sum_{i=i_1<\cdots<i_n=j} D_{i_1\cdots i_n}\,, \tag{2.31}$$

De même, définissons $S_{ij}(k)$ comme la somme sur les chemins qui partent de i, montent au plus jusqu'à j puis redescendent à k :

$$S_{ij}(k) = \sum_{\substack{i=i_1<\cdots<i_n\leq j\\ i_n>i_{n+1}>\cdots>i_m=k}} D_{i_1\cdots i_m}\,. \tag{2.32}$$

Avec ces notations, nous pouvons écrire simplement les deux encadrements que nous avons obtenus dans cette section :

$$\sigma\underline{\rho_0}\underline{C_{02}} \leq x_2 \leq \sigma\overline{\rho_0}\overline{C_{02}} + \overline{D_{32}} + \overline{D_{12}D_{21}}\,,$$

et

$$\sigma\underline{\rho_0}\underline{S_{02}}(1) \leq x_1 \leq \sigma\overline{\rho_0}\overline{S_{02}}(1) + \overline{D_{31}} + \overline{D_{321}} + \overline{D_{121}D_{21}}\,,$$

Voyons que D_{32} et $D_{12}M_{21}$ sont tous deux d'ordre q, de même les termes D_{31}, D_{321} et $D_{121}M_{21}$ sont tous d'ordre q^2.

4.1 Retour sur les master sequences

Revenons à la proportion de master sequences, l'équation qui permet de calculer la proportion de master sequence à l'équilibre est

$$-(\sigma-1)x_0^2+(\sigma-1)r_0x_0+x_1M_{10}+x_2M_{20}+\sum_{k\geq 3}^{\ell}x_kM_{k0}=0\,.$$

Cette fois, nous encadrons le dernier terme par

$$0\leq\sum_{k\geq 3}^{\ell}x_kM_{k0}\leq M_{30}\leq q^3\,.$$

Nous appliquons alors le même raisonnement qu'à la section (3.3) : d'après le lemme 5, la proportion de master sequence est minorée par la solution positive du trinôme

$$(\sigma-1)X^2-\Big((\sigma-1)r_0+\sigma M_{10}\underline{S_{02}}(1)+\sigma M_{20}\underline{C_{02}}\Big)X=0\,.$$

La proportion de master sequence est majorée par la solution positive du trinôme

$$(\sigma-1)X^2-\Big((\sigma-1)r_0+\sigma M_{10}\overline{S_{02}}(1)+\sigma M_{20}\overline{C_{02}}\Big)X+R=0\,,$$

où R est un terme d'ordre q^3. Nous appliquons maintenant le lemme 6 à cette équation et nous obtenons

$$\begin{aligned}\left(r_0+\frac{\sigma}{\sigma-1}\Big(M_{10}\underline{S_{02}}(1)+M_{20}\underline{C_{02}}\Big)\right)_+ &\\ \leq x_0\leq &\\ \left(r_0+\frac{\sigma}{\sigma-1}\Big(M_{10}\overline{S_{02}}(1)+M_{20}\overline{C_{02}}\Big)\right)_+ &+O\Big(\frac{q^3}{r_0+q}\Big)\,,\end{aligned}\tag{2.33}$$

cette inégalité nous permet de déuire le théorème suivant :

Théorème 7. *Dans le régime asymptotique, la proportion de master sequence x_0 admet le développement suivant*

$$x_0=\left(r_0+\frac{\sigma}{\sigma-1}\frac{M_{10}M_{01}}{1-M_{11}}\right)_+ +O(qr_0+q^2)\,.$$

Nous connaissons donc maintenant les premiers termes du développement de la proportion de master sequence, ce développement est valable pour tous les régimes de q et r_0. Nous pouvons faire ce raisonnement pour un nombre de classe fixé. Mais avant cela, vérifions si le développement que nous avons obtenu pour la proportion de master sequence est bien en accord avec le développement de la fitness moyenne λ que nous avons effectuée dans le premier chapitre.

4.2 Le même développement !

Comme nous travaillons dans le paysage à un pic, la fitness moyenne et la proportion de master sequences sont directement reliées au travers de l'égalité

$$\lambda - 1 = (\sigma - 1)\rho_0 \,.$$

Nous allons maintenant vérifier que les deux développements que nous avons obtenus par des méthodes complètement différentes coïncident bien. Nous avons montré que

$$(\sigma - 1)\rho_0 = \left(\sigma(1-q)^\ell - 1 + \frac{\sigma(1-q)^{2(\ell-1)}\ell q^2}{1-(1-q)^\ell + (\sigma-1)r_0}\right)_+ + o(q)\,.$$

Rappelons-nous que

$$\sigma(1-q)^\ell = 1 - (\ell q - \ln\sigma) - \frac{1}{2}q\ln\sigma + o\Big((\ell q - \ln\sigma) + q\Big)\,.$$

Ainsi, en remplaçant ces quantités, nous obtenons

$$(\sigma - 1)\rho_0 = \left(- (\ell q - \ln\sigma) - \frac{1}{2}q\ln\sigma + \frac{q\ln\sigma}{\sigma - 1}\right)_+ + o\Big(q + (\ell q - \ln\sigma)\Big)\,.$$

C'est-à-dire finalement

$$(\sigma - 1)\rho_0 = \left(- \frac{\sigma - 3}{2(\sigma-1)}q\ln\sigma - (\ell q - \ln\sigma)\right)_+ + o\Big(q + (\ell q - \ln\sigma)\Big)\,.$$

Rappelons que l'équation (1.28) donnait

$$-\frac{(\sigma-1)^2}{2^\ell\sigma(\lambda-1)} + \lambda - 1 = -\frac{\sigma-3}{2(\sigma-1)}q\ln\sigma - (\ell q - \ln\sigma) + o\Big(q + (\ell q - \ln\sigma) + (\lambda - 1)\Big)\,.$$

Si la quantité à l'intérieur de la parenthèse est positive, nous obtenons exactement le même développement. Cette condition correspond au régime quasiespèce. Si la quantité est négative, c'est que nous sommes dans le régime neutre, ces encadrements ne nous permettent pas d'estimer la proportion de master sequences.

5 Les master sequences et les classes 1 jusqu'à L

Fixons-nous un entier L qui ne dépend pas des paramètres. Le raisonnement général sera le suivant. Nous allons estimer les proportions des individus à l'équilibre jusqu'à la classe L en tenant compte uniquement des classes inférieures à L. Nous les reprendrons ensuite pour améliorer nos estimations précédentes. Nous appliquerons ensuite le lemme 5 puis le lemme 6. Nous poussons le raisonnement et nous estimons toutes les classes jusqu'à L sans prendre en compte aucune mutation de retour. Considérons la proportion d'individus dans la classe $k < L$, l'équation (2.2) conduit à

$$0 = -\Big(1 - M_{kk} + (\sigma-1)x_0\Big)x_k + \sigma M_{0k}x_0 + \sum_{j=1}^{k-1} M_{jk}x_j + \sum_{j>k}^{\ell} x_j M_{jk}\,.$$

Nous pouvons à nouveau majorer le terme dépendant des classes supérieures :

$$0 \leq \sum_{j>k}^{\ell} x_j M_{jk} \leq M_{k+1,k}\,.$$

Nous obtenons, pour toute classe intermédiaire $k \leq L$,

$$\sigma \underline{\rho_0 D_{0k}} + \sum_{j=1}^{k-1} \underline{D_{jk}\rho_j} \leq x_k \leq \sigma \overline{\rho_0} \overline{D_{0k}} + \sum_{j=1}^{k-1} \overline{D_{jk}\rho_j} + R\,,$$

où R est d'ordre q.

Lemme 8. *Nous pouvons aussi l'écrire comme*

$$\sigma \underline{C_{0k}\rho_0} \leq x_k \leq \sigma \overline{C_{0k}} \overline{\rho_0} + R\,,$$

où R est d'ordre q et C_{0k} est la somme définie à la formule (2.31)

Démonstration. Ces formules se démontrent par récurrence. Si la proportion d'individus dans les classes 0 à $k-1$ est donnée par les formules ci-dessus, la proportion d'individus dans la classe k est donnée par

$$x_k = \sigma\Big(D_{0k} + \sum_{j=1}^{k-1} C_{0,j} D_{jk}\Big) x_0 + R\,,$$

où R est d'ordre q. La première quantité entre parenthèses est la somme sur tous les chemins entre 0 et k :

$$D_{0k} + \sum_{j=1}^{k-1} C_{0,j} D_{jk} = C_{0,k}\,.$$

□

Une fois que nous avons estimé toutes ces quantités, nous pouvons redescendre et réestimer les classes $k \in \{1, \cdots, L\}$ par

$$x_k = \sigma D_{0k} x_0 + \sigma \sum_{j=1, j\neq k}^{L} D_{jk} x_j + R\,,$$

où R est d'ordre q^{L+1-k}. Nous pouvons aussi écrire ce développement sous la forme

$$\sigma \underline{S_{0L}(k)\rho_0} \leq x_k \leq \sigma \overline{S_{0L}(k)} \overline{\rho_0} + q^{L+1-k}\,,$$

où $S_{0L}(k)$ est définie par son expression (2.32). Ces formules se montrent de la même façon que précédemment par récurrence. Une fois que nous avons toutes ces estimations, nous pouvons revenir à la proportion de master sequence.

5.1 Retour à la proportion de master sequences

La proportion de master sequences est solution de l'équation

$$-(\sigma+1)X^2+\Big((\sigma-1)r_0\Big)X+\sum_{k=1}^{L}M_{k0}\rho_k+\sum_{j>L}^{\ell}x_jM_{jk}=0\,.$$

Nous encadrons maintenant une dernière fois ce dernier terme :

$$0\leq\sum_{j>L}^{\ell}x_jM_{jk}\leq M_{L+1,0}\leq q^{L+1}\,.$$

D'après le lemme 6, nous obtenons que la proportion de master sequence est estimée par la racine de

$$(\sigma-1)X^2-\Big((\sigma-1)r_0+\sigma\sum_{k=1}^{L}M_{k0}S_{0L}(k)\Big)X-R=0\,,$$

où R est d'ordre q^{L+1}. Nous en déduisons

$$x_0=r_0+\sigma\sum_{k=1}^{L}M_{k0}S_{0L}(k)+O\Big(\frac{q^{L+1}}{r_0+q}\Big)\,.$$

En remplaçant $S_{0L}(k)$ par son expression et en réordonnant les termes, nous obtenons

$$\begin{aligned}x_0=r_0+\frac{\sigma}{\sigma-1}\Bigg(&C_{01}M_{10}+C_{02}\Big(C_{21}M_{10}+M_{20}\Big)\\&+C_{03}\Big(C_{31}M_{10}+C_{32}M_{20}+M_{30}\Big)+\cdots\Bigg)+O\Big(\frac{q^{L+1}}{r_0+q}\Big)\,.\end{aligned}\tag{2.34}$$

Deuxième partie

Interlude

Chapitre 3

Une population avec un seul individu

Dans le modèle d'Eigen que nous avons étudié au cours des chapitres précédents, nous avons travaillé avec des proportions d'individus. Pour cela, nous avions considéré que notre population était infinie. La suite de cette thèse est consacrée à des modèles pour lesquels la population est finie. Avant de considérer une population avec un nombre arbitraire d'individus, commençons par examiner un cas plus simple : une population composée d'un seul individu.

Nous nous plaçons dans le même cadre que précédemment : les individus sont caractérisés par leur génotype qui est une suite finie de longueur ℓ composée de zéros et de uns. A chaque pas de temps, l'unique individu de la population se reproduit et donne naissance à un nouvel individu qui le remplace. Le génôme de ce nouvel individu est une copie de celui de son parent soumise aux mutations : chaque digit est changé indépendamment avec une probabilité q. Comme la population est constituée d'un unique individu, il n'y a pas de compétition entre les individus, le modèle qui suit ne comporte donc pas de sélection. Notre but ici est de comprendre au mieux la dynamique de ce modèle, par exemple en déterminant le nombre de générations nécessaires pour observer la chaîne composée uniquement de zéros pour la première fois. Nous suivons donc les génotypes de la lignée issue d'un individu. Nous fixons la probabilité de mutation $q \in (0,1)$. Nous commençons avec une chaîne X_0 dans $\{0,1\}^\ell$ qui représente le génotype du premier individu. Appelons X_n le génotype du n-ième individu dans la lignée. Au temps n, pour chaque digit de X_n, nous lançons une pièce de monnaie de paramètre q pour décider si une mutation arrive sur ce digit, dans ce cas, il est changé en son complémentaire. Toutes ces pièces de monnaie sont prises indépendantes, ainsi, la probabilité qu'aucune mutation ne se produise est $(1-q)^\ell$. Comme nous l'avons fait précédemment, nous allons partitionner l'espace des chaînes possibles en classes de Hamming : nous agrégeons dans une même classe les chaînes qui ont le même nombre de uns. La classe de Hamming i consiste en la réunion des chaînes qui ont i digits égaux à 1 et $\ell - i$ digits égaux à 0. Nous définissons alors un nouveau processus $(Y_n)_{n\geq 0}$ en posant

$$Y_n \;=\; \text{nombre de digits de } X_n \text{ égaux à } 1\,.$$

Ce processus définit une nouvelle chaîne de Markov sur l'espace d'état $\{0,\ldots,\ell\}$. Nous allons essayer de comprendre et d'exprimer les espérances de divers temps d'atteinte de cette chaîne Y_n.

Nous avons défini une marche aléatoire $(X_n)_{n\geq 0}$ sur l'hypercube $\{0,1\}^\ell$ pour laquelle la probabilité de transition entre deux états u et v ne dépend que de la distance de Hamming entre les chaînes u et v. Arrêtons-nous un instant pour introduire un modèle très classique, qui est aussi une marche sur l'hypercube : le modèle d'Ehrenfest. Nous constaterons au cours de ce chapitre de nombreuses analogies avec ce modèle.

Le modèle d'Ehrenfest. Considérons deux urnes et ℓ boules. Initialement, toutes les boules sont dans la seconde urne. A chaque pas de temps, une boule est choisie au hasard et déplacée dans l'autre urne. La question centrale du modèle est la suivante :

> En moyenne, combien de temps faut-il attendre
> pour que le système revienne à son état initial ?

En 1947, Mar Kac publie un article [Kac47] qui répond à cette question de manière très simple. Il suit l'évolution du nombre de boules dans la première urne, ce processus est une chaîne de Markov sur $\{0,\cdots,\ell\}$, Mark Kac a montré pour ce processus d'Ehrenfest que, partant de 0, le temps moyen pour atteindre à nouveau cet état initial est égal à 2^ℓ. Il a aussi montré que, partant de l'état j, le temps moyen pour retourner à cet état j est égal à $2^\ell/\binom{j}{\ell}$.

Le processus de Kac est assez différent de notre processus $(Y_n)_{n\geq 0}$. Par exemple, ses incréments possibles sont seulement $-1, 0$ ou $+1$. Dans notre chaîne, il est possible que tous les digits changent en un pas de temps. De façon assez surprenante, nous allons obtenir exactement les mêmes formules pour les temps moyens de retour dans notre processus. Définissons τ^* le temps d'atteinte de la master sequence $0\cdots 0$, c'est-à-dire,

$$\tau^* \;=\; \inf\left\{\, n\geq 1 : X_n = 0\cdots 0 \,\right\}.$$

Nous allons calculer l'espérance de cette variable aléatoire τ^*. La valeur moyenne de τ^* est particulièrement intéressante pour plusieurs raisons. Elle représente le temps nécessaire à un mutant pour découvrir la master sequence. En outre, cette quantité joue un rôle crucial pour comprendre le comportement d'équilibre du modèle de quasiespèce d'Eigen dans le contexte de populations finies [CD18] comme nous allons le voir dans les prochains chapitres. Pour $0\leq j\leq \ell$, nous définissons aussi le temps d'atteinte de la classe de Hamming j par

$$\tau_j \;=\; \inf\left\{\, n\geq 1 : Y_n = j \,\right\}.$$

Stratégie. Nous allons employer la même méthode que celle utilisée par Marc Kac pour le modèle d'Ehrenfest : nous allons calculer les quantités $E\big(\tau_j \,\big|\, Y_0 = i\big)$ pour $0\leq i,j\leq \ell$. Pour cela, nous allons calculer les fonctions génératrices des événements $\{\, Y_n = j \,\}$ et de la variable aléatoire τ_j, et nous allons les relier à travers une équation fonctionnelle bien connue que nous développerons dans la prochaine section. Nous donnerons finalement une formule générale pour les temps $E\big(\tau_j \,\big|\, Y_0 = i\big)$ pour $0\leq i,j\leq \ell$, basée sur la théorie du potentiel pour les chaînes de Markov.

1 Une identité probabiliste

Dans tous les calculs qui vont suivre, nous noterons P_i la probabilité conditionnée à l'évennement $\{Y_0 = i\}$. Choisissons $0 \leq i, j \leq \ell$. Pour $n \geq 1$, nous calculons la probabilité $P_i(Y_n = j)$ en décomposant l'évenement $\{\, Y_n = j \,\}$ selon le premier instant où le point j est atteint,

$$\begin{aligned} P_i(Y_n = j) &= \sum_{k=1}^{n} P_i(Y_n = j,\, \tau_j = k) \\ &= \sum_{k=1}^{n} P_i\big(Y_1 \neq j, \ldots, Y_{k-1} \neq j, Y_k = j, Y_n = j\big)\,. \end{aligned}$$

En conditionnant la somme, nous nous ramenons à

$$\begin{aligned} P_i(Y_n = j) = \sum_{k=1}^{n} & P_i\big(Y_1 \neq j, \ldots, Y_{k-1} \neq j, Y_k = j\big) \\ & \times P_i\big(Y_n = j \,\big|\, Y_1 \neq j, \ldots, Y_{k-1} \neq j, Y_k = j\big)\,. \end{aligned}$$

D'après la propriété de Markov, nous avons

$$P_i\big(Y_n = j \,\big|\, Y_1 \neq j, \ldots, Y_{k-1} \neq j, Y_k = j\big) = P_i\big(Y_n = j \,\big|\, Y_k = j\big)\,.$$

Notre chaîne de Markov étant homogène, nous pouvons aussi écrire que

$$P_i\big(Y_n = j \,\big|\, Y_k = j\big) = P_j\big(Y_{n-k} = j\big)\,.$$

Après avoir réintroduit les deux égalités dans la somme, nous obtenons finalement

$$P_i(Y_n = j) = \sum_{k=1}^{n} P_i\big(\tau_j = k\big) P_j\big(Y_{n-k} = j\big)\,. \tag{3.1}$$

Nous allons reconnaître dans cette identité un produit de Cauchy de deux séries, commençons par introduire les séries génératrices associées à l'évenement $\{\, Y_n = j \,\}$ et à la variable aléatoire τ_j :

$$\begin{aligned} F_{ij}(z) &= \sum_{n\geq 1} P_i(Y_n = j)\, z^n\,, \\ G_{ij}(z) &= \sum_{n\geq 1} P_i(\tau_j = n)\, z^n\,. \end{aligned}$$

Dans l'équation (3.1), nous isolons le terme correspondant à $k = n$, et nous multiplions par z^n pour obtenir

$$P_i(Y_n = j) z^n = P_i\big(\tau_j = n\big) z^n + \sum_{k=1}^{n-1} P_i\big(\tau_j = k\big) z^k\, P_j\big(Y_{n-k} = j\big) z^{n-k}\,.$$

Nous reconnaissons maintenant le produit de Cauchy des deux séries F_{ij} et G_{jj}. En sommant cette identité sur $n \geq 1$, nous avons finalement que

$$F_{ij}(z) = G_{ij}(z) + F_{ij}(z)\, G_{jj}(z)\,. \tag{3.2}$$

Cette équation fonctionnelle bien connue se trouve par exemple dans le livre d'Aldous et Fill [AF02], au lemme 2.25. Notre stratégie est maintenant la suivante. Nous allons calculer les fonctions F_{ij} pour notre processus. Nous utiliserons ensuite l'égalité ci-dessus pour obtenir l'expression des fonctions G_{ij}. Nous pourrons enfin calculer les espérances des temps d'atteinte à partir des séries G_{ij} en prenant la dérivée à gauche au point 1 :

$$E\big(\tau_j \,\big|\, Y_0 = i\big) \;=\; \sum_{n\geq 1} nP_i(\tau_j = n) \;=\; G'_{ij}(1)\,.$$

Pour calculer les séries F_{ij}, il nous faut travailler sur les probabilités $P_i(X_n = j)$ et donc étudier la chaîne X_n. Dans notre modèle, les mutations se produisent indépendamment à chaque site. Une conséquence importante de cette hypothèse structurelle est que les composantes de X_n, $(X_n(i), 1 \leq i \leq \ell)$, sont elles-mêmes des chaînes de Markov, ces chaînes Markov sont en outre indépendantes. Commençons par étudier leur dynamique. Nous pourrions procéder de la même façon que dans le chapitre 1, avec des produits tensoriels, cependant nous choisissons ici de présenter une manière différente.

2 Pour un seul nucléotide

Le processus $\big(X_n(1)\big)_{n\geq 0}$ est la chaîne de Markov d'espace d'état $\{0, 1\}$ et de matrice de transition

$$M \;=\; \begin{pmatrix} 1-q & q \\ q & 1-q \end{pmatrix}.$$

Les valeurs propres de M sont 1 et $1-2q$. Pour $n \geq 1$, nous calculons

$$M^n \;=\; \frac{1}{2}\begin{pmatrix} 1+(1-2q)^n & 1-(1-2q)^n \\ 1-(1-2q)^n & 1+(1-2q)^n \end{pmatrix}.$$

Voyons un moyen simple de comprendre l'expression de la n-ième puissance M^n. Posons $(\varepsilon_n)_{n\geq 1}$ une suite de variables de Bernoulli indépendantes de paramètre q. A chaque pas de temps, nous utilisons la variable ε_n pour décider si $X_n(1)$ doit changer ou non. Plus précisément, nous posons

$$X_n(1) \;=\; \begin{cases} X_{n-1} & \text{si} \quad \varepsilon_n = 0\,, \\ 1 - X_{n-1} & \text{si} \quad \varepsilon_n = 1\,. \end{cases}$$

L'événement $X_n(1) = X_0(1)$ est réalisé si et seulement si le nombre total de mutations qui se sont produites avant le temps n est pair, c'est-à-dire

$$P\big(X_n(1) = X_0(1)\big) \;=\; P\big(\varepsilon_1 + \cdots + \varepsilon_n \text{ est pair}\big)\,.$$

Posons maintenant

$$S_n \;=\; \varepsilon_1 + \cdots + \varepsilon_n\,.$$

Une astuce pour calculer la probabilité que S_n est pair est d'exprimer de deux manières différentes l'espérance de $(-1)^{S_n}$. Nous avons en effet que

$$\begin{aligned} E\big((-1)^{S_n}\big) \;&=\; \Big(E\big((-1)^{\varepsilon_1}\big)\Big)^n \;=\; \big(-q+1-q\big)^n \;=\; \big(1-2q\big)^n \\ &=\; P\big(S_n \text{ est pair}\big) \;-\; P\big(S_n \text{ est impair}\big)\,. \end{aligned}$$

Bien sûr, nous avons également que

$$P\big(S_n \text{ est pair}\big) \,+\, P\big(S_n \text{ est impair}\big) \,=\, 1\,,$$

Ainsi, nous obtenons que

$$P\big(X_n(1) = X_0(1)\big) \,=\, P\big(S_n \text{ est pair}\big) \,=\, \frac{1}{2}\Big(1 + (1-2q)^n\Big)\,.$$

De cette façon, nous reconnaissons l'expression des coefficients diagonaux de la matrice M^n. Définissons maintenant

$$p_n \,=\, \frac{1}{2}\Big(1 + (1-2q)^n\Big)\,. \tag{3.3}$$

A partir des calculs ci-dessus, nous affirmons la chose suivante : conditionnellement à $X_0(1) = 1$, $X_n(1)$ est une variable de Bernoulli avec paramètre p_n, c'est-à-dire,

$$P\big(X_n(1) = 1 \,\big|\, X_0(1) = 1\big) \,=\, p_n\,, \qquad P\big(X_n(1) = 0 \,\big|\, X_0(1) = 1\big) \,=\, 1 - p_n\,.$$

De manière analogue, conditionnellement à $X_0(1) = 0$, $X_n(1)$ est une variable de Bernoulli avec paramètre $1 - p_n$.

3 Plusieurs nucléotides

Nous considérons maintenant la chaîne X_n en entier. Les résultats de la section précédente nous permettent de déduire explicitement la distribution de Y_n : Y_n est le nombre de uns dans la chaîne X_n, c'est-à-dire

$$Y_n = \sum_{k=1}^{\ell} X_n(k)\,.$$

En effet, supposons que nous partions de $Y_0 = i$. Cela signifie que i digits de X_0 sont égaux à 1 et que $\ell - i$ digits sont égaux à 0. Au temps n, les i digits de X_n qui étaient initialement égaux à 1 sont répartis selon une loi de Bernoulli de paramètre p_n, les autres sont distribués selon une loi de Bernoulli de paramètre $1 - p_n$. L'évolution des nucléotides étant indépendante, ces variables de Bernoulli sont indépendantes, donc leur somme est distribuée comme la somme de deux variables aléatoires binomiales indépendantes :

$$Y_n \,\sim\, \mathrm{Bin}(i, p_n) + \,\mathrm{Bin}(\ell - i, 1 - p_n)\,.$$

Ainsi, nous pouvons écrire que

$$\begin{aligned} P_i(Y_n = j) \,&=\, \sum_{\substack{0\le k\le i\\ 0\le j-k\le \ell-i}} P\big(\mathrm{Bin}(i,p_n) = k\big) P\Big(\mathrm{Bin}(\ell - i, 1-p_n) = j - k\Big) \\ &= \sum_{\substack{0\le k\le i\\ 0\le j-k\le \ell-i}} \binom{i}{k}\binom{\ell-i}{j-k} (1-p_n)^{i+j-2k} (p_n)^{\ell-i-j+2k}\,. \end{aligned} \tag{3.4}$$

Cette formule est assez compliquée. Elle devient pourtant particulièrement simple dans les cas où un des deux états i ou j est égal à 0 ou ℓ. En effet, nous avons, pour $0 \leq i \leq \ell$,

$$\begin{aligned} P_i(Y_n = 0) &= (1-p_n)^i (p_n)^{\ell-i}\,, \\ P_i(Y_n = \ell) &= (p_n)^i (1-p_n)^{\ell-i}\,, \end{aligned} \tag{3.5}$$

et pour $0 \leq j \leq \ell$,

$$P_0(Y_n = j) = \binom{\ell}{j} (1-p_n)^j (p_n)^{\ell-j}\,, \tag{3.6}$$

$$P_\ell(Y_n = j) = \binom{\ell}{j} (p_n)^j (1-p_n)^{\ell-j}\,. \tag{3.7}$$

Pour une fois, et c'est assez surprenant, ces deux cas sont aussi les plus pertinents pour les applications génétiques, nous les traitons donc en premier. Nous allons en effet comparer ces cas extrêmes à la chaîne générale et en déduire une estimation sur le temps de découverte de la master sequence et les temps de retour aux différentes classes.

4 Le temps de découverte de la master sequence

Il est important que nous comprenions le temps nécessaire pour découvrir la master sequence en partant de n'importe quel chaîne x_0. Nous pouvons cependant encadrer ce temps comme

$$E\big(\tau^* \,\big|\, Y_0 = 0\big) \leq E\big(\tau^* \,\big|\, X_0 = x_0\big) \leq E\big(\tau^* \,\big|\, Y_0 = \ell\big)\,. \tag{3.8}$$

En effet, le temps de découverte de la master sequence est majoré par le temps qu'il faut pour atteindre la classe 0 en partant de la classe ℓ. Cela correspond à la situation où, commençant avec uniquement des uns, nous attendons jusqu'à obtenir une chaîne avec uniquement des zéros. Ce temps est minoré par le temps de retour moyen à 0, en partant de la classe 0, que nous calculerons ensuite.

Notre but est donc maintenant de calculer ces deux bornes. Nous allons mettre en œuvre la stratégie expliquée à la fin de la première section. Notre premier objectif est de calculer la fonction génératrice

$$F_{\ell 0}(z) = \sum_{n \geq 1} P_\ell(Y_n = 0)\, z^n\,.$$

D'après les formules (3.5) et (3.3), nous avons

$$P_\ell(Y_n = 0) = (1-p_n)^\ell = \left(\frac{1-(1-2q)^n}{2}\right)^\ell.$$

Nous utilisons le binôme de Newton pour développer la puissance ℓ :

$$P_\ell(Y_n = 0) = \frac{1}{2^\ell} \sum_{k=0}^{\ell} \binom{\ell}{k} (-1)^k (1-2q)^{nk}. \tag{3.9}$$

Remarquons que $P_\ell\big(Y_0 = 0\big) = 0$. Pour plus de commodité, nous commençons la somme définissant $F_{\ell 0}$ à $n = 0$ et nous reconnaissons des séries géométriques :

$$\begin{aligned} F_{\ell 0}(z) &= \sum_{n\geq 0} P_\ell(Y_n = 0)\, z^n \\ &= \sum_{n\geq 0} \frac{1}{2^\ell} \sum_{k=0}^{\ell} \binom{\ell}{k} (-1)^k (1-2q)^{nk} z^n \\ &= \frac{1}{2^\ell} \sum_{k=0}^{\ell} \binom{\ell}{k} \frac{(-1)^k}{1-(1-2q)^k z}\,. \end{aligned}$$

Il nous faut maintenant calculer la quantité $1 + F_{00}$ pour pouvoir utiliser la formule (3.2) dans le but de pouvoir exprimer la série $G_{\ell 0}$. Notre prochain objectif est donc de calculer la série génératrice

$$F_{00}(z) = \sum_{n\geq 1} P_0(Y_n = 0)\, z^n\,.$$

D'après les formules (3.6) et (3.3), nous avons, après avoir développé la puissance,

$$\begin{aligned} P_0(Y_n = 0) &= (p_n)^\ell = \Big(\frac{1+(1-2q)^n}{2}\Big)^\ell \\ &= \frac{1}{2^\ell} \sum_{k=0}^{\ell} \binom{\ell}{k} (1-2q)^{nk}\,. \end{aligned} \tag{3.10}$$

Cette fois, nous avons $P_0\big(Y_0 = 0\big) = 1$. En additionnant ce terme avec F_{00}, nous obtenons encore une jolie série géométrique

$$\begin{aligned} 1 + F_{00}(z) &= \sum_{n\geq 0} P_0(Y_n = 0)\, z^n \\ &= \sum_{n\geq 0} \frac{1}{2^\ell} \sum_{k=0}^{\ell} \binom{\ell}{k} (1-2q)^{nk} z^n \\ &= \frac{1}{2^\ell} \sum_{k=0}^{\ell} \binom{\ell}{k} \frac{1}{1-(1-2q)^k z}\,. \end{aligned}$$

Pour $0 \leq k \leq \ell$, nous introduisons les fonctions auxiliaires

$$\phi_k(z) = \binom{\ell}{k} \frac{1}{1-(1-2q)^k z}\,,$$

nous réécrivons alors les expressions de $F_{\ell 0}$ et $1 + F_{00}$ comme

$$\begin{aligned} F_{\ell 0}(z) &= \frac{1}{2^\ell} \sum_{k=0}^{\ell} (-1)^k \phi_k(z)\,, \\ 1 + F_{00}(z) &= \frac{1}{2^\ell} \sum_{k=0}^{\ell} \phi_k(z)\,. \end{aligned} \tag{3.11}$$

A partir de l'identité probabiliste (3.2), nous obtenons

$$G_{\ell 0}(z) \;=\; \frac{F_{\ell 0}(z)}{1+F_{00}(z)}\,.$$

Notre but est maintenant de calculer la dérivée à gauche de $G_{\ell 0}$ au point 1. Les fonctions ϕ_k sont régulières autour de 1, sauf la première, ϕ_0, en effet,

$$\phi_0(z) \;=\; \frac{1}{1-z}\,.$$

Pour obtenir $G'_{\ell 0}(1)$, nous effectuons un développement limité de $G_{\ell 0}$ autour de 1, comme suit :

$$\begin{aligned} G_{\ell 0}(z) \;&=\; \frac{\dfrac{1}{1-z}+\displaystyle\sum_{k=1}^{\ell}(-1)^k\phi_k(z)}{\dfrac{1}{1-z}+\displaystyle\sum_{k=1}^{\ell}\phi_k(z)} \\ &=\; 1+(1-z)\sum_{k=1}^{\ell}\big((-1)^k-1\big)\phi_k(z)+o(z-1)\,. \end{aligned}$$

Ce développement donne facilement la valeur de la dérivée à gauche de $G_{\ell 0}$ au point 1 :

$$G'_{\ell 0}(1) \;=\; \sum_{k=1}^{\ell}\big(1-(-1)^k\big)\phi_k(1)\,.$$

En remplaçant $\phi_k(1)$ par sa valeur, nous obtenons la formule suivante.

Théorème 9. *Le temps moyen pour atteindre la master sequence en partant d'une chaîne composée uniquement de* 1 *est*

$$E\big(\tau^* \,\big|\, Y_0=\ell\big) \;=\; \sum_{k=1}^{\ell}\binom{\ell}{k}\frac{1-(-1)^k}{1-(1-2q)^k}\,.$$

Nous procédons de la même manière pour calculer l'espérance du temps de retour à la master sequence. En fait, nous devons calculer $G'_{00}(1)$, l'identité probabiliste (3.2) donne

$$G_{00}(z) \;=\; \frac{F_{00}(z)}{1+F_{00}(z)} \;=\; 1-\frac{1}{1+F_{00}(z)}\,.$$

Nous avons déjà calculé $1+F_{00}(z)$ dans la formule (3.11). Nous utilisons cette expression et nous développons autour de $z=1$:

$$G_{00}(z) \;=\; 1-\frac{2^\ell}{\displaystyle\sum_{k=0}^{\ell}\phi_k(z)} \;=\; 1-2^\ell(1-z)+o(1-z)\,.$$

Ce développement montre que $G'_{00}(1) = 2^\ell$. Nous en déduisons que l'espérance du temps de retour à la classe 0 est donné par

$$E\big(\tau^* \,\big|\, Y_0 = 0\big) \;=\; 2^\ell\,. \tag{3.12}$$

Ce résultat peut aussi être vu comme une conséquence du fait que la mesure invariante d'une chaîne de Markov s'exprime facilement en fonction des espérances de ses temps de retour. Ici, la mesure invariante du processus est la mesure uniforme sur $\{0,1\}^\ell$. C'est aussi le cas pour le résultat de la section suivante : Sur le processus agrégé où nous considérons les classes de Hamming, la mesure invariante est une loi binomiale. L'espérance du temps de retour à la classe j est donc l'inverse de la masse de la classe j : $\binom{\ell}{j}/2^\ell$. Nous ne détaillerons pas ces affirmations, et nous allons obtenir ce résultat avec la méthode de Kac. Il ne semble pas que les autres formules que nous avons obtenues soient aussi des conséquences directes de ce fait.

Nous pouvons maintenant revenir à notre encadrement, nous obtenons, pour toute chaîne de départ x_0,

$$2^\ell \;\le\; E\big(\tau^* \,\big|\, X_0 = x_0\big) \;\le\; \sum_{k=1}^{\ell} \binom{\ell}{k} \frac{1-(-1)^k}{1-(1-2q)^k} \;\le\; \frac{2^\ell}{q}\,.$$

Ces inégalités montrent que le temps de découverte de la master sequence est d'ordre 2^ℓ. Ce qui signifie qu'en moyenne, il faut attendre autant de générations que le nombre total de chaînes pour retrouver la master sequence.

5 Le temps de retour à la classe j

Nous allons pouvoir retrouver l'analogue du théorème de Kac sur le temps d'atteinte moyen de la classe j lorsque le processus commence à la classe j. Nous écrivons la formule (3.4) avec $i = j$, nous réindexons la somme en posant $l = j - k$ et nous développons les deux puissances comme suit :

$$\begin{aligned}
P_j(Y_n = j) \;&=\; \sum_{\substack{0\le k\le j\\ 0\le j-k\le \ell-j}} \binom{j}{k}\binom{\ell-j}{j-k}\Big(\frac{1-(1-2q)^n}{2}\Big)^{2j-2k}\Big(\frac{1+(1-2q)^n}{2}\Big)^{\ell-2j+2k}\\
&=\; \sum_{l=0}^{j\wedge(\ell-j)} \binom{j}{l}\binom{\ell-j}{l}\Big(\frac{1-(1-2q)^n}{2}\Big)^{2l}\Big(\frac{1+(1-2q)^n}{2}\Big)^{\ell-2l}\\
&=\; \sum_{l=0}^{j\wedge(\ell-j)} \frac{1}{2^\ell}\binom{j}{l}\binom{\ell-j}{l}\sum_{\alpha=0}^{2l}\sum_{\beta=0}^{\ell-2l}\binom{2l}{\alpha}\binom{\ell-2l}{\beta}(-1)^\alpha(1-2q)^{(\alpha+\beta)n}\,.
\end{aligned}$$

Pour $n = 0$, nous avons $P_j(Y_0 = j) = 1$, ainsi, après avoir calculé la série géométrique, nous obtenons

$$\begin{aligned}
1 + F_{jj}(z) \;&=\; \sum_{n\ge 0} P_j(Y_n = j)\, z^n\\
&=\; \sum_{l=0}^{j\wedge(\ell-j)} \frac{1}{2^\ell}\binom{j}{l}\binom{\ell-j}{l}\sum_{\alpha=0}^{2l}\sum_{\beta=0}^{\ell-2l}\binom{2l}{\alpha}\binom{\ell-2l}{\beta}\frac{(-1)^\alpha}{1-(1-2q)^{\alpha+\beta}z}\,.
\end{aligned}$$

Un développement de cette fonction autour du point $z = 1$ nous donne

$$\begin{aligned} 1 + F_{jj}(z) &= \sum_{l=0}^{j\wedge(\ell-j)} \frac{1}{2^\ell} \binom{j}{l} \binom{\ell-j}{l} \frac{1}{1-z} + O(1) \\ &= \frac{1}{2^\ell} \binom{\ell}{j} \frac{1}{1-z} + O(1)\,, \end{aligned}$$

où nous avons utilisé l'identité combinatoire énoncée dans le lemme suivant

Lemme 10. *Pour $0 \le j \le \ell$, nous avons*

$$\sum_{l=0}^{j\wedge(\ell-j)} \binom{j}{l} \binom{\ell-j}{l} = \binom{\ell}{j}\,.$$

Démonstration. Fixons j dans $\{0, \cdots, \ell\}$, et considérons un ensemble E ayant cardinalité ℓ. Nous fixons également un sous-ensemble A de E ayant j éléments. Nous classons les sous-ensembles de E ayant la cardinalité j selon la cardinalité de leur intersection avec A et nous obtenons facilement la formule du lemme. □

A partir de l'identité probabiliste (3.2), nous obtenons

$$G_{jj}(z) = 1 - \frac{1}{1 + F_{jj}(z)}\,,$$

ainsi, $G_{jj}(z)$ admet le développement suivant autour de $z = 1$:

$$G_{jj}(z) = 1 + \frac{2^\ell}{\binom{\ell}{j}}(z-1) + o(z-1)\,.$$

A partir de ce développement, nous obtenons que

$$E\big(\tau_j \,\big|\, Y_0 = j\big) = G'_{jj}(1) = \frac{2^\ell}{\binom{\ell}{j}}\,.$$

Nous obtenons donc encore exactement les mêmes formules que Kac pour le modèle d'Ehrenfest : Pour $1 \le j \le \ell$, le temps de retour moyen à la classe j est

$$E\big(\tau_j \,\big|\, Y_0 = j\big) = \frac{2^\ell}{\binom{\ell}{j}}\,.$$

Les calculs suivants pourraient également être effectués à l'aide des polynômes orthogonaux dit de Krawtchouk [DG12], de la même manière que cela est fait dans [Hes54] pour le modèle d'Ehrenfest. Cependant, cela ne semble pas simplifier les calculs ni donner de meilleurs résultats de manière significative.

6 Un soupçon de théorie du potentiel

Il a été observé il y a longtemps que les équations définissant la mesure invariante, ou le temps de retour à un ensemble pour une marche aléatoire, ou plus généralement une chaîne de Markov, sont formellement équivalentes aux équations de la théorie du potentiel, si nous interprétons les probabilités de transition comme des conductances (voir le très beau livre [DS84]).

Nous nous attaquons ici au cas général, c'est-à-dire que nous cherchons une formule pour le temps d'atteinte de la classe j lorsque le processus part de la classe i. Commençons par un cas particulier, en partant de la formule du théorème 9, que nous appliquons dans le cas $i = \ell$ et $j = 0$. En développant le dénominateur comme une série géométrique, nous obtenons

$$E\big(\tau^* \,\big|\, Y_0 = \ell\big) \;=\; \sum_{k=1}^{\ell} \sum_{n\geq 0} \binom{\ell}{k} \Big((1-2q)^{nk} - (-1)^k (1-2q)^{nk} \Big) .$$

Nous échangeons l'ordre de sommation et en utilisant les formules (3.9) et (3.10), nous obtenons

$$E\big(\tau^* \,\big|\, Y_0 = \ell\big) \;=\; 2^\ell \sum_{n\geq 0} \Big(P_0\big(Y_n = 0\big) - P_\ell\big(Y_n = 0\big) \Big) .$$

A partir de la formule (3.12), nous avons également que $E\big(\tau^* \,\big|\, Y_0 = 0\big) = 2^\ell$. La formule ci-dessus est en fait un cas particulier d'une formule plus générale valable pour une large classe de chaînes de Markov, que nous nous proposons d'implémenter pour notre processus. Désignons par P la matrice de transition du processus :

$$\forall i, j \in \{\, 0, \ldots, \ell \,\} \quad \forall n \geq 0 \qquad P(i,j) \;=\; P(Y_{n+1} = j \,|\, Y_n = i) .$$

Les arguments présentés ci-dessous sont en fait valables pour une classe générale de chaînes de Markov avec un espace d'état fini. Par exemple, il suffit que P, ou l'une de ses puissances, ait toutes ses entrées positives. Dans cette situation, le théorème ergodique classique pour les chaînes de Markov assure l'existence et l'unicité d'une mesure de probabilité invariante et nous avons la convergence suivante :

$$\forall i, j \in \{\, 0, \ldots, \ell \,\} \qquad \lim_{n\to\infty} P^n(i,j) \;=\; \frac{1}{E\big(\tau_j \,\big|\, Y_0 = j\big)} . \tag{3.13}$$

A partir de maintenant, fixons j dans $\{\, 0, \ldots, \ell \,\}$ et calculons l'espérance $E\big(\tau_j \,\big|\, Y_0 = i\big)$. L'idée est d'étudier le comportement de la chaîne de Markov jusqu'au temps τ_j. Pour ce faire, nous introduisons la matrice compagnon G définie par

$$\forall i, k \in \{\, 0, \ldots, \ell \,\} \qquad G(i,k) \;=\; E_i\Big(\sum_{n=0}^{\tau_j - 1} 1_{\{Y_n = k\}} \Big) .$$

La matrice G est appelée matrice potentielle associée à la restriction de P à l'ensemble $\{\, 0, \ldots, \ell \,\} \setminus \{j\}$. La quantité $G(i,k)$ représente le nombre moyen de visites

en k avant d'atteindre l'état j en partant de i. Nous introduisons aussi la matrice H donnée par

$$\forall i,k \in \{\,0,\dots,\ell\,\} \qquad H(i,k) \,=\, \begin{cases} 1 & \text{if} \quad k=j \\ 0 & \text{if} \quad k\neq j \end{cases} .$$

La matrice H décrit la loi du point de sortie de l'ensemble $\{\,0,\dots,\ell\,\}\setminus\{\,j\,\}$. Dans notre cas, il s'agit nécessairement de la mesure de Dirac sur j, mais dans le cas général, la matrice H est plus compliquée. Les trois matrices P, G, H sont reliées à travers une idendité assez simple.

Lemme 11. *En notant I la matrice identité, nous avons*

$$GP = H + G - I\,.$$

Démonstration. La matrice G encode le comportement du processus jusqu'à ce qu'il atteigne j. En multipliant à droite la matrice G par la matrice de transition P, nous faisons faire une étape supplémentaire au processus. Ce pas supplémentaire peut soit rester à l'intérieur de $\{\,0,\dots,\ell\,\}\setminus\{\,j\,\}$, auquel cas nous retrouvons la matrice $G-I$, ou alors il peut arriver au point j, c'est dans ce cas que la matrice H entre en jeu. Rendons cet argument rigoureux : nous devons vérifier que

$$\forall i,k \in \{\,0,\dots,\ell\,\} \qquad GP(i,k) \,=\, H(i,k) + G(i,k) - I(i,k)\,.$$

Pour $i,k \in \{\,0,\dots,\ell\,\}$, nous calculons

$$\begin{aligned} GP(i,k) \,&=\, \sum_{0\leq l\leq \ell} G(i,l)\,P(l,k) \\ &= \sum_{0\leq l\leq \ell} E_i\Big(\sum_{n\geq 0} 1_{\{\tau_j>n\}} 1_{\{Y_n=l\}}\Big)\,P(l,k) \\ &= \sum_{0\leq l\leq \ell}\sum_{n\geq 0} P_i\Big(\tau_j>n,\,Y_n=l\Big)\,P\big(Y_{n+1}=k\,\big|\,Y_n=l\big) \\ &= \sum_{n\geq 0}\sum_{0\leq l\leq \ell} P_i\Big(\tau_j>n,\,Y_n=l,\,Y_{n+1}=k\Big) \\ &= P(i,k) + \sum_{n\geq 1} P_i\Big(\tau_j>n,\,Y_{n+1}=k\Big)\,. \end{aligned}$$

Considérons maintenant deux cas. Si $k=j$, la formule devient

$$GP(i,j) \,=\, \sum_{n\geq 0} P_i\big(\tau_j=n+1\big) \,=\, 1 \,=\, H(i,j) + G(i,j) - I(i,j)\,.$$

Si $k\neq j$, la formule devient

$$\begin{aligned} GP(i,k) \,&=\, P(i,k) + \sum_{n\geq 1} P_i\Big(\tau_j>n+1,\,Y_{n+1}=k\Big) \\ &=\, G(i,k) - I(i,k)\,. \end{aligned}$$

Ceci termine la preuve, puisque $H(i,k)=0$ dans ce cas. □

En multipliant la formule du lemme 11 par P^n et en sommant de 0 à m, nous obtenons

$$G - GP^{m+1} = \sum_{n=0}^{m} \left(P^n - HP^n\right).$$

Concentrons-nous sur les coefficients (i, j) des matrices et faisons tendre m vers l'infini :

$$\lim_{m\to\infty} \left(G(i,j) - GP^m(i,j)\right) = \sum_{n\geq 0} \left(P^n(i,j) - P^n(j,j)\right).$$

Maintenant, $G(i, j) = 0$ et à partir de la convergence (3.13), nous avons

$$\lim_{m\to\infty} GP^m(i,j) = \Big(\sum_{k=0}^{\ell} G(i,k)\Big) \times \frac{1}{E\big(\tau_j \,\big|\, Y_0 = j\big)}.$$

Remarquons que

$$\sum_{k=0}^{\ell} G(i,k) = E\big(\tau_j \,\big|\, Y_0 = i\big),$$

En rassemblant les identités ci-dessus, nous obtenons le résultat classique suivant, que l'on trouve par exemple dans le livre d'Aldous et Fill [AF02] au lemme 2.12. Pour tous entiers i, j distincts dans $\{0, \ldots, \ell\}$, nous avons

$$E\big(\tau_j \,\big|\, Y_0 = i\big) = E\big(\tau_j \,\big|\, Y_0 = j\big)\Big(\sum_{n\geq 0}\Big(P_j\big(Y_n = j\big) - P_i\big(Y_n = j\big)\Big)\Big).$$

Dans le cas spécifique de notre modèle, nous pouvons remplacer les quantités contenues dans la formule précédente pour obtenir une expression exacte des temps d'atteinte à l'aide de la formule (3.4).

Théorème 12. *Pour $1 \leq j \leq \ell$, le temps moyen d'atteinte de la classe j en partant de la classe i est donné par*

$$E\big(\tau_j \,\big|\, Y_0 = i\big) = \frac{1}{\binom{\ell}{j}} \sum_{n\geq 0}\sum_{k=0}^{\ell} \left(\binom{j}{k}\binom{\ell-j}{k} - \binom{i}{j-k}\binom{\ell-i}{k}\Big(\frac{1-(1-2q)^n}{1+(1-2q)^n}\Big)^{i-j}\right) \Big(1-(1-2q)\Big)^{2k}\Big(1+(1-2q)\Big)^{\ell-2k}. \quad (3.14)$$

Troisième partie

Le modèle de Moran

Chapitre 4

Un modèle pour une population finie

Nous introduisons le modèle de Moran pour généraliser le modèle d'Eigen à une population finie. Nous discutons de la bonne définition du seuil d'erreur avec cette nouvelle contrainte. Nous estimons le temps de survie des master sequences et nous en déduisons différents critères pour le paramètre critique.

1 Un modèle pour une population finie.

Dans les chapitres précédents, nous avons discuté de modèles disposant d'une population infinie ou dont la population est réduite à un unique individu. Pour ces deux derniers chapitres, nous allons considérer une population finie d'individu et où le nombre d'individu restera constant au cours du temps. Le nombre d'individus dans la population, noté m dans la suite, est un nouveau paramètre et nous travaillerons das le régime asymptotique avec la convergence simultanée des paramètres :

$$m \to \infty\,, \quad \ell \to \infty\,, \quad q \to 0\,. \tag{4.1}$$

Nous nous plaçons encore une fois dans le régime critique

$$\ell q \to \ln \sigma\,. \tag{4.2}$$

Nous devons supposer que r_0 est positif et que

$$\sqrt{m} r_0 >> m^{2\varepsilon}\,. \tag{4.3}$$

Nous considérons une population de m individus dont le matériel génétique est codé avec une chaîne de ℓ caractères choisis dans $\{0, \cdots, \kappa - 1\}$. Tous les κ^ℓ génotypes ont une fitness égale à 1 sauf une chaîne, disons $0, \cdots 0$, qui a une fitness égale à σ, avec $\sigma > 1$. La séquence $0 \cdots 0$ est appelée la master sequence. Le temps est discret : à chaque génération, la population est soumise à des événements de mutation et de sélection. La façon de passer d'une génération à une autre va devoir généraliser les mécanismes du modèle d'Eigen, et elle pourrait être implémentée de différentes façons. Décrivons le modèle avec lequel nous travaillerons : Le modèle de Moran. A chaque pas de temps un individu est choisi pour être un parent, ce choix n'est pas uniforme : les master sequences ont un avantage sélectif. C'est lors de ce choix

qu'intervient la sélection dans le modèle. Les master sequences ont σ fois plus de chances d'être choisies que toutes les autres chaînes. Si nous notons fit la fitness moyenne de la population, une master séquence est choisie avec probabilité $\sigma/\,fit$, un individu qui n'est pas une master sequences est choisi avec probabilité $1/\,fit$. L'individu choisi est ensuite dupliqué, mais cette copie est sujette aux mutations, chaque bit de son génôme est changé indépendamment avec probabilité q en l'une des $\kappa-1$ autres lettres choisie uniformément. La nouvelle génération est alors formée à partir de ce nouvel individu et de tous les individus de la génération précédente sauf un, choisi uniformément au hasard. De cette façon, la taille de la population reste constante au cours du temps, égale à m. Comme nous l'avons expliqué dans l'introduction, nous suivons l'évolution du nombre N_t de master sequences au cours du temps.

2 Le modèle de vie et de mort

Nous suivons la stratégie de [NS89] pour simplifier le processus initial, à savoir que nous séparons les chaînes en deux classes. La première classe T_0 regroupe toutes les master sequences, toutes les autres séquences sont mises dans la seconde classe T_1, que nous avions appelé le nuage de mutants. Ecrivons P_{ij} pour la probabilité pour un individu de type T_i de donner naissance à un individu de type T_j, pour $i,j \in \{0,1\}$. Certaines de ces probabilités peuvent être calculées immédiatement, par exemple, P_{00} est la probabilité qu'une master sequence donne naissance à une master sequence, pas un seul bit ne doit être changé, donc $P_{00} = M_{00} = (1-q)^\ell$, et bien sûr, $P_{01} = 1-M_{00}$. Cependant, la probabilité pour un individu de type T_1 de donner naissance à une master sequence dépend du nombre de bits de son génome qui sont différents de 0. Cette probabilité est donc hors de portée si l'on ne fait aucune hypothèse sur la répartition de la population dans les différentes classes de Hamming. C'est pour cette quantité que Nowak et Schuster ont supposé une distribution uniforme de tous les génotypes, en prenant

$$P_{10} = \sum_{k=1}^{\ell} \frac{\binom{\ell}{k}}{\kappa^\ell - 1} q^k (1-q)^{\ell-k} = \frac{1-(1-q)^\ell}{\kappa^\ell - 4}\,.$$

En fait, les individus qui ne sont pas des master sequences sont génétiquement proches des master sequences. Selon Eigen, les mutations de T_1 à T_0 jouent un rôle mineur dans la phase de quasiespèce, mais elles contribuent quand même à sa stabilité. Dans cette phase, l'hypothèse de Nowak et Schuster sur P_{10} est assez forte car elle sous-estime largement la probabilité qu'un individu qui n'est pas une master sequence donne naissance à une master sequence. Ils rendent cette probabilité d'ordre $1/\kappa^\ell$, les développements des chapitres précédents, elle semblerait être d'ordre $r_0 q$. Dans ce texte, nous ne ferons aucune hypothèse, nous allons majorer puis minorer la probabilité P_{10} de la même manière que nous l'avons fait dans le chapitre 2. Cela nous conduira à définir deux processus qui conduiront à la même estimation pour les premiers termes de l'espérance du temps de persistance.

• Comme nous étudions un régime où il y a peu de mutations ($q \to 0$), il est plus facile d'obtenir des master sequences si les individus de type T_1 n'ont besoin

que d'une seule mutation pour devenir une master sequence. Dans ce régime, le nombre de master sequences sera plus important que dans le processus initial et nous obtiendrons ainsi un temps de persistance plus long. Nous majorons donc

$$P_{10} \leq M_{10} \leq q\,.$$

• Au contraire, si les individus de type T_1 ne peuvent jamais devenir des master sequences, les master sequences s'éteindront plus rapidement : $P_{10} > 0$. Selon les hypothèses précédentes, dans les processus simplifiés, le nombre de master sequences N_t peut changer en une étape d'au plus une unité. On dit que N_t évolue selon un processus de naissance et de mort sur l'espace d'état $\{0, \cdots, m\}$. Comme dans les chapitres précédents, la fitness moyenne de la population va jouer un rôle particulier, si x est la proportion de master sequences, nous notons cette fitness

$$\mathrm{fit}(x) = 1 + (\sigma - 1)x\,.$$

Pour k entre 0 et $m-1$, nous désignons par δ_k la probabilité que N_t saute de k à $k+1$:

$$\forall t \in \mathbb{N} \quad \forall k \in \{0, \cdots, m-1\} \quad \delta_k = P\Big(N_{t+1} = k+1 \,\Big|\, \mathcal{F}_t\Big)\,.$$

Pour que la quantité N_t augmente de 1, il faut que le nouvel individu soit une master sequence et remplace un individu de classe T_1.

$$\delta_k = P\left(\begin{array}{c}\text{Remplacer un individu}\\ \text{non master sequence}\end{array}\right) P\left(\begin{array}{c}\text{Le nouvel individu est une}\\ \text{master sequence}\end{array}\right).$$

Nous conditionnons alors selon la classe du parent du nouvel individu :

$$\delta_k = \Big(1 - \frac{k}{m}\Big)\left[P\left(\begin{array}{c}\text{Le nouvel individu est une}\\ \text{master sequence}\end{array}\ \text{et}\ \begin{array}{c}\text{son parent est une}\\ \text{master sequence}\end{array}\right)\right.$$
$$\left. + P\left(\begin{array}{c}\text{Le nouvel individu est une}\\ \text{master sequence}\end{array}\ \text{et}\ \begin{array}{c}\text{son parent n'est pas}\\ \text{une master sequence}\end{array}\right)\right].$$

La première probabilité ne dépend que du nombre de master sequences et nous pourrons donc la calculer :

$$P\left(\begin{array}{c}\text{Le nouvel individu est une}\\ \text{master sequence}\end{array}\ \text{et}\ \begin{array}{c}\text{son parent est une}\\ \text{master sequence}\end{array}\right) = \frac{\sigma}{\mathrm{fit}\left(\frac{k}{m}\right)}\frac{k}{m}M_{00}\,.$$

La seconde probabilité dépend en revanche du nombre d'individus dans chaque classe. Afin d'obtenir deux processus qui ne dépendent pas de la répartition de toute la population, nous encadrons la dernière probabilité. Pour obtenir une borne inférieure, il nous suffit de minorer cette probabilité par 0, ce qui revient à négliger les mutations de retour des classes 1 et supérieures. Pour la borne supérieure, nous

majorons cette probabilité : nous faisons comme si les individus qui ne sont pas des master sequences sont tous de classe 1.

$$P\Bigg(\begin{array}{c}\text{Le nouvel individu est une}\\ \text{master sequence}\end{array}\text{ et }\begin{array}{c}\text{son parent n'est pas}\\ \text{une master sequence}\end{array}\Bigg)\leq$$
$$\frac{1-\frac{k}{m}}{\text{fit}\left(\frac{k}{m}\right)}P\Bigg(\begin{array}{c}\text{Le nouvel individu issu d'un parent de classe 1}\\ \text{est une master sequence}\end{array}\Bigg)\leq\frac{M_{10}}{\text{fit}\left(\frac{k}{m}\right)},$$

où nous avons majoré $1-\frac{k}{m}$ par 1. Notre encadrement est satisfaisant puisque la différence entre les deux bornes tend vers 0 comme q. Nous construirons donc la borne inférieure en remplaçant cette probabilité par 0 et la borne supérieure en la remplaçant par M_{10}. Pour k entre 1 et m, nous notons γ_k la probabilité que N_t saute de k à $k-1$:

$$\forall t\in\mathbb{N}\quad\forall k\in\{1,\cdots,m\}\quad\gamma_k=P\Big(N_{t+1}=k-1\,\Big|\,\mathcal{F}_t\Big),$$

Pour que le nombre N_t diminue de 1, le nouvel individu ne doit pas être une master sequence et remplacer une master sequence, ce qui se produit avec probabilité

$$\gamma_k=\frac{k}{m}\Bigg[P\Bigg(\begin{array}{c}\text{Le nouvel individu n'est pas}\\ \text{une master sequence}\end{array}\text{ et }\begin{array}{c}\text{son parent est}\\ \text{une master sequence}\end{array}\Bigg)$$
$$+P\Bigg(\begin{array}{c}\text{Le nouvel individu n'est pas}\\ \text{une master sequence}\end{array}\text{ et }\begin{array}{c}\text{son parent n'est pas}\\ \text{une master sequence}\end{array}\Bigg)\Bigg].$$

Comme pour δ_k, la première probabilité ne dépend pas de la répartition de la population et vaut

$$P\Bigg(\begin{array}{c}\text{Le nouvel individu n'est pas}\\ \text{une master sequence}\end{array}\text{ et }\begin{array}{c}\text{son parent est}\\ \text{une master sequence}\end{array}\Bigg)=\frac{\sigma}{\text{fit}\left(\frac{k}{m}\right)}\frac{k}{m}(1-M_{00}).$$

Occupons-nous maintenant de la deuxième probabilité, qui elle dépend de la répartition de toute la population. Comme γ_k est la probabilité de perdre une master sequence, majorer γ_k conduira à un processus plus petit. Cette probabilité est inférieure à la probabilité de choisir un parent non master sequence, donc

$$P\Bigg(\begin{array}{c}\text{Le nouvel individu n'est pas}\\ \text{une master sequence}\end{array}\text{ et }\begin{array}{c}\text{son parent n'est pas}\\ \text{une master sequence}\end{array}\Bigg)\leq\frac{1-\frac{k}{m}}{\text{fit}\left(\frac{k}{m}\right)}.$$

Pour minorer cette probabilité, le nouvel individu a moins de chance de ne pas être une master sequence si le génôme de son parent est dans la classe 1, d'où

$$P\Bigg(\begin{array}{c}\text{Le nouvel individu n'est pas}\\ \text{une master sequence}\end{array}\text{ et }\begin{array}{c}\text{son parent n'est pas}\\ \text{une master sequence}\end{array}\Bigg)\geq$$
$$\frac{1-\frac{k}{m}}{\text{fit}\left(\frac{k}{m}\right)}P\Bigg(\begin{array}{c}\text{Le nouvel individu issu d'un parent de classe 1}\\ \text{n'est pas une master sequence}\end{array}\Bigg)\geq$$
$$\frac{1-\frac{k}{m}}{\text{fit}\left(\frac{k}{m}\right)}(1-M_{10})\geq\frac{1-\frac{k}{m}-M_{10}}{\text{fit}\left(\frac{k}{m}\right)},$$

où la dernière minoration consiste à développer et minorer $M_{10}\frac{k}{m}$ par 0.

Introduisons maintenant les notations spécifiques à ce processus en reprenant le schéma général. Pour la borne supérieure, nous posons

$$\overline{\psi}(x) = M_{10} + \sigma M_{00}\, x\,,$$

et

$$\overline{\phi}(x) = 1 - M_{10} + \Big(\sigma - 1 - \sigma M_{00}\Big)\, x\,.$$

Pour la borne inférieure, nous posons

$$\underline{\psi}(x) = \sigma M_{00}\, x\,,$$

et

$$\underline{\phi}(x) = 1 + \Big(\sigma - 1 - \sigma M_{00}\Big)\, x\,.$$

Nous avons finalement encadré les probabilités du vrai processus comme

$$\Big(1 - \frac{k}{m}\Big)\frac{\underline{\psi}(k/m)}{\mathrm{fit}\big(\frac{k}{m}\big)} \;\le\; \delta_k \;\le\; \Big(1 - \frac{k}{m}\Big)\frac{\overline{\psi}(k/m)}{\mathrm{fit}\big(\frac{k}{m}\big)}\,,$$

et

$$\frac{k}{m}\frac{\underline{\phi}(k/m)}{\mathrm{fit}\big(\frac{k}{m}\big)} \;\ge\; \gamma_k \;\ge\; \frac{k}{m}\frac{\overline{\phi}(k/m)}{\mathrm{fit}\big(\frac{k}{m}\big)}\,.$$

Nous pouvons donc définir deux processus de vie et de mort qui encadrent le vrai processus.

• Le premier, qui majore le vrai processus, est un processus de vie et de mort sur $\{0, \cdots, m\}$ dont les probabilités de transition sont

$$\overline{\delta}_k = \Big(1 - \frac{k}{m}\Big)\frac{\overline{\psi}(k/m)}{\mathrm{fit}\big(\frac{k}{m}\big)}\,, \tag{4.4}$$

et

$$\overline{\gamma}_k = \frac{k}{m}\frac{\overline{\phi}(k/m)}{\mathrm{fit}\big(\frac{k}{m}\big)}\,. \tag{4.5}$$

• Le second, qui minore le vrai processus et dont les probabilités de transition sont

$$\underline{\delta}_k = \Big(1 - \frac{k}{m}\Big)\frac{\underline{\psi}(k/m)}{\mathrm{fit}\big(\frac{k}{m}\big)}\,, \tag{4.6}$$

et

$$\underline{\gamma}_k = \frac{k}{m}\frac{\underline{\phi}(k/m)}{\mathrm{fit}\big(\frac{k}{m}\big)}\,. \tag{4.7}$$

Dans la suite, nous effectuerons les calculs une seule fois pour les deux bornes : nous détaillerons la façon de faire cela un peu plus loin. Notre but est donc maintenant d'estimer la mesure invariante de ces deux processus pour en déduire un encadrement sur la mesure invariante du vrai processus.

Nous nous intéressons seulement ici au régime quasiespèce, nous supposerons qu'il existe toujours au moins une master sequence dans la population, formellement,

pour simplifier les futures formules, nous imposons $\delta_0 = 1$. Nous noterons dans la suite L_q la quantité

$$L_q = \sigma - 1 - \sigma M_{00}\,. \tag{4.8}$$

Remarquons que

$$L_q = \sigma - 2 - \big(\sigma M_{00} - 1\big)\,,$$

puisque nous nous sommes placés dans un régime où M_{00} tend vers $1/\sigma$, nous en déduisons que L_q tend vers $\sigma - 2$. De plus, comme

$$\frac{1 + \frac{1}{\sigma}}{2} < M_{00} < 1\,,$$

nous en déduisons

$$-1 < \frac{\sigma - 3}{2} < L_q < \sigma - 2\,. \tag{4.9}$$

De sorte que dans le cas où $\sigma \neq 2$, L_q tend vers $\sigma - 2$ et n'est donc pas nul si q est assez petit. Dans le cas où $\sigma = 2$, nous avons toujours $L_q < 0$. Dans les deux cas, $L_q \neq 0$.

Pour ces processus de naissance et de mort, il existe une formule explicite pour la mesure invariante du processus. Nous fixons $\pi_0 = 1$ et

$$\forall i \in \{1, \cdots, m\} \qquad \pi_i = \frac{\delta_1 \cdots \delta_i}{\gamma_1 \cdots \gamma_i}\,. \tag{4.10}$$

Afin d'en déduire un critère pour le point critique, Nowak et Schuster ont considéré la mesure invariante pour un processus analogue à temps continu, nous avons discuté de leur résultat dans la section (4.3) de l'introduction. Commençons par estimer le temps de persistance.

2.1 Une formule pour le temps de persistance

Pour ce type de processus, il existe également des formules explicites pour l'espérance du temps nécessaire pour atteindre l'état 0 à partir de 1, qui est exactement le temps de persistance τ_0. La formule est énoncée dans le lemme suivant et peut être trouvée dans les manuels classiques, par exemple [KT81].

Lemme 13. *L'espérance du temps de persistance τ_0 commencé à $N_0 = 1$ est donnée par*

$$E\big(\tau_0\big) = \sum_{i=1}^{m-1} \frac{1}{\delta_i}\,\pi_i + \frac{\pi_{m-1}}{\gamma_m}\,.$$

Le dernier terme de cette expression sera traité comme un terme de reste. Nous n'aboutirons pas à une expression exacte de cette somme. Nous allons souvent estimer des quantités et majorer uniformément le terme de reste que nous obtiendrons. Nous ne les garderons pas tout au long du calcul, souvent, il sera noté R. Pour estimer la somme, nous commencerons par travailler sur $\ln(\delta_k/\gamma_k)$, nous nous concentrerons ensuite sur $\ln \pi_i$ que nous estimerons par des sommes de Riemann et une comparaison astucieuse avec une intégrale. Enfin, nous additionnerons les quantités $\exp(\ln \pi_i)/\delta_i$, et mettrons en œuvre la méthode de Laplace pour estimer le temps de persistance.

2.2 Une expression pour $\ln \delta_k/\gamma_k$

Nous allons définir deux processus de vie et de mort qui vont encadrer notre processus. Commençons par poser des notations avec les deux fonctions affines ψ et ϕ, ces notations nous suivront jusqu'à la fin de notre dernier chapitre. Notre but avec ces fonctions est de pouvoir écrire le rapport des probabilités de transition, de la même façon pour les deux bornes comme

$$\frac{\delta_k}{\gamma_k} = \frac{1-\frac{k}{m}}{\frac{k}{m}} \frac{\psi\left(\frac{k}{m}\right)}{\phi\left(\frac{k}{m}\right)}.$$

Pour la borne inférieure, nous nous réfèrerons aux définitions (4.6) et (4.7). Pour la borne supérieure, aux définitions (4.4) et (4.5).

Pour ne pas avoir à écrire deux fois chaque calcul, nous allons introduire une notation qui nous permettra d'écrire une seule formule pour les deux bornes, représentée par la fonction f. Pour les formules qui suivent, remplacer f par la fonction identité donnera des formules pour la borne supérieure, alors que $f = 0$ donnera des formules pour la borne inférieure. Par exemple, nous allons pouvoir noter les fonctions ϕ et ψ à l'aide de cette fonction et oublier les barres horizontales :

$$\psi(x) = f(q) + \sigma M_{00} x\,, \tag{4.11}$$

et

$$\phi(x) = 1 - f(q) + L_q x\,. \tag{4.12}$$

Remarquons que, dans les deux cas, les fonctions ψ et ϕ sont strictement positives et non constantes. Avec ces notations, le rapport δ_k/γ_k peut être réécrit comme nous le voulions :

$$\frac{\delta_k}{\gamma_k} = \frac{1-\frac{k}{m}}{\frac{k}{m}} \frac{\psi\left(\frac{k}{m}\right)}{\phi\left(\frac{k}{m}\right)}.$$

Notre but est d'estimer le produit π_i de l'expression (4.10). Il est plus facile de travailler avec des sommes qu'avec des produits, c'est pourquoi nous commençons par l'écrire sous la forme

$$\ln \frac{\delta_k}{\gamma_k} = \ln\left(\frac{1-\frac{k}{m}}{\frac{k}{m}}\right) + \ln \psi\left(\frac{k}{m}\right) - \ln \phi\left(\frac{k}{m}\right). \tag{4.13}$$

Soit i un entier dans $\{1, \cdots, m-1\}$, en sommant les équations (4.13) entre 1 et i, nous obtenons

$$\ln \pi_i = \sum_{k=1}^{i} \ln \frac{\delta_k}{\gamma_k} \tag{4.14}$$

$$= \sum_{k=1}^{i} \ln\left(\frac{1-\frac{k}{m}}{\frac{k}{m}}\right) + \sum_{k=1}^{i} \ln \psi\left(\frac{k}{m}\right) - \sum_{k=1}^{i} \ln \phi\left(\frac{k}{m}\right). \tag{4.15}$$

Nous allons étudier chacun de ces trois termes séparément. La première somme conduit au coefficient binomial,

$$\sum_{k=1}^{i} \ln\left(\frac{1-\frac{k}{m}}{\frac{k}{m}}\right) = \ln\left(\frac{1}{i!}\prod_{k=1}^{i}(m-k)\right) = \ln\binom{m}{i} + \ln\left(1-\frac{i}{m}\right). \tag{4.16}$$

Nous allons pouvoir décrire le comportement de $\ln \binom{m}{i}$ grâce à une comparaison série intégrale astucieuse. Le terme $\ln(1-i/m)$ se simplifiera avec la division par δ_i dans le lemme 13. Notons τ la fonction

$$\tau(x) = -(1-x)\ln(1-x) - x\ln x - \frac{1}{2m}\ln\Big(mx(1-x)\Big)\,.$$

Cette fonction est une très bonne approximation de notre coefficient binomial, comme le montre le lemme suivant.

Lemme 14. *Pour tout $i \in \{1, \cdots, m-1\}$, nous avons*

$$\left|\ln\binom{m}{i} - m\tau\Big(\frac{i}{m}\Big)\right| \leq 2\,.$$

Démonstration. La preuve repose sur une comparaison astucieuse entre une série et une intégrale dérivée par Robbins [Rob55] qui conduit aux inégalités suivantes :

$$\forall n \geq 1 \qquad \frac{1}{12n+1} < \ln n! - n\ln n + n - \frac{1}{2}\ln(2\pi n) < \frac{1}{12n}\,.$$

En utilisant cette inégalité trois fois, avec i, m et $m-i$ à la place de n, nous obtenons l'approximation voulue pour $\ln \binom{m}{i}$ ainsi qu'une estimation de l'erreur. Pour tout $i \in \{1, \cdots, m-1\}$,

$$\frac{1}{12m+1} - \frac{1}{12i} - \frac{1}{12(m-i)} \leq \ln\binom{m}{i} - m\tau\Big(\frac{i}{m}\Big) + \frac{1}{2}\ln(2\pi) \leq \frac{1}{12m} - \frac{1}{12i+1} - \frac{1}{12(m-i)+1}\,.$$

Nous obtenons alors la borne uniforme :

$$\forall i \in \{1, \cdots, m-1\} \qquad \left|\ln\binom{m}{i} - m\tau\Big(\frac{i}{m}\Big)\right| \leq \frac{1}{6} + \frac{1}{2}\ln(2\pi) \leq 2\,.$$

□

Grâce à ce lemme, nous remplacerons le logarithme du coefficient binomial par, pour tout i entre 1 et $m-1$,

$$\ln\binom{m}{i} = -m\frac{i}{m}\ln\frac{i}{m} - m\Big(1-\frac{i}{m}\Big)\ln\Big(1-\frac{i}{m}\Big) - \frac{1}{2}\ln\left(m\frac{i}{m}\Big(1-\frac{i}{m}\Big)\right) + R\,, \quad (4.17)$$

et nous oublierons R dans les formules qui suivent. Pour les deux autres sommes de (4.14), nous reconnaissons des sommes de Riemann des fonctions $\ln\psi$ et $\ln\phi$. Les formules de Taylor nous permettront d'en avoir des estimations.

Notons Ψ une primitive de $\ln\psi$ et Φ une primitive de $\ln\phi$. Comme ψ et ϕ sont affines, nous avons que

$$\Psi(x) = \frac{\psi(x)}{\psi'}\ln\psi(x) - x\,,$$

et

$$\Phi(x) = \frac{\phi(x)}{\phi'}\ln\phi(x) - x\,.$$

Lemme 15. *Nous avons les formules*

$$\sum_{k=1}^{i} \ln \phi\Big(\frac{k}{m}\Big) = m\Phi\Big(\frac{i}{m}\Big) - m\Phi(0) + R_1\,,$$

et

$$\sum_{k=1}^{i} \ln \psi\Big(\frac{k}{m}\Big) = m\Psi\Big(\frac{i}{m}\Big) - m\Psi\Big(\frac{1}{m}\Big) + \frac{1}{2}\bigg(\ln \psi\Big(\frac{i}{m}\Big) + \ln \psi\Big(\frac{1}{m}\Big)\bigg) + R_2\,,$$

où les quantités R_1 et R_2 sont encadrées entre deux constantes.

Démonstration. Pour toute fonction h de classe C^2 sur l'intervalle $[0,1]$, nous avons la formule suivante à notre disposition. Nous laissons la preuve en appendice, à la section 6.1

$$\sum_{k=1}^{i} h\Big(\frac{k}{m}\Big) = m\int_0^{\frac{i}{m}} h(s)\,\mathrm{d}s + \frac{1}{2}\bigg(h\Big(\frac{i}{m}\Big) - h(0)\bigg) + R_1\,, \tag{4.18}$$

si la fonction f a un comportement trop extrême en 0, nous avons aussi une formule analogue qui reste éloignée de 0.

$$\sum_{k=1}^{i} h\Big(\frac{k}{m}\Big) = m\int_{\frac{1}{m}}^{\frac{i}{m}} h(s)\,\mathrm{d}s + \frac{1}{2}\bigg(h\Big(\frac{i}{m}\Big) + h\Big(\frac{1}{m}\Big)\bigg) + R_2\,, \tag{4.19}$$

où nous pouvons majorer les termes de reste uniformément en i, pour la première formule, par

$$|R_1| \leq \frac{1}{m^2}\sum_{k=1}^{m} \sup_{[\frac{k-1}{m},\frac{k}{m}]} |h''|\,, \tag{4.20}$$

et pour la deuxième, le reste est majoré par

$$|R_2| \leq \frac{1}{m^2}\sum_{k=2}^{m} \sup_{[\frac{k-1}{m},\frac{k}{m}]} |h''|\,. \tag{4.21}$$

Comme $\psi(0) = f(q)$ tend vers 0, nous chercherons à rester loin de 0 et nous appliquerons toujours la seconde formule à la fonction $\ln\psi$, la première formule sera utile pour la fonction $\ln\phi$. Nous remplaçons donc

$$\sum_{k=1}^{i} \ln \phi\Big(\frac{k}{m}\Big) = m\Phi\Big(\frac{i}{m}\Big) - m\Phi(0) + \frac{1}{2}\bigg(\ln \phi\Big(\frac{i}{m}\Big) - \ln\phi(0)\bigg) + R_\phi\,,$$

et

$$\sum_{k=1}^{i} \ln \psi\Big(\frac{k}{m}\Big) = m\Psi\Big(\frac{i}{m}\Big) - m\Psi\Big(\frac{1}{m}\Big) + \frac{1}{2}\bigg(\ln \psi\Big(\frac{i}{m}\Big) + \ln \psi\Big(\frac{1}{m}\Big)\bigg) + R_\psi\,.$$

Occupons nous maintenant des **termes de restes**.

• Pour la fonction ϕ une majoration grossière suffira. Comme la fonction ϕ est affine et positive, nous avons que

$$R_\phi \leq \frac{1}{m^2} \sum_{k=1}^{m} \sup_{[\frac{k-1}{m}, \frac{k}{m}]} \frac{\phi'^2}{\phi^2} < \frac{\phi'^2}{m\phi^2(0)} + \frac{\phi'^2}{m\phi^2(1)} \,. \tag{4.22}$$

En remplaçant ϕ par son expression, nous obtenons

$$R_\phi \leq \frac{{L_q}^2}{m\big(1 - f(q)\big)^2} + \frac{{L_q}^2}{m\big(1 + L_q\big)^2} \,.$$

D'après l'encadrement (4.9) sur L_q, ce reste tend toujours vers 0.

• Pour la fonction ψ, il nous faudra être plus subtil, comme ψ est croissante et positive, nous avons que

$$\sup_{[\frac{k-1}{m}, \frac{k}{m}]} \frac{\psi'^2}{\psi^2} = \frac{\psi'^2}{\psi^2\big(\frac{k-1}{m}\big)} \,.$$

Pour majorer le reste R_ψ, nous sortons le premier terme de la somme et nous comparons ensuite la somme restante avec une intégrale

$$R_\psi \leq \frac{1}{m^2} \sum_{k=2}^{m} \sup_{[\frac{k-1}{m}, \frac{k}{m}]} \frac{\psi'^2}{\psi^2} \leq \frac{1}{m^2} \frac{\psi'^2}{\psi^2(\frac{1}{m})} + \frac{1}{m} \int_{1/m}^{1} \frac{\psi'^2}{\psi^2(s)} \,\mathrm{d}s \,. \tag{4.23}$$

Nous pouvons calculer l'intégrale car nous connaissons une primitive de la fonction intégrée :

$$\int_{1/m}^{1} \frac{\psi'^2}{\psi^2(s)} \,\mathrm{d}s = -\frac{\psi'}{\psi(1)} + \frac{\psi'}{\psi(1/m)} \,.$$

Ainsi, après avoir majoré le premier terme par 0, nous obtenons une majoration du reste R_ψ :

$$R_\psi \leq \left(\frac{\psi'}{m\psi\big(\frac{1}{m}\big)}\right)^2 + \frac{\psi'}{m\psi\big(\frac{1}{m}\big)} \,. \tag{4.24}$$

Or,

$$m\psi\Big(\frac{1}{m}\Big) = mf(q) + \sigma M_{00} > \frac{1}{2} \,,$$

le reste R_ψ est donc borné par une constante. □

3 Le temps de persistance

Nous pouvons maintenant nous occuper du temps de persistance, rappelons-nous que le lemme 13 donne

$$E\big(\tau_0\big) = \sum_{i=1}^{m-1} \frac{\mathrm{fit}(i/m)}{\big(1 - \frac{i}{m}\big)\psi(i/m)} \exp\big(\ln \pi_i\big) + \frac{\pi_{m-1}}{\gamma_m} \,. \tag{4.25}$$

Notre but est donc maintenant de rassembler la formule (4.14), et les lemmes 14 et 15. Cette façon de regrouper les termes peut sembler arbitraire, mais elle permet d'obtenir une première approximation avec les bornes que nous nous sommes

fixées. Nous verrons une autre façon de rassembler les termes au chapitre suivant. Commençons par rassembler les termes constants :

$$K = \exp\left(m\Phi(0) - m\Psi\Big(\frac{1}{m}\Big) + \frac{1}{2}\ln\psi\Big(\frac{1}{m}\Big) + \frac{mf(q)}{L_q}\right), \tag{4.26}$$

nous avons introduit le terme $mf(q)/L_q$, nous le retrancherons à la fonction G définie ci-dessous. Ensuite, nous rassemblons les termes principaux, ceux qui dicteront le comportement de la somme, ces termes seront facteurs de m dans la suite

$$F(x) = -(1-x)\ln(1-x) + x\ln\big(\sigma M_{00}\big) - \Big(\frac{1}{L_q} + x\Big)\ln\big(1 + L_q x\big), \tag{4.27}$$

Le premier terme provient de la fonction τ, nous avons introduit les deux suivants et nous les retrancherons eux aussi à la fonction G. La fonction G contient finalement tous ceux que nous n'avons pas encore considérés, ainsi que les compensations pour les termes que nous avons introduits.

$$\begin{aligned} G(x) = m\Big(\Psi(x) - x\ln\big(\sigma M_{00}\big)\Big) - mx\ln x - \frac{1}{2}\ln\Big(mx(1-x)\Big) \\ - \frac{1}{2}\ln\psi(x) - m\Big(\Phi(x) - \Big(\frac{1}{L_q} + x\Big)\ln\big(1 + L_q x\big) + \frac{f(q)}{L_q}\Big). \end{aligned} \tag{4.28}$$

Avec ces notations, la formule (4.25) devient

$$E\big(\tau_0\big) = K\sum_{i=1}^{m-1}\exp\left(mF\Big(\frac{i}{m}\Big) + G\Big(\frac{i}{m}\Big) + R\right) + T, \tag{4.29}$$

avec R un terme de reste, borné uniformément par une constante et T est le terme de reste

$$T = \frac{\pi_{m-1}}{\gamma_m}.$$

La fonction F dans la somme (4.29) est multipliée par m, et m tend vers l'infini. Si nous arrivons à encadrer uniformément la fonction G sur l'intervalle $[\frac{1}{m}, 1 - \frac{1}{m}]$, nous pourrons en déduire que les indices qui compteront le plus dans la somme seront ceux qui sont proches du maximum de la fonction F.

3.1 Majoration uniforme de la fonction G

Le but de cette section est de prouver le lemme suivant qui borne uniformément la fonction G. Nous utiliserons cette borne supérieure plusieurs fois dans la suite.

Lemme 16. *Dans le régime asymptotique (4.1), avec les conditions (4.2), nous avons*

$$\sup_{[\frac{1}{m}, \frac{m-1}{m}]} |G(x)| \leq C\Big(1 + mf(q)\Big)\ln m + m\frac{f(q)^2}{|L_q|},$$

pour une certaine constante C.

Bien sûr, dans le cas où $\sigma \neq 2$, L_q ne tend pas vers 0 de sorte que le dernier terme est plus petit que le premier, mais dans le cas $\sigma = 2$, où L_q tend vers 0, nous ne savons pas encore lequel est le plus grand.

Démonstration. Étudions les termes qui apparaissent dans l'expression (4.28) de la fonction G. Soit x dans l'intervalle $[\frac{1}{m}, 1 - \frac{1}{m}]$. Remplacer ψ par son expression conduit à

$$\Psi(x) - x \ln\big(\sigma M_{00}\big) - x \ln x = \\ - x + \frac{m f(q)}{\sigma M_{00}} \ln\Big(f(q) + \sigma M_{00} x\Big) + mx \ln\Big(1 + \frac{f(q)}{\sigma M_{00} x}\Big)\,.$$

Bornons donc ces deux termes :
• Pour le premier, et puisque $\sigma M_{00} > 1$ et $x > 1/m$, nous avons

$$\left| \frac{f(q)}{\sigma M_{00}} \ln\Big(f(q) + \sigma M_{00} x\Big) \right| \leq f(q) \ln m\,.$$

• Pour le second,

$$\left| x \ln\left(1 + \frac{f(q)}{\sigma M_{00} x}\right) \right| \leq f(q)\,.$$

Puisque la fonction $x \mapsto x(1-x)$ est toujours plus petite que $1/4$, alors le troisième terme de la fonction G peut être contrôlé par

$$\left| -\frac{1}{2} \ln\Big(mx(1-x)\Big) \right| \leq \ln m\,.$$

Comme la fonction ψ est croissante et inférieure à 1, nous avons que

$$\Big| -\frac{1}{2} \ln \psi(x) \Big| \leq \Big| \frac{1}{2} \ln \psi\Big(\frac{1}{m}\Big) \Big| \leq \ln m\,.$$

Finalement, pour le dernier terme

$$\Phi(x) - \Big(\frac{1}{L_q} + x\Big) \ln\big(1 + L_q x\big) + \frac{f(q)}{L_q} = \\ - x - \frac{f(q)}{L_q} \ln \phi(x) + \Big(\frac{1}{L_q} + x\Big) \ln\Big(1 - \frac{f(q)}{1 + L_q x}\Big) + \frac{f(q)}{L_q}\,. \quad (4.30)$$

• Pour le premier de ces deux termes,

$$|\ln \phi| \leq |\ln \phi(0)| + |\ln \phi(1)| \leq \frac{f(q)}{2} + \Big| \ln\Big(1 - f(q) + L_q\Big) \Big|\,.$$

Pour borner ce second terme, nous considérons deux cas. Si L_q est positif, alors L_q ne tend pas vers 0, donc nous avons

$$\Big| \ln\Big(1 - f(q) + L_q\Big) \Big| \leq L_q - f(q)\,.$$

Si L_q est négatif, nous utilisons l'inégalité $-\ln(1-u) \le 2u$ valable si u est positif, elle conduit à

$$\Big|\ln\Big(1-f(q)+L_q\Big)\Big| \le 2\big(f(q)-L_q\big)\,.$$

Dans les deux cas, nous avons que

$$\Big|-\frac{f(q)}{L_q}\ln\phi(x)\Big| \le 3\frac{f(q)^2}{|L_q|}+f(q)\,.$$

• Intéressons-nous maintenant aux deuxième et troisième termes de l'expression (4.30), comme $1+L_qx$ est positif pour tout x,

$$\Big(1+L_qx\Big)\ln\Big(1-\frac{f(q)}{1+L_qx}\Big)+f(q)\le 0\,.$$

L'inégalité $-\ln(1-x)\le x+x^2$ est valable si $x\le 1/2$, elle donne

$$\Big|\Big(\frac{1}{L_q}+x\Big)\ln\Big(1-\frac{f(q)}{1+L_qx}\Big)+\frac{f(q)}{L_q}\Big| \le \frac{f(q)^2}{(\sigma+1)|L_q|}\,.$$

Ainsi,

$$\Big|\Phi(x)-\Big(\frac{1}{L_q}+x\Big)\ln\big(1+L_qx\big)+\frac{f(q)}{L_q}\Big| \le \frac{(\sigma+1)f(q)^2}{|L_q|}+f(q)\,.$$

En rassemblant les majorants précédents, nous obtenons, pour tout x dans l'intervalle $[\frac{1}{m},1-\frac{1}{m}]$,

$$\big|G(x)\big| \le C\Big(1+mf(q)\Big)\ln m + C\frac{mf(q)^2}{|L_q|}\,,$$

pour une certaine constante C. □

3.2 Le maximum de la fonction F

Nous cherchons le maximum de la fonction F sur $[0,1]$. Ce point correspond aux termes de la somme qui conduiront aux contributions majeures. Nous allons retrouver une quantité familière du chapitre 2. La fonction F définie dans l'expression (4.27) admet pour première dérivée

$$F'(x)=\ln(1-x)+\ln\Big(\sigma(1-q)^\ell\Big)-\ln\Big(1+L_qx\Big)\,,$$

et pour dérivée seconde

$$F''(x)=-\frac{1}{1-x}-\frac{L_q}{1+L_qx}\,.$$

Comme $1-x<1+L_qx$, nous avons

$$F''(x)=-\frac{1}{1-x}-\frac{L_q}{1+L_qx}\le -\frac{1+L_q}{1+L_qx}\,. \tag{4.31}$$

• Si $L_q > 0$, la majoration précédente de F'' est croissante, donc $F'' \leq F''(1) \leq -1$.
• Si $L_q > 0$, la majoration précédente de F'' est décroissante, donc

$$F'' \leq F''(0) \leq -(1 + L_q) \leq -\frac{\sigma}{2}\,.$$

Dans les deux cas, F'' est strictement inférieure à $-1/2$, la fonction F est donc concave, Son point critique vérifie l'équation suivante :

$$\ln(1-x) + \ln\big(\sigma M_{00}\big) - \ln\big(1 + L_q x\big) = 0\,. \tag{4.32}$$

Par conséquent nous obtenons que r_0 est le point critique de la fonction F.

$$r_0 = \frac{\sigma M_{00} - 1}{\sigma - 1}\,.$$

La quantité $1 + L_q r_0$ apparaîtra souvent dans la suite, l'équation (4.32) fournit l'expression suivante pour cette quantité :

$$1 + L_q r_0 = \sigma M_{00}(1 - r_0)\,. \tag{4.33}$$

Dans la suite, nous allons remplacer $F(i/m)$ dans la somme (4.29) par son développement de Taylor autour du point r_0, nous aurons donc besoin de calculer $F(r_0)$.

$$F(r_0) = -(1-r_0)\ln(1-r_0) + r_0 \ln\big(\sigma M_{00}\big) - \Big(\frac{1}{L_q} + r_0\Big)\ln\left(1 + L_q r_0\right)\,.$$

L'équation (4.33) donne

$$\Big(\frac{1}{L_q} + r_0\Big)\ln\left(1 + L_q r_0\right) = \Big(\frac{1}{L_q} + r_0\Big)\ln\Big(\sigma M_{00}(1 - r_0)\Big)\,.$$

En séparant le terme logarithmique, nous obtenons

$$F(r_0) = -\frac{1 + L_q}{L_q}\ln(1 - r_0) - \frac{1}{L_q}\ln\big(\sigma M_{00}\big)\,.$$

En remplaçant L_q par son expression (4.8) nous arrivons finalement à

$$F(r_0) = \frac{\sigma(1 - M_{00})\ln\frac{\sigma(1-M_{00})}{\sigma-1} + \ln\big(\sigma M_{00}\big)}{1 - \sigma\big(1 - M_{00}\big)}\,.$$

Ainsi, la quantité $F(r_0)$ est égale à

$$F(r_0) = \varphi\big(M_{00}\big)\,.$$

où la fonction φ est définie par

$$\varphi(x) = \frac{\sigma(1-x)\ln\frac{\sigma(1-x)}{\sigma-1} + \ln(\sigma x)}{1 - \sigma(1-x)}\,. \tag{4.34}$$

4 Implémentation de la méthode de Laplace

Introduisons maintenant une notation pour la somme (4.29) : nous posons

$$S_m = \sum_{i=1}^{m-1} \exp\left(mF\Big(\frac{i}{m}\Big) + G\Big(\frac{i}{m}\Big)\right). \tag{4.35}$$

Les contributions principales dans cette somme proviendront des termes dont les indices sont proches de mr_0, nous allons donc dans un premier temps estimer la somme tronquée autour d'un certain voisinage de ce point. Nous choisissons

$$\delta = m^{2/3}, \tag{4.36}$$

et nous posons $[i_-, i_+]$ l'intervalle sur lequel nous allons sommer, avec

$$i_- = \max\Big(\lfloor mr_0 - \delta \rfloor, 0\Big) + 1\,,$$

$$i_+ = \lfloor mr_0 + \delta \rfloor\,.$$

Puisque $i_- \geq 1$ et $r_0 < 1/2$ dans le régime asymptotique puisque r_0 tend vers 0, l'intervalle $[i_-, i_+]$ est strictement inclus dans $[1, m-1]$. Notre but est donc maintenant d'estimer la somme

$$S_m(\delta) = \sum_{i=i_-}^{i_+} \exp\left(mF\Big(\frac{i}{m}\Big) + G\Big(\frac{i}{m}\Big)\right). \tag{4.37}$$

Rappelons-nous que r_0 est le maximum de la fonction F, la somme tronquée est liée à l'expression (4.35) par les inégalités

$$S_m(\delta) \leq S_m \leq S_m(\delta) + m \exp\Big(mF(r_0) + \sup_{[\frac{1}{m}, \frac{m-1}{m}]} |G|\Big)\,.$$

Nous obtenons ainsi, selon le lemme 16,

$$S_m = S_m(\delta) + \exp\Big(mF(r_0) + O\big(\sup_{[\frac{1}{m}, \frac{m-1}{m}]} |G|\big)\Big)\,. \tag{4.38}$$

La formule de Taylor-Lagrange à l'ordre 2 pour F nous permet d'estimer l'expression (4.37) de $S_m(\delta)$:

$$S_m(\delta) = \sum_{i=i_-}^{i_+} \exp\left(m\left(F(r_0) + \Big(\frac{i}{m} - r_0\Big)^2 \frac{F''(\eta_i)}{2}\right) + G\left(\frac{i}{m}\right)\right), \tag{4.39}$$

où pour tout i tel que $i_- \leq i \leq i_+$, la quantité η_i est un nombre réel entre i_-/m et i_+/m.

Nous n'aurons pas besoin d'une expression exacte de $F''(r_0)$, mais seulement de savoir que F'' est strictement plus petit que $-1/2$, comme nous l'avons montré avec l'expression (4.31). La fonction G sera une fois de plus uniformément bornée par

le lemme 16. Avec ce développement, l'estimation de la somme $S_m(\delta)$ se réduit à l'estimation de la quantité $T_m(\delta)$, où

$$T_m(\delta) = \sum_{i=i_-}^{i_+} \exp\left(m\Big(\frac{i}{m} - r_0\Big)^2 \frac{F''(\eta_i)}{2} \right). \tag{4.40}$$

A partir de l'expression (4.39), nous avons

$$S_m(\delta) = \exp\Big(mF(r_0)\Big)\, T_m(\delta)\, \exp\Big(O\big(\sup_{[\frac{1}{m}, \frac{m-1}{m}]} |G|\big)\Big), \tag{4.41}$$

Nous n'aurons besoin que d'une approximation grossière sur $T_m(\delta)$. Comme $F''(r_0)$ est négatif, nous constatons d'abord que

$$T_m(\delta) \leq m\,.$$

Afin de minorer $T_m(\delta)$, nous avons besoin d'une borne inférieure sur $F''(\eta_i)$, nous considérons deux cas.

• Si L_q est positif, alors

$$F''(\eta_i) \geq F''(0) \geq -2\sigma\,.$$

• Si L_q est négatif, alors

$$F''(\eta_i) \geq F''\Big(\frac{i_+}{m}\Big) \geq F''\Big(\frac{1}{2}\Big) \geq -6\,.$$

Dans les deux cas, $F''(\eta_i)$ est supérieur à $-2(\sigma+3)$. Nous pouvons aussi minorer $T_m(\delta)$ par l'un de ses termes : par exemple, le terme d'indice $1 + \lfloor mr_0 \rfloor$

$$T_m(\delta) \geq \exp\left(-(\sigma+3)m\Big(\frac{1+\lfloor mr_0\rfloor}{m} - r_0\Big)^2 \right).$$

En revanche, nous avons que

$$mr_0 - 1 \leq \lfloor mr_0 \rfloor \leq mr_0\,,$$

donc

$$T_m(\delta) \geq \exp\Big(\frac{\sigma+3}{m}\Big),$$

et cette borne tend vers 1 Par conséquent, nous avons pour m assez grand

$$\frac{1}{2} \leq T_m(\delta) \leq m\,.$$

Avec la formule (4.41), nous pouvons réécrire la somme (4.37) comme

$$S_m(\delta) = \exp\Big(mF(r_0)\Big)\, R_5\, e^{O\left(\sup_{[\frac{1}{m}, \frac{m-1}{m}]} |G|\right)}, \tag{4.42}$$

où

$$\frac{1}{2} \leq R_5 \leq m\,.$$

Grâce à l'estimation (4.38), nous obtenons finalement

$$S_m = e^{mF(r_0)} e^{O\left(\sup_{[\frac{1}{m}, \frac{m-1}{m}]} |G|\right)}\,.$$

5 Retour au temps de survie

L'espérance du temps de survie est

$$E\big(\tau_0\big) = KS_m\, e^R + T\,, \tag{4.43}$$

où R est bornée par une certaine constante et T est le terme de reste

$$T = \frac{\pi_{m-1}}{\gamma_m}\,.$$

Puisque γ_m tend vers une constante finie, le reste T est au plus du même ordre que le terme principal. Nous avons donc

$$E\big(\tau_0\big) = K \exp\left(mF\left(r_0\right) + O\Big(\sup_{[\frac{1}{m},\frac{m-1}{m}]} |G|\Big)\right). \tag{4.44}$$

Nous allons maintenant développer cette expression, rappelons-nous que l'expression (4.26) de K donne

$$K = \exp\left(\Big(\frac{mf(q)}{\sigma M_{00}} + \frac{1}{2}\Big)\Big(\ln m - \ln\Big(\sigma P_{11} + mf(q)\Big)\Big) + m\frac{1-f(q)}{L_q}\ln\big(1-f(q)\big) + \frac{mf(q)}{L_q}\right).$$

Tous les termes de K sont d'ordre plus petits que $\big(1+mf(q)\big)\ln m$, nous les incluons donc dans le terme de reste et nous obtenons, selon (4.34),

$$E\big(\tau_0\big) = \exp\left(m\varphi\big(M_{00}\big) + O\Big(\big(1+mf(q)\big)\ln m + \frac{mf(q)^2}{L_q}\Big)\right). \tag{4.45}$$

avec

$$\varphi(x) = \frac{\sigma\big(1-x\big)\ln\frac{\sigma(1-x)}{\sigma-1} + \ln(\sigma x)}{1-\sigma(1-x)}\,.$$

Nous allons maintenant effectuer un développement de la fonction φ au point $1/\sigma$. Pour cela, nous avons besoin de connaître la première dérivée non nulle de cette fonction en ce point. Nous écrivons

$$\varphi(x) = \varphi\bigg(\Big(x - \frac{1}{\sigma}\Big) + \frac{1}{\sigma}\bigg)\,,$$

Nous nous servons de l'expression de φ et nous écrivons

$$\varphi(x) = \frac{\Big(\sigma - (\sigma x - 1) - 1\Big)\ln\frac{\sigma-(\sigma x-1)-1}{\sigma-1} + \ln\Big((\sigma x-1)+1\Big)}{1-\sigma(1-x)}\,.$$

En développant cette expression selon les puissances de $\sigma x - 1$, nous obtenons

$$\varphi(x) = \frac{-(\sigma x - 1) + \Big(\sigma - 1 - \frac{1}{2}(\sigma - 1)\Big)\Big(\frac{\sigma x - 1}{\sigma - 1}\Big)^2 + O\Big((\sigma x - 1)^3\Big)}{1 - \sigma(1 - x)} + \frac{\sigma x - 1 - \frac{1}{2}(\sigma x - 1)^2 + O\Big((\sigma x - 1)^3\Big)}{1 - \sigma(1 - x)}.$$

Ainsi, la fonction φ et sa première dérivée s'annulent au point $1/\sigma$ et

$$\varphi''\Big(\frac{1}{\sigma}\Big) = 2\sigma^2 \frac{\sigma - 1 - \frac{1}{2}(\sigma - 1) - \frac{1}{2}(\sigma - 1)^2}{(\sigma - 1)^2(1 - \sigma + 1)} = \sigma^2 \frac{(\sigma - 1)(2 - \sigma)}{(\sigma - 1)^2(2 - \sigma)} = \frac{\sigma^2}{\sigma - 1}.$$

Nous obtenons donc

$$m\varphi\Big((1 - q)^\ell\Big) = m\Big((1 - q)^\ell - 1/\sigma\Big)^2 \frac{\sigma^2}{2(\sigma - 1)} + O\Big(m\big(\sigma(1 - q)^\ell - 1\big)^3\Big).$$

En écrivant cette expression à l'aide de la notation r_0 et en remplaçant le premier terme avec nos estimées ci-dessus, nous obtenons

$$m\varphi\Big((1 - q)^\ell\Big) = m{r_0}^2 \frac{\sigma - 1}{2} + O(m{r_0}^3).$$

Finalement, nous obtenons le théorème suivant.

Théorème 17. *Le temps de persistance admet le développement suivant*

$$E\big(\tau_0\big) = \exp\left(\frac{\sigma - 1}{2}\, m{r_0}^2 + O\Big((1 + mq)\ln m + m{r_0}^3 + \frac{mq^2}{\sigma - 2 + r_0}\Big)\right). \tag{4.46}$$

5.1 Une condition supplémentaire

Dans la section suivante, nous verrons que le développement asymptotique du paramètre critique dépend largement du choix des paramètres ℓ et m. Cependant, notre borne inférieure et notre borne supérieure diffèrent d'un facteur mq, nos résultats seront donc beaucoup plus précis si nous ajoutons l'hypothèse

$$\frac{m}{\ell} \to 0,$$

ce que nous supposerons pour la prochaine section. Nous relacherons cette hypothèse dans le chapitre suivant. Cette condition permet aussi de déterminer la nature du rapport r_0/q, nous avons en effet, en utilisant le développement limité de $\ln(1 - q)$,

$$(\sigma - 1)r_0 = (\ln\sigma - \ell q) - \frac{\ln\sigma}{2}q + o\Big((\ln\sigma - \ell q) + q\Big).$$

Ainsi,

$$(\sigma - 1)^2 m r_0^2 = m(\ln\sigma - \ell q)^2 + o(1).$$

nous avons le théorème suivant.

Théorème 18. *Si le régime asymptotique est tel que le rapport m/ℓ tend vers 0 et sous la condition que $\sigma \neq 2$, l'espérance du temps de persistance est d'ordre*

$$E\big(\tau_0\big) = P(m) \exp\left(\frac{m(\ell q - \ln \sigma)^2}{2(\sigma - 1)}\right), \tag{4.47}$$

où $P(m)$ est un terme qui croît au plus comme un polynome en m.

6 Appendices

6.1 Les sommes de Riemann

Lemme 19. *Pour toute fonction f de classe C^2 sur l'intervalle $[0,1]$ et pour tout $i \in \{1, \cdots, m\}$, nous avons*

$$\sum_{k=1}^{i} f\Big(\frac{k}{m}\Big) = m \int_0^{\frac{i}{m}} f(s)\,\mathrm{d}s + \frac{1}{2}\bigg(f\Big(\frac{i}{m}\Big) - f(0)\bigg) + R\,,$$

où R est borné par

$$R \leq \frac{1}{m} \sup_{[0,1]} |f''|\,.$$

A cause du fait que $\psi(0)$ tend vers 0, nous aurons besoin d'une autre formule, qui reste loin de 0.

Lemme 20. *Pour toute fonction f de classe C^2 sur l'intervalle $[0,1]$ et pour tout $i \in \{1, \cdots, m\}$, nous avons*

$$\sum_{k=1}^{i} f\Big(\frac{k}{m}\Big) = m \int_{1/m}^{\frac{i}{m}} f(s)\,\mathrm{d}s + \frac{1}{2}\bigg(f\Big(\frac{i}{m}\Big) + f\Big(\frac{1}{m}\Big)\bigg) + R\,,$$

où R est borné par

$$R \leq \frac{1}{m^2} \sum_{k=2}^{m} \sup_{[\frac{k-1}{m}, \frac{k}{m}]} |f''|\,.$$

Pour toute fonction f de classe C^2 sur l'intervalle $[0,1]$ et pour tout $k \in \{1, \cdots, m\}$, la formule de Taylor-Lagrange appliquée à f entre les points $s \in [\frac{k-1}{m}, \frac{k}{m}]$ et $\frac{k}{m}$ donne

$$\exists \eta_s^k \in \Big]s, \frac{k}{m}\Big[\qquad f(s) = f\Big(\frac{k}{m}\Big) + \Big(s - \frac{k}{m}\Big) f'\Big(\frac{k}{m}\Big) + \Big(s - \frac{k}{m}\Big)^2 \frac{f''(\eta_s^k)}{2}\,.$$

Nous intégrons cette inégalité entre les points $\frac{k-1}{m}$ et $\frac{k}{m}$ et nous obtenons

$$\int_{\frac{k-1}{m}}^{\frac{k}{m}} f(s)\ \mathrm{d}s = \frac{1}{m} f\Big(\frac{k}{m}\Big) - \frac{1}{2m^2} f'\Big(\frac{k}{m}\Big) + \int_{\frac{k-1}{m}}^{\frac{k}{m}} \Big(s - \frac{k}{m}\Big)^2 \frac{f''(\eta_s^k)}{2}\ \mathrm{d}s\,. \tag{4.48}$$

Le dernier terme de cette équation sera un terme de reste, notons-le R_2 :

$$R_2(k) = \int_{\frac{k-1}{m}}^{\frac{k}{m}} \Big(s - \frac{k}{m}\Big)^2 \frac{f''(\eta_s^k)}{2}\, \mathrm{d}s\,. \tag{4.49}$$

Nous sommons l'expressio (4.48) pour k que varie entre 1 et i et nous obtenons

$$\int_0^{\frac{i}{m}} f(s)\,\mathrm{d}s = \frac{1}{m}\sum_{k=1}^{i} f\Big(\frac{k}{m}\Big) - \frac{1}{2m^2}\sum_{k=1}^{i} f'\Big(\frac{k}{m}\Big) + \sum_{k=1}^{i} R_2(k)\,. \tag{4.50}$$

De la même façon, en appliquant la formule de Taylor-Lagrange à l'ordre 1 à la fonction f', après avoir intégré et sommé, nous sommes ramenés à

$$\int_0^{\frac{i}{m}} f'(s)\,\mathrm{d}s = \frac{1}{m}\sum_{k=1}^{i} f'\Big(\frac{k}{m}\Big) + \sum_{k=1}^{i}\int_{\frac{k-1}{m}}^{\frac{k}{m}} \Big(s - \frac{k}{m}\Big) f''(\zeta_s^k)\,\mathrm{d}s\,, \tag{4.51}$$

pour des certains ζ_s^k entre s et $\frac{k}{m}$. Posons maintenant

$$R_1(k) = \int_{\frac{k-1}{m}}^{\frac{k}{m}} \Big(s - \frac{k}{m}\Big) f''(\zeta_s^k)\,\mathrm{d}s\,.$$

En combinant les deux formules (4.50) et (4.51), nous obtenons

$$\sum_{k=1}^{i} f\Big(\frac{k}{m}\Big) = m\int_0^{\frac{i}{m}} f(s)\,\mathrm{d}s + \frac{1}{2}\int_0^{\frac{i}{m}} f'(s)\,\mathrm{d}s - \frac{1}{2}\sum_{k=1}^{i} R_1(k) - m\sum_{k=1}^{i} R_2(k)\,.$$

Pour pouvoir contrôler les termes de reste, nous majorons uniformément la dérivée de la fonctio f :

$$\Big|R_1(k)\Big| \leq \frac{1}{2m^2} \sup_{[\frac{k-1}{m},\frac{k}{m}]} |f''|\,,$$

et pour $R_2(k)$ à partir de l'expression (4.49), nous avons

$$\Big|R_1(k)\Big| \leq \frac{1}{6m^3} \sup_{[\frac{k-1}{m},\frac{k}{m}]} |f''|\,.$$

Finallement, le reste total est borné par

$$\Big|\sum_{k=1}^{i} R_1(k) + m\sum_{k=}^{i} R_2(k)\Big| \leq \frac{1}{m^2}\sum_{k=1}^{m} \sup_{[\frac{k-1}{m},\frac{k}{m}]} |f''|\,.$$

Puisque f est une primitive de f', nous obtenonsle lemme 20. Pour toute fonction f de classe C^2 sur $[0,1]$ et pour tout $i \in \{1,\cdots,m\}$, nous obtenons

$$\sum_{k=1}^{i} f\Big(\frac{k}{m}\Big) = m\int_0^{\frac{i}{m}} f(s)\,\mathrm{d}s + \frac{1}{2}\Big(f\Big(\frac{i}{m}\Big) - f(0)\Big) + R\,,$$

où R est majoré par

$$R \leq \frac{1}{m} \sup_{[\frac{k-1}{m},\frac{k}{m}]} |f''|\,.$$

En majorant uniformément la fonction $|f''|$, nous obtenons le lemme.

6.2 Le polynôme de degré 3

Cette section vient justifier les affirmations de la section 4 de l'introduction. Nous reprenons les calculs de Nowak et Schuster dans le cadre de notre processus à temps discret. Nous allons étudier la fonction ζ :

$$\zeta\Big(\frac{k}{m}\Big) = \delta_{k-1} - \gamma_k\,.$$

Pour notre modèle, les probabilités ajoutent un dénominateur qui va compliquer les choses. Sous les mêmes dénominateurs, que nous n'avons pas écrits, nous avons

$$\zeta(x) = \Big((\sigma-1)x+1\Big)\Big(1-x+\frac{1}{m}\Big)\Big(f(q)+\sigma M_{00}(x-\frac{1}{m})\Big) \\ -\Big((\sigma-1)(x-\frac{1}{m})+1\Big)x\Big(1-f(q)+L_q x\Big)\,.$$

En factorisant, nous obtenons

$$\zeta(x) = \Big((\sigma-1)x+1\Big)\Big(-(\sigma-1)x^2+\Big(-1+\sigma M_{00}\Big(1+\frac{2}{m}\Big)\Big)x \\ +\Big(1+\frac{1}{m}\Big)\Big(f(q)-\frac{\sigma M_{00}}{m}\Big)\Big)+\frac{\sigma-1}{m}x\Big(1-f(q)+L_q x\Big)\,.$$

Après avoir développé, nous sommes ramenés à

$$\zeta(x) = -(\sigma-1)^2x^3+(\sigma-1)\Big(-1+(\sigma-1)r_0+\frac{\sigma}{m}+R_2\Big)x^2 \\ +\Big((\sigma-1)r_0+\frac{2}{m}+(\sigma-1)f(q)+R_1\Big)x+f(q)-\frac{1}{m}+R_0\,,$$

avec

$$R_2 = (\sigma-1)\frac{r_0}{m}\,,$$

$$R_1 = \frac{r_0}{m}-\frac{r_0+1}{m^2}\,,$$

et

$$R_0 = -\frac{r_0}{m}+\frac{1}{m}\Big(f(q)-\frac{\sigma M_{00}}{m}\Big)\,.$$

Maintenant que nous avons l'équation, nous suivons la stratégie et les notations de [Nic93]. Tout d'abord, nous calculons le centre de symétrie x_N du polynôme. Si $P = ax^3+bx^2+cx+d$, nous avons $x_N = -b/3a$, dans notre cas, nous écrivons

$$x_N = \frac{1}{3}\Big(-\frac{1}{\sigma-1}+r_0+\frac{\sigma}{m(\sigma-1)}\Big)+o\Big(r_0+\frac{1}{m}\Big)\,. \tag{4.52}$$

Nous aurons besoin de l'image de ce point : $y_N = P(x_N)$

$$y_N = P(x_N) = \frac{2}{3^3}\frac{b^3}{a^2}-\frac{bc}{3a}+d\,.$$

Calculons chaque terme indépendamment.

$$\frac{2}{3^3}\frac{b^3}{a^2} = \frac{2}{3^3(\sigma-1)}\Big(-1+3(\sigma-1)r_0+\frac{3\sigma}{m}-3(\sigma-1)^2r_0^2+o\big(r_0^2+\frac{1}{m}\big)\Big)\,,$$

et

$$\frac{bc}{3a} = \frac{r_0}{3}-\frac{\sigma-1}{3}r_0^2+\frac{2}{3m(\sigma-1)}+\frac{f(q)}{3}+o(r_0^2+\frac{1}{m}+q)\,.$$

Après avoir factorisé, nous obtenons

$$y_N = -\frac{2}{3^3(\sigma-1)}\Big(1+\frac{3(\sigma-1)}{2}r_0+3\frac{7\sigma-3}{2}\frac{1}{m}-3^2(\sigma-1)f(q) \\ -\frac{3(\sigma-1)^2}{2}{r_0}^2+o\big(r_0^2+\frac{1}{m}+q\big)\Big)\,.$$

Une autre quantité dont nous aurons besoin est δ, où

$$\delta^2 = \frac{b^2-3ac}{9a^2}\,.$$

Occupons-nous des deux termes séparément

$$\frac{b^2}{9a^2} = x_N^2 = \frac{1}{9}\Big(\frac{1}{(\sigma-1)^2}-2\frac{r_0}{\sigma-1}-\frac{2\sigma}{m(\sigma-1)^2}+r_0^2\Big)\,,$$

et

$$-\frac{3ac}{9a^2} = -\frac{c}{3a} = \frac{1}{3(\sigma-1)}\Big(r_0+\frac{2}{m(\sigma-1)}+f(q)\Big)\,.$$

Ainsi,

$$\delta^2 = \frac{1}{3^2(\sigma-1)^2}\Big(1+(\sigma-1)r_0+\frac{-2\sigma+6}{m} \\ +3(\sigma-1)f(q)+(\sigma-1)^2r_0^2+o\big(r_0^2+\frac{1}{m}+q\big)\Big)\,.$$

Finalement, δ vérifie

$$\delta = \frac{1}{3(\sigma-1)}\Big(1+\frac{\sigma-1}{2}r_0+\frac{-\sigma+3}{m} \\ +\frac{3(\sigma-1)}{2}f(q)+\frac{3(\sigma-1)^2}{8}r_0^2+o\big(r_0^2+\frac{1}{m}+q\big)\Big)\,.$$

Nous aurons aussi besoin de la quantité $h = 2a\delta^3$. A partir de ce qui précède, nous déduisons

$$h = -\frac{2}{3^3(\sigma-1)}\Big(1+\frac{3}{2}(\sigma-1)r_0+\frac{-3\sigma+9}{m} \\ +\frac{9(\sigma-1)}{2}f(q)+\frac{15(\sigma-1)^2}{8}r_0^2+o\big(r_0^2+\frac{1}{m}+q\big)\Big)\,.$$

Si $y_N > h$, l'équation admet 3 racines réelles. Définissons maintenant θ grâce à l'équation $\cos(3\theta) = -\frac{y_N}{h}$,

$$\cos(3\theta) = -1 - \frac{y_N - h}{h}\,.$$

où

$$y_N - h = -\frac{1}{m} + f(q) + \frac{(\sigma-1)}{4}r_0^2 + o\big(r_0^2 + \frac{1}{m} + q\big)\,,$$

Cette différence nous permet déjà de déduire combien le polynôme admet de racines réelles. Si $mr_0^2 > 1$, alors $y_N - h$ est positive ce qui implique que $y_N^2 < h^2$, dans ce cas, il y a trois racines distinctes. Dans l'autre cas, il n'y a qu'une seule racine. Plaçons-nous dans le premier cas et calculons les racines. Nous avons que

$$\cos(3\theta) = -1 + \frac{3^3(\sigma-1)}{2}\Big(-\frac{1}{m} + f(q) + \frac{(\sigma-1)}{4}r_0^2 + o\big(r_0^2 + \frac{1}{m} + q\big)\Big)\,.$$

Puisque $\cos(3\theta)$ est proche de -1, le paramètre θ doit être proche de π, posons alors $\theta = \frac{\pi}{3} + u$. Nous écrivons donc

$$\cos(3\theta) = -1 + \frac{3^2u^2}{2}\,,$$

ce qui signifie

$$u = \frac{\sqrt{3}(\sigma-1)}{2}r_0 - \frac{\sqrt{3}}{mr_0} + \frac{\sqrt{3}f(q)}{r_0}\,.$$

Nous avons maintenant tout ce qu'il faut pour exprimer les solutions, la première est $\alpha = x_N + 2\delta\cos\theta$. Commençons par développer

$$\cos\theta = \frac{1}{2} - \frac{\sqrt{3}}{2}u\,.$$

Le produit $2\delta\cos\theta$ est équivalent à

$$2\delta\cos\theta = \delta\Big(1 - \sqrt{3}u + O(u^2)\Big) = \frac{1}{3(\sigma-1)} - \frac{r_0}{3} + \frac{1 - mf(q)}{(\sigma-1)mr_0}\,.$$

Par conséquent,

$$\alpha = \frac{1 - mf(q)}{(\sigma-1)mr_0}\,.$$

La deuxième racine est $\beta = x_N + 2\delta\cos(\theta + \frac{2\pi}{3})$, le développement du cosinus est alors

$$\cos\Big(\theta + \frac{2\pi}{3}\Big) = \frac{1}{2} + \frac{\sqrt{3}}{2}u\,.$$

Dans ce cas, le produit $2\delta\cos(\theta + \frac{4\pi}{3})$ est équivalent à

$$2\delta\cos\theta = \delta\Big(1 + \sqrt{3}u + O(u^2)\Big) = \frac{1}{3(\sigma-1)} + \frac{2r_0}{3} - \frac{1 - mf(q)}{(\sigma-1)mr_0}\,,$$

ainsi,

$$\beta = r_0 - \frac{1 - mf(q)}{(\sigma - 1)mr_0}\,.$$

La troisième et dernière racine est $\gamma = x_N + 2\delta\cos(\theta + \frac{2\pi}{3})$, pour laquelle nous avons

$$\cos\left(\theta + \frac{2\pi}{3}\right) = -1 + \frac{u^2}{2}\,,$$

ainsi

$$\gamma = -\frac{1}{\sigma - 1}\,.$$

Chapitre 5

Processus Encadrants pour Moran

Comme nous l'avons dit dans l'introduction, ce chapitre n'aboutit pas à une nouvelle estimation du temps de disparition des master sequences. Nous avons implémenté les processus encadrants itératifs du chapitre 2 dans le modèle de Moran. Cependant, en l'état actuel, ces processus ne permettent pas d'approcher la mesure invariante du processus général. Pour cela, il reste un travail à effectuer. Il faudrait en effet montrer que la dynamique du processus global est de type lent-rapide, les master sequences représentant la composante lente. Pour cela, il faudrait probablement obtenir des estimées sur les trajectoires typiques des processus encadrants. L'avantage des processus encadrants est que les interactions entre les individus des différentes classes sont considérablement plus simples que dans le processus de Moran. Le programme consisterait donc à obtenir des estimées sur les temps d'entrée et de sortie des attracteurs pour chaque processus, qui soient suffisamment simples pour qu'il soit possible de les propager d'un processus à l'autre, et suffisamment précises pour permettre d'approcher la mesure invariante comme souhaité. Une version rudimentaire de ce programme a déjà été implémentée pour obtenir les formules exactes de la distribution de la quasi-espèce pour le modèle de Moran (voir la section 4 de [CD16b]). Clairement, le résultat souhaité ici est beaucoup plus délicat, et il n'est pas certain que ce programme aboutisse, en tout cas il demandera beaucoup de travail supplémentaire. Les mesures que nous construisons dans ce chapitre ne sont malheureusement pas reliées directement à la vraie mesure invariante du processus de Moran, mais elles permettent d'essayer cette stratégie et d'espérer que ce programme soit réalisable.

Nous allons travailler dans un régime asymptotique, dans toute la suite, lorsque nous utiliserons les notations O ou parlerons de constantes, ces quantités ne dépendront jamais de quantités du régime asymptotique ou pourront être encadrées par deux quantités strictement positives (ou strictement négatives) qui ne dépendent pas du régime asymptotique. Plus précisément, nous nous plaçons toujours dans le cas où

$$m \to \infty \quad \ell \to \infty \quad q \to 0\,.$$

Ce régime vérifie encore les conditions asymptotiques

$$\ell q \to \ln \sigma.$$

Nous notons dans toute la suite,

$$r_0 = \frac{\sigma(1-q)^\ell - 1}{\sigma - 1}\,. \tag{5.1}$$

Dans ce régime général, nous ne fixons aucune hypothèse sur la nature du rapport m/ℓ. Nous devons ajouter deux hypothèses sur cette quantité, nous supposons que

$$\sqrt{m} r_0 >> m^{2\varepsilon}\,. \tag{5.2}$$

$$\ell\sqrt{r_0} >> m^{3\varepsilon}\,. \tag{5.3}$$

La première hypothèse va nous permettre d'effectuer les méthodes de Laplace de façon symétrique, la seconde nous servira à encadrer les termes de reste lorsque nous intégrerons contre les mesures de nos processus. Les calculs que nous allons réaliser se placeront dans le cas où r_0 tend vers 0.

1 Motivations

Ce chapitre a pour but d'implémenter les processus encadrants du chapitre 2 pour le modèle de Moran. Ces processus vont nous permettre d'estimer les mesures invariantes de chaque classe de Hamming. Nous pourrons aussi en déduire une estimation du temps de disparition des master sequences. Pour implémenter cette stratégie, nous allons tout d'abord écrire un résultat général, que nous appliquerons à chaque processus et qui nous permettra d'obtenir une estimation de sa mesure invariante, les autres classes étant fixées.

2 Un résultat général

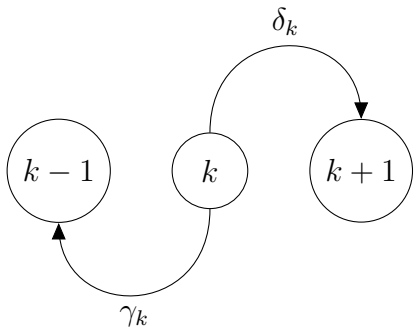
2.1 Une hypothèse sur le processus

Nous nous intéressons à une chaîne de Markov $\big(N_t\big)_{t\in\mathbb{N}}$ qui suit un processus de vie et de mort sur les entiers $\{0, \cdots, m\}$ dont nous fixerons certaines caractéristiques. Pour k entre 0 et $m-1$, nous notons δ_k la probabilité que N passe en une étape de k à $k+1$:

$$\forall t \geq 0 \quad \forall k \in \{0, \cdots, m-1\} \quad \delta_k = P\Big(N_{t+1} = k+1 \,|\, N_t = k\Big)\,. \tag{5.4}$$

Nous notons de même γ_k la probabilité que N passe en une étape de k à $k-1$:

$$\forall t \geq 0 \quad \forall k \in \{1, \cdots, m\} \quad \gamma_k = P\Big(N_{t+1} = k-1 \,|\, N_t = k\Big)\,. \tag{5.5}$$



Nous nous restreindrons à une certaine classe de processus, ceux vérifiant l'hypothèse suivante.

Hypothèse (H). Il existe deux fonctions affines ψ et ϕ positives ainsi qu'un nombre réel Δ encadré par deux constantes strictement positives tels que les probabilités de transition du processus s'écrivent

$$\delta_k = \Big(1 - \frac{k}{m}\Big)\,\psi\Big(\frac{k}{m}\Big)\Delta\,,$$

et

$$\gamma_k = \frac{k}{m}\,\phi\Big(\frac{k}{m}\Big)\Delta\,,$$

Remarquons que $(1-x)\psi(x)$ et $x\phi(x)$ sont deux fonctions polynomiales de degré 2, au point $x = 0$ la première est positive et la seconde est nulle, et au point $x = 1$ la première est nulle et la seconde est positive. Nous déduisons de cela l'existence et l'unicité du point ρ, le point d'intersection de ces deux fonctions sur l'intervalle $[0, 1]$. Les fonctions ϕ, et ψ et donc ρ sont formées par des quantités qui dépendent de q, ℓ et m. Dans le régime asymptotique, $\psi(0)$ tend vers 0 et ψ' est supérieur à $1/2$. Nous supposons enfin que la fonction ϕ est encadrée entre 2 constantes strictement positives sans être elle-même constante.

Nous imposons enfin la valeur de la somme des deux fonctions ψ et ϕ, et nous distinguerons à partir d'ici deux hypothèses qui nous conduiront à étudier deux cas distincts.

• Dans le premier cas, que nous appellerons **(MS)**, et qui nous sera utile pour l'étude du nombre de master sequences, nous aurons toujours

$$\psi' = \sigma M_{00}\,, \tag{5.6}$$

et aussi

$$\psi\Big(\frac{k}{m}\Big) + \phi\Big(\frac{k}{m}\Big) = 1 + (\sigma - 1)\frac{k}{m}\,. \tag{5.7}$$

• Dans le second cas, que nous appellerons **(Mut)**, et qui servira pour l'étude du nombre d'individus dans une certaine classe de Hamming différente de la master sequence, nous aurons toujours que ψ' est une constante strictement inférieure à 1, que $\psi(0)$ est supérieure à une constante multipliée par r_0. La somme, elle, ne dépendra pas de k :

$$\psi\Big(\frac{k}{m}\Big) + \phi\Big(\frac{k}{m}\Big) = \zeta\,, \tag{5.8}$$

où ζ est supérieur à 1, et inférieur à σ.

Nous avons maintenant assez d'hypothèses pour en déduire des résultats sur une mesure invariante du processus. Fixons quelques notations, comme les fonctions ϕ et ψ sont affines, nous écrirons

$$\phi(x) = \phi(0) + x\phi' \,,$$

sans préciser l'argument de la fonction constante ϕ', et de même

$$\psi(x) = \psi(0) + x\psi' \,.$$

Une fonction va jouer un rôle central dans la suite, il s'agit de la fonction F définie par

$$F(x) = -(1-x)\ln(1-x) - x\ln x + \frac{\psi(x)}{\psi'}\ln\psi(x) - \frac{\phi(x)}{\phi'}\ln\phi(x)\,.$$

Définissons une quantité η qui sera notre fenêtre de sommation : la taille de l'intervalle autour duquel se concentre la mesure invariante. Nous prenons, pour $\varepsilon > 0$ défini comme en (5.2),

$$\eta = \frac{m^\varepsilon}{\sqrt{m|F''(\rho)|}}\,.$$

Théorème 21. *La mesure de probabilité invariante du processus N vérifiant l'hypothèse **(H)** vérifie la formule suivante : Si k/m est entre $\rho-\eta$ et $\rho+\eta$, dans le cas **(MS)**,*

$$\nu(k) = \alpha\frac{\rho}{\sqrt{m}}\,\frac{e^{mF(\frac{k}{m})-mF(\rho)}}{k/m}\,, \tag{5.9}$$

*et dans le cas **(Mut)**,*

$$\nu(k) = \beta\sqrt{\frac{\rho}{m}}\,\frac{e^{mF(\frac{k}{m})-mF(\rho)}}{k/m}\,, \tag{5.10}$$

où α et β sont des constantes positives et bornées dans le régime asymptotique. Dans les deux cas, la masse de la mesure ν en dehors de l'intervalle $[\rho-\eta, \rho+\eta]$ tend exponentiellement vite vers 0.

$$\exists\alpha>0 \quad \exists\varepsilon>0 \quad \nu\Big([0,1]\setminus\big[\rho-\eta,\,\rho+\eta\big]\Big) \le \exp(-\alpha m^\varepsilon)\,. \tag{5.11}$$

Le reste de cette section est dédiée à une preuve de ce résultat.

2.2 Une formule pour la mesure invariante

Nous reprenons les notations du chapitre précédent et nous répétons les premières étapes du calcul dans ce cadre plus général. Pour les chaînes de vie et de mort en général, il existe une formule explicite pour exprimer la mesure invariante du processus. Posons $\pi_0 = 1$ et

$$\forall i \in \{1,\cdots,m\} \qquad \pi_i = \frac{\delta_1\cdots\delta_i}{\gamma_1\cdots\gamma_i}\,, \tag{5.12}$$

une mesure invariante du processus est alors donnée par

$$\begin{aligned} \forall i \in \{1, \cdots, m\} \qquad \mu(i) &= \delta_0 \frac{\pi_i}{\delta_i}\,, \\ \mu(0) &= \quad 1\,. \end{aligned} \tag{5.13}$$

Notre but est d'estimer cette mesure invariante, et il nous sera donc utile de travailler sur π_k. Nous écrivons π_i comme

$$\pi_i = \exp\Big(\sum_{k=1}^{i} \ln \frac{\delta_k}{\gamma_k}\Big)\,, \tag{5.14}$$

et nous nous ramènerons à l'étude du rapport δ_k/γ_k. Sous l'hypothèse **(H)**, le rapport des probabilités de transition du processus N a la forme suivante :

$$\frac{\delta_k}{\gamma_k} = \frac{1-\frac{k}{m}}{\frac{k}{m}}\ \frac{\psi\big(\frac{k}{m}\big)}{\phi\big(\frac{k}{m}\big)}\,. \tag{5.15}$$

Reprenons l'expression (5.14) de π_i, nous écrivons

$$\pi_i = \exp\left(\sum_{k=1}^{i} \ln\left(\frac{1-\frac{k}{m}}{\frac{k}{m}}\right) + \sum_{k=1}^{i} \ln\psi\Big(\frac{k}{m}\Big) - \sum_{k=1}^{i} \ln\phi\Big(\frac{k}{m}\Big)\right). \tag{5.16}$$

Comme dans le chapitre précédent, nous allons étudier chacun de ces trois termes séparément. La première somme conduit encore au coefficient binomial,

$$\sum_{k=1}^{i} \ln\left(\frac{1-\frac{k}{m}}{\frac{k}{m}}\right) = \ln\left(\frac{1}{i!}\prod_{k=1}^{i}(m-k)\right) = \ln\binom{m}{i} + \ln\Big(1-\frac{i}{m}\Big)\,. \tag{5.17}$$

Nous reprenons la fonction τ du chapitre précédent

$$\tau(x) = -(1-x)\ln(1-x) - x\ln x - \frac{1}{2m}\ln\Big(mx(1-x)\Big)\,.$$

Nous allons nous servir une nouvelle fois du lemme 14, qui permet d'approcher le coefficient binomial. Nous remplacerons donc le logarithme du coefficient binomial par, pour tout i entre 1 et $m-1$,

$$\ln\binom{m}{i} = -m\frac{i}{m}\ln\frac{i}{m} - m\Big(1-\frac{i}{m}\Big)\ln\Big(1-\frac{i}{m}\Big) - \frac{1}{2}\ln\left(m\frac{i}{m}\Big(1-\frac{i}{m}\Big)\right) + R\,, \tag{5.18}$$

et nous oublierons R dans les formules qui suivent. Avec cette simplification, la mesure s'écrit

$$\mu(i) = \frac{\psi(0)}{\psi(i/m)}\exp\left(\ln\binom{m}{i} + \sum_{k=0}^{i}\ln\psi\big(k/m\big) - \ln\phi\big(k/m\big)\right).$$

Pour les deux autres sommes de (5.16), nous allons à nouveau les exprimer à l'aide de sommes de Riemann des fonctions $\ln\psi$ et $\ln\phi$. Nous notons encore Ψ une primitive de $\ln\psi$ et Φ une primitive de $\ln\phi$. Comme ψ et ϕ sont affines, nous avons que

$$\Psi(x) = \frac{\psi(x)}{\psi'}\ln\psi(x) - x\,, \tag{5.19}$$

et

$$\Phi(x) = \frac{\phi(x)}{\phi'}\ln\phi(x) - x\,. \tag{5.20}$$

Nous avons les mêmes formules que dans le chapitre précédent au lemme 22

Lemme 22. *Nous pouvons changer les sommes en intégrale grâce aux formules suivantes :*

$$\sum_{k=1}^{i}\ln\phi\Big(\frac{k}{m}\Big) = m\Phi\Big(\frac{i}{m}\Big) - m\Phi(0) + R_1\,,$$

et

$$\sum_{k=1}^{i}\ln\psi\Big(\frac{k}{m}\Big) = m\Psi\Big(\frac{i}{m}\Big) - m\Psi\Big(\frac{1}{m}\Big) + \frac{1}{2}\bigg(\ln\psi\Big(\frac{i}{m}\Big) + \ln\psi\Big(\frac{1}{m}\Big)\bigg) + R_2\,,$$

où les quantités R_1 et R_2 sont encadrées entre deux constantes.

Démonstration. La preuve est identique au lemme analogue du chapitre précédent. Il nous suffit de vérifier que les termes de reste sont bien majorés par une constante. Comme ϕ est encadrée entre deux constantes strictement positives, le reste associé à la fonction ϕ tend toujours vers 0. Pour la fonction ψ, l'expression (4.24) donnait

$$R_\psi \leq \bigg(\frac{\psi'}{m\psi\big(\frac{1}{m}\big)}\bigg)^2 + \frac{\psi'}{m\psi\big(\frac{1}{m}\big)}\,. \tag{5.21}$$

Ici, nous minorons seulement

$$m\psi\Big(\frac{1}{m}\Big) = m\psi(0) + \psi' > \psi' > \frac{1}{2}\,,$$

le reste R_ψ est donc borné par une constante. □

Revenons maintenant à la mesure invariante et à la formule (5.16), après avoir remplacé les sommes par leur estimée et divisé le tout par δ_i, nous regroupons certains de ces termes et nous introduisons de nouvelles notations. Cette fois, nous regroupons tous les termes qui sont facteurs de m dans la fonction F :

$$F(x) = -x\ln x - (1-x)\ln(1-x) + \Psi(x) - \Phi(x)\,. \tag{5.22}$$

Nous regroupons ensuite les termes qui ne dépendent pas du point où nous calculons la mesure :

$$\ln T = -m\Psi\Big(\frac{1}{m}\Big) + m\Phi(0) + \frac{1}{2}\bigg(\ln\psi\Big(\frac{1}{m}\Big) + \ln\phi(0)\bigg) - \frac{1}{2}\ln m\,, \tag{5.23}$$

et enfin les autres termes, ceux que nous pourrons encadrer ou estimer uniformément

$$g(x) = \exp\left(-\frac{1}{2}\ln\psi(x) - \frac{1}{2}\ln\phi(x) - \frac{1}{2}\ln\Big(x(1-x)\Big)\right). \tag{5.24}$$

Un terme $-\ln\psi(x)$ provient de la division par δ_i dans la formule (5.13). A l'aide de ces notations et de la formule (5.13), nous pouvons exprimer la mesure invariante à une constante multiplicative près :

$$\forall i \in \{1, \cdots, m-1\} \qquad \mu(i) = \delta_0\, T\, g\Big(\frac{i}{m}\Big)\, e^{mF(i/m)}. \tag{5.25}$$

Ces estimations ne sont pas valables au point m, pour estimer la mesure en ce point, nous pourrons utiliser

$$\mu(m) = \frac{\delta_{m-1}}{\gamma_m}\mu(m-1) = \frac{1}{m}\frac{\psi\big(1-\frac{1}{m}\big)}{\phi(1)}\mu(m-1).$$

Dans la suite, il sera important de connaître non seulement une mesure invariante du processus N mais aussi une probabilité invariante de ce processus. Notre but est donc maintenant d'estimer la masse totale de la mesure μ.

2.3 La constante de normalisation

Lemme 23. *La masse totale de la mesure μ est donnée par*

$$\mu\big([0,1]\big) = \delta_0\, T\, g(\rho) e^{mF(\rho)} \sqrt{\frac{m}{-F''(\rho)}}.$$

Démonstration. Notons Z la masse de la mesure μ, d'après (5.25), nous avons

$$Z = \sum_{k=0}^{m} \mu(k) = 1 + \delta_0\, T \sum_{k=1}^{m-1} g\Big(\frac{k}{m}\Big) e^{mF(k/m)} + \mu(m).$$

Les termes aux extrémités seront négligeables devant la somme centrale, appelons S cette somme,

$$S = \sum_{k=1}^{m-1} g\Big(\frac{k}{m}\Big) e^{mF(k/m)}, \tag{5.26}$$

et concentrons-nous sur son estimation. Quand m devient grand, c'est la fonction F qui va dicter le comportement de la somme. Nous allons appliquer le principe de la méthode de Laplace pour montrer que les termes qui comptent le plus dans la somme sont ceux qui sont proches du maximum de la fonction F. Nous ne pouvons pas appliquer directement la méthode ici car, lorsque m grandit, la fonction F change : elle dépend elle-même de quantités liées au régime asymptotique. D'après sa définition (5.22), la dérivée de la fonction F vaut

$$F'(x) = \ln\left(\frac{(1-x)\psi(x)}{x\phi(x)}\right),$$

cette dérivée s'annule si

$$-x\Big(\psi(x)+\phi(x)\Big)+\psi(x)=0\,.$$

Cette équation admet bien une racine positive : le membre de gauche est positif en 0 et négatif en 1. Comme ce membre de gauche est un trinôme dont le coefficient dominant est négatif, cette racine est de plus un maximum de la fonction F. Nous retrouvons ainsi ρ. Nous déduisons que F croît jusqu'à la valeur ρ puis décroît. Nous obtenons aussi une relation sur le point ρ qui nous servira beaucoup par la suite,

$$(1-\rho)\psi(\rho)=\rho\phi(\rho)\,. \tag{5.27}$$

Les deux cas que nous avons distingués vont donner lieu à deux expressions différentes de ρ.

• Dans le cas **(MS)**, la relation (5.7), utilisée avec (5.27), conduit au trinôme suivant, dont ρ est la solution positive :

$$-(\sigma-1)x^2+(\psi'-1)x+\psi(0)=0\,. \tag{5.28}$$

A l'aide de la formule (5.6), nous pouvons introduire r_0,

$$-x^2+r_0\,x+\frac{\psi(0)}{\sigma-1}=0\,.$$

De cette relation, nous pouvons déduire que ρ est plus grand que r_0.

• Dans le cas **(Mut)**, il n'y a pas de termes quadratiques, d'après la relation (5.8), le maximum ρ s'exprime simplement comme

$$\rho=\frac{\psi(0)}{\zeta-\psi'}\,. \tag{5.29}$$

De cette expression et de nos hypothèses sur le cas **(Mut)**, nous déduisons que ρ est supérieur à une certaine constante multipliée par r_0.

La méthode de Laplace

Nous appliquons le schéma de la méthode de Laplace pour estimer la somme (5.26) autour du point ρ. Dans la méthode de Laplace classique, c'est-à-dire lorsque la fonction est indépendante du paramètre qui tend vers l'infini, il suffit, pour avoir un équivalent de la somme, de considérer les termes proches de ρ à une certaine distance η près. Dans la méthode classique, η peut-être choisie comme $m^{-1/2+\varepsilon}$. Ici, la fonction F dépend de q et ℓ qui varient avec m. Dans ce cas, il est possible d'appliquer la méthode avec une autre fenêtre, si par exemple la fonction F est très concentrée autour du maximum, on pourra prendre une fenêtre plus petite, c'est la valeur de la dérivée seconde de la fonction au point ρ qui va fixer l'ordre de la fenêtre. Nous allons en effet remplacer les fonctions par leur polynôme de Taylor à l'ordre 2 et nous allons obtenir

$$mF(x)\sim mF(\rho)+m\frac{(x-\rho)^2}{2}F''(\rho)\,,$$

il faudra alors que ce dernier terme tende vers l'infini. Une bonne fenêtre sera par exemple

$$\eta = \frac{m^{\varepsilon}}{\sqrt{m|F''(\rho)|}}\,, \tag{5.30}$$

pour un certain ε.

Lemme 24. *Dans le cas **(MS)**, la quantité $F''(\rho)$ vérifie les inégalités suivantes*

$$0 < C_1 < -F''(\rho) < C_2 < \infty\,.$$

*Dans le cas **(Mut)**, cette dérivée seconde est d'ordre $-1/\rho$,*

$$0 < C_1 < -\rho F''(\rho) < C_2 < \infty\,.$$

Démonstration. Commençons par calculer la dérivée seconde de la fonction F. Nous pouvons calculer une expression de F'' :

$$F''(x) = -\frac{1}{1-x} - \frac{1}{x} + \frac{\psi'}{\psi(x)} - \frac{\phi'}{\phi(x)}\,. \tag{5.31}$$

En mettant les fractions de F'' sous certains même dénominateurs, nous obtenons que

$$F''(x) = \frac{-\psi(x) + (1-x)\psi'}{(1-x)\psi(x)} + \frac{-\phi(x) - x\phi'}{x\phi(x)}\,. \tag{5.32}$$

En rassemblant les dénominateurs, nous obtenons

$$F''(x) = \frac{x\phi(x)\Big(-\psi(x) + (1-x)\psi'\Big) - (1-x)\psi(x)\Big(\phi(x) + x\phi'\Big)}{x(1-x)\psi(x)\phi(x)}\,. \tag{5.33}$$

La formule (5.27) nous donne déjà

$$F''(\rho) = \frac{-\Big(\phi(\rho) + \psi(\rho)\Big) + \psi' - \rho(\psi' + \phi')}{\rho\phi(\rho)}\,.$$

En reprenant nos deux cas, nous pouvons exprimer plus simplement $F''(\rho)$, les hypothèses (5.7) et (5.8) nous permettent d'obtenir les expressions suivantes :

- Dans le cas **(MS)**,

$$F''(\rho) = \frac{\psi' - 1 - 2(\sigma-1)\rho}{\rho\phi(\rho)}\,.$$

En utilisant (5.6), nous retrouvons r_0,

$$F''(\rho) = (\sigma-1)\frac{r_0 - 2\rho}{\rho\phi(\rho)}\,. \tag{5.34}$$

Nous pouvons obtenir une expression de cette quantité en résolvant le trinôme (5.28), cela donne

$$r_0 - 2\rho = -\sqrt{r_0^2 + 4\frac{\psi(0)}{\sigma-1}}\,. \tag{5.35}$$

En remplaçant dans $F''(\rho)$, nous obtenons donc

$$F''(\rho) = -2\frac{(\sigma-1)}{\phi(\rho)} \frac{\sqrt{r_0^2 + 4\frac{\psi(0)}{\sigma-1}}}{r_0 + \sqrt{r_0^2 + 4\frac{\psi(0)}{\sigma-1}}}\,.$$

Comme $\psi(0)$ est positif, nous avons

$$\frac{1}{2} \leq \frac{\sqrt{r_0^2 + 4\frac{\psi(0)}{\sigma-1}}}{r_0 + \sqrt{r_0^2 + 4\frac{\psi(0)}{\sigma-1}}} \leq 1\,,$$

la quantité $F''(\rho)$ est donc négative et inférieure à une certaine constante strictement négative et bornée.

- Dans le cas **(Mut)**, nous avons

$$F''(\rho) = \frac{\psi' - \zeta}{\rho\phi(\rho)}\,. \tag{5.36}$$

Comme ζ est supérieur à 1, et ψ' strictement inférieur à 1, la quantité $F''(\rho)$ est négative et d'ordre $-1/\rho$. □

Nos hypothèses nous permettent maintenant de nous assurer que η/ρ tend vers 0, en effet

$$\frac{\eta}{\rho} = \frac{m^\varepsilon}{\rho\sqrt{-F''(\rho)m}}. \tag{5.37}$$

Distinguons les cas. Dans le cas **(MS)**, nous avons montré que ρ est supérieur à r_0. De plus, nous avons choisi r_0 pour que $\sqrt{m}r_0$ soit dominant devant $m^{2\varepsilon}$. Ainsi,

$$\frac{\eta}{\rho} \leq \frac{1}{m^\varepsilon\sqrt{-F''(\rho)}}\,.$$

Comme $F''(\rho)$ est borné dans le cas **(MS)**, ce rapport tend effectivement vers 0. Dans le cas **(Mut)**, cette fois, d'après le lemme 24, $F''(\rho)$ est d'ordre $1/\rho$, nous pouvons donc écrire

$$\frac{\eta}{\rho} \leq C\frac{m^\varepsilon}{\sqrt{m\rho}}\,.$$

De plus, ρ est supérieur à une constante multipliée par r_0 donc

$$\frac{\eta}{\rho} \leq C'\frac{m^\varepsilon}{m^{1/4}(\sqrt{m}r_0)^{1/2}} \leq C'\frac{m^\varepsilon}{m^{1/4}m^\varepsilon}\,.$$

et ce rapport tend effectivement vers 0.

Nous considèrerons donc la somme S entre $m(\rho-\eta)$ et $m(\rho+\eta)$. Reprenons la somme (5.26), nous la décomposons comme

$$\sum_{k=1}^{m-1} = \sum_{|\frac{k}{m}-\rho|\leq\eta} + \sum_{|\frac{k}{m}-\rho|>\eta}\,. \tag{5.38}$$

La première somme donnera le terme principal, nous y remplacerons les fonctions par leur développement de Taylor et nous estimerons l'erreur commise. La seconde somme sera constituée de restes, son estimation sera réalisée à l'aide de majorations. Au cours de ces estimations, nous aurons besoin de comprendre comment la fonction F se comporte autour de son maximum. Pour cela, nous commençons par le lemme suivant

Lemme 25. *Pour tout h dans l'intervalle $[-\eta, \eta]$, nous avons*

$$F''(\rho+h) = F''(\rho)\Big(1 + O\Big(\frac{\eta}{\rho}\Big)\Big)\,. \tag{5.39}$$

Démonstration. Pour montrer ce lemme, nous allons calculer explicitement $F''(\rho+h)$. En considérant tous les termes de cette quantité, une simplification va apparaître et donner le bon terme de reste.

Reprenons l'expression (5.33) de F'' appliquée en $x = \rho + h$ et commençons par calculer le numérateur de $F''(\rho+h)$, que nous appelons N.

$$N = (\rho+h)\phi(\rho+h)\Big(-\psi(\rho+h) + (1-\rho-h)\psi'\Big) \\ - (1-\rho-h)\psi(\rho+h)\Big(\phi(\rho+h) + (\rho+h)\phi'\Big)$$

Rappelons-nous que les fonctions ψ et ϕ sont affines, nous écrivons donc

$$\phi(\rho+h) = \phi(\rho) + h\phi'\,,$$

et

$$\psi(\rho+h) = \psi(\rho) + h\psi'\,.$$

Le numérateur de $F''(\rho+h)$ peut donc s'écrire

$$N = (\rho+h)\Big(\phi(\rho) + h\phi'\Big)\Big(-\psi(\rho) + (1-\rho-2h)\psi'\Big) \\ - (1-\rho-h)\Big(\psi(\rho) + h\psi'\Big)\Big(\phi(\rho) + (\rho+2h)\phi'\Big)\,.$$

Nous allons maintenant tout développer et rassembler les termes en puissances de h. Le terme facteur de h^0 conduira à $F''(\rho)$, le terme facteur de h^1 est constitué de 6 termes :

$$\phi(\rho)\Big(-\psi(\rho) + (1-\rho)\psi'\Big) + \phi'\rho\Big(-\psi(\rho) + (1-\rho)\psi'\Big) - 2\psi'\rho\phi(\rho) \\ + \psi(\rho)\Big(\phi(\rho) + \rho\phi'\Big) - \psi'(1-\rho)\Big(\phi(\rho) + \rho\phi'\Big) - 2\phi'(1-\rho)\psi(\rho)\,,$$

après développement, la plupart de ces termes se simplifient, il reste un seul terme et ce terme contient le facteur ρ :

$$-2\rho\,\phi(\rho)\big(\phi' + \psi'\big)\,.$$

Ce facteur va venir se simplifier avec le terme ρ présent au dénominateur pour donner un terme d'ordre h/ρ. Nous ne nous attarderons pas sur les termes dont les

puissances de h sont supérieures car comme nous allons le voir, le dénominateur est au plus d'ordre ρ^2.

Concentrons-nous maintenant sur le dénominateur D de l'expression (5.33) appliquée en $x = \rho + h$

$$D = (\rho + h)(1 - \rho - h)\psi(\rho + h)\phi(\rho + h)\,.$$

Comme $|h|$ est inférieur à η, qui est lui-même négligeable devant ρ, ce produit est encadré comme

$$C_1\rho\psi(\rho) \leq D \leq C_2\rho\psi(\rho)\,,$$

où les constantes C_1 et C_2 ne dépendent pas des paramètres et sont strictement positives. De plus, comme $\psi(\rho) \geq \psi'\rho$, alors $1/D$ est d'ordre au plus ρ^2. Finalement, nous avons que

$$F''(\rho + h) = \frac{N}{D}$$

est encadré par

$$F''(\rho + h) = F''(\rho)\Big(1 + O\Big(\frac{h}{\rho}\Big)\Big) + O\Big(\frac{h}{\rho}\Big)\,.$$

Comme $F''(\rho)$ ne tend pas vers 0, nous avons donc

$$F''(\rho + h) = F''(\rho)\Big(1 + O\Big(\frac{\eta}{\rho}\Big)\Big)\,. \tag{5.40}$$

□

Le terme de reste

Le but de cette section est de majorer le terme de reste dans la somme (5.26). Commençons par nous intéresser à la seconde somme, nous pouvons la majorer par le nombre de termes, lui-même inférieur à m, multiplié par les maxima des fonctions sont

$$\Big|\sum_{|\frac{k}{m}-\rho|>\eta} g\Big(\frac{k}{m}\Big)e^{mF\left(\frac{k}{m}\right)}\Big| \leq m \sup_{[\frac{1}{m},1-\frac{1}{m}]} |g| \exp\Big(m \sup_{|x-\rho|>\eta} F(x)\Big)\,. \tag{5.41}$$

Comme la fonction F atteint son maximum au point ρ, et que η tend vers 0, nous avons que

$$\sup_{|x-\rho|>\eta} F(x) = \max\Big(F(\rho-\eta),\, F(\rho+\eta)\Big)\,.$$

Nous pouvons alors développer F autour de ρ pour estimer le facteur exponentiel de (5.41). La formule de Taylor donne

$$F\big(\rho \pm \eta\big) = F(\rho) + \frac{\eta^2}{2}F''(\rho + h)\,,$$

où h est un point entre $-\eta$ et η. Nous allons estimer cette quantité uniformément en h. Ainsi, à la suite de (5.41) et grâce au lemme 25, nous écrivons

$$\Big|\sum_{|\frac{k}{m}-\rho|>\eta} g\Big(\frac{k}{m}\Big)e^{mF\left(\frac{k}{m}\right)}\Big| \leq m \sup_{[\frac{1}{m},1-\frac{1}{m}]} |g| \exp\left(mF(\rho)+m\eta^2\left(F''(\rho)\Big(1+O\Big(\frac{\eta}{\rho}\Big)\Big)\right)\right)\,.$$

Intéressons-nous tout d'abord à la fonction g, d'après (5.24), elle s'exprime comme

$$g(x) = \Big(x(1-x)\phi(x)\psi(x)\Big)^{-1/2}.$$

Nous reprenons le fait que ϕ soit encadrée entre deux constantes et nous majorons uniformément la fonction g par

$$\sup_{[\frac{1}{m},1-\frac{1}{m}]} |g| \leq c\sqrt{\frac{m}{\psi\big(\frac{1}{m}\big)}} \leq c'm\,,$$

où c et c' sont des constantes. La seconde quantité qui apparaît dans cette majoration est, d'après la définition de η

$$m\eta^2 F''(\rho) = -m^{2\varepsilon}\,.$$

Enfin, le rapport η/ρ tend toujours vers 0 d'après (5.37), il est donc toujours négligeable devant $F''(\rho)$.

D'après la définition (5.30) de η, et comme $F''(\rho)$ est négatif nous avons finalement

$$\left|\sum_{|\frac{k}{m}-\rho|>\eta} g\Big(\frac{k}{m}\Big)e^{mF\left(\frac{k}{m}\right)}\right| \leq\ cm^2\,\exp\left(mF(\rho) - \Big(m^{2\varepsilon}\big(1+o(1)\big)\Big)\right).$$

Nous en déduisons la majoration

$$\mu\Big([0,1]\setminus[\rho_0-\sqrt{m},\,\rho_0+\sqrt{m}]\Big) \leq \delta_0 T e^{mF(\rho)}\exp\big(-m^{2\varepsilon}\big)\,. \tag{5.42}$$

De cette manière, lorsque nous aurons factorisé par $e^{mF(\rho)}$, nous dirons que cette quantité tend vers 0 de manière exponentielle.

Le terme principal

Le but de cette section est d'estimer le terme principal de (5.26). Intéressons-nous maintenant à la première somme de (5.38) qui va donner le terme principal de S, nous allons remplacer les fonctions par leur développement de Taylor autour de ρ. Occupons-nous tout d'abord de la fonction g, d'après son expression (5.24), nous avons que, pour tout h entre $-\eta$ et η,

$$g(\rho+h) = \Big((\rho+h)(1-\rho-h)\phi(\rho+h)\psi(\rho+h)\Big)^{-1/2}.$$

A l'aide ce que nous avons fait précédemment, et comme h/ρ tend vers 0,

$$g(\rho+h) = g(\rho)\Big(1+O\Big(\frac{\eta}{\rho}\Big)\Big)\,. \tag{5.43}$$

Nous pouvons donc remplacer g par son estimée (5.43), pour la fonction F, le lemme (25) nous permet d'écrire

$$F(\rho+h) = F(\rho) + h^2\left(F''(\rho)\Big(1+O\Big(\frac{\eta}{\rho}\Big)\Big)\right).$$

Ces deux estimations nous donnent que

$$\sum_{|\frac{k}{m}-\rho|<\eta} g\Big(\frac{k}{m}\Big)e^{mF\left(\frac{k}{m}\right)} = g(\rho)\Big(1+O\Big(\frac{\eta}{\rho}\Big)\Big)\sum_{|\frac{k}{m}-\rho|<\eta}\exp\left(mF(\rho)+m\big(\rho-\frac{k}{m}\big)^2F''(\rho)\Big(1+O\big(\frac{\eta}{\rho}\big)\Big)\right).$$

Comme η/ρ tend vers 0, nous nous ramenons à la somme

$$\sum_{|\frac{k}{m}-\rho|<\eta} g\Big(\frac{k}{m}\Big)e^{mF\left(\frac{k}{m}\right)} \sim g(\rho)e^{mF(\rho)}\sum_{|\frac{k}{m}-\rho|<\eta}\exp\Big(m(\rho-\frac{k}{m})^2F''(\rho)\Big).$$

Nous allons comparer cette somme à l'intégrale correspondante, pour cela il suffit de constater que la fonction

$$e(x) = e^{m(x-\rho)^2F''(\rho)}, \tag{5.44}$$

croît jusque ρ puis décroît (rappelons que $F''(\rho)$ est négatif). Nous allons maintenant comparer la somme avec l'intégrale correspondante, que nous pourrons calculer avec un changement de variables. Afin de simplifier les formules, notons

$$i_- = \frac{\lfloor m(\rho-\eta)\rfloor}{m}, \qquad i_0 = \frac{\lfloor m\rho\rfloor}{m}, \qquad i_+ = \frac{\lfloor m(\rho+\eta)\rfloor}{m}.$$

Nous devons évaluer la somme

$$\sum_{k=mi_-+1}^{mi_+} e\Big(\frac{k}{m}\Big), \tag{5.45}$$

pour cela, nous allons traiter séparément les deux intervalles $[i_-, i_0]$ et $[i_0, i_+]$.
• Tout d'abord, sur l'intervalle $[i_-,\, i_0]$, la fonction e est croissante donc pour tout k entre mi_- et mi_0

$$e\Big(\frac{k-1}{m}\Big) \le m\int_{(k-1)/m}^{k/m} e(s)\,\mathrm{d}s \le e\Big(\frac{k}{m}\Big).$$

Sommer ces inégalités conduit d'une part à

$$m\int_{i_-}^{i_0} e(s)\,\mathrm{d}s \le \sum_{k=mi_-+1}^{mi_0} e\Big(\frac{k}{m}\Big). \tag{5.46}$$

et d'autre part

$$\sum_{k=mi_-+1}^{mi_0} e\Big(\frac{k}{m}\Big) - e(i_0) \le m\int_{i_-+1/m}^{i_0} e(s)\,\mathrm{d}s. \tag{5.47}$$

• Considérons maintenant l'intervalle $[i_0,\ i_+]$. Cette fois, la fonction e est décroissante donc pour tout k entre mi_0 et mi_+

$$e\Big(\frac{k}{m}\Big) \leq m \int_{(k-1)/m}^{k/m} e(s)\,\mathrm{d}s \leq e\Big(\frac{k-1}{m}\Big)\,.$$

Nous sommons à nouveau les ingélités et nous obtenons

$$m \int_{i_0+1/m}^{i_++1/m} e(s)\,\mathrm{d}s \leq \sum_{k=mi_0+1}^{mi_+} e\Big(\frac{k}{m}\Big)\,. \tag{5.48}$$

et pour la majoration

$$\sum_{k=mi_0+1}^{mi_+} e\Big(\frac{k}{m}\Big) - e\Big(i_0+\frac{1}{m}\Big) \leq m \int_{i_0+1/m}^{i_+} e(s)\,\mathrm{d}s\,. \tag{5.49}$$

Nous pouvons maintenant encadrer notre somme, sommer les inégalités (5.46) et (5.48) conduisent à

$$m \int_{i_-}^{i_++1/m} e(s)\,\mathrm{d}s - m \int_{i_0}^{i_0+1/m} e(s)\,\mathrm{d}s \leq \sum_{k=mi_-+1}^{mi_+} e\Big(\frac{k}{m}\Big) \tag{5.50}$$

D'autre part en sommant (5.47) et (5.49), nous avons

$$\sum_{k=mi_-+1}^{mi_+} e\Big(\frac{k}{m}\Big) \leq m \int_{i_-+1/m}^{i_+} e(s)\,\mathrm{d}s - m \int_{i_0}^{i_0+1/m} e(s)\,\mathrm{d}s + e(i_0) + e\Big(i_0+\frac{1}{m}\Big)\,. \tag{5.51}$$

Comme la fonction e est positive et inférieure à 1, nous allons montrer que ces termes de bords sont négligeables devant l'intégrale, nous avons, d'après les inégalités (5.51) et (5.50)

$$m \int_{i_-}^{i_++1/m} e(s)\,\mathrm{d}s - 1 \leq \sum_{k=mi_-+1}^{mi_+} e\Big(\frac{k}{m}\Big) \leq m \int_{i_-+1/m}^{i_+} e(s)\,\mathrm{d}s + 2\,.$$

Nous pouvons nous ramener à l'intégrale sur l'intervalle $[\rho-\eta, \rho+\eta]$:

$$m \int_{\rho-\eta}^{\rho+\eta} e(s)\,\mathrm{d}s - 2 \leq \sum_{k=mi_-+1}^{mi_+} e\Big(\frac{k}{m}\Big) \leq m \int_{\rho-\eta}^{\rho+\eta} e(s)\,\mathrm{d}s + 3\,.$$

Nous pouvons maintenant estimer cette intégrale grâce au changement de variable

$$u = \sqrt{-mF''(\rho)}(x-\rho)\,.$$

Par définition de η, nous avons

$$\sqrt{-mF''(\rho)}\eta = m^{\varepsilon}\,.$$

Le changement de variable donne alors, d'après l'expression (5.44) de la fonction e,

$$m\int_{\rho-\eta}^{\rho+\eta} e(s)\,\mathrm{d}s = \sqrt{\frac{m}{-F''(\rho)}}\int_{-m^{\varepsilon}}^{m^{\varepsilon}} e^{-u^2}\,\mathrm{d}u\,.$$

L'intégrale restante est l'intégrale de Gauss et converge vers $\sqrt{\pi}$, finalement

$$m\int_{\rho-\eta}^{\rho+\eta} e(s)\,\mathrm{d}s \sim \sqrt{\pi}\sqrt{\frac{m}{-F''(\rho)}}\,.$$

La somme (5.45) ci-dessus vérifie donc

$$\sum_{k=mi_-+1}^{mi_+} e\Big(\frac{k}{m}\Big) \sim \sqrt{\pi}\sqrt{\frac{m}{-F''(\rho)}}\,.$$

Nous pouvons maintenant revenir à la mesure invariante, la masse de la mesure μ est équivalente à

$$Z = C\delta_0\, T\, g(\rho) e^{mF(\rho)}\sqrt{\frac{m}{-F''(\rho)}}\,,$$

où C est une certaine quantité constante dans le régime asymptotique. Ceci conclue la preuve du lemme 23 sur le calcul de la constante de normalisation. □

Fin de la preuve du théorème 21

Nous pouvons simplifier encore $g(\rho)$, d'après la relation (5.27),

$$g(\rho) = \frac{1}{\rho\phi(\rho)}\,,$$

et comme $\phi(\rho)$ est encadré entre deux constantes,

$$g(\rho) = \frac{C}{\rho}\,.$$

Nous pouvons donc déduire une mesure de probabilité invariante :

$$\forall k \in \{1,\cdots,m\} \quad \nu(k) = \rho\sqrt{\frac{-F''(\rho)}{m}}\, g\Big(\frac{k}{m}\Big) e^{mF(\frac{k}{m})-mF(\rho)}\,. \tag{5.52}$$

$$\nu(0) = \sqrt{\frac{-F''(\rho)}{m}}\frac{\rho}{\delta_0\, T\, e^{mF(\rho)}}\,.$$

Si nous nous restreignons aux points autour de ρ, nous pouvons encore simplifier g en écrivant

$$g(x) = \frac{1}{x}\left((1-x)\frac{\psi(x)}{x}\phi(x)\right)^{-1/2}\,.$$

La quantité entre parenthèses est alors encadrée entre 2 constantes strictement positives. Notre probabilité invariante devient, si k est entre $\rho-\eta$ et $\rho+\eta$, dans le cas **(MS)**,

$$\nu(k) = \frac{\rho}{\sqrt{m}}\,\frac{e^{mF(\frac{k}{m})-mF(\rho)}}{k/m}\,. \tag{5.53}$$

et dans le cas **(Mut)**,

$$\nu(k) = \sqrt{\frac{\rho}{m}}\,\frac{e^{mF(\frac{k}{m})-mF(\rho)}}{k/m}\,. \tag{5.54}$$

Dans les deux cas, la majoration (5.42) sur la mesure μ conduit à une estimation sur la masse de la mesure ν en dehors de l'intervalle proche de ρ :

$$\nu\Big([0,1]\setminus\Big[\rho-\eta,\,\rho+\eta\Big]\Big)\leq R\,, \tag{5.55}$$

et R tend exponentiellement vite vers 0.

3 Le processus $|Z_0$

Nous allons commencer par reprendre les processus majorant et minorant de la proportion de master sequences du chapitre précédent. Nous noterons encore dans la suite

$$L_q = \sigma - 1 - \sigma M_{00}\,. \tag{5.56}$$

Cette quantité est la différence entre $\sigma(1-M_{00})$ qui est reliée à l'autodestruction des master sequences, et 1 qui est la probabilité pour une non master sequences d'engendrer une non master sequences. Le signe de L_q indique comment les master sequences meurent. Si la probabilité de mutation est petite, les master sequences ont peu de chances de s'autodétruire, donc L_q négatif. En reprenant les estimations du chapitre précédent, nous avons toujours que L_q ne s'annule jamais. Ainsi, en définissant les fonctions

$$\psi(x) = f(M_{10}) + \sigma M_{00}\,x\,, \tag{5.57}$$

et

$$\phi(x) = 1 - f(M_{10}) + L_q\,x\,. \tag{5.58}$$

Le rapport des probabilités de transition s'écrit

$$\frac{\delta_k}{\gamma_k} = \frac{1-\frac{k}{m}}{\frac{k}{m}}\;\frac{\psi\big(\frac{k}{m}\big)}{\phi\big(\frac{k}{m}\big)}\,.$$

Les deux fonctions ϕ et ψ vérifient bien l'hypothèse **(H)** : La quantité $\phi(1)$ est bien positive puisque $1+L_q$ tend vers $\sigma-1$ et M_{10} tend vers 0.

Des deux cas que nous avons distingués dans la section générale, nous nous trouvons dans le cas **(MS)**. La fonction centrale pour ce processus est donc

$$F_0(x) = -(1-x)\ln(1-x) - x\ln x + \frac{\psi(x)}{\psi'}\ln\psi(x) - \frac{\phi(x)}{\phi'}\ln\phi(x)\,.$$

Reprenons la relation (5.28) nous permettant de calculer ρ, le maximum de la fonction F. Dans notre borne inférieure, $\psi(0)=0$, il est donc beaucoup plus simple de calculer ρ :

$$\underline{\rho_0} = \frac{\sigma M_{00}-1}{\sigma-1} = r_0\,.$$

Pour la borne supérieure en revanche, il faut résoudre le trinôme, la résolution donne

$$\overline{\rho_0} = r_0\left(\frac{1}{2} + \frac{1}{2}\sqrt{1 + 4\frac{M_{10}}{(\sigma-1)(r_0)^2}}\right).$$

Nous pouvons résumer ces deux cas en une seule égalité :

$$\overline{\rho_0} = r_0\left(\frac{1}{2} + \frac{1}{2}\sqrt{1 + 4\frac{f(M_{10})}{(\sigma-1)(r_0)^2}}\right). \tag{5.59}$$

A priori, nous ne connaissons pas le comportement du rapport q/r_0^2 et ne pouvons donc pas développer la racine. Nous voulons obtenir le développement du point maximal de la mesure invariante, nous ne pouvons pas conclure ici. En effet les deux bornes ne coïncident pas, et nous allons devoir ruser pour que la borne supérieure se rapproche du vrai développement, donné par la borne inférieure. C'est plutôt normal, notre borne supérieure considère que toute la population hormis les master sequences sont dans la classe 1, cette simplification est trop grossière. Nous suivons les calculs faits dans le cas général et nous obtenons donc la probabilité invariante grâce à (5.52) : en posant

$$\eta_0 = \frac{m^\varepsilon}{\sqrt{m}},$$

nous obtenons, pour tout k compris entre $\rho_0 - \eta_0$ et $\rho_0 + \eta_0$

$$\nu_0(k) = \rho_0 \frac{\exp\left(mF_0(k/m) - mF_0(\rho_0)\right)}{\sqrt{m}\frac{k}{m}}.$$

4 Le processus Z_1

Essayons maintenant d'estimer le nombre d'individus de la classe 1 en utilisant nos estimations sur le nombre de master sequences. Notons z_0 la proportion de master sequences.

$$z_0 = \frac{1}{m}\text{nombre de master sequences à l'instant } t\,. \tag{5.60}$$

Nous voyons ici z_0 comme un paramètre fixé, en particulier, z_0 ne dépend pas du temps. Evaluons δ_k (il y a donc dans la population k individus de classe 1) :

$$\delta_k = P\begin{pmatrix}\text{Remplacer un individu}\\ \text{de classe autre que 1}\end{pmatrix} P\begin{pmatrix}\text{Enfant est}\\ \text{une classe 1}\end{pmatrix}.$$

Nous conditionnons alors selon la classe du parent.

$$\begin{aligned}\delta_k = \left(1 - \frac{k}{m}\right)\Bigg[&P\begin{pmatrix}\text{Enfant est}\\ \text{une classe 1}\end{pmatrix} \text{et} \begin{pmatrix}\text{son parent est}\\ \text{une master sequence}\end{pmatrix}\\ &+ P\begin{pmatrix}\text{Enfant est}\\ \text{une classe 1}\end{pmatrix} \text{et} \begin{pmatrix}\text{son parent est}\\ \text{une classe 1}\end{pmatrix}\\ &+ P\begin{pmatrix}\text{Enfant est}\\ \text{une classe 1}\end{pmatrix} \text{et} \begin{pmatrix}\text{son parent est}\\ \text{de classe } J > 1\end{pmatrix}\Bigg].\end{aligned}$$

Les deux premières probabilités ne dépendent que du nombre de master sequences et de classes 1, nous pouvons les calculer :

$$P\left(\begin{array}{c}\text{Enfant est}\\ \text{une classe 1}\end{array}\text{ et }\begin{array}{c}\text{son parent est}\\ \text{une master sequence}\end{array}\right)=\frac{\sigma z_0}{\text{fit}(z_0)}M_{01}\,,$$

et pour la deuxième,

$$P\left(\begin{array}{c}\text{Enfant est}\\ \text{une classe 1}\end{array}\text{ et }\begin{array}{c}\text{son parent est}\\ \text{une classe 1}\end{array}\right)=\frac{1}{\text{fit}(z_0)}\frac{k}{m}M_{11}\,.$$

Comme nous l'avons fait pour le nombre de master sequences, nous allons maintenant encadrer la dernière probabilité de transition par au-dessus et par en-dessous car elle dépend de la répartition de la population dans les différentes classes. Tout d'abord, nous obtenons une minoration en remarquant que la dernière probabilité est positive, ce qui revient à négliger les mutations de retour des classes 2 et supérieures. Pour la majoration, on considère que le génôme du parent est une classe 2 :

$$P\left(\begin{array}{c}\text{Enfant est}\\ \text{une classe 1}\end{array}\text{ et }\begin{array}{c}\text{son parent est}\\ \text{de classe } J>1\end{array}\right)\leq \frac{1-\frac{k}{m}}{\text{fit}(z_0)}P\left(\begin{array}{c}\text{Enfant issu d'un parent de classe 2}\\ \text{est une classe 1}\end{array}\right)\leq\frac{M_{21}}{\text{fit}(z_0)}\,.$$

Notre encadrement est satisfaisant puisque la différence entre les deux bornes tend vers 0 comme q. Nous construirons donc la borne inférieure en remplaçant cette probabilité par 0 et la borne supérieure en la remplaçant par M_{21}. De même, pour la probabilité de perdre un individu de classe 1, nous avons

$$\gamma_k=\frac{k}{m}\left[P\left(\begin{array}{c}\text{Enfant n'est pas}\\ \text{une classe 1}\end{array}\text{ et }\begin{array}{c}\text{son parent est}\\ \text{une master sequence}\end{array}\right)+P\left(\begin{array}{c}\text{Enfant n'est pas}\\ \text{une classe 1}\end{array}\text{ et }\begin{array}{c}\text{son parent est}\\ \text{une classe 1}\end{array}\right)+P\left(\begin{array}{c}\text{Enfant n'est pas}\\ \text{une classe 1}\end{array}\text{ et }\begin{array}{c}\text{son parent est}\\ \text{de classe } J>1\end{array}\right)\right].$$

Comme pour δ, nous pouvons calculer les deux premières probabilités :

$$P\left(\begin{array}{c}\text{Enfant n'est pas}\\ \text{une classe 1}\end{array}\text{ et }\begin{array}{c}\text{son parent est}\\ \text{une master sequence}\end{array}\right)=\frac{\sigma z_0}{\text{fit}(z_0)}(1-M_{01})\,,$$

$$P\left(\begin{array}{c}\text{Enfant n'est pas}\\ \text{une classe 1}\end{array}\text{ et }\begin{array}{c}\text{son parent est}\\ \text{une classe 1}\end{array}\right)=\frac{1}{\text{fit}(z_0)}\frac{k}{m}(1-M_{11})\,.$$

En revanche, la dernière probabilité dépend de la répartition de la population, nous devons donc maintenant encadrer cette dernière probabilité. Comme γ_k est la probabilité de perdre un individu de classe 1, majorer γ_k conduira à un processus plus

petit. Nous majorons cette probabilité par la probabilité de choisir un parent de classe $J > 1$:

$$P\left(\begin{array}{c}\text{Enfant n'est pas}\\ \text{une classe 1}\end{array}\text{ et }\begin{array}{c}\text{son parent est}\\ \text{de classe } J>1\end{array}\right) \leq \frac{1-z_0-\frac{k}{m}}{\mathrm{fit}(z_0)}\,.$$

Pour minorer cette probabilité, l'enfant a moins de chance de ne pas être une classe 1 si le génôme de son parent est dans la classe 2 :

$$P\left(\begin{array}{c}\text{Enfant n'est pas}\\ \text{une classe 1}\end{array}\text{ et }\begin{array}{c}\text{son parent est}\\ \text{de classe } J>1\end{array}\right) \geq \frac{1-z_0-\frac{k}{m}}{\mathrm{fit}(z_0)} P\left(\begin{array}{c}\text{Enfant issu d'un parent de classe 2}\\ \text{n'est pas une classe 1}\end{array}\right) \geq \frac{1-z_0-\frac{k}{m}}{\mathrm{fit}(z_0)}(1-M_{21}) \geq \frac{1-z_0-\frac{k}{m}-M_{21}}{\mathrm{fit}(z_0)}\,.$$

Pour simplifier les formules, nous posons

$$L_q^1 = \sigma - 1 - \sigma M_{01}\,.$$

Introduisons maintenant les notations habituelles, notre borne inférieure conduit à poser

$$\underline{\psi}^{z_0}(x) = \sigma z_0 M_{01} + M_{11} x\,,$$

et

$$\underline{\phi}^{z_0}(x) = 1 + L_q^1 z_0 - M_{11} x\,.$$

Pour la borne supérieure

$$\overline{\psi}^{z_0}(x) = \sigma z_0 M_{01} + M_{21} + M_{11} x\,,$$

et

$$\overline{\phi}^{z_0}(x) = 1 - M_{21} + L_q^1 z_0 - M_{11} x\,.$$

Nous avons finalement encadré les probabilités du vrai processus comme suit :

$$\Big(1-\frac{k}{m}\Big)\frac{\underline{\psi}(k/m)}{\mathrm{fit}\big(\frac{k}{m}\big)} \;\leq\; \delta_k \;\leq\; \Big(1-\frac{k}{m}\Big)\frac{\overline{\psi}(k/m)}{\mathrm{fit}\big(\frac{k}{m}\big)}\,,$$

et

$$\frac{k}{m}\frac{\underline{\phi}(k/m)}{\mathrm{fit}\big(\frac{k}{m}\big)} \;\geq\; \gamma_k \;\geq\; \frac{k}{m}\frac{\overline{\phi}(k/m)}{\mathrm{fit}\big(\frac{k}{m}\big)}\,.$$

Nous traiterons les deux bornes de la même manière dans la suite. Pour cela, nous les mettons sous la forme de l'équation (5.15), écrivons, en adaptant la même notation que précédemment pour les deux bornes

$$\frac{\delta_k}{\gamma_k} = \frac{1-\frac{k}{m}}{\frac{k}{m}}\;\frac{\psi\big(\frac{k}{m}\big)}{\phi\big(\frac{k}{m}\big)}\,. \tag{5.61}$$

avec

$$\psi^{z_0}(x) = \sigma z_0 M_{01} + f(M_{21}) + M_{11} x \,,$$

et

$$\phi^{z_0}(x) = 1 - f(M_{21}) + L_q^1 z_0 - M_{11} x \,.$$

Ces deux fonctions vérifient bien les hypothèses nécessaires pour que le processus se trouve dans la classe **(H)**. Pour le moment, nous ne pouvons pas nous assurer que $\psi(0)$ est supérieur à une constante multipliée par r_0 comme nous l'avions annoncé puisqu'a priori, z_0 peut prendre toutes les valeurs. Cependant, lorsque nous intégrerons contre la mesure du processus des master sequences, nous nous restreindrons à un intervalle dont les bornes vérifierons cette condition.

Aussi, nous sommes dans le cas **(Mut)**. Introduisons les notations nécessaires à l'expression de la mesure invariante :

$$F_1^{z_0}(x) = -(1-x)\ln(1-x) - x\ln x + \frac{\psi^{z_0}(x)}{\psi'} \ln \psi^{z_0}(x) - \frac{\phi^{z_0}(x)}{\phi'} \ln \phi^{z_0}(x) \,,$$

Pour estimer la mesure invariante de ce processus, nous allons transférer les estimées que nous avons sur le nombre de master sequences.

4.1 Transfert des estimées

Nous avons défini un processus dont la mesure invariante $\nu_1^{z_0}$ dépend du nombre de master sequences z_0.

Pour obtenir une mesure invariante qui ne dépend pas de ce nombre, nous aimerions pouvoir intégrer cette mesure contre la mesure ν_0, cependant, cette opération n'a a priori aucune raison d'être reliée à la mesure invariante du processus du processus de Markov en dimension 2. Nous allons tout de même définir une mesure en intégrant ces mesures conditionnées et nous espérons pouvoir déduire des estimations sur la mesure invariante dans un travail futur. Nous définissons

$$\overline{\nu_1} = \int_{z_0} \overline{\nu_1^{z_0}}\, \overline{\nu_0}(\mathrm{d}z_0) \,,$$

pour la borne supérieure. De même, pour la borne inférieure, nous posons

$$\underline{\nu_1} = \int_{z_0} \underline{\nu_1^{z_0}}\, \underline{\nu_0}(\mathrm{d}z_0) \,.$$

Nous oublierons les barres pour la suite et noterons simplement

$$\nu_1 = \int_{z_0} \nu_1^{z_0}\, \nu_0(\mathrm{d}z_0) \,.$$

Nous savons que la mesure ν_0 est concentrée autour de ρ_0 avec une fenêtre que nous noterons

$$\eta_0 = \frac{m^\varepsilon}{\sqrt{m}} \,,$$

nous découpons donc cette intégrale et nous écrivons

$$\nu_1 = \int_{|z_0-\rho_0|\leq\eta_0} \nu_1^{z_0}\,\nu_0(\mathrm{d}z_0) + \int_{|z_0-\rho_0|>\eta_0} \nu_1^{z_0}\,\nu_0(\mathrm{d}z_0)\,.$$

Comme $\nu_1^{z_0}$ est une mesure de probabilité, nous avons une borne sur la deuxième intégrale :

$$\int_{|z_0-\rho_0|>\eta_0} \nu_1^{z_0}\nu_0(\mathrm{d}z_0) \leq \nu_0\Big([0,1] \setminus \Big[\rho_0-\eta_0,\, \rho_0+\eta_0\Big] \Big)\,.$$

D'après notre majoration (5.11), cette quantité tend exponentiellement vite vers 0. Nous nous concentrons donc sur la première intégrale, c'est-à-dire que nous plaçons dans la situation où z_0 est compris entre $\rho_0-\eta_0$ et $\rho_0+\eta_0$. D'après (5.29), la fonction F_1 atteint son maximum au point

$$\rho_1(z_0) = \frac{\sigma M_{01} z_0 + f(M_{21})}{1 - M_{11} + (\sigma-1)z_0}\,. \tag{5.62}$$

Nous pouvons maintenant justifier le fait que $\rho_1(z_0)$ est supérieur à une constante multipliée par r_0. Comme z_0 est supérieur à $\rho_0-\eta_0$ qui est supérieur à $r_0-\eta_0$, et que η_0 est négligeable devant r_0 grâce à l'hypothèse (5.2) donnée au début du chapitre., il existe bien une constante positive C telle que $\rho_1(z_0) > Cr_0$.

D'après le lemme 24, la dérivée seconde de la fonction $F_1^{z_0}$ au point $\rho_1(z_0)$ est d'ordre $1/\rho_1(z_0)$. Nous prendrons donc comme fenêtre de sommation

$$\eta_1 = \frac{\sqrt{\rho_1(z_0)}}{\sqrt{m}} m^{\varepsilon}\,. \tag{5.63}$$

Ainsi, si k est compris dans l'intervalle $\big[\rho_1(z_0)-\eta_1,\, \rho_1(z_0)+\eta_1\big]$,

$$\nu_1^{z_0}(k) = \sqrt{\frac{\rho_1(z_0)}{m}}\,\frac{\exp\Big(mF_1^{z_0}(k/m) - mF_1^{z_0}\big(\rho_1(z_0)\big)\Big)}{\frac{k}{m}}\,. \tag{5.64}$$

Malheureusement, le méthode de Laplace ne nous permet pas de calculer l'intégrale

$$\nu_1 \sim \int_{|z_0-\rho_0|\leq\eta_0} \nu_1^{z_0}\nu_0(\mathrm{d}z_0)\,, \tag{5.65}$$

à cause des termes de restes. Nous conservons donc la définition de la mesure ν_1 comme cette intégrale, nous ne pourrons jamais la calculer mais cela ne nous empêchera pas de calculer la mesure analogue $||\nu_0$, que nous définissons dans la prochaine section.

5 Le processus $||Z_0$

En l'état actuel, nous n'avons pas a priori d'estimation sur le nombre d'individus dans la classe 1, nous espérons cependant pouvoir déduire une telle estimation à partir de l'intégrale (5.65). Si nous déduisons de ce qui précède une première estimation

du nombre d'individus de classe 1, nous pourrons alors en déduire une estimation plus précise du nombre de master sequences. Notons z_1 la proportion d'individus dans la classe 1 :

$$z_1 = \frac{1}{m}\text{nombre d'individus dans la classe 1 à l'instant } t\,. \tag{5.66}$$

Evaluons δ_k, la probabilité de passer en une étape de k à $k+1$ master sequences.

$$\delta_k = P\begin{pmatrix}\text{Remplacer un individu}\\ \text{non-master sequence}\end{pmatrix} P\begin{pmatrix}\text{Enfant est}\\ \text{une master sequence}\end{pmatrix}.$$

Nous conditionnons alors selon la classe du parent.

$$\begin{aligned}\delta_k = \Big(1-\frac{k}{m}\Big)\Bigg[&P\begin{pmatrix}\text{Enfant est}\\ \text{une master sequence}\end{pmatrix} \text{et} \begin{pmatrix}\text{son parent est}\\ \text{une master sequence}\end{pmatrix}\\ &+P\begin{pmatrix}\text{Enfant est}\\ \text{une master sequence}\end{pmatrix} \text{et} \begin{pmatrix}\text{son parent est}\\ \text{une classe 1}\end{pmatrix}\\ &+P\begin{pmatrix}\text{Enfant est}\\ \text{une master sequence}\end{pmatrix} \text{et} \begin{pmatrix}\text{son parent est}\\ \text{de classe } J>1\end{pmatrix}\Bigg].\end{aligned}$$

Nous pouvons calculer les deux premières probabilités. Afin d'obtenir deux processus qui ne dépendent pas de la répartition de toute la population, nous encadrons la dernière probabilité. Pour obtenir une borne inférieure, il nous suffit de minorer cette probabilité par 0, ce qui revient à négliger les retours des classes 2 et supérieures. Pour la borne supérieure, nous majorons cette probabilité en faisant comme si le génôme du parent était de classe 1 :

$$\begin{aligned}&P\begin{pmatrix}\text{Enfant est une}\\ \text{master sequence}\end{pmatrix} \text{et} \begin{pmatrix}\text{son parent est}\\ \text{de classe } J>1\end{pmatrix} \leq\\ &\qquad\frac{1-\frac{k}{m}-z_1}{\text{fit}\left(\frac{k}{m}\right)}P\begin{pmatrix}\text{Enfant issu d'un parent de classe 2}\\ \text{est une master sequence}\end{pmatrix} \leq \frac{M_{20}}{\text{fit}\left(\frac{k}{m}\right)}\,.\end{aligned}$$

Notre encadrement est satisfaisant puisque la différence entre les deux bornes tend vers 0 comme q. Nous construirons donc la borne inférieure en remplaçant cette probabilité par 0 et la borne supérieure en la remplaçant par M_{20}. De même, pour γ_k, on perd une master sequence si

$$\begin{aligned}\gamma_k = \frac{k}{m}\Bigg[&P\begin{pmatrix}\text{Enfant n'est pas}\\ \text{une master sequence}\end{pmatrix} \text{et} \begin{pmatrix}\text{son parent est}\\ \text{une master sequence}\end{pmatrix}\\ &+P\begin{pmatrix}\text{Enfant n'est pas}\\ \text{une master sequence}\end{pmatrix} \text{et} \begin{pmatrix}\text{son parent est}\\ \text{une classe 1}\end{pmatrix}\\ &+P\begin{pmatrix}\text{Enfant n'est pas}\\ \text{une master sequence}\end{pmatrix} \text{et} \begin{pmatrix}\text{son parent est}\\ \text{de classe } J>1\end{pmatrix}\Bigg].\end{aligned}$$

Comme pour δ_k, nous pouvons calculer les deux premières probabilités et nous devons maintenant encadrer cette dernière probabilité. Comme γ est la probabilité de perdre une master sequence, majorer γ conduira à un processus plus petit. Nous majorons cette probabilité par la probabilité de choisir un parent de classe $J > 1$:

$$P\left(\begin{array}{c}\text{Enfant n'est pas}\\ \text{une master sequence}\end{array}\text{ et }\begin{array}{c}\text{son parent est}\\ \text{de classe } J>1\end{array}\right) \leq \frac{1-\frac{k}{m}-z_1}{\mathrm{fit}\left(\frac{k}{m}\right)}\,.$$

Pour minorer cette probabilité, l'enfant a moins de chance de ne pas être une master sequence si le génôme de son parent est dans la classe 2, d'où

$$P\left(\begin{array}{c}\text{Enfant n'est pas}\\ \text{une master sequence}\end{array}\text{ et }\begin{array}{c}\text{son parent est}\\ \text{de classe } J>1\end{array}\right) \geq \frac{1-\frac{k}{m}-z_1}{\mathrm{fit}\left(\frac{k}{m}\right)} P\left(\begin{array}{c}\text{Enfant issu d'un parent de classe 2}\\ \text{n'est pas une master sequence}\end{array}\right) \geq \frac{1-\frac{k}{m}-z_1}{\mathrm{fit}\left(\frac{k}{m}\right)}(1-M_{20}) \geq \frac{1-\frac{k}{m}-z_1-M_{20}}{\mathrm{fit}\left(\frac{k}{m}\right)}\,.$$

Comme précédemment, Nous définissons les deux fonctions

$$\psi^{z_1}(x) = \sigma M_{00}\, x + z_1 M_{10} + f(M_{20})\,,$$

et

$$\phi^{z_1}(x) = 1 + L_q\, x - M_{10} z_1 - f(M_{20})\,.$$

Nous nous trouvons donc à nouveau dans le cas **(MS)**, et nous posons

$$F_0^{z_1}(x) = -(1-x)\ln(1-x) - x\ln x + \frac{\psi^{z_1}(x)}{\psi^{z_1\prime}}\ln\psi^{z_1}(x) - \frac{\phi^{z_1}(x)}{\phi^{z_1\prime}}\ln\phi^{z_1}(x)\,,$$

Nous ne pouvons pas continuer sans transférer les estimées que nous avons obtenues sur le nombre de classe 1.

Comme ce que nous avons fait plus haut, nous allons intégrer $||\nu_0^{z_1}$ contre la mesure ν_1.

$$||\nu_0 = \int_{z_1} ||\nu_0^{z_1}\nu_1(\mathrm{d}z_1)\,.$$

La mesure ν_1 s'exprime elle aussi avec une intégrale (5.65). Nous allons la remplacer par son expression puis échanger l'ordre d'intégration, commençons par écrire

$$||\nu_0 \sim \int_{|z_0-\rho_0|\leq\eta_0}\int_{z_1} ||\nu_0^{z_1}\nu_1^{z_0}(\mathrm{d}z_1)|\nu_0(\mathrm{d}z_0)\,. \tag{5.67}$$

Maintenant, nous allons à nouveau séparer l'intégrale en deux

$$\int_{z_1} ||\nu_0^{z_1}|\nu_1^{z_0}(\mathrm{d}z_1) = \int_{|z_1-\rho_1(z_0)|\leq\eta_1} ||\nu_0^{z_1}\nu_1^{z_0}(\mathrm{d}z_1) + \int_{|z_1-\rho_1(z_0)|>\eta_1} ||\nu_0^{z_1}\nu_1^{z_0}(\mathrm{d}z_1)\,.$$

Nous pouvons alors utiliser notre estimation (5.11) sur la masse de la mesure $\nu_1^{z_0}$ en dehors de l'intervalle $[\rho_1(z_0) - \eta_1,\, \rho_1(z_0) + \eta_1]$:

$$\int_{|z_1-\rho_1(z_0)|>\eta_1} \big|\big|\nu_0^{z_1}\nu_1^{z_0}(\mathrm{d}z_1) \leq \exp(-m^\varepsilon)\,.$$

Ainsi, nous nous ramenons seulement à l'intégrale autour de ce maximum :

$$\big|\big|\nu_0 \sim \int_{|z_0-\rho_0|\leq\eta_0} \int_{|z_1-\rho_1(z_0)|\leq\eta_1} \big|\big|\nu_0^{z_1}\nu_1^{z_0}(\mathrm{d}z_1)|\nu_0(\mathrm{d}z_0)\,. \tag{5.68}$$

Nous nous intéressons donc maintenant à cette intégrale et supposerons que z_1 est encadré entre $\rho_1(z_0)-\eta_1$ et $\rho_1(z_0)+\eta_1$. Maintenant que nous avons une idée de l'ordre de grandeur de z_1, nous pouvons calculer explicitement la mesure $\nu_0^{z_1}$ en suivant les calculs généraux de la première section. Nous en déduisons la probabilité invariante à la constante près au voisinage de $\rho_0(z_1)$,

$$\big|\big|\nu_0^{z_1}(k) = \rho_0(z_1)\frac{e^{mF_0^{z_1}(k/m)-mF_0^{z_1}(\rho_0(z_1))}}{\sqrt{m}\,k/m}\,.$$

5.1 Intégrer

Reprenons dans cette section des notations générales, nous allons montrer un résultat qui nous servira plusieurs fois par la suite. Nous supposons ici que les fonctions ϕ et ψ dépendent d'une variable z et nous allons intégrer par rapport à cette variable. La quantité z vit dans un certain intervalle autour de z^*. Nous notons

$$\psi(x) = \psi(x,z) = \psi_1(x) + \psi_2(z)\,,$$

De manière analogue pour ϕ, nous posons

$$\phi(x) = \phi(x,z) = \phi_1(x) + \phi_2(z)\,.$$

Les fonctions ϕ_1 et ψ_1 sont toujours affines, mais ce ne sera pas forcément le cas pour les fonction ϕ_2 et ψ_2. Le maximum ρ de la mesure ν dépend lui aussi de z, nous noterons $\rho(z)$ dans la suite.

Lemme 26. *Dans le cas **(MS)**, nous avons l'égalité*

$$\nu^z(x) = \nu^{z^*}(x)\big(1+R_1\big)\exp(R_2)\,,$$

avec

$$|R_1| \leq \sup|\psi_2'|\frac{|z-z^*|}{(\sigma-1)r_0}\,,$$

et

$$|R_2| \leq 2m\frac{\sup\big|\psi_2'\big|}{\psi\big(\rho(z),z\big)}\big|x-\rho(z)\big|\big|z-z^*\big|\,.$$

Démonstration. Pour montrer cela, nous partons de l'expression de F dans la mesure ν et nous écrivons

$$F(x,z) = -x\ln x - (1-x)\ln(1-x) + \frac{\psi(x,z)}{{\psi_1}'}\ln\psi(x,z) - \frac{\phi(x,z)}{{\phi_1}'}\ln\phi(x,z)\,.$$

Nous allons remplacer les fonctions par leur développement de Taylor, il nous faut donc calculer la dérivée de $F(x,z)$ par rapport à la variable z ainsi que celle de $F(\rho(z),z)$. Avec ces notations,

$$\Psi(x,z) = \frac{\psi(x,z)}{\psi_1'}\ln\psi(x,z) - x\,.$$

Commençons par calculer la dérivée de Ψ par rapport à la variable z,

$$\frac{\partial}{\partial z}\Psi(x,z) = \frac{\psi_2'(z)}{{\psi_1}'}\Big(\ln\psi(x,z)+1\Big)\,,$$

et de même pour la fonction Φ. Ainsi,

$$\frac{\partial}{\partial z}F(x,z) = \frac{\psi_2'(z)}{{\psi_1}'}\Big(\ln\psi(x,z)+1\Big) - \frac{\phi_2'(z)}{{\phi_1}'}\Big(\ln\phi(x,z)+1\Big)\,.$$

Nous avons aussi besoin de la dérivée par rapport à la variable z de $F\big(\rho(z),z\big)$. A nouveau, commençons par calculer la dérivée de la fonction Ψ :

$$\Psi\big(\rho(z),z\big) = \frac{\psi_1\big(\rho(z)\big)+\psi_2(z)}{\psi_1'}\ln\psi\big(\rho(z),z\big) - x\,.$$

Ainsi, nous avons

$$\frac{\partial}{\partial z}\Psi\big(\rho(z),z\big) = \frac{1}{\psi_1'}\Big(\psi_1'\frac{\partial}{\partial z}\rho(z) + \psi_2'(z)\Big)\Big(\ln\psi\big(\rho(z),z\big)+1\Big)\,.$$

D'autre part, par dérivée d'une composée,

$$\frac{\partial}{\partial z}\Big(-\rho(z)\ln\rho(z) - \big(1-\rho(z)\big)\ln\big(1-\rho(z)\big)\Big) = \Big(\ln\big(1-\rho(z)\big) - \ln\rho(z)\Big)\frac{\partial}{\partial z}\rho(z)\,.$$

En utilisant la relation (5.27), tous les termes facteurs de $\frac{\partial}{\partial z}\rho(z)$ se simplifient, nous avons alors

$$\frac{\partial}{\partial z}F\big(\rho(z),z\big) = \frac{\psi_2'(z)}{\psi_1'}\Big(\ln\psi\big(\rho(z),z\big)+1\Big) - \frac{\phi_2'(z)}{\phi_1'}\Big(\ln\phi\big(\rho(z),z\big)+1\Big)\,.$$

Nous écrivons alors

$$\frac{\partial}{\partial z}\Big(F(x,z) - F\big(\rho(z),z\big)\Big) = \frac{\psi_2'(z)}{\psi_1'}\ln\frac{\psi(x,z)}{\psi\big(\rho(z),z\big)} - \frac{\phi_2'(z)}{\phi_1'}\ln\frac{\phi(x,z)}{\phi\big(\rho(z),z\big)}\,.$$

Il nous suffit maintenant de simplifier ces fractions : comme la fonction ψ_1 est affine, nous écrivons

$$\psi(x,z) = \psi_1(x) + \psi_2(z) = \psi\big(\rho(z),z\big) + \big(x-\rho(z)\big)\psi_1'\,,$$

et en utilisant l'inégalité $\ln(1+u) \leq u$, nous obtenons

$$\left|\frac{\partial}{\partial z}\Big(F(x,z) - F\big(\rho(z),z\big)\Big)\right| \leq \psi_2'(z)\,\frac{x-\rho(z)}{\psi\big(\rho(z),z\big)} + \phi_2'(z)\,\frac{x-\rho(z)}{\phi\big(\rho(z),z\big)}\,.$$

Nous appliquons maintenant la formule de Taylor :

$$F(x,z) - F\big(\rho(z),z\big) = F(x,z^*) - F\big(\rho(z^*),z^*\big) + (z-z^*)\frac{\partial}{\partial z}\Big(F(x,z) - F\big(\rho(z),z\big)\Big)\Big|_{\xi}\,,$$

Pour un certain ξ entre z et z^*. Nous majorons cette dérivée par

$$\left|\frac{\partial}{\partial z}\Big(F(x,z) - F\big(\rho(z),z\big)\Big)\right| \leq \left(\frac{\big|\psi_2'(z)\big|}{\psi\big(\rho(z),z\big)} + \frac{\big|\phi_2'(z)\big|}{\phi\big(\rho(z),z\big)}\right)\big|x-\rho(z)\big|\,.$$

Comme la fonction ϕ est encadrée par des constantes non nulle, et que $\psi(0)$ tend vers 0, nous avons aussi

$$\left|\frac{\partial}{\partial z}\Big(F(x,z) - F\big(\rho(z),z\big)\Big)\right| \leq 2\frac{\big|\psi_2'(z)\big|}{\psi\big(\rho(z),z\big)}\big|x-\rho(z)\big|\,.$$

Dans l'expression de la mesure ν, il y a aussi un facteur $\rho(z)$, La relation (5.35) nous donne

$$\Big(\rho(z) - \frac{r_0}{2}\Big)^2 = \frac{r_0^2}{4} + \frac{\psi(0,z)}{\sigma-1}\,.$$

Ainsi, en soustrayant cette égalité et la même pour z^*, nous obtenons

$$\Big(\rho(z) - \rho(z^*)\Big)\Big(\rho(z) + \rho(z^*) - r_0\Big) = \frac{\psi_2(z) - \psi_2(z^*)}{\sigma-1}\,.$$

Comme ρ est supérieur à r_0 d'après (5.28), nous en déduisons

$$\big|\rho(z) - \rho(z^*)\big| \leq \sup|\psi_2'|\frac{|z-z^*|}{(\sigma-1)r_0}\,.$$

Ce qu'il nous fallait démontrer. □

5.2 Intégrons contre la mesure de Z_1

Reprenons l'expression intégrale (5.68) de la mesure ν_0. Nous allons tout d'abord calculer l'intégrale centrale à z_0 fixé. Pour cela, nous nous servons du lemme 26 pour écrire

$$\big|\big|\nu_0^{z_1}(x) = \big|\big|\nu_0^{\rho_1(z_0)}(x)\big(1+R_1\big)\exp(R_2)\,,$$

avec, comme $M_{10} \leq q$,

$$|R_1| \leq \eta_1\frac{M_{10}}{(\sigma-1)r_0}\,,$$

et

$$|R_2| \leq m\frac{M_{10}}{\psi\big(\rho_0(z_1)\big)}\eta_1\,\eta_0\,.$$

Vérifions que ces restes tendent bien vers zéro. D'après l'expression de η_1, le premier reste est d'ordre

$$|R_1| \leq \frac{q\sqrt{r_0}}{(\sigma-1)\sqrt{m}r_0} m^{\varepsilon} .$$

Comme $\sqrt{m}r_0$ est dominant devant $m^{2\varepsilon}$ par hypothèse, ce reste tend bien vers 0. Pour le second, nous avons $\psi\big(\rho_0(z_1)\big)$ est supérieur à une constante multipliée par r_0, donc

$$|R_2| \leq \frac{q}{\sqrt{r_0}} m^{2\varepsilon} .$$

D'après l'hypothes sur le régime asymptotique (5.3), cette quantité tend effectivement vers 0. Nous écrivons alors

$$||\nu_0 = \int_{|z_0-\rho_0|\leq m^{1/2+\varepsilon}} \nu_0(\mathrm{d}z_0) \int_{|z_1-\rho_1(z_0)|\leq\eta_1} ||\nu_0^{\rho_1(z_0)} \nu_1^{z_0}(\mathrm{d}z_1) .$$

Comme l'intégrande ne dépend plus de z_1, et que la masse de la mesure $\nu_1^{z_0}$ est concentrée autour de $\rho_1(z_0)$, nous avons

$$\int_{|z_1-\rho_1(z_0)|\leq\eta_1} ||\nu_0^{\rho_1(z_0)} \nu_1^{z_0}(\mathrm{d}z_1) = ||\nu_0^{\rho_1(z_0)} \big(1-R\big) ,$$

avec, d'après notre estimation (5.11),

$$R \leq \exp(-m^{\varepsilon}) .$$

Nous pouvons donc simplifier l'intégrale en écrivant

$$||\nu_0 = \int_{|z_0-\rho_0|\leq\eta_0} \nu_0(\mathrm{d}z_0)\, ||\nu_0^{\rho_1(z_0)} .$$

Nous devons donc maintenant nous occuper de $||\nu_0^{\rho_1(z_0)}$. Nous allons utiliser le même lemme que pour la mesure précédente avec les fonctions

$$\psi_1(x) = \sigma M_{00} x + f\big(M_{20}\big) ,$$

et

$$\psi_2(z) = \rho_1(z_0) M_{10} .$$

Avec les mêmes manipulations que précédemment, nous arrivons à

$$||\nu_0^{\rho_1(z_0)} = ||\nu_0^{\rho_1(\rho_0)} \big(1+R_1\big) e^{R_2} ,$$

avec les termes de reste qui sont du même ordre :

$$|R_1| \leq \frac{q\sqrt{r_0}}{(\sigma-1)\sqrt{m}r_0} m^{\varepsilon} .$$

et

$$|R_2| \leq \frac{q}{\sqrt{r_0}} m^{2\varepsilon} .$$

Par ce qui précède, ces restes tendent vers 0 et nous obtenons

$$\nu_0(k) = \frac{\rho_0\Big(\rho_1\big(\rho_0\big)\Big)}{\sqrt{m}\frac{k}{m}} \exp\Big(mF_0^{\rho_1}(k/m) - mF_0^{\rho_1}(\rho_0)\Big)\,,$$

où nous avons simplement noté

$$\rho_1 = \rho_1\big(\rho_0\big)\,.$$

Nous retrouvons les points du chapitre 2 en prenant le maximum de la fonction F. Arrêtons-nous un instant pour voir une analogie avec le chapitre 2. Nous avions un premier encadrement de la mesure invariante de la proportion de master sequences. Nous en avons déduit un encadrement sur la mesure invariante de la proportion d'individus dans la première classe. Forts de ce nouvel encadrement, nous avons pu revenir sur la mesure de la proportion de master sequences avec un encadrement meilleur que le premier. Comme nous l'avons fait dans le chapitre 2, nous allons pouvoir répéter le procédé une infinité de fois, et nous intéresser aux limites des processus minorants et majorants.

6 Itérer le procédé

Jusque maintenant, nous avons estimé grossièrement la proportion de master sequences. A partir de cette estimation, nous avons montré comment nous aurions pu obtenir une estimation sur la proportion d'individus dans la classe 1. Enfin, nous avons montré comment en déduire une meilleure approximation sur la proportion de master sequences. Nous allons comment nous pourrions appliquer à nouveau ce procédé à partir de cette meilleure estimation pour obtenir une estimation encore plus précise. Nous pourrions appliquer ce procédé une infinité de fois et nous intéresser à la limite de ces processus. Adoptons de nouvelles notations pour cette section uniquement. Appelons $\nu_0^{(0)}$ la première mesure invariante de la proportion de master sequence que nous avons obtenue. Nous entendons par cette notation les deux mesures invariantes, $\underline{\nu_0}^{(0)}$ pour la borne supérieure et $\overline{\nu_0}^{(0)}$ pour la borne inférieure. A partir de cette mesure, nous avons tiré une estimation sur la mesure invariante de la proportion d'individus dans la classe 1, que nous notons $\nu_1^{(0)}$. Nous en avons déduit la mesure $||\nu_0$ que nous noterons dans cette section $\nu_0^{(1)}$. Comme décrit précédemment, nous allons donc construire une mesure $\nu_1^{(1)}$, puis $\nu_0^{(2)}$, etc ...

Supposons que nous avons construit le processus $Z_0^{(n)}$, pour un certain n qui estime la proportion de master séquence avec une mesure invariante centrée autour de ${\rho_0}^{(n)}$ avec une fenêtre de η_0. Nous construisons alors le processus $Z_1^{(n)}$ à partir des fonctions

$$\psi_1^{z_0}(x) = \sigma z_0 M_{01} + f\big(M_{21}\big) + M_{11}x\,,$$

et

$$\phi_1^{z_0}(x) = 1 - f\big(M_{21}\big) + L_q^1 z_0 - M_{11}x\,.$$

Le processus issu de ces fonctions atteint son maximum au point $\rho_1(z_0)$ vérifiant

$$\rho_1(z_0) = \frac{\sigma M_{01}\, z_0 + f\big(M_{21}\big)}{1 - M_{11} + (\sigma - 1)z_0}\,.$$

La mesure invariante de ce processus serait reliée à l'intégrale contre la mesure $\nu_0^{(n)}$

$$\nu_1^{(n)} = \int_{z_0} \nu_1^{z_0}\, \nu_0^{(n)}(\mathrm{d}z_0)\,.$$

Une fois de plus, nous allons pouvoir restreindre cette intégrale au voisinage de $\rho_0^{(n)}$ grâce à notre estimation (5.11) :

$$\nu_1^{(n)} = \int_{|z_0-\rho_0^{(n)}|\leq\eta_0} \nu_1^{z_0}\, \nu_0^{(n)}(\mathrm{d}z_0)\,.$$

Nous pouvons alors construire le processus $Z_0^{(n+1)}$ à partir des fonctions

$$\psi_0^{z_1}(x) = \sigma M_{00}\, x + z_1 M_{10} + f(M_{20})\,,$$

et

$$\phi_0^{z_1}(x) = 1 + L_q\, x - M_{10} z_1 - f(M_{20})\,.$$

Le point critique $\rho_0^{(n+1)}$ de ce processus est la solution positive du trinôme

$$-(\sigma-1)X^2 + r_0 X + M_{10} z_1 + f(M_{20}) = 0$$

Introduisons à nouveau les notations qui nous ont servies dans le chapitre 2 : pour tout entier i, j, nous posons

$$\overline{D_{ij}} = \frac{M_{ij}}{1 - M_{jj} + (\sigma-1)(\rho_0^{(0)} - \eta_0)}\,,$$

et

$$\underline{D_{ij}} = \frac{M_{ij}}{1 - M_{jj} + (\sigma-1)(\rho_0^{(0)} + \eta_0)}\,.$$

Comme dans toutes notations précédentes, nous noterons D_{ij} et intérprêterons les formules pour les deux bornes à la fois. Nous remplaçons les fonctions ψ et ϕ par leur expression et nous obtenons que le point critique est solution de l'équation

$$-(\sigma-1)X^2 + (\sigma M_{00} - 1)X + \sigma D_{01}\, M_{10}\rho_0^{(n)} + f\big(D_{21}M_{10}\big) + f(M_{20}) = 0 \quad (5.69)$$

La fonction f est continue et croissante, la suite u_n est donc monotone, elle est aussi bornée, elle converge donc vers la solution positive du trinôme

$$-(\sigma-1)X^2 + \Big(\sigma M_{00} - 1 + \sigma D_{01}\, M_{10}\Big)X + f\big(D_{10}M_{21}\big) + f(M_{20}) = 0 \quad (5.70)$$

Nous utilisons à nouveau notre lemme 6 pour estimer la racine positive de cette équation et nous obtenons Comme r_0 est très grand devant q, les premiers termes de $\overline{\rho}_0^{(\infty)}$ sont donc

$$\rho_0^{(\infty)} = r_0 + \frac{\sigma D_{01}\, M_{10}}{(\sigma-1)(1-M_{11})} + O\Big(\frac{q^2}{r_0+q}\Big) \quad (5.71)$$

Nous retrouvons ici le même développement (2.26) obtenu au chapitre 2.

7 Encore plus d'étapes

Rédigeons la récurrence, supposons que le développement de ρ_0 soit connu jusqu'à l'ordre q^k, et expliquons comment trouver le terme d'ordre q^{k+1}. Nous avons une estimation sur la mesure invariante du nombre de master sequences, nous allons utiliser cette estimation pour construire un processus sur le nombre de classes 1, puis le nombre de classes 2, et ceci jusqu'au nombre de classe L. Nous pourrons ensuite utiliser cette estimée du nombre de classe L pour estimer le nombre de classes $L-1$, puis le nombre de classe $L-2$, etc jusqu'à revenir au nombre de master sequence. Nous utilisons l'estimation précédente sur le nombre de master sequences. Rédigeons la partie descendante. Il s'agit des processus qui prennent en compte seulement les classes qui leur sont inférieures. Nous suivons les individus de la classe K, nous supposons que le nombre d'individus des classes $0, 1, \cdots, K-1$ est connu et fixé, nous les notons $z_0, \cdots, z_{K-1}$. Nous connaissons de plus les mesures invariantes associées à ces processus : le nombre d'individus de la classe J est concentré autour de $[m(\rho_J - \eta_J),\, m(\rho_J + \eta_J)]$, nous ne pourrons sans doutes pas calculer ρ_J. η_J vaut $m^{-1/2+\varepsilon}$ si J vaut 0 et $\sqrt{r_0/m}m^{\varepsilon}$ sinon.

Considérons δ_k (il y a donc dans la population k individus de classe K) :

$$\delta_k = P\left(\begin{array}{c}\text{Remplacer un}\\ \text{individu non } K\end{array}\right) P\left(\begin{array}{c}\text{Enfant est}\\ \text{de type } K\end{array}\right).$$

Nous conditionnons alors selon la classe du parent.

$$\begin{aligned}\delta_k = \left(1-\frac{k}{m}\right)\Bigg[&\sum_{J<K} P\left(\begin{array}{c}\text{Enfant est}\\ \text{de type } K\end{array} \text{ et } \begin{array}{c}\text{son parent est}\\ \text{de type } J\end{array}\right)\\ &+ P\left(\begin{array}{c}\text{Enfant est}\\ \text{de type } K\end{array} \text{ et } \begin{array}{c}\text{son parent est}\\ \text{de type } K\end{array}\right)\\ &+ P\left(\begin{array}{c}\text{Enfant est}\\ \text{de type } K\end{array} \text{ et } \begin{array}{c}\text{son parent est}\\ \text{de type } J>K\end{array}\right)\Bigg]. \quad (5.72)\end{aligned}$$

Une fois encore, les deux premières probabilités dépendent de quantités que l'on connait, et il nous suffit d'encadrer la dernière pour pouvoir calculer une mesure invariante du processus. Cette probabilité est positive, ce qui nous donne une borne inférieure. Cela revient à négliger les retours des classes $K+1$ et plus. Pour la borne supérieure, nous faisons comme si le génôme du parent était de classe $K+1$:

$$P\left(\begin{array}{c}\text{Enfant est}\\ \text{de type } K\end{array} \text{ et } \begin{array}{c}\text{son parent est}\\ \text{de type } J>K\end{array}\right) \leq \frac{M_{K+1,K}}{\text{fit}(z_0)}.$$

De même, pour la probabilité γ_k :

$$\gamma_k = \frac{k}{m}\left[\sum_{J<K} P\left(\begin{array}{c}\text{Enfant est}\\ \text{de type non } K\end{array}\text{ et }\begin{array}{c}\text{son parent est}\\ \text{de type } J\end{array}\right) + P\left(\begin{array}{c}\text{Enfant est}\\ \text{de type } K\end{array}\text{ et }\begin{array}{c}\text{son parent est}\\ \text{de type } K\end{array}\right) + P\left(\begin{array}{c}\text{Enfant est}\\ \text{de type non } K\end{array}\text{ et }\begin{array}{c}\text{son parent est}\\ \text{de type } J>K\end{array}\right)\right]. \quad (5.73)$$

Encadrer la dernière probabilité pour obtenir un processus qui ne dépend pas de toute la population. Cette probabilité est inférieure si nous n'imposons aucune condition sur l'enfant :

$$P\left(\begin{array}{c}\text{Enfant est}\\ \text{de type non } K\end{array}\text{ et }\begin{array}{c}\text{son parent est}\\ \text{de type } J>K\end{array}\right) \leq \frac{1-\sum_{j=0}^{K-1} z_j - x}{\text{fit}(z_0)}$$

Pour la borne inférieure, nous avons

$$P\left(\begin{array}{c}\text{Enfant est}\\ \text{de type non } K\end{array}\text{ et }\begin{array}{c}\text{son parent est}\\ \text{de type } J>K\end{array}\right) \geq \frac{1-\sum_{j=0}^{K-1} z_j - \frac{k}{m}}{\text{fit}(z_0)}\left(1-M_{K+1,K}\right) \geq \frac{1-\sum_{j=0}^{K-1} z_j - \frac{k}{m} - M_{K+1,K}}{\text{fit}(z_0)}.$$

Tout d'abord, nous descendons : nous construisons les processus permettant de décrire l'évolution du nombre d'individus dans la classe K. Appelons cette chaîne $\left|Z_K^{z_0,\cdots,z_{K-1}}\right.$. Introduisons les notations habituelles :

$$\psi(x) = \sigma M_{0K} z_0 + \sum_{j=1}^{K-1} z_j M_{jK} + x M_{KK} + f(M_{K+1,K}), \quad (5.74)$$

et

$$\phi(x) = 1 + (\sigma-1)z_0 - \sigma M_{0K} z_0 - \sum_{j=1}^{K-1} z_j M_{jK} - x M_{KK} - f(M_{K+1,K}). \quad (5.75)$$

Afin de ne pas avoir à écrire toutes les dépendances à chaque étape, nous regroupons les quantités $z_0, \cdots, z_{K-1}$ dans le vecteur t_K. La fonction $F_K^{z_0,\cdots,z_{K-1}}$ sera simplement noté $F_K^{t_K}$, elle atteint son maximum au point $\rho_K(t_K)$. Ce point vérifie

$$\rho_K(t_K) = \frac{\sigma M_{0K} z_0 + \sum_{j=1}^{K-1} z_j M_{jK} + f(M_{K+1,K})}{1 - M_{KK} + (\sigma-1)z_0} \quad (5.76)$$

Nous avons aussi une mesure invariante du processus en suivant le théorème 21 :

$$|\nu_K^{t_K} = \sqrt{\frac{\rho_K^*}{m}} \frac{\exp\left(m|F_K^{t_K}(\frac{k}{m}) - m|F_K^{t_K}(\rho_K)\right)}{\frac{k}{m}}$$

Pour obtenir la mesure sans les dépendances, il faudrait ensuite intégrer contre les mesures des processus $Z_0, \cdots, Z_{K-1}$. Rappelons cependant qu'il reste à montrer comment relier cette mesure définie par une intégrale à la mesure invariante du processus.

$$|\nu_K(k) = \int_{z_0} \cdots \int_{z_{K-1}} |\nu_K^{t_K}|\nu_0(\mathrm{d}z_0) \cdots |\nu_{K-1}(\mathrm{d}z_{K-1})\,. \tag{5.77}$$

De la même manière que nous n'avons pas pu calculer l'intégrale de la mesure ν_1 à la section 4, nous ne pourrons pas calculer ces intégrales avec la méthode que nous avons décrite.

7.1 En redescendant

Nous allons maintenant construire une deuxième approximation de tous les processus précédents. En partant de $z_0, \cdots, z_L$ approchés par les processus précédents, nous construisons un processus permettant d'estimer le nombre d'individus dans la classe $L-1$, avec cette nouvelle estimation, nous construirons un processus qui approche le nombre d'individus de la classe $L-2$ etc .. Nous rédigeons seulement l'étape de récurrence. Nous supposons maintenant que le nombre d'individus des classes $0, \cdots, L$ est connu et fixé noté $z_0, \cdots, z_L$. Nous connaissons de plus les mesures invariantes associées à ces processus : le nombre d'individus de la classe J est concentré autour de $[m(\rho_J - \eta_J),\, m(\rho_J + \eta_J)]$, nous ne pourrons sans doutes pas calculer ρ_J. η_J vaut $m^{-1/2+\varepsilon}$ si J vaut 0 et $\sqrt{r_0/m}m^{\varepsilon}$ sinon. Considérons δ_k, il y a donc dans la population k individus de classe K.

$$\delta_k = P\begin{pmatrix} \text{Remplacer un} \\ \text{individu non } K \end{pmatrix} P\begin{pmatrix} \text{Enfant est} \\ \text{de type } K \end{pmatrix}.$$

Nous conditionnons alors selon la classe du parent.

$$\delta_k = \left(1 - \frac{k}{m}\right)\left[\sum_{J \le L} P\begin{pmatrix} \text{Enfant est} \\ \text{de type } K \end{pmatrix} \text{et} \begin{pmatrix} \text{son parent est} \\ \text{de type } J \end{pmatrix} + P\begin{pmatrix} \text{Enfant est} \\ \text{de type } K \end{pmatrix} \text{et} \begin{pmatrix} \text{son parent est} \\ \text{de type } J > L \end{pmatrix}\right]. \tag{5.78}$$

Encore une fois, toutes les probabilités de cette expression dépendent de quantités que nous connaissons, il suffit donc d'encadrer la dernière probabilité pour obtenir un processus qui ne dépend pas de la répartition de la population. En la minorant par 0, ce qui revient à supprimer les retours des classes $L+1$ et plus, nous obtenons une borne inférieure. Pour la borne supérieure, nous pouvons majorer par

$$P\begin{pmatrix} \text{Enfant est} \\ \text{de type } K \end{pmatrix} \text{et} \begin{pmatrix} \text{son parent est} \\ \text{de type } J > L \end{pmatrix} \le \frac{M_{L+1,K}}{\mathrm{fit}(z_0)}\,.$$

De même, pour la probabilité γ_k :

$$\gamma_k = \frac{k}{m}\left[\sum_{J\leq L} P\begin{pmatrix}\text{Enfant est}\\ \text{de type non } K\end{pmatrix}\text{ et }\begin{pmatrix}\text{son parent est}\\ \text{de type } J\end{pmatrix}\right.$$
$$\left. + P\begin{pmatrix}\text{Enfant est}\\ \text{de type non } K\end{pmatrix}\text{ et }\begin{pmatrix}\text{son parent est}\\ \text{de type } J > L\end{pmatrix}\right]. \quad (5.79)$$

Encadrer la dernière probabilité pour obtenir un processus qui ne dépend pas de toute la population. Nous commençons par la majorer par

$$P\begin{pmatrix}\text{Enfant est}\\ \text{de type non } K\end{pmatrix}\text{ et }\begin{pmatrix}\text{son parent est}\\ \text{de type } J > L\end{pmatrix} \leq \frac{1-\sum_{\substack{J=0\\J\neq K}}^{L} z_j - x}{\text{fit}(z_0)}\,.$$

Nous pouvons aussi la minorer par

$$P\begin{pmatrix}\text{Enfant est}\\ \text{de type non } K\end{pmatrix}\text{ et }\begin{pmatrix}\text{son parent est}\\ \text{de type } J > L\end{pmatrix} \geq \frac{1-\sum_{\substack{J=0\\J\neq K}}^{L} z_j - x - M_{L+1,K}}{\text{fit}(z_0)}\,.$$

Nous rassemblons les deux sommes et nous définissons donc les fonctions

$$\psi(x) = \sigma M_{0K} z_0 + \sum_{\substack{J=0\\J\neq K}}^{L} z_J M_{JK} + x M_{KK} + f(M_{L+1,K})\,,$$

et

$$\phi(x) = 1 + (\sigma-1)z_0 - \sigma M_{0K} z_0 - \sum_{\substack{J=0\\J\neq K}}^{L} z_J M_{JK} - x M_{KK} - f(M_{L+1,K})\,.$$

Notons encore t_L la collection des variables $z_0, \cdots, z_L$, cette fois, pas besoin de z_K. La fonction $F_K^{t_L}$ atteint son maximum au point $|\rho_K(t_L)$ qui s'écrit comme

$$|\rho_K(t_L) = \frac{\sigma M_{0K} z_0 + \sum_{j=1, j\neq L}^{K} z_j M_{jK} + f(M_{L+1,K})}{1 - M_{KK} + (\sigma-1)z_0\,.} \quad (5.80)$$

Nous avons aussi une mesure invariante du processus

$$||\nu_K^{t_L} = \sqrt{\frac{|\rho_K^*(t_L)}{m}} \frac{\exp\left(mF_K^{t_L}(\frac{k}{m}) - mF_K^{t_L}\left(|\rho_K(t_L)\right)\right)}{\frac{k}{m}}\,.$$

Il suffirait ensuite d'intégrer cette mesure contre les mesures des processus qui décrivent les autres classes

$$|\nu_K(k) = \int_{z_0} \cdots_K \int_{z_L} |\nu_K^{t_L}| \nu_0(z_0) \cdots_K |\nu_L(z_L)\,. \quad (5.81)$$

7.2 L'intégrale contre les mesures

Toutes les fonctions précédentes dépendent du nombre d'individus dans chaque classe $z_0, \cdots, z_L$. A cette étape, il faudrait donc intégrer contre les mesures de ces processus et majorer les termes qui sont loin des ρ, pour ν_0, il faut une fenêtre de sommation $\sqrt{m}$ pour les autres, $\sqrt{mr_0}$.

$$\nu_K = \int_{z_0} \cdots \int_{z_L} \nu_K^{z_0,\cdots,z_L} \nu_0(\mathrm{d}z_0) \cdots \nu_L(\mathrm{d}z_L)\,.$$

Nous allons montrer que le terme principal de cette intégration viendra des termes où tous les z_J sont proches des points critiques ρ_J, pour ce faire, nous majorons simplement par 1 la mesure que l'on intègre dans les intégrales où l'un des z_J est loin de $m\rho_J$. Nous nous ramenons ainsi à

$$\nu_K = \int_{|z_0-\rho_0|\leq\eta_0} \cdots \int_{|z_L-\rho_L|\leq\eta_L} \nu_K^{z_0,\cdots,z_L} \nu_0(\mathrm{d}z_0) \cdots \nu_L(\mathrm{d}z_L)\,.$$

nous nous plaçons donc dans le cas où tous les nombres $z_0, \cdots, z_L$ apparaissant dans la mesure ν_K sont proches de leur $m\rho_K$.

8 Retour aux master sequences

Maintenant que nous avons une idée du nombre d'individus dans chaque classe, nous pouvons nous intéresser au nombre de master sequences. Nos fonctions ψ et ϕ s'écrivent

$$\psi(x) = \sigma M_{00} x + \sum_{J=1}^{L} z_J M_{J0} + f(M_{L+1,0})\,,$$

et

$$\phi(x) = 1 + L_q x - \sum_{J=1}^{L} z_J M_{J0} - f(M_{L+1,0})\,.$$

8.1 Intégrale contre les mesures

Une fois que nous avons une estimation sur le nombre d'individus dans les classes jusqu'à L, nous intégrons la mesure du processus des master sequences

$$||\nu_0 = \int_{z_1,\cdots,z_L} ||\nu_0^{z_1,\cdots,z_L}\, \nu_1(\mathrm{d}z_1) \cdots \nu_L(\mathrm{d}z_L)\,.$$

Rappelons une dernière fois qu'il manque un argument pour relier cette mesure à la mesure invariante du processus du nombre de master sequences.

9 Une piste pour le temps de disparition

Si nous arrivons à relier nos mesures définies par des intégrales aux mesures invariantes des processus, alors nous aurons un moyen pour estimer le temps de

disparition. Il suffira de considérer l'inverse de la masse du singleton $\{0\}$ de la mesure invariante pour le processus du nombre de master sequences. Nous espérons pouvoir effectuer ce développement dans un travail futur.

Bibliographie

[ADL04] Jon P. Anderson, Richard Daifuku et Lawrence A. Loeb. « Viral Error Catastrophe by Mutagenic Nucleosides ». en. *Annual Review of Microbiology* 58.1 (oct. 2004), p. 183-205.

[AF02] David Aldous et Jim Fill. *Reversible Markov chains and random walks on graphs*. 2002.

[AF10] Julia Alonso et Hugo Fort. « Error catastrophe for viruses infecting cells : analysis of the phase transition in terms of error classes ». *Philosophical Transactions : Mathematical, Physical and Engineering Sciences* 368.1933 (2010), p. 5569-5582.

[AF96] Domingos Alves et Jose Fernando Fontanari. « Population genetics approach to the quasispecies model ». *Physical Review E* 54.4 (1996), p. 4048.

[AF97] Domingos Alves et JF Fontanari. « Error threshold in the evolution of diploid organisms ». *Journal of Physics A : Mathematical and General* 30.8 (1997), p. 2601.

[Ast+13] Elizabeth Aston et al. « Critical mutation rate has an exponential dependence on population size in haploid and diploid populations ». *PLoS One* 8.12 (2013), e83438.

[BB11] F. Bagnoli et M. Bezzi. « Eigen's Error Threshold and Mutational Meltdown in a Quasispecies Model ». en. *International Journal of Modern Physics C* (nov. 2011).

[BG00] Ellen Baake et Wilfried Gabriel. « Biological evolution through mutation, selection, and drift : An introductory review ». *Annual Reviews of Computational Physics* 7 (2000), p. 203-264.

[BG07] Ellen Baake et Hans-Otto Georgii. « Mutation, selection, and ancestry in branching models : a variational approach ». *Journal of mathematical biology* 54.2 (2007), p. 257-303.

[BK83] Chikafusa Bessho et Naoki Kuroda. « A note on a more general solution of Eigen's rate equation for selection ». *Bulletin of mathematical biology* 45.1 (1983), p. 143-149.

[BK98] D Bonnaz et A J Koch. « Stochastic Model of Evolving Populations ». en. *Journal of Physics A : Mathematical and General* 31.2 (jan. 1998), p. 417-429.

[BMS06] Yisroel Brumer, Franziska Michor et Eugene I. Shakhnovich. « Genetic instability and the quasispecies model ». *Journal of Theoretical Biology* 241.2 (2006), p. 216-222.

[BNS19] Alexander S Bratus, Artem S Novozhilov et Yuri S Semenov. « Rigorous mathematical analysis of the quasispecies model : From Manfred Eigen to the recent developments ». *Advanced Mathematical Methods in Biosciences and Applications* (2019), p. 27-51.

[Bür98] Reinhard Bürger. « Mathematical properties of mutation-selection models ». *Genetica* 102 (1998), p. 279-298.

[Cad16] Yannis Cadieu. « La théorie des quasi-espèces : concepts, application à la dynamique des populations de virus a ARN, implications biologiques et limites ». Thèse de doct. 2016.

[CCA01] Shane Crotty, Craig E. Cameron et Raul Andino. « RNA Virus Error Catastrophe : Direct Molecular Test by Using Ribavirin ». en. *Proceedings of the National Academy of Sciences* 98.12 (juin 2001), p. 6895-6900.

[CD16a] Raphaël Cerf et Joseba Dalmau. « Quasispecies on class-dependent fitness landscapes ». *Bulletin of mathematical biology* 78.6 (2016), p. 1238-1258.

[CD16b] Raphaël Cerf et Joseba Dalmau. « The distribution of the quasispecies for a Moran model on the sharp peak landscape ». *Stochastic Processes and their Applications* 126.6 (2016), p. 1681-1709.

[CD18] Raphaël Cerf et Joseba Dalmau. « The Quasispecies for the Wright–Fisher Model ». *Evolutionary Biology* 45.3 (sept. 2018), p. 318-323.

[Cer15] Raphaël Cerf. *Critical Population and Error Threshold on the Sharp Peak Landscape for a Moran Model*. en. T. 233. Memoirs of the American Mathematical Society. American Mathematical Society, jan. 2015.

[CF98] P. R. A. Campos et J. F. Fontanari. « Finite-Size Scaling of the Quasispecies Model ». en. *Physical Review E* 58.2 (août 1998), p. 2664-2667.

[CF99] P R A Campos et J F Fontanari. « Finite-Size Scaling of the Error Threshold Transition in Finite Populations ». en. *Journal of Physics A : Mathematical and General* 32.1 (jan. 1999), p. L1-L7.

[CK65] James F. Crow et Motoo Kimura. « Evolution in Sexual and Asexual Populations ». *The American Naturalist* 99.909 (1965), p. 439-450.

[CK70] J. F. Crow et M. Kimura. « An Introduction to Population Genetics Theory. » English. *An introduction to population genetics theory.* (1970).

[CN12] Irene A Chen et Martin A Nowak. « From prelife to life : How chemical kinetics become evolutionary dynamics ». *Accounts of chemical research* 45.12 (2012), p. 2088-2096.

[Dal14] Joseba Dalmau. « Convergence of a Moran Model to Eigen's Quasispecies Model ». *arXiv :1404.2133 [math, q-bio]* (avr. 2014).

[Dal16] Joseba Dalmau. « La distribution de la quasi-espèce pour une population finie ». Thèse de doct. Université Paris-Saclay, 2016.

[Dal18] Joseba Dalmau. « Asymptotic behavior of Eigen's quasispecies model ». *Bulletin of mathematical biology* 80.7 (2018), p. 1689-1712.

[DG12] Persi Diaconis et Robert Griffiths. « exchangeable pairs of bernoulli random variables, krawtchouck polynomials, and ehrenfest urns ». *Australian & New Zealand Journal of Statistics* 54.1 (2012), p. 81-101.

[Dom+78] Esteban Domingo et al. « Nucleotide Sequence Heterogeneity of an RNA Phage Population ». *Cell* 13.4 (avr. 1978), p. 735-744.

[Dom02] Esteban Domingo. « Quasispecies Theory in Virology ». en. *Journal of Virology* 76.1 (jan. 2002), p. 463-465.

[DS84] Peter G Doyle et J Laurie Snell. *Random walks and electric networks*. T. 22. American Mathematical Soc., 1984.

[DSP12] Esteban Domingo, Julie Sheldon et Celia Perales. « Viral quasispecies evolution ». *Microbiology and Molecular Biology Reviews* 76.2 (2012), p. 159-216.

[DSS85] Lloyd Demetrius, Peter Schuster et Karl Sigmund. « Polynucleotide evolution and branching processes ». *Bulletin of mathematical biology* 47.2 (1985), p. 239-262.

[DSV12] Narendra M Dixit, Piyush Srivastava et Nisheeth K Vishnoi. « A finite population model of molecular evolution : Theory and computation ». *Journal of Computational Biology* 19.10 (2012), p. 1176-1202.

[EHM99] Ágoston E Eiben, Robert Hinterding et Zbigniew Michalewicz. « Parameter control in evolutionary algorithms ». *IEEE Transactions on evolutionary computation* 3.2 (1999), p. 124-141.

[Eig71] Manfred Eigen. « Selforganization of Matter and the Evolution of Biological Macromolecules ». en. *Die Naturwissenschaften* 58.10 (1971), p. 465-523.

[EMS07] Manfred Eigen, John McCaskill et Peter Schuster. « The Molecular Quasi-Species ». en. *Advances in Chemical Physics*. Sous la dir. d'I. Prigogine et Stuart A. Rice. Hoboken, NJ, USA : John Wiley & Sons, Inc., 2007, p. 149-263.

[EMS88] M. Eigen, J. McCaskill et P. Schuster. « The Molecular Quasispecies. » *J. Phys. Chem.* 92 (1988), p. 6881-6891.

[ES78] Manfred Eigen et Peter Schuster. « The hypercycle ». *Naturwissenschaften* 65.1 (1978), p. 7-41.

[Fel74] Joseph Felsenstein. « The Evolutionary Advantage of Recombination ». *Genetics* 78.2 (oct. 1974), p. 737-756.

[Fis58] Ronald Aylmer Fisher. *The genetical theory of natural selection*. 1958.

[FOC96] Marcus W Feldman, Sarah P Otto et Freddy B Christiansen. « Population genetic perspectives on the evolution of recombination ». *Annual review of genetics* 30.1 (1996), p. 261-295.

[FP97] Silvio Franz et Luca Peliti. « Error threshold in simple landscapes ». *Journal of Physics A : Mathematical and General* 30.13 (1997), p. 4481.

[Fur97] Hiroshi Furutani. « A Method for Estimating Mean First-Passage Time in Genetic Algorithms ». en. *Complex Systems* (1997), p. 24.

[FZB11] Jozef Fekiač, Ivan Zelinka et Juan C Burguillo. « A review of methods for encoding neural network topologies in evolutionary computation » (2011), p. 410-416.

[Gal97] Stefano Galluccio. « Exact solution of the quasispecies model in a sharply peaked fitness landscape ». *Physical Review E* 56.4 (1997), p. 4526.

[Gil83] John H Gillespie. « A simple stochastic gene substitution model ». *Theoretical population biology* 23.2 (1983), p. 202-215.

[Gòm+99] J Gòmez et al. « Hepatitis C viral quasispecies ». *Journal of viral hepatitis* 6.1 (1999), p. 3-16.

[Hal27] John Burdon Sanderson Haldane. « A mathematical theory of natural and artificial selection, part V : selection and mutation ». *Mathematical Proceedings of the Cambridge Philosophical Society*. T. 23. 7. Cambridge University Press. 1927, p. 838-844.

[Hes54] F. G. Hess. « Alternative Solution to the Ehrenfest Problem ». *The American Mathematical Monthly* 61.5 (1954), p. 323-328.

[HF15] Gregory R Hart et Andrew L Ferguson. « Error catastrophe and phase transition in the empirical fitness landscape of HIV ». *Physical Review E* 91.3 (2015), p. 032705.

[Hig95] Paul G. Higgs. « Frequency Distributions in Population Genetics Parallel Those in Statistical Physics ». en. *Physical Review E* 51.1 (jan. 1995), p. 95-101.

[HM+14] Peter Hegarty, Anders Martinsson et al. « On the existence of accessible paths in various models of fitness landscapes ». *Annals of Applied Probability* 24.4 (2014), p. 1375-1395.

[HM91] Jürgen Hesser et Reinhard Männer. « Towards an Optimal Mutation Probability for Genetic Algorithms ». en. *Parallel Problem Solving from Nature*. Sous la dir. d'Hans-Paul Schwefel et Reinhard Männer. T. 496. Berlin/Heidelberg : Springer-Verlag, 1991, p. 23-32.

[JER76] Billy L Jones, Richard H Enns et Sadanand S Rangnekar. « On the theory of selection of coupled macromolecular systems ». *Bulletin of Mathematical Biology* 38.1 (1976), p. 15-28.

[JN06] Martin Nilsson Jacobi et Mats Nordahl. « Quasispecies and recombination ». *Theoretical population biology* 70.4 (2006), p. 479-485.

[Jon77] Billy L Jones. « Analysis of Eigen's equations for selection of biological molecules with fluctuating mutation rates ». *Bulletin of mathematical biology* 39.3 (1977), p. 311-316.

[Kac47] Mark Kac. « Random Walk and the Theory of Brownian Motion ». *American Mathematical Monthly* 54.7 (1947), p. 369-391.

[KS02] Michal Kolář et František Slanina. « How the quasispecies evolution depends on the topology of the genome space ». *Physica A : Statistical Mechanics and its Applications* 313.3-4 (2002), p. 549-568.

[KT81] Samuel Karlin et Howard E Taylor. *A second course in stochastic processes*. Elsevier, 1981.

[Leu86] Ira Leuthäusser. « An exact correspondence between Eigen's evolution model and a two-dimensional Ising system ». *The Journal of chemical physics* 84.3 (1986), p. 1884-1885.

[Luq03] Bartolo Luque. « An Introduction to Physical Theory of Molecular Evolution ». *Open Physics* 1.3 (2003), p. 516-555.

[Mar+12] Arturo Marín et al. « Characteristic time in quasispecies evolution ». *Journal of Theoretical Biology* 303 (2012), p. 25-32.

[MDL10] Susanna C. Manrubia, Esteban Domingo et Ester Lázaro. « Pathways to extinction : beyond the error threshold ». *Philosophical Transactions : Biological Sciences* 365.1548 (2010), p. 1943-1952.

[Mor58] P. A. P. Moran. « Random Processes in Genetics ». en. *Mathematical Proceedings of the Cambridge Philosophical Society* 54.01 (1958), p. 60.

[Nic93] Richard WD Nickalls. « A new approach to solving the cubic : Cardan's solution revealed ». *The Mathematical Gazette* 77.480 (1993), p. 354-359.

[Now02] Martin A Nowak. « From quasispecies to universal grammar ». *Zeitschrift für physikalische Chemie* 216.1 (2002), p. 5.

[Now06] Martin A Nowak. « Five rules for the evolution of cooperation ». *science* 314.5805 (2006), p. 1560-1563.

[NS02] Martin Nilsson et Nigel Snoad. « Quasispecies evolution on a fitness landscape with a fluctuating peak ». *Physical Review E* 65.3 (2002), p. 031901.

[NS89] M. Nowak et P. Schuster. « Error Thresholds of Replication in Finite Populations Mutation Frequencies and the Onset of Muller's Ratchet ». eng. *Journal of Theoretical Biology* 137.4 (avr. 1989), p. 375-395.

[OF10] Benedikt Obermayer et Erwin Frey. « Error thresholds for self- and cross-specific enzymatic replication ». *Journal of theoretical biology* 267.4 (2010), p. 653-662.

[OH97] Gabriela Ochoa et Inman Harvey. « Recombination and error thresholds in finite populations ». *Foundations of genetic algorithms* 5 (1997), p. 245-264.

[OHB] Gabriela Ochoa, Inman Harvey et Hilary Buxton. « Optimal Mutation Rates and Selection Pressure in Genetic Algorithm ». en (), p. 8.

[OHB99] Gabriela Ochoa, Inman Harvey et Hilary Buxton. « Error Thresholds and Their Relation to Optimal Mutation Rates ». en. *Advances in Artificial Life*. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 1999, p. 54-63.

[Pel96] Luca Peliti. « Fitness landscapes and evolution ». *Physics of biomaterials : Fluctuations, selfassembly and evolution*. Springer, 1996, p. 287-308.

[PMD10] Jeong-Man Park, Enrique Munoz et Michael W Deem. « Quasispecies theory for finite populations ». *Physical Review E* 81.1 (2010), p. 011902.

[Prü97] Adam Prügel-Bennett. « Modelling Evolving Populations ». *Journal of Theoretical Biology* 185.1 (mar. 1997), p. 81-95.

[Rob55] Herbert Robbins. « A Remark on Stirling's Formula ». *The American Mathematical Monthly* 62.1 (1955), p. 26-29.

[Saa+06] David B Saakian et al. « Quasispecies theory for multiple-peak fitness landscapes ». *Physical Review E* 73.4 (2006), p. 041913.

[SBN14] Yuri S Semenov, Alexander S Bratus et Artem S Novozhilov. « On the behavior of the leading eigenvalue of Eigen's evolutionary matrices ». *Mathematical biosciences* 258 (2014), p. 134-147.

[Sch97] Peter Schuster. « Genotypes with phenotypes : Adventures in an RNA toy world ». *Biophysical chemistry* 66.2-3 (1997), p. 75-110.

[SD04] Ricard V. Solé et Thomas S. Deisboeck. « An Error Catastrophe in Cancer ? » eng. *Journal of Theoretical Biology* 228.1 (mai 2004), p. 47-54.

[SDH12] David B. Saakian, Michael W. Deem et Chin Kun Hu. « Finite Population Size Effects in Quasispecies Models with Single-Peak Fitness Landscape ». *EPL (Europhysics Letters)* 98.1 (avr. 2012), p. 18001.

[SH06] David B. Saakian et Chin-Kun Hu. « Exact Solution of the Eigen Model with General Fitness Functions and Degradation Rates ». *Proceedings of the National Academy of Sciences of the United States of America* 103.13 (2006), p. 4935-4939.

[SS82] Jörg Swetina et Peter Schuster. « Self-replication with errors : A model for polvnucleotide replication ». *Biophysical chemistry* 16.4 (1982), p. 329-345.

[TH07] Nobuto Takeuchi et Paulien Hogeweg. « Error-threshold exists in fitness landscapes with lethal mutants ». *BMC Evolutionary Biology* 7.1 (2007), p. 1-12.

[TM74] Colin J Thompson et John L McBride. « On Eigen's theory of the self-organization of matter and the evolution of biological macromolecules ». *Mathematical biosciences* 21.1-2 (1974), p. 127-142.

[Tri+12] Kushal Tripathi et al. « Stochastic Simulations Suggest That HIV-1 Survives Close to Its Error Threshold ». *PLoS Computational Biology* 8.9 (sept. 2012).

[TS04] Emmanuel Tannenbaum et Eugene I Shakhnovich. « Solution of the quasispecies model for an arbitrary gene network ». *Physical Review E* 70.2 (2004), p. 021903.

[WAE10] Bartłomiej Waclaw, Rosalind J Allen et Martin R Evans. « Dynamical phase transition in a model for evolution with migration ». *Physical review letters* 105.26 (2010), p. 268101.

[Wie97] Thomas Wiehe. « Model Dependency of Error Thresholds : The Role of Fitness Functions and Contrasts between the Finite and Infinite Sites Models ». en. *Genetical Research* 69.2 (avr. 1997), p. 127-136.

[Wil05] Claus O. Wilke. « Quasispecies Theory in the Context of Population Genetics ». *BMC Evolutionary Biology* 5.1 (août 2005), p. 44.

[Wri49] Sewall Wright. « The genetical structure of populations ». *Annals of eugenics* 15.1 (1949), p. 323-354.

[Xia+07] Feng Xiao-Li et al. « Error thresholds in single-peak Gaussian distributed fitness landscapes ». *Communications in Theoretical Physics* 48.4 (2007), p. 763.

[Zha97] Yi-Cheng Zhang. « Quasispecies Evolution of Finite Populations ». en. *Physical Review E* 55.4 (avr. 1997), R3817-R3819.

RÉSUMÉ

Nous nous intéressons à l'évolution une population d'individus. Lorsqu'un individu se reproduit, le matériel génétique de son enfant n'est pas une copie parfaite de celui de son parent, certains gènes sont changés aléatoirement. Ces erreurs sont le premier ingrédient de l'évolution: ce sont les mutations. Le second ingrédient est la sélection: le nombre moyen d'enfants qu'engendre un individu dépend de son matériel génétique. Il s'ensuit une interaction entre les mutations qui tendent à découvrir des comportements nouveaux, et la sélection qui favorise les individus qui se reproduisent davantage. Nous étudions le modèle d'Eigen, un modèle mathématique qui rend compte de ce phénomène. Ce modèle fait apparaître une transition de phase, c'est-à-dire qu'il existe une valeur critique des paramètres qui sépare deux régimes distincts. Dans le premier, l'individu le plus adapté envahit la population, dans le second, la population est complètement aléatoire. Nous calculons le développement asymptotique de la proportion de l'individu le plus adapté à l'équilibre. Cela nous permet de prolonger le développement du paramètre critique. Nous nous posons ensuite les mêmes questions dans le modèle de Moran, qui est un modèle avec une population finie d'individus. Nous nous interrogeons sur la bonne définition du paramètre critique dans ce modèle et proposons plusieurs critères pour le caractériser.

MOTS CLÉS

Probabilités discrètes, Génétique, Chaîne de Markov

ABSTRACT

We are interested in the evolution of a population of individuals. When an individual reproduces, the genetic material of the offspring is not a perfect copy of its parent's, some genes are changed randomly. These errors are the first ingredient of evolution: mutation. The second ingredient is selection: the average number of children an individual produces depends on its genetic material. There is an interaction between mutations that tend to discover new behaviours, and selection that favours individuals that reproduce more. We study the Eigen model, a mathematical model that accounts for this phenomenon. This model shows a phase transition, which is a critical value of the parameters that separates two distinct regimes. In the first regime, the most adapted individual invades the population, in the second regime, the population is completely random. We calculate the asymptotic development of the proportion of the most adapted individual at equilibrium. This allows us to extend the development of the critical parameter. We then ask the same questions in the Moran model, which is a model with a finite population of individuals. We discuss the correct definition of the critical parameter in this model and propose several criteria to characterise it.

KEYWORDS

Discrete probability, Genetics, Markov chains